Statistical Methods for Weather-related Insurance Claims

Christian Rohrbeck, Dipl.-Math., M.Res.

Submitted for the degree of Doctor of Philosophy

at Lancaster University

March 2017

Abstract

Severe weather events, for instance, heavy rainfall, snow-melt or droughts, cause large losses of lives and money every year. Insurance companies offer some form of protection against such undesirable outcomes, and decision makers want to take precautions to prevent future catastrophes. Both, decision makers and insurance companies, are hence interested to understand which weather events induce a high risk. This information then allows the insurance companies to set premiums for their policies by predicting future losses. Further, the relationship between damages and weather is also important to assess the impact of climate change.

Several aspects have to be considered in the statistical modelling of this relationship. For instance, some regions in the world are more used to severe rainfall events than others and, hence, presumably less vulnerable to small amounts of rainfall than others. Spatial statistics provides a statistical framework which allows for a spatially varying relationship while accounting for certain similarities for areas which are geographically close. Further, damages, especially large losses, are rather rare and the statistical analysis is hence usually based on a low number of observations. Methods from extreme value theory consider the modelling of such events and may hence be beneficial.

This thesis aims to develop statistical models for the relationship between damages, in particular property insurance claims, and weather events, based on daily Norwegian insurance and weather data. To improve existing models, new methodology is introduced which allows for substantial flexibility of the statistical model. The risk induced by certain weather events is assumed to be spatially varying across Norway but with neighbouring regions exhibiting similar vulnerability. To account for certain non-linear effects, the class

of monotonic regression functions is considered. Specifically, this work is the first to define flexible dependence structures for such functions. In particular, the first approach considers a Bayesian framework and estimates are obtained by Markov chain Monte Carlo algorithms while the second approach is optimization-based. The last part of the thesis derives extreme value models for discrete data and estimates them in a Bayesian framework. In particular, a mixture model which allows for a flexible tail behaviour is motivated by an exploratory analysis of the highest claims in the data. Additionally, the data are restructured based on spatial and temporal patterns and then combined with the proposed extreme value mixture model. All these approaches, monotonic regression and extreme value analysis, lead to an improved model fit and a better understanding of the relationship between insurance claims and weather events.

Acknowledgements

There are many people I would like to thank for their support over the past years. First and foremost my supervisors, Deborah Costain, Jonathan Tawn, Emma Sherlock and Arnoldo Frigessi, for their time, guidance and tireless efforts throughout the project. You have been a constant source of encouragement and supported me wherever possible by, for instance, contributing so much time when you were busy, reading through my numerous drafts and by introducing me to many influential people in the statistical community. I'm very much looking forward to work with you on future projects. Thanks also to Ida Scheel and Ola Haug for helpful discussions and providing the data considered in this thesis.

I would also like to thank the students and staff of the STOR-i Centre for Doctoral Training for providing such a stimulating and enjoyable research environment. You have each contributed to this memorable experience via conversations, lunches, meeting groups, forums and innumerable fun activities such as dinners, pub quizzes, football, hikes, etc.

I am very grateful for the financial support provided by the EPSRC. Thanks also to Statistics for Innovation and the Norwegian Research Council for funding my four trips to Oslo, including my 3 months research visit in 2014.

Abschließend möchte ich noch meiner Familie und meinen Freunden für ihre Unterstützung danken. Im Besonderen danke ich meinen Eltern und Großeltern, ohne deren unermüdlichen Einsatz und großartige Unterstützung ich keinesfalls in der Lage gewesen wäre diesen Weg zu beschreiten.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of another degree.

Chapter 4 has been submitted to *Biometrika* as Rohrbeck, C., Costain D. and Frigessi, A. (2016). Bayesian Spatial Monotonic Multiple Regression.

Chapter 6 has been submitted to *Annals of Applied Statistics* as Rohrbeck, C., Eastoe E. F., Frigessi, A. and Tawn, J.A. (2016). Extreme Value Modelling of Water-related Insurance Claims.

Christian Rohrbeck

Contents

Chapter 1

Introduction

1.1 Motivation

Large parts of society and economy face uncertainties, such as illness, unemployment or commercial crises, in their decision processes. This led to the introduction of national insurance systems in several countries, including the United Kingdom and Germany, as well as the establishment of private insurers. Nowadays, the insurance industry is an important economical factor and insurance companies offer protection against a wide range of undesirable future events, for instance, car accidents, disability and property damages. An insurance company guarantees compensation up to a sum assured for a specified type of damage in return for a fixed monthly or yearly payment (insurance premium). These premiums are pooled together to compensate damages. Consequently, the risk of damages and losses is shared between insurance policy holders in the portfolio of the insurance company. This is the fundamental principle of insurance. However, there still exists the risk that the promised payouts exceed the capital assets of the insurance company, e.g. due to catastrophic events. Hence, reinsurance companies offer contracts to achieve a higher level of risk sharing and reinsurance agreements are an important risk management tool.

Quantification of the risk of an insured object or event is required in order to set an adequate insurance premium. Actuarial science is the discipline which applies mathematical and statistical methods to assess this risk in probabilistic terms. Properties of the probability distribution of the claims are often estimated based on historical claim data and used to produce insurance products. Several principles have been developed to derive the insurance premiums based on these properties.

In this thesis, weather is considered as a specific risk factor because weather events such as droughts or rainfall can cause severe damages and losses. Droughts and heatwaves can cause both crop shortfall and casualties due to heat exhaustion. Intense rainfall, on the other hand, may lead to localized flooding and destroy whole neighbourhood areas in the process. In order to take adequate precautions, e.g. by building flood defenses or reservoirs, accurate weather and climate models are required. Though weather forecasts have improved recently, there still remains high uncertainty on future weather events and on the future distribution of the weather, which we call climate.

The insurance industry offers two weather-related policies: property and weather index insurances. In order to estimate the risk distribution, accurate models for both the weather and the relationship between weather and claims are required. Here, interest lies in the latter, in particular, the detection of weather events which induce a high claim risk. The Norwegian meteorological institute predicts that the yearly precipitation will increase by more than 30% for some regions in Norway by 2100 (`www.senorge.no`). A model which explains the impact of certain weather events on the claim risk may hence help decision makers to take adequate actions.

The statistical modelling is challenging due to the spatial and temporal variation of the weather. In other words, some regions are expected to exhibit higher vulnerability to certain weather events. This motivates the application of spatial statistical models which allow a separate analysis for each region while accounting for potential similarities between them to reduce model uncertainty. Here, a Norwegian data set which consider property insurances between 1997 and 2006 is investigated in order to develop such models. In particular, the data provide information on claims caused by small-scale weather events such as rainfall and snow-melt. The data have already been analyzed by Haug et al. (2011) and Scheel et al. (2013) and this thesis extends their approaches and contributes new methodology to the statistical areas of spatial statistics, monotonic regression and extreme value theory.

The remainder of this introductory chapter is organized as follows: Section 1.2 describes the existing insurance policies against weather-related damages and corresponding pricing methods. Further, a summary on the research in actuarial science with respect to climate change is provided. Section 1.3 introduces the Norwegian insurance and weather data which are considered throughout this thesis. Additionally, the results of an explanatory analysis of the data are presented. Section 1.4 summarizes and discusses the work by Haug et al. (2011) and Scheel et al.

(2013) which consider the insurance and weather data. The limitations of their models motivate this thesis whose aim and structure are described in Section 1.5.

1.2 Weather-related Insurances

1.2.1 Property and Weather Index Insurances

Insurance companies offer two products with which individuals and companies can get protection against damages and monetary losses caused by undesirable weather events: property insurances and weather index insurance. The former covers the costs of the damage (up to a sum assured) while the latter's payoff depends on a weather index, e.g. the temperature or the amount of rainfall. Property insurances originated from fire insurances which were extended to include additional perils (Dickson and Steele, 1986). Nowadays, most of the property insurances cover several perils such as flooding, storms, falling trees, subsidence or riots.

In contrast, weather index insurances promise payment based on the difference between a weather index and an agreed strike value. For instance, the owner of an outdoor swimming pool may want to negotiate an insurance against too many rainy days to hedge against a low income. Studies imply that these insurances can be a valuable risk management tool in agribusiness for developed countries (Turvey et al., 2006) and newly industrialized countries (Heimfarth and Musshoff, 2011). Other areas of applications include the energy industry, the construction industry (Jewson et al., 2005) and the tourism industry (Bank and Wiesner, 2011). Richards et al. (2004) state that these insurances are quite beneficial since no damage has to be assessed. This feature is especially important to farmers as it is quite expensive to estimate crop losses. Further, there is only a low risk of moral hazard effects as the payoff is based on an objective measure. As such, farmers are motivated to harvest as much as possible. Barnett and Mahul (2007) further claim that there is little potential for adverse selection since both the insurance policy holders and the insurers have a similar knowledge on the weather. Fuchs and Wolff (2011), however, argue that weather index insurances may also lead to potential overspecialization and monoculture. For both property and weather index insurances, there exists a range of pricing approaches using historical actuarial data and these are described in the following.

Pricing of Property Insurances

Insurance companies aim to set the insurance premium such that it covers both the expected payout and the service costs associated with the policy. Typically, the expected payout per policy is estimated based on recorded and estimated loss data and then *loaded*, that is, the expected claim size is adjusted by service costs, risk covering or profit (Williams et al., 1995; Malinovski, 2008). To load the expected claim size, the risk (variance) has to be estimated and this is one of the objectives in actuarial science. Policy holders are usually classified based on several risk factors and the expected claims size is derived for each class separately. The insurance premium is then calculated by using, for instance, premium calculation principles to load the risk.

Several standard premium principles are described by Goovaerts and Haezendonck (1984). The simplest one is the net premium principle which does not load for risk and is equal to the expected claim size. However, this approach appears only suitable for a large number of policies to ensure validity of the law of large numbers. Extensions of this basic approach are the expected value, the variance and the standard deviation principles. The expected value principle includes a risk load proportional to the expected claim size while the variance premium principle loads a risk proportional to the variance of the claim distribution. Another approach is the equivalent utility principle (Moore and Young, 2003; Xiao, 2011) whose solution can be interpreted as the minimum premium such that the insurance company accepts to insure the expected claim size. Further popular premium principles are the exponential (Gerber, 1974), Esscher (Bühlmann, 1980) and Wang's premium principle (Wang, 1996). Dickson (2005) states five preferable properties which premium principles should satisfy. However, with the exception of the net premium principle, none of the described principles above fulfills all these properties (Young, 2006). More recently developed approaches include the weighted premium principle (Furman and Zitikis, 2008) which contains, e.g., both the net and Esscher principles as special cases (Kaluszka et al., 2012).

The principles outlined above require the expected claim size, the variance of the claim size or even the claim distribution. As a consequence, it is necessary to estimate the required statistical information based on the historical, actuarial data. Eshita (1977) states that the claim distribution should be analyzed via two components: the claim frequency distribution and the claim size distribution. Further, Eshita (1977) motivates the modelling of the claim frequency by a Binomial or a Poisson distribution and this approach is considered in later sections.

The application of these approaches in order to set premiums for property insurances is complicated as the number of factors that influence the claim distribution is very high, e.g., spatial proximity to a river or construction design. Furthermore, the temporal variation of the weather leads to a non-stationary claim distribution. For instance, a day with heavy rainfall is expected to induce a higher average claim frequency than a sunny day. Additional to the temporal variation, weather and climate vary in space and induce a spatially varying relationship between claims and weather. In other words, two locations with distinct yearly rainfall levels are likely to exhibit a different vulnerability to the same amount of rainfall. Prahl et al. (2012) also consider the dependence between the wind speed and the claim distribution. Due to this variety of risk factors, the classification of the actuarial data with respect to economical, hydrological and meteorological factors could lead to a small number of cases in each class. Consequently, the insurance companies would have to load for high uncertainty in the estimates, leading to high insurance premiums. In summary, good reliable statistical approaches to estimate the relationship between multiple risk factors and claims are required in order set adequate insurance premiums.

Pricing of Weather Index Insurances

Weather index insurance contracts are usually defined by five attributes: (i) contract period T, (ii) location of the measurement station, (iii) weather index, (iv) pay-off function which converts the index into cash flow and (v) the premium paid by the policy holder (Jewson et al., 2005). Weather measures of interest are typically temperature and precipitation, but snowfall, hail and sunshine hours also feature. In the case of temperature, the indexes used are cooling degree day (CDD), heating degree day (HDD) and cumulative average temperature (CAT). For example, the CDD index is defined as the cumulative amount of degrees above a threshold c. Mathematically, CDD can be expressed as

$$\sum_{t \in T} \max\{C(t) - c, 0\},$$

where $C(t)$ is the daily average temperature. In the market, the threshold c is equal to 18°C (Benth and Šaltytė Benth, 2011). Pricing techniques for weather index insurances are similar to the ones for weather derivatives which are described in the following.

Weather derivatives are traded financial instruments which can be used by companies but

also private households to reduce the risk associated with weather events. They differ from weather index insurances in some contract details (Jewson et al., 2005) but payout depends on the same weather indexes mentioned above. Nowadays, weather derivatives are, for instance, traded at the Chicago Mercantile Exchange (Dorfleitner and Wimmer, 2010). Pricing techniques for weather derivatives are based on the statistical modelling of historical weather data. *Burn analysis* evaluates how a contract would have performed over the training period to estimate the expected payout. Špička (2011) performs parametric bootstrapping (Efron, 1979) within the burn analysis to quantify the uncertainty. Another approach is *index modelling* which fits a statistical distribution to the historical weather data and provides some information on the values outside the range of observed values (Taib and Benth, 2012).

Publications considering weather derivatives often focus on temperature indexes. Several papers model the temperature dynamics first and derive the prices in a second step. Richards et al. (2004) model the temperature data from Fresno County, California, by a mean-reverting Brownian motion with log-normal jumps and time-varying volatility. They also state that the temperature is more volatile in winter than in summer and capture this property by a first-order ARCH process. Similarly, Benth and Šaltytė Benth (2011) model the temperature as the sum of a seasonal mean function and a continuous-time autoregressive process. The design of precipitation-related weather derivatives are, for instance, considered by Stoppa and Hess (2003). Paulson et al. (2010) apply spatial kriging and MCMC techniques to estimate rainfall histories and to derive insurance rates.

1.2.2 The Impact of Climate Change

Climate change makes the application of the pricing techniques described in the previous Section 1.2.1 difficult, because the relationship between claims and weather events varies in time, requiring more complicated statistical models. Consider, for instance, an increased frequency of intense rainfall which causes more cases of household flooding than currently observed. This may motivate decision makers to improve the infrastructure, e.g. extend the drainage system, and house owners to invest in new construction designs to deal with these events. Hence, the risk induced by certain amounts of precipitation is reduced. Furthermore, the temporal variation in the risk structure and the climate implies higher uncertainty which is passed on to the policy holders in the form of higher insurance premiums. The aspect of insuring and pricing the uncertainty on climate change is discussed by Tol (1998) and Litterman (2011).

Actuarial research focuses on the impact of climate change with respect to the intensity and occurrence of natural disasters, e.g. floods, storms or heat waves. These extreme events generally imply high expenses for the insurance companies; their effect on economic growth is investigated by Loayza et al. (2012). Some countries have implemented a special pool that settles the natural disaster damage compensation. In Norway, for instance, all policyholders who enter a fire insurance have to pay a fee to the Norwegian Natural Perils Pool. Though this thesis considers claims caused by non-extreme weather events, properties found on natural disasters may be similar and thus considerable. Hence, the results are summarized in the following.

Several authors investigate a potential temporal trend in the monetary losses due to natural disasters based on estimated loss data. However, the availability and correctness of these estimates is limited, though it has improved over the past decades (Downton and Pielke, 2005). Uncertainty on the loss data may be reduced by considering insured losses instead since these are estimated with greater precision (Barthel and Neumayer, 2012). Furthermore, the data has to be normalized with respect to economic wealth in order to be comparable from year to year. Pielke and Landsea (1998) normalize the losses with respect to inflation, population and growing wealth. Neumayer and Barthel (2011) extend this normalization as it ignores a potential spatial variation of the wealth within a country. Nevertheless, this alternative requires information on the insured assets potentially at risk in any given area, a condition which is usually infeasible (Barthel and Neumayer, 2012).

The normalized loss data are then used to explore a potential global temporal trend. Miller et al. (2008) consider weather-related natural disasters between 1950 and 2005 from several developed and developing countries and find an annual upward trend of 2% per year since 1970. Nevetheless, Miller et al. (2008) state that the results are highly affected by Hurricane Katrina with an estimated insured loss of $41.1 billion and a total loss of $108 billion in 2005 (Knabb et al., 2005). Further, no significant temporal trend is found if the US hurricane season 2004-2005 and the flood damages between 1970 and 2005 in China are left out. Barthel and Neumayer (2012) find no global trend from 1990 to 2008. However, the considered period is short in terms of climate modelling and the authors also do not split between geophysical, e.g. earthquakes and tsunamis, and weather-related natural disasters. In summary, there is no clear consensus on a positive global trend in terms of the monetary losses related to natural disasters.

1.3 The Insurance and Weather Data

1.3.1 Description

The insurance data are kindly provided by Gjensidige (`www.gjensidge.no`) and include observations for all 430 Norwegian municipalities between 1st January 1997 and 31st December 2006. These records cover all insured private buildings over this period and give information on claims which are caused by non-catastrophic related weather events. Natural disasters are excluded as these are covered via a governmental pool and not recorded by Gjensidige. For each municipality, the data constitute the daily number of claims due to damages caused by precipitation, surface water, snow melting, undermined drainage, sewage back-flow or blocked pipes. Additionally, the monthly number of policy holders per municipality is reported too. In what follows, $N_{k,t}$ and $A_{k,t}$ denote the number of claims and policies, respectively, on day t for municipality k, $k = 1, \ldots, 430$.

The meteorological and hydrological data are produced by the Norwegian Meteorological Institute (`www.met.no`), together with the Norwegian Water Resources and Energy Directorate (`www.nve.no`). Four daily measurements are provided for each municipality: (i) Mean temperature $C_{k,t}$, (ii) Precipitation $R_{k,t}$ (iii) Drainage run-off $D_{k,t}$ and (iv) Snow-water equivalent $S_{k,t}$. The covariate $D_{k,t}$ provides information on surface water while $S_{k,t}$ corresponds to the amount of water which is stored in form of snow on the ground.

These weather data are generated as follows: Temperature and precipitation measurements are recorded for more than 200 weather stations across Norway and then spatially interpolated to a high-resolution grid with 1×1 km cells. The precipitation measurement on day t is recorded at 6am in the morning, i.e. the rain observed on day t is mostly related to events on day $t-1$. Since the weather may vary considerably within a municipality and damages are supposedly linked to local events, $C_{k,t}$ and $R_{k,t}$ are derived by weighted averaging over the most densely populated grid cells within the municipality only. The values for $D_{k,t}$ and $S_{k,t}$ are then derived via a gridded water balance model. In the following, the covariate value $R_{k,t}$ refers to the precipitation measurement at 6am on day $t+1$ and this notation deviates from the one by Haug et al. (2011) and Scheel et al. (2013).

Table 1.3.1: Provided and derived weather covariates which are considered in Section 1.3.2.

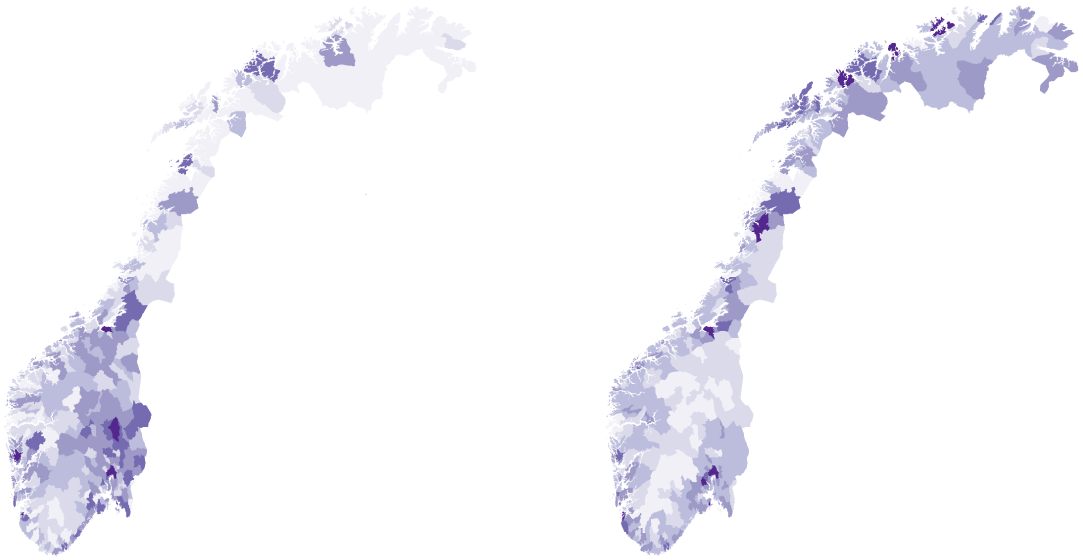
Variable	Description	Unit
$R_{k,t}$	Total amount of precipitation in day t (Between 6am on day t to 6am on day $t+1$)	mm
$R_{k,t-1}$	Total amount of precipitation in day t (Between 6am on day $t-1$ to 6am on day t)	mm
$C_{k,t}$	Mean temperature in day t	°C
$D_{k,t}$	Drainage run-off in day t	mm
$S_{k,t}$	Snow-water equivalent in day t	mm
$\Delta S_{k,t}$	Difference in snow-water equivalent $S_{k,t-1} - S_{k,t}$	mm

1.3.2 Exploratory Data Analysis

Since this thesis considers modelling the dependence of the claims and weather events, an exploratory analysis is performed in this section. Both the weather and insurance data are firstly assessed separately with respect to temporal and geographical variations. Additionally, a potential geographical correlation of the claim numbers for adjacent municipalities is examined. Finally, the relationship of $N_{k,t}$ and the weather covariates is investigated. Table 1.3.1 summarizes the six covariates which are considered, the rain on the previous day $R_{k,t-1}$ and the difference in the snow-water equivalent $\Delta S_{k,t}$ are derived as additional covariates to the four mentioned in Section 1.3.1. Positive values in $\Delta S_{k,t}$ correspond to a drop in the amount of water stored in form of snow and, hence, imply a period of snow-melt.

Spatial and Temporal Analysis of the Insurance Data

Prior to $N_{k,t}$, the spatial and temporal variation in $A_{k,t}$ is explored. Figure 1.3.1a illustrates the average number of policies for each municipality over the 10-year period. The scale of these figures is dropped in order not to disclose industrial information. Nevertheless, there is no loss of understanding of the qualitative properties of these plots. The highest values of $A_{k,t}$ are recorded for the large cities of Oslo and Bergen while rural municipalities observe substantially lower values. This geographical variation in $A_{k,t}$ is consistent with the one in the total population, Oslo counts more than 500,000 inhabitants while other municipalities have a population of less than 1,000. Therefore, it is assumed that the portfolios of property insurances for the different municipalities are comparable. On average, the total number of policies per month over all municipalities, $\sum_{k=1}^{430} A_{k,t}$, is about 400,000 and some temporal variation is found.



(a) Average number of policies **(b)** Standardized average claim rate

Figure 1.3.1: (a) Average number of policies and (b) Average claim rate standardized with respect to the highest average claims per day per policy holder ratio for all 430 municipalities between 1997 and 2006. A darker colour corresponds to a higher value.

Considering $N_{k,t}$, the insurance data exhibit a large frequency of zero claims, so on many days no damages are reported within a municipality. Averaged across all 430 municipalities, no claims, $N_{k,t} = 0$, are reported on about 98% of days and $N_{k,t} > 1$ with a frequency of less than 0.2%. These findings are, however, non-homogeneous with respect to the individual municipalities. While the large cities of Oslo and Bergen each record more than 300 days with $N_{k,t} > 1$, no such event is observed for 172 municipalities over the 10-year period. Since $A_{k,t}$ is higher for densely populated municipalities (Figure 1.3.1a), claims are presumably more frequent. However, Figure 1.3.1b also indicates that, especially in southern and central Norway, cities exhibit a larger average claim rate per policy holder than rural areas. In particular, the cities of Drammen, Oslo, Bærum and Trondheim are among the highest. The lack of natural drainage in the cities potentially leads to a higher risk related to surface water. Furthermore, the largest daily observations are also recorded for cities, 135 in Bærum, followed by Trondheim with 59 and Oslo with 57. Note, the maximums in each municipality are distinctively higher than all other observations. For instance, the second highest observations of $N_{k,t}$ for Bærum and Oslo are 27 and 18, respectively.

Temporal dependence for the time series $N_{k,t}$, $k = 1, \ldots, 430$, is examined next. Figure 1.3.2 illustrates that very high observations are not recorded every year even for large cities;

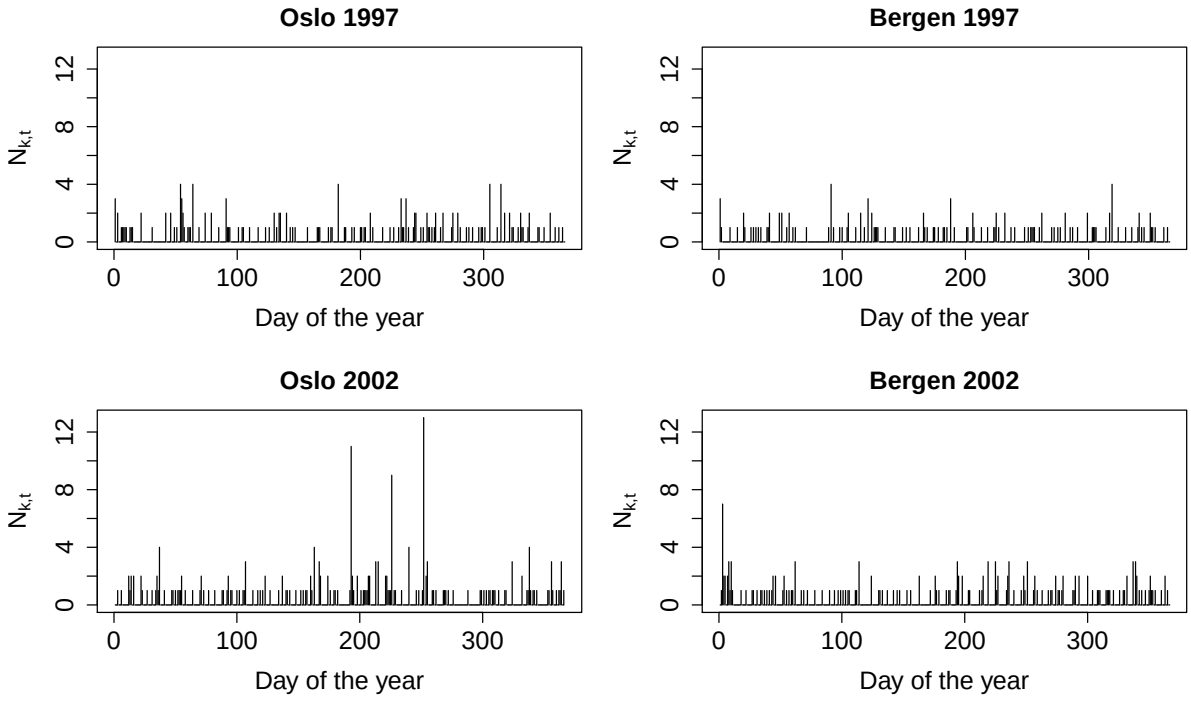
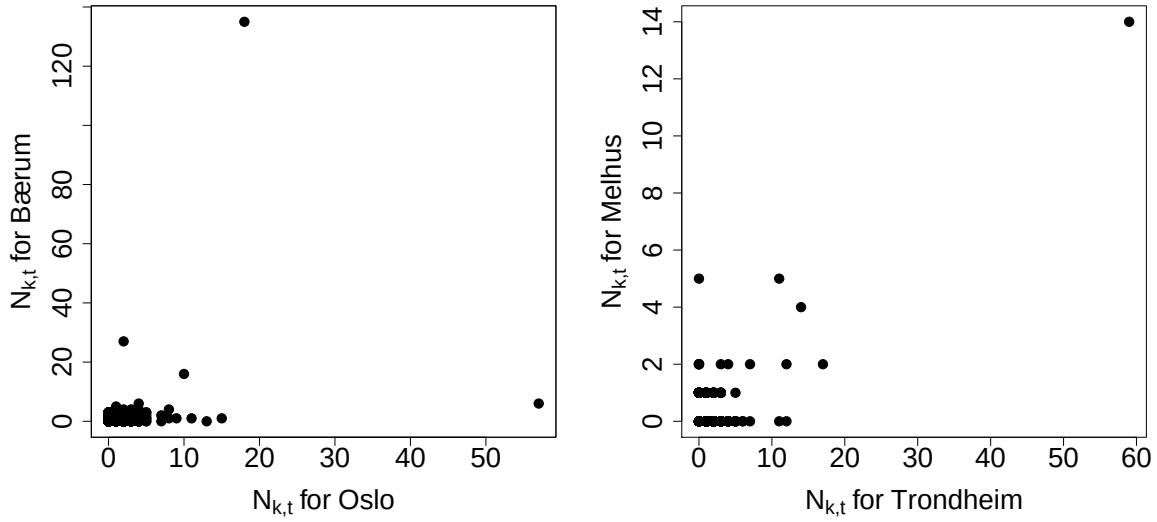


Figure 1.3.2: Daily number of claims for Oslo (Column 1) and Bergen (Column 2) for the years 1997 and 2002.

the maximum daily observed number of claims in 1997 is $N_{k,t} = 4$ for both Oslo and Bergen. Further, there is little, or no, seasonality in the data and high claim numbers occur in both summer and winter. Note, the highest observations are generally recorded over the summer. Additional to seasonality in the claim numbers, temporal dependence is also explored based on a daily basis. If at least one claim occurs in Oslo, then the empirical probability for at least one claim on the next day, $\mathbb{P}(N_{k,t} > 0 \mid N_{k,t-1} > 0)$, is 0.65, as opposed to 0.62 on any day. More formally, the data exhibits a slight positive correlation for claim occurrences for Oslo. Similar but smaller levels of positive correlation are also found for Bærum and Trondheim but not for Bergen. These findings may indicate two hidden processes. Firstly, weather events may affect claim dynamics over consecutive days, leading to higher vulnerability and an increased risk. Secondly, there exists a potential lag in the claim recording process since $N_{k,t}$ refers to the claims reported to the insurance company on the day and these are not necessarily identical to the ones which occurred on the day. For instance, a weather event may cause two damages and one is reported on the same day while the second one the day after.

Spatial correlation of claims is explored for two pairs of adjacent densely populated municipalities. Figure 1.3.3a shows some dependence for high numbers of claims for Oslo and Bærum in south-east Norway. The observation of 135 claims for Bærum coincides with 18 recorded claims



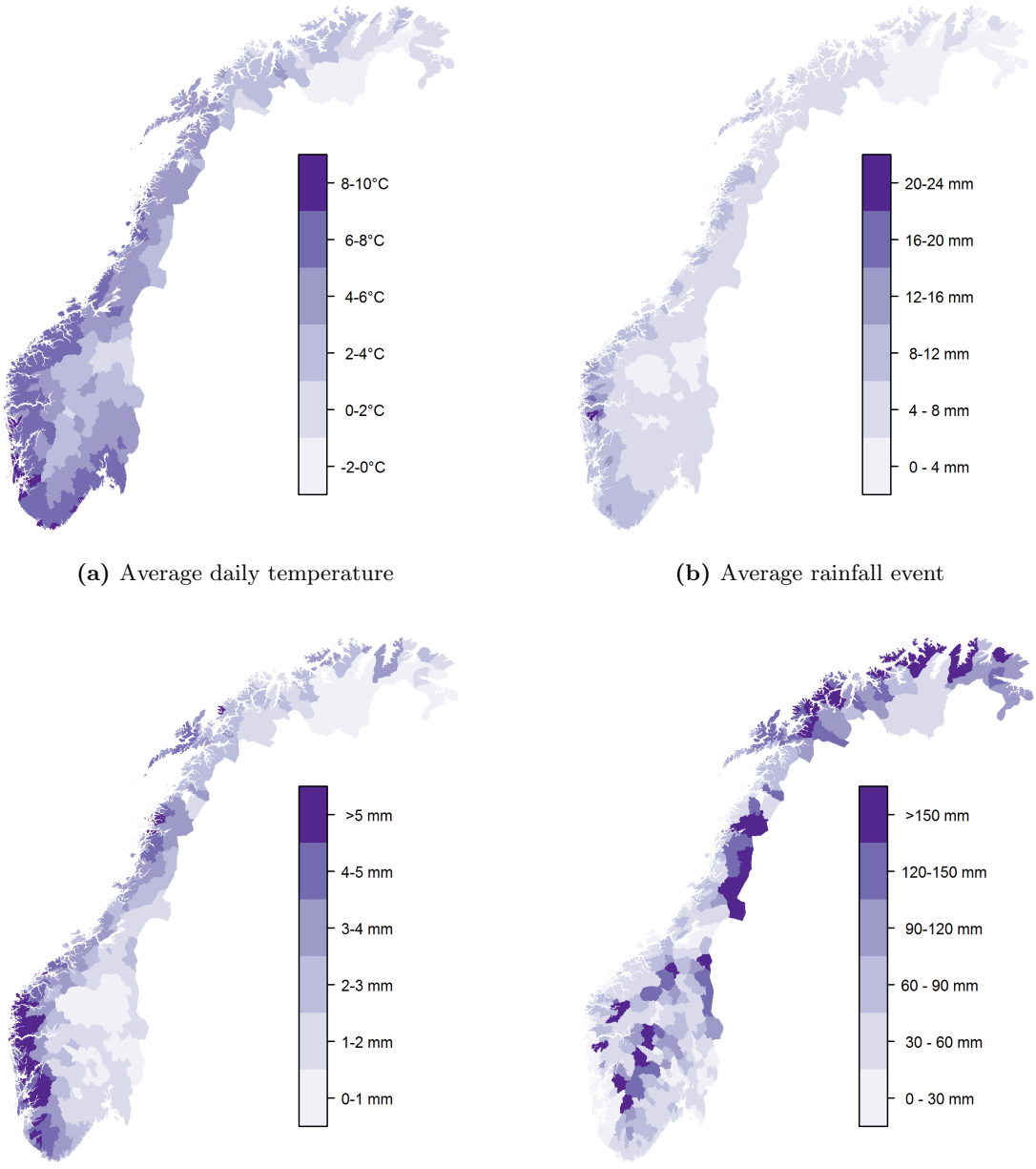
(a) Oslo and Bærum **(b)** Trondheim and Melhus

Figure 1.3.3: Number of claims reported on day, $N_{k,t}$, for two pairs of adjacent Norwegian municipalities.

for Oslo which is the second highest observation for the latter. Similar conclusions are found for the municipalities of Melhus and Trondheim in north-west Norway in Figure 1.3.3b, high observations of $N_{k,t}$ coincide on several days. The Pearson's correlation coefficient for the two pairs is calculated too and yields to 0.28 for Oslo and Bærum, and 0.64 for Trondheim and Melhus. Consequently, these results are consistent with the findings from Figure 1.3.3. This dependence is presumably due to adjacent municipalities often being exposed to similar weather conditions. Hence, a severe weather event in one place is likely to affect the adjacent municipalities too. This relationship supposedly also holds vice versa for low-risk weather events.

Spatial and Temporal Analysis of the Weather Data

Norway's geographical extent, it spans about 13 degrees in latitude, leads to distinct spatial differences in the climate. Figure 1.3.4a illustrates that the highest average temperatures of about 8°C in the data are recorded for the south-western municipalities around Bergen and Stavanger. Conversely, lower average temperatures of around 0°C are observed for northern and easterly municipalities and in the Scandinavian Mountains. The mild temperatures for coastal areas are due to their exposure to the North Atlantic Current and Gulf Stream. Indeed, the warm ocean currents lead to high temperatures, compared to Alaska, Greenland and Siberia which have similar latitude. The northern and easterly municipalities have a more continental



(a) Average daily temperature

(b) Average rainfall event

(c) Average drainage run-off

(d) Average snow-water equivalent

Figure 1.3.4: Averages across Norway between 1997 and 2006 for (a) Temperature $C_{k,t}$ (b) Rainfall event $R_{k,t}|R_{k,t} > 0$ (c) Drainage run-off $D_{k,t}$ and (d) Snow-water equivalent $S_{k,t}|S_{k,t} > 0$.

climate with cold winters and hence observe lower average temperatures.

The exposure to the ocean does also affect the frequency and intensity of rainfall events. Figure 1.3.4b shows that the highest precipitation levels are observed around Bergen with, on average, more than 20 mm while most inland municipalities exhibit average rainfall events between 5 and 10 mm. The large observations for coastal areas are caused by orographic (relief) and frontal precipitation. Conversely, eastern municipalities lie in the rain shadow and the humid

Atlantic air is blocked by the Scandinavian Mountains. In terms of the frequency, $R_{k,t} > 0$ is observed on about 70% of days for Bergen while only on 50% for Oslo. Furthermore, $R_{k,t}$ varies temporally with western areas observing the highest rainfall events during autumn and winter while summer is the wettest season for Oslo; time series for 4 municipalities for 1998 are provided in Appendix A.1. This difference in the rainfall levels transfers to the drainage run-off with the highest averages being observed for coastal municipalities (Figure 1.3.4c).

In terms of the snow-water equivalent, Figure 1.3.4d illustrates that high averages are recorded for northern municipalities and the Scandinavian Mountains while observations are generally lower for the south-west. Furthermore, differences between adjacent municipalities are larger for $S_{k,t}$ than for the previous covariates $C_{k,t}$, $R_{k,t}$ and $D_{k,t}$. The large differences across Norway are linked to those for $C_{k,t}$ since colder areas generally exhibit longer periods with temperatures below 0°C. Consequently, snow accumulates over several days and higher values for $S_{k,t}$ occur while, as mentioned by Scheel et al. (2013), snow is quite rare on the west coast. Finally, the intensity of the snow-melt is explored by considering days with $\Delta S_{k,t} > 0$ only and little spatial variation is found; a plot is provided in Appendix A.2. Consequently, periods of snow-melt vary spatially across Norway with respect to their duration rather than in terms of their intensity.

Dependence between Insurance and Weather Data

In this thesis, the main interest lies in the relationship between the weather covariates and the daily number of claims. The following analysis considers the two most densely populated municipalities of Oslo and Bergen since they record the most claims and thus provide the most insight. Furthermore, weather events leading to a higher number of claims may be investigated better and compared to weather events leading to a smaller number of claims. Additional to a separate analysis for the dependence in each covariate, potential combined effects between pairs of covariates are examined.

The covariates $R_{k,t}$ and $R_{k,t-1}$ capture the risk induced in form of both rain and snowfall on the day itself and the previous one, respectively. Figure 1.3.5 indicates that the highest claim numbers are generally related to higher observations for $R_{k,t}$ or $R_{k,t-1}$. However, this association is mostly observed for days with no snow since higher responses usually occur for $S_{k,t} = 0$. Further, the plots indicate that Oslo and Bergen have a different vulnerability to the same amount of precipitation. While very high responses in Oslo are already observed for

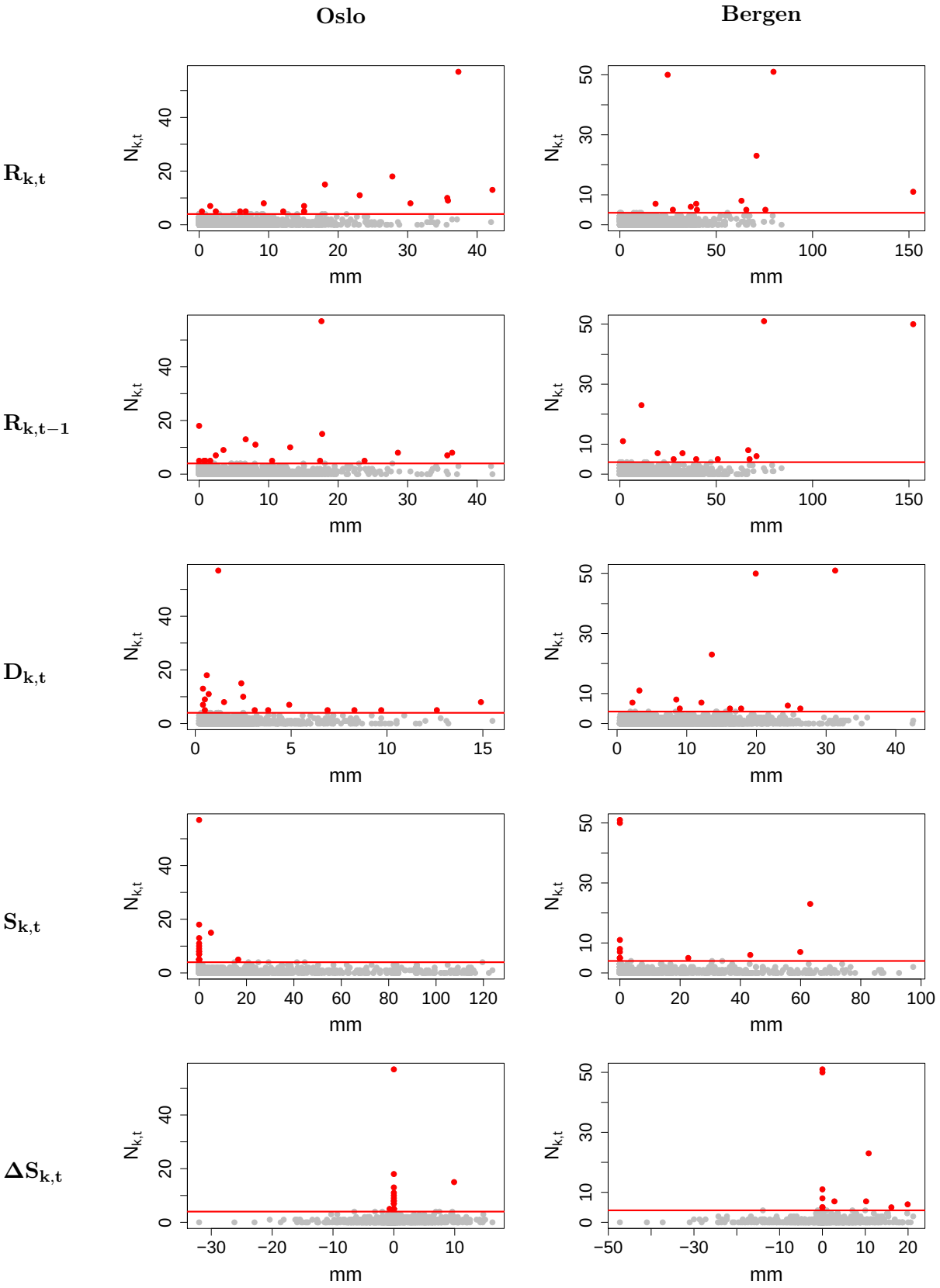
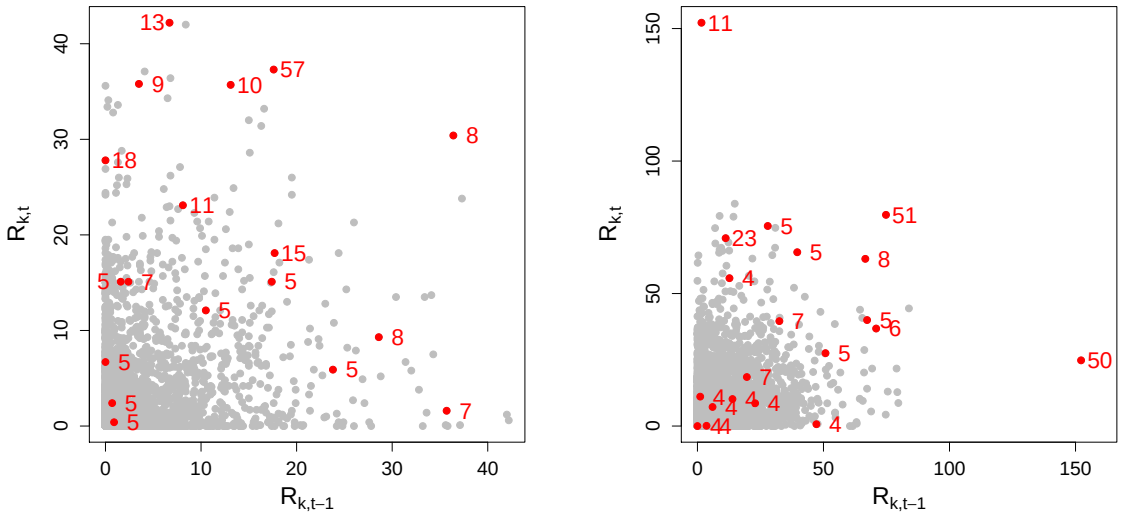


Figure 1.3.5: Recorded number of claims $N_{k,t}$ in dependence on the observed covariates $R_{k,t}$, $R_{k,t-1}$, $D_{k,t}$, $S_{k,t}$ and $\Delta S_{k,t}$ for the municipalities of Oslo (Column 1) and Bergen (Column 2). The red points refer to the observations with $N_{k,t} > 4$.



(a) Oslo **(b)** Bergen

Figure 1.3.6: Observed covariate values for $R_{k,t}$ and $R_{k,t-1}$ for Oslo (left panel) and Bergen (right panel). Days with $N_{k,t}$ exceeding 4 and 3 are highlighted for Oslo and Bergen, respectively.

rainfall levels exceeding 15 mm, this only occurs in Bergen for much higher amounts of 30-50 mm. Consequently, the plots indicate that houses in Bergen are designed to withstand more severe rainfall events than Oslo which is expected since the west coast exhibits much higher precipitation levels. Nevertheless, high observations for one of the two precipitation covariates do not necessarily imply high responses. Figure 1.3.6 illustrates this aspect since even rainfall levels exceeding 30mm are not leading to high observations of $N_{k,t}$ for Oslo. Conversely, the combined effect of moderate precipitation levels on both days t and $t-1$ may also cause higher claim numbers.

The risk induced by $R_{k,t}$ and $R_{k,t-1}$ may depend on whether the precipitation is in form of rain or snow. As mentioned previously, high responses are usually observed for $\Delta S_{k,t} = 0$. In order to explore this aspect, observations of the mean daily temperature $C_{k,t}$ are used. More precisely, we examine the correlation between $R_{k,t}$ and $N_{k,t}$, conditional on $C_{k,t}$. Figure 1.3.7 shows that the correlation is generally negative or close to zero for days with $C_{k,t} < 0°\text{C}$ while being positive for warmer days. Hence, snow may have a non-significant effect on the claim dynamics for the same day while rain does. Snow over a period with negative values for $C_{k,t}$ accumulates and, hence, additional snow does not necessarily increase the risk of water damages on the same day. However, additional snow may still lead to a higher risk in terms of the next snow-melting period. In order to account for this behaviour in the covariates, one may set $R_{k,t}$

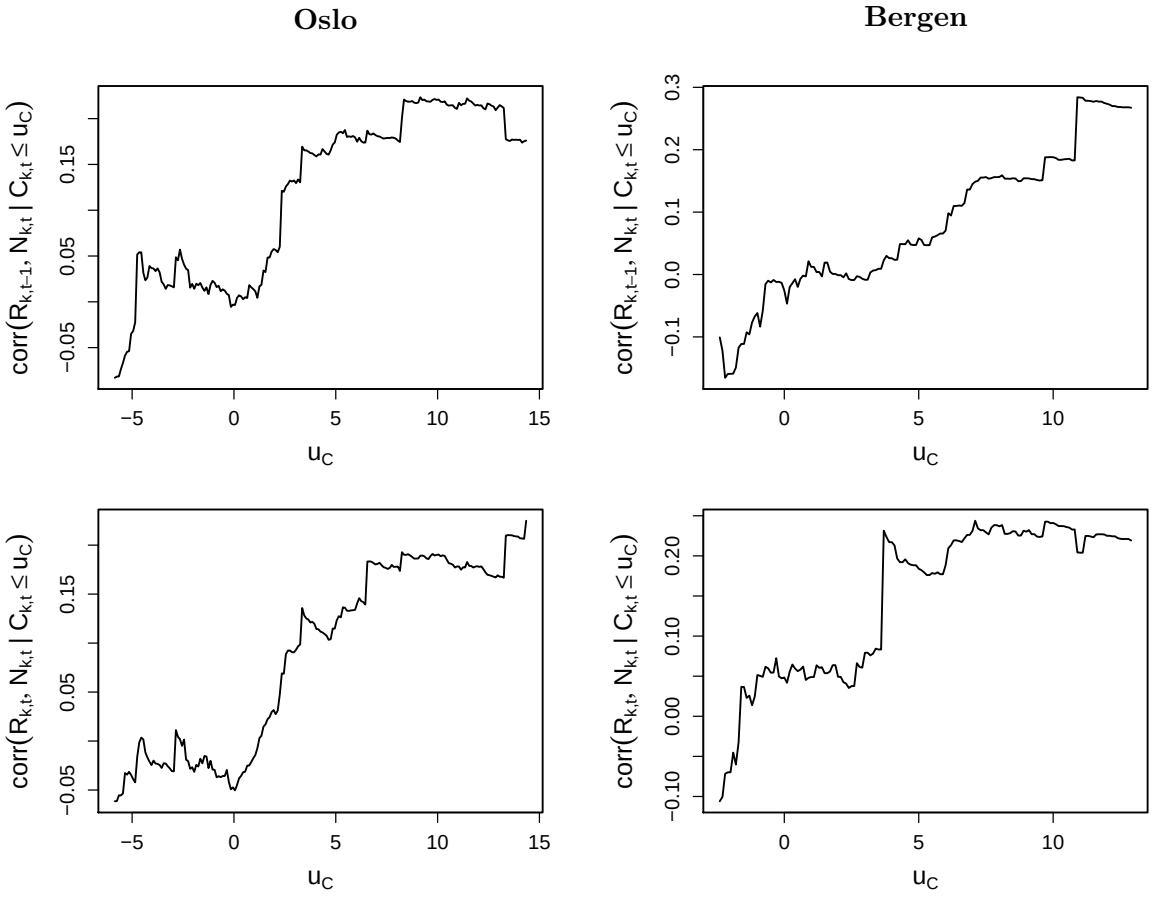


Figure 1.3.7: Pearson's correlation coefficient between number of claims $N_{k,t}$ and amounts of precipitation $R_{k,t-1}$ and $R_{k,t}$ in dependence on the mean daily temperature $C_{k,t}$. The functional level corresponds to the correlation between $N_{k,t}$ and $R_{k,t-1}$ (Row 1), and $N_{k,t}$ and $R_{k,t}$ (Row 2) for Oslo (Column 1) and Bergen (Column 2) conditional on $C_{k,t}$ being smaller or equal u_C.

equal to 0 for days with $C_{k,t} \leq 0$. Otherwise, each rainfall covariate may be split into two covariates to account for a potential difference between snow and rain. Additional to Pearson's correlation coefficient, Kendall's and Spearman's correlation coefficient have been considered and the results are consistent with the ones in Figure 1.3.7; plots are provided in Appendix A.3.

Similarly to $R_{k,t}$ and $R_{k,t-1}$, Figure 1.3.5 shows that high values in $D_{k,t}$, $S_{k,t}$ or $\Delta S_{k,t}$ do not automatically imply high responses. With respect to $D_{k,t}$, higher claim numbers are observed for both lower and higher covariate values. This is particularly the case for Oslo where the highest responses coincide with $D_{k,t} \leq 5$. The snow-water equivalent $S_{k,t}$ is considered next and the plot indicates no apparent dependence. Days with higher number of claims, conditional on $S_{k,t} > 0$, generally coincide with rain, $R_{k,t} > 0$, or snow-melt, $\Delta S_{k,t} > 0$. For both Oslo and Bergen, the highest response, conditional on $S_{k,t} > 0$, occurs on a day where both $R_{k,t}$ and

$\Delta S_{k,t}$ are positive. Similarly to $R_{k,t}$ and $R_{k,t-1}$, the correlation coefficient is derived. While the correlation coefficient between $N_{k,t}$ and $\Delta S_{k,t}$, conditional on $\Delta S_{k,t} > 0$, takes a value of about 0.10, it is zero or slightly negative for days with $\Delta S_{k,t} < 0$. Consequently, there exist a potential difference between rain and snow which neither Haug et al. (2011) nor Scheel et al. (2013) discuss. In summary, the exploratory analysis indicates a dependence between the claims on the day, $N_{k,t}$, and both rainfall, $R_{k,t}$ and $R_{k,t-1}$, as well as snow-melt, $\Delta S_{k,t} > 0$.

1.4 Existing Claim Models for the Insurance and Weather Data

1.4.1 Haug et al. (2011)

Interest lies in modelling the total amount of insured losses on a day for each of the 430 Norwegian municipalities. Additional to the daily number of claims $N_{k,t}$ and the monthly number of policies $A_{k,t}$, the mean claim sizes $\overline{M}_{k,t}$ on day t for municipality k are observed. Similarly to Eshita (1977), the claim distribution is considered via two separate factors: the claim numbers $N_{k,t}$ and the average claim size $\overline{M}_{k,t}$. The claim model is then applied to assess the impact of climate change with respect to claim frequency and claim sizes.

Claim Model

The number of claims $N_{k,t}$ is modelled via a Binomial distribution with overdispersion to account for the high variability in the data. Hence, the distribution on day t for municipality k has mean $\mathbb{E}(N) = A_{k,t}\, p_{k,t}$ and variance $\mathrm{Var}(N) = \phi\, A_{k,t}\, p_{k,t}\, (1-p_{k,t})$, where ϕ is the dispersion parameter and $p_{k,t}$ denotes the claim probability. Dependence between $N_{k,t}$ and the weather data described in Section 1.3 is defined via $p_{k,t}$. In particular, $N_{k,t}$ is modelled via a generalized linear model (GLM) (McCullagh and Nelder, 1989) with logit link function.

Haug et al. (2011) perform further analysis prior to the estimation of the GLM to obtain a suitable set of explanatory variables. Firstly, additional covariates, such as the aggregated rain over the previous five days, $R_{k,t-2}$ to $R_{k,t-6}$, are derived. Secondly, seasonal components are introduced to account for a potential temporal trend due to unknown processes and macroeconomic factors. Finally, potential parametric forms for the weather covariates, e.g. $C^2_{k,t}$, are examined based upon their generalized additive model fit (Hastie and Tibshirani, 1990). Note, the difference in the snow-water equivalent $\Delta S_{k,t}$ is not considered in their approach. To obtain a parsimonious set of explanatory variables, variable selection is performed on the set of can-

didates via the Bayesian Information Criterion (BIC) (Schwarz, 1978) which is modified as in Burnham and Anderson (2002) to handle overdispersion. The derived set contains 13 variables, including the weather observations described in Section 1.3 but with the drainage run-off $D_{k,t}$ being on log-scale. The GLM is then fitted county-wise with covariate effects being taken as constant for each municipality within a county. Geographical variation of the baseline risk, as found in Section 1.3, is modelled via a mean county-level and a geographically varying correction term for each municipality.

The average claim size $\overline{M}_{k,t}$ is modelled similarly via a GLM with log-link function. Average claim sizes are assumed to be Gamma distributed and the mean varies with the set of explanatory variables. The expectation of $\overline{M}_{k,t}$ is further assumed to be independent of $N_{k,t}$ but the variance decreases in $N_{k,t}$, i.e. $N_{k,t}$ is used as a weight. Again, a parsimonious set of explanatory variables is derived similarly to $N_{k,t}$ but based on the original BIC. The final set contains ten explanatory variables but the snow-water equivalent $S_{k,t}$ and the temperature $C_{k,t}$ are found to be uninformative. In terms of model parameters, larger geographical entities (regions) are considered since days with no claims are uninformative in terms of the claim sizes. Equivalently to $N_{k,t}$, covariate effects are constant within each of these regions. Further, the baseline levels are the same for each municipality within a county and geographical variation is only allowed for between counties.

Results

The statistical model is estimated for all 430 municipalities and results for the counties of Hordaland, Akershus and Buskerud are provided; the remaining counties are omitted by Haug et al. (2011) due to confidentiality reasons. Dependence between the responses and the covariates is examined via normalized factor curves. The effect of $R_{k,t}$ and $R_{k,t-1}$ is not explored via separate factor curves but merged. Precipitation levels are found to be significant and positively correlated for both $N_{k,t}$ and $\overline{M}_{k,t}$. Further, the easterly county of Akershus appears more vulnerable to rainfall than the westerly county of Hordaland. This is consistent with the results in Section 1.3; the west coast presumably exhibits a smaller vulnerability for the same amount of precipitation, as compared to other municipalities. The observed snow-water equivalent $S_{k,t}$ is only significant and positively correlated in terms of $N_{k,t}$. Compared to precipitation, the induced risk is smaller and Hordaland shows less vulnerability than the other two counties. Similar conclusions are found for $D_{k,t}$ which is only significant for Akershus and Buskerud.

Finally, the factor curve for the mean temperature $C_{k,t}$ is U-shaped with peaks at large positive and negative values for all three counties. The U-shape of the curve is due to a second order term of the temperature contained in the model. Equivalently to $D_{k,t}$ and $S_{k,t}$, $C_{k,t}$ is only significant for the distribution of the claim numbers $N_{k,t}$ but not the average claim sizes $\overline{M}_{k,t}$.

Haug et al. (2011) assess the model fit by comparing the observed and estimated average claim frequencies and claim sizes. These averages are more or less equal for the claim frequency and quite close for the claim sizes for all three counties. The estimated claim model is applied to predict the impact of climate change on the insured losses for the future period 2071-2100; see Haug et al. (2011) for technical details on how these estimates are derived. In terms of the claim frequency, results indicate an 30% increase for Hordaland and the north of Oslo but a smaller one of about 5-10% for the more rural municipalities in Buskerud. Furthermore, the impact of climate change appears similar for adjacent municipalities. Similar geographical patterns are found with respect to the change in mean claim size which is relatively small, compared to the increase in claim frequency. For example, the increase is 1-3% for all considered municipalities in Buskerud. Since this thesis considers the modelling of $N_{k,t}$ in dependence on the weather covariates, the reader is referred to Haug et al. (2011) for more detailed results on their effect study of climate.

1.4.2 Scheel et al. (2013)

In contrast to Haug et al. (2011), their work only considers the dependence between the daily claim numbers $N_{k,t}$ and the weather data for $K = 319$ municipalities in central and southern Norway. Interest lies in both the detection of weather events which may lead to claims and the prediction of the number of claims, given a particular weather event. In the context of the covariates described in Section 1.3, the aim is to derive which of these are important for the claim dynamics in the individual municipalities.

Claim Model

Scheel et al. (2013) argue that a Binomial or Poisson distribution for $N_{k,t}$ ignores important features of the data. Firstly, the frequency of zero claims is larger than expected from fitting a Poisson distribution. Secondly, the mechanisms leading to claims may be different from the ones for the number of claims, given a claim occurred. In order to account for these effects, they propose a Poisson hurdle model (Mullahy, 1986) which consists of two components. Hurdle

models are widely applied, for instance in agricultural economics (Ricker-Gilbert et al., 2011) and medical statistics (Neelon et al., 2013). Zero-inflated Poisson models (Lambert, 1992) would be another approach to account for the excessive occurrence of zero claims. Nevertheless, the Poisson hurdle model appears more appropriate due to the possibility that the mechanism of whether a claim occurs is different from the mechanism leading to the number of claims.

The first component, a Bernoulli distribution, models whether the number of claims is either zero, $N_{k,t} = 0$, or strictly positive, $N_{k,t} > 0$. A zero-truncated Poisson distribution is defined as second component for the strictly positive counts $N_{k,t} \mid N_{k,t} > 0$. Let $\alpha_{k,t} \in [0,1]$ denote the probability for no claims on day t for municipality k while $\lambda_{k,t} > 0$ refers to the rate of the zero-truncated Poisson distribution. The probability mass function for $N_{k,t}$ is then formally given as

$$\mathbb{P}(N_{k,t} = n \mid \alpha_{k,t}, \lambda_{k,t}) = \begin{cases} \alpha_{k,t} & \text{if } n = 0 \\ (1 - \alpha_{k,t}) \, \dfrac{\lambda_{k,t}^n}{n! \, \{\exp(\lambda_{k,t}) - 1\}} & \text{if } n > 0. \end{cases} \tag{1.4.1}$$

Both parameters $\alpha_{k,t}$ and $\lambda_{k,t}$ vary in dependence on the weather data. Additional to the six covariates explored in Section 1.3 ($R_{k,t}, R_{k,t-1}, S_{k,t}, D_{k,t}, C_{k,t}$ and $\Delta S_{k,t}$), the aggregated rainfall over the previous three days, $R_{k,t-1} + R_{k,t-2} + R_{k,t-3}$, is derived. Let $\mathbf{X}_{k,t} = (X_{k,t,1}, \ldots, X_{k,t,7})$ denote the vector of the seven covariate observations for municipality k on day t. Separate models are then specified for $\alpha_{k,t}$ and $\lambda_{k,t}$, conditional on $\mathbf{X}_{k,t}$. While the covariate effects are permissibly different across municipalities and no dependence structure is defined on their regression coefficient, the locations of the municipalities are used to perform geographically smoothed variable selection. The model expresses the belief that, given covariate j, $j = 1, \ldots, 7$, influences the claim dynamics for municipality k, covariate j is presumably also important for the processes in its adjacent municipalities. The models for $\alpha_{k,t}$ and $\lambda_{k,t}$ are detailed in the following.

In case of the Bernoulli component with parameter $\alpha_{k,t}$, a logit link function is used for the transformation of the linear predictor. In order to perform geographically smoothed variable selection, a set of latent binary variables is introduced. Define $\boldsymbol{\gamma}_k^\alpha = (\gamma_{k,1}^\alpha, \ldots, \gamma_{k,7}^\alpha)$, $k \in \{1, \ldots, 319\}$ such that $\gamma_{k,j}^\alpha = 1$ if the jth covariate enters the model for $\alpha_{k,t}$ and $\gamma_{k,j}^\alpha = 0$ otherwise. The parameter $\alpha_{k,t}$, conditional on $\mathbf{X}_{k,t}$ and $\boldsymbol{\gamma}_k^\alpha$, is formally given as

$$\operatorname{logit}(\alpha_{k,t}) = \nu_{k,0} + \sum_{\{j \,:\, \gamma_{k,j}^\alpha = 1\}} \nu_{k,j} X_{k,t,j}. \tag{1.4.2}$$

Gaussian priors are defined for the baseline risk $\nu_{k,0}$ and the set of covariate effects $\nu_{k,j}$ for which the associated $\gamma^{\alpha}_{k,j} = 1$; see Scheel et al. (2013) for details.

Geographically smoothed variable selection is performed by specifying, a priori, a dependence model for the vectors of binary variables $\widetilde{\boldsymbol{\gamma}}^{\alpha}_{j} = \left(\gamma^{\alpha}_{1,j}, \ldots, \gamma^{\alpha}_{K,j}\right),\ j = 1, \ldots, 7$. Specifically, an Ising prior distribution is assumed for each covariate whose probability mass function is given as

$$p(\widetilde{\boldsymbol{\gamma}}^{\alpha}_{j} \mid \omega_j) \propto \exp\left\{\omega_j \sum_{1 \leq k < k' \leq K} d_{k,k'} \mathbb{1}\left(\gamma^{\alpha}_{k',j} = \gamma^{\alpha}_{k,j}\right)\right\}, \tag{1.4.3}$$

where $\mathbb{1}(C)$ is equal to 1 if C is true and 0 otherwise and the constant $d_{k,k'}$ is equal to 1 if municipalities k and k' are adjacent and 0 otherwise. The hyperparameter ω_j controls the spatial smoothness and an uniform prior distribution is defined, $\omega_j \sim \text{Uniform}\,(0, \omega_{\max})$, where $\omega_{\max}$ is fixed. Note, this prior specification assumes that the vectors $\widetilde{\boldsymbol{\gamma}}^{\alpha}_{j}$ and $\widetilde{\boldsymbol{\gamma}}^{\alpha}_{j^*}$, $j, j^* = 1, \ldots, 7,\ j \neq j^*$, are statistically independent.

For the zero-truncated Poisson component, a log-linear model with Gaussian overdispersion is defined. Inference is performed on days with positive claim numbers only as observations of $N_{k,t} = 0$ are uninformative. As for the Bernoulli component, variable selection is performed via a set of binary random variables which define whether a covariate enters the model or not. Let $\boldsymbol{\gamma}^{\lambda}_{k} = (\gamma^{\lambda}_{k,1}, \ldots, \gamma^{\lambda}_{k,7})$, $k \in \{1, \ldots, 319\}$ be a vector of binary variables such that $\gamma^{\lambda}_{k,j} = 1$ if the jth covariate enters the model for $\lambda_{k,t}$ and $\gamma^{\lambda}_{k,j} = 0$ otherwise. The dependence between $\lambda_{k,t}$ and $\mathbf{X}_{k,t}$, conditional on $N_{k,t} > 0$, is formally specified as

$$\log(\lambda_{k,t}) \sim \text{Normal}\left(\beta_{k,0} + \sum_{\{j\,:\,\gamma_{k,j}=1\}} \beta_{k,j} X_{k,t,j} - \log(A_{k,t}),\ \sigma^2_k\right), \tag{1.4.4}$$

where $A_{k,t}$ is the observed number of policies. A Gaussian prior is defined on the vector of covariate effects for which $\gamma^{\lambda}_{k,j} = 1$ and its covariance structure is in the form of a g-prior (Zellner, 1986). Further, a conjugate inverse-Gamma prior is defined for σ^2_k while an improper prior is specified for the baseline risk, $\beta_{k,0} \propto 1$; for more details on the prior specification see Section 3 in Scheel et al. (2013). Geographically smoothed variable selection for each covariate is performed equivalently to the Bernoulli component via an Ising prior distribution as in (1.4.3) with hyperparameter ω^{λ}_{j}.

Due to the statistical framework formulated above, inference on the parameters for $\lambda_{k,t}$ can be performed separately from $\alpha_{k,t}$ as the two parameters are conditionally independent given

the data. In the case of the zero-truncated Poisson component, the covariate effects $\boldsymbol{\beta}_k$ and the variance σ_k^2 can be integrated out from the prior; see Scheel et al. (2011) for details. Samples from the posterior distribution for the zero-truncated Poisson component are then obtained via Gibbs sampling and an adaptive Metropolis algorithm (Roberts and Rosenthal, 2009). For the Bernoulli component, samples of the posterior distribution are drawn by a reversible jump MCMC (Green, 1995).

Results

Parameters are estimated based on the data for all years except 2001 via the MCMC algorithm outlined above. The zero-truncated Poisson component is not estimated for all municipalities as some of them do not observe any days with $N_{k,t} > 1$. Since interest lies in the detection of claim-driving weather events, analysis focuses on the posterior distributions of $\gamma_{k,j}^{\alpha}$ and $\gamma_{k,j}^{\lambda}$ which provide insight into the importance of covariate j for the claim dynamics in municipality k. Results show that the posterior means vary geographically but also between the two model components.

Only the covariates $R_{k,t},\ R_{k,t-1}$ and $D_{k,t}$ appear to be important for the Bernoulli component. Posterior mean plots indicate that $R_{k,t}$ and $R_{k,t-1}$ enter the model for $\alpha_{k,t}$ in more than 60% of the samples for most of the western coast and in south-east Norway but not for the mountainious areas in central Norway. Scheel et al. (2013) explain these differences via the vegetation and the soil absorbing water in the rural, mountainous municipalities, as opposed to the urbanized areas with asphalt-covered streets. Further, $D_{k,t}$ has a strong impact on the claim dynamics for south-east Norway where the landscape is flat and water cannot escape as easily as in the mountainous municipalities in western Norway. Figure 3 in Scheel et al. (2011) further shows that the posterior mode for $\boldsymbol{\gamma}_{k,j}^{\alpha} = \mathbf{0}$ for municipalities in central Norway where $A_{k,t}$ is relatively small, i.e. any weather covariate hardly enters the model. Compared to Haug et al. (2011), both approaches conclude that the drainage run-off is not important for western municipalities.

For the zero-truncated Poisson component, plots indicate that $R_{k,t},\ R_{k,t-1},\ S_{k,t}$ and $\Delta S_{k,t}$ are important but not $D_{k,t}$. Results for $R_{k,t}$ and $R_{k,t-1}$ are similar to the ones of the Bernoulli component. The covariates $S_{k,t}$ and $\Delta S_{k,t}$ appear important, with varying degree, across the considered municipalities. While $S_{k,t}$ enters the model for $\lambda_{k,t}$ most often along the coast, the binary factor associated to $\Delta S_{k,t}$ has highest posterior mean in Oppland and Hedmark.

Table 1.4.1: Posterior predictive median, 95% prediction interval and actual observation of the weekly-aggregated claim numbers for (a) the four weeks with the highest observations, (b) the four weeks with maximum total precipitation for the Binomial and Poisson Hurdle model with proposed covariates for Oslo and Bergen.

Period	Results for Oslo			Results for Bergen		
	Median	95% prediction interval	Truth	Median	95% prediction interval	Truth
(a)	4	(0,14)	11	3	(0,8)	7
	4	(1,11)	11	3	(0,7)	7
	3	(0,8)	8	2	(0,6)	6
	3	(0,7)	7	2	(0,7)	6
(b)	3	(0,8)	3	2	(0,7)	2
	3	(0,7)	3	2	(0,7)	2
	3	(0,7)	3	3	(0,8)	2
	3	(0,7)	3	2	(0,6)	2
(c)	5	(1,13)	5	4	(0,10)	5
	4	(0,14)	11	4	(0,12)	1
	4	(0,11)	6	3	(0,9)	3
	4	(1,11)	11	3	(0,9)	3
(d)	3	(0,8)	6	3	(0,7)	0
	3	(0,8)	3	3	(0,6)	2
	3	(0,8)	1	2	(0,7)	3
	3	(0,8)	3	3	(0,6)	3

Predictive performance is assessed on a weekly basis for the year 2001. In particular, the weekly-aggregated claim numbers are classified with respect to three types: (i) zero claims, (ii) one to three claims and (iii) more than 3 claims. The predicted type is the one with the highest posterior predictive probability. On average, the Poisson Hurdle model predicts the correct type in 89% of the cases but performance differs strongly between municipalities. While predictions are correct in more than 90% of the weeks for most rural municipalities, the success rate is about 70% for the largest cities and only 46% for Sarpsborg in south-east Norway.

Additional to the claim type prediction, the predictive performance is considered for Oslo and Bergen and the results from Scheel et al. (2013) are provided in Table 1.4.1. While the predictions are relatively good for weeks with medial claim numbers and precipitation levels, the results show a clear tendency to underpredict high claim numbers. Furthermore, the model also shows limitations in capturing the effect of rainfall since the observed claim numbers are sometimes at the upper end of the 95% prediction intervals.

1.4.3 Limitations

Both claim models by Haug et al. (2011) and Scheel et al. (2013) appear to have potential limitations in terms of capturing the dependence structure between the insurance and weather data analyzed in Section 1.3. Haug et al. (2011) estimate a Binomial model with overdispersion for $N_{k,t}$ but assess its model fit solely by comparing the observed and predicted average numbers of claims over the whole 10-year period. Such an examination does not provide much insight into the model performance for days with high claim numbers. Furthermore, the modelling framework is based upon the restrictive assumption of common covariate effects for all municipalities within a county. For instance, the county of Buskerud has a diverse topology, western areas are mountainous and rural while the east is rather flat and densely populated with the city of Drammen. The model estimates by Scheel et al. (2013) also show geographical differences in the importance of the covariate within the region; $R_{k,t}$ is found to usually enter the model for Drammen while this is hardly the case for western municipalities in Buskerud. While the approach assumes common vulnerability to the weather covariates across the county, any potential geographical structure between counties is not considered. Finally, Haug et al. (2011) allow for non-linear parametric forms of the single covariates but they do not account for a non-linear combined effect of, e.g. snow-melt and precipitation.

Scheel et al. (2013) argue that a Binomial distribution cannot handle the high frequency of zero claims which may be another potential limitation of Haug et al. (2011). They hence propose a Bayesian Poisson Hurdle (BPH) model whose parameters depend on seven covariates. However, Table 1.4.1 shows that the BPH model underperforms in weeks with high claim numbers. Consequently, the dynamics which induce high claim numbers are not well modelled although the proposed distribution accounts for the high frequency of zero claims. Consequently, the log-linear model with overdispersion and zero-truncated Poisson distribution is not flexible enough to fit the strictly positive claim numbers. Equivalently to Haug et al. (2011), any non-linear combined effects of the covariates are not considered. Both approaches also ignore potential non-continuous threshold effects. For instance, the infrastructure of a city is generally able to cope with low amounts of precipitation, i.e. the risk changes only slightly, but the claim probability may 'jump' at a certain precipitation level and then increase much stronger beyond this threshold.

1.5 Thesis Aims and Structure

This thesis aims to improve the approaches by Haug et al. (2011) and Scheel et al. (2013) and introduces new statistical models for the dependence between the number of claims and the weather data. Section 1.2 outlined that good models are required as decision makers want to take efficient actions against weather events which cause severe damages while the insurance companies want to set adequate premiums. These models are based upon the new methodology in spatial statistics, monotonic regression and extreme value theory developed in this thesis. Bayesian inference and optimization are applied to estimate the introduced models. The thesis is split into several chapters that consider different aspects and handle the limitations discussed in Section 1.4.3. It includes an introduction to the statistical areas of spatial statistics, monotonic regression and extreme value theory as well as a summary of existing research (Chapter 2).

Chapter 3 compares the performance of the Binomial and the Bayesian Poisson Hurdle model with respect to the insurance and weather data. A Bayesian hierarchical model is introduced for both approaches separately and defines a geographical dependence structure for the covariate effects. The dependence model is more flexible than the one by Haug et al. (2011), as it estimates covariate effects municipality-wise, but defines a higher degree of dependence between municipalities than Scheel et al. (2013). Equivalently to Scheel et al. (2013), the predictive performance is assessed on a weekly basis. The following chapters then introduce the new methodologies which increase the flexibility of the modelling framework in Chapter 3 and are structured in form of papers. Hence, they can be read as separate entities. As such, some aspects, e.g. the description of the insurance and weather data, are repeated.

Chapter 4 introduces *Bayesian Spatial Monotonic Multiple Regression* (BSMMR), a new methodology to perform monotonic, multiple regression for a set of contiguous regions (lattice data). The regression functions permissibly vary between regions and exhibit geographical structure. Bayesian non-parametric methodology is developed which allows for both continuous and discontinuous functional shapes and which are estimated using marked point processes and reversible jump Markov Chain Monte Carlo techniques (Green, 1995). Geographical dependence is incorporated by a flexible prior distribution; the parametrization allows the dependence to vary with functional level. The approach is tuned using Bayesian global optimization and cross-validation. Estimates enable variable selection, threshold detection and prediction as well as the extrapolation of the regression function. Performance and flexibility of the approach are

illustrated by simulation studies and an application to a subset of the Norwegian insurance data set explored in Section 1.3.

Chapter 5 introduces an alternative approach to BSMMR which also considers monotonic, multiple regression for lattice data. This methodology is motivated by the high computational cost of the BSMMR approach. Monotonic functions are estimated via optimization and can be combined with all other optimization-based approaches considering a single monotonic function. The approach is slightly less flexible than BSMMR, but also less computationally demanding. A simulation study is performed to illustrate the performance and benefits.

Chapter 6 introduces a new approach to model the dependence between weather events, e.g. rainfall or snow-melt, and the number of water-related property insurance claims. Similarly to Scheel et al. (2013), the model accounts for a spatial variation of the underlying dynamics due to differences in topology, construction designs and climate. The new model is in particular motivated by the lack of model fit for high numbers of claims and methods of extreme value theory are used. More precisely, the statistical framework is based on both mixture and extremal mixture modelling, with the latter being based on a discretized generalized Pareto distribution. Further, a temporal clustering algorithm is proposed and new covariates are derived which lead to a better understanding of the association between claims and weather events. The modelling of the claims, conditional on the locally observed weather events, both fits the marginal distributions well and captures the spatial dependence between locations. To demonstrate its benefits, the methodology is applied to the three cities of Oslo, Bergen and Bærum. The thesis concludes with a summary and discussion on future research in Chapter 7.

Chapter 2

Literature Reviews

2.1 Statistical Models for Lattice Data

2.1.1 Overview

Spatial data arise in several application areas, including meteorology (Handcock and Wallis, 1994) and public health (Wakefield, 2007). Let $D \subset \mathbb{R}^d,\ d \geq 1$, denote the spatial domain, that is, the set of locations which is considered in the analysis. Spatial statistics considers inference about a stochastic process (random field) $\mathcal{Z}$ over D which defines a random variable $\mathcal{Z}(s)$ at each spatial location $s \in D$. Cressie (1993) classifies spatial data into three categories based on the nature of D: *geostatistical data*, *point pattern data* and *lattice data*.

For geostatistical data, D is a continuous set and observations are recorded for a finite set of fixed spatial locations, $s_1, \ldots, s_n \in D$. Interest lies in the spatial interpolation of the data to predict $\mathcal{Z}$ at an unobserved spatial locations $s^* \in D$. Such data emerge, for instance, in meteorology as the temperature is recorded for a finite set of weather stations and then derived for each location via spatial interpolation. Note, the weather covariates described in Section 1.3 are generated by such an approach. Gaussian process models are widely discussed and applied in geostatistics (Kim et al., 2005; Banerjee et al., 2008); see Rasmussen and Williams (2006) for a detailed overview on Gaussian processes. Conditional on such a model, the interpolated value at a location corresponds to the mean of the estimated Gaussian process model at this point in space.

Point pattern data are also observed over a continuous spatial domain D but the spatial locations are stochastic. Each observation is a countable set of points which are a realization of the spatial stochastic process $\mathcal{Z}$. Consequently, inference for point pattern data considers the

distribution of the spatial locations while methods for geostatistical data analyze the spatial dependence between the observed values. For instance, the point patterns could be realizations from a Poisson process with non-stationary spatial density function. Point patterns arise in epidemiology and are used to estimate the spatial variation in risk of infection (Diggle, 2013). In practice, point pattern data derive from case-control studies where the exact spatial location for each individual is known (Costain, 2009). Such data then allow to detect areas which exhibit an elevated or reduced risk of infection. For more details on the statistical methodology for geostatistical and point pattern data, the reader is referred to Cressie (1993), Schabenberger and Gotway (2004) and Gelfand et al. (2010).

Lattice data refer to spatial data for which the number of locations in D is finite and examples include pixel data and areal unit data. The latter generally occur when individual observations are locally aggregated due to practicality or confidentiality concerns (Paiva et al., 2014). Note, the areal units may be irregular both in shape and size. Let K denote the number of areal units. The stochastic process $\mathcal{Z}$ then corresponds to a random vector $\mathbf{Z} = (Z_1, \ldots, Z_K)$, where $Z_k,\ k = 1, \ldots, K$ refers to the kth areal unit. In applications, an adjacency (neighbourhood) structure is generally defined either via an adjacency matrix $\mathbf{B} \in \{0,1\}^{K \times K}$ or an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ refer to vertex and edge set, respectively. An adjacency matrix $\mathbf{B}$ has zeros on its diagonal and off-diagonal entry $B_{k,k'} = 1,\ k, k' = 1, \ldots, K,\ k \neq k'$, if the areal units k and k' are adjacent, and 0 otherwise. With respect to the graph representation, each node $v \in \mathcal{V}$ corresponds to an areal unit and an edge $e \in \mathcal{E}$ between two nodes corresponds to the associated areal units being adjacent.

The Norwegian insurance and weather data explored in Section 1.3 and used throughout the thesis is 'areal' in structure and the spatial domain consists of the $K = 430$ municipalities (areal units). Therefore, the following literature review considers this data type only and summarizes the existing research. Section 2.1.2 details two classical models for lattice data: the Ising model and Gaussian Markov random fields. Section 2.1.3 describes how these models are generally applied and provides some examples from the literature, including the aspect of spatially varying regression functions.

2.1.2 Ising Model and Gaussian Markov Random Field

Interest lies in the specification of a spatial dependence model for the K-dimensional random vector $\mathbf{Z} = (Z_1, \ldots, Z_K) \in \mathbb{R}^K$. Scheel et al. (2013) apply the Ising model to perform spatially

varying variable selection. The spatial variation in the baseline risk as modelled by Haug et al. (2011), on the other hand, may be defined in terms of a Gaussian Markov random field. These two spatial models are detailed in the following:

Ising Model

Let $\mathbf{Z}$ be a vector of binary random variables which take values 0 or 1 and consider an adjacency matrix $\mathbf{B}$ which defines the spatial structure. Note, the following model can also be defined more generally for any symmetric matrix with non-negative entries. The Ising model for $\mathbf{Z}$ based on $\mathbf{B}$ then defines a joint density which allocates higher probability to outcomes of $\mathbf{Z}$ for which adjacent areal units have the same binary outcome. Formally, the joint density $\pi(\mathbf{z})$ of $\mathbf{Z}$ is, up to a normalizing constant, given as

$$\pi(\mathbf{z}) \propto \exp\left\{-\omega \sum_{\{(k,k') \,:\, B_{k,k'}=1\}} d_{k,k'} \mathbb{1}(z_k = z_{k'})\right\}, \tag{2.1.1}$$

where $d_{k,k'} \geq 0$ are prespecified constants. The parameter $\omega \geq 0$ defines the degree of spatial smoothing, that is, the probability mass function has greater mass around the outcomes $\mathbf{z} = 1$ and $\mathbf{z} = 0$ with increasing ω. Such dependence models are, for instance, considered in image analysis (Smith and Fahrmeir, 2007).

The spatial model in (2.1.1) can readily be integrated into a MCMC scheme, in particular, a Gibbs sampling scheme. Consider the update of $z_k,\ k = 1, \ldots, K$, and let $\mathcal{D}$ denote the data. Further, define $\mathbf{Z}_{-k}$ as the set of binary random variables excluding Z_k and $\mathbf{z}_{-k}$ its current values. The full conditional posterior for Z_k with likelihood function $f(\mathcal{D} \mid \mathbf{z})$ is then given as

$$\pi(z_k \mid \mathbf{z}_{-k}, \mathcal{D}) \propto f(\mathcal{D} \mid \mathbf{z})\, \pi(z_k \mid \mathbf{z}_{-k}) = f(\mathcal{D} \mid \mathbf{z}) \exp\left\{-\omega \sum_{\{k' \,:\, B_{k,k'}=1\}} d_{k,k'} \mathbb{1}(z_k = z_{k'})\right\},$$

in which the sum is taken over the areal units adjacent to area k.

While the update of the binary random variables, $\mathbf{z}$, is straightforward, the smoothing parameter ω is more complex. In order to sample ω, the normalizing constant of the density function in (2.1.1) is required. For small K, one may evaluate the sum in (2.1.1) for all possible combinations of the binary random variables and hence derive the normalizing constant exactly. However, this approach is infeasible for large K. For example, consider: the number of possible

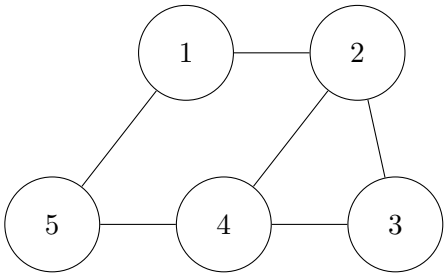


Figure 2.1.1: Example of an undirected graph with $K = 5$ nodes.

outcomes of $\mathbf{Z}$ for the insurance data is $2^{430} \approx 3 \times 10^{129}$. Since the normalizing constant is intractable for large K, approximative approaches are generally applied. For instance, Smith and Fahrmeir (2007) and Scheel et al. (2013) apply the thermodynamic integration approach by Green and Richardson (2002) which assumes $\omega > 0$ and approximations are derived for a finite set of values of ω, e.g. defined via a grid. While this approach requires additional off-line computation, Møller et al. (2006) propose the introduction of an auxiliary variable which requires no additional step. The latter also relates to importance sampling; see Møller et al. (2006) for details.

Gaussian Markov Random Fields

Prior to introducing the concept of a Gaussian Markov Random field, a more general introduction to Markov random fields is provided. Consider the K-dimensional vector $\mathbf{Z} = (Z_1, \ldots, Z_K) \in \mathbb{R}^K$ and let the dependence structure between the areal units be specified via an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as described in Section 2.1.1. Hence, the kth node corresponds to Z_k and an edge $e \in \mathcal{E}$ between two nodes corresponds to the associated areal units being adjacent. Figure 2.1.1 illustrates an example of a dependence structure for a set of $K = 5$ areal units. The random vector $\mathbf{Z}$ is then termed a *Markov random field* with respect to $\mathcal{G}$ if it satisfies certain Markov properties:

Pairwise Markov property

Any two components which are not connected via an edge $e \in \mathcal{E}$ are conditionally independent given all the other components. With respect to Figure 2.1.1, this property implies that $Z_1 \perp\!\!\!\perp Z_4 \mid (Z_2, Z_3, Z_5)$.

Local Markov property

A variable is conditionally independent of all other components given all the ones to

which an edge exists. Considering Figure 2.1.1, the local Markov property implies that $Z_1 \perp\!\!\!\perp (Z_3, Z_4) \mid (Z_2, Z_5)$.

Global Markov property

Any two subsets of variables are conditionally independent given a non-empty separating subset. For Figure 2.1.1, the global Markov property implies that $(Z_1, Z_5) \perp\!\!\!\perp Z_3 \mid (Z_2, Z_4)$.

For a Markov random field, the Hammersley-Clifford theorem provides a factorization of the joint density of $\mathbf{Z}$, $\pi_{\mathbf{Z}}(\mathbf{z})$, based on $\mathcal{G}$ and the concept of a *clique*. A subset $\widetilde{\mathcal{V}} \subseteq \mathcal{V}$ is called a clique if all vertexes in $v \in \widetilde{\mathcal{V}}$ are connected to each other. For instance in Figure 2.1.1, the subgraph $(2, 3, 4)$ is a clique while $(1, 2, 5)$ is not a clique since there exists no edge between the 2nd and 5th vertex. The Hammersley-Clifford theorem then states that for all outputs $\mathbf{z}$ of $\mathbf{Z}$ with $\pi(\mathbf{z}) > 0$, the joint density factorizes over the cliques of the graph; see Besag (1974) for a proof.

The approach in spatial statistics for lattice data is then to define a model on $\mathbf{Z}$ via a set of full conditional distributions $Z_k \mid (\mathbf{Z}_{-k} = \mathbf{z}_{-k})$, where $\mathbf{Z}_{-k}$ refers to $\mathbf{Z}$ without Z_k. Due to the local Markov property, it is sufficient to specify the full conditional for Z_k based on the adjacent areal units only. The other attractive property of such an approach is the ease of integrability into a MCMC algorithm. Similarly to the Ising model, estimates of the posterior distribution of $\mathbf{Z}$ can be sampled in a Gibbs sampling scheme as the full conditionals are easily accessible. Hence the full conditional posterior density $\pi(z \mid \mathcal{D}, \mathbf{z}_{-k})$ of $Z_k \mid (\mathcal{D}, \mathbf{z}_{-k})$ is given as

$$\pi(z_k \mid \mathcal{D}, \mathbf{z}_{-k}) \propto f(\mathcal{D} \mid z_k, \mathbf{z}_{-k})\, \pi(z_k \mid \mathbf{z}_{-k}).$$

However, not every set of full conditionals defines a valid joint distribution for $\mathbf{Z}$. In the following, the special case of $\mathbf{Z}$ being Gaussian distributed is considered and criteria on the full conditionals are specified.

Let $\mathbf{Z}$ be Gaussian distributed with positive definite precision matrix $\mathbf{Q} \in \mathbb{R}^{K \times K}$ and whose dependence structure can be represented via an undirected graph $\mathcal{G}$. Then $Q_{k,k'}$ is equal to 0 if, and only if, there exists no edge between the nodes k and k', that is, Z_k and $Z_{k'}$ are conditionally independent given all other components. Besag (1974) generally defines a set of full conditionals $Z_k \mid \mathbf{z}_{-k}$, termed conditional autoregressive (CAR) models, such that the joint density of $\mathbf{Z}$ is

Gaussian distributed. Assume that $\mathbf{Z}$ has mean 0 and let $Z_k \mid (\mathbf{Z}_{-k} = \mathbf{z}_{-k})$ be defined as

$$Z_k \mid (\mathbf{Z}_{-k} = \mathbf{z}_{-k}) \sim \text{Normal}\left(\sum_{k' \neq k} d_{k,k'} z_{k'},\ \tau_k^{-1}\right), \tag{2.1.2}$$

where $d_{k,k'} \geq 0$ is constant and $\tau_k > 0$. If $\tau_k\ d_{k,k'} = \tau_{k'}\ d_{k',k}$ for all pairs $k \neq k'$, the model specified by expression (2.1.2) results in a multivariate Gaussian distribution for $\mathbf{Z}$, $\mathbf{Z} \sim \text{MVN}\left(0, \mathbf{Q}^{-1}\right)$; subject to $\mathbf{Q}$ being positive definite. Specifically, the precision matrix $\mathbf{Q}$ has diagonal entries $Q_{k,k} = \tau_k$ and off-diagonal entries $Q_{k,k'} = -\tau_k d_{k,k'}$.

In several applications, a specification by Besag et al. (1991) which results in $\mathbf{Q}$ being positive semidefinite and not having full rank is considered. Rue and Held (2005) refer to this approach as intrinsic autoregressive (IAR) while Wakefield (2007) terms it intrinsic conditional autoregressive (ICAR). Formally, the full conditionals are defined as

$$\pi\left(z_k \mid \mathbf{z}_{-k}\right) \sim \text{Normal}\left(\frac{\sum_{k' \neq k} d_{k,k'} z_{k'}}{\sum_{k' \neq k} d_{k,k'}}, \frac{1}{\omega \sum_{k' \neq k} d_{k,k'}}\right). \tag{2.1.3}$$

Similarly to the Ising model in (2.1.1), the parameter $\omega > 0$ controls the spatial smoothness, with higher values for ω implying less spatial variation and hence higher spatial dependence. The constants $d_{k,k}$ impose a weighting of the areal units and are often adjacency-defined, that is, $d_{k,k} = 1$ if the areal units k and k' are adjacent and 0 otherwise. An alternative definition of $d_{k,k'}$ considers the distance between the centroids of the areal units.

For the conditional model specification in (2.1.3) with adjacency-based weights and $\mathcal{G}$ being a connected graph, the resulting joint distribution is improper and has probability density

$$\begin{aligned} \pi(\mathbf{z}) \ &\propto\ \omega^{\frac{K-1}{2}} \exp\left\{-\frac{\omega}{2} \sum_{1 \leq k < k' \leq K} d_{k,k'} \left(z_k - z_{k'}\right)^2\right\} \\ &=\ \omega^{\frac{K-1}{2}} \exp\left\{-\frac{\omega}{2} \mathbf{z}^\top \mathbf{Q} \mathbf{z}\right\}. \end{aligned} \tag{2.1.4}$$

Here, $\mathbf{Q}$ has diagonal entry $Q_{k,k}$ equal to the number of areal units adjacent to k, and off-diagonal entries $Q_{k,k'} = -1$ if the areal units k and k' are adjacent and 0 otherwise. This distribution has no mean since it is defined upon pairwise differences and hence the right-hand side is invariant to translation of $\mathbf{z}$ by adding a single constant to each component of $\mathbf{z}$. While Besag et al. (1991) suggest a factor of $\omega^{K/2}$ in (2.1.4), several authors agree on $\omega^{(K-1)/2}$, as advocated by Knorr-Held (2003) and Hodges et al. (2003), and which is based on the precision matrix having rank

$K-1$. If $\mathcal{G}$ is unconnected, then the factor depends on the number of unconnected subgraphs. Alternative approaches to the IAR in expression (2.1.4) are, for instance, proposed by Cressie (1993) or Leroux et al. (1999). Lee (2011) performs a comparative study and finds that all approaches produce close to unbiased estimates and that the prior model by Leroux et al. (1999) appears to be best in terms of overall performance. These conditional autoregressive models are also implemented in the `CARBayes R` package by Lee (2013).

A specified improper prior model, as in expression (2.1.4), results in a proper posterior distribution if the data allow for the estimation of an overall mean of $\mathbf{Z}$. As later described in Section 2.1.3, the overall mean of $\mathbf{Z}$ is defined as an additional parameter and a constraint is imposed on $\mathbf{Z}$ to ensure identifiability. Let $\widetilde{\mathbf{Z}} = \mathbf{Z} - \mathbb{E}(\mathbf{Z})$. The constraint then corresponds to

$$\sum_{k=1}^{K} \widetilde{Z}_k = 0. \tag{2.1.5}$$

While this condition appears restrictive, Gelfand and Sahu (1999) state that it is sufficient to replace the estimates $\widetilde{\mathbf{z}}$ by $\widetilde{\mathbf{z}} - \overline{\widetilde{\mathbf{z}}}$ after each iteration of the Gibbs sampler. Finally, a conjugate Gamma prior is generally defined for the spatial smoothing parameter ω; the reader is referred to Rue and Held (2005) for more details on Gaussian Markov random fields.

2.1.3 Statistical Models

The statistical models described in Section 2.1.2 permit incorporation of spatial dependence in a wide range of applications. In environmental epidemiology, also referred to as *disease mapping* or spatial epidemiology (Lawson, 2013), the baseline risk is typically assumed to be spatially varying across the areal units and the CAR or IAR model is then used. Assume, for example, that both the number of infected and susceptible are observed for a set of K areal units. Then the number of infected for areal unit k may be modelled by a Binomial distribution with the probability on logit scale being defined as

$$\text{logit}\,(p_k) = \beta_0 + u_k + \epsilon_k. \tag{2.1.6}$$

The parameter β_0 denotes the average baseline risk of infection over all areal units, ϵ_k is a local random effect with mean 0 and u_k is a spatial random effect. In a Bayesian modelling framework as described by Besag et al. (1991), the dependence structure for u_k is defined via an IAR model with adjacency-based weights while Gaussian priors are defined for β_0 and ϵ_k.

In order to ensure identifiability of β_0, the constraint (2.1.5) is imposed on the spatial random effects $\mathbf{u} = (u_1, \ldots, u_K)$. Alternatively to the Binomial distribution, a Poisson model may be considered since the number of infections is typically sufficiently small in comparison to the population size.

Further, additional explanatory variables are often observed and included in (2.1.6) via a linear term or more generally via a non-parametric regression function (Hughes and Haran, 2013; Neelon et al., 2013). In these cases, the regression function is usually assumed to be identical for all areal units and any spatial variation in the process is captured via differences in the observed values of the explanatory variables. In the context of disease mapping, this approach often appears reasonable since factors such as temperature affect the risk of infection independently of the spatial location. However, Bell et al. (2004) find a spatially heterogeneous effect of air pollution levels with respect to mortality across the major cities in the United States. They explain these differences with potentially city-specific differences in pollution characteristics and socioeconomic factors.

Approaches for spatial varying regression functions usually assume a linear shape of these functions. Brunsdon et al. (1998) and Fotheringham et al. (2002) introduce the concept of *geographically weighted regression* (GWR). In principle, GWR considers the locations in the data sequentially and fits a linear regression at each location using weighted least squares methodology. The weights are generally positively correlated to the spatial proximity, in particular, the most weight is given to the locations which are the closest spatially, while locations which are too distant have weight 0. A Bayesian approach based upon the CAR prior specification in expression (2.1.2) is considered, for instance, by Assunção (2003) and Congdon (2003). The former terms this approach a *geographically varying coefficient* (GVC) model. Similarly to (2.1.6), the covariate effect for each areal unit is defined as the sum of an average covariate effect and a spatial random effect, with latter being modelled via a conditional autoregressive model. Waller et al. (2007) compares the two approaches in the context of alcohol and violence data and the results indicate that they are qualitatively similar. Haug et al. (2011) and Scheel et al. (2013) consider a spatial variation in the covariate structure too and define two quite different models. Haug et al. (2011) assume no spatial variation within a Norwegian region but covariate effects are assumed to vary across regions and no spatial structure is imposed. The approach by Scheel et al. (2013) defines spatial dependence on the covariate effects via the latent binary indicator variables detailed in Section 1.4. More flexible regression shapes are considered by Congdon

(2006) who proposes a non-parametric approach based on generalized additive modelling.

The conditional autoregressive approach can also be applied in a spatio-temporal modelling framework. Similarly to the modelling of a spatial random effect as in expression (2.1.6) via, for instance, an IAR model, a temporal random effect is introduced for each areal unit. Potential temporal dependence between time points is again incorporated via a conditional autoregressive model. Knorr-Held (2000) introduces a second independently distributed random effect, resulting in a total of 4 random effects for each observation. Such spatio-temporal models are also applied by Mercer et al. (2015) and Lee and Lawson (2016); for more details on spatio-temporal models see Cressie and Wikle (2015).

2.2 Monotonic Regression

2.2.1 Overview

Monotonic regression, also termed *isotonic regression*, considers the estimation of an unknown regression function $\lambda : \mathcal{X} \subseteq \mathbb{R}^m \to \mathcal{Y} \subseteq \mathbb{R}$ subject to the constraint of monotonicity. More precisely, the constraint states that an ordering in the input set is preserved or reversed in the output set and is defined here in terms of the Euclidean ordering $\preceq$. Formally, the constraint

$$\mathbf{u} \preceq \mathbf{v} \;\Rightarrow\; \lambda(\mathbf{u}) \leq \lambda(\mathbf{v}),\; \mathbf{u}, \mathbf{v} \in \mathcal{X}, \tag{2.2.1}$$

implies that λ is monotonically non-decreasing, called *isotonic*. The notation $u \preceq v$ for two vectors $\mathbf{u}, \mathbf{v} \in \mathcal{X}$ corresponds to $\mathbf{u} \leq \mathbf{v}$ component-wise. Each monotonic function can be transformed to an isotonic one by reversing some or all of the coordinate axis. Hence, it is sufficient to consider settings with constraint (2.2.1) in the following. Note, the term *monotone (isotone) regression* generally refers to the special case $m = 1$.

Shape constraints such as (2.2.1) are imposed in several applications in which linearity is too restrictive. The monotonicity assumption is, for instance, imposed on dose-response relationships in the pharmaceutical industry. Dose-response curves are generally monotone and model the response, or effect, as a function of the concentration, or dose. Other application areas are considered by Royston (2000), Hyndman and Ullah (2007), Farah et al. (2013) and Wilson et al. (2014).

In the following, interest lies in obtaining an estimate $\widehat{\lambda}$ for λ based on a set of observations

$\mathcal{D} = \{(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathcal{X} \ : \ i = 1, \ldots, n\}$, subject to the monotonicity constraint. Such problems were originally considered by Ayer et al. (1955) and Brunk (1955) in the context of parameter estimation. Several approaches have since been developed in both Operations Research and Statistics, and these are generally semi- or non-parametric. Most methods impose one or more additional assumptions such as continuity, smoothness or boundedness of λ.

The following sections consider these approaches in detail and summarize the existing research on monotonic regression. Methods for monotonic regression which are based on (i) optimization, (ii) generalized additive models, (iii) Bayesian non-parametrics and (iv) functional data analysis are detailed, respectively, in Sections 2.2.2 to 2.2.4. The section concludes with a summary, including benefits and limitations in Section 2.2.5.

2.2.2 Optimization

Monotonic Regression can be formalized via an optimization problem with constraints as defined in (2.2.1). In this context, λ is estimated locally at the observed points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $\widehat{y}_i$ denote the functional level of the estimated function $\widehat{\lambda}$ at $\mathbf{x}_i$, $\widehat{y}_i = \widehat{\lambda}(\mathbf{x}_i)$. The aim is to find values $\widehat{y}_1, \ldots, \widehat{y}_n$ which minimize the residuals $|y_i - \widehat{y}_i|$, $i = 1, \ldots, n$. Formally, the optimization problem is defined as

$$\begin{aligned} &\min_{\widehat{y}_1, \ldots, \widehat{y}_n} \left[\sum_{i=1}^{n} \ w_i \ |y_i - \widehat{y}_i|^p \right]^{\frac{1}{p}}, \ 1 \leq p \leq \infty, \\ &\text{subject to } \mathbf{x}_i \preceq \mathbf{x}_j \ \Rightarrow \ \widehat{y}_i \leq \widehat{y}_j, \ \forall i, j \in \{1, \ldots, n\}, \end{aligned} \tag{2.2.2}$$

where the objective function is the L_p norm of the vector of residuals with specified constants $w_1, \ldots, w_n \geq 0$. The ordering induced by the constraints in (2.2.2) on the data points in $\mathcal{D}$ can be represented via a directed acyclic graph (DAG), $G = (V, E)$, where the vertex set V represents the n observations and an edge $e \in E$ corresponds to a constraint between two observations. Optimization problems of the form (2.2.2) are, for instance, considered in supply chain management (Maxwell and Muckstadt, 1985), medicine (Schell and Singh, 1997), biology (Obozinski et al., 2008), genetics (Luss et al., 2012) and economics (Keshvari and Kuosmanen, 2013).

The optimization problem in (2.2.2) is generally tractable. If $1 < p < \infty$, the objective function is a sum of strictly convex functions which is also strictly convex. Since the space of potential solutions $(\widehat{y}_1, \ldots, \widehat{y}_n)$ is convex too, (2.2.2) has an unique solution for $1 < p < \infty$. For

the remaining cases $p = 1$ and $p = \infty$, there exists one, not necessarily unique, solution (Khattree et al., 1999; Stout, 2012). Note, most theoretical results hold for any strictly convex loss function $\Phi(\widehat{y}_1, \ldots, \widehat{y}_n)$ and are not limited to the L_p norm in (2.2.2). Furthermore, some publications in the optimization literature examine variations or extensions of 2.2.2. For instance, Liu and Ubhaya (1997) impose the additional constraint of each estimate $\widehat{y}_i,\ i \in \{1, \ldots, n\}$, being integer while Sasabuchi et al. (1983) and Sasabuchi et al. (1992) consider a multivariate analogue.

The solution to (2.2.2) is of piecewise-constant form (Barlow and Brunk, 1972) and can be derived using *Convex Optimization* in the case $p > 1$; see Boyd and Vandenberghe (2004) for a detailed overview on convex optimization. In practice, algorithms split $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ into a set of disjoint blocks $B_1, \ldots, B_J$ with all data points in one block having the same functional level: $i, i' \in B_j \ \Rightarrow\ \widehat{y}_i = \widehat{y}_{i'}$. For instance for $p = 2$, the functional level of block $B_j,\ j = 1, \ldots, J$ corresponds to the weighted average of the points y_i with $i \in B_j$ since this minimizes (2.2.2). Multiple approaches to derive $B_1, \ldots, B_J$ are available in the literature; recent developments include the work by Stout (2015) and Kyng et al. (2015). Three iterative algorithms, (i) *Pool Adjacent Violators*, (ii) *Minimum Lower Set* and (iii) *Isotonic Recursive Partitioning*, are outlined in the following:

Pool Adjacent Violators Algorithm (PAVA)

The algorithm is introduced by Ayer et al. (1955) and Miles (1959) and is restricted to the estimation of univariate monotonic (monotone) functions, i.e. $m = 1$. Hence, the Euclidian ordering induces a total order on the input space $\mathcal{X}$ and the observations $\mathbf{x}_1$ to $\mathbf{x}_n$ can be ordered. For notational simplicity, let $\mathbf{x}_i < \mathbf{x}_{i+1},\ \forall i \in \{1, \ldots, n-1\}$ which implies $\widehat{y}_i \leq \widehat{y}_{i+1}$.

Initially, all n data points are considered as individual blocks, $B_j = \{j\},\ j = 1, \ldots, n$, with assigned level $\widehat{y}_i = y_i$. Since this setup violates, in general, the monotonicity constraint, adjacent blocks are pooled until the monotonic constraint is fulfilled. More precisely, the algorithm executes the following operations recursively:

1. If $\widehat{y}_i > \widehat{y}_{i+1}$ for any $i \in \{1, \ldots, n-1\}$, pool the blocks containing $\widehat{y}_i$ and $\widehat{y}_{i+1}$.

2. Derive new functional level $\widehat{y}_i = \widehat{y}_{i+1}$ for of all points in the pooled block as the one which minimizes the objective function in (2.2.2).

3. Stop if $\widehat{y}_i \leq \widehat{y}_{i+1},\ \forall i \in \{1, \ldots, n-1\}$, else continue with Step 1.

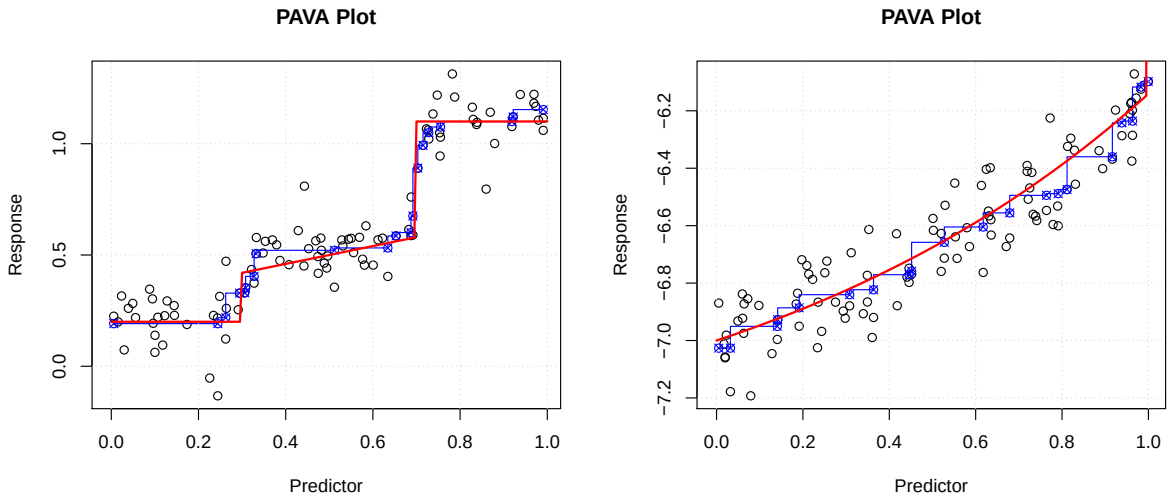


Figure 2.2.1: Function estimates obtained by the PAVA in the `isotone` package for two data sets with 100 Normally distributed observations, $y_i|x_i \sim \text{Normal}(\lambda(x_i), 0.1)$ and uniformly distributed input values x_i, $x_i \sim \text{Uniform}(0,1)$, $i = 1, \ldots, 100$. The values for the optimization are set to $p = 2$ and $w_i = 1$, $i = 1, \ldots, 100$. The plots provide (——) the true underlying function λ, (——) the estimate $\widehat{\lambda}$ and the sampled observations.

De Leeuw et al. (2009) introduce the `R` package `isotone` which provides functions for the PAVA. Figure 2.2.1 shows that the PAVA yields reasonable results for both smooth and discontinuous monotone functions. The algorithm is an active set method (Best and Chakravarti, 1990; Best et al., 2000) and several modifications have been proposed. For instance, Yeganova and Wilbur (2009) modify the PAVA in order to obtain a continuous estimate for $\widehat{\lambda}$, based upon λ fulfilling the Lipschitz condition. For more details on the algorithm see Robertson et al. (1988) and De Leeuw et al. (2009).

Minimum Lower Set Algorithm (MLSA)

The approach is based on Brunk (1955) and Brunk et al. (1957) and uses the concept of lower sets. A set $L \subseteq B$ is a lower set if it is downward closed, i.e. $a \in L,\ b \in B,\ b \preceq a \ \Rightarrow\ b \in L$. In contrast to the PAVA, the MLSA only requires a partial order on $\mathcal{X}$, i.e. the algorithm is applicable for any dimension $m \geq 1$ of the input space $\mathcal{X}$.

All observations are initially in one block B_1, i.e. $\widehat{y}_1 = \cdots = \widehat{y}_n$, with functional level set to the one which minimizes the objective function in (2.2.2). The algorithm then recursively subtracts lower sets $L_1, L_2, \ldots$ from the initial set B_1 as follows:

1. Derive lower set $L_j \subseteq B_j$ which provides the lowest functional level.

2. Set $B_{j+1} = B_j \backslash L_j$.

Estimates $\widehat{y}_1, \ldots, \widehat{y}_n$ are then derived via the lower sets $L_1, L_2, \ldots$. Modifications of the MLSA for more general optimization problems than (2.2.2) are proposed by Robertson and Wright (1980) and Qian (1992).

Isotonic Recursive Partitioning (IRP)

The principle idea for IRP is introduced by Maxwell and Muckstadt (1985) and later considered by Spouge et al. (2003) and Luss et al. (2012). For notational simplicity, let $p = 2$ and set the constants in (2.2.2) as $w_1 = \ldots = w_n = 1$. Equivalently to the MLSA, all data points are initially in one block B_1 with functional levels $\widehat{y}_1^{(1)} = \cdots = \widehat{y}_n^{(1)} = \overline{y}$. An optimal partition for the initial block B_1 is derived by solving the best cut problem

$$\max_{L} \left\{ \sum_{i \in B_1 \backslash L} (y_i - \overline{y}) - \sum_{i \in L} (y_i - \overline{y}) \right\}, \tag{2.2.3}$$

where L is a lower set. In the next iteration, the algorithm solves the optimization problem (2.2.3) for both blocks L and $B_1 \setminus L$ separately. This is required since the optimal lower set L in (2.2.3) is not necessarily identical to the optimal minimum lower set in the MLSA. Best cut problems of type (2.2.3) are then recursively solved until no further partition is optimal, i.e. the lower set L is empty or the whole set. The difference between IRP and MLSA can also be described in terms of the aforementioned DAG representation. While the MLSA separates a connected subgraph from the original graph, IRP splits the graph into two potentially disconnected graphs in each iteration. The advantage of IRP is that the optimization problem in (2.2.3) can be efficiently solved by linear programming; for more details see Luss et al. (2012). The generalization of the IRP approach is derived in Luss and Rosset (2014).

2.2.3 Generalized Additive Models

We provide a short introduction to the generalized additive modelling (GAM) framework prior to its consideration in the context of monotonic regression. Hastie and Tibshirani (1986) introduce GAM by combing generalized linear and additive models, the latter are established by Friedman and Stuetzle (1981) as a tool for non-parametric regression. More precisely, the linear predictor is replaced by a sum of smooth functions $f_1, f_2, \ldots$, i.e. a higher-dimensional function is expressed

in terms of lower-dimensional functions. Formally, a general GAM is of the form

$$g\left[\mathbb{E}(Y)\right] = \beta_0 + f_1(x_1) + \cdots + f_m(x_m) + f_{m+1}(x_1, x_2) + \cdots, \tag{2.2.4}$$

where Y is the response variable with a distribution function from the exponential family. Equivalently to the generalized linear case, the link function g is invertible; for theoretical details see Hastie and Tibshirani (1990). GAM frameworks as in (2.2.4) are widely applied, e.g. in medicine (Hawn et al., 2013) and risk analysis (Chavez-Demoulin et al., 2016).

The smooth functions $f_1, f_2, \ldots$ can be estimated from the data $\mathcal{D}$ non-parametrically via the backfitting algorithm, Hastie and Tibshirani (1986) refer to it as the local scoring algorithm. In order to ensure identifiability, it is often assumed that

$$\sum_{i=1}^{n} f_j(\mathbf{x}_{i,j}) = 0, \ \forall j = 1, 2, \ldots, m, m+1, \ldots,$$

where $\mathbf{x}_{i,j}, \ j = 1, \ldots, m$ denotes the jth element of the input vector $\mathbf{x}_i$. Alternative methods are usually semi-parametric and these are based on penalized regression splines (Eilers and Marx, 1996; Ruppert, 2002; Wood, 2006). An approach for large data sets is introduced by Wood et al. (2015).

Monotonic regression in an additive setting is introduced by Cunningham (1982) and termed *additive monotonic (isotonic) modeling*. The framework postulates $f_1, f_2 \ldots$ in (2.2.4) to be monotonic instead of smooth. Most publications on additive monotonic isotonic models consider a model with univariate monotonic functions only, i.e. $f_j = 0, \ \forall j > m$. Then, similarly to (2.2.2), the estimation problem for $f_1, \ldots, f_m$ can be formulated via a convex optimization problem of the form

$$\begin{aligned} \min_{\widehat{f}_1, \ldots, \widehat{f}_m} \ & \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{m} \widehat{f}_i(x_{i,j}) \right]^p, \\ & \text{subject to } \widehat{f}_1, \ldots, \widehat{f}_m \text{ being isotonic,} \end{aligned} \tag{2.2.5}$$

but the fit may also be defined in terms of the likelihood.

Bacchetti (1989) proposes the non-parametric cyclic pool adjacent violators (CPAV) algorithm to obtain the estimates $\widehat{f}_1, \ldots, \widehat{f}_m$ in (2.2.5). The CPAV algorithm combines the PAVA (Section 2.2.2) and the alternating conditional expectations algorithm (Breiman and Friedman, 1985) in order to obtain estimates $\widehat{f}_1, \ldots, \widehat{f}_m$ as follows:

1. Set initial values $\widehat{f}_j(x_{i,j}) = \widehat{y}_{i,j},\ i \in \{1,\dots,n\},\ j \in \{1,\dots,m\}$.

2. Optimize the fit with respect to $\widehat{f}_1$ via the PAVA while keeping $\widehat{f}_2,\dots,\widehat{f}_m$ fixed. Then optimize $\widehat{f}_2$ subject to $\widehat{f}_1,\widehat{f}_3,\dots,\widehat{f}_m$ being fixed, followed by $\widehat{f}_3$ and so on. The initial setup for each update of $\widehat{f}_j$ is as in the PAVA: $\widehat{f}_j(x_{i,j}) = \widehat{y}_{i,j}\ i \in \{1,\dots,n\}$.

3. Apply Step 2 until convergence is reached.

Morton-Jones et al. (2000) extend the CPAV algorithm to settings with additional linear predictors. The authors apply their algorithm in an epidemiological framework in order to model the relationship between risk and exposure. Such regression problems with both linear and smooth functions are termed *semi-parametric additive monotonic* and are widely studied; see Cheng (2009); Cheng et al. (2012); Rueda (2013); Yu (2014); Chen and Samworth (2016) and references therein.

A substantially different class of approaches considers the estimation of $f_1,\dots,f_m$ via monotonic regression splines. Hence, the estimates $\widehat{f}_1,\dots,\widehat{f}_m$ in (2.2.5) are constrained to be both monotonic and continuous (smooth). The additional constraint of continuity is plausible in some applications, e.g. for the estimation of dose-response curves. In such cases, the application of regression splines may be preferred to the CPAV algorithm which estimates $f_1,\dots,f_m$ in form of step functions. Ramsay (1988) proposes to fit monotone regression splines using integrated splines (I-splines) which are derived using M-splines (Curry and Schoenberg, 1966). The modelling approach is applied in several areas, e.g. toxicology (De Boer et al., 2002), and is used in a setting with multiple covariates by Tutz and Leitenstorfer (2007).

Kelly and Rice (1990) and He and Shi (1998) consider monotone regression using B-splines with each estimate $\widehat{f}_j,\ j = 1,\dots,m$ being of the form

$$\widehat{f}_j = \sum_{s=1}^{S} \alpha_{j,s} B_{j,s}(x), \tag{2.2.6}$$

subject to $\alpha_{j,s}\ \forall\ j \in \{1,\dots,m\},\ s \in \{1,\dots,S\}$ preserving monotonicity. A similar setting to (2.2.6) is also considered by Wang and Small (2015) and Wang and Xue (2015), and applied by Leitenstorfer and Tutz (2007). While these approaches estimate $f_1,\dots,f_m$ using optimization methods, Neelon and Dunson (2004) and Cai and Dunson (2007) propose a Bayesian setting for single and multiple outcomes, respectively.

In recent publications, additive monotonic regression models are also considered for high-

dimensional problems with the number of covariates, m, exceeding the number of observations, n. In a linear regression setting, the *least absolute shrinkage and selection operator* (LASSO) (Tibshirani, 1996) is a well-known technique to minimize the number of non-zero covariate effects. The approach by Fang and Meinshausen (2012) is motivated by the LASSO and extends (2.2.5) by adding a penalty for high variations of $\widehat{f}_1, \ldots, \widehat{f}_m$. Estimates $\widehat{f}_1, \ldots, \widehat{f}_m$ are then obtained via a modified CPAV algorithm and Fang and Meinshausen (2012) term their method the LASSO Isotone (LISO). While LISO yields step-wise constant functions $\widehat{f}_1, \ldots, \widehat{f}_m$, Bergersen et al. (2014) propose an alternative method which derives estimates $\widehat{f}_1, \ldots, \widehat{f}_m$ using I-splines.

2.2.4 Bayesian Nonparametrics

Bayesian nonparametric methods typically refer to Bayesian models with a very large or infinite number of parameters. Several monotonic regression approaches in Bayesian nonparametrics are based on the Dirichlet process by Ferguson (1973). A Dirichlet process is defined by a base distribution F_0 and a concentration parameter α, $\mathrm{DP}(F_0, \alpha)$, and its realizations are probability distributions. The sampled distributions are almost surely discrete even though F_0 may be continuous. Dirichlet processes are widely applied in Bayesian nonparametrics (Antoniak, 1974; Jain and Neal, 2004; Kim et al., 2006; Zhou et al., 2012), e.g. for clustering, infinite mixture models and hidden Markov models (Hjort et al., 2010). Several extensions are proposed in the literature, e.g. Teh et al. (2006) and Rodríguez et al. (2008) introduce a hierarchical and nested modelling framework, respectively. Gelfand et al. (2005) and Duan et al. (2007) define spatial Dirichlet process models from a geostatistical perspective.

In the context of monotonic regression, several authors propose a Dirichlet process based approach. Gelfand and Kuo (1991) consider the estimation of monotone dose-response or potency curves and propose an ordered Dirichlet process prior. The authors further propose a second prior which, in contrast to the Dirichlet process prior, is conjugate and based on a product of Beta distributions. Additionally, Gelfand and Kuo (1991) propose an inference procedure for two functions λ_1, λ_2 with $\lambda_1(\mathbf{x}) \leq \lambda_2(\mathbf{x}),\ \forall\ \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}$. Lavine and Mockus (1995) and Kottas et al. (2002) extend this approach, e.g. the former incorporates the estimation of an unknown error density via a second Dirichlet process. The models considered in these three publications are, however, slightly limited since the estimated functions are strictly increasing. Therefore, Dunson (2005) introduce a mixture prior which allows for flat areas of the unknown regression function.

A different set of approaches consider monotonic regression under the constraint of λ being differentiable. Hence, λ can be formally represented by

$$\lambda(x) = \beta_0 + \beta_1 \int_0^x f(u)\,\mathrm{d}u, \tag{2.2.7}$$

where $f(x)$ is non-negative, $\beta_0 \in \mathbb{R}$ and $\beta_1 \geq 0$. Ramsay (1998) consider the case of f in (2.2.7) being the exponential of an unconstrained function g. Ramsay and Silverman (2005) then apply this approach to model the tibia data collected by Hermanussen et al. (1998). While Ramsay and Silverman (2005) estimate g via B-splines, Shively et al. (2009) define a Wiener process as prior distribution and estimate f via a MCMC algorithm. Similarly to Gelfand and Kuo (1991), the estimated functions of the form (2.2.7) are strictly monotone. Bornkamp and Ickstadt (2009) consider a setting as in (2.2.7) and set f as a probability density function of a continuous bounded random variable. The distribution function associated to f is then modelled via a mixture of parametric probability distributions, with a general discrete random measure as a prior distribution.

Many other approaches have been developed and are briefly mentioned here. Shively et al. (2009) propose, additionally to a framework as in (2.2.7), a fixed-knot regression spline approach, allowing the first derivative, $\lambda'(x)$, to take the value 0. Shively et al. (2011) then extend this spline approach and introduce a free-knot regression spline model. Riihimäki and Vehtari (2010) estimate monotonic functions using Gaussian processes and results indicate a good performance for smooth surfaces. Lin and Dunson (2014) introduce a monotonic regression method which uses projections of Gaussian processes and optimization based approaches, e.g. the PAVA in Section 2.2.2. Finally, Bornkamp et al. (2010) and Wang and Dunson (2011) consider monotonicity in the wider framework of density functions.

In this work, we focus on the approaches by Holmes and Heard (2003) and Saarela and Arjas (2011), the latter extending the monotone modelling framework of the former to higher dimensions. The principal idea lies in defining the monotone (monotonic) function via a marked point process with constraints which ensure monotonicity. Points are then iteratively added, deleted or shifted using the reversible jump MCMC (RJMCMC) algorithm by Green (1995) which allows to traverse models of varying dimension. Each sampled marked point process then corresponds to a piecewise constant function via imposition of a monotonic relation on the functional space. In conclusion, this method is a Bayesian analogue to the optimization based

approaches in 2.2.2 in the sense that the estimated functions are piecewise constant. Specific details on the marked point process formulation and the RJMCMC algorithm are omitted here since they are detailed in Section 4.2.

2.2.5 Summary

The publications mentioned in the previous Sections 2.2.2 through 2.2.4 illustrate the variety of methodologies and applications in which monotonicity of the underlying regression function λ is considered. Approaches have been proposed to obtain an estimate $\widehat{\lambda}$ of either continuous or non-continuous shape. Conditional on the application, smooth or discontinuous functions may be preferred. While the optimization based approaches in 2.2.2 generally lead to a piecewise constant estimate, smooth estimates for λ are obtained via the monotonic regression splines in Sections 2.2.3 and 2.2.4.

Monotonic regression has two very positive aspects. Firstly, it provides a valuable alternative if multiple linear regression leads to a poor model fit. Since each linear model with non-negative covariate effects fulfills the monotonic constraint in expression (2.2.1), it can be replaced by a monotonic model without the requirement of additional assumptions. Secondly, the assumption of monotonicity can be tested using the methods by Bowman et al. (1998) and Ghosal et al. (2000). Furthermore, Scott et al. (2015) introduce a test for monotonicity based on Bayesian nonparametric statistics. From an applied perspective, aspects of monotonicity may also be discussed with operators and engineers. If, however, there exists uncertainty on the monotonicity of λ, Tibshirani et al. (2011) introduce a modelling framework which penalizes non-monotonicity but does not preclude it.

Nevertheless, there exist some limitations to the monotonic regression framework. Most importantly, extrapolation of the regression function is hardly possible since functions are generally estimated locally and non-parametrically. This issue does, however, not solely occur in monotonic regression framework but exists for flexible non-parametric regression models in general. Furthermore, the constraint of monotonicity is generally less restrictive for higher dimensions and leads to overfitting, unless λ is expressed in terms of an additive model. Consider two independent samples, $\mathbf{u}, \mathbf{v}$ from the input space $\mathcal{X}$, $\mathbf{u}, \mathbf{v} \sim Z_{\mathcal{X}}$, where $Z_{\mathcal{X}}$ refers to the distribution of the covariates. The monotonic constraint (2.2.1) only applies if $\mathbf{u} \leq \mathbf{v}$ component-wise or vice versa. It can be easily verified that for a given dimension m, the probability for a monotonic relationship between two covariate observations $\mathbf{u}, \mathbf{v} \in \mathcal{X}$ is given by $\left(\frac{1}{2}\right)^{m-1}$. Therefore, the

number of points, n, required to obtain a sufficiently good fit increases exponentially with the dimension of the covariate space.

2.3 Extreme Value Theory

2.3.1 Overview

The modelling of rare (extreme) events is of general interest in several application areas, for instance, in sports analysis (Stephenson and Tawn, 2013), to assess equity or storm risks (Embrechts et al., 2013; Economou et al., 2014) and to model river flows (Asadi et al., 2015). Focus lies then on the lower and upper tails of the distribution. In order to predict the occurrence rate of future extremes, there is also the desire to explore the tail behaviour beyond the observed minima and maxima, i.e. extrapolation of the tails is required. However, model estimates are mainly driven by the body rather than the tails since data are concentrated towards the centre of the distribution. Furthermore, different models that fit the body well can have very different extrapolations. The statistical area of extreme value theory provides an asymptotically justified modelling framework to estimate the tail behaviour of an unknown distribution.

This section provides an overview of the extreme value methodology for univariate random variables used in this thesis. Sections 2.3.2 and 2.3.3 detail the extremal modelling of the block maxima and the exceedances above a threshold, respectively, for a continuous random variable. Section 2.3.4 then considers the research in extreme value theory for discrete (integer-valued) random variables. The section concludes with a summary of extreme-value mixture models in Section 2.3.5.

2.3.2 Block Maxima Approach

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed (IID) continuous random variables with distribution function F. Define further the random variable M_n as the maximum of $X_1, \ldots, X_n$, i.e.

$$M_n = \max(X_1, \ldots, X_n). \tag{2.3.1}$$

Although interest lies in the modelling of both the lower and upper tail, the following theory focuses on M_n in (2.3.1) and hence on the upper tail. This is not restrictive due to the relation

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n).$$

If F is known, the distribution of M_n can be derived exactly as

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x) \\ &= [F(x)]^n .\end{aligned}$$

The result above is usually not very helpful since F is generally unknown in applications. Instead the asymptotic behaviour of F^n as $n \to \infty$ is studied. Note, M_n converges in probability to the upper end point of F, x^F, as $n \to \infty$; the asymptotic distribution of M_n is degenerate. If there exist sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$, such that, as $n \to \infty$

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x), \tag{2.3.2}$$

for some non-degenerate distribution G, then G belongs to the family of extreme value distributions. The Extremal Types Theorem (ETT) by Leadbetter et al. (1983) states that G takes one of three limiting distributions (Gumbel, Fréchet and Negative Weibull). For instance, if $X_1, \ldots, X_n$ are Normally distributed, G results in the Gumbel distribution, or in other words, the Normal distribution function lies in the max-domain of attraction of the Gumbel distribution. A detailed proof of the ETT is provided in Leadbetter et al. (1983). Note, the ETT does neither guarantee the existence of a non-degenerate limit G nor does it imply the type of limiting distribution. Furthermore, the Gumbel, Fréchet and Negative Weibull are the only distributions which satisfy the property of max-stability, that is, the maximum of a set of IID random variables has, up to type, the same distribution as each individual sample. More formally in terms of the distribution function G, there exist constants $A_n > 0$ and B_n for every $n > 0$ such that

$$G(A_n x + B_n) = [G(x)]^n .$$

A parametrization which unifies the Gumbel, Fréchet and Negative Weibull distributions is commonly used instead of the three distinct types in the ETT. The generalized extreme value (GEV) distribution, $\mathrm{GEV}(\mu, \sigma, \xi)$, has cumulative distribution function

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, \tag{2.3.3}$$

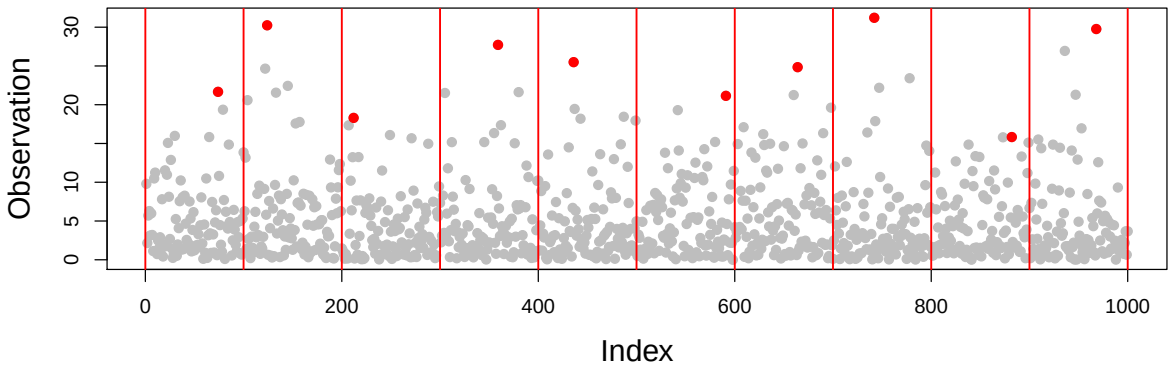


Figure 2.3.1: Example of a partition of a data set with 1,000 observations and a block size $n = 100$. The block maxima are highlighted in red.

where $y_+ = \max\{y, 0\}$, $\sigma > 0$ and $\mu, \xi \in \mathbb{R}$. The three parameters (μ, σ, ξ) are termed the location, scale and and shape parameters, respectively. The tail behaviour and limiting distribution are determined via ξ, where

- $\xi > 0$ corresponds to a Fréchet distribution which has a heavy upper tail.
- $\xi = 0$ corresponds to the Gumbel distribution which has an exponential tail.
- $\xi < 0$ corresponds to the Negative Weibull distribution which has a finite upper limit.

The GEV distribution in (2.3.3) is used to model block maxima based on observed data $\mathcal{D}$. In order to apply the asymptotic theory outlined above, it is assumed that the left-hand side in expression (2.3.2) is approximately GEV distributed for some finite value of n. The appropriateness of this approach then relies on the choice of a sufficiently high block size n. If n is too low, the asymptotic argument may not apply while a higher value for n leads to less observations for M_n and a higher variance of the estimated parameters. In other words, we aim to find the best trade-off between the block size n and the number of observed blocks. Figure 2.3.1 illustrates a potential partition for a simulated data set. Estimates for the parameters (μ, σ, ξ) are then obtained via likelihood or Bayesian inference. The reader is referred to Coles (2001) for more details on the GEV.

2.3.3 Threshold Exceedance Approach

The block maxima approach can lead to inefficient statistical procedures for the modelling of extreme values. More precisely, any statistical information from other extremal observations,

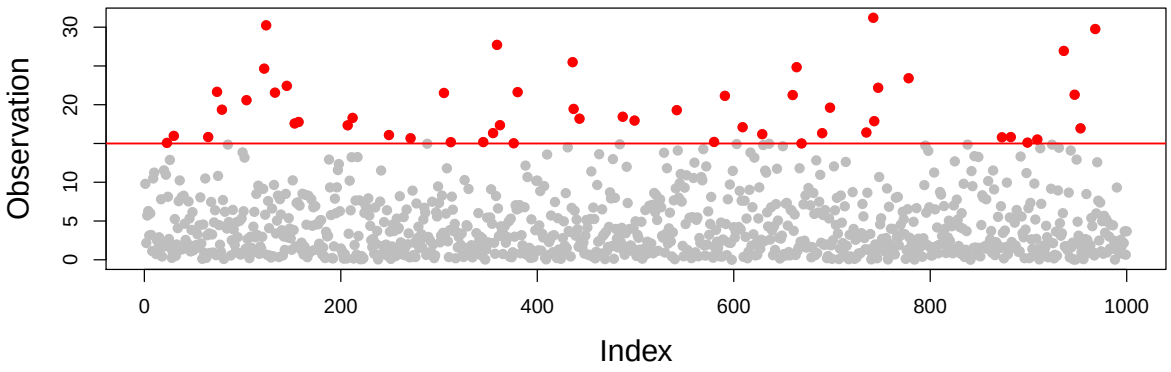


Figure 2.3.2: A sample of 1,000 simulated data points with the observations exceeding the threshold $u = 15$ being highlighted in red.

apart from the block maxima, are discarded. This motivates the derivation of extreme value models which incorporate all observations which are large enough to be called extreme. In contrast to the block maxima setting, the approach detailed in the following is based on all observations which exceed a sufficiently high threshold u and does not require the data to be split into blocks. Figure 2.3.2 illustrates this aspect for the simulated data from Figure 2.3.1.

Consider a set of IID random variables $X_1, \ldots, X_n$ whose distribution function F lies in the domain of attraction of a GEV distribution with shape parameter $\xi \in \mathbb{R}$. Let further a_n and b_n denote the normalizing constants as in (2.3.2). A sequence of point processes $P_1, P_2, \ldots$ on $[0,1] \times \mathbb{R}$ is then constructed by

$$P_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) \; : \; i = 1, \ldots, n \right\} \tag{2.3.4}$$

and the limit process as $n \to \infty$ is examined. The limit process is non-degenerate since $(M_n - b_n)/a_n$ is non-degenerate. Smaller points are normalized to the same value b_l as $n \to \infty$ while larger points are retained in the limit process. Under the construction in (2.3.4), the point process P_n converges to a non-homogeneous Poisson process P on $[0,1] \times (b_l, \infty)$ whose intensity function is determined by ξ.

The asymptotic description of the limit Poisson process motivates the modelling of large values. Define $u_n(v) = a_n v + b_n$, where $v > b_l$, and note that $u_n(v)$ tends to the upper endpoint x^F of the distribution. Pickands (1975) shows that for $X \sim F$ and $x > 0$, the Poisson process

limit P implies that

$$\mathbb{P}\left(X > u_n(v) + a_n x \mid X > u_n(v)\right) \to \left[1 + \xi \frac{x}{\sigma_v}\right]_+^{-\frac{1}{\xi}}, \tag{2.3.5}$$

where $\sigma_v = 1 + \xi v$. The right-hand side in (2.3.5) corresponds to the survival function of a generalized Pareto distribution (GPD) with scale $\sigma_v > 0$ and shape ξ. Consequently, the limiting distribution for excesses above a threshold converges to a GPD as the threshold tends to x^F.

Based upon the assumption that the limit in (2.3.5) holds exactly for a sufficiently high threshold u and by absorbing a_n into the scale parameter, for any $x > u$

$$\mathbb{P}\left(X > x \mid X > u\right) = \left[1 + \xi \frac{(x-u)}{\sigma_u}\right]_+^{-\frac{1}{\xi}}, \tag{2.3.6}$$

i.e. $(X - u) \mid (X > u) \sim \mathrm{GPD}\left(\sigma_u, \xi\right)$. Note, the shape parameter ξ in (2.3.6) is equal to the one in (2.3.3). Hence, ξ characterizes the tail behaviour of the GPD in the same way as for the GEV distribution, e.g., $\xi < 0$ implies that the GPD is short-tailed with a finite upper end point.

To apply the threshold exceedance approach, it is necessary to set a sufficiently high threshold u. If u is too low, the GPD may not fit well since the asymptotic argument may not apply. Conversely, the amount of data for inference shrinks with increasing u and a too high threshold leads to a high variance in the parameter estimates. Several approaches to select u are proposed in the literature and the two most commonly used selection diagnostics are outlined here: the threshold stability plot and the mean residual life (MRL) plot.

The first diagnostic is based on the threshold-stability property of the GPD, that is, if $(X - u) \mid (X > u) \sim \mathrm{GPD}\left(\sigma_u, \xi\right)$, then for any higher threshold $v > u$

$$(X - v) \mid (X > v) \sim \mathrm{GPD}\left(\sigma_u + \xi(v - u), \xi\right). \tag{2.3.7}$$

Hence, the shape parameter ξ is constant with increasing threshold but the scale parameter is not (Davison and Smith, 1990). In order to assess parameter stability in the scale parameter too, the modified scale $\sigma^* = \sigma_v - \xi v$ is considered instead of σ_v since it is threshold invariant. Threshold diagnostics is then based upon the examination of the threshold-stability plots, that is, the behaviour of σ^* and ξ in dependence on u. The threshold u is then chosen as the smallest value for which both parameters remain constant (excluding variability). Since some uncertainty

and subjectivity is hence involved in the selection of u, test-based approaches have been recently introduced (Wadsworth and Tawn, 2012; Northrop and Coleman, 2014; Wadsworth, 2016).

The MRL approach considers the mean excess. If $(X - u) \mid (X > u) \sim \text{GPD}(\sigma_u, \xi)$, the mean excess over $v > u$ for $\xi < 1$ results as

$$\mathbb{E}(X - v \mid X > v) = \frac{\sigma_u + \xi(v - u)}{1 - \xi}. \tag{2.3.8}$$

This property implies that the mean excess in (2.3.8) is linear in v with gradient $\xi/(1-\xi)$. Hence, a threshold diagnostic is performed by plotting the sample mean excess against the threshold and selecting u such that the MRL is a straight line above u.

2.3.4 Extreme Value Theory for Discrete Data

While extreme value modelling is well established for observations from a continuous random variable, comparatively little research has been done for discrete data. Most publications consider the limiting behaviour of the block maxima M_n for a positive integer-valued random variable N with cumulative distribution function F. Anderson (1970) derives several limit laws for M_n, conditional on N having infinite support, which are based upon

$$\lim_{n \to \infty} \frac{1 - F(n)}{1 - F(n+1)}. \tag{2.3.9}$$

The first result by Anderson (1970) implies that the block maxima M_n almost surely takes one of two consecutive integers if, and only if, the limit (2.3.9) tends to infinity. One important example satisfying this condition is the Poisson distribution. Anderson (1970) further proves limiting bounds on the error induced by approximating the discrete data by a continuous distribution, conditional on (2.3.9) being finite and greater than one. This condition holds, for instance, for the geometric distribution. Finally, Anderson (1970) shows that the limiting distribution of M_n is infinitely dispersed if (2.3.9) takes a value equal to 1. These results are later considered by McCormick and Park (1992) and Athreya and Sethuraman (2001).

Anderson (1980) then explores the link between the block maxima behaviour of N and the max-domains of attraction described in Section 2.3.2. If N has a right upper bound in the sense that all higher integer values are observed with zero probability, M_n converges geometrically fast to this point. Hence, no discrete random variable can belong to the max-domain of attraction of the Negative Weibull distribution. Furthermore, Anderson (1980) derives conditions on F

which imply that M_n has a Gumbel or Fréchet distribution.

The Poisson distribution does not satisfy the conditions in Anderson (1980) but Anderson et al. (1997) argue that extreme value modelling can nevertheless be used for Poisson maxima if the rate parameter λ is sufficiently high. Their approach is based upon the approximative Normal behaviour of the Poisson random variable for large λ. In particular, the main result by Anderson et al. (1997) states that the limiting behaviour of M_n is that of a Gumbel distribution, if λ grows with n in a suitable manner. Coles and Pauli (2001) then extend this result to the bivariate case. Nadarajah and Mitov (2002, 2004) derive similar results for other univariate integer-valued distributions, such as, the discrete Uniform and Binomial distributions. Anderson et al. (1997) further consider discrete triangular arrays and specify conditions which imply that the row-wise maximum lies in the max-domain of attraction of a Gumbel distribution. Dkengne et al. (2016) extend this theory and apply it in the context of avalanches.

Shimura (2012) investigates the extremal behaviour of discrete distributions, conditional on their continuous analogue being in any max-domain of attraction. The discretization of a continuous random variable X is defined as the minimal integer not less than X. As mentioned previously, M_n converges to a single point if the discrete random variable has finite support. Hence, the block maxima M_n of a discrete analogue has a degenerate distribution if its continuous analogue lies in the max-domain of attraction of the negative Weibull distribution. Further, in case the limiting behaviour of the continuous distribution is of Fréchet type, its discretization lies in the same max-domain of attraction. This result implies that, for instance, the block maxima M_n sampled from either a t-distribution or its discrete analogue are both Fréchet distributed. Finally, if the block maxima of a continuous random variable is Gumbel distributed, then the discretized analogue is so too if, and only if, the limit in (2.3.9) takes value one. Shimura (2012) states that, for instance, the log-Normal distribution satisfies this condition. These results are align with Anderson (1970). For instance, the geometric distribution does not belong do any max-domain of attraction even though the exponential distribution, its continuous analogue, lies in the max-domain of attraction of the Gumbel distribution.

Additional to block maxima, threshold exceedances of a discrete random variable N are considered too. Prieto et al. (2014) define a discretized GPD based upon the cumulative distribution function of the GPD. If $Y \sim \text{GPD}(\sigma, \xi)$, the discretized analogue $\widetilde{Y}$ has probability mass function

$$\mathbb{P}\left(\widetilde{Y} = y\right) = \mathbb{P}(Y > y) - \mathbb{P}(Y > y + 1), \; y = 1, 2, \dots \; . \tag{2.3.10}$$

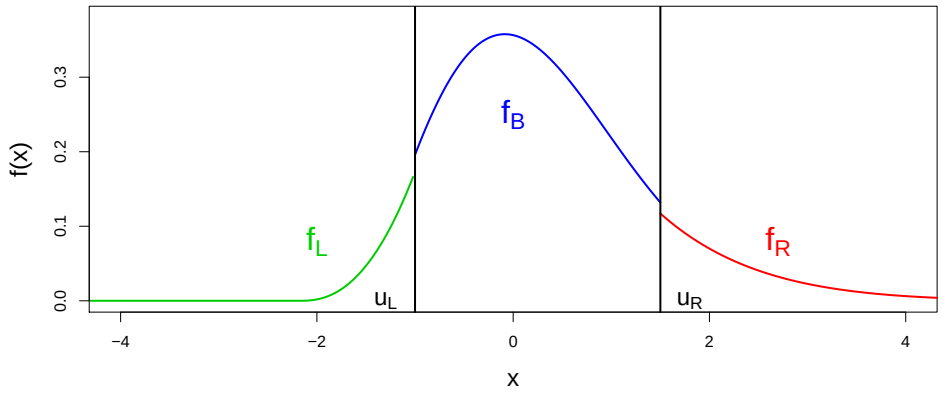


Figure 2.3.3: A density function defined via three components f_L, f_B and f_R for the lower tail, body and upper tail of the distribution, respectively.

Prieto et al. (2014) further derive properties of this discretized GPD, such as its mean and hazard function. The distribution defined in (2.3.10) is then applied to model Spanish road traffic data, in particular, the number of accidents in blackspots.

2.3.5 Extreme-value Mixture Models

The previous sections focused solely on the modelling of the tails of a distribution by considering either block maxima or threshold exceedances. In many applications, however, both the body of a distribution and its tails are of interest. Extreme-value mixture models embed the extreme value methodology and consider the simultaneous estimation of the body and the tails. These models are generally considered for a continuous random variable X with unknown distribution function as in Sections 2.3.2 and 2.3.3.

In principle, the defined model consists of three components which each consider a different domain of F, that is, lower tail, body and upper tail. By defining these components via their density functions f_L, f_B and f_R, the resulting density function f can be formally expressed as

$$f(x) = \mathbb{P}(X < u_L) f_L(x) + \mathbb{P}\left(u_L \leq X \leq u_R\right) f_B(x) + \mathbb{P}\left(X > u_R\right) f_R(x). \tag{2.3.11}$$

Here the support of f_L, f_B and f_R lies within $(-\infty, u_L), [u_L, u_R]$ and (u_R, ∞), respectively, and F_L, F_B and F_R are the associated distribution functions. Further u_L and u_R denote respectively the lower and upper thresholds. The unknown distribution function F_L and F_R are commonly modelled as GPD. Figure 2.3.3 shows an example for a density function f defined as in (2.3.11).

In what follows, several approaches to estimate such models are described, some of these impose a continuity constraint for f at the thresholds u_L and u_R. For notational simplicity, the model consists of two components only, the body f_B and the upper tail f_R.

Behrens et al. (2004) introduce a Bayesian modelling framework which treats the threshold as unknown and estimates it together with the remaining parameters. Specifically, the approach considers a parametric form of F_B, e.g. a Normal or Weibull distribution, and a GPD for F_R. Hence, the distribution function $F(x)$ is formally given as

$$F(x) = \begin{cases} F_B(x) & x \leq u \\ F_B(u) + [1 - F_B(u)]\, F_R(x) & x > u. \end{cases} \tag{2.3.12}$$

This model specification generally results in a discontinuity of f at u. Behrens et al. (2004) state that uncertainty in u increases with a decreasing size of the discontinuity. De Melo Mendes and Lopes (2004) consider a similar model specification with F_B being a Normal distribution and different models for lower and upper tail. Instead of incorporating uncertainty in the thresholds, they optimize the proportion of data used to model the lower and upper tail. Carreau and Bengio (2009) propose a hybrid Pareto model which splices a Normal density with the density of a GPD while preserving smoothness of $f(x)$ at the threshold.

While the former parametric approaches consider a strict partition as in (2.3.11), Frigessi et al. (2002) propose a dynamically weighted mixture model of a Weibull distribution and a GPD. In particular, the specified model does not require the estimation of a threshold. The density function is then formally given by

$$f(x) = \frac{[1 - p(x)]\, f_B(x) + p(x) f_R(x)}{C},$$

where the mixing function $p(x)$ satisfies $p(x) \to 1$ as $x \to x^F$, and f_B and f_R denote the density functions of a Weibull distribution and a GPD respectively. Further, C is a normalising constant which depends on the parameters. This model formulation implies that the GPD dominates the tail behaviour.

Additional to the parametric approaches, the non-parametric estimation of f_B is considered in the literature. Tancredi et al. (2006) define f_B as piece-wise constant with an unknown number of steps and u is taken to be unknown. Estimates are then obtained via in reversible jump MCMC scheme (Green, 1995). MacDonald et al. (2011) introduce a highly flexible kernel-

density based approach which also uses the Poisson process formulation in Section 2.3.3.

Chapter 3

Modelling Insurance Claims by Spatially Varying Regression

3.1 Introduction

Eshita (1977) models the claim frequencies via a Binomial or Poisson model and, as outlined in Section 2.1, such approaches are often also applied in disease mapping. Scheel et al. (2013) argue, however, that these distributions are unsuitable for the insurance and weather data explored in Section 1.3 as these are incapable of capturing the high frequency of zero claims. They hence propose a Bayesian Poisson hurdle (BPH) model; see Section 1.4 for details. In this chapter, a comparative study is performed in order to assess the degree of improvement obtained by the Poisson hurdle model approach, as compared to one based on the Binomial distribution. Here, this comparison is based upon all $K = 430$ municipalities.

Scheel et al. (2013) estimate the baseline risk and covariate effects in the BPH model individually for each municipality and spatial information is used for variable selection (Section 1.4). The modelling framework introduced in this chapter also estimates the covariate effects municipality-wise but defines a dependence structure which, a priori, assumes covariate effects of adjacent municipalities to be more similar. In particular, the approach is based on the geographically varying coefficient (GVC) model (Assunção, 2003; Congdon, 2003) described in Section 2.1. Hence, estimates are obtained in a Bayesian framework rather than via weighted least squares methodology as in the geographically weighted regression approach (Brunsdon et al., 1998; Fotheringham et al., 2002).

A spatially varying modelling approach is suitable to examine the dependence between the

daily number of claims $N_{k,t}$ and the weather data. Firstly, Section 1.3 shows a spatially varying vulnerability to the covariates, e.g. the same amount of precipitation affects Oslo and Bergen differently. Secondly, adjacent municipalities exhibit strong similarities in the weather covariates and, hence, have a, presumably, similar vulnerability. The GVC modelling framework allows the explicit specification of a dependence model which shares statistical information between adjacent municipalities.

In addition to this model comparison, the potential adjustments to the covariates discussed in Section 1.3 are investigated. The exploratory data analysis indicates that the effect of rainfall and snowfall are different, in particular, the correlation of snowfall and claim numbers is close to zero. Conversely, $N_{k,t}$ and $R_{k,t}$ are positively correlated, conditional on a positive daily mean temperature, $C_{k,t} > 0^\circ$ Celsius. In other words, the analysis implies that rain affects the claim dynamics on the day stronger than snow. Hence, the original covariates $R_{k,t}$ and $R_{k,t-1}$ may be replaced by $\widetilde{R}_{k,t}$ and $\widetilde{R}_{k,t-1}$ which are defined as

$$\widetilde{R}_{k,t} = R_{k,t} \ \mathbb{1}_{\{C_{k,t}>0\}}. \tag{3.1.1}$$

Consequently, the amount of snow fall on the day has no effect on the claim dynamics on the same day.

Section 1.3 also discusses the difference in snow-water equivalent, $\Delta S_{k,t}$. The exploratory data analysis indicates that $N_{k,t}$ and $\Delta S_{k,t}$ are positively correlated, conditional on $\Delta S_{k,t} > 0$, while little or no dependence is found otherwise. Similarly to $R_{k,t}$, the covariate $\Delta S_{k,t}$ may be replaced by $\widetilde{\Delta S}_{k,t}$ which is defined by

$$\widetilde{\Delta S}_{k,t} = \Delta S_{k,t} \ \mathbb{1}_{\{\Delta S_{k,t}>0\}}. \tag{3.1.2}$$

In conclusion, $\widetilde{\Delta S}_{k,t}$ corresponds to the amount of snow-melt rather than the difference in snow-water equivalent. Additionally to these potential refinements in the covariates, Section 1.3 indicates that cities exhibit higher vulnerability than rural areas. This potential difference is included via an additional binary factor Z_k which takes value $Z_k = 1$ if the average number of policies over the 10-year period exceeds 2,000 and $Z_k = 0$ otherwise. To derive the threshold of 2,000, the average claim rate per policy holder for municipalities below and above a range of potential thresholds was derived first. The examination of the ratio of the average claim rate then indicated that a threshold at around 2,000 yields the highest difference in the average claim

rates.

In summary, interest lies in the comparison of the Binomial and Poisson hurdle approaches, as well as, the examination of the performance of the proposed covariates $\widetilde{R}_{k,t}$ and $\widetilde{\Delta S}_{k,t}$ in expressions (3.1.1) and (3.1.2), respectively. The drainage run-off $D_{k,t}$ and the snow-water equivalent $S_{k,t}$ are included additionally to the three covariates discussed above. In the following, three competing models are examined: (i) Binomial distribution and original covariates (ii) Binomial distribution and proposed covariates and (iii) Poisson hurdle model and proposed covariates. A fourth possible setting is a Poisson-Hurdle model with the original covariates as done by Scheel et al. (2013). Instead of fitting this model, we compare the three settings to the original model by Scheel et al. (2013) which imposes less spatial structure on the covariate effects.

The remainder of this chapter is organized as follows: Section 3.2 details the modelling framework for both the Binomial and Poisson hurdle distribution and describes the estimation of the model parameters using Markov chain Monte Carlo (MCMC) techniques. Results for the three models are then compared and discussed in Section 3.3. Equivalently to Scheel et al. (2013), the predictive performance is assessed on a weekly basis. The chapter concludes with a discussion in Section 3.4.

3.2 Modelling and Inference

The statistical framework for both the Binomial distribution and the Poisson hurdle model are specified via a Bayesian hierarchical model. Such models are widely applied, for instance in atmospheric modelling (McBride et al., 2007) and marine ecology (Qian et al., 2009), and are based on three levels: a data, a process and a parameter model. Interest lies in the distribution of the daily claim numbers $N_{k,t}$ conditional on the weather covariates for each municipality $k = 1, \ldots, K$. Both distributions assume that the individual insurance policies are IID Bernoulli distributed, given the covariates. While this assumption appears unlikely as some dependence probably exists for adjacent households, it may provide a reasonable approximation.

In the following, the weather covariates on day t for municipality k are denoted by $\mathbf{X}_{k,t}$; this notation refers to both the original and proposed covariate set. Let $T_k \subseteq \{1, \ldots, 3651\}$ denote the set of days on which both $N_{k,t}$ and $\mathbf{X}_{k,t}$ are recorded for municipality k (as some missing values exist). Sections 3.2.1 and 3.2.2 define the modelling frameworks and Section 3.2.3

describes the MCMC algorithm to obtain estimates of the model parameters.

3.2.1 Binomial Model

Since the number of policies $A_{k,t}$ per month and municipality is known, the number of claims can be modelled via a Binomial distribution. The daily number of claims $N_{k,t}$ for municipality k on day t is hence modelled via a Binomial distribution with number of samples $A_{k,t}$ and claim probability $p_{k,t}$. Formally, the data model for municipality k, $k = 1, \ldots, K$, on day $t \in T_k$ is

$$N_{k,t} \sim \text{Binomial}\left(A_{k,t}, p_{k,t}\right). \tag{3.2.1}$$

Alternatively, a Poisson distribution may be considered as $N_{k,t}$ is very small in comparison to $A_{k,t}$.

Similarly to Haug et al. (2011) and Scheel et al. (2013), the dependence of $N_{k,t}$ on the weather data $\mathbf{X}_{k,t}$ is specified via $p_{k,t}$. Since the claim probability takes values between 0 and 1 only, a linear model with covariates $\mathbf{X}_{k,t}$ and binary factor Z_k is defined for $p_{k,t}$ on the logit scale. The process model is then formally given by

$$\text{logit}\left(p_{k,t}\right) = \beta_{k,0} + \mathbf{X}_{k,t}^T \boldsymbol{\beta}_k + \rho_p Z_k, \tag{3.2.2}$$

where ρ_p, the baseline $\beta_{k,0}$ and the covariate effects $\boldsymbol{\beta_k} = \left(\beta_{k,1}, \ldots, \beta_{k,5}\right)$, $k = 1, \ldots, K$, are the parameters of interest. This model specification for municipality k with claim numbers $\mathbf{N}_k = \{N_{k,t} : t \in T_k\}$, conditional on $\mathbf{X}_k = \{\mathbf{X}_{k,t} : t \in T_k\}$ and Z_k, results in a likelihood function which is formally given as

$$f\left(\mathbf{N}_k \mid \mathbf{X}_k, Z_k, \beta_{k,0}, \boldsymbol{\beta}_k, \rho_p\right) = \prod_{t \in T_k} \binom{A_{k,t}}{N_{k,t}} \frac{\left[\exp\left(\beta_{k,0} + \mathbf{X}_{k,t}^T \boldsymbol{\beta}_k + \rho_p Z_k\right)\right]^{N_{k,t}}}{\left[1 + \exp\left(\beta_{k,0} + \mathbf{X}_{k,t}^T \boldsymbol{\beta}_k + \rho_p Z_k\right)\right]^{A_{k,t}}}. \tag{3.2.3}$$

The Bayesian hierarchical model is concluded by defining the parameter model via prior distributions for the parameters. Here, the baseline risk and the covariate effects of adjacent municipalities are assumed to be similar. Further, no dependence is assumed, a priori, between the parameters associated with different covariates. In other words, a dependence model is defined separately for $\beta_{k,0}$ through to $\beta_{k,5}$. In particular, a geographical structure in each parameter is induced via the specification of an intrinsic conditional autoregressive (ICAR) model (Besag

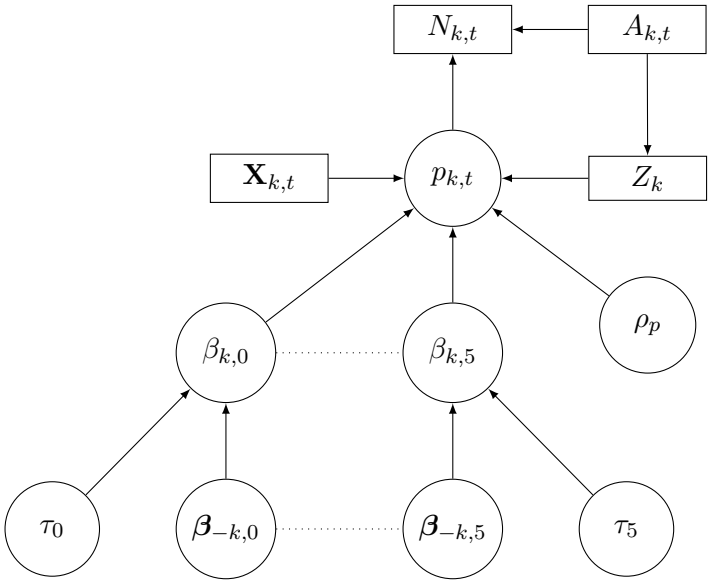


Figure 3.2.1: Network representing the dependence of the parameters for the Binomial model detailed in in Section 3.2.1.

et al., 1991; Rue and Held, 2005). Hence, the joint prior distribution for the parameter vector $\pi\left(\beta_{1,j},\ldots,\beta_{K,j}\right),\ j=0,\ldots,5$, is specified via its full conditionals $\beta_{k,j}|\boldsymbol{\beta}_{-k,j},\ k=1,\ldots,K$, where $\boldsymbol{\beta}_{-k,j}=(\beta_1,j,\ldots,\beta_{k-1,j},\beta_{k+1,j},\ldots,\beta_{K,j})$ denotes the vector of covariate effects for all municipalities except k. Here, the full conditional $\beta_{k,j}|\boldsymbol{\beta}_{-k,j}$ is Normally distributed and of the form

$$\beta_{k,j}\mid\boldsymbol{\beta}_{-k,j}\sim\text{Normal}\left(\frac{\sum_{k'=1}^{K}d_{k,k'}\beta_{k',j}}{\sum_{k'=1}^{K}d_{k,k'}}\ ,\ \frac{1}{\tau_j\sum_{k'=1}^{K}d_{k,k'}}\right),\tag{3.2.4}$$

where $\tau_j\geq 0$ is a hyperparameter related to the conditional variance of $\beta_{k,q}$ given the covariate effects of the other municipalities. More precisely, the degree of dependence between the covariate effects increases with τ_j. The constants $d_{k,k'}$ describe the weighting of the municipalities and are specified here as

$$d_{k,k'}=\begin{cases}1 & \text{if municipalities } k \text{ and } k' \text{ are adjacent}\\ 0 & \text{otherwise.}\end{cases}\tag{3.2.5}$$

In conclusion, the prior specified in (3.2.4) and (3.2.5) favours $\beta_{k,j}$ to be close to a weighted average over $\boldsymbol{\beta}_{-k,j}$. Further, the variance of the Gaussian distribution in (3.2.4) decreases with the number of adjacent municipalities.

The modelling framework is completed by the specification of priors on the remaining parameters ρ_p and $\tau_j,\ j=0,\ldots,5$. Figure 3.2.1 illustrates the dependence in the model parameters

via a directed acyclic graph (DAG). Weakly-informative Gamma priors are set on the hyper-paramaters

$$\tau_j \sim \text{Gamma}\,(1, 0.01) \tag{3.2.6}$$

and a standard Gaussian prior is placed on ρ_p

$$\rho_p \sim \text{Normal}\,(0, 1)\,. \tag{3.2.7}$$

3.2.2 Poisson Hurdle Model

The BPH model consists of two components, one of which is a Bernoulli distribution modelling whether $N_{k,t}$ is zero or positive while the other models the number of claims, conditional on at least one claim being recorded: $N_{k,t}|N_{k,t} > 0$, by a count distribution. Here, the parameter setup of the Bernoulli component is defined slightly differently to that of in Section 1.4. In particular, the parameter $\alpha_{k,t}$ is replaced by $(1 - q_{k,t})^{A_{k,t}}$ in order to aid comparability to the Binomial model. If the estimates for the Binomial distribution and the BPH model are quite different, the estimated parameters for $p_{k,t}$ and $q_{k,t}$ should be so too. Consequently, the data model is then formally defined by

$$\mathbb{P}\left(N_{k,t} = n \mid \mathbf{X}_{k,t}\right) = \begin{cases} (1 - q_{k,t})^{A_{k,t}} & \text{if } n = 0 \\ \left[1 - (1 - q_{k,t})^{A_{k,t}}\right] \dfrac{\lambda_{k,t}^n}{n!\left[\exp(\lambda_{k,t}) - 1\right]} & \text{if } n > 0 \end{cases}, \tag{3.2.8}$$

The parameters of interest are $q_{k,t}$ and $\lambda_{k,t}$ which vary in dependence on the weather covariates $\mathbf{X}_{k,t}$. As for $p_{k,t}$ in (3.2.2), a linear model is defined for the probability $q_{k,t}$ on the logit scale. Formally,

$$\text{logit}\left(q_{k,t}\right) = \gamma_{k,0} + \mathbf{X}_{k,t}^T \boldsymbol{\gamma}_k + \rho_q Z_k. \tag{3.2.9}$$

The rate parameter $\lambda_{k,t}$ on log scale is modelled via a linear model (similarly to Section 1.4) and yields

$$\log\left(\lambda_{k,t}\right) = \delta_{k,0} + \mathbf{X}_{k,t}^T \boldsymbol{\delta}_k + \rho_\lambda Z_k + \log\left(A_{k,t}\right),\ k = 1, \ldots, K. \tag{3.2.10}$$

Consequently, we obtain a likelihood function for each municipality k for the set of model parameters specified in (3.2.8) to (3.2.10), denoted $f\left(\mathbf{N}_k \mid \mathbf{X}_k, Z_k, \gamma_{k,0}, \boldsymbol{\gamma}_k, \delta_{k,0}, \boldsymbol{\delta}_k, \rho_q, \rho_\lambda\right)$.

The Bayesian hierarchical model is completed by specifying priors for the parameters. Similarly to the Binomial setting in Section 3.2.1, the geographical dependence in the baselines and

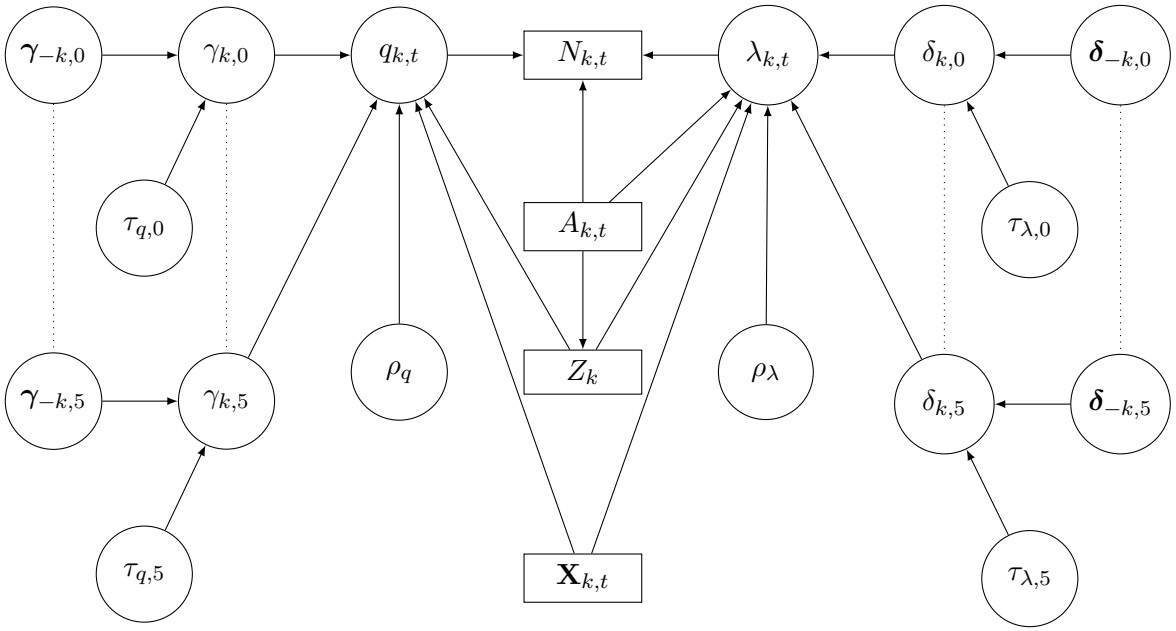


Figure 3.2.2: Network representing the dependence of the parameters for the Poisson Hurdle model.

covariate effects is specified via conditional autoregressive models $\gamma_{k,j}|\boldsymbol{\gamma}_{-k,j}$ and $\delta_{k,j}|\boldsymbol{\delta}_{-k,j}$. The weights $d_{k,k'}$ are set as in (3.2.5) and let $\boldsymbol{\tau}_q = (\tau_{q,0}, \ldots, \tau_{q,5})$ and $\boldsymbol{\tau}_\lambda = (\tau_{\lambda,0}, \ldots, \tau_{\lambda,5})$ denote the hyperparameters in the conditional autoregressive models for the parameters $\boldsymbol{\gamma}_k$ and $\boldsymbol{\delta}_k$, respectively. Figure 3.2.2 illustrates the dependence of the model parameters for $N_{k,t}$ and the DAG shows that $q_{k,t}$ and $\lambda_{k,t}$ are conditionally independent given the observations $\mathbf{X}_{k,t}$, $A_{k,t}$ and Z_k. Finally, standard Gaussian priors as in (3.2.7) are specified for ρ_q and ρ_λ, and Gamma priors for the hyperparameters in the ICAR model as in (3.2.6).

3.2.3 Inference

The posterior distribution given by the specified likelihood function and the prior distributions is of non-standard form for both the Binomial and the BPH model. Realizations from the posterior are therefore sampled via a block Metropolis-Hastings algorithm, also known as Metropolis within Gibbs. Parameter values are then updated iteratively by sampling either directly from the full conditional, if it is of standard form, or via a Metropolis-Hastings step as introduced by Metropolis et al. (1953) and Hastings (1970). The parameters of the municipalities are updated following their order of appearance in the data set. In the following, the updates for the two modelling frameworks are detailed:

Binomial Model

Baseline levels $\beta_{k,0}$ and covariate effects $\boldsymbol{\beta}_k$ are updated via a Random-Walk Metropolis step. Consider the baseline $\beta_{k,0},\ k = 1, \ldots, K$. A proposal $\beta^*_{k,0}$ is sampled from a Normal distribution with mean $\beta_{k,0}$ and standard deviation 0.05, $\beta^*_{k,0} \sim \text{Normal}\left(\beta_{k,0}, 0.05\right)$. Based on the likelihood function in (3.2.3) and the ICAR prior in (3.2.4), the acceptance probability $\alpha\left(\beta_{k,0}, \beta^*_{k,0}\right)$ then takes the form

$$\alpha\left(\beta_{k,0}, \beta^*_{k,0}\right) = \min\left\{1\ ,\ \prod_{t \in T_k} \frac{f\left(\mathbf{N}_k \mid \mathbf{X}_k, Z_k, \beta^*_{k,0}, \boldsymbol{\beta}_k, \rho_p\right)}{f\left(\mathbf{N}_k \mid \mathbf{X}_k, Z_k, \beta_{k,0}, \boldsymbol{\beta}_k, \rho_p\right)} \times \frac{\pi\left(\beta^*_{k,0} \mid \boldsymbol{\beta}_{-k,0}\right)}{\pi\left(\beta_{k,0} \mid \boldsymbol{\beta}_{-k,0}\right)}\right\} \tag{3.2.11}$$

To improve numerical stability, the acceptance probability is derived on the log-scale. The baseline $\beta_{k,0}$ is set to $\beta^*_{k,0}$ if $\log\left[\alpha\left(\beta_{k,0}, \beta^*_{k,0}\right)\right] > \log(u)$, where u is sampled from a standard uniform distribution: $u \sim \text{Uniform}(0, 1)$. The covariate effects $\beta_{k,j},\ j = 0, \ldots, 5,\ k = 1, \ldots, K$ and the parameter ρ_p are then updated similarly with proposal distributions taking the form $\beta^*_{k,j} \sim \text{Normal}\left(\beta_{k,j}, 0.05\right)$ and $\rho^*_p \sim \text{Normal}\left(\rho_p, 0.05\right)$, respectively.

The hyperparameters $\boldsymbol{\tau}_q$ are updated via Gibbs sampling. As detailed in Section 2.1, several authors advocate that the joint prior $\pi\left(\beta_{1,j}, \ldots, \beta_{K,j}\right),\ j = 0, \ldots, 5$, specified via the full conditionals $\beta_{k,j}|\boldsymbol{\beta}_{-k,j},\ k = 1, \ldots, K$, in (3.2.4) and (3.2.5) takes the form

$$\pi\left(\beta_{1,j}, \ldots, \beta_{K,j}\right) \propto \tau_p^{\frac{K-1}{2}} \exp\left(-\frac{\tau_p}{2}\left(\beta_{1,p}, \ldots, \beta_{K,p}\right)^T Q\left(\beta_{1,p}, \ldots, \beta_{K,p}\right)\right). \tag{3.2.12}$$

Here, Q is a $K \times K$ matrix with non-diagonal entries $Q_{k,k'} = -1$ if municipalities k and k' are adjacent and diagonal entries $Q_{k,k}$ are equal to the number of municipalities adjacent to k. Since Gamma priors are set on each component in $\boldsymbol{\tau}_p$, the full conditional, $\pi\left(\tau_{p,j}|\cdot\right)$, of the posterior is a Gamma distribution. Hence, the parameter $\tau_{p,j}$ is updated via a Gibbs step which samples directly from

$$\tau_{p,j} \sim \text{Gamma}\left(\frac{K-1}{2} + 1, \left(\beta_{1,j}, \ldots, \beta_{K,j}\right)^T Q\left(\beta_{1,j}, \ldots, \beta_{K,j}\right) + 0.01\right). \tag{3.2.13}$$

Poisson Hurdle Model

Parameter values are updated similarly to the Binomial modelling framework. Proposals for the baselines $\gamma_{k,0}$ and $\delta_{k,0}$, the covariate effects $\boldsymbol{\gamma}_k$ and $\boldsymbol{\delta}_k$, and the parameter ρ_q and ρ_λ are updated via a Random Walk Metropolis with proposals being sampled from a Normal distribution

with standard deviation 0.05. The acceptance probability is then as in (3.2.11) but with likelihood function $f\left(\mathbf{N}_k \mid \mathbf{X}_k, Z_k, \gamma_{k,0}, \boldsymbol{\gamma}_k, \delta_{k,0}, \boldsymbol{\delta}_k, \rho_q, \rho_\lambda\right)$. For the parameters of the zero-truncated Poisson component, only observations with $N_{k,t} > 0$ have to be considered for the update. Furthermore, if for a municipality k, no event $N_{k,t} > 1$ is recorded in the data, then the parameters are solely updated via the full conditional prior specification. Finally, the hyperparameters in $\boldsymbol{\tau}_q$ and $\boldsymbol{\tau}_\lambda$ are updated as in (3.2.13).

3.3 Results

Equivalently to Scheel et al. (2013), the observations for 2001 are stored as a test data set and parameter estimates are based on the remaining nine years. Covariates are scaled across all 430 municipalities to have mean 0 and variance equal to 1, hence allowing comparison of the associated vulnerability across municipalities. Samples for the three model specifications are obtained via the MCMC algorithms outlined in Section 3.2.3. For each model, 10,000 iterations are performed after a burn-in of 1,000 iterations.

Convergence is verified via sampling two additional Markov chains with 4,000 iterations for each model and performing Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998), as well as, investigating the trace plots. These diagnostic tools indicate that the Markov chains have converged after the burn-in period. Furthermore, only every tenth sample is considered for analysis, leading to a collection of 1,000 samples for the analysis. The mixing is assessed for the four cities of Oslo, Bergen, Trondheim and Bærum and the effective sample size obtained for the 1,000 samples lies between 200 and 1,000 for most of the parameters. Trace plots of the marginal posterior samples for Oslo and Bergen are provided in Appendix B.1.

Section 3.3.1 summarizes and compares the estimated parameter values for the three models. Furthermore, uncertainty in the estimates is considered too and we explore which covariates are important for which municipality. Section 3.3.2 then assesses the predictive performance of the estimated models.

3.3.1 Parameter Estimates

The difference between the original and proposed covariates, $\mathbf{X}_{k,t}$ and $\widetilde{\mathbf{X}}_{k,t}$ respectively, is explored first by considering the estimates of the two Binomial models. Results indicate that the posterior mean estimates of the two models are very similar. Figure 3.3.1 illustrates the results

for the $\mathbf{X}_{k,t}$ and plots for $\widetilde{\mathbf{X}}_{k,t}$ are provided in Appendix B.2. South-eastern and northern Norway are estimated to have the highest baseline risk while central municipalities have the lowest one. These results are consistent with the empirical average claim rate in Figure 1.3.1 in Section 1.3. Further, high correlation is found between the parameter ρ_p and the parameter $\beta_{k,0}$ for the municipalities with $Z = 1$; hence the combined baseline risk is considered. Nevertheless, the samples obtained via the MCMC algorithm for $\beta_{k,0} + Z_k\rho_p$ show good mixing properties.

The estimated covariate effects for $R_{k,t}$ and $R_{k,t-1}$ are generally positive and indicate that inland municipalities are more vulnerable than coastal areas. Since western municipalities exhibit, on average, higher precipitation levels, the buildings are expected to be designed such that they can withstand more severe rainfall events. Furthermore, the two covariate effects typically have a similar value, that is, if the covariate effect for $R_{k,t}$ is high, so is the one for $R_{k,t-1}$ and vice versa. Similarly, the covariate effects associated to $D_{k,t}$ indicate a higher risk for eastern and northern regions, as compared to coastal areas. The snow-water equivalent $S_{k,t}$ appears to have the largest effect on the claim dynamics for northern municipalities and these are also the ones which observe the highest covariate values. With respect to $\Delta S_{k,t}$, the municipalities exhibiting highest vulnerability are in eastern Norway. When compared to the rainfall covariates, the remaining covariates appear to often have a smaller impact as manifest by the smaller posterior mean estimates. As for the baseline claim risk, the sampled Markov chains show good mixing.

Considering the BPH model, the estimates for the hurdle component in Figure 3.3.2 exhibit similar spatial patterns to the estimates of the Binomial model. Larger differences are only found for western coastal municipalities with respect to $S_{k,t}$ and also for $\Delta S_{k,t}$ for northern Norway. Again, urban municipalities exhibit a higher baseline risk than rural municipalities in central Norway. In comparison to the estimated baseline risk for the Binomial model, the posterior mean estimates are slightly lower. Figure 3.3.2 further shows that the estimated covariate effects for $q_{k,t}$ are generally lower, as compared to the Binomial model, and take negative values for a few municipalities. This is, in particular, the case for $\Delta S_{k,t}$. These results indicate that the frequency of zero claims may indeed be higher than the estimated Binomial model suggests.

Next, the Poisson component is examined by the spatial posterior mean plots for $\delta_k, 0$ through to $\delta_{k,5}$ in Figure 3.3.3. Compared to the hurdle component, the estimates are spatially more variable. The baseline risk is also higher than for the hurdle component with the highest levels found for northern Norway. The rainfall covariates $R_{k,t}$ and $R_{k,t-1}$ appear to be important

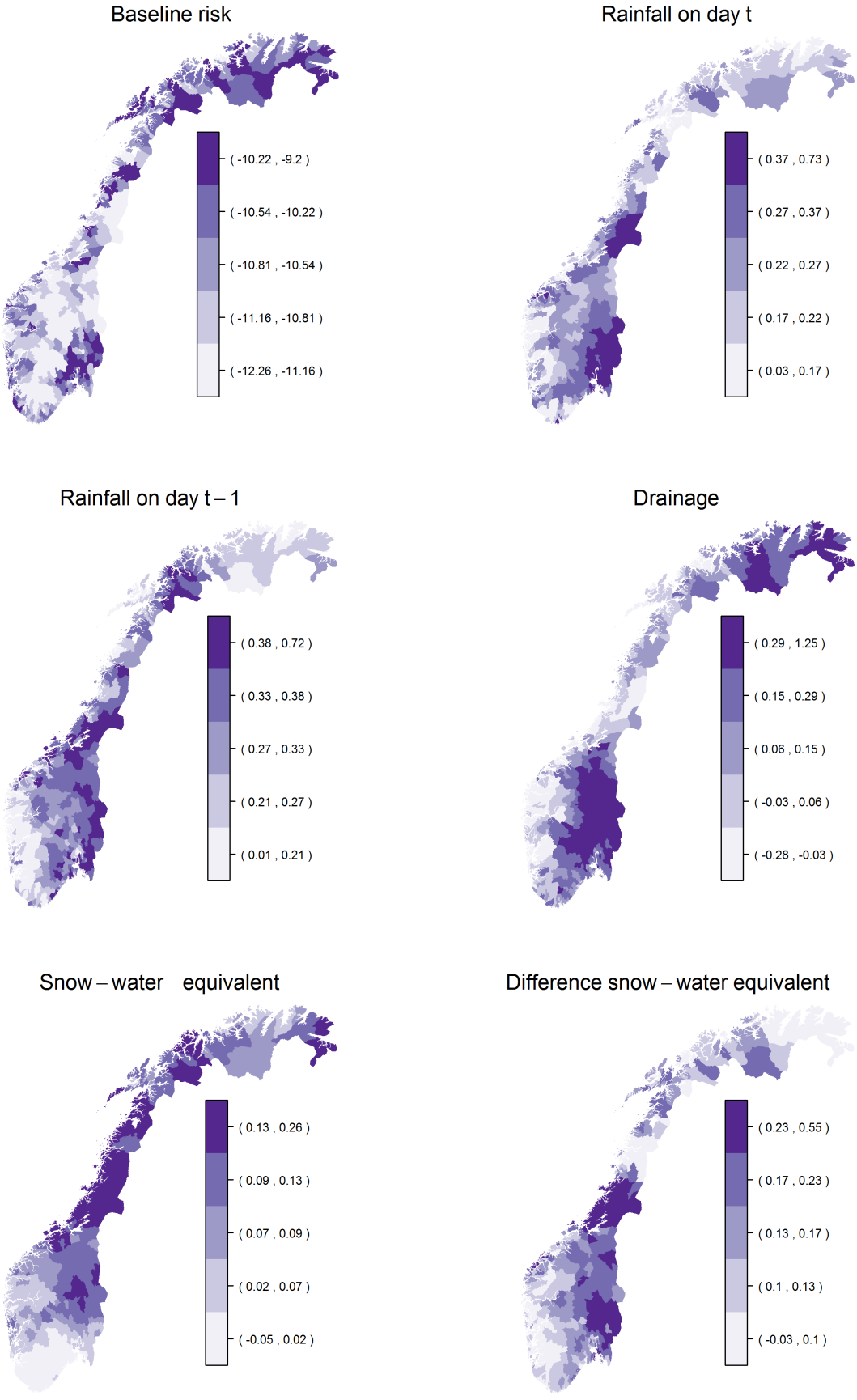


Figure 3.3.1: Posterior mean estimates of the baseline risk $\beta_{k,0}+\rho_p Z_k$ and the covariate effects obtained for the Binomial model with the original covariates.

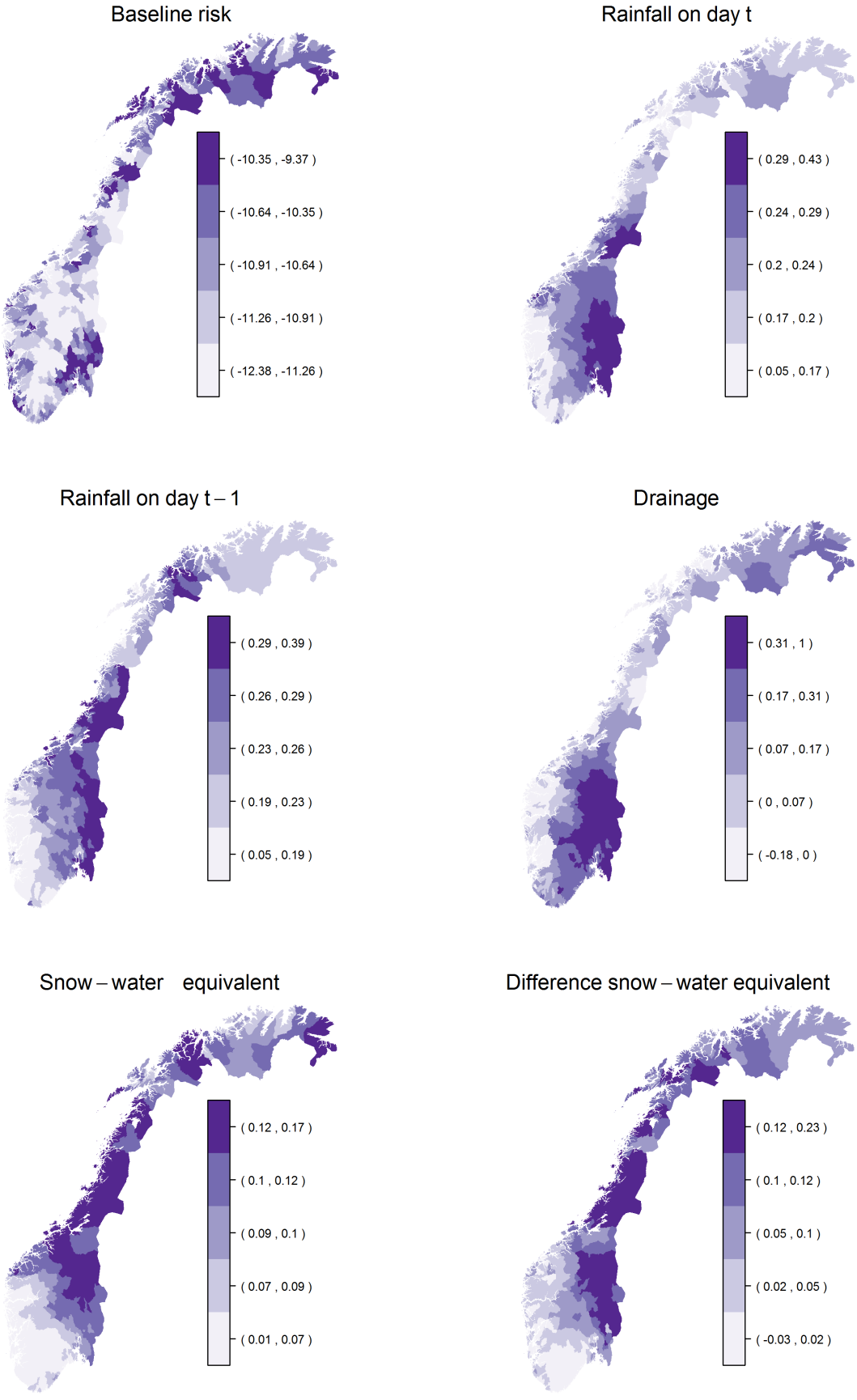


Figure 3.3.2: Posterior mean estimates of the baseline risk $\gamma_0 + \rho_q Z$ and the covariate effects $\gamma_1, \ldots, \gamma_5$ obtained for the Binomial model with the proposed data.

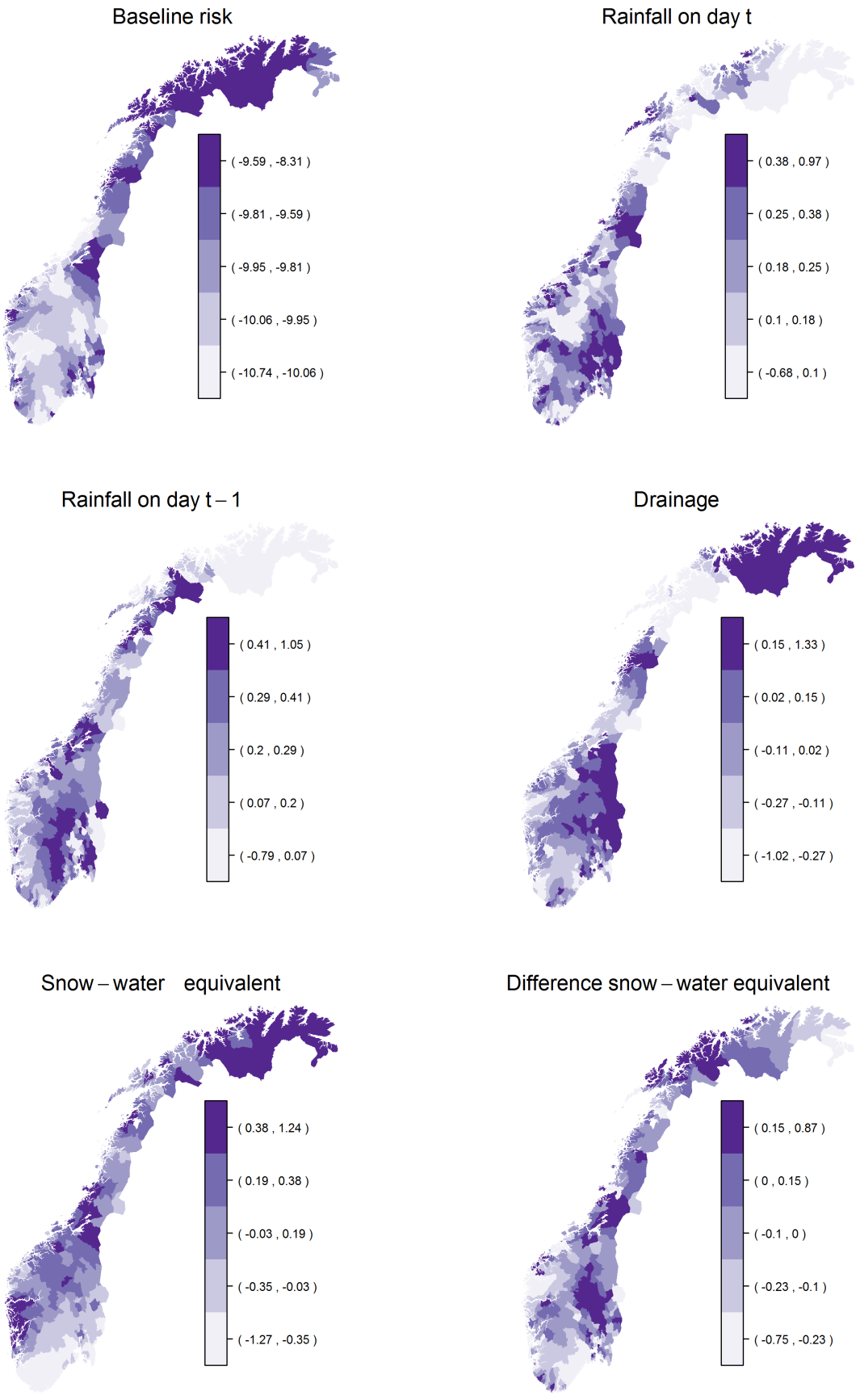


Figure 3.3.3: Posterior mean estimates of the baseline risk $\delta_0 + \rho_\lambda Z$ and the covariate effects $\delta_1, \ldots, \delta_5$ obtained for the Poisson Hurdle model with the proposed data.

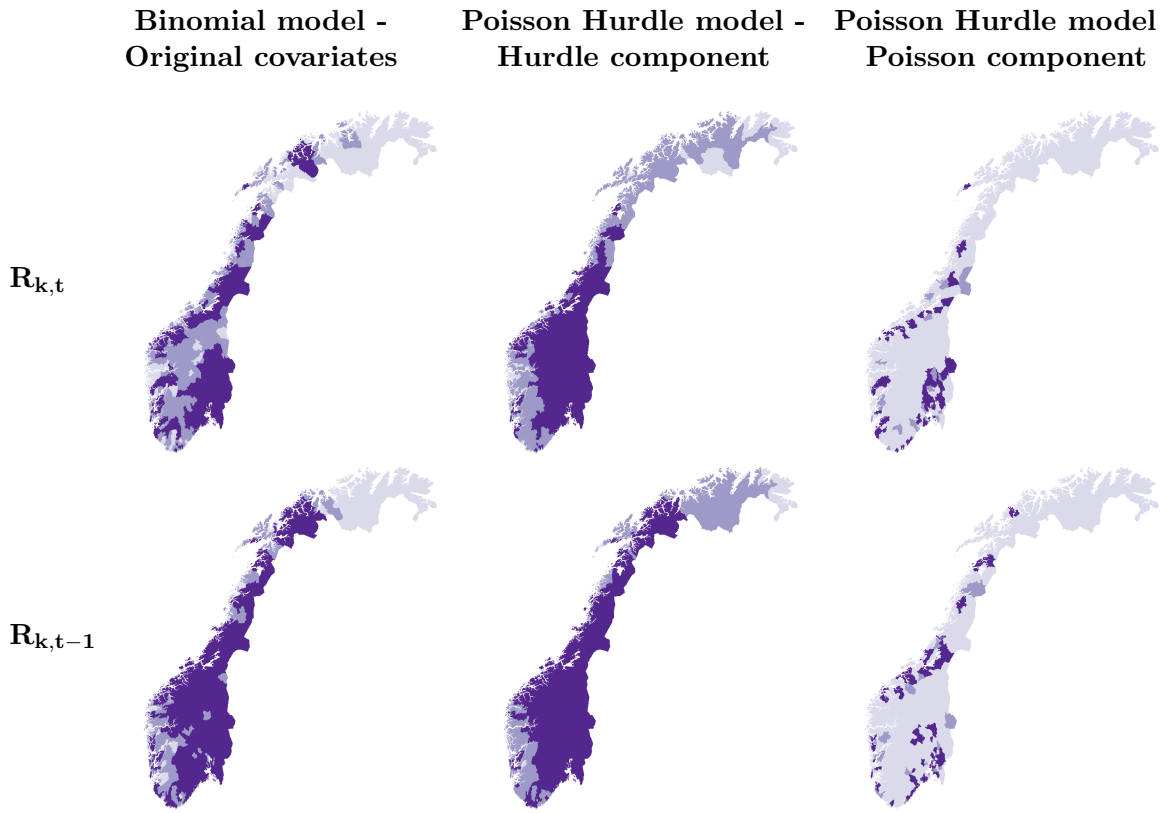


Figure 3.3.4: Classification of the 95% credibility interval of the covariate effects associated to $R_{k,t}$ and $R_{k,t-1}$ into three categories. For municipalities with the darkest colour, the central 95% credibility interval does not intersect with the interval $(-0.1, 0.1)$ while zero is contained in the 95% credibility interval for municipalities with a light colouring.

for the claim dynamics for most, but not all, municipalities. Further, the covariate effects for $\delta_{k,3}$ through $\delta_{k,5}$ for $D_{k,t}$, $S_{k,t}$ and $\Delta S_{k,t}$, respectively, take negative values for about half of the municipalities. This may partly be due to the low frequency of higher claims for most municipalities and, hence, no high claim numbers being recorded for days with high values for $\Delta S_{k,t}$ and $D_{k,t}$.

After the investigation of the spatial variation in the posterior mean, interest lies in the detection of covariates which are important for the claim dynamics in each municipality. Therefore, the uncertainty is considered in addition to the posterior mean plots in (Figures 3.3.1 through to 3.3.3) and the central 95% credibility intervals are derived. A covariate is considered important if a covariate value of zero does not lie within the derived credibility interval. Firstly, the two rainfall covariates are examined in Figure 3.3.4. Again, the results for the Binomial model with proposed covariate set $\widetilde{\mathbf{X}}_{k,t}$ are very similar to the ones for the original covariates and, hence, omitted. Figure 3.3.4 indicates a high degree of certainty for the Binomial model and the hurdle component on the covariates effects associated to $R_{k,t}$ and $R_{k,t-1}$ being non-zero for almost all

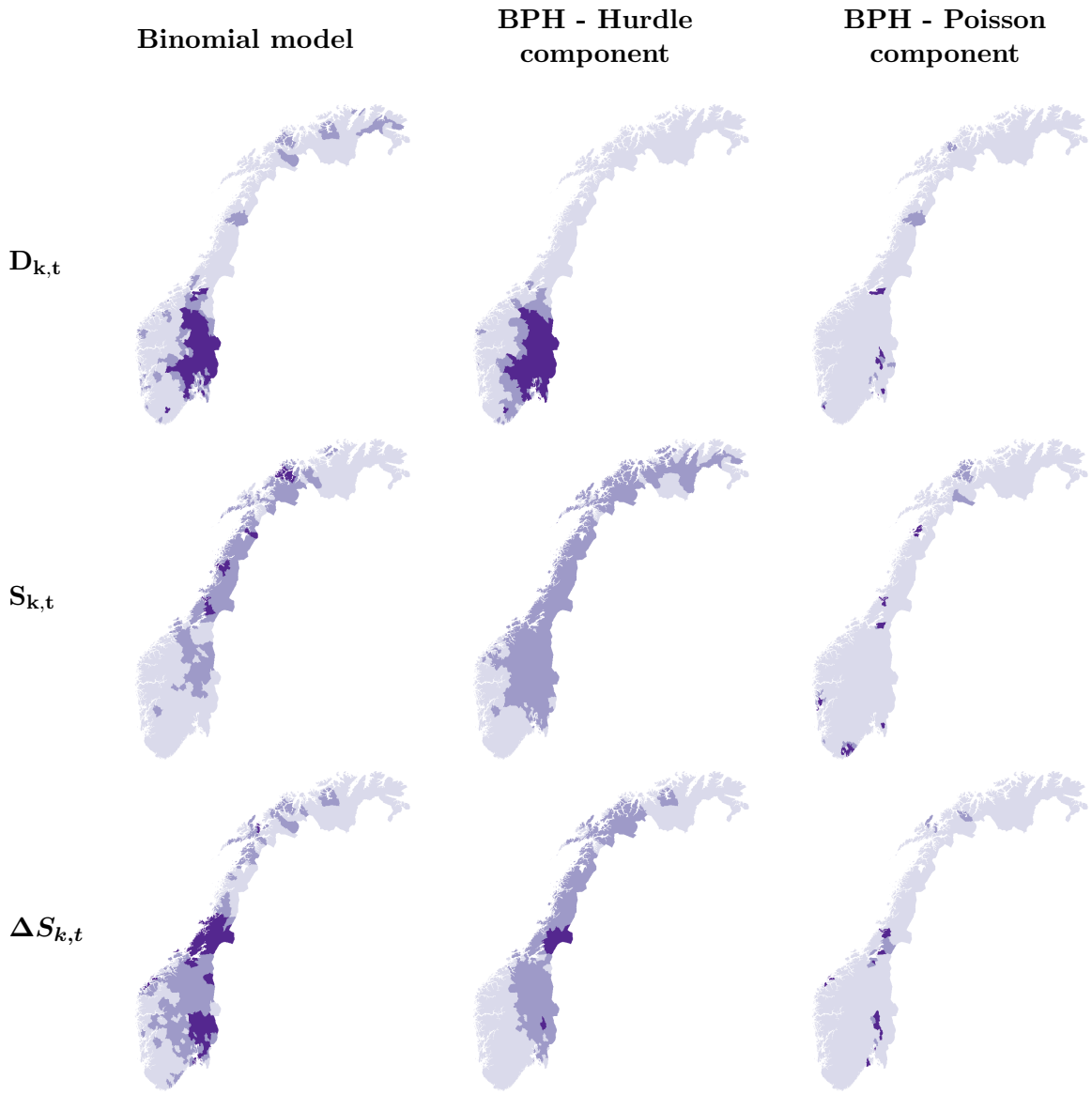


Figure 3.3.5: Classification of the 95% credibility interval of the covariate effects associated to $D_{k,t}$, $S_{k,t}$ and $\Delta S_{k,t-1}$ into three categories. For municipalities with the darkest colour, the central 95% credibility interval does not intersect with the interval $(-0.1, 0.1)$ while zero is contained in the 95% credibility interval for municipalities with a light colouring.

municipalities. The only exceptions are a few coastal municipalities and the most northern ones. With respect to the Poisson component, zero is contained in the central 95% credibility interval for several municipalities, especially in northern and central Norway. Since central and northern municipalities are less populated, estimates are often solely based on the prior and, hence, more uncertain. Note, the municipalities with the highest certainty on the covariate effects being non-zero are similar to those of Scheel et al. (2013).

While the amount of precipitation appears to have a general impact on the claim dynamics, Figure 3.3.5 indicates that the remaining covariates are only important for some municipalities.

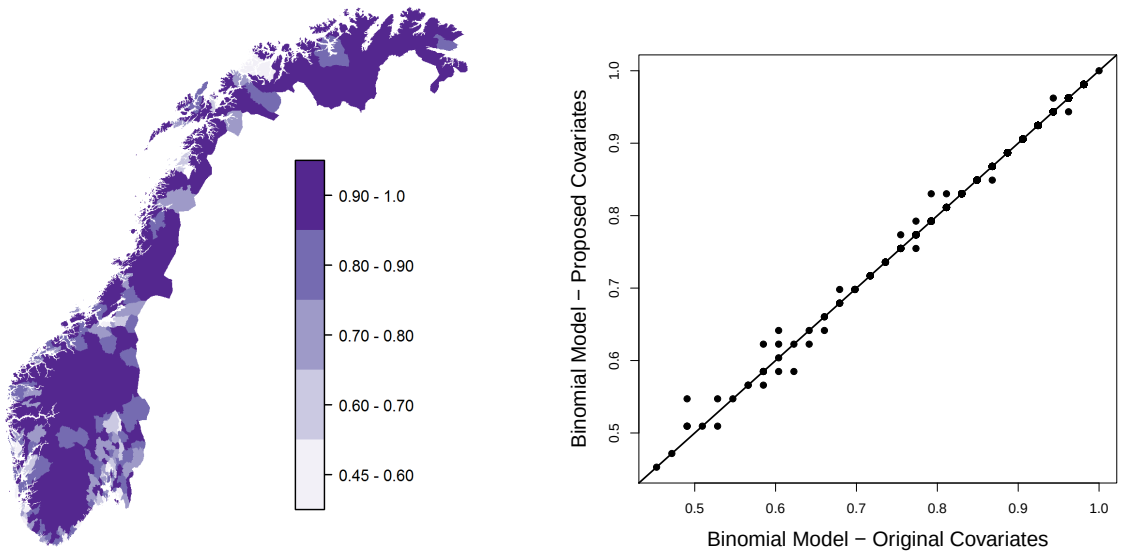


Figure 3.3.6: Ratio of correctly predicted claims type for the Binomial model with original covariates (left panel) and comparison of the predictive performance between the Binomial models fitted for original and proposed covariates (right panel).

High certainty on the covariate effect associated to the drainage run-off $D_{k,t}$ being non-zero is only found for eastern Norway for both the Binomial model and the hurdle component. Furthermore, zero is contained within the central 95% credibility interval for almost all municipalities with respect to $\delta_{k,3}$ the Poisson component. Scheel et al. (2013) also state that their results also indicate that drainage is important for the hurdle component but not the Poisson one. Next, the snow-water related covariates $S_{k,t}$ and $\Delta S_{k,t}$ have an impact for northern and eastern municipalities only. Again, zero is contained in the 95% credibility interval for most regions for the zero-truncated Poisson component, and the municipalities for which the results indicate an importance are rather spatially isolated. Precipitation on the day itself and the previous day appear the most important factor for the claim dynamics. Furthermore, the other three covariates considered here are important for some regions, especially eastern municipalities. Finally, the estimated models do not indicate a significant difference between the original and proposed covariates.

3.3.2 Predictive Performance

The predictive performance is assessed for the observations in 2001 which were excluded from the model estimation. Interest lies in predicting the number of claims within each week for each municipality. Equivalently to Scheel et al. (2013), each week is classified as one of three types:

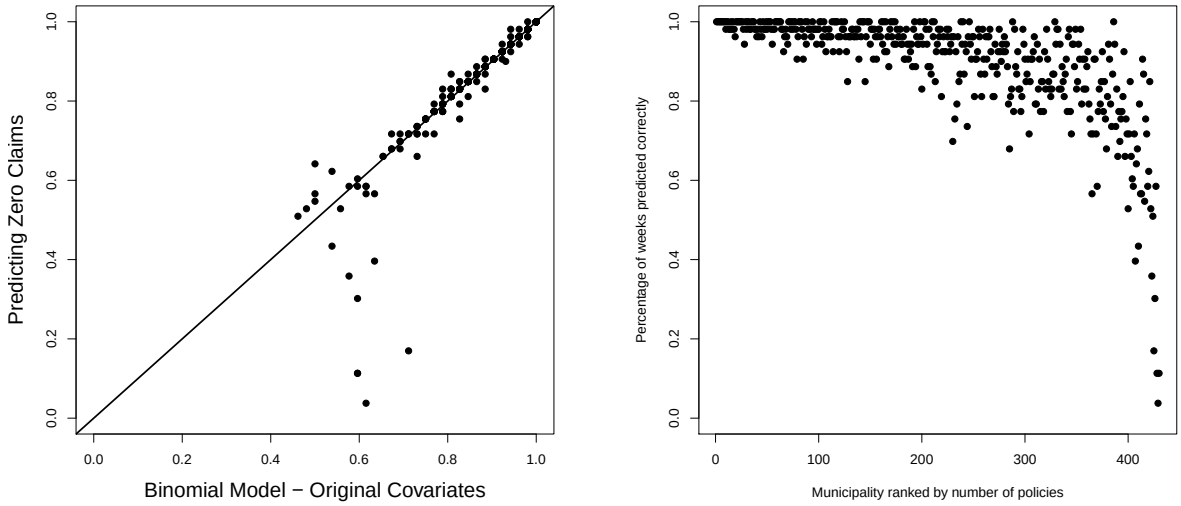


Figure 3.3.7: Comparison of the predictive performance for the Binomial model with original covariates and a weekly prediction of zero claims (left panel). The right panel illustrates the performance of predicting zero claims with respect to the rank of the municipality in terms of number of policies.

(i) No claims, (ii) 1-3 claims and (iii) more than four claims. The predicted type for a week yields to the one with highest posterior probability.

For the Binomial models, Figure 3.3.6 shows a good overall performance. On average, the predicted claim type is correct for 89% of the weeks but large differences between municipalities exist. While the predictive performance is generally above 95% for rural municipalities, it is below 50% for the more densely populated cities of Sarpsborg and Fredrikstad in south-east Norway. Comparison of the original and proposed covariates shows only small differences and the proposed covariates perform slightly better. In comparison to Scheel et al. (2013), the predictive performance is similar too and they find that the model yields poor predictions for Sarpbsborg in 2001.

The good performance for rural municipalities is due to the high occurrence of weeks with zero claims. To illustrate this aspect, the performance for a constant prediction of zero claims is compared to the Binomial model fir for the original covariates. The left plot in Figure 3.3.7 shows that a prediction of zero claims performs similarly if the Binomial model fits well too but has a lower rate of correct predictions otherwise. The municipalities for which the Binomial model yields better predictions are the ones with the highest number of policies (right panel in Figure 3.3.7). In addition to a prediction of zero claims, the performance of an approach based on the average daily number of claims over the training period has been considered; the plots

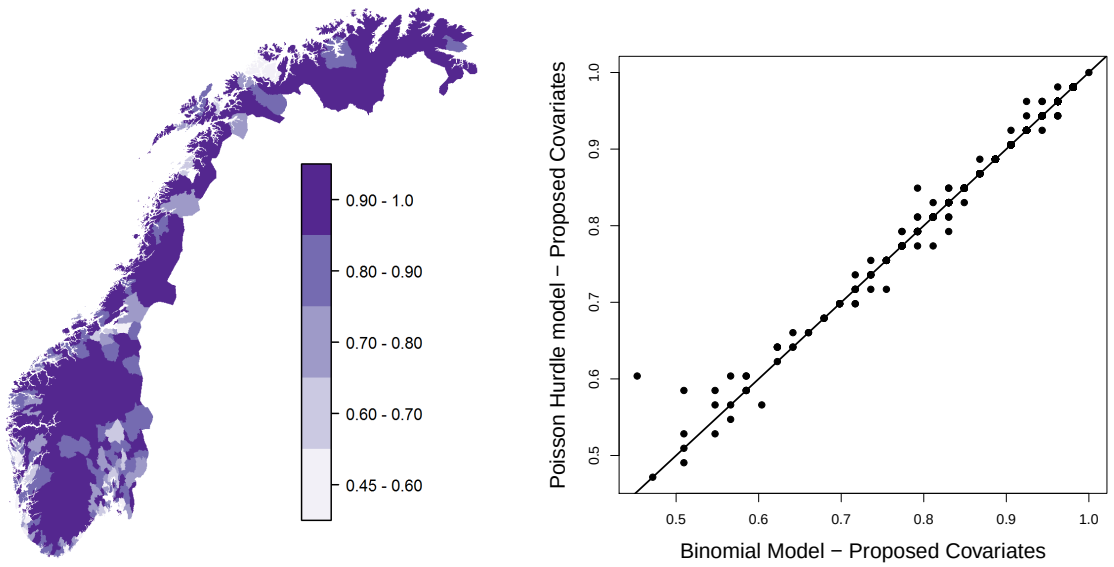


Figure 3.3.8: Ratio of correctly predicted claims type for the Poisson hurdle model with proposed covariates (left panel) and comparison of the predictive performance between the Binomial and Poisson hurdle model both with proposed covariates (right panel).

are provided in Appendix B.3.

Next, the results are compared to the fitted Poisson hurdle model. Figure 3.3.8 indicates that the model also performs quite well. Again, the predictive performance is better for rural municipalities than for urban ones. Furthermore, the right panel in Figure 3.3.8 shows that the differences between the Poisson hurdle and Binomial model are quite small for most municipalities, with the exception of two municipalities which exhibit a much better performance for the Poisson hurdle model. In particular for Fredrikstad, the Poisson hurdle model predicts the correct type for 32 weeks while the Binomial models do so for just 25 weeks. The results are also quite similar to the ones by Scheel et al. (2013).

In order to gain more insight, the municipalities of Oslo and Bergen are examined in more detail. Equivalently to Scheel et al. (2013), the weeks with the highest claim numbers and the highest precipitation levels in 2001 are considered and the 95% prediction intervals are derived. Table 3.3.1 provides the results obtained for all three models. The two Binomial models performed very similarly and differ only in small details. Note, the week in which 8 claims are observed for Oslo corresponds to a week with small amounts of snowfall and hence the predictions differ. With respect to the Poisson hurdle model, estimates are slighty better but no large differences are found. In conclusion, similarly to Scheel et al. (2013), both the Binomial and Poisson hurdle model tend to underpredict the number of claims.

Table 3.3.1: Posterior predictive median, 95% prediction interval and actual observation of the weekly-aggregated claim numbers for (a) the four weeks with the highest observations, (b) the four weeks with maximum total precipitation for the three different spatially varying regression models Oslo and Bergen.

Municipality	Period	Truth	Binomial Original	Binomial Proposed	Poisson hurdle Proposed
Oslo	(a)	11	6 (2,11)	6 (2,11)	6 (1,12)
		11	5 (1,10)	5 (1,10)	5 (1,11)
		8	3 (0,7)	2 (0,6)	3 (0,7)
		7	3 (0,7)	3 (0,7)	3 (0,7)
	(b)	11	6 (2,11)	6 (2,11)	6 (1,12)
		5	7 (2,12)	7 (2,12)	6 (1,13)
		6	5 (1,10)	5 (1,9)	4 (1,10)
		11	5 (1,10)	5 (1,10)	5 (1,11)
Bergen	(a)	7	3 (0,7)	3 (0,7)	3 (0,7)
		7	2 (0,6)	2 (0,6)	2 (0,7)
		6	2 (0,6)	2 (0,6)	2 (0,6)
		6	2 (0,6)	2 (0,5)	2 (0,6)
	(b)	5	6 (2,11)	6 (2,11)	4 (1,10)
		1	6 (2,12)	6 (2,12)	5 (1,11)
		3	4 (1,9)	4 (1,9)	3 (0,8)
		4	4 (1,9)	4 (1,9)	4 (0,9)

3.4 Discussion

This chapter performed a comparative study of a Binomial model, as used in many disease mapping approaches, and a Poisson hurdle model which accounts for the high frequency of zero claims in the insurance data. Note, both models are based upon the assumption that claims occur independently both within and across municipalities, conditional on the weather covariates. Additionally, modified covariates for the amount of precipitation and the difference in the snow-water equivalent were considered too. Similarly to Scheel et al. (2013), the parameters were assumed to vary spatially across municipalities. In order to borrow statistical information between municipalities, a ICAR prior (Section 2.1) was specified in order to define a spatial dependence of the model parameters. Results showed that the two considered sets of covariates perform very similarly in terms of both parameter estimates and predictive performance. Compared to the more complex Poisson hurdle model, results showed a slight improvement with respect to the predictive performance. However, all three models performed poorly in terms of predicting weeks with high numbers of claims which are of particular interest to the insurance companies.

The results in this chapter motivate the new methodology for monotonic regression and extreme value theory in the following chapters. One potential limitation is the assumption of linearity in the model setup. Since observations of $N_{k,t} > 1$ are very rare, the parameter estimates are mostly dominated by days with zero or one claim. However, the days with high claims are much more important in order to detect the weather events which induce the highest claim risk. A simple extension of the considered model may introduce a weighting in order to focus the analysis on days with high claim numbers. Nevertheless, there are further limitations to the linear model. As discussed in Section 1.4, the combined effect of the covariates may be non-linear. Further, the model also does not allow for any jumps in the regression function, also termed threshold effects. These arguments motivate the consideration of a more flexible monotonic regression approach (Section 2.2) in the following Chapters 4 and 5. However, its application is not straightforward since there exists no statistical model which defines a dependence structure on such functions. Hence, such a model is introduced in Chapter 4 and then embedded and estimated in a Bayesian framework. Chapter 5 then proposes an alternative version which considers the optimization-based approaches detailed in Section 2.2.2.

While the approaches in Chapters 4 and 5 introduce a more flexible process model, in terms of the Bayesian hierarchical modelling approach in this chapter, Chapter 6 considers an improved data model. While almost all days observe values of $N_{k,t}$ between 0 and 5, there only exist 11 days with more than 5 claims for Oslo and some are very large. Hence, the Poisson component may lack flexibility in terms of fitting these extremes. This motivates the application of extreme value models (Section 2.3). Chapter 6 considers an extension of the zero-truncated Poisson component in the Poisson hurdle model based upon the generalized Pareto distribution. Additional to the model fit with respect to the highest observations, the data structure itself is considered too. Firstly, the potential lag in the recording process of the claims discussed in Section 1.3 may partly cause the tendency to underpredict weeks with higher claims. Secondly, the considered covariate set is limited in terms of exploiting spatial and temporal patterns in the covariate set. This leads to the derivation of a temporal cluster algorithm with respective covariates which is also detailed in Chapter 6.

Finally, the results in the chapter indicate that the measurement for predictive performance applied by Scheel et al. (2013) has limitations. Although they argue that the year 2001 is generally representative, days with $N_{k,t} > 4$ are observed for neither Oslo nor Bergen in this period. Hence, it is difficult to examine model differences with respect to the highest claims.

Further, the aggregation over several days may make the assessment of certain aspects, such as the performance for days with snow-melt, difficult. Therefore, the model fit is assessed differently in the rest of this thesis. Chapter 4 assesses predictive performance on a daily basis while Chapter 6 considers both the Bayesian (Schwarz, 1978) and Deviance (Spiegelhalter et al., 2002) Information Criterion. Application of the former to the models considered in this chapter reveals that the Poisson hurdle model performs better than the Binomial models for the most populated municipalities. On the other hand, the added flexibility leads to only small improvements in the likelihood fit for more rural municipalities.

Chapter 4

Bayesian Spatial Monotonic Multiple Regression

4.1 Introduction

Geospatial data are considered in several areas, including ecology (Guttorp, 1991), forestry (Penttinen et al., 1992) and epidemiology (Waller and Gotway, 2004). Data in a locally aggregated form, *lattice data* (Cressie, 1993), are common due to practicality or confidentiality concerns and are typically over an irregular lattice. Statistical methods for such area-level data model associations between a response variable and a set of explanatory variables via a regression function, whilst accounting for potential spatial dependence in the model parameters. To introduce spatial dependence, a neighbourhood structure, often based upon the arrangement of the areal units (regions) on a map, is typically defined in form of an adjacency matrix.

Most modelling frameworks in spatial statistics assume the regression function to have the same shape across all regions (Waller and Gotway, 2004; Wakefield, 2007; Waller and Carlin, 2010). Spatial variation is then typically accommodated via a spatially structured random effect on the intercept (baseline level) with dependence being, for instance, defined by an intrinsic conditional autoregressive (ICAR) prior (Besag, 1974; Besag et al., 1991; Rue and Held, 2005). Some applications, however, need to allow for a spatial varying association between response and explanatory variables (Bell et al., 2004; Zhang and Shi, 2004; Cahill and Mulligan, 2007; Waller et al., 2007) and, hence, require a more flexible modelling framework. Statistical methods for such scenarios are available for generalized linear (Brunsdon et al., 1998; Fotheringham et al., 2002; Assunção, 2003; Congdon, 2003; Scheel et al., 2013) and additive models (Congdon,

2006). Whilst more flexible, these approaches are limited in terms of recovering discontinuities (threshold effects) as continuity of the regression function is assumed. More precisely, abrupt changes in the regression surface are not captured unless these are explicitly included; negligence of such effects may result in a bias due to oversmoothing (Bowman and Azzalini, 1997).

Since the assumption of continuity may be inappropriate, we substitute it by one of monotonicity, an important assumption in several applications (Royston, 2000; Farah et al., 2013; Wilson et al., 2014). Whilst continuity can, in general, not be verified, tests of monotonicity for the underlying process are available (Bowman et al., 1998; Ghosal et al., 2000; Scott et al., 2015). Conditional on the monotonicity constraint, we develop methodology which estimates the form of the association between the response and explanatory variables for each region; whilst exploiting any neighbourhood structure.

The estimation of a single multivariate, monotonic function is considered in several statistical areas and is usually referred to as *isotonic regression*. Early publications discuss inference on parameter values under monotonic constraints (Ayer et al., 1955; Brunk, 1955; Barlow and Brunk, 1972) and solution algorithms are available in the optimization literature (Brunk et al., 1957; Luss et al., 2012). Isotonic regression is further considered for additive (Bacchetti, 1989; Morton-Jones et al., 2000; Tutz and Leitenstorfer, 2007) and high-dimensional models (Fang and Meinshausen, 2012; Bergersen et al., 2014), in functional data analysis (Ramsay, 1998; Ramsay and Silverman, 2005) and Bayesian nonparametrics (Holmes and Heard, 2003; Shively et al., 2009; Saarela and Arjas, 2011; Lin and Dunson, 2014). To apply these techniques for the modelling of spatially varying regression functions, a dependence model for monotonic functions, without the additional assumption of continuity, is required. However, little, to no, research exists on dependence models for the class of discontinuous functions.

We introduce new Bayesian, non-parametric, methodology, *Bayesian Spatial Monotonic Multiple Regression* (BSMMR), which allows for dependence modelling of regression functions under the assumptions of monotonicity and boundedness. The regional (areal) monotonic functions are each represented by a set of marked point processes, permitting both smooth contours and threshold effects in the regression surface. Potential spatial dependence is modelled via the specification of a novel joint prior distribution on the monotonic functions, which is constructed based upon a flexible pair-wise discrepancy measure. The defined prior allows the functional dependence to be either constant, increasing or decreasing with an increasing functional level. In order to tune the prior, we propose a new algorithm, *EGO-CV*, which combines the concepts

of cross-validation and Bayesian global optimization. Realizations of the posterior are obtained by a reversible jump MCMC (RJMCMC) algorithm (Green, 1995) and facilitate analysis with regard to threshold effects, variable selection, prediction and extrapolation.

The remainder of this chapter is organized as follows: Section 4.2 details the model specification and inferential framework of the BSMMR approach, including the newly defined dependence model and the *EGO-CV* algorithm. Performance and sensitivity of our approach is assessed via multiple simulation studies in Section 4.3. In Section 4.4, the methodology is applied to Norwegian insurance and meteorological data, with a view to investigating weather related claim dynamics over an area. The chapter concludes with a summary and discussion in Section 4.5.

4.2 Modelling and Inference

4.2.1 Probability Model, Notation and Outlook

Consider a set of K (contiguous) regions whose neighbourhood structure is given by an adjacency matrix or a lattice graph (regular or irregular). Let $y_k \in \mathbb{R}$ and $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,m}) \in \mathbb{R}^m$ denote the response and explanatory variables, respectively, for region $k = 1, \ldots, K$. The probability model (likelihood) is defined as

$$f\left(y_k \mid \lambda_k(\mathbf{x}_k), \boldsymbol{\theta}_k\right), \tag{4.2.1}$$

where $\lambda_k : \mathbb{R}^m \to [\delta_{\min}, \delta_{\max}]$ refers to the monotonic regression function for region k; a mapping for which the functional level $\lambda_k(\mathbf{x}_k)$ is assumed to lie within a prespecified interval $[\delta_{\min}, \delta_{\max}]$. Monotonicity is here defined in terms of the partial Euclidean ordering $\preceq$, that is, $\forall\ \mathbf{u}, \mathbf{v} \in \mathbb{R}^m$ such that $\mathbf{u} \leq \mathbf{v}$ component-wise, then $\lambda_k(\mathbf{u}) \leq \lambda_k(\mathbf{v})$. The vector $\boldsymbol{\theta}_k$ denotes additional, potentially spatially varying, model parameters which are a priori independent of $\lambda_1, \ldots, \lambda_K$.

In what follows, we perform inference on λ_1 through λ_K while accounting for potential spatial structure in these functions. Each λ_k, $k = 1, \ldots, K$, is estimated over an associated closed set $X_k \subset \mathbb{R}^m$ which is permissibly different across regions. In applications, X_k and the boundaries, $\delta_{\min}$ and $\delta_{\max}$, may be defined in terms of the observed explanatory variables and responses, respectively. For instance, $\delta_{\min}$ may be set to the minimum observed response across the K regions if y_k follows a Gaussian distribution with mean $\lambda_k(\mathbf{x}_k)$. Sections 4.2.2 and 4.2.3 complete the Bayesian framework by defining a joint prior on $\lambda_1, \ldots, \lambda_K$ while Sections 4.2.4

and 4.2.5 detail the estimation procedure.

4.2.2 A Spatial Dependence Model for Monotonic Functions

Interest lies in the imposition of a spatial structure on the K monotonic functions $\lambda_1, \ldots, \lambda_K$ defined in Section 4.2.1. In a Bayesian framework, beliefs on spatial dependence in the model parameters are typically accommodated via the specification of a prior distribution. Since little research exists on dependence models for monotonic functions, we derive a joint density $\pi(\lambda_1, \ldots, \lambda_K)$ which favours similarity of λ_1 through λ_K by penalizing differences in their functional levels. For notational simplicity, $\lambda_k(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^m,\ k = 1, \ldots, K$, is assumed to be non-negative since, in general, one would naturally consider $\lambda_k(\mathbf{x}) - \delta_{\min}$ instead of $\lambda_k(\mathbf{x})$.

In the first step, we introduce a pair-wise discrepancy measure for function pairs λ_k and $\lambda_{k'}$ which evaluates their functional difference over a set $W_{k,k'} \subset \mathbb{R}^m$. Section 4.3.3, later, demonstrates that $W_{k,k'}$ may be defined such that it permits borrowing of statistical information for extrapolation, in particular, in case that the observation spaces for λ_k and $\lambda_{k'}$ differ. The discrepancy measure should be minimal if, and only if, λ_k and $\lambda_{k'}$ are equal, and increase with an increasing difference in the functional levels. An intuitive choice which satisfies these properties is the integrated squared difference

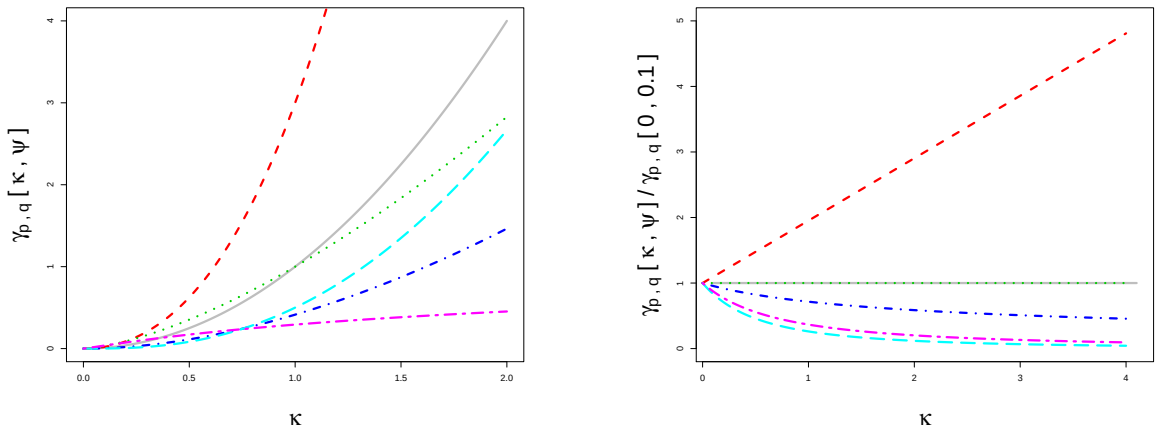
$$\int_{W_{k,k'}} [\lambda_k(\mathbf{x}) - \lambda_{k'}(\mathbf{x})]^2 \ \mathrm{d}\mathbf{x}. \tag{4.2.2}$$

The defined measure in (4.2.2) is, however, rather inflexible as it penalizes functional differences regardless of the exact functional levels. In certain cases, differences in the lower, or higher, functional levels should be particularly downweighted, or avoided. For example, measurement errors in the response may be expected to increase with the values of the explanatory variables.

To achieve greater flexibility, we substitute the squared distance $[\lambda_k(\mathbf{x}) - \lambda_{k'}(\mathbf{x})]^2$ in (4.2.2) by

$$\gamma_{p,q}\left[\lambda_k(\mathbf{x}), \lambda_{k'}(\mathbf{x})\right] := \Big|\, [1 + \lambda_k(\mathbf{x})]^p - [1 + \lambda_{k'}(\mathbf{x})]^p \,\Big| \times |\lambda_k(\mathbf{x}) - \lambda_{k'}(\mathbf{x})|^q \ ,\ p \in \mathbb{R}\ ,\ q \geq 0. \tag{4.2.3}$$

The functional levels in the first modulus term are increased by 1 to ensure numerical stability for the case $p < 0$, as $\lambda_k(\mathbf{x})$ and $\lambda_{k'}(\mathbf{x})$ may be close or equal to 0. In other words, $\gamma_{p,q}\left[\lambda_k(\mathbf{x}), \lambda_{k'}(\mathbf{x})\right]$ is bounded for every choice of p and q by shifting both functional levels by 1. Note, the setting $p = q = 1$ results in the squared distance in expression (4.2.2).



(a) $\psi = 0$ **(b)** $\psi = \kappa + 0.1$

Figure 4.2.1: Behaviour of $\gamma_{p,q}[\kappa, \psi]$ for (——) $p = 1,\ q = 1$, (- - -) $p = 2,\ q = 1$, (·····) $p = 1,\ q = 0.5$, (-·-·-) $p = 0.5,\ q = 1$, (-·--) $p = -1,\ q = 2$ and (-·-·-) $p = -0.5,\ q = 0.1$. Both plots examine $\gamma_{p,q}[\kappa, \psi]$ with respect to κ, subject to (a) $\psi = 0$ is fixed and (b) $\psi = \kappa + 0.1$ is fixed. Note, the curves of the first and third setting are identical in (b).

We illustrate, Figure 4.2.1, the behaviour of $\gamma_{p,q}[\lambda_k(\mathbf{x}), \lambda_{k'}(\mathbf{x})]$ at a fixed point $\mathbf{x} \in \mathbb{R}^m$ for different settings of p and q. For notational brevity, let $\kappa := \lambda_k(\mathbf{x})$ and $\psi := \lambda_{k'}(\mathbf{x})$. In Figure 4.2.1a, $\psi = 0$ and, as desired, $\gamma_{p,q}[\kappa, \psi]$ increases with κ for all values of p and q. Figure 4.2.1b then examines the dependence of $\gamma_{p,q}[\kappa, \psi]$ on κ, subject to $\psi = \kappa + 0.1$ being fixed. The plot shows that the fixed difference, $\psi - \kappa = 0.1$, is penalized more heavily for higher κ if $p > 1$ while being penalized less heavily for $p < 1$. A constant penalty is induced for $p = 1$. For instance, the setting $p = 2, q = 1$ leads to a five-fold increase in $\gamma_{p,q}[\kappa, \psi]$ when $\kappa = 4$ compared to when $\kappa = 0$. In conclusion, p allows the penalty for the functional difference between λ_k and $\lambda_{k'}$ to vary with the functional levels. Since $\gamma_{p,q}[\cdot, \cdot]$ in (4.2.3) fulfills the desired properties, we now formally define the discrepancy measure as

$$D_{p,q}(\lambda_k, \lambda_{k'}) := \int_{W_{k,k'}} \gamma_{p,q}[\lambda_k(\mathbf{x}), \lambda_{k'}(\mathbf{x})]\ \mathrm{d}\mathbf{x}\ ,\ p \in \mathbb{R}\ ,\ q \geq 0. \tag{4.2.4}$$

The dependence model for $\lambda_1, \ldots, \lambda_K$ is then defined as a Gibbs distribution with the measure $D_{p,q}$ in (4.2.4) as a pair-potential. Formally, the joint density for the K-set of monotonic functions is given by

$$\pi(\lambda_1, \ldots, \lambda_K \mid \omega) \propto \prod_{1 \leq k < k' \leq K} \exp\Big[-\omega \cdot d_{k,k'} \cdot D_{p,q}(\lambda_k, \lambda_{k'}) \Big],\ \omega \geq 0, \tag{4.2.5}$$

where the product is defined over all pairs of regions. The non-negative constant $d_{k,k'}$ describes

our belief on the degree of similarity of λ_k and $\lambda_{k'}$. A natural choice is $d_{k,k'} = 1$ if regions k and k' are adjacent (share a border) and 0 otherwise, a common setting in Ising or ICAR models. Such a choice reduces the computational cost as the integral in (4.2.4) needs only to be evaluated for pairs of adjacent regions. The degree of dependence increases in ω, with $\omega = 0$ corresponding to $\lambda_1, \ldots, \lambda_K$ being independent. Further, p allows the dependence to vary with the functional levels; sensitivity on p and q is explored in Section 4.3.2. In the next Section 4.2.3, we specify an individual prior for each $\lambda_k,\ k = 1, \ldots, K$, which, combined with (4.2.5), results in a composite prior.

4.2.3 Marked Point Process Prior Formulation

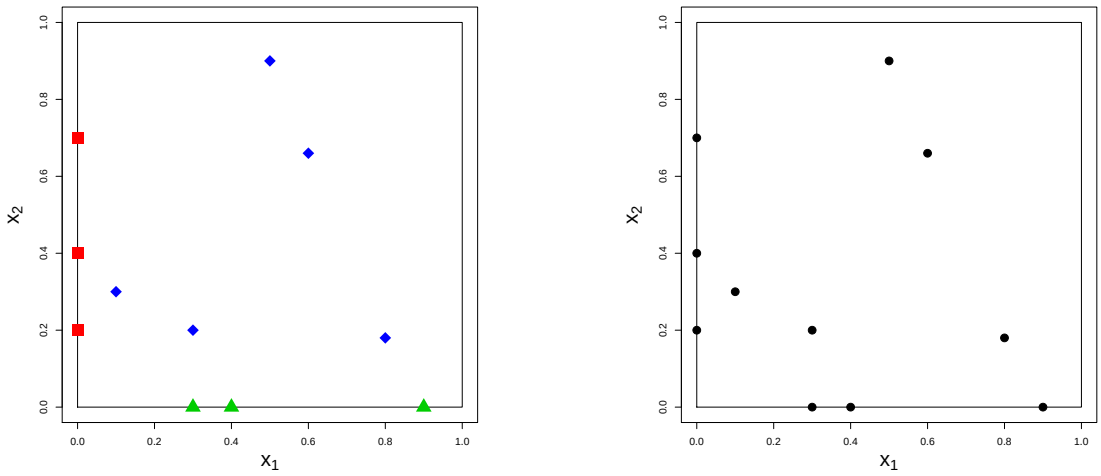
We specify an individual prior model for each $\lambda_k : X_k \to [\delta_{\min}, \delta_{\max}]\,,\ k = 1, \ldots, K$. Several prior distributions for a single monotonic function are proposed in the literature, for instance, an ordered Dirichlet process (Gelfand and Kuo, 1991) or a constrained spline model (Shively et al., 2009). We define a similar prior to that of Saarela and Arjas (2011). Specifically, λ_k is postulated to be a step function, that is, λ_k is monotonic and piecewise constant, with $\lambda_k(\mathbf{x}) \in [\delta_{\min}, \delta_{\max}]\,,\ \forall\ \mathbf{x} \in X_k$. This prior setting is highly flexible as any monotonic, bounded function can be approximated up to a desired degree of accuracy by increasing the number of steps. Furthermore, the posterior mean, induced by the likelihood in expression (4.2.1), may be a smooth function, given the model permits variability in the number, locations and heights of the steps (Heikkinen and Arjas, 1998; Heikkinen, 2003). Consequently, both smooth and discontinuous functional shapes can be recovered.

The step function λ_k is represented via its characteristics, namely the location and height of the jumps, which define a marked point process on X_k. Following Saarela and Arjas (2011), we consider a set of I marked point processes, $\boldsymbol{\Delta}_k = (\Delta_{k,1}, \ldots, \Delta_{k,I})$, instead of a single marked point process; the benefits of this approach are discussed later in this section. Here, the marked point processes $\Delta_{k,1}, \ldots, \Delta_{k,I}$ are defined on non-empty subsets $X_{k,1}, \ldots, X_{k,I}$, respectively, where $\bigcup_{i=1}^{I} X_{k,i} = X_k$. In what follows, the representation of λ_k via $\boldsymbol{\Delta}_k$ is formalized.

Consider the set $\boldsymbol{\Delta}_k = (\Delta_{k,1}, \ldots, \Delta_{k,I})$ and let the marked point process $\Delta_{k,i},\ i = 1, \ldots, I$, be of the form

$$\Delta_{k,i} = \{(\boldsymbol{\xi}_{k,i,j}, \delta_{k,i,j}) \ :\ j = 1, \ldots, n(\Delta_{k,i})\}\,, \tag{4.2.6}$$

where $\boldsymbol{\xi}_{k,i,j}$ and $\delta_{k,i,j}$ refer to a location and associated mark, respectively, and $n(\Delta_{k,i})$ is the



(a) Representation of λ_k via $\boldsymbol{\Delta}_k$ **(b)** Representation of λ_k via Γ_k

Figure 4.2.2: Locations for a marked point process representation of λ_k on $X_k = [0,1]^2$ via (a) $\boldsymbol{\Delta}_k = \{\Delta_{k,1}, \Delta_{k,2}, \Delta_{k,3}\}$ defined in terms of the non-empty subsets of the covariate set $\{1,2\}$ and (b) a single marked point process Γ_k. The processes $\Delta_{k,1}$ (▲) and $\Delta_{k,2}$ (■) are defined on the one-dimensional covariate subsets, $\{1\}$ and $\{2\}$, respectively while $\Delta_{k,3}$ (◆) is defined on $\{1,2\}$.

number of points in $\Delta_{k,i}$. Functional monotonicity is preserved via the imposition of a monotonic constraint on the marks. Specifically, if $\boldsymbol{\xi}_{k,i,j} \preceq \boldsymbol{\xi}_{k,i',j'}$, then $\delta_{k,i,j} \leq \delta_{k,i',j'}$, $i, i' \in \{1,\ldots,I\}$, $j \in \{1,\ldots,n(\Delta_{k,i})\}$, $j' \in \{1,\ldots,n(\Delta_{k,i'})\}$. The functional level $\lambda_k(\mathbf{x})$ is then defined by $\boldsymbol{\Delta}_k$ as the highest mark $\delta_{k,i,j}$ such that $\mathbf{x}$ imposes a monotonic constraint on the associated location $\boldsymbol{\xi}_{k,i,j}$. Formally, $\lambda_k(\mathbf{x})$ is given by

$$\lambda_k(\mathbf{x}) = \max_{i,j} \{\delta_{k,i,j} \ : \ \boldsymbol{\xi}_{k,i,j} \preceq \mathbf{x}\}. \tag{4.2.7}$$

While there is no restriction on the number, I, of marked point processes or the associated subsets, we define $X_{k,1},\ldots,X_{k,I}$ based on the non-empty subsets of the covariate set $\{1,\ldots,m\}$. For instance, in the case $m = 2$ and $X_k = [0,1]^2$, this yields $I = 3$ processes (Figure 4.2.2a) with, for instance, $\Delta_{k,1}$ containing the locations with the second component being 0: $\boldsymbol{\xi}_{k,1,j} = (\cdot, 0)$, $j = 1,\ldots,n(\Delta_{k,1})$. Although λ_k may also be represented via a single marked point process Γ_k (Figure 4.2.2b), the proposed approach has indeed benefits in terms of variable selection, and these are explained in the following:

Assume X_k is scaled to $X_k = [0,1]^m$ and suppose that, for instance, the explanatory variable $x_{k,1}$ is redundant. Hence, the regression function λ_k is constant with increasing $x_{k,1}$, that is, $\lambda_k(\mathbf{x}) = \lambda_k(\mathbf{x} + \boldsymbol{\epsilon}_1)$, $\forall\ \mathbf{x} \in X_k$, where $\boldsymbol{\epsilon}_1 = (\epsilon, 0, \ldots, 0)$ has positive first component and is zero

otherwise. As the points in both Γ_k and $\boldsymbol{\Delta}_k$ represent the jumps of λ_k, the locations in Γ_k and $\boldsymbol{\Delta}_k$ are 0 in the first component. For instance in the case $m = 2$, the redundancy of $x_{k,1}$ implies that locations lie on the vertical x_2-axis in Figure 4.2.2. In terms of $\boldsymbol{\Delta}_k$, all points are thus contained in $\Delta_{k,2}$ while $\Delta_{k,1}$ and $\Delta_{k,3}$ are empty. Consequently, $n(\Delta_{k,i}),\ i = 1, \ldots, I$, provides an indicator of the redundancy of explanatory variables, in addition to examination of the functional shape.

The defined association between λ_k and $\boldsymbol{\Delta}_k$ results in a well-defined mapping between the space of step functions and that of marked point processes with monotonic constraints. Thus, we can set a prior for λ_k via a prior specification for $\boldsymbol{\Delta}_k$. The number of steps of λ_k, $n(\boldsymbol{\Delta}_k) = \sum_{i=1}^{I} n(\Delta_{k,i})$, is taken to be geometrically distributed with probability $1/\eta,\ \eta > 1$, where $n(\boldsymbol{\Delta}_k) = 0$ corresponds to λ_k being constant. This setting favours parsimony: λ_k to have a small number of steps. Given $n(\boldsymbol{\Delta}_k)$, the vector $[n(\Delta_{k,1}), \ldots, n(\boldsymbol{\Delta}_{k,I})]$ is uniformly distributed over the set of possibilities of allocating $n(\boldsymbol{\Delta}_k)$ points to the I processes. For $\Delta_{k,i},\ i = 1, \ldots, I$, the $n(\Delta_{k,i})$ locations $\boldsymbol{\xi}_{k,i,1}, \ldots, \boldsymbol{\xi}_{k,i,n(\Delta_{k,i})}$ are a priori uniformly distributed on $X_{k,i}$. The set of marks $\boldsymbol{\delta}_k = \{\delta_{k,i,j} : j = 1, \ldots, n(\Delta_{k,i}),\ i = 1, \ldots, I\}$ is then uniformly distributed on $[\delta_{\min}, \delta_{\max}]$, subject to the monotonic constraints imposed by the associated set of locations $\boldsymbol{\xi}_k = \{\boldsymbol{\xi}_{k,i,j} : j = 1, \ldots, n(\Delta_{k,i}),\ i = 1, \ldots, I\}$. Based on this Bayesian hierarchical model, the prior for $\boldsymbol{\Delta}_k$ yields to

$$\phi(\boldsymbol{\Delta}_k \mid \eta) = \pi[\boldsymbol{\delta}_k \mid \boldsymbol{\xi}_k] \times \prod_{i=1}^{I} \pi[\boldsymbol{\xi}_{k,i} \mid n(\Delta_{k,i})] \times \pi[n(\Delta_{k,1}), \ldots, n(\Delta_{k,I}) \mid n(\boldsymbol{\Delta}_k)] \times \pi[n(\boldsymbol{\Delta}_k) \mid \eta]. \tag{4.2.8}$$

The density $\phi(\boldsymbol{\Delta}_k \mid \eta)$ is proper and analytically tractable; see Appendix C.1. A conjugate Beta prior may be specified to perform inference on $1/\eta$.

We impose a spatial structure on $\boldsymbol{\Delta}_1$ through $\boldsymbol{\Delta}_K$ by combining $\phi(\boldsymbol{\Delta}_k \mid \eta)$ in (4.2.8) with the dependence model in (4.2.5). The joint prior for the K-set $(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K)$ is then formally given as

$$\pi(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K \mid \omega, \eta) \propto \prod_{1 \le k < k' \le K} \exp\Big[-\omega \cdot d_{k,k'} \cdot D_{p,q}(\lambda_k, \lambda_{k'}) \Big] \times \prod_{k=1}^{K} \phi(\boldsymbol{\Delta}_k \mid \eta), \tag{4.2.9}$$

where λ_k and λ'_k refer to the step functions represented by $\boldsymbol{\Delta}_k$ and $\boldsymbol{\Delta}_{k'}$, respectively. Note, the prior for $\boldsymbol{\Delta}_k$ in (4.2.9) converges to (4.2.8) as $\omega \to 0$ and is proper since the first term lies within $(0, 1]$ and the second term is a proper density function. Further, $D_{p,q}(\lambda_k, \lambda_{k'})$ can be computed

efficiently as λ_k and λ'_k are step functions and, hence, the integral simplifies to a sum.

The likelihood (4.2.1) and prior (4.2.9) fully specify a posterior density for $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$ which is proportional to

$$\left[\prod_{k=1}^{K}\prod_{t=1}^{T_k} f\left(y_{k,t} \mid \lambda_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)\right] \times \pi\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K \mid \omega, \eta\right), \tag{4.2.10}$$

where T_k denotes the number of observations for region k and λ_k is the step function represented by $\boldsymbol{\Delta}_k$. Section 4.2.4 details the sampling of $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$ from the posterior density in (4.2.10).

In a fully Bayesian framework, we would want to specify priors for the parameters η and ω in order to infer on these too. However, the full conditional posterior density for η and ω is not obtainable as the normalizing constant in (4.2.9) is intractable. This issue led to our novel inferential approach for ω which is detailed in Section 4.2.5. To select η, one may first consider the posterior density resulting from the likelihood (4.2.1) and the prior $\phi\left(\boldsymbol{\Delta}_k \mid \eta\right)$. In this case, the normalizing constant of the conditional posterior density for η is analytically tractable. Hence, we can sample realizations from this density and then set η to the posterior mean when inferring on ω and $\boldsymbol{\Delta}_1$ through $\boldsymbol{\Delta}_K$. Performance of this posterior mean approach is explored in Section 4.3.

4.2.4 Inference and Analysis of the Marked Point Processes

We outline a RJMCMC algorithm to sample realizations of $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$ from the posterior density in (4.2.10). Each point $(\boldsymbol{\xi}_{k,i,j}, \delta_{k,i,j})$ in $\boldsymbol{\Delta}_k$ is considered as one parameter with the number of points, hence the dimension of the parameter space, unknown. Initially, $\boldsymbol{\Delta}_1 \ldots, \boldsymbol{\Delta}_K$ are empty and λ_1 through λ_K are defined as constant with level δ_{min}. The sets $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$ are then updated sequentially with moves being defined similarly to Saarela and Arjas (2011). More precisely, one of three moves, implying local changes in the regression surface, is randomly proposed for one of the processes $\Delta_{k,1}, \ldots, \Delta_{k,I}$, for region k, $k = 1, \ldots, K$, in turn.

Assume that the process $\Delta_{k,i}$ has been sampled to be updated. The first move, *Birth*, adds a point $(\boldsymbol{\xi}^*, \delta^*)$ to $\Delta_{k,i}$, where $\boldsymbol{\xi}^*$ is sampled uniformly on $X_{k,i}$. Given $\boldsymbol{\xi}^*$, δ^* is sampled uniformly, subject to monotonicity being preserved. A *Death* removes a point from the current process, maintaining reversibility. The last move, *Shift*, leads to a 'local' change in both the location and level of an existing point in $\Delta_{k,i}$, subject to the monotonic structure of the locations being maintained. See Appendix C.2 for more details on the algorithm.

Sampling $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$ via this RJMCMC algorithm has one small limitation. If, for instance, $X_k = [0,1]^m$, then $\lambda_k(0,\ldots,0) = \delta_{\min}$ as $\xi^* = (0,\ldots,0)$ is proposed with probability 0. To address this, one may define the decomposition $\lambda_k(\mathbf{x}_k) := \mu_k + \widetilde{\lambda}_k(\mathbf{x}_k)$ and then infer on $\mu_k \in \mathbb{R}$ and $\widetilde{\lambda}_k$ separately; this approach is applied in Sections 4.3 and 4.4. Further alternatives are (i) definition of an Ith$+1$ process $\Delta_{k,0} = \{((0,0), \delta_{\min})\}$ for which only changes in $\delta_{\min}$ are proposed and (ii) estimation of λ_k on an extended set, such as $X_k = [-0.1, 1.0]^m$. The decomposition is, however, more flexible as it allows for a similar functional behaviour with respect to the covariates but different baseline levels; this aspect will be particularly useful in Section 4.4.

Realizations from the posterior distribution are rich and facilitate detailed analysis of the functions $\lambda_1, \ldots, \lambda_K$. Thinning is performed on the sampled Markov chains in order to reduce autocorrelation and for storage reasons. Posterior mean estimates $\widehat{\lambda}_k$ of λ_k are obtained by averaging over the stored realizations. The mean and quantiles of the posterior distribution are accessible for any $\mathbf{x} \in X_k$ by deriving the functional level $\lambda_k^{(r)}(\mathbf{x})$ for each sample r, $r = 1, \ldots, R$. Further, the sampled marked point processes facilitate the detection of discontinuities; see Appendix C.3.

4.2.5 Estimation of the Prior Parameter ω

Performance of our approach relies on a suitable ω in (4.2.9). If ω is too high, the prior dominates the posterior distribution and spatial variation in $\boldsymbol{\Delta}_1$ to $\boldsymbol{\Delta}_K$ is oversmoothed. Otherwise, the data may be overfitted if ω is too small. In a Bayesian framework, ω should ideally be updated within the RJMCMC algorithm in Section 4.2.4, for instance, via an additional Gibbs step. However, the normalizing constant of (4.2.9), which depends on $\delta_{\min}$, $\delta_{\max}$, ω, and η, is intractable. We considered several approaches to handle intractable normalizing constants, including Beaumont et al. (2002), Møller et al. (2006) and Andrieu and Roberts (2009). Nevertheless, these approaches cannot be adapted since efficient sampling from the prior distribution in (4.2.9) is infeasible. Approximate Bayesian computation, for instance, would require multiple samples from (4.2.9) for each update of ω. Hence, we estimate ω via a separate approach, prior to inferring on $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K$.

One possible approach to find a suitable value for ω is s-fold cross-validation, that is, the data for each of the K regions are split into s subsets of equal size. The RJMCMC in Section 4.2.4 is then performed s times with varying training and test data. Parameter values are then, for instance, compared by the mean squared error (MSE) of the posterior predictive functional mean

of the test data points derived by Monte Carlo integration. Nevertheless, the number of evaluated values for ω should be as small as possible since the RJMCMC algorithm is computationally expensive.

We reduce the number of evaluations by applying Bayesian optimization, in particular, the *efficient global optimization (EGO)* algorithm (Jones et al., 1998). Despite having potential to reduce the number of evaluations substantially, this concept has, to the best of our knowledge, never been applied in combination with cross-validation. In the following, we outline a new algorithm, termed *EGO-CV*, which combines the two concepts and aims to reduce the computational time.

The *EGO* concept postulates a sequential design strategy to detect global extrema of a black-box function g. *EGO* is widely applied in simulations if g is costly to evaluate and the parameter space is relatively small (Roustant et al., 2012). The rationale is to model g by a Gaussian process G which is updated sequentially with proposals being based on the expected improvement. More formally, the expected improvement at an arbitrary point z given G and the current optimum g_{opt} of the unknown g is defined as

$$\mathbb{E}\left[\max\left(g_{opt} - G(z), 0\right)\right] \tag{4.2.11}$$

and represents the potential of $g(z)$ being smaller than g_{opt}. Proposals are considered until the expected improvement falls below a critical value for all z, corresponding to the current g_{opt} being presumably close to the unknown global minimum of g. As *EGO* balances between a local exploration of the values likely to provide 'good model fit' and a global search (to avoid a local but not global minimum), a suitable solution is generally found after a reasonable number of evaluations.

In the context of estimating ω, interest lies in the global minimum of the unknown cross-validation function, $\mathrm{CV}(\omega)$, and a general layout of our *EGO-CV* approach is given in Algorithm 4.1. First, an upper bound is derived as *EGO* can only be applied to a closed set. Hence, an initial bound ω_u is increased until the associated MSE is sufficiently greater than the one for $\omega = 0$. More clarity is provided in lines 2 to 7 in Algorithm 4.1. An upper bound based on $\beta = 2$ in Algorithm 4.1 proved reasonable in all simulations. Once the bound is fixed, an initial proposal $\omega^* \in [0, \omega_u]$ is made, guaranteeing that the process G in (4.2.11) is fitted with at least 3 data points. After performing cross-validation for ω^*, *EGO* is performed until the expected

improvement falls below the critical value α. The value ω_{opt} providing the lowest MSE is then used in the conclusive RJMCMC algorithm. In this work, *EGO* is performed by the `DiceOptim` `R` package by Roustant et al. (2012). To reduce dependence on the split, multiple repetitions with the same value for ω are performed and the variance of the MSE across the repetitions is used, additional to the average MSE, to fit G.

Algorithm 4.1 *Efficient Global Optimization within Cross-Validation (EGO-CV)*

Require: Parameter settings for both the RJMCMC algorithm and cross-validation
Require: Initial upper bound ω_u, critical value α, factor β
1: Initialize expected improvement `max_EI`$> \alpha$
2: Perform cross-validation for $\omega = 0$ and ω_u, and store results CV(0) and CV(ω_u)
3: **while** CV(ω_u) $< \beta$ CV(0) **do**
4: Increase upper bound ω_u
5: Perform cross-validation for new ω_u and store CV(ω_u)
6: **end while**
7: Set initial proposal ω^*, e.g. $\omega^* = \omega_u/2$
8: **while** `max_EI` $> \alpha$ **do**
9: Perform cross-validation for ω^* and store CV(ω^*)
10: Perform *EGO* on the interval $[0, \omega_u]$ and update ω^* and `max_EI`
11: **end while**
12: **return** Parameter value ω_{opt} which provides smallest error for the potential values

4.3 Simulation Study

4.3.1 Introduction

We aim to demonstrate that our approach is highly flexible, in terms of reconstructing a wide range of regression surfaces, and to appraise the value for sharing statistical information spatially between regions. Multiple simulations studies are performed in order to

1. Illustrate that BSMMR in combination with the *EGO-CV* algorithm improves estimates if similarities between functional shapes exist, and is also robust if the functions are dissimilar.

2. Examine sensitivity on the prior parameters p, q and η in expression (4.2.9).

Improvements in the estimates are discussed with respect to $\omega = 0$, a setting which imposes no dependence. Performance is evaluated via the posterior mean estimate $\widehat{\lambda}_k$ of the true function λ_k. Specifically, the absolute difference $\left|\lambda_k - \widehat{\lambda}_k\right|$ and the standard deviation of $\lambda_k - \widehat{\lambda}_k$ are derived based upon a 100×100 grid on X_k. Here, X_k is defined as the square spanned by the

lowest and highest observed values in each component. While this definition of X_k has some computational benefits, function estimates at the boundaries of X_k are solely based on the prior, raising the issue of extrapolation. To focus analysis on the data supported areas of X_k, only grid points contained in the convex hull of the observed values of $\mathbf{x}_k$ are considered for comparison.

EGO-CV is applied with $\beta = 2$, $\alpha = \min(\texttt{cv\_MSE})/1000$ and an initial $\omega_u = 50$ which is increased by factor 10 until the condition in line 3 of Algorithm 4.1 is satisfied. For each proposal w_u and w^*, 5 repetitions of a 10-fold cross validation are performed, where a fold consists of 50,000 iterations and every 100th sample being stored after a burn-in period of 25,000. In addition to the expected improvement criterion, *EGO-CV* stops if 30 values for ω have been considered. Since smoothing is more sensitive on lower than upper values of ω, *EGO* is performed on a transformed scale with $\widetilde{\omega} = \sqrt{\omega/50}$ which provided increased robustness. Alternatively, *EGO* may be applied on a transformed log scale, for example. *Births*, *Deaths* and *Shifts* are proposed with probabilities 0.3, 0.3 and 0.4, respectively. The conclusive MCMC algorithm runs with ω_{opt} for 2,500,000 iterations after a burn-in period of 500,000 and every 1000th sample is stored for analysis. Convergence of the Markov chains for region k is checked by first sampling some points over X_k and then investigating the associated trace plots of the functional levels; see Appendix C.4 for examples.

Section 4.3.2 considers $K = 2$ contiguous regions with Gaussian response data and performs sensitivity analysis on the prior parameters η, p and q. A more complex spatial networks with Binomial response data and regionally varying sample spaces is considered in Section 4.3.3.

4.3.2 Gaussian Data

Observations for region $k = 1, 2$ are simulated independently from a Normal distribution

$$y_k \sim \text{Normal}\left(\lambda_k(\mathbf{x}_k), \sigma_k^2\right), \tag{4.3.1}$$

where $\mathbf{x}_k \in [0,1]^2$ with a varying distribution across the simulations; details are provided in the respective sections. Instead of inferring on λ_k directly, the decomposition described in Section 4.2.4 is applied. Hence, we define $\lambda_k(\mathbf{x}) := \mu_k + \widetilde{\lambda}_k(\mathbf{x})$ and the considered probability model (4.2.1) is then given as

$$f\left(y_k \mid \widetilde{\lambda}_k(\mathbf{x}_k), \mu_k, \sigma_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{1}{2\sigma_k^2}\left(y_k - \mu_k - \widetilde{\lambda}_k(\mathbf{x}_k)\right)^2\right]. \tag{4.3.2}$$

Table 4.3.1: Absolute difference $\times 10^{-2}$ (standard deviation of the difference $\times 10^{-2}$) between true function and posterior mean estimate for the five pairs of functions (λ_1, λ_2) in Studies 1 to 5 for three settings of η. The last column refers to $\omega = 0$ and η being updated within the MCMC scheme. The setting providing the lowest combined absolute difference is printed in bold.

Study	Function	$\eta = 10$	$\eta = 1000$	$\eta = \widehat{\eta}$	$\omega = 0$
1	λ_1	1.8 (2.8)	1.7 (2.6)	**1.7 (2.5)**	1.8 (2.7)
	λ_2	3.1 (4.5)	2.8 (4.1)	**2.7 (3.9)**	4.5 (7.3)
2	λ_1	**1.6 (3.5)**	1.6 (3.6)	1.6 (3.5)	1.6 (3.8)
	λ_2	**2.9 (4.1)**	3.0 (4.4)	3.0 (4.5)	4.7 (7.2)
3	λ_1	1.3 (1.8)	**1.1 (1.6)**	1.2 (1.6)	1.1 (1.5)
	λ_2	1.9 (2.4)	**1.7 (2.2)**	1.8 (2.3)	2.4 (3.5)
4	λ_1	3.0 (9.3)	**2.8 (8.3)**	2.8 (8.1)	2.8 (8.3)
	λ_2	4.1 (9.6)	**4.0 (8.8)**	4.1 (9.0)	5.9 (12.4)
5	λ_1	1.4 (1.9)	**1.3 (1.7)**	1.3 (1.8)	1.3 (1.8)
	λ_2	2.3 (3.6)	**2.3 (3.4)**	2.3 (3.4)	2.4 (3.6)

The prior in Section 4.2.3 is then set for $\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2\right)$ with $d_{1,2} = 1$. An ICAR prior is defined for (μ_1, μ_2), imposing a spatial structure, and these are updated separately via a Random-Walk Metropolis step. The hyperparameter in the ICAR prior is updated via Gibbs sampling as proposed by Knorr-Held (2003). Inverse-Gamma priors are set for σ_1^2 and σ_2^2 and these parameters are updated via Gibbs sampling. The first set of simulations performs sensitivity analysis for η while the second set does so for p and q.

Sensitivity Analysis for Prior Parameter η

The five considered pairs of λ_1 and λ_2, ranging from smooth curves through to discontinuous surfaces with several threshold effects, are illustrated in Column 1 and 2, respectively, of Figure 4.3.1. The functional levels $\lambda_k(\mathbf{x}_k) \in [0, 2]\,,\ k = 1, 2$, across all studies. For each pair, 1,000 and 100 data points are sampled for regions 1 and 2, respectively, with $\sigma_k^2 = 0.05^2$ and $\mathbf{x}_k \sim \text{Unif}\left([0, 1]^2\right)\,,\ k = 1, 2$. This setting explores, in particular, the potential benefits of borrowing statistical information from region 1 when estimating λ_2. To explore sensitivity with respect to η, three parameter settings are considered: (i) $\eta = 10$, (ii) $\eta = 1000$ and (iii) $\eta = \widehat{\eta}$, In (iii), $\widehat{\eta}$ is the posterior mean estimate based on 150,000 iterations after a burn-in period of 50,000 for the case $\omega = 0$ (Section 4.2.3). The other parameters are fixed to $p = q = 1,\ \delta_{\min} = -1.0$ and $\delta_{\max} = 4.0$ across all studies.

Study 1 and 2 consider cases with $\lambda_1 = \lambda_2$ and Table 4.3.1 shows that both error measures

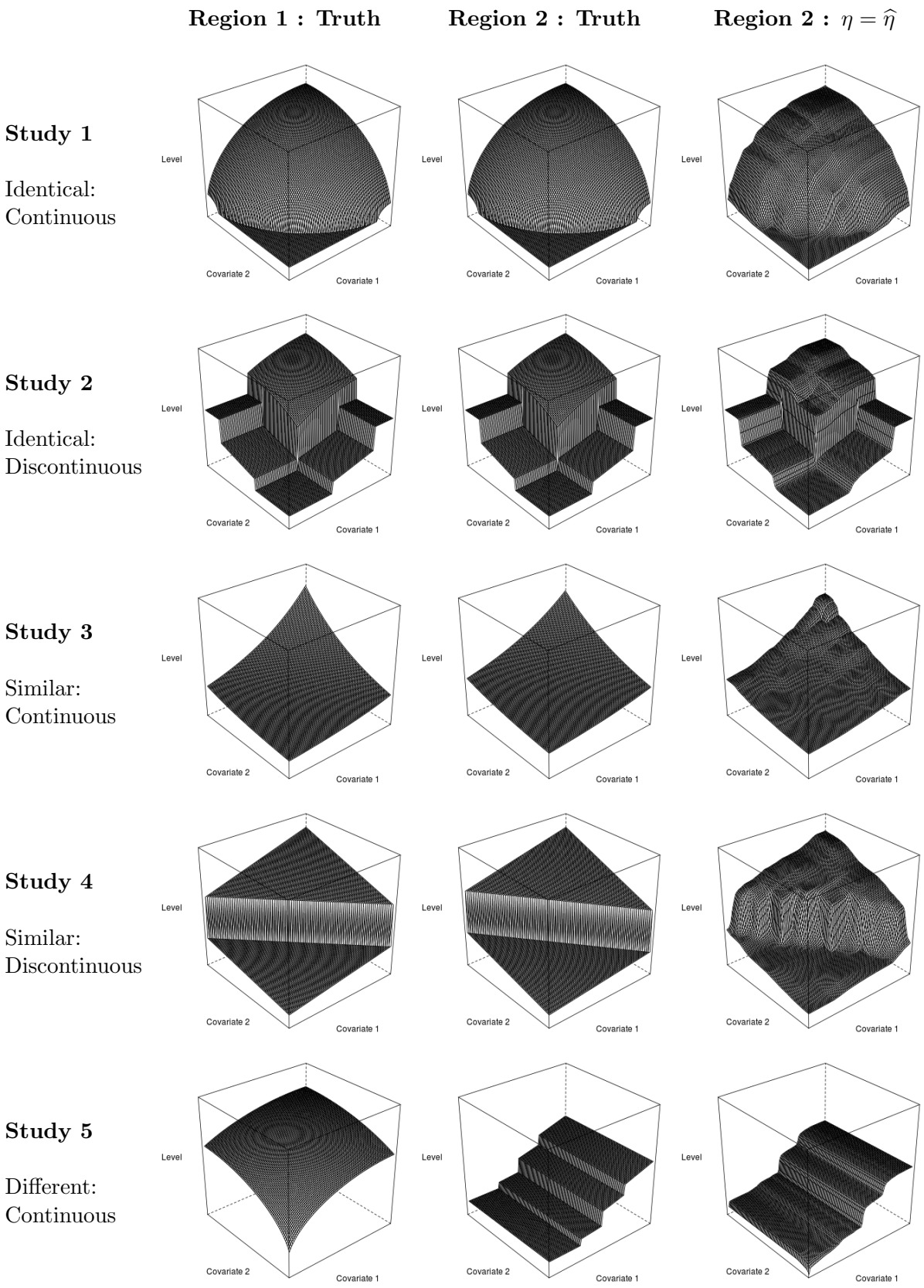


Figure 4.3.1: True functions λ_1 (Column 1) and λ_2 (Column 2), and the posterior mean estimate $\widehat{\lambda}_2$ obtained by BSMMR and *EGO-CV* with $\eta = \widehat{\eta}$ (Column 3) for the five function pairs in Section 4.3.2.

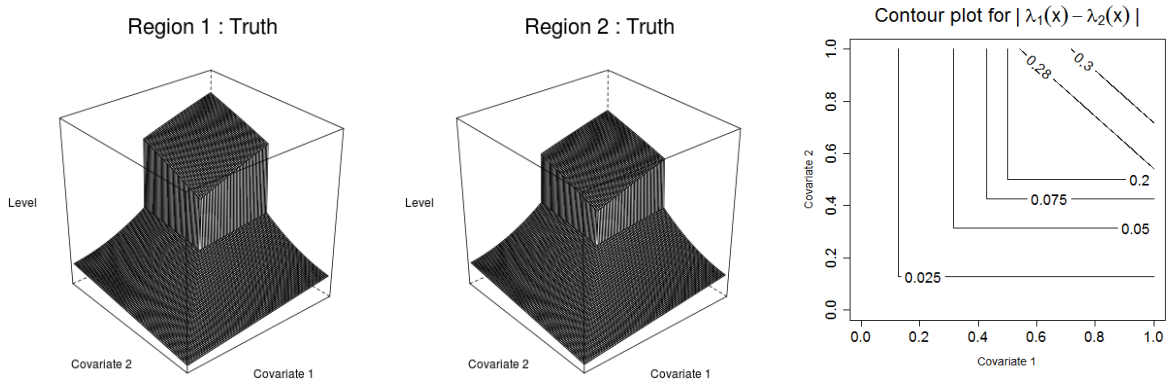


Figure 4.3.2: True functions λ_1 and λ_2 in Section 4.3.2 and contour plot of the absolute difference in their functional levels.

are reduced, in particular, for region 2. The posterior mean plots for λ_2 and $\eta = \widehat{\eta}$ in Figure 4.3.1 illustrate that both smooth surfaces and discontinuities are captured well. In Study 3 and 4, λ_1 and λ_2 are similar and the conclusions are consistent with those for Study 1 and 2.

Finally, Study 5 applies BSMMR to a pair of substantially different functions and Table 4.3.1 shows no worsening for both regions. The capacity for variable selection, described in Section 4.2.3, has been tested for λ_2 in Study 5, a function for which $\lambda_2(\mathbf{x})$ only depends on $x_{2,1}$. Almost all sampled points of $\boldsymbol{\Delta}_2$ are contained in $\Delta_{2,1}$ (results not shown) and the samples, hence, indicate that $x_{2,2}$ is redundant.

Table 4.3.1 further indicates a small sensitivity with respect to η. In particular, $\eta = 10$ leads to slightly worse results than $\eta = 1000$ or $\eta = \widehat{\eta}$ if λ_1 and λ_2 are smooth. As higher η allow, on average, for a higher number of process points, the smooth surfaces are fitted better due to the samples having more, but smaller, jumps. All posterior mean plots are provided in Appendix C.4. As the setting $\eta = \widehat{\eta}$ performs generally well, it is applied in the following simulations.

Sensitivity Analysis for Prior Parameters p and q

When considering p and q, the setup differs from the previous set of simulations in that the distribution of $\mathbf{x}_k$ varies across the studies, whereas λ_1 and λ_2 are fixed. This setting explores the performance of our approach subject to relatively more or less data points being observed in areas with similar functional levels. Figure 4.3.2 illustrates the true function pair (λ_1, λ_2), which is fixed across studies, and a contour plot of the difference in their functional levels. Both functions exhibit a discontinuity at $(0.5, 0.5)$ and the lower functional levels, referring to $\mathbf{x}_k \in [0,1]^2 \setminus [0.5, 1.0]^2$, are more similar than the upper levels, $\mathbf{x}_k \in [0.5, 1.0]^2$, $k = 1, 2$.

Table 4.3.2: Absolute difference $\times 10^{-2}$ (standard deviation of the difference $\times 10^{-2}$) between true function and posterior mean estimate for the three simulation setups in Section 4.3.2 and different values for the prior parameters p and q. The setting providing the lowest combined absolute difference is printed in bold.

Study	Function	$p = 1.0$ $q = 1.0$	$p = 1.0$ $q = 2.0$	$p = 0.1$ $q = 1.0$	$p = -1.0$ $q = 1.0$	$p = 2.0$ $q = 1.0$	$\omega = 0$
1	λ_1	3.7 (6.7)	3.4 (5.4)	3.4 (6.2)	**3.1 (5.7)**	3.6 (6.3)	3.6 (6.4)
	λ_2	3.3 (6.3)	3.2 (5.8)	3.0 (5.5)	**2.9 (5.3)**	3.4 (6.5)	3.7 (7.8)
2	λ_1	4.5 (7.8)	4.2 (6.8)	4.2 (7.1)	**4.2 (6.9)**	4.5 (8.0)	5.6 (12.5)
	λ_2	4.3 (6.7)	4.2 (5.8)	4.0 (6.0)	**4.0 (6.0)**	4.4 (6.5)	4.3 (6.2)
3	λ_1	3.5 (7.6)	3.5 (8.1)	3.6 (7.4)	**3.3 (6.9)**	3.6 (6.3)	3.5 (7.6)
	λ_2	4.0 (7.3)	3.3 (6.0)	3.5 (6.1)	**3.2 (5.3)**	3.5 (5.1)	4.3 (8.4)

For each study, 300 data points are simulated for each region with $\sigma_1^2 = \sigma_2^2 = 0.1^2$. The three studies considered in the following vary with respect to the number of observations sampled on $[0.5, 1.0]^2$, the area for which the levels of λ_1 and λ_2 are quite different. Study 1 considers the case $\mathbf{x}_k \sim \text{Unif}\left([0,1]^2\right),\ k = 1, 2$. In Study 2, 200 data points are sampled uniformly from $[0.5, 1.0]^2$ for each region while only 40 observations are sampled from this area in Study 3. The remaining data points, 100 and 260, respectively, are sampled uniformly from $[0,1]^2 \setminus [0.5, 1.0]^2$.

Five settings for p and q are compared in order to explore sensitivity on these parameters. The first two settings: (1) $p = 1,\ q = 1$ and (2) $p = 1,\ q = 2$ impose a constant degree of dependence between λ_1 and λ_2. Settings (3) $p = 0.1,\ q = 1$ and (4) $p = -1,\ q = 1$ allow for stronger dependence in the lower functional levels while the last setting (5) $p = 2,\ q = 1$ imposes increased dependence for higher levels. The functional level boundaries are set to $\delta_{\min} = 0.0$ and $\delta_{\max} = 3.0$.

Table 4.3.2 shows that all settings for p and q improve upon $\omega = 0$, with $p = -1,\ q = 1$ performing generally best. Consequently, the imposition of a dependence structure is beneficial despite the upper functional levels being rather different. The setting $p = -1,\ q = 1$ performs best as it effectively borrows statistical information for the lower functional levels without inducing a large bias on the upper functional levels. Table 4.3.2 further indicates a small sensitivity with respect to q, in particular, setting (2) outperforms setting (1) across the three studies. An increase in q implies that less dependence is imposed on the smaller functional differences, relative to the highest functional level difference. Hence, a higher value for q allows here for a better mixing of the functional levels at the discontinuity. Since the true function is strictly increasing, this leads to the setting $p = 1,\ q = 2$ performing better than $p = 1,\ q = 1$. All

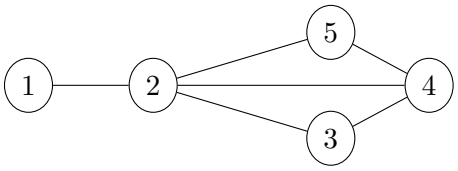


Figure 4.3.3: Spatial network of the 5 regions considered in Section 4.3.3.

posterior mean plots are provided in Appendix C.5.

4.3.3 Binomial Data

We consider the spatial network in Figure 4.3.3 with the $K = 5$ regions having between 1 and 4 neighbours. Responses y_k are sampled independently from a Binomial distribution of the form

$$y_k \sim \text{Binomial}\left(A_k,\ \lambda_k\left(\mathbf{x}_k\right)\right), \tag{4.3.3}$$

where the number of trials $A_k = 100$ is fixed $\forall k$ and $\lambda_k\left(\mathbf{x}_k\right) \in [0,1]$ refers to the success probability with $\mathbf{x}_k \in [0,1]^2$. Here, the sample space for the first explanatory variable $x_{k,1}$ is regionally varying. Specifically, $\mathbf{x}_k \in [0.0, 0.7] \times [0,1]$ for region $k = 1, 3, 5$, and observed on $[0.2, 1.0] \times [0,1]$ and $[0.1, 0.9] \times [0,1]$ for regions 2 and 4, respectively. Column 1 in Figure 4.3.4 illustrates the true functions $\lambda_1, \ldots, \lambda_5$ over the defined spaces. The number of observations generated for regions 1 through 5 are 100, 500, 200, 300 and 200, respectively, with $\mathbf{x}_k,\ k = 1, \ldots, 5$, being uniformly distributed on the defined sample spaces.

In addition to inferring on $\lambda_k,\ k = 1, \ldots, 5$, over its associated observation space, we borrow statistical information spatially for extrapolation. Let $\widetilde{X}_k$ denote the square spanned by the lowest and highest observed values for $x_{k,j},\ j = 1, 2$, and set $d_{k,k'} = 1$ in (4.2.9) if regions k and k' are adjacent, and $d_{k,k'} = 0$ otherwise. We then extrapolate λ_k by defining the domain $W_{k,k'}$ in (4.2.4) as the union of $\widetilde{X}_k$ and $\widetilde{X}_{k'}$, that is, $W_{k,k'} := \widetilde{X}_k \cup \widetilde{X}_{k'}$. Consequently, there exists statistical information to estimate λ_k on the set $X_k := \bigcup_{\{k' : d_{k,k'} = 1\}} \left\{\widetilde{X}_k \cup \widetilde{X}_{k'}\right\}$. In the simulation setting above, $X_k,\ k = 1, \ldots, 5$, then corresponds approximately to the unit square. Note, the extrapolation assumes similar functional forms for adjacent regions. If this assumption appears too strong, inference can be restricted to $\widetilde{X}_k$ by defining $W_{k,k'}$ as the intersection of $\widetilde{X}_k$ and $\widetilde{X}_{k'}$: $W_{k,k'} := \widetilde{X}_k \cap \widetilde{X}_{k'}$. Both the union and intersection settings for $W_{k,k'}$ are applied to the simulated data with prior parameters $p = 1,\ q = 1$ and $\eta = \widehat{\eta}$ (Section 4.3.2).

Table 4.3.3 indicates that both settings for $W_{k,k'}$ perform similarly in terms of the considered

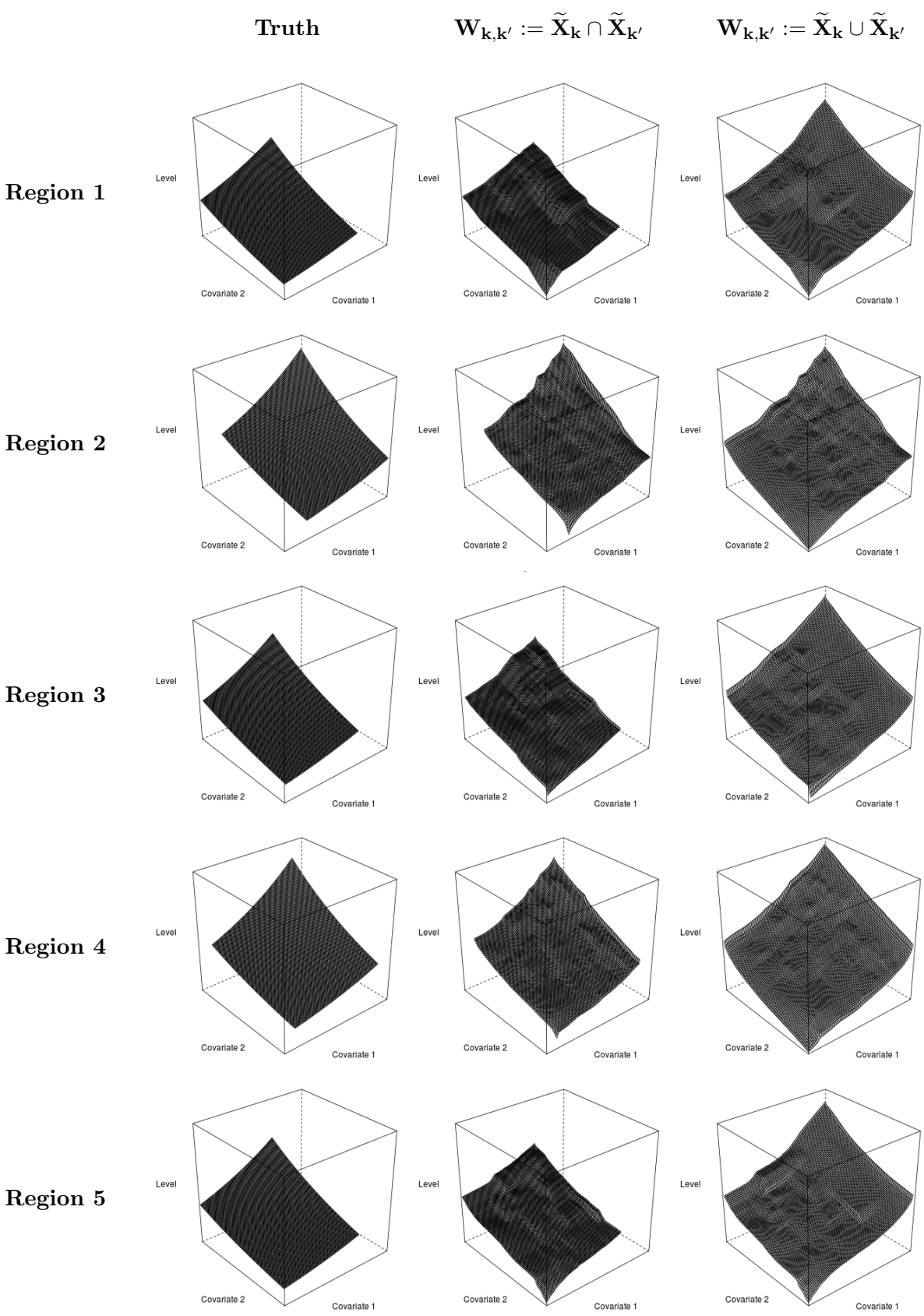


Figure 4.3.4: True functions λ_1 through λ_5 (Column 1) and estimated posterior means for the settings $W_{k,k'} := \widetilde{X}_k \cap \widetilde{X}_{k'}$ (Column 2) and $W_{k,k'} := \widetilde{X}_k \cup \widetilde{X}_{k'}$ (Column 3) in Section 4.3.3.

Table 4.3.3: Average absolute $\times 10^{-2}$ (standard deviation $\times 10^{-2}$) of the difference between true function and posterior mean estimate for the five functions λ_1 through λ_5 in Section 4.3.3 with two different settings for $W_{k,k'}$ and the case $\omega = 0$.

Function	$W_{k,k'} := \widetilde{X}_k \cap \widetilde{X}_{k'}$	$W_{k,k'} := \widetilde{X}_k \cup \widetilde{X}_{k'}$	$\omega = 0$
λ_1	1.4 (1.4)	1.3 (1.4)	1.9 (2.2)
λ_2	0.8 (0.9)	0.8 (0.9)	0.8 (1.1)
λ_3	1.0 (1.2)	1.1 (1.3)	1.5 (1.9)
λ_4	0.9 (1.0)	0.9 (1.2)	1.3 (1.8)
λ_5	0.9 (1.2)	0.9 (1.3)	1.6 (2.2)

error measurements and also improve upon $\omega = 0$. Specifically, the results suggest that, for instance, region 1 borrows statistical information from region 2. Figure 4.3.4 then shows that our approach recovers the smooth functions λ_1 through λ_5 well. Further, the extrapolated functions in Column 3 of Figure 4.3.4 show that our approach allows for effective borrowing of statistical information. For instance, we borrow from region 2 to extrapolate the functions λ_1, λ_3, λ_4 and λ_5 which mimic the functional shape of $\widehat{\lambda}_2$ in the extrapolated areas.

4.4 Case Study

We consider the insurance and weather data used by Haug et al. (2011) and Scheel et al. (2013). The data provide the daily number of insurance claims caused by precipitation, surface water, snow melt, undermined drainage, sewage back-flow or blocked pipes for all 430 Norwegian municipality from 1997 to 2006. Further, the average number of policies held per month and multiple daily weather metrics are recorded municipality-wise. Here, we explore the effect of the amount of precipitation on the current, t, and previous, $t-1$, day on the daily number of claims. Scheel et al. (2013) found these to be the most informative explanatory variables. Intuitively, the assumption that the average claim risk per property increases with the amount of precipitation appears reasonable and, hence, motivates the application of our BSMMR methodology. Analysis is performed for a contiguous set of $K = 11$ municipalities around the Oslofjord (Figure 4.4.1). The notation and modelling framework is formalized in the following.

Let $N_{k,t}$ and $A_{k,t}$ denote the number of claims and policies, respectively, on day t for municipality k. Further, $R_{k,t}$ and $R_{k,t-1}$ refer to the amount of precipitation on day t and $t-1$, respectively, for municipality k. We then model $N_{k,t}$ via a Binomial distribution with the claim probability on day t for municipality k, $p_{k,t}$, on the logit scale being given as $\lambda_k(R_{k,t}, R_{k,t-1})$. Since exploratory data analysis indicates that differences exist in the average claim rate per pol-

Figure 4.4.1: Map of the 11 municipalities considered in Section 4.4.

icy holder across the 11 municipalities, we apply the decomposition described in Section 4.2.4 and define $\lambda_k\left(R_{k,t}, R_{k,t-1}\right) := \mu_k + \widetilde{\lambda}_k\left(R_{k,t}, R_{k,t-1}\right)$. Formally, the claim model is then given by

$$\begin{aligned} N_{k,t} &\sim \text{Binomial}\left(A_{k,t}, p_{k,t}\right) \\ \text{logit}\left(p_{k,t}\right) &= \mu_k + \widetilde{\lambda}_k\left(R_{k,t}, R_{k,t-1}\right). \end{aligned} \tag{4.4.1}$$

As presented in Section 4.3.2, an ICAR prior is defined for the intercepts $\mu_1, \ldots, \mu_{11}$.

To assess the predictive performance, observations for 2001 and 2003 are stored as test data and the monotonic functions and intercepts are estimated based on the remaining 8 years. In addition to our BSMMR approach, we consider two competing models:

1. The number of claims on each day is simply predicted as the average over the training data set: $\overline{N_{k,t}}$.

2. A spatially varying coefficient (SVC) model (Assunção, 2003) which defines $p_{k,t}$ on the logit scale as a linear combination of $R_{k,t}$ and $R_{k,t-1}$. Separate ICAR priors are defined for the intercepts and the two covariate effects.

BSMMR is applied with prior parameters $p = -1$, $q = 1$, $\eta = \widehat{\eta}$ and $d_{k,k'} = 1$ if municipalities k and k' share a border and 0, otherwise. The selection $p = -1$ is motivated by the high occurrence of days with little or no precipitation. More specifically, relatively more data points are available to model the lower functional levels (lower rainfall) compared to the number of days with high amount of precipitation. The functional level boundaries are set to $\delta_{\min} = 0$ and

Table 4.4.1: Sum of squared prediction errors of the daily number of claims for the years 2001 and 2003 based on the model fitted with explanatory variables $R_{k,t}$ and $R_{k,t-1}$ for the remaining 8 years between 1997 and 2006. For each municipality, the model performing best is in bold type face.

Municipality	$\omega = \omega_{opt}$	$\omega = 0$	$\overline{N}_{k,t}$	SVC
Ås	14.0	14.0	**13.9**	14.3
Asker	360.8	361.4	372.5	**331.0**
Bærum	**296.0**	318.2	915.1	679.3
Frogn	**8.2**	**8.2**	8.5	12.3
Hurum	17.4	17.5	17.7	**17.1**
Nesodden	20.6	20.9	20.5	**20.2**
Oppegård	31.6	33.9	**26.2**	27.6
Oslo	445.8	444.0	**412.2**	452.3
Røyken	57.6	57.5	63.5	**53.3**
Ski	39.0	39.0	**38.2**	38.8
Vestby	18.6	18.6	**18.5**	18.9
Σ	**1309.6**	1333.2	1906.8	1665.1

$\delta_{\max} = 10$, and $W_{k,k'}$ in (4.2.4) is defined as $W_{k,k'} := \widetilde{X}_k \cup \widetilde{X}_{k'}$, where $\widetilde{X}_k$ refers to the square spanned by the observed minima and maxima of $R_{k,t}$ and $R_{k,t-1}$ for municipality k. Due to positive skew, the transformed variables $\sqrt{R_{k,t}}$ and $\sqrt{R_{k,t-1}}$ are considered. Alternatively, one may transform the data using the empirical distribution function, resulting in a modelling of $N_{k,t}$ in dependence on the precipitation-quantiles. After deriving the posterior mean estimate $\widehat{\eta}$ and the prior parameter ω_{opt}, $\lambda_1, \ldots, \lambda_{11}$ are estimated by performing 1,000,000 iteration steps of the MCMC algorithm and storing every 500th sample after a burn-in period of 200,000. The SVC model is fitted with the two covariate effects by performing 10,000 iteration steps with a burn-in of 1,000.

Table 4.4.1 shows that BSMMR performs the best in terms of the overall predictive error Σ; reducing it to $\Sigma = 1309.6$, compared to $\Sigma = 1333.2$ for the setting $\omega = 0$ and $\Sigma = 1665.1$ for the fitted SVC model. Slight improvements are achieved by accounting for spatial structure in the regression functions. The small scale of improvement from $\omega = 0$ to $\omega = \omega_{opt}$ can be explained by the high number of training data points (≈ 3000) for each municipality. Hence, important structures in the regression surface are likely to be captured without borrowing statistical information from adjacent municipalities. Posterior mean plots for the municipalities of Oslo and Hurum are provided in Appendix C.6. For confidentiality reasons, no information is given on the estimated functional levels.

The largest improvement is achieved for Bærum, the municipality which observes the highest

daily count over the test period. Hence, the results indicate that BSMMR is better at predicting higher claim number than the competing models. A possible explanation may be the existence of a threshold effect which induces an elevated claim risk if precipitation levels exceed a certain level. The high frequency of days with low precipitation levels then effects the linear model fit for higher precipitation levels stronger than it does for our more flexible approach.

The results for the other 10 municipalities are similar for the different models which relates to there being zero high-claim days observed over the test period. This occurrence of zero high-claim days can be further seen in the results; the predictive squared error obtained for model 1 is low for most municipalities. BSMMR performs slightly worse than the SVC model (model 2) for Asker due to one day with a high claim number $N_{k,t}$ that is not captured well as days with similar precipitation in the training data observe no daily count $N_{k,t}$ of this magnitude.

4.5 Discussion

We developed new non-parametric Bayesian methodology for modelling and estimation of a spatially varying regression function under the assumptions of monotonicity and boundedness. To impose a spatial structure on the monotonic functions, we constructed a flexible pair-wise discrepancy measure which allows the degree of dependence to vary with the functional levels via a tuning parameter p. We further postulated the functions to be step functions and which are represented via marked point processes. The conclusive joint prior was then defined on the marked point processes and incorporated the pair-wise defined dependence model. A RJMCMC scheme was formulated to sample from the posterior distribution. As the normalizing constant of the prior was intractable, we developed the *EGO-CV* algorithm, combining the concepts of cross-validation and Bayesian global optimization, to derive a robust value for the smoothing parameter ω.

We applied our methodology to several simulated data sets to illustrate its benefits. The results show that BSMMR can recover both smooth and discontinuous surfaces and that statistical information is shared effectively. In particular, our methodology substantially improved upon existing approaches if the functional shapes were similar. These conclusions were irrespective of the functional shapes and the distribution of the explanatory variables. We also considered a pair of monotonic functions with very different functional forms in order to illustrate the robustness of our *EGO-CV* approach. In this chapter, we further demonstrated that the BSMMR

methodology allows for extrapolation and variable selection. Finally, we applied our methodology to explore the association between precipitation levels and water-related insurance claims. Again, our approach performed better than its competitors.

From a general perspective, BSMMR provides a useful modelling approach which allows for both smooth and discontinuous functional forms and which may not be captured if a linear or additive form is assumed. Also, the approach can be applied generally for network and dependence models and is not limited to a spatial context. Our simulations further provide some practical guidance for selecting the parameters p, q and η. A reliable η can be derived from an initial MCMC algorithm with $\omega = 0$; a setting for which posterior realizations of η can be sampled via an additional Gibbs sampling step. To select p and q, the function estimates obtained from the MCMC algorithm used to tune η may indicate functional differences and, hence, allow, for instance, to infer on the choice of $p < 1$, $p = 1$ or $p > 1$.

This chapter considered spatial variation in the regression function, per se, but the methodology can be extended to a spatio-temporal context. Assume that the effect of the explanatory variables is temporally stationary but that the baseline level changes between observations. The monotonic regression function $\lambda_{k,t}$ at time t for region k could then be defined as

$$\lambda_{k,t}(\mathbf{x}) = \alpha_{k,t} + \widetilde{\lambda}_k(\mathbf{x}), \tag{4.5.1}$$

where $\widetilde{\lambda}_k$ is estimated as proposed in this chapter. To impose temporal structure, for instance, an ICAR prior could be set on $(\alpha_{k,t-1}, \alpha_{k,t})$. The modelling framework is also expandable to a spatio-temporal setting for which the regional regression functions change at specified time points. Temporal structure is then imposed analogously to the spatial structure using time-adjacency.

An aspect not discussed here is the selection of the number of subprocesses. Since we only considered $m = 2$ explanatory variables, the number of subprocesses was $I = 3$. In higher dimensions, however, one may want to restrict I. Assume there exists prior knowledge on a continuous explanatory variable $x_{k,j}$, $j = 1, \ldots, m$, being informative and let $X_k = [0,1]^m$. Then the marked point processes which the jth entry of $\boldsymbol{\xi}_{k,i,j}$ being 0 might be ignored.

Our methodology is well-suited for lower-dimensional regression problems, and we find BSMMR to perform best for models with two to five explanatory variables. However, as for many other flexible approaches, e.g. generalize additive models, some issues arise for higher

dimensions. Firstly, the computational cost for calculating the prior density scales exponentially with the number of explanatory variables. Secondly, the monotonic constraint becomes less restrictive with increasing dimensions, leading to a potential overfit of the data. As, for instance, discussed by Bergersen et al. (2014), larger sets of explanatory variables can be accommodated by imposing an additive or semi-parametric structure on the regression function. The lower-dimensional monotonic functions could then be estimated in-turn. In conclusion, our methodology can be applied for higher dimensional regression problems, but we would recommend a pre-analysis in order to reduce the computational time substantially.

Computationally, BSMMR is quite demanding, depending on the neighbourhood structure and the number of both explanatory variables and data points. Since proposals change the regression surface only locally, the computational time for computing the prior ratio is reduced by firstly deriving the area affected by the proposal and then evaluating $D_{p,q}$ over it. However, further splits into smaller areas are usually required as the functions associated to the adjacent regions are likely to be non-constant over this area. To reduce the computational time, in particular for the *EGO-CV* algorithm, parallelization techniques could be used which allocate the folds to multiple processors. The `C++` implementation and `R` files for Section 4.3 are available from `www.lancaster.ac.uk/pg/rohrbeck/BSMMR`.

Our work can be extended in several ways. From a theoretical perspective, interest may lie in the construction of a different discrepancy measure, for instance, one based on the Kullback-Leibler divergence. Further, one may want to derive p and q directly from the functional estimates for $\omega = 0$, rather than selecting these manually. The estimation would then be a three-step process which first estimates the parameters p, q and η based on the setting $\omega = 0$, followed by the*EGO-CV* algorithm to infer on ω and, finally, the conclusive RJMCMC algorithm. An aspect we have not considered yet is the selection of the number of folds s which we arbitrarily fixed to $s = 10$. As the value for ω also depends on the number of data points, a larger number of folds may return a more robust estimate. The additional computational time may be handled by parallelized computing techniques.

Chapter 5

Modelling Functional Dependence in an Isotonic Regression Framework

5.1 Introduction

The association of a response Y and a set of predictors $X_1, \ldots, X_m$ is of interest in several applications. Often it can be assumed that the mean of Y is non-decreasing with increases in $X_i,\ i = 1, \ldots, m$. This is typically the case, for instance, for dose-response relationships. Analysis under this constraint is termed isotonic, or monotonic, regression and is considered in several statistical areas, including optimization (Ayer et al., 1955; Maxwell and Muckstadt, 1985; Luss and Rosset, 2014), generalized additive modelling (Bacchetti, 1989; Leitenstorfer and Tutz, 2007; Bergersen et al., 2014) and Bayesian non-parametrics (Gelfand and Kuo, 1991; Dunson, 2005; Saarela and Arjas, 2011). In this chapter, the optimization-based approaches are considered in particular.

Consider a set of observed responses and predictors $\{(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^m \ : \ i = 1, \ldots, n\}$ and let $\preceq$ define a partial order on $\mathbb{R}^m$, e.g. the Euclidian order. Barlow and Brunk (1972) consider Gaussian distributed responses and derive estimates $\widehat{y}_1, \ldots, \widehat{y}_n \in \mathbb{R}$ at $\mathbf{x}_1, \ldots, \mathbf{x}_n$ via the solution of an optimization problem of the form

$$\min_{\widehat{y}_1, \ldots, \widehat{y}_n} \sum_{i=1}^{n} w_i \left| y_i - \widehat{y}_i \right|^2, \tag{5.1.1}$$

subject to the constraint $\widehat{y}_i \leq \widehat{y}_j$ when $\mathbf{x}_i \preceq \mathbf{x}_j,\ \forall i, j = 1, \ldots, n$ and where $w_i \geq 0$ are fixed constants which impose a weighting of the observations. Note, the optimization problem (5.1.1)

is not limited to Gaussian data and can, for instance, also be applied to Binomial or Poisson data. Several algorithms to solve this optimization problem exist, for instance, the pool adjacent violators algorithm (PAVA) (Ayer et al., 1955), for the case $m = 1$, or isotonic recursive partitioning (Luss et al., 2012). To predict the response at a point $\mathbf{z} \in \mathbb{R}^m$, a monotonic function $\widehat{\lambda} : \mathbb{R}^m \to \mathbb{R}$ can be derived based on the solution $\widehat{y}_1, \ldots, \widehat{y}_n \in \mathbb{R}$ via interpolation. For instance, by defining

$$\widehat{\lambda}(\mathbf{z}) = \max_{i=1,\ldots,n} \{\widehat{y}_i : \mathbf{x}_i \preceq \mathbf{z}\}, \tag{5.1.2}$$

for which the resulting monotonic function $\widehat{\lambda}$ is piecewise constant with $\widehat{\lambda}(\mathbf{x}_i) = \widehat{y}_i,\ i = 1, \ldots, n$.

An extended modelling framework is developed here, where data are observed for a fixed number of subgroups rather than for a single group. The association between response and predictor is assumed to be monotonic but potentially varying across subgroups. Such scenarios have generic relevance, for instance, in medical applications; patients may exhibit individually varying dose-response curves and hence be classified into different subgroups based on an objective criterion. Another application area concerns spatial (Cressie, 1993) and spatio-temporal (Cressie and Wikle, 2015) lattice data, where the association between response and predictors may vary either temporally or spatially. Statistical methods considering spatio-temporal data generally aim to improve estimates by borrowing statistical information from subgroups (areal units or geographical subregions) which exhibit a similar functional structure. In this spirit, certain similarities between subgroups may exist and interest lies in exploiting these in order to estimate the underlying monotonic relationships between response and predictors for the different subgroups.

While some methodology exists to model a spatial variation in the dependence between response and predictors (Fotheringham et al., 2002; Assunção, 2003; Scheel et al., 2013), little research has been done in the context of isotonic regression. Chapter 4 introduced a Bayesian approach, termed Bayesian spatial monotonic multiple regression (BSMMR), which defines a dependence structure on the functional levels via a prior distribution. However, BSMMR is computationally expensive and requires multiple iterations of a reversible jump Markov Chain Monte Carlo algorithm. This chapter introduces a computationally more efficient approach which is motivated by the optimization problem in (5.1.1). This increased efficiency comes at the cost of reduced flexibility, in particular, functions are estimated at the observed data points only and then derived via interpolation, for instance, as in expression (5.1.2).

The remainder of this chapter is organized as follows: Section 5.2 details the modelling framework which consists of a general optimization problem, the optimization algorithm and an approach to estimate the pair-wise similarity between subgroups at each observed data point. Performance of the method is then assessed for simulated Gaussian data in Section 5.3. The chapter concludes with a summary and discussion in Section 5.4.

5.2 Methodology

5.2.1 The Optimization Problem

Consider a fixed number of K subgroups and let $Y_k \in \mathbb{R}$ and $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,m}) \in \mathbb{R}^m$ denote the response and predictor for subgroup $k = 1, \ldots, K$, respectively. Further, let the association between Y_k and $\mathbf{X}_k,\ k = 1, \ldots, K$ be defined via an unknown monotonic function $\lambda_k : \mathbb{R}^m \to \mathbb{R}$, for instance, $\mathbb{E}(Y_k) = \lambda_k(\mathbf{X}_k)$. Without loss of generality, assume that λ_k is monotonic increasing with respect to a partial order $\preceq$, that is,

$$\mathbf{u} \preceq \mathbf{v} \ \Rightarrow\ \lambda_k(\mathbf{u}) \leq \lambda_k(\mathbf{v}),\ \mathbf{u}, \mathbf{v} \in \mathbb{R}^m. \tag{5.2.1}$$

Interest lies in estimating the K monotonic functions $\lambda_1, \ldots, \lambda_K$ based on the collection of K data sets, $\mathcal{D}_1, \ldots, \mathcal{D}_K$, where $\mathcal{D}_k = \{(y_{k,i}, \mathbf{x}_{k,i}) \in \mathbb{R} \times \mathbb{R}^m \ :\ i = 1, \ldots, n_k\},\ k = 1, \ldots, K$. Note, the number of data points, n_k, is potentially varying between subgroups.

If $\lambda_1, \ldots, \lambda_K$ are assumed to be independent, each function λ_k can be estimated separately at $\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{n_k}$ by solving the optimization problem in (5.1.1) with respect to the data set $\mathcal{D}_k$. However, $\lambda_1, \ldots, \lambda_K$ may exhibit pair-wise similarities in their functional shapes. Hence, observations in one subgroup, k, may increase the efficiency when estimating another monotonic function $\lambda_{k'},\ k' \neq k$. The idea is to extend the optimization problem (5.1.1) such that statistical information may be shared between subgroups. For notational simplicity, the case $K = 2$ is considered first and later generalized.

Assume that λ_1 and λ_2 are similar with respect to their functional levels and, hence, $\lambda_1(\mathbf{x}_{k,i})$ and $\lambda_2(\mathbf{x}_{k,i}),\ i = 1, \ldots, n_k,\ k = 1, 2$, are close. In this case, the estimates for λ_2 at $\mathbf{x}_{2,i}\ i = 1, \ldots, n_2$, provide additional statistical information to estimate λ_1. Similarly, the values of λ_1 at the data points in $\mathcal{D}_1$ are useful to estimate λ_2. Consequently, λ_1 and λ_2 can be estimated over the combined set of covariate observations, $\mathcal{E} = \{\mathbf{x}_{1,i} : i = 1, \ldots, n_1\} \cup \{\mathbf{x}_{2,i} : i = 1, \ldots, n_2\}$,

rather than separately over $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. This approach is consistent with the monotonic regression framework in (5.1.1) which estimates the function locally at the observed data points only.

The arguments above motivate the estimation of λ_1 based upon the data $\mathcal{D}_1$ and λ_2, and vice versa for λ_2. Hence, an optimization problem is defined, based upon (5.1.1), which mediates between the observed data and the expected similarity of λ_1 and λ_2. To incorporate belief in the similarity of λ_1 and λ_2, a discrepancy measure is introduced which penalizes differences in the functional levels at each point $\mathbf{x} \in \mathcal{E}$. Similarly to (5.1.1), the squared difference may be considered. Let, further, $\{\widetilde{w}_{k,i}\}$, $i = 1, \ldots, n_k$, $k = 1, 2$ denote non-negative weights which represent the similarity of λ_1 and λ_2 at $\mathbf{x}_{k,i}$; the estimation of such weights is considered later in Section 5.2.2. If the discrepancy measure for the functional levels is defined as the squared difference, the optimization problem results in

$$\begin{aligned} \min_{\widehat{\lambda}_1,\, \widehat{\lambda}_2} \sum_{i=1}^{n_1} & \left\{ w_{1,i} \left[y_{1,i} - \widehat{\lambda}_1 \left(\mathbf{x}_{1,i} \right) \right]^2 \; + \; \widetilde{w}_{1,i} \left[\widehat{\lambda}_1 \left(\mathbf{x}_{1,i} \right) - \widehat{\lambda}_2 \left(\mathbf{x}_{1,i} \right) \right]^2 \right\} \; + \\ & \sum_{i=1}^{n_2} \left\{ w_{2,i} \left[y_{2,i} - \widehat{\lambda}_2 \left(\mathbf{x}_{2,i} \right) \right]^2 \; + \; \widetilde{w}_{2,i} \left[\widehat{\lambda}_1 \left(\mathbf{x}_{2,i} \right) - \widehat{\lambda}_2 \left(\mathbf{x}_{2,i} \right) \right]^2 \right\}, \end{aligned} \tag{5.2.2}$$

where $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are estimated over $\mathcal{E}$ and satisfy the monotonicity constraint in (5.2.1). The constants $\{w_{k,i}\} \geq 0$ are prespecified as in (5.1.1). To ensure uniqueness of the solution $\widehat{\lambda}_1$, one may restrict its estimation to the set of points with positive weight only. Furthermore, the solution is unique under this restriction since the objective function is a sum of strictly convex functions and the set of potential solutions is convex too. In principle, the estimates obtained via (5.2.2) are weighted averages over the monotonic functions obtained by applying the optimization problem in (5.1.1) separately to each data set.

To extend the optimization problem (5.2.2) to K subgroups, the union of the K covariate sets, $\mathcal{E} = \bigcup_{k=1}^{K} \{\mathbf{x}_{k,i} : i = 1, \ldots, n_k\}$, is considered. Furthermore, the approach is extended to a more general class of discrepancy measures than the squared distance, in particular, non-negative, strictly convex, differentiable loss functions. This class of functions contains, for instance, the negative Poisson log-likelihood or the negative-Bernoulli log-likelihood. Let $\Phi_{k,i}$ denote the loss function for $y_{k,i}$ and the fitted functional level $\widehat{\lambda}_k \left(\mathbf{x}_{k,i} \right)$ while $\widetilde{\Phi}_{k,k',i}$ refers to the discrepancy measure between $\widehat{\lambda}_k$ and $\widehat{\lambda}_{k'}$ at $\mathbf{x}_{k,i}$. Again, knowledge on the similarity of the functions is

expressed via a set of weights $\{\widetilde{w}_{k,k',i}\}$. The general optimization problem is then of the form

$$\min_{\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K} \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\{ w_{k,i}\Phi_{k,i}\left[y_{k,i},\widehat{\lambda}_k\left(\mathbf{x}_{k,i}\right)\right] + \sum_{k'\neq k}\widetilde{w}_{k,k',i}\widetilde{\Phi}_{k,k',i}\left[\widehat{\lambda}_k\left(\mathbf{x}_{k,i}\right),\widehat{\lambda}_{k'}\left(\mathbf{x}_{k,i}\right)\right]\right\}, \quad (5.2.3)$$

subject to $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K$ satisfying the monotonicity constraint in (5.2.1). Since the loss functions are strictly convex and differentiable, the objective function in (5.2.3) is convex too. Similarly to the optimization problem (5.2.2), the set of feasible solution for 5.2.3 is also convex. Consequently, the optimization problem is convex and has an unique solution. In applications, a constant loss function $\widetilde{\Phi}_{k,k',i} = \widetilde{\Phi}$ would be considered instead of specifying a different one for each data point. Such a scenario is applied in Section 5.3.

5.2.2 Deriving the Optimal Solution

Since the optimization problem is convex, any local minimum found via a solution algorithm is guaranteed to be the global optimum. Therefore, a sensitivity analysis with respect to the initial setting is unnecessary. To solve the optimization problem (5.2.3), a cyclic algorithm is applied which is motivated by Bacchetti (1989) who considers the estimation of a generalized additive model, Hastie and Tibshirani (1990), under the monotonic constraint in expression (5.2.1). Specifically, the functions $\widehat{\lambda}_1$ to $\widehat{\lambda}_K$ are optimized in-turn while keeping the others fixed. Hence, each optimization step corresponds to the estimation of a single isotonic function and existing isotonic regression solution algorithms can be used. If $\lambda_1,\ldots,\lambda_K$ are univariate, the PAVA can be applied and this is done in Section 5.3. In case the isotonic functions are multivariate, optimization can be performed, for instance, via a minimum lower set algorithm (Brunk, 1955) or generalized isotonic recursive partitioning (Luss and Rosset, 2014). The cyclic optimization procedure stops if none of the current estimates $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K$ changes and this corresponds to the objective function being minimal. Note, if all constants $\widetilde{w}_{k,k',i} = 0$, the algorithm is guaranteed to converge after one iteration as the functions are estimated independently based on the respective data sets.

To conclude the modelling framework, the constants $\widetilde{w}_{k,k',i}$ in (5.2.3) have to be specified. Since there usually exists little knowledge on the functional similarity between functions in applications, an approach to estimate these is proposed in the following. Initially, set $\widetilde{w}_{k,k',i} = 0$ which implies that all functions are independent and they can thus be estimated separately by an existing solution algorithm for monotonic regression problems. Let $\widehat{\lambda}_1^0$ to $\widehat{\lambda}_K^0$ denote the

solutions obtained by such an algorihtm for the subgroups 1 to K, respectively. Consider the point $\mathbf{x}_{k,i}$ at which $\widehat{\lambda}_k^0$ has been estimated and consider a different function $\widehat{\lambda}_{k'}^0$. Based upon an interpolation such as (5.1.2), the functional level $\widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})$ can be derived. If $\widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})$ is close to $\widehat{\lambda}_k^0(\mathbf{x}_{k,i})$, the weight $\widetilde{w}_{k,k',i}$ should be positive since estimates indicate similarity in the functional levels. Conversely, $\widetilde{w}_{k,k',i} = 0$ if the difference is too large.

The approach described above for a single point is applied to all points $\mathbf{x}_{k,i} \in \mathcal{E}$. One possibility is to define $\widetilde{w}_{k,k',i}$ to decay linearly with an increasing distance between $\widehat{\lambda}_k^0(\mathbf{x}_{k,i})$ and $\widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})$. This approach formally results in

$$\widetilde{w}_{k,k',i} = \begin{cases} 1 - \dfrac{\left|\widehat{\lambda}_k^0(\mathbf{x}_{k,i}) - \widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})\right|}{\epsilon} & \text{if } \left|\widehat{\lambda}_k^0(\mathbf{x}_{k,i}) - \widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})\right| \leq \epsilon \\ 0 & \text{if } \left|\widehat{\lambda}_k^0(\mathbf{x}_{k,i}) - \widehat{\lambda}_{k'}^0(\mathbf{x}_{k,i})\right| > \epsilon, \end{cases} \tag{5.2.4}$$

where $\epsilon > 0$ is a fixed constant. Other settings are possible, for instance, quadratic or logarithmic decay. The choice of an upper bound of 1 for $\widetilde{w}_{k,k',i}$ in expression (5.2.4) is not compulsory. However, if $w_{k,i}$ is bounded between 0 and 1, then a value of $\widetilde{w}_{k,k',i} = 1$ corresponds to $\widehat{\lambda}_{k'}(\mathbf{x}_{k,i})$ being considered as an additional observation with maximum possible weight to estimate $\widehat{\lambda}_k$. Performance of the weights derived by (5.2.4) depends on the selection of ϵ. If ϵ is too large, strong similarity is assumed between functions which may actually be quite different. Conversely, only little statistical information is used from other data sets if ϵ is too small. Sensitivity with respect to ϵ is considered in Section 5.3.

5.3 Simulation Study

The methodology detailed in Section 5.2 is applied to simulated Gaussian data in order to examine its performance. Interest lies in the overall improvement, the sensitivity of the results with respect to ϵ in expression (5.2.4) and the computational cost. The underlying regression functions $\lambda_1, \ldots, \lambda_K$ are univariate and the distribution of the response Y_k, conditional on the predictor X_k, is formally given as

$$Y_k \sim \text{Normal}\left(\lambda_k(X_k), \sigma_k^2\right). \tag{5.3.1}$$

The loss functions $\Phi_{k,i}$ and $\widetilde{\Phi}_{k,k',i}$ in the optimization problem (5.2.3) are defined as the squared distance, that is, $\Phi(u,v) = \widetilde{\Phi}(u,v) = (u-v)^2$. Further, the weights $\{w_{k,i}\}$ in (5.2.3)

are set equal to 1 for all observations. Since λ_k is univariate, the PAVA can be applied in the described cyclic optimization routine in Section 5.2. Here, the `R` package `isotone` is used to obtain the individual updates of the monotonic functions. The functional levels $\widehat{\lambda}_{k'}(\mathbf{x}_{k,i})$ in (5.2.4) are obtained via linear interpolation. Let $x_l = \max\{x_{k',i} : x_{k',i} \leq x_{k,i}\}$ and $x_u = \min\{x_{k',i} : x_{k',i} \geq x_{k,i}\}$. The functional level of $\widehat{\lambda}_{k'}$ at z is then defined as

$$\widehat{\lambda}_{k'}(z) = \widehat{\lambda}_{k'}(x_l) + \frac{\widehat{\lambda}_{k'}(x_u) - \widehat{\lambda}_{k'}(x_l)}{x_u - x_l}(z - x_l). \tag{5.3.2}$$

If no point x_l exists, $\widehat{\lambda}_{k'}$ is set to the smallest estimated functional level and, conversely, to the highest estimated level if no x_u exists. As discussed in Section 5.2, the function λ_k, $k = 1, \ldots, K$ is only estimated over the covariate values in $\mathcal{D}_k$ and the set of points for which $\widetilde{w}_{k',k,i}$, $k' \neq k$, is positive in order to ensure uniqueness of the solution. This set depends on the value of ϵ and is denoted by $\mathcal{E}_k$ in the following

To assess the proposed approach, estimates of λ_k are also obtained classically by applying an optimization routine, the PAVA in this case, to the observations in $\mathcal{D}_k$, $k = 1, \ldots, K$. Results are compared with respect to the average bias over the set of points on which λ_k is estimated, that is, $\mathcal{E}_k$ for the new method and $\{\mathbf{x}_{k,i} : i = 1, \ldots, n_k\}$ otherwise. Let $\widehat{\lambda}_k$ and $\widehat{\lambda}_k^0$ denote the estimated functions obtained via the new and an existing algorithm, respectively. The model fit for $\widehat{\lambda}_k$ and $\widehat{\lambda}_k^0$ is then compared based on the ratio

$$C\left(\widehat{\lambda}_k, \widehat{\lambda}_k^0\right) = \frac{\sum_{i=1}^{n_k} \left|\lambda_k(\mathbf{x}_{k,i}) - \widehat{\lambda}_k^0(\mathbf{x}_{k,i})\right|}{\sum_{\mathbf{x} \in \mathcal{E}_k} \left|\lambda_k(\mathbf{x}) - \widehat{\lambda}_k(\mathbf{x})\right|}, \tag{5.3.3}$$

where a value of $C\left(\widehat{\lambda}_k, \widehat{\lambda}_k^0\right)$ greater than 1 corresponds to the new method performing better.

In the first simulation study, a collection of $K = 2$ subgroups is considered. The number of data points in $\mathcal{D}_1$ and $\mathcal{D}_2$ is set to 200 and 50, respectively. Covariate values for both sets are simulated uniformly from the interval 0 to 3 and the standard deviation in expression (5.3.1) is set to $\sigma_1 = \sigma_2 = 1$. Row 1 in Figure 5.3.1 illustrates the underlying monotonic regression functions and the sampled data points. Functional levels of λ_1 and λ_2 are similar up to the occurrence of a discontinuity for λ_2 which leads to quite different upper levels.

The algorithm is performed for a range of values and the ratio in (5.3.3) is derived for both data sets (Figure 5.3.2). The plots indicate that there exists values of ϵ for which the model fit is improved for both functions. Row 2 in Figure 5.3.1 provided the estimated functions $\widehat{\lambda}_1$

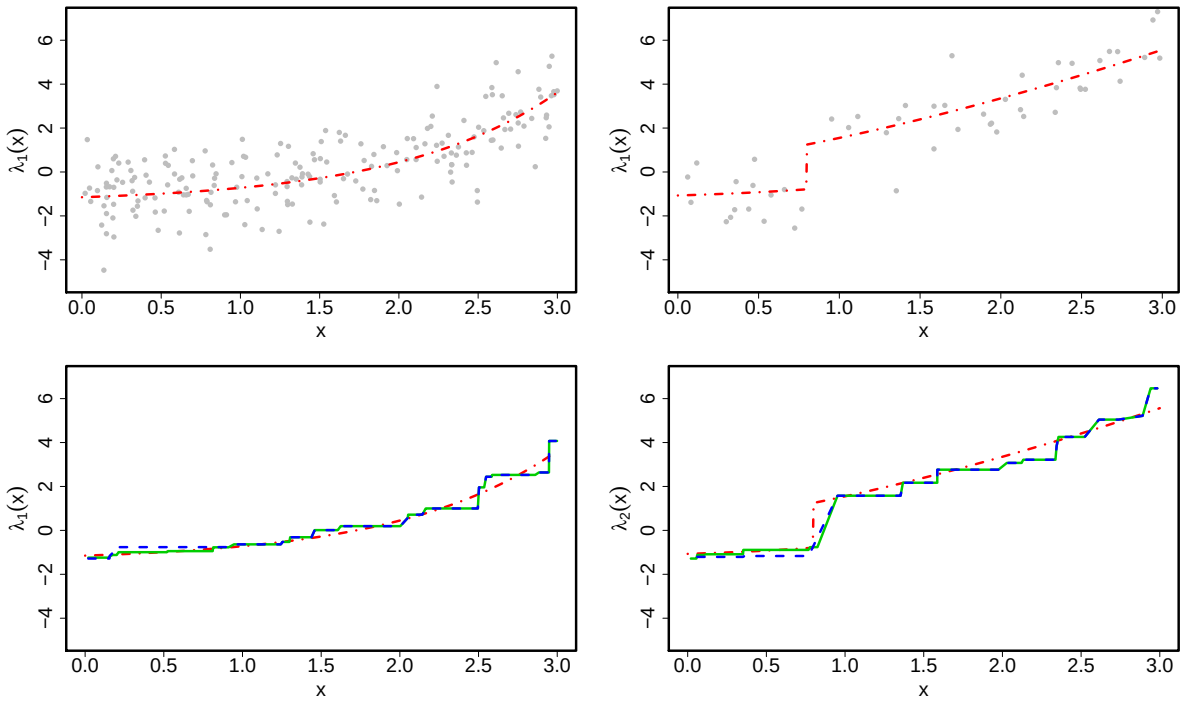


Figure 5.3.1: Simulated data points and underlying regression functions (-·-·-) used in Study 1 (Row 1). Row 2 illustrates the true underlying function λ_k (-·-·-), the estimate $\widehat{\lambda}_k$ obtained by the proposed algorithm (——) and the PAVA estimate $\widehat{\lambda}_k^0$ (- - - -).

and $\widehat{\lambda}_2$ for $\epsilon = 0.55$ which provided the highest combined value $C\left(\widehat{\lambda}_1, \widehat{\lambda}_1^0\right) + C\left(\widehat{\lambda}_2, \widehat{\lambda}_2^0\right)$. The plots illustrate that the proposed method leads to a better fit at the lower functional levels, as compared to the PAVA estimates. However, Figure 5.3.2 shows sensitivity with respect to ϵ since a worse model fit is obtained for small and very high values of ϵ. The latter is due to all points being considered informative and, thus, the difference in the upper levels leads to a poor fit. For small ϵ, the behaviour may be caused by a few points which are potentially not beneficial for estimating the true underlying regression function or possibly due to the points in the set $\mathcal{E}_k$ which depend on ϵ. Additionally, with respect to ϵ, the data set with less observations is more sensitive as the ratio of data points in $\mathcal{D}_2$ to points from $\widehat{\lambda}_1$ is higher. It is further found that the number of iterations until convergence increases in ϵ. While the algorithm converges in about 30 iteration steps for small ϵ, 250 iteration steps are required for $\epsilon = 2.0$. Nevertheless, the computational time is very small and the algorithm took less than 1 second for each considered value of ϵ.

In the second simulation study, a setting with $K = 5$ subgroups is considered. The distribution of the predictor X_k as well as the standard deviation in (5.3.1) varies across the subgroups but is otherwise constant. Table 5.3.1 summarizes the values for σ_k and the distribution G_k

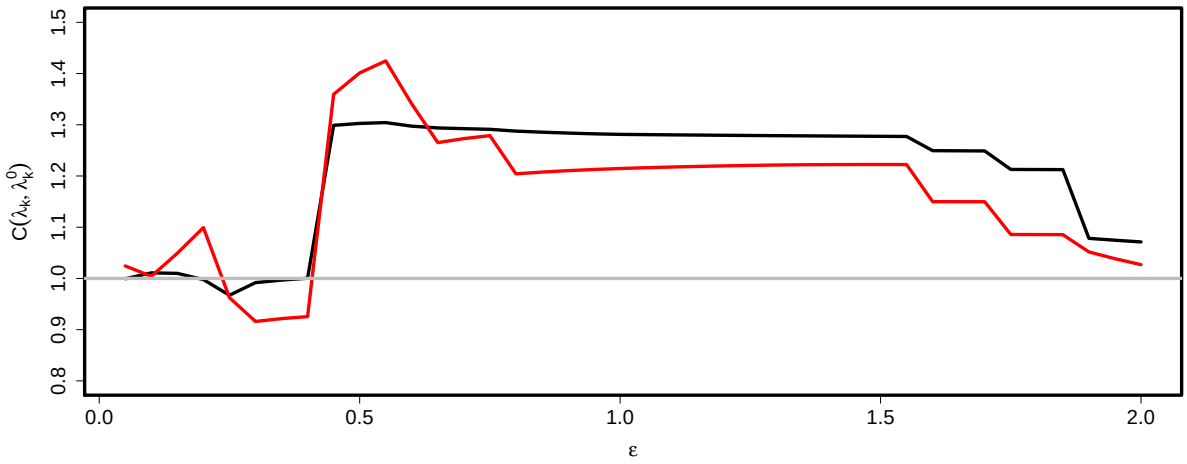


Figure 5.3.2: Dependence between the parameter ϵ in (5.2.4) and the ratio in (5.3.3) for Study 1 for (——) $\widehat{\lambda}_1$ and (——) $\widehat{\lambda}_2$

Table 5.3.1: Setup for Study 2 with a set of $K = 5$ subgroups and Gaussian distributed data.

Subgroup	1	2	3	4	5
G_k	Gamma(1.1,1)	Gamma(3,2)	Uniform(0,5)	Exp(0.8)	Weibull(2,2)
σ_k	0.4	0.3	0.3	0.3	0.3
n_k	300	200	100	50	25

of X_k. As in Study 1, a different number of data points is sampled for each subgroup. In particular, n_k for Δ_1 through Δ_5 is set to 300, 200, 100, 50 and 25, respectively. Column 1 in Figure 5.3.3 illustrates the true underlying monotonic functions and the sampled data points. The plots illustrate that λ_1, λ_3 and λ_4 exhibit a similar functional shape which is different from λ_2 and λ_5 with respect to the upper functional levels. Furthermore, all five functions exhibit a similar discontinuity around $x = 1.2$.

Again, a range of values is considered for the fixed constant ϵ in expression (5.2.4). Similarly to Study 1, there exists a range of values which lead to an improvement in the overall model fit. The selected value of ϵ leads to good improvements with respect to the estimates for λ_2, λ_4 and λ_5. On the other hand, very small improvements are obtained for λ_1 and λ_3 as the ratio is 1.01 for both. As for Study 1, the number of iterations until convergence increases with ϵ and the algorithm requires about 700 iteration steps for $\epsilon = 1.0$. Column 2 in Figure 5.3.3 shows the estimates of the monotonic functions obtained for the optimal ϵ and indicates that the true underlying regression function is fitted well. The plots further show that the proposed method also allow for extrapolation of the functions. In terms of the computational time, the algorithm

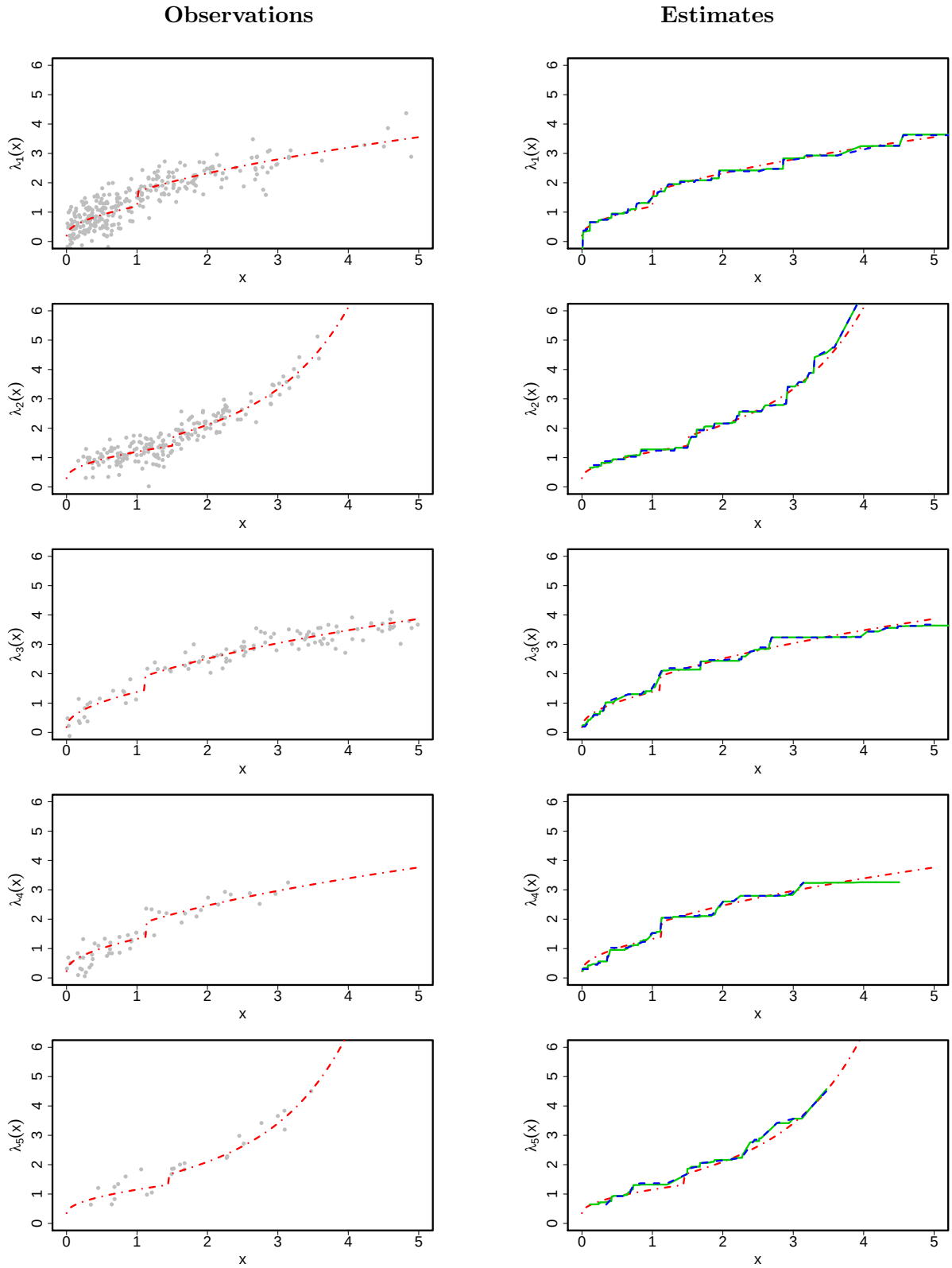


Figure 5.3.3: Simulated data points and underlying regression functions (-·-·-) used in Study 1 (Column 1). Column 2 illustrates the true underlying function λ_k (-·-·-), the estimate $\widehat{\lambda}_k$ obtained by the proposed algorithm (——) and the PAVA estimate $\widehat{\lambda}_k^0$ (- - -).

takes about 4-5 seconds per considered value of ϵ.

5.4 Discussion

This chapter introduced an alternative, optimization-based approach to the computationally expensive BSMMR methodology in Chapter 4. Similarly to BSMMR, the proposed algorithm considers a fixed number of subgroups and allows for sharing of statistical information between subgroups to improve the estimates. The methodology extends the classical isotonic regression problem by introducing additional compounds to penalize differences in the functional levels. Since the optimization problem is convex, its solution can be computed via a cyclic routine which updates each function while keeping the remaining ones fixed. Estimates are derived at the observed covariate values and the monotonic functions are then obtained via interpolation. This approach is less flexible than BSMMR and depends on the interpolation approach. To avoid a poor model fit, an approach to estimate weights based upon an initial isotonic regression fit is considered and requires the specification of one parameter. The algorithm has been applied to simulated Gaussian data and results show an improvement in the overall model fit of the monotonic functions. Further, some sensitivity is found with respect to the specified parameter.

The research done in this chapter can be extended in several ways and some aspects will be considered in future work. Here, only univariate functions are considered in the simulation study. However, the methodology is much more general and interest may lie to examine its performance for higher dimensions: isotonic recursive partitioning (Luss et al., 2012) or its generalization (Luss and Rosset, 2014) may then be used instead of the restrictive PAVA. Further, the specification of ϵ remains an open problem and has to be considered in future research. Some information may potentially be derived based on the variance of the residuals obtained for the initial estimates or via cross-validation. While the comparison to the BSMMR approach has only been done in terms of the computational cost, a more thorough analysis has to be performed which also compares the model fit. Finally, the algorithm has to be applied to the Norwegian insurance and weather data to assess its predictive performance.

Chapter 6

Extreme Value Modelling of Insurance Claims

6.1 Introduction

Since large parts of society and the economy are weather-sensitive, insurances against undesirable weather events have become an important economical factor. Mills (2005) state that the payout by insurance companies for weather related disasters in developing countries is three times higher than the international aid. In order to set premiums correctly, the insurance companies require accurate models. Thus, it is necessary to understand which weather events are responsible for damages. While natural disasters such as Hurricane Katrina, which caused damages of over $100 billion in 2005 (Knabb et al., 2005), lead to large monetary losses, the majority of insured losses are related to small scale weather events (Mills, 2005; Botzen and Van Den Bergh, 2008). Damages caused by precipitation are, for instance, studied by Schuster et al. (2006) and Kubilay et al. (2013). In this chapter, interest lies in the impact of small-scale weather events, e.g. heavy rain or snow-melt, and we aim to explain which weather events induce a high claim risk and to predict the number of claims related to those events. This approach is also important in the context of the current climate change which will affect both society and economy (Sanders and Phillipson, 2003; Jenkins et al., 2008; Botzen and van den Bergh, 2012).

We consider the insurance and weather data analyzed by Haug et al. (2011) and Scheel et al. (2013). The insurance data provide the daily number of claims caused by either precipitation, surface water, snow melt, undermined drainage, sewage back-flow or blocked pipes for all Norwegian municipalities between 1997 and 2006. Let $N_{k,t}$ denote the number of claims on day t

Table 6.1.1: Weather covariates provided by the Norwegian Meteorological Institute and the Norwegian Water Resources and Energy Directorate.

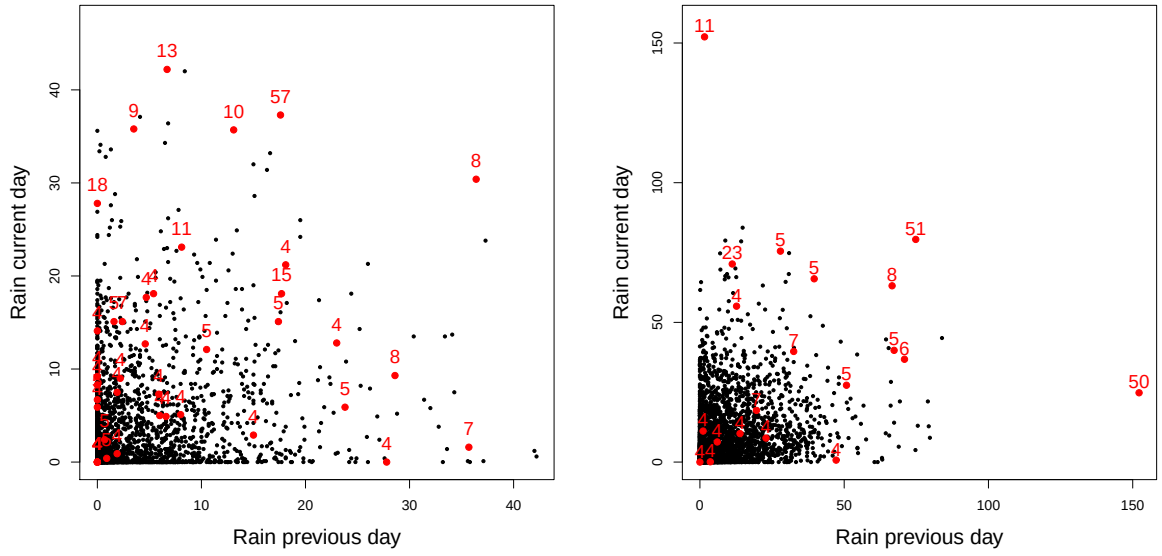
Variable	Description	Unit
$R_{k,t}$	Total amount of precipitation in day t (Between 6am on day t to 6am on day $t+1$)	mm
$C_{k,t}$	Mean temperature in day t	°C
$D_{k,t}$	Drainage run-off in day t	mm
$S_{k,t}$	Snow-water equivalent in day t (Amount of water in form of snow)	mm

for municipality k. Table 6.1.1 shows the set of meteorological and hydrological covariates $\mathbf{X}_{k,t}$ which are either empirical or model generated with a single value for each covariate representing day t and municipality k; see Section 1.3, Haug et al. (2011) and Scheel et al. (2013) for details. The weather data are derived by spatial interpolation, weighted proportionally to the population density within the municipality. Norway's climate varies due to the country's large geographical extent and the input of the Gulf Stream, inducing differences in the distribution of the weather observations. Comparison across municipalities shows, for instance, that western coastal areas observe relatively mild temperatures and large amounts of rainfall while central (inland) areas such as Oslo are drier and have more of a continental climate. These differences are likely to lead to a spatial variation of the claim dynamics and have to be accounted for in the modelling framework.

Scheel et al. (2013) propose a Bayesian Poisson hurdle (BPH) model for the dependence of $N_{k,t}$ on $\mathbf{X}_{k,t}$, since the high frequency of zero claims, $N_{k,t} = 0$, limits the applicability of a Poisson or Binomial distribution for $N_{k,t}$. Further, mechanisms leading to any claim in a region may be different from mechanisms for the number of claims given damage occurred. They also derive additional simple covariates from $\mathbf{X}_{k,t}$. Formally, their probability model is then given by

$$\mathbb{P}\left(N_{k,t} = n \mid \mathbf{X}_{k,t}\right) = \begin{cases} \alpha_{k,t} & \text{if } n = 0 \\ (1-\alpha_{k,t}) \dfrac{\lambda_{k,t}^{n}}{n!\left[\exp(\lambda_{k,t}) - 1\right]} & \text{if } n > 0, \end{cases} \tag{6.1.1}$$

where both $\lambda_{k,t} > 0$ and $\alpha_{k,t} \in [0,1]$ depend on $\mathbf{X}_{k,t}$ and the latter also depends on the number of policies. According to distribution (6.1.1), $\alpha_{k,t}$ corresponds to the frequency of zero claims while $\lambda_{k,t}$ is the rate of a zero-truncated Poisson distribution for the number of claims, given at least one claim is reported. The parameters $\lambda_{k,t}$ and $\alpha_{k,t}$ are separately modelled since they are



(a) Oslo **(b)** Bergen

Figure 6.1.1: Observed covariate values for $R_{k,t}$ and $R_{k,t-1}$ for the original data by Scheel et al. (2013) for Oslo (left panel) and Bergen (right panel). Days with the number of claims exceeding 3, $N_{k,t} > 3$, are highlighted.

conditionally independent, given the data; see Scheel et al. (2013) for details.

Scheel et al. (2013) assess the predictive performance of the BPH model on a weekly basis and the results are generally positive. Table 2 in Scheel et al. (2013) (Table 1.4.1 in Section 1.4) indicates, however, that the model substantially underperforms in weeks with high numbers of claims and underpredicts the impact of high precipitation levels, especially for Oslo. In conclusion, the distribution (6.1.1) is limited in its ability to model high numbers of claims, which is the most important feature of the model.

Figure 6.1.1 provides some insight into the causes of the lack of model fit for the BPH model. The plots illustrate the dependence between claims on day t and the amount of precipitation on the claim day t and the previous day $t-1$ for two of the municipalities. Firstly, higher claim numbers are not always associated to high precipitation levels on either day. Some claims associated with weak rainfall coincide with snow-melt but others cannot be linked to the weather covariates. The latter may be caused by localized weather events which are not recorded by any measurement station. Further, blocked pipes or sewage back-flow, which are also contained in the data, are not necessarily related to the weather on the same day. Ignoring such effects may influence the estimated model and lead to biased estimation of the covariate effects. Finally, while claim numbers for Oslo lie between zero and three claims on about 97% of days, much

higher numbers occur and these are generally related to high precipitation levels, sometimes in combination with snow-melt. A Poisson distribution is incapable of fitting these extremes while accounting for the high frequency of lower claims.

This chapter introduces several new methods in order to improve the model fit which have generic relevance to the modelling of insurance claim data. Interest lies, in particular, in the days with high numbers of claims. We extend the zero-truncated Poisson component in the BPH approach using extreme value and mixture models. Extreme value models such as the generalized Pareto distribution (GPD) are widely applied to estimate the tail of a random variable (Holmes and Moriarty, 1999; Coles, 2001; Li et al., 2005). Here, a discretized analogue of the GPD is defined since $N_{k,t}$ takes non-negative integer values only. Mixture models are considered in several areas, including finance (Lin et al., 2007) and medicine (Ozenne et al., 2015), and are also applied for the extreme value modelling in risk analysis (Smith and Goodman, 2000; Bottolo et al., 2003). Additional to advancing the statistical model, the input data are considered too. This leads to the derivation of new covariates which exploit temporal and spatial patterns in the weather covariates. These covariates are based on an exploratory analysis of the data for Oslo in Figure 6.1.1a. Furthermore, we introduce a temporal clustering algorithm to obtain periods of consecutive days which are exposed to the same weather event for each municipality. The distributions of clustered claims, conditional on covariates, over different municipalities are used to derive the marginal distribution of clustered claims. They also show that claims over different municipalities appear independent conditionally on the covariates, indicating that our model has captured the key meteorological factors that explain water-related insurance claims in Norway. The usefulness of these approaches is validated with respect to the data for the municipalities of Oslo (Figure 6.1.1a), Bergen (Figure 6.1.1b) and Bærum, and the likelihood of future extremes is predicted under the assumption of no climate change.

The remainder of this chapter is organized as follows: Section 6.2 details our extensions of the zero-truncated Poisson distribution and introduces an approach to optimize tail dependency for additional covariates. Section 6.3 defines the new covariates and introduces the temporal clustering algorithm. The extended model is then applied to the three Norwegian municipalities in Sections 6.4 and 6.5 and results are provided. Finally, the chapter concludes with a summary and discussion in Section 6.6.

6.2 Extension of the Bayesian Poisson Hurdle Model

This section details our extensions to the zero-truncated Poisson distribution in expression (6.1.1) to obtain a better model for claim occurrences $N_{k,t} \mid (\mathbf{X}_{k,t}, N_{k,t} > 0)$. For notational simplicity, the indexes k and t are dropped in the following. Section 6.2.1 introduces a mixture model while Section 6.2.2 defines an integer-valued GPD and combines it with the zero-truncated Poisson distribution via an extremal mixture model. Section 6.2.3 details a general methodology to optimize the tail dependence between a response and a family of predictors which is later applied in Section 6.3.

6.2.1 Mixture Modelling

Figure 6.1.1, coupled with exploratory analysis, indicates that claim dynamics are mainly driven by the observed precipitation and snow-melt levels but some claims may also be caused by additional processes which are not captured via the provided weather data. Information on the precise cause of damage, e.g. snow-melt or sewage back-flow, may allow the fit of a separate model for each cause but these are not available.

We propose a two-component mixture distribution with discrete positive-valued random variables Y and Z for $N \mid (\mathbf{X}, N > 0)$ to accommodate a potentially varying weather-dependence of these claim types. The first model component Y captures the dependence of N on the weather covariates $\mathbf{X}$ while the second component Z considers the claims which are caused by unobserved processes. All reported claims on a day are assumed to come from exactly one of the two components. The distribution function of $N \mid (\mathbf{X}, N > 0)$ is then formally given by

$$\mathbb{P}\left(N = n \mid \mathbf{X}, N > 0\right) \;= p\, \mathbb{P}\left(Y = n \mid \mathbf{X}\right) \;+\; (1-p)\, \mathbb{P}\left(Z = n\right),\; n \geq 1, \tag{6.2.1}$$

where p denotes the probability of $N \mid (\mathbf{X}, N > 0)$ being distributed according to $Y \mid \mathbf{X}$. Here, the component Z is defined as a zero-truncated Poisson distribution with rate parameter $\kappa > 0$

$$\mathbb{P}\left(Z = n\right) = \frac{\kappa^n}{n!\left[\exp(\kappa) - 1\right]},\; n \geq 1. \tag{6.2.2}$$

Note, the case $p = 0$ in distribution (6.2.1) corresponds to the BPH model in (6.1.1) without covariate structure. The choice of only two components is due to parsimony and the results in Section 6.4 show that this number is sufficient.

6.2.2 Extremal Mixture Modelling

Defining the mixture component Y in (6.2.1) as a zero-truncated Poisson distribution leads to a poor fit of the extreme claim numbers for Oslo and Bergen in Figure 6.1.1. Hence, we extend the model in order to allow for a more flexible tail behaviour. In particular, the lower claim numbers are still distributed according to a zero-truncated Poisson model but the highest observations are fitted using extreme value models. First, a distribution for the extremes of a discrete random variable is derived without the consideration of covariates. The zero-truncated Poisson model is then combined with this distribution and covariates are included.

Consider the modelling of $Y \mid Y > u$, where $u \in \mathbb{R}$ is a certain sufficiently high threshold. The discrete variable Y can be considered as $Y = \lfloor H \rfloor$, where H is a continuous random variable. In an extreme value modelling framework, the distribution of H above threshold u is generally modelled by a GPD with scale parameter σ_u and shape parameter ξ (Coles, 2001), a model that has asymptotic justification as u tends to the upper endpoint of H. For a large enough choice of u, the distribution of $H \mid H > u$ is then approximately given by

$$\mathbb{P}\left(H \leq h + u \mid H > u\right) = 1 - \left(1 + \frac{\xi h}{\sigma_u}\right)_+^{-\frac{1}{\xi}}, \; h > 0, \tag{6.2.3}$$

where $x_+ = \max(x, 0)$, $\sigma_u > 0$ and $\xi \in \mathbb{R}$, with the value for $\xi = 0$ interpreted as the limit as $\xi \to 0$. We then derive a discretized GPD to model $Y \mid Y > u$ via a GPD for H above threshold $\lfloor u \rfloor$. The probability mass function for $Y \mid Y > u$, for $n > u$, is then formally given by

$$\begin{aligned} \mathbb{P}\left(Y = n \mid Y > u\right) &= \mathbb{P}\left(H \leq n \mid H > \lfloor u \rfloor\right) - \mathbb{P}\left(H \leq n - 1 \mid H > \lfloor u \rfloor\right) \\ &= \begin{cases} \left[1 + \dfrac{\xi(n-1)}{\sigma_u}\right]_+^{-\frac{1}{\xi}} - \left[1 + \dfrac{\xi n}{\sigma_u}\right]_+^{-\frac{1}{\xi}} & \xi \neq 0 \\ \exp\left(-\dfrac{n-1}{\sigma_u}\right) - \exp\left(-\dfrac{n}{\sigma_u}\right) & \xi = 0. \end{cases} \end{aligned} \tag{6.2.4}$$

In the following, the distribution (6.2.4) is termed an *integer-valued Generalized Pareto distribution*, IGPD(σ_u, ξ, u), above threshold u with scale σ_u and shape ξ. Interpretation of the shape parameter ξ is equivalent to that of the GPD: a negative shape parameter $\xi < 0$ corresponds to the distribution being short-tailed, with upper bound. Conversely, $\xi > 0$ indicates a power-law tail, much heavier than a Poisson distribution.

Prieto et al. (2014) consider a similar formulation to expression (6.2.4) but they do not exam-

ine its properties with varying threshold, i.e., how the distribution changes as the threshold is increased to $v > u$. The GPD has the threshold stability, that is, if $H - u \mid H > u \sim \text{GPD}\,(\sigma_u, \xi)$, then for any higher threshold $v > u$, $H - v \mid H > v \sim \text{GPD}\,(\sigma_u + \xi(v-u), \xi)$. As such, ξ is constant with increasing threshold while the scale parameter σ_u is not. An equivalent property also holds for the defined IGPD. In particular, if $Y \mid Y > u \sim \text{IGPD}(\sigma_u, \xi, u)$, then for $v > u$ $Y \mid Y > v \sim \text{IGPD}(\sigma_u + \xi(\lfloor v \rfloor - \lfloor u \rfloor), \xi)$; see Appendix D.1 for the proof. This property is important since it allows the selection of a threshold u for the IGPD via a threshold stability plot, the same technique applied for a GPD (Coles, 2001).

Note, the Poisson distribution above a high threshold u does not follow an IGPD and there also exists no limiting generalized extreme value (GEV) distribution for the maximum of Poisson variables (Anderson, 1970, 1980). However, Anderson et al. (1997) show that asymptotically the Poisson follows a GEV distribution with shape parameter $\xi = 0$ as the rate parameter tends to infinity. Since the interpretation of the shape parameter is identical for GEV and GPD, an estimate of ξ that is statistically significantly different from zero for the IGPD indicates that the tail of the underlying distribution is not Poisson.

The IGPD (6.2.4) is combined with the zero-truncated Poisson distribution to form an extremal mixture distribution, i.e. a distribution with different forms below and above a threshold u. Such mixtures have been widely studied in a continuous variable setting (Coles and Tawn, 1991; Frigessi et al., 2002; Behrens et al., 2004; Carreau and Bengio, 2009; MacDonald et al., 2011) and the estimation of the threshold u is considered too. Here, observations smaller than or equal to u are zero-truncated Poisson distributed while being IGPD otherwise. Formally, the probability mass function for $Y \mid (\mathbf{X}, Y > 0)$ is then given by

$$\mathbb{P}\,(Y = n \mid \mathbf{X}, Y > 0) = \begin{cases} \dfrac{\lambda^n}{n!\,[\exp(\lambda) - 1]} & 1 \leq n \leq u \\ B_u\,\mathbb{P}\,(Y = n \mid \mathbf{X}, Y > u) & n > u, \end{cases} \tag{6.2.5}$$

where B_u denotes the probability of a zero-truncated Poisson distribution with parameter λ exceeding u and $\mathbb{P}\,(Y = n \mid \mathbf{X}, Y > u)$ is given by model (6.2.4). The parameters λ and σ_u both vary with the covariates $\mathbf{X}$.

6.2.3 Optimizing Tail Dependency of New Covariates

The generalized linear modelling framework by Scheel et al. (2013) has limited ability to account for the interaction effect of multiple risk factors; e.g. snow-melt and rainfall. This is due to a range of reasons, these include: simple interaction terms not capturing the non-linearity of the known physical properties of the relationship, parsimony, and a lack of weight given to extreme events when the signal to noise ratio is at its greatest. These weaknesses motivate our approach to construct an additional covariate, based upon $\mathbf{X}$, which overcomes these deficiencies and is tuned using extreme event data. In particular, a new covariate X^* is derived non-linearly from $\mathbf{X}$, as $X^* = f(\mathbf{X}, \boldsymbol{\theta})$, with unknown parameters $\boldsymbol{\theta}$ and the function f is selected based on the context of the problem. Such additional covariates are later defined in Section 6.3. Since X^* is motivated by the extreme claim numbers, $\boldsymbol{\theta}$ should be selected such that the tail dependence between X^* and N is optimized.

To achieve this aim, we adapt the approach by Russell et al. (2016) which is based on the following result. For identically distributed random variables V_1 and V_2 with unit Fréchet margins it follows, under a weak assumption of bivariate regular variation (Resnick, 1987), that for any Borel set B and $v \geq 1$

$$\lim_{t\to\infty} \mathbb{P}(V_1 + V_2 > tv,\ V_1/(V_1+V_2) \in B \mid V_1 + V_2 > t) = v^{-1}\Psi(\{B\}), \tag{6.2.6}$$

where Ψ is known as the spectral distribution, corresponding to a $[0,1]$ random variable with mean $\frac{1}{2}$. Expression (6.2.6) presents bivariate regular variation for an L_1 norm, though in practice if it holds for one norm, it holds for any norm, so the choice of the L_1 norm is without loss of generality. The weakest tail behaviour between V_1 and V_2 occurs when $\Psi(\{0\}) = \Psi(\{1\}) = 1/2$ and the strongest when $\Psi(\{\frac{1}{2}\}) = 1$, the former and latter corresponding to asymptotic independence (Ledford and Tawn, 1996) and perfect dependence respectively. Thus the greater the mass that the spectral measure places close to $\frac{1}{2}$ the stronger the tail dependence. There is no unique way to define the closeness of the spectral measure to $\frac{1}{2}$. A classic way of measuring dependence in extremes is via the coefficient of asymptotic dependence, χ, defined as

$$\chi = \lim_{t\to\infty} \mathbb{P}(V_2 > t \mid V_1 > t), \tag{6.2.7}$$

with larger values of χ corresponding to strong extremal dependence (Coles et al 1999). In terms

of Ψ we can write

$$\chi = 2\int_0^1 \min(w, 1-w)\, d\Psi(w). \tag{6.2.8}$$

This χ measure does not measure well the deviation of $W := V_1/(V_1+V_2)$ from $\frac{1}{2}$, as it does not distinguish between cases of extreme values of V_1 occurring with large V_2 (W near $\frac{1}{2}$) or typical values of V_2 (W near 1). We want a measure which strongly penalizes departures of W from $\frac{1}{2}$, with the penalty symmetric about $\frac{1}{2}$. Although we could have taken the penalty as $|w-\frac{1}{2}|$, its empirical performance was not strong in our experience. Instead we use

$$D_\epsilon = \int_0^1 \min\left\{ \left|\log\left(\frac{w}{1-w}\right)\right| , \left|\log\left(\frac{\epsilon}{1-\epsilon}\right)\right| \right\} \mathrm{d}\Psi(w). \tag{6.2.9}$$

Here the $|\log\{w/(1-w)\}|$ term penalizes departures from $\frac{1}{2}$ more strongly. The value of ϵ is assumed to be fixed and sufficiently small. With the penalty function as in expression (6.2.9), $D_\epsilon \geq 0$ with $D_\epsilon = 0$ occurring when $\Psi(\{\frac{1}{2}\}) = 1$ and D_ϵ is increasing as Ψ places mass further from $\{\frac{1}{2}\}$ with $D_\epsilon \to \left|\log\left(\frac{\epsilon}{1-\epsilon}\right)\right|$ as asymptotic independence is approached. Here $\left|\log\left(\frac{w}{1-w}\right)\right|$ is bounded by the minimum term in the integrand as $w \to 0$ and $w \to 1$ to avoid cases where Ψ puts any mass at $\{0\}$ and $\{1\}$ leading to $D_\epsilon \to \infty$.

To apply the asymptotic property of bivariate regular variation in practice when we have observations $(V_{1,i}, V_{2,i})$ for $i = 1, \ldots, m$, we need to assume that limit (6.2.6) holds for a finite t, i.e.,

$$\mathbb{P}(V_1/(V_1+V_2) \leq w \mid V_1 + V_2 > t) = \Psi(w). \tag{6.2.10}$$

To ensure that the limit holds, t needs to be large enough to give the conditional independence of variables $V_1 + V_2$ and $V_1/(V_1+V_2)$ for the limit (6.2.6) to factorize. For the lowest choice of t for which conditional independence is a reasonable assumption, Ψ can be estimated using the set of points $(V_{1,i}, V_{2,i})$ with $V_{1,i} + V_{2,i} > t$, denoted by Q_t, by

$$\tilde{\Psi}_t(w) = \frac{1}{|Q_t|}\sum_{i=1}^{n} \mathbb{1}(V_{1,i} + V_{2,i} > t \;\&\; V_{1,i}/(V_{1,i}+V_{2,i}) \leq w), \tag{6.2.11}$$

with $\mathbb{1}$ being the indicator function. The dependence measure D_ϵ is then approximated using

$$\begin{aligned}\tilde{D}_{\epsilon,t} &= \int_0^1 \min\left\{ \left|\log\left(\frac{w}{1-w}\right)\right| , \left|\log\left(\frac{\epsilon}{1-\epsilon}\right)\right| \right\} \mathrm{d}\tilde{\Psi}(w) \\ &= \frac{1}{|Q_t|}\sum_{i=1}^{n}\left|\log\left(\frac{w_i}{1-w_i}\right)\right| \mathbb{1}(V_{1,i}+V_{2,i}>t), \\ &= \frac{1}{|Q_t|}\sum_{i=1}^{n}\left|\log\left(\frac{V_{1,i}}{V_{2,i}}\right)\right| \mathbb{1}(V_{1,i}+V_{2,i}>t)\end{aligned} \tag{6.2.12}$$

where $w_i = V_{1,i}/\left(V_{1,i}+V_{2,i}\right)$, with the second equality holding when $\epsilon < w_i < 1-\epsilon$ for all i with $V_{1,i}+V_{2,i} > t$. To apply this dependence measure to construct the covariate X^*, we transform the observations of X^* and N to Fréchet margins (V_1, V_2) and then select $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^* = \operatorname{argmin} \tilde{D}_{\epsilon,t}$ and set $X^* = f\left(\mathbf{X}, \boldsymbol{\theta}^*\right)$.

6.3 Restructuring the Data

This section introduces our algorithm to obtain clusters of consecutive days which are exposed to the same severe weather event. Prior to the algorithm, we derive additional covariates in Sections 6.3.1 to 6.3.3 based upon the assumption that adjacent municipalities may provide additional insight. In particular, these covariates account for spatial and temporal patterns of snow-melt and rainfall and are partly set as inputs in the clustering algorithm. Section 6.3.4 details the clustering algorithm which is based on the weather covariates. Covariates summarizing the weather events over the derived cluster periods are defined in Section 6.3.5. Finally, the event-based covariates are tuned to increase their ability to describe the occurrence of the largest numbers of claims in Section 6.3.6. In the following, the notation $k' \sim k$ refers to municipalities k and k' being adjacent.

6.3.1 Snow-melt

Long periods of snow-melt, or rapid melts of large volumes of snow, can give flood levels that are comparable to large rainfall events. Hence, periods of high temperatures or rain, conditional on snow being on the ground, may affect the claim dynamics and induce a higher risk for property damages. Information on the level of snow-melt is derived via the daily observed mean temperature $C_{k,t}$ and the snow-water equivalent $S_{k,t}$. Scheel et al. (2013) consider the difference in the snow-water equivalent over a day, i.e., $S_{k,t-1} - S_{k,t}$, which for positive values represents an additional source of water for properties to deal with. Estimates indicate a positive correlation

of this difference with respect to the claim risk for several municipalities. However, negative values of $S_{k,t-1} - S_{k,t}$ do not affect the water-related claim dynamics on the day since these only correspond to a rise of the amount of snow on the ground. Positive values of the difference $S_{k,t-1} - S_{k,t}$ will only approximate the true amount of snow-melt in municipality k. Certain topological factors are likely to be ignored since observations are weighted according to the population density. Consider a city which lies at the foot of a mountain range. The buildings are then affected by the snow-melt both within the city and on higher ground while $S_{k,t-1} - S_{k,t}$ captures the former only.

We use the observations for the adjacent municipalities to introduce a new snow-melt covariate $\Delta S_{k,t}$ as a spatially weighted average. In particular, our formulation for $\Delta S_{k,t}$ varies from Scheel et al. (2013) as $\Delta S_{k,t} > S_{k,t-1} - S_{k,t}$ if an adjacent municipality exhibits higher levels of snow-melt. Formally, $\Delta S_{k,t}$ is defined by

$$\Delta S_{k,t} = \frac{1}{1+\omega_k^S}\left[S_{k,t-1} - S_{k,t} + \omega_k^S \max_{m\in\{k,k'\sim k\}} (S_{m,t-1} - S_{m,t})\right] \mathbb{1}_{\{C_{k,t}\geq 0\}}, \tag{6.3.1}$$

with weight $\omega_k^S \geq 0$. The maximum term in (6.3.1) is derived over the set of adjacent municipalities $k' \sim k$ and k itself. Note, $\Delta S_{k,t} = S_{k,t-1} - S_{k,t}$ if snow-melt in municipality k exceeds snow-melt in its neighbours and $C_{k,t} > 0$, or if $\omega_k^S = 0$. The indicator function is set in order to ensure that no snow-melt occurs for temperatures $C_{k,t}$ below 0°C.

6.3.2 Surface Water

An increased claim risk is induced by the interaction of multiple weather events or the duration of one event over consecutive days. Scheel et al. (2013) attempt to account for such processes via the values of two covariates: the drainage run-off $D_{k,t}$ and the aggregated rain on the previous three days, denoted by $R_{k,3t}$. Their results indicate that both $R_{k,3t}$ and $D_{k,t}$ have a small effect on the distribution of $N_{k,t} \mid N_{k,t} > 0$. However, $R_{k,3t}$ and $D_{k,t}$ are limited in their potential to explain interaction effects. Values for $D_{k,t}$ change very slowly from day to day, that is, $D_{k,t}$ may be high despite the last rain being several days ago. Further, $R_{k,3t}$ cannot distinguish whether high amounts of rainfall were recorded two or three days ago. The derivation of new covariates appears advisable.

To help our construction of a new covariate, we consider a highly idealized model of the ability of the infrastructure to handle surface water. Assume that a maximum c_k mm of water

drains off within a day. Here, the value c_k may correspond to a certain quantile of the observed rain and be linked to the capacity of the drainage system. The amount of water left in the system on day t, $W_{k,t}$, is then given by

$$W_{k,t} = (W_{k,t-1} + R_{k,t-1} + \Delta S_{k,t-1} - c_k)_+ . \tag{6.3.2}$$

A value of $W_{k,t}$ greater than 0 implies that the previous weather events affect the risk induced by the weather on day t, for instance, in form of surface water. Further, $W_{k,t}$ is assumed to influence the claim dynamics if, and only if, $R_{k,t} + \Delta S_{k,t} > c_k$ since the value $W_{k,t}$ in (6.3.2) decreases otherwise, implying that no additional properties are threatened by surface water. This results in the definition of a new *amplifier* covariate,

$$G_{k,t} = W_{k,t} \mathbb{1}_{\{R_{k,t}+\Delta S_{k,t} > c_k\}}, \tag{6.3.3}$$

which captures the risk induced by heavy rainfall in combination with high surface water levels.

6.3.3 Rainfall Intensity

Since the covariate $R_{k,t}$ corresponds to the aggregated precipitation measurements over 24 hours, it provides little insight into the peak-daily intensity. High values of $R_{k,t}$ can be due to either short-term intense or longer-term moderate rainfall but the former is likely to induce a higher risk for property flooding. We attempt to derive additional information from the spatial variation of $\{R_{k,t}\}$ on day t. To achieve this, we assume that the intensity correlates with the difference in the precipitation levels of adjacent municipalities. Further, an intense rainfall within a municipality is also taken to affect the claim dynamics of the adjacent municipalities, though on a smaller scale.

These considerations result in our definition of the covariate *intensity*, $I_{k,t}$, which is based on the spatial pattern of $\{R_{k,t}\}$ at day t. Let $\widetilde{k}$ be the municipality, adjacent to municipality k, with the highest level of precipitation, i.e.,

$$\widetilde{k} = \underset{k' \sim k}{\operatorname{argmax}} \; R_{k',t}.$$

If $R_{k,t}$ is larger than $R_{\widetilde{k},t}$, the centre of the rainfall event lies within municipality k and, hence, may be rather intense. Similarly, if $R_{\widetilde{k},t} > R_{k,t}$, we consider the adjacent municipalities $k' \sim \widetilde{k}$

to explore whether the rainfall event leads to the highest precipitation levels in municipality $\widetilde{k}$. In order to represent the impact of a rainfall event at municipality $\widetilde{k}$ for municipality k, we introduce a weight $\omega_k^R \in [0,1]$ to downscale the intensity. Finally, if the rainfall is centred in neither of these municipalities, the rainfall is considered as not intense. The covariate value $I_{k,t}$ is then defined as

$$I_{k,t} = \begin{cases} R_{k,t} - R_{\widetilde{k},t} & \text{if } R_{k,t} > R_{\widetilde{k},t} \\ \omega_k^R \left(R_{\widetilde{k},t} - \max\limits_{k' \sim \widetilde{k}} R_{k',t} \right) & \text{if } R_{\widetilde{k},t} > \max\limits_{k' \sim \widetilde{k}} R_{k',t} \\ 0 & \text{otherwise.} \end{cases} \tag{6.3.4}$$

Note, the last case in (6.3.4) corresponds to the municipalities k and $\widetilde{k}$ observing lower precipitation levels than at least one of their adjacent neighbours. The upper bound for ω_k^R is justified since $I_{k,t}$ should not be higher than $I_{k',t}$ if the highest precipitation levels are recorded for municipality k'. Similarly to $G_{k,t}$, $I_{k,t}$ only affects the claim dynamics for high rainfall levels, $R_{k,t} > c_k$, since the intensity of the rainfall is presumably not important for the claim dynamics otherwise.

6.3.4 Cluster Definition

In many cases, the observations of consecutive day claim numbers $N_{k,t}$ and $N_{k,t+1}$, are dependent as they are consequences of the same weather event. Figure 6.1.1 indicates that high numbers of claims caused by heavy rain $R_{k,t}$ are typically reported on the current or next day. For instance, the highest rainfall level in Figure 6.1.1b results in observations of 11 and 50 claims. Nevertheless, some claims may be reported on later days and the covariate observations on the day of the claim would then provide little insight. Therefore, we derive periods of consecutive days, for each municipality individually, in order to cluster days which are exposed to a higher claim risk due to the same severe weather event and hence also reduce the effects of claim lag in the recording process.

Interest lies in the derivation of cluster periods $\{(\alpha_{k,j}, \beta_{k,j}), j = 1, \ldots, J_k\}$ for municipality k, based upon the weather covariates $\mathbf{X}_{k,t}$, with $\alpha_{k,j},\ \beta_{k,j}$ representing the start and end point of the jth of J_k clusters in municipality k respectively. While the daily claims within a cluster period $(\alpha_{k,j}, \beta_{k,j})$ are assumed to depend on the same severe weather event, the claims in two different clusters are considered as temporally independent. In particular, the claim dynamics

on day $\alpha_{k,j}$ are solely dependent on the weather events on the same day, irrespective of the weather on day $\beta_{k,j-1}$.

Our approach to identify cluster start points $\alpha_{k,j}$ is based upon two prespecified trigger events which affect the claim dynamics on subsequent days: rain on the current day exceeds c_k, $R_{k,t} > c_k$, and snow-melt occurs, $\Delta S_{k,t} > 0$. The first trigger event is motivated by the discussion in Section 6.3.2 while the second trigger reflects our expectation that snow-melt in combination with rainfall induces a high claim risk over several days. These events then initialize clusters of length greater than one day. The main criterion for the end of a cluster considers the change in the drainage run-off, i.e., $\Delta D_{k,t} = D_{k,t} - D_{k,t-1}$. In particular, a cluster period ends if $\Delta D_{k,t}$ drops below a certain value d_k. Additionally, clusters triggered by snow-melt also end if no snow is left on the ground. Algorithm 6.1 summarizes our cluster approach which is based on the covariates $R_{k,t}$, $D_{k,t}$ and $\Delta S_{k,t}$.

Algorithm 6.1 Derive clusters for municipality k

Require: Weather covariates $\Delta S_{k,t}$, $\Delta D_{k,t}$, $R_{k,t}$, and thresholds c_k and d_k

Go to first time point $t = 1$
while Unclustered observations left **do**
  if $\Delta S_{k,t} > 0$ **then**
    Set start point $\alpha = t$ and initial end point $\beta = t + 1$
    while $\Delta D_{k,\beta} > d_k$ AND $\Delta S_{k,\beta} > 0$ **do**
      Shift end point $\beta \leftarrow \beta + 1$
    end while
  else if $R_{k,t} > c_k$ **then**
    Set start point $\alpha = t$ and initial end point $\beta = t + 1$
    while $\Delta D_{k,\beta} > d_k$ **do**
      Shift end point $\beta \leftarrow \beta + 1$
    end while
  else
    Set start and end point to $\alpha = \beta = t$
  end if
  Store start and end points of cluster period (α, β)
  Go to next time point $t = \beta + 1$
end while
return Cluster periods

6.3.5 Cluster Data

The daily data have to be adapted to the cluster periods derived by Algorithm 6.1. Consider the jth cluster period for municipality k with start and end point $\alpha_{k,j}$ and $\beta_{k,j}$, respectively.

The number of claims over the jth cluster period, $\widetilde{N}_{k,j}$, is then

$$\widetilde{N}_{k,j} = \sum_{t=\alpha_{k,j}}^{\beta_{k,j}} N_{k,t}. \tag{6.3.5}$$

Snow-melt for cluster j is summarized via the accumulated amount over the cluster period

$$\Delta S_{k,j}^{\Sigma} = \sum_{t=\alpha_{k,j}}^{\beta_{k,j}} \Delta S_{k,t}, \tag{6.3.6}$$

where $\Delta S_{k,t}$ is defined via (6.3.1).

To capture both the maximum and aggregated rainfall, covariates $R_{k,j}^{\max}$ and $R_{k,j}^{\Sigma}$ are defined, respectively. While $R_{k,j}^{\max}$ focuses on a single day over the cluster period, $R_{k,j}^{\Sigma}$, takes the amount of precipitation over all days into account. Let γ_j denote the day over the period $\alpha_{k,j}$ to $\beta_{k,j}$ with highest value $R_{k,t}$. Then

$$R_{k,j}^{\max} = \eta_k \; G_{k,\gamma_j} + R_{k,\gamma_j} \exp\left(\rho_k \; I_{k,\gamma_j}\right), \tag{6.3.7}$$

where G_{k,γ_j} and I_{k,γ_j} are defined as in (6.3.3) and (6.3.4), respectively. The parameters η_k and ρ_k are selected to optimize the tail dependence of $R^{\max}$ and $\widetilde{N}$, details are given in Section 6.3.6. The non-linear structure of expression (6.3.7) aims to account for two separate claim processes which are associated to rainfall. In particular, the first additive component accounts for the risk in terms of surface water induced by previous rainfall events while the second component considers the rainfall on the day. The impact of the rainfall on the day for claims depends on both the rainfall and its intensity. Our arguments for the construction of the covariates $G_{k,t}$ and $I_{k,t}$ suggests that $\eta_k \in [0,1]$ and $\rho_k \geq 0$. Covariate $R_{k,j}^{\Sigma}$ is

$$R_{k,j}^{\Sigma} = \sum_{t=\alpha_{k,j}}^{\beta_{k,j}} R_{k,t} - R_{k,\gamma_j}, \tag{6.3.8}$$

i.e., the aggregation of the rainfall, except the highest, in the cluster. Note, $R_{k,j}^{\Sigma}$ takes value zero if the jth cluster is of length 1.

6.3.6 Selection of Parameter Values

The covariates introduced in this work depend on several parameters whose tuning is considered in this section. First, the parameter ω_k^S in (6.3.1) is selected based upon a simple generalized linear model fit for the original daily data for municipality k. The parameter ω_k^S has to be estimated prior to the cluster algorithm since it is important to gain insight into whether $\omega^S = 0$ or not. The maximum likelihood estimator of ω_k^S is found using the model

$$N_{k,t} \sim \text{Poisson}\left(\exp[\phi_0 + \phi_1 \Delta S_{k,t}(\omega_k^S)] \right).$$

The parameter may be estimated again after the clustering algorithm but the results in Section 6.4 are obtained without this additional step.

Next, the thresholds c_k and d_k in Algorithm 6.1 are specified. An explanatory analysis for Oslo indicates that periods with high numbers usually coincide with rainfall events exceeding the 80% quantile. Similarly, claim occurrences are observed if the change in drainage levels exceeds the 80% quantile. Hence, the thresholds are specified via quantiles as $c_k = q_{0.8}\left(R_{k,t} \mid R_{k,t} > 0\right)$ and $d_k = q_{0.8}\left(\Delta D_{k,t}\right)$. While this choice is motivated based on Oslo only, Section 6.4 indicates that it works for other municialities too. Algorithm 6.1 is then applied to the data.

The vector of covariate observations of the maximum rainfall covariate in expression (6.3.7), $\mathbf{R}_k^{\max}$, depends on the parameters ρ_k, η_k and also on the weight ω_k^R via $I_{k,t}$. Since $I_{k,t}$ and $G_{k,t}$ are predominately designed with respect to the high numbers of claims, ρ_k, η_k and ω_k^R are selected such that the tail dependency between $\mathbf{R}_k^{\max}$ and $\widetilde{\mathbf{N}}_k$ in expression (6.3.5) is maximized. Here, we adapt the approach detailed in Section 6.2.3 with $X^* = R^{\max}$ with $f(\mathbf{X}, \boldsymbol{\theta})$ given by expression (6.3.7) and the optimization is over a set of candidates for $\boldsymbol{\theta}_k = \left(\eta_k,\ \rho_k, \omega_k^R\right)$. This involves first transforming the data to Fréchet margins, selecting a threshold t above which the conditional independence property (6.2.6) holds, then estimating $\tilde{\Psi}_t(w)$ and finally deriving the distance measure $\tilde{D}_{\epsilon,t}$ for each candidate. Combining these ideas leads to the following selection process for the optimal candidate:

1. Derive the covariate values $\mathbf{R}_k^{\max}\left(\boldsymbol{\theta}^*\right)$ for each candidate $\boldsymbol{\theta}^*$ on a grid.

2. Use the empirical distribution functions and the probability integral transform to transform

$\mathbf{R}_k^{\max}$ and $\widetilde{\mathbf{N}}_k$ to have Fréchet margins

$$\mathbf{N}^* = -\left\{\log\left[\frac{rank\left(\widetilde{\mathbf{N}}_k\right)}{m+1}\right]\right\}^{-1} \quad \text{and} \quad \mathbf{R}^* = -\left\{\log\left[\frac{rank\left(\mathbf{R}_k^{\max}\right)}{m+1}\right]\right\}^{-1}.$$

3. The threshold t in (6.2.11) is chosen as a 99.5% quantile of the set $\{\mathbf{N}^* + \mathbf{R}^*\}$. Further set

$$Q_t = \{i = 1, \ldots, m \ : \ N_i^* + R_i^* \ > \ t := q_{0.995}\left(\mathbf{N}^* + \mathbf{R}^*\right)\}.$$

4. Derive the distance measure as outlined in Section 6.2.3. Via the substitutions $V_{1,i} = N_i^*$ and $V_{2,i} = R_i^*$, the distance measure in (6.2.12) yields

$$\tilde{D}_{\epsilon,t} = \frac{1}{|Q_t|}\sum_{i \in Q_t} |\log(N_i^*) - \log(R_i^*)|\, \mathbb{1}(N_i^* + R_i^* > t).$$

5. The optimal set of parameters $\boldsymbol{\theta}^*$ is then the one which provides the minimum $\tilde{D}_{\epsilon,t}$.

6.4 Application to the Insurance Data

Performance of the approaches in Sections 6.2 and 6.3 is assessed by applying them to three Norwegian municipalities: Oslo, Bærum and Bergen, where the first two are adjacent and the latter approximately 300 miles away from them. Oslo and Bergen were chosen since these are the municipalities with the highest number of policies and the model by Scheel et al. (2013) is limited in terms of capturing their highest claims. Bærum was selected based on its spatial proximity to Oslo and due to the highest observation over the 10 year period for Norway being observed for this municipality. Section 6.4.1 defines the statistical model and details the estimation of the model parameters. Section 6.4.2 then summarizes the estimates and investigates the model fit.

6.4.1 Statistical Model

Each of *the* three municipalities is considered separately and the indexes k, j and t for the municipality, the cluster period and the day, respectively, are dropped for notational simplicity. The optimisation approach in Section 6.3.6 derives that $\rho = 0$ for Bærum and Bergen; the latter is plausible since Bergen is surrounded by mountain ranges, leading to potentially high but uninformative values for the intensity I. Table 6.4.1 shows that about one third of days are

Table 6.4.1: Occurrence of cluster lengths for three Norwegian municipalities.

Cluster length	1	2	3	4	5	6	> 6
Oslo	2091	254	57	98	43	23	17
Bærum	2453	105	43	92	46	19	18
Bergen	1868	340	55	131	39	23	11

allocated to clusters of length greater than 1. It also shows that clusters are almost always less than 7 days, which is the window that the insurance industry typically treats as a single event for reinsurance purposes. Figure 6.4.1 illustrates that, post clustering, most of the high number of claims coincide with high values for $R^{\max}$ and R^{Σ}, suggesting that our methods of Section 6.3 for constructing justifiable covariates and their relationship to claims has been successful.

The statistical models in Section 6.2 are applied to model the dependence between periods with positive numbers of claims, $\widetilde{N} > 0$, and the covariates $\widetilde{\mathbf{X}} = \left(R^{\Sigma}, \Delta S^{\Sigma}, R^{\max}\right)$. Despite clustering, there still appear to be some modest claims which do not coincide with large values of the covariates. Hence, the mixture model in Section 6.2.1 is applied. The extremal mixture model in Section 6.2.2 is applied too as the highest numbers of claims are not well captured via a zero-truncated Poisson distribution. Hence, $\widetilde{N} \mid \left(\widetilde{\mathbf{X}}, \widetilde{N} > 0\right)$ is modelled via a two-component mixture with a covariate-dependent component $\widetilde{Y}$ and a random component $\widetilde{Z}$.

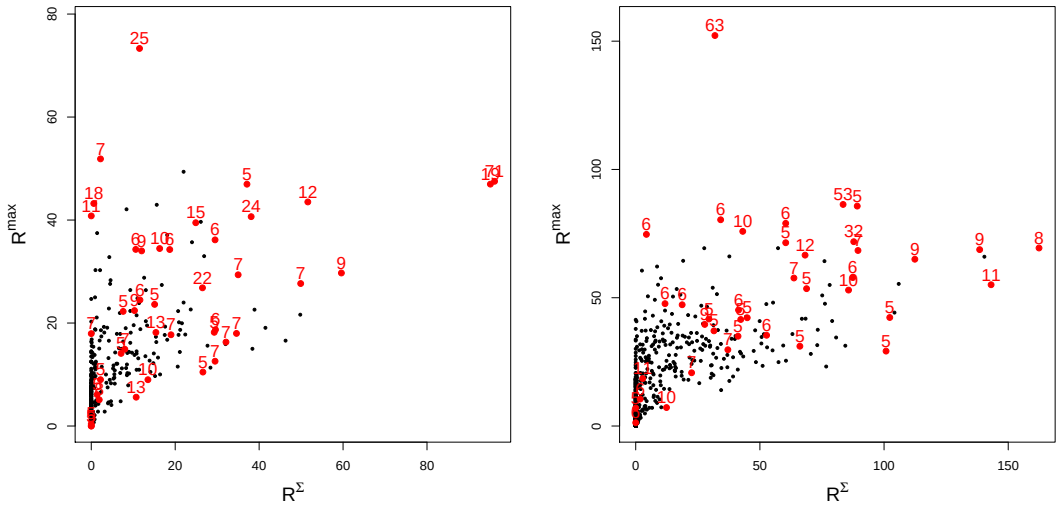
The mixture component $\widetilde{Y}$ in (6.2.1) is defined according to (6.2.5) as an extremal mixture of a zero-truncated Poisson and an IGPD while $\widetilde{Z}$ is defined as a zero-truncated Poisson distribution with constant rate parameter κ. Formally, the model for $\widetilde{N} \mid \left(\widetilde{\mathbf{X}}, \widetilde{N} > 0\right)$ is

$$\mathbb{P}\left(\widetilde{N} = n \mid \widetilde{\mathbf{X}}, \widetilde{N} > 0\right) = p\, \mathbb{P}\left(\widetilde{Y} = n \mid \widetilde{\mathbf{X}}\right) + (1-p)\, \mathbb{P}\left(\widetilde{Z} = n\right),\ n \geq 1. \tag{6.4.1}$$

Adapting the approach in Section 6.2.2, the probability mass function of $\widetilde{Y}$ is then given by

$$\mathbb{P}\left(\widetilde{Y} = n \mid \widetilde{\mathbf{X}}\right) = \begin{cases} \dfrac{\lambda^n}{n!\left[\exp(\lambda) - 1\right]} & 1 \leq n \leq u \\ \mathbb{P}\left(\widetilde{Y} > u \mid \widetilde{\mathbf{X}}\right)\ \mathbb{P}\left(\widetilde{Y} = n \mid \widetilde{\mathbf{X}}, \widetilde{Y} > u\right) & n > u, \end{cases} \tag{6.4.2}$$

where $\mathbb{P}\left(\widetilde{Y} = n \mid \widetilde{\mathbf{X}}, \widetilde{Y} > u\right)$ is of IGPD-form as in expression (6.2.4). We complete the model



(a) Oslo **(b)** Bergen

Figure 6.4.1: Dependence between the aggregated rain R^{Σ} and the maximum rain within a day $R^{\max}$ for the cities of Oslo (left panel) and Bergen (right panel). Periods with $\widetilde{N} > 4$ are highlighted.

by specifying the dependence of λ and σ_u on R^{Σ}, ΔS^{Σ} and $R^{\max}$ via linear models of the form

$$\begin{aligned} \log \sigma_u &= \beta_0 + \beta_1 R^{\Sigma} + \beta_2 \Delta S^{\Sigma} + \beta_3 R^{\max} \\ \log \lambda &= \delta_0 + \delta_1 R^{\Sigma} + \delta_2 \Delta S^{\Sigma} + \delta_3 R^{\max}. \end{aligned} \tag{6.4.3}$$

Note, the tail of the distribution defined by (6.4.1) and (6.4.2) is a mixture of zero-truncated Poisson and an IGPD. Similarly to the IGPD, threshold stability in the tails can also be proven for this mixture distribution; see Appendix D.2 for details. Using this threshold stability property, the threshold u is set to 4, 2 and 4 for Oslo, Bærum and Bergen, respectively.

6.4.2 Results

The statistical model in expressions (6.4.1) to (6.4.3) is specified by 11 parameters which are now estimated via Bayesian inference. Specifically, a Metropolis-Gibbs algorithm is used; see Appendix D.3 for details. Alternatively, estimates may also be obtained via an Expectation-Maximization algorithm. However, we found that this led to poor estimates since the support of the component $\widetilde{Y}$ varies in the shape parameter ξ, for $\xi < 0$. The MCMC algorithm runs for 100,000 iterations and every 50th sample is stored for analysis after a burn-in of 25,000. Convergence is checked by investigation of the trace plots and Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998) based on three sampled chains. Our `R` implementation took about

Table 6.4.2: Posterior mean estimates, lower 5% quantile ($q_{0.05}$) and upper 5% quantile ($q_{0.95}$) of the model parameters for the municipalities of Oslo, Bærum and Bergen with thresholds $u_k = 4, 2$ and 4, respectively.

City	Statistic	p	β_0	β_1	β_2	β_3	ξ	δ_0	δ_1	δ_2	δ_3	κ
Oslo	Mean	0.90	0.12	0.21	0.23	0.76	-0.32	-0.16	0.42	0.32	0.71	2.21
	$q_{0.05}$	0.83	-0.61	0.09	0.10	0.46	-0.76	-0.30	0.29	0.22	0.57	1.65
	$q_{0.95}$	0.96	0.81	0.33	0.35	1.03	0.15	-0.03	0.56	0.42	0.86	2.93
Bærum	Mean	0.83	-1.80	0.15	0.35	1.31	0.16	-0.87	0.44	0.33	0.89	1.14
	$q_{0.05}$	0.67	-2.92	-0.01	0.18	0.87	-0.40	-1.31	0.18	0.20	0.58	0.70
	$q_{0.95}$	0.95	-0.90	0.34	0.53	1.70	0.82	-0.56	0.88	0.50	1.30	1.79
Bergen	Mean	0.88	-0.61	-0.03	0.19	0.37	0.53	-0.52	0.15	0.13	0.41	1.23
	$q_{0.05}$	0.79	-1.64	-0.17	0.05	0.17	0.10	-0.74	0.09	0.07	0.33	0.65
	$q_{0.95}$	0.95	0.30	0.12	0.33	0.57	1.10	-0.35	0.20	0.20	0.49	1.99

20 minutes per chain on a 2.80-GHz Intel Core i7 processor.

Through the posterior distributions of p, Table 6.4.2 indicates that $80 - 90\%$ of the observations are estimated to be related to the defined covariates. With respect to the shape parameter, 0 is contained in the 90% credibility interval for only 2 of the 3 municipalities. Hence, there is evidence that the tail behaviour is not Poisson for Bergen. The covariate effects for Bergen are generally lower than for Oslo and Bærum. Since Bergen exhibits higher precipitation levels than Oslo and Bærum, the buildings are presumably designed to withstand more severe rainfall events than Oslo. The posterior estimates further show that covariate effects are non-negative except for β_1. Hence, the risk induced by the accumulated rainfall R^{Σ} is mainly captured via δ_1. Collectively, this indicates that a increase in R^{Σ} results in more claims above 4 in Bergen but a reduction in the variability of these claims over 4. The municipalities of Oslo and Bærum exhibit similar covariate effects for R^{Σ} and ΔS^{Σ} which correlates with their spatial proximity. Further, the estimates for the non-weather related rate κ differ by a factor of 2 for Oslo and Bærum, which is consistent with the number of policies in Oslo being about twice that of Bærum. The large difference for β_3 posteriors between Oslo and Bærum is mainly driven by one large observation of 143 claims. Indeed, β_3 has posterior mean 0.75 and 0.81 for Oslo and Bærum, respectively, when leaving their highest response out; see Appendix D.4.

The behaviour of the probability distribution for $\widetilde{N} \mid (\widetilde{X}, \widetilde{N} > 0)$ is further investigated in Figure 6.4.2 which shows changes in the mean claim for each covariate at each municipality. In general, the probability for high number of claims increases with increasing values for each of the three covariates, with $R^{\max}$ being the main risk factor for high number of claims. Further,

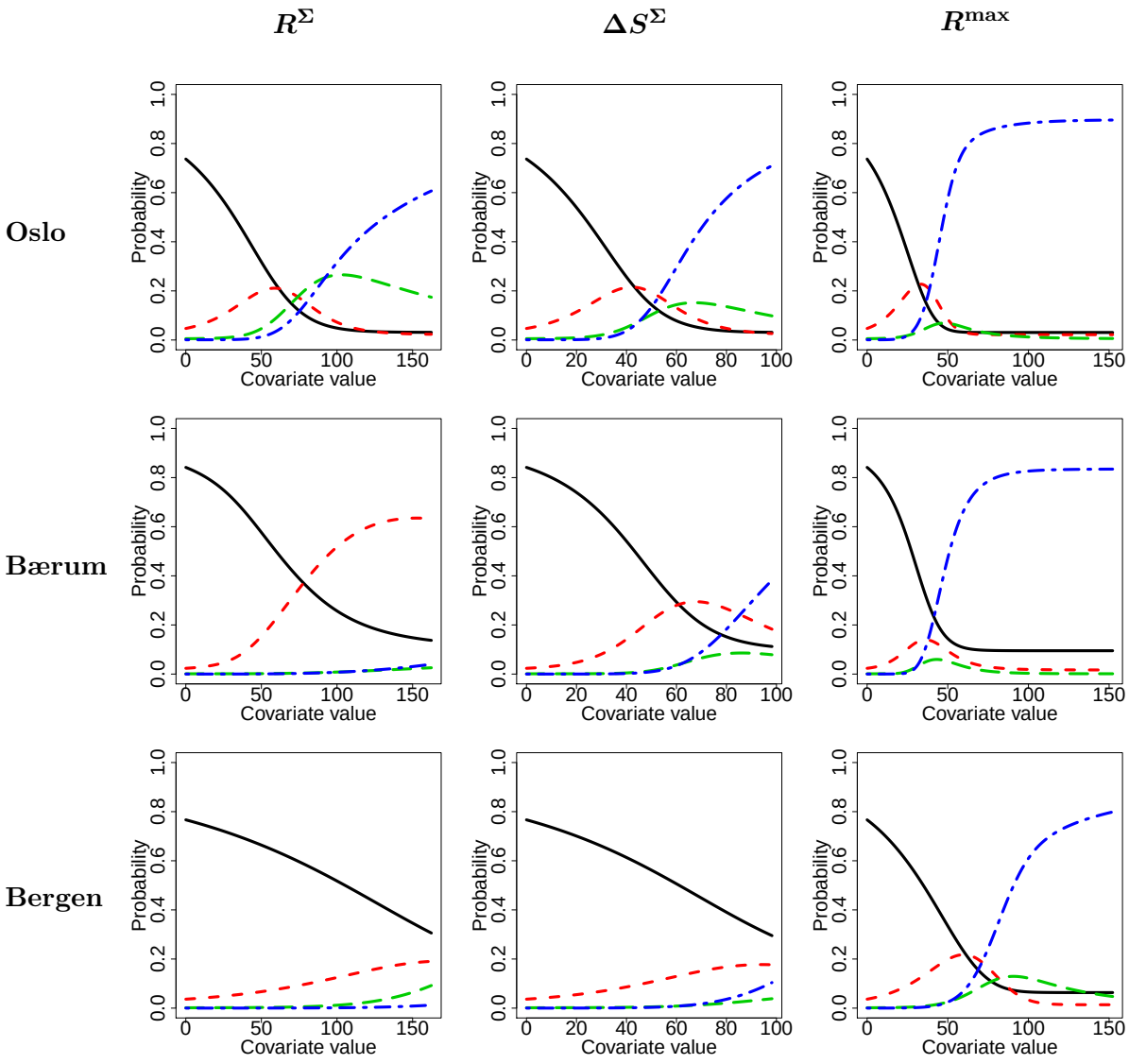


Figure 6.4.2: Probability for certain events of $\widetilde{N} \mid \left(\widetilde{X}, \widetilde{N} > 0\right)$ for Oslo, Bærum and Bergen varying with each of the covariates R^{Σ}, ΔS^{Σ} and $R^{\max}$. The events are $\widetilde{N} = 1$ (——), $\widetilde{N} = 3$ (- - - -), $\widetilde{N} = 5$ (– – –) and $\widetilde{N} > 6$ (– · –). In the first column, the probability is considered with respect to R^{Σ} while the remaining covariates are fixed at their minimum value. Equivalently, the second column and third column consider ΔS^{Σ} and $R^{\max}$, respectively.

the risk for very high number of claims increases more strongly for Oslo and Bærum than for Bergen. For instance, a covariate value of $R^{\max} = 50$ results in a probability of 60% for observing more than 6 claims in Oslo while it is only about 10% in Bergen. These findings are consistent with previous arguments that properties in Bergen are likely to be designed to withstand higher precipitation levels than in Oslo.

The fit of the estimated model is verified separately for observations below and above the threshold. For observations $\widetilde{N} \leq u$, the estimated and empirical frequencies are compared

Table 6.4.3: Posterior mean and empirical frequencies both times $\times 10^2$ for the number of claims between 1 and 4 for different rainfall settings for Oslo, Bærum and Bergen. For the empirical frequency, central 95% confidence intervals are given in parentheses. The rainfall settings are (1) $R^{\max} = 0$, (2) $0 < R^{\max} \leq q_{0.5}\,(R^{\max}|R^{\max} > 0)$ and (3) $R^{\max} > q_{0.5}\,(R^{\max}|R^{\max} > 0)$.

$\widetilde{N}$	$R^{\max}$	Oslo		Bærum		Bergen	
		estimated	empirical	estimated	empirical	estimated	empirical
1	(1)	72	75 (72,79)	83	84 (80,87)	77	79 (74,83)
	(2)	69	74 (69,80)	81	83 (78,90)	74	76 (72,81)
	(3)	40	34 (28,41)	49	47 (39,57)	44	45 (40,51)
2	(1)	20	18 (14,21)	13	13 (9,17)	19	16 (12,21)
	(2)	22	17 (12,23)	15	14 (9,21)	21	18 (13,22)
	(3)	25	24 (17,30)	25	22 (14,31)	26	24 (18,30)
3	(1)	5	5 (1, 8)			4	4 (0,9)
	(2)	6	7 (2,14)			4	3 (0,8)
	(3)	14	12 (6,19)			13	13 (7,19)
4	(1)	2	2 (0, 6)			1	0 (0,5)
	(2)	2	0 (0, 6)			1	2 (0,6)
	(3)	9	14 (7,21)			7	8 (2,13)
$> u_k$	(1)	1	0 (0, 4)	3	4 (0,8)	0	0 (0,4)
	(2)	1	1 (0, 7)	4	2 (0,9)	0	1 (0,5)
	(3)	13	16 (10,23)	25	31 (23,40)	10	11 (6,17)

in Table 6.4.3. In order to examine the performance more thoroughly, observations are split into three subsets with respect to $R^{\max}$. For instance, the posterior frequency for n claims, conditional on $R^{\max} = 0$, is given as

$$\mathbb{P}\left(\widetilde{N} = n \mid R^{\max} = 0, \widetilde{N} > 0\right) = \int \mathbb{P}\left(\widetilde{N} = n \mid R^{\max} = 0, \widetilde{N} > 0, \widetilde{\mathbf{x}}\right) \pi\left(\widetilde{\mathbf{x}} \mid R^{\max} = 0, \widetilde{N} > 0\right) d\widetilde{\mathbf{x}}.$$

The frequency for each posterior sample is derived by Monte Carlo integration via sampling from the empirical density $\pi\left(\widetilde{\mathbf{x}} \mid R^{\max} = 0, \widetilde{N} > 0\right)$. In particular, the sampled covariate values for $R^{\max}$ and R^{Σ} are from the same cluster periods while ΔS^{Σ} can be from any cluster period. The joint sampling of the rainfall covariates is required since the joint occurrence of $R^{\max} = 0$ and $R^{\Sigma} > 0$ is impossible, and hence high dependence exists. In contrast, snow-melt is approximately independent of the rainfall covariates. Confidence intervals are obtained by considering observations as realizations of a multinomial distribution with 5 possible outcomes for Oslo and Bergen and 3 for Bærum. Table 6.4.3 illustrates that the model-based estimated frequency for $\widetilde{N}$ lies within the central empirical 95% confidence interval in all cases. The model fit for the tails is verified via the posterior QQ plots provided in Column 1 of Figure 6.4.3. These are generated

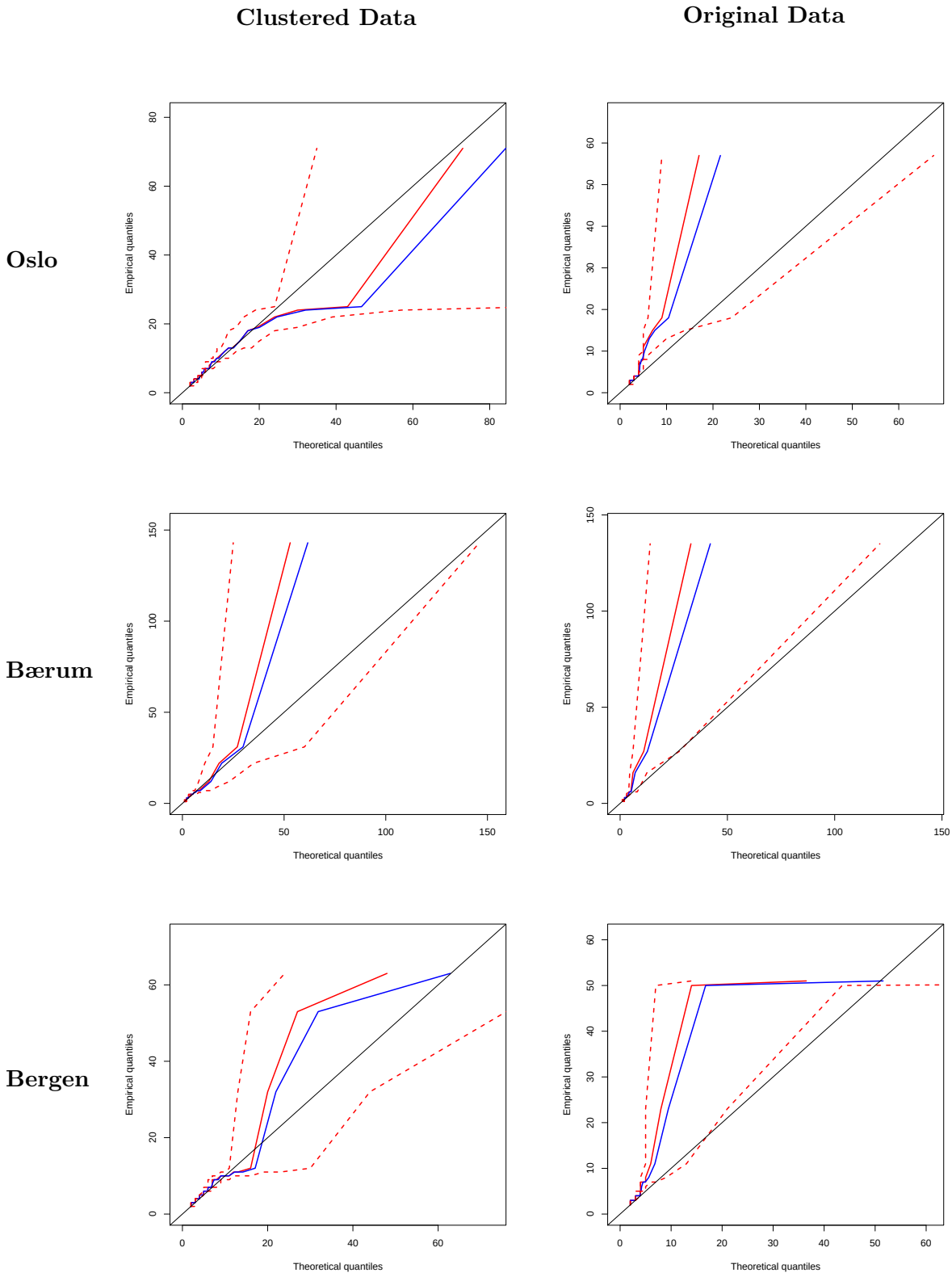


Figure 6.4.3: Posterior Quantile-Quantile for Oslo, Bærum and Bergen obtained by the full model. Column 1 provides the results for the clustered data while Column 2 considers the original daily data. The lines in each plot represent (——) Posterior mean, (——) Posterior median and (- - - -) Central 95% posterior interval.

Table 6.4.4: Average Bayesian Information Criterion (BIC) and Deviance Information criterion (DIC) for several competing models considering the distribution of $\widetilde{N} \mid (\widetilde{\mathbf{X}}, \widetilde{N} > 0)$ for Oslo, Bærum and Bergen. The best model fit for each municipality is highlighted.

Model	City	$\overline{\text{BIC}}$	DIC
Poisson	Oslo	2158	4.06
	Bærum	1079	3.96
	Bergen	2005	**3.92**
Poisson-Mixture	Oslo	2137	5.74
	Bærum	963	6.21
	Bergen	1977	5.63
Poison-IGPD	Oslo	2088	8.26
	Bærum	939	8.69
	Bergen	1937	8.75
Poisson-IGPD-Mixture	Oslo	**1779**	**3.18**
	Bærum	**596**	**−0.62**
	Bergen	**1632**	4.72

by deriving an individual QQ plot for each posterior sample and then deriving the mean and quantiles over this set of QQ plots. The plots indicate that a good model fit is obtained for each municipality as the diagonal lies within the 95% credibility interval. A slightly poor fit is found for Oslo around 20 claims which is due to the occurrence of three claim periods with 22-25 claims while there exists two with 16-21 claims. For Bærum, the highest observation is not fitted prefectly due to it being by far the highest observation over the 10-year period. However, it is still consistent with our model when uncertainty is accounted for.

Finally, we consider whether a similar performance may have been achieved with a different model. In order to verify the improvement obtained via clustering, the full model is fitted for the daily data with a lower threshold of $u = 3$ for Oslo and Bergen and u being unchanged for Bærum. The modification of the threshold is required since the frequency of higher number of claims is lower in the daily data than in the clustered data. For instance, the daily data for Oslo only contain 41 days with $N > 3$ while the clustering algorithm results in a total of 86 periods with $\widetilde{N} > 3$. Column 2 in Figure 6.4.3 shows a much worse model fit for the daily data, in particular for the medium to large claim numbers. Next, we compare the full model to three less-complex models for $\widetilde{N} \mid (\widetilde{X}, \widetilde{N} > 0)$: zero-truncated Poisson as in (6.1.1), Poisson-mixture without the extremal mixture model for $\widetilde{Y}$ and an extremal mixture model without the component $\widetilde{Z}$. Table 6.4.4 gives the Bayesian Information Criterion (BIC) (Schwarz, 1978) averaged over all posterior samples and results indicate that the full model performs better than the competing models.

The municipalities are similar in showing evidence that the additional flexibility offered by both our mixture and tail modelling components leads to substantial improvements. This conclusion is largely supported using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). In conclusion, both the full model and the clustering of the data are beneficial and provide the best model fit within subclasses of the model formulation.

6.5 Geographical Dependence and Prediction of Extremes

Section 6.4 illustrates that our approaches in Sections 6.2 and 6.3 lead to a better model fit than previous models. However, the statistical framework is based upon the assumption that claims in different municipalities are independent, conditional on the weather covariates. We investigate to which extent the fitted model captures the geographical dependence by considering the adjacent municipalities of Oslo and Bærum. Furthermore, the observation $\widetilde{N} = 143$ for Bærum is the highest number of claims reported over the ten year period. It is of interest to predict the frequency of such extreme events for Oslo, Bærum and Bergen irrespective of weather covariates. Here, we estimate the probability of $\widetilde{N}$ exceeding a large level v. Section 6.5.1 examines the spatial dependence while Section 6.5.2 derives $\mathbb{P}\left(\widetilde{N} > v\right)$.

6.5.1 Geographical Dependence

Interest lies in the spatial dependence of the clustered claims for different municipalities, in particular, for the adjacent municipalities of Bærum and Oslo. Figure 6.5.1 shows that high claim numbers tend to be observed for both municipalities over the same cluster period. Here, we examine to which extent the derived weather covariates capture this dependence. Since the cluster periods are not identical for both municipalities, some simplification is required. From the site-specific sets of cluster start days, the subset of days on which clusters are identified to have started at both sites is found. From this new set of start days, the time periods between consecutive start dates are then examined, and the cluster period with the largest number of claims during this period is identified separately for each site. While this approach appears very restrictive, only 10% of the cluster periods for Bærum are discarded and most of these observe zero claims. Furthermore, we discard samples for which $\widetilde{N}$ is equal to zero for Oslo or Bærum since our model considers cluster periods with a positive number of claims only. About 60% of the cluster periods for Bærum with a positive number of claims remain for analysis. This

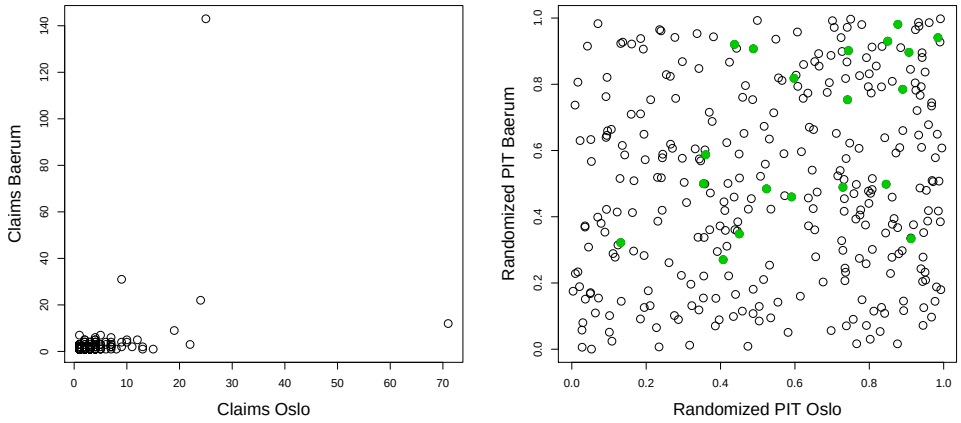


Figure 6.5.1: Plots of simultaneous clustered claims for the municipalities of Oslo and Bærum (left panel) and of the randomized probability integral transformed samples using the estimated conditional distributions of claims given weather at each municipality (right panel). Observations for which simultaneously more than 4 claims are observed for Oslo and more than 2 for Bærum are highlighted.

reduced data set is appropriate since the main interest lies in the dependence of the high claims.

In order to examine dependence of these observations, conditional on the weather covariates, the randomized probability integral transform (PIT), as described by Smith (1985) and Brockwell (2007) is applied. The randomized PIT approach is chosen since observations are discrete. Consider the observations $\{(\widetilde{n}_i, \widetilde{\mathbf{x}}_i) \ : \ i = 1, \ldots, m\}$. Conditional on the estimated model in Section 6.4 being correct, the set $\{v_i \ : \ i = 1 \ldots, m\}$ sampled via

$$v_i \sim \text{Uniform}\left[\mathbb{P}\left(\widetilde{N} \leq \widetilde{n}_i - 1 \mid \widetilde{\mathbf{x}}_i, \widetilde{N} > 0\right), \mathbb{P}\left(\widetilde{N} \leq \widetilde{n}_i \mid \widetilde{\mathbf{x}}_i, \widetilde{N} > 0\right)\right] \tag{6.5.1}$$

is uniformly distributed. Here, the probabilities in expression (6.5.1) are set to the corresponding average posterior probability. Consequently, the two data sets for Oslo and Bærum are transformed according to (6.5.1) and the resulting sets are denoted by $\mathbf{v}^O$ and $\mathbf{v}^B$ for Oslo and Bærum, respectively. If there is no claim dependence for the two municipalities, conditional on the weather covariates, the point process $\left\{\left(\mathbf{v}_i^O, \mathbf{v}_i^B\right) \ : \ i = 1, \ldots, V\right\}$ is uniform. This approach verifies the model fit as claims in different municipalities are only related through similar weather.

Dependence of the claims for the two municipalities, given the weather covariates, is hence examined by plotting $\mathbf{v}^O$ versus $\mathbf{v}^B$. Figure 6.5.1 shows that the points are relatively uniformly distributed on the unit square and hence independent. With respect to the largest claim num-

bers, the associated points are rather uniformly concentrated in the upper right corner of the unit square. This is due to the non-weather related component $\widetilde{Z}$ in the modelling framework which induces a lower bound on $\mathbb{P}\left(\widetilde{N} \leq \widetilde{n}_i \mid \widetilde{\mathbf{x}}_i, \widetilde{N} > 0\right)$ for large $\widetilde{n}_i$, irrespective of the covariate values.

6.5.2 Probability of Large Claims

The probability $\mathbb{P}\left(\widetilde{N} > v\right)$ for $v > u$ is approximated by a series of estimated probabilities. In the first step, $\mathbb{P}\left(\widetilde{N} > v\right)$ is split into the distribution of the positive claim numbers $\mathbb{P}\left(\widetilde{N} > v \mid \widetilde{N} > 0\right)$ and one empirical part for claim occurrences $\mathbb{P}\left(\widetilde{N} > 0\right)$. The former is then derived via the modelling framework considered in the previous sections. Formally,

$$\begin{aligned}
\mathbb{P}\left(\widetilde{N} > v\right) &= \mathbb{P}\left(\widetilde{N} > v \mid \widetilde{N} > 0\right) \mathbb{P}\left(\widetilde{N} > 0\right) \\
&= \left[p\, \mathbb{P}\left(\widetilde{Y} > v\right) + (1-p)\, \mathbb{P}\left(\widetilde{Z} > v\right)\right] \mathbb{P}\left(\widetilde{N} > 0\right) \\
&\approx \left\{ \frac{1}{J} \sum_{j=1}^{J} \left[p^{(j)} \mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}\right) + \left(1 - p^{(j)}\right)\, \mathbb{P}\left(\widetilde{Z} > v \mid \boldsymbol{\theta}^{(j)}\right)\right] \right\} \mathbb{P}\left(\widetilde{N} > 0\right),
\end{aligned} \tag{6.5.2}$$

where $\boldsymbol{\theta}^{(j)}$ refers to the jth of J samples obtained via the MCMC algorithm in Section 6.4. The probability $\mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}\right)$ captures the dependence on the weather and more work is required to obtain it, see (6.5.3) below. Conversely, the remaining components can be derived quite straightforwardly. Equivalently to the hurdle component of the BPH model in expression (6.1.1), the random variable $\widetilde{N} > 0$ is assumed to be Bernoulli distributed. Column 4 in Table 6.5.1 provides the posterior mean and central 90% credibility intervals of this probability obtained via Bayesian inference with an uniform prior. Further, the probability $\mathbb{P}\left(\widetilde{Z} > v \mid \boldsymbol{\theta}^{(j)}\right)$ is independent of the covariates $\widetilde{X}$ and hence directly accessible.

For the covariate-driven component $\widetilde{Y}$, additional steps are necessary since we require the marginal $\mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}\right)$. Formally, the probability $\mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}\right)$ can be expressed via

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}\right) = \mathbb{P}\left(\widetilde{Y} > v \mid \boldsymbol{\theta}^{(j)}, \widetilde{Y} > u\right) \times \mathbb{P}\left(\widetilde{Y} > u \mid \boldsymbol{\theta}^{(j)}\right) \\
&= \int \mathbb{P}\left(\widetilde{Y} > v \mid \widetilde{\mathbf{x}}, \boldsymbol{\theta}^{(j)}, \widetilde{Y} > u\right) \pi\left(\widetilde{\mathbf{x}} \mid \widetilde{Y} > u\right) \mathrm{d}\widetilde{\mathbf{x}} \times \int \mathbb{P}\left(\widetilde{Y} > u \mid \widetilde{\mathbf{x}}, \boldsymbol{\theta}^{(j)}\right) \pi\left(\widetilde{\mathbf{x}}\right) \mathrm{d}\widetilde{\mathbf{x}}.
\end{aligned} \tag{6.5.3}$$

In order to evaluate these integrals, the probability density functions $\pi\left(\widetilde{\mathbf{x}}\right)$ and $\pi\left(\widetilde{\mathbf{x}} \mid \widetilde{Y} > u\right)$ are

Table 6.5.1: Estimated scale ν and shape η for the distribution $R^{\max}|\widetilde{Y} > u \sim \text{GPD}(\nu, \eta)$ and standard errors. Column 3 provides the posterior mean and central 90% credibility intervals of the probability that $\widetilde{N}$ exceeds 100 conditional on $\widetilde{N} > 0$. Column 4 gives the empirical maximum likelihood estimate and central 90% confidence intervals of the frequency for $\widetilde{N} > 0$.

Municipality	ν	η	$\mathbb{P}\left(\widetilde{N} > 100 \mid \widetilde{N} > 0\right)$	$\mathbb{P}\left(\widetilde{N} > 0\right)$
Oslo	37.6	-0.48	0.00029	0.391
	(6.7)	(0.11)	$\left(6.3 \times 10^{-7}, 0.00096\right)$	(0.376, 0.407)
Bærum	27.73	-0.47	0.00044	0.209
	(4.8)	(0.11)	$\left(5.1 \times 10^{-5}, 0.00122\right)$	(0.197, 0.222)
Bergen	67.34	-0.40	0.00052	0.393
	(12.0)	(0.10)	$\left(4.8 \times 10^{-5}, 0.00148\right)$	(0.377, 0.409)

required. The former is approximated by the empirical density function since a sufficient number of observations is available. Due to low amount of observations greater than u, $\pi\left(\widetilde{\mathbf{x}} \mid \widetilde{Y} > u\right)$ is estimated parametrically and some simplifications are required. As our conditional model for claims given weather shows that extreme claims are strongly associated with extreme $R^{\max}$ only, $\pi\left(\widetilde{\mathbf{x}} \mid \widetilde{Y} > u\right)$ is approximated as $\pi\left(R^{\max} \mid \widetilde{Y} > u\right)$ and the remaining values are set to their average observed values, conditional on the number of claims exceeding u. Larger $R^{\max}$ than those observed need to be accounted for as they are critical when considering extreme $\widetilde{Y}$. The one-dimensional density $\pi\left(R^{\max} \mid \widetilde{Y} > u\right)$ is estimated via an extremal mixture model.

In order to select a threshold, the mean residual life plots in Figure 6.5.2 are considered. A threshold of $u_R = 0.1$, which corresponds to smallest positive amount of rainfall, seems suitable and, hence, we fit a GPD with point mass at the minimum value since a few claims exceeding u occur for $R^{\max} = 0$. The distribution is fitted separately for each city via maximum likelihood and estimates and standard errors for the scale parameter ν and shape parameter η for the GPD as in expression (6.2.3) are provided in Table 6.5.1. The estimated shape parameter η is negative for all three municipalities, that is, the associated GPD is short-tailed with an upper end point. Figure 6.5.2 shows that the fit is good for Oslo and Bærum while being slightly off for Bergen.

The case $v = 100$, $\mathbb{P}\left(\widetilde{N} > 100 \mid \widetilde{N} > 0\right)$, is then derived using Monte Carlo integration for the expressions in (6.5.2) and (6.5.3). Table 6.5.1 shows a varying behaviour for the three municipalities for $\mathbb{P}\left(\widetilde{N} > 0\right)$ and $\mathbb{P}\left(\widetilde{N} > 100 \mid \widetilde{N} > 0\right)$. The results indicate that about 1 in 5000 events for Bergen will cause more than 100 claims. Considering that about 2,500 events were observed over a 10 year horizon, that corresponds to one occurrence every 20 years on

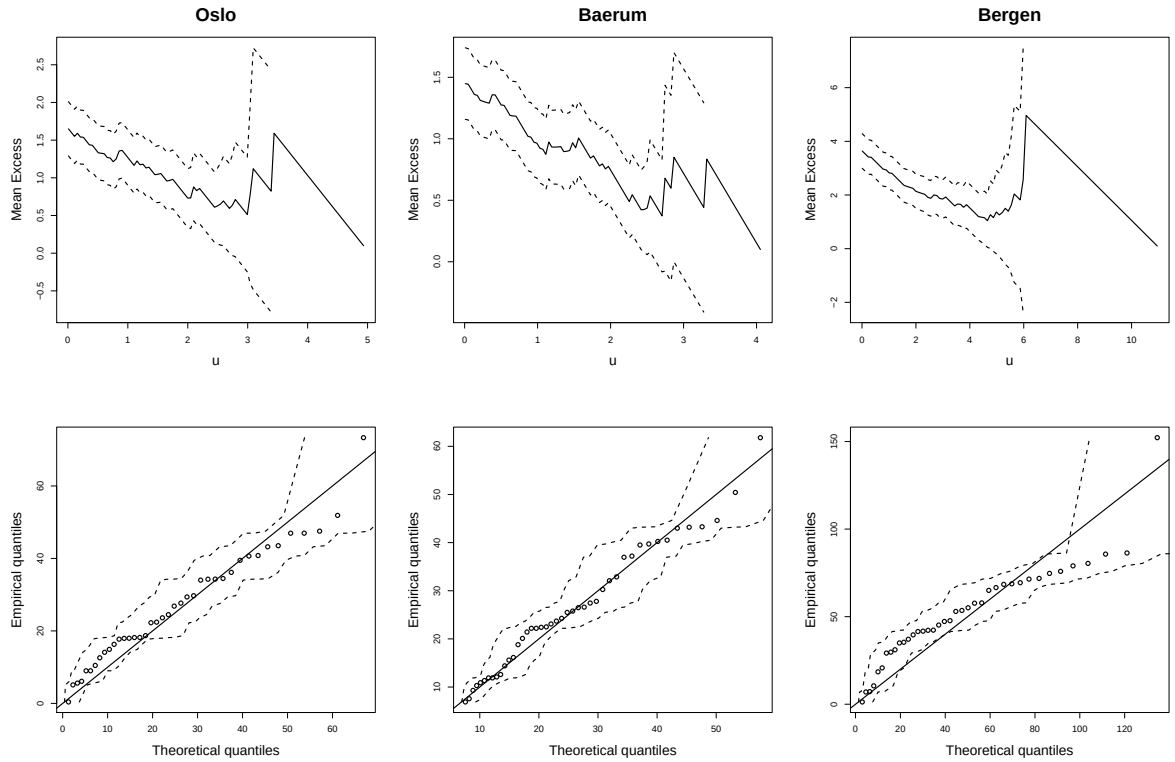


Figure 6.5.2: Mean residual life plots (Row 1) and Quantile-Quantile plots with central 95% confidence intervals of the fitted GPD distribution (Row 2) for $\pi\left(R^{\max} \mid \widetilde{Y} > u\right)$ for the municipalities of Oslo, Bærum and Bergen.

average. Furthermore, the same approach implies that such an event happens every 30-40 years for Oslo and Bærum. Hence, the observation of 143 claims for Bærum is a very rare event.

6.6 Discussion

We extended the modelling framework by Haug et al. (2011) and Scheel et al. (2013) in order to improve the model fit for higher number of claims. Additional information was gained by analyzing the spatial and temporal patterns with respect to snow-melt and precipitation. A temporal cluster algorithm, based solely on the observed weather covariates, was introduced in order to reduce the effects of potential lags in the recording process and to account for weather events which affect the claim dynamics on consecutive days. The original daily data were then adapted to the respective cluster periods and one covariate was tuned to maximize its relevance to large claims.

A mixture model with an extremal mixture component was applied to model the number of claims over the cluster periods. Results have shown good performance for lower as well

as higher numbers of claims. Furthermore, the spatial dependence between claims in different municipalities appears to be accounted for conditional on the derived weather covariates. Finally, the estimated model also facilitates the investigation of the probability for very rare events.

The derived model can also be applied to assess the impact of climate change. Haug et al. (2011) use the daily data and perform an effect study, subject to the insurance portfolio of properties of future periods being close in value and quality to the one of the model fitting period. Their results indicate an increase in the claim frequency for all municipalities. In order to perform a similar study with our new model, it is necessary to simulate weather observations for cluster periods rather than single days.

There are various way to extend the model presented in this chapter. Firstly in the model fitting of the extremal mixture model for claims, the distribution can be restricted to a unimodal form by excluding parameter settings which induce

$$\mathbb{P}\left(\widetilde{Y} = \lfloor u \rfloor - 1 \mid \widetilde{\mathbf{X}}\right) > \mathbb{P}\left(\widetilde{Y} = \lfloor u \rfloor \mid \widetilde{\mathbf{X}}\right) < \mathbb{P}\left(\widetilde{Y} = \lfloor u \rfloor + 1 \mid \widetilde{\mathbf{X}}\right).$$

This set of inequalities imposes additional constraints on the parameters λ, σ_u and ξ. This chapter focused on the periods with $\widetilde{N} > 0$ but there is interest for all periods. We considered a Poisson-IGPD mixture with the same parameter values as for the zero-truncated Poisson-IGPD mixture in Section 4 and found that the model underpredicts the frequency of periods with zero claims $\widetilde{N} = 0$. Hence, the model could be extended via a hurdle component as in the BPH. Furthermore, Figure 6.4.2 shows that the event $\widetilde{N} = 1$ has a probability of about 0.10 even for very high values of $R^{\max}$ due to the non-weather related mixture component. One may argue that such predictions are unrealistic since extreme precipitation levels over a day should lead to large damages, regardless of their intensity. Therefore, the mixture probability p could be modelled as a function of the covariate $R^{\max}$.

Further research can also be undertaken from a spatial perspective. Spatial dependence of the parameters of the conditional distribution of $\widetilde{N} \mid \left(\widetilde{X}, \widetilde{N} > 0\right)$ may be introduced to allow for a better model fit similarly to Scheel et al. (2013). For instance, the threshold $u = 2$ for Bærum may be too low for the extremal mixture model but there are not enough observations to raise it to $u = 3$. Additional information may be borrowed from the adjacent municipalities, in particular Oslo, in order to achieve this. Spatial dependence could be modelled via a conditional autoregressive prior (Besag, 1974; Besag et al., 1991) on $(\beta_1, \beta_2, \beta_3)$ in (6.4.3).

Chapter 7

Discussion

7.1 Summary

This thesis explored the association between property insurance claims and weather events. The relationship is of general interest as, for instance, insurance companies have to set premiums. To derive adequate models, daily insurance and weather data for all Norwegian municipalities were considered. An exploratory data analysis indicated that the degree of vulnerability with respect to the weather covariates varies spatially and, additionally, a higher average claim risk was found for cities, as compared to rural municipalities. Since the existing models exhibit limitations, in particular for days with high claim numbers, novel methodologies in spatial statistics, monotonic regression and extreme value theory were introduced. In particular, this thesis contributed approaches to model the dependence of monotonic functions and to perform extreme value analysis of discrete random variables. These approaches were designed to address certain properties of the claim dynamics and increased the flexibility of the statistical models. Results showed that the model fit improved which, in conclusion, provided a better understanding of the claim processes. In the following, the contributions of each chapter are outlined individually.

Chapter 3 defined a Bayesian hierarchical modelling framework which allowed for a spatially varying relationship between the number of claims and the weather covariates while assuming a similar claim process for adjacent municipalities. A comparative study was performed to assess the difference between two data models and partly differing covariates. Estimates indicated a spatial variation in the claim process, for instance, a certain rainfall amount affects the claim dynamics stronger for inland municipalities than for coastal areas. As for the existing models, the overall predictive performance was good but the higher number of claims were generally

underestimated. While the models performed similarly in terms of the considered performance measure, a difference was found with respect to the BIC. Specifically, a Bayesian Poisson hurdle model outperformed a Binomial model for densely populated municipalities. Consequently, results indicated that the performance measure by Scheel et al. (2013) is limited in terms of assessing model fit.

The Bayesian hierarchical model in Chapter 3 was then extended in Chapter 4 by substituting the linear predictor in the process model by a monotonic function, leading to the concept of Bayesian spatial monotonic multiple regression (BSMMR). Since the dependence modelling of monotonic functions has not been considered in the literature, a flexible prior distribution was introduced which allowed the sharing of statistical information across municipalities. Further, the hyperparameter in the prior distribution was estimated via cross-validation and Bayesian global optimization as the normalizing constant of the prior is intractable. A reversible jump MCMC algorithm, as proposed by Saarela and Arjas (2011), was implemented to obtain the function estimates. The simulation studies performed illustrated the benefits of the new methodology. Finally, the approach was also applied to a subset of the insurance and weather data and yielded to an improvement of the daily predictive performance.

Chapter 5 then introduced an alternative approach to BSMMR. The method is optimization-based and was motivated by the high computational cost of the BSMMR methodology. Dependence between functions was incorporated via the addition of penalty terms in the original optimization problem. Since the resulting optimization problem is convex, estimates can be obtained via a cyclic algorithm which updates one function while keeping the remaining ones fixed. Statistical information on the dependence structure was derived by treating the functions as initially independent and then examining the individual function estimates. Simulations show good results at a low computational cost. The performance partly depends on a selected interpolation routine as the functions are estimated at a finite set of points only while BSMMR estimates them over a continuous space.

Chapter 6 considered the application of extreme value models to improve the fit for higher numbers of claims. An analysis of the specific data for Oslo motivated the definition of a two-component mixture model; a covariate-dependent and a random distribution. Further, the tail of the covariate-dependent distribution was modelled via a discretized generalized Pareto distribution. Additional to the statistical model, a temporal clustering algorithm was introduced which aggregated periods of consecutive days based on the observed covariate values. The defined

cluster covariates, associated to the cluster periods, represented the amount of precipitation and snow-melt. One covariate was tuned to maximize its relevance to the highest claims. The combination of the clustering approach and the mixture model led to a large improvement in terms of the model fit for the municipalities of Oslo, Bærum and Bergen. Further, the estimated mixture distribution captured the spatial dependence of the claims and allowed the derivation of the probability of high claims in the future, conditional on no climate change.

7.2 Future Work and Possible Extensions

7.2.1 Combining the Different Models and Reducing Computational Time

The approaches in Chapters 4 and 6 have been applied separately to model the number of claims. However, these approaches can be combined in a Bayesian hierarchical model. The assumption of linearity in the mixture model in Chapter 6 may be violated as it does not account for any potential threshold effects. Therefore, the linear predictors for the model parameters may be replaced by monotonic ones to account for these potential effects. Such an approach can also be considered for the daily covariates and claim numbers.

Nevertheless, the assumption of linearity may be plausible for some of covariates. Both approaches to model dependence in a monotonic regression framework can be extended to a mixture of linear and monotonic functions. Such a model would reduce the computational time substantially as the BSMMR algorithm is very costly for higher dimensions. However, it is too computationally expensive to consider all possible combinations of linear and monotonic factors. Therefore, statistical tools are required in order to decide which covariates may be assumed to be linear. These guidelines would then have to be assessed via an exhaustive simulation study.

Additional to this approach, parallelization techniques may be used to reduce the computational cost of the BSMMR algorithm. The simplest approach may be to run several cross-validations for one proposed spatial smoothing parameter in parallel. For instance, the simulation studies in Chapter 4 performed 50 cross-validations for each proposal and these could be run independently on 50 processors. Alternatively, the reversible jump MCMC may be parallelized itself to some extent. For one update, the current implementation samples a proposal, updates the likelihood and computes the prior ratio before considering the next monotonic function. Since the prior is defined via pair-wise differences, some improvements are possible. Further, proposals for all regions may be sampled jointly at the beginning of the iteration step and the

likelihood ratio may be computed in parallel for each municipality.

By reducing the computational time, the statistical models may then be derived for a larger set of municipalities in a reasonable amount of time. The new methods for monotonic regression and extreme value analysis were applied to single or small sets of municipalities only but interest lies in the estimation of a claim model for all municipalities. Further, the mixture model in Chapter 6 may be extended by defining a spatial structure for the model parameters. Finally, the temporal variation in the number of claims was ignored in Chapter 6 and may also be incorporated.

7.2.2 Compound Poisson Distribution

This thesis assumed smaller claims to be Binomially distributed which implies independence of the claims within a municipality. Haug et al. (2011) consider the overdispersed Binomial distribution to achieve higher flexibility. Here, a compound Poisson distribution is discussed as an alternative. The number of claims is then the sum of Poisson random variables with the same rate parameter, and the number of components is assumed to be Poisson distributed too. Formally, the distribution of N is given as

$$\begin{aligned} N \mid (M = m, \theta) &\sim \text{Poisson}\,(m\theta + \theta) \\ M \mid \lambda &\sim \text{Poisson}(\lambda). \end{aligned} \tag{7.2.1}$$

From an applied perspective, $M + 1$ is the number of areas within the municipality for which claims may be observed and θ is the expected number of claims for each area. A simulation study with 200 samples and parameter values $\theta = 3$ and $\lambda = 8$ is performed to assess this approach. The parameters θ and λ are estimated via a Metropolis-within-Gibbs algorithm and every 500th sample is considered for analysis. Figure 7.2.1 illustrates the sampled Markov chains and the plots indicate a poor mixing and high correlation between the samples.

To reduce the dependence of the parameters, they are transformed based on the mean and variance. Using conditional probabilities, the mean of the compound Poisson distribution in expression (7.2.1) yields to

$$\begin{aligned} \mathbb{E}\,[N] &= \mathbb{E}\,[\mathbb{E}\,(N \mid M)] \\ &= \mathbb{E}\,[M\theta + \theta] \\ &= \theta(\lambda + 1). \end{aligned} \tag{7.2.2}$$

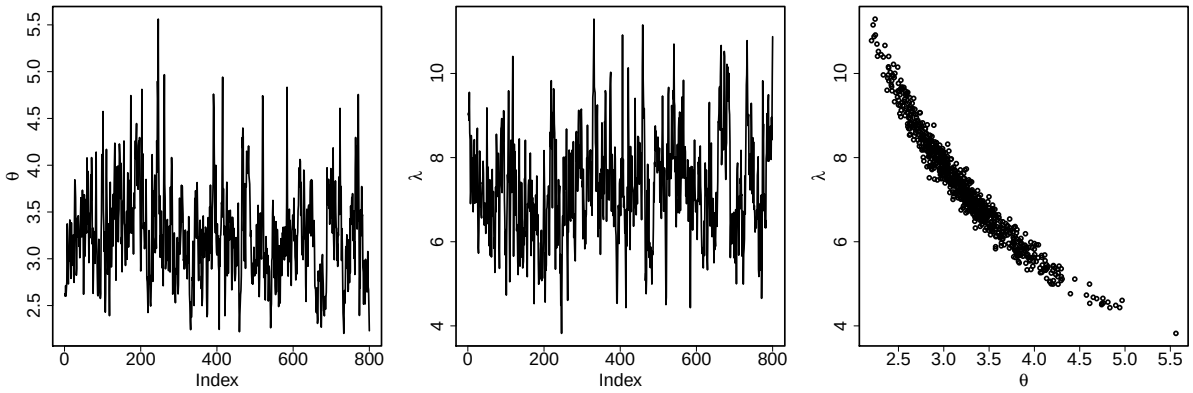


Figure 7.2.1: Sampled Markov chains for the parameters θ (Plot 1) and λ (Plot 2) of the compound Poisson distribution. Plot 3 illustrates the sampled pairs of parameters θ and λ.

Similarly, the variance of the distribution is given as

$$\begin{aligned}
\text{Var}\,[N] &= \mathbb{E}\left[N^2\right] - \mathbb{E}\,[N]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left(N^2 \mid M\right)\right] - \theta^2(\lambda+1)^2 \\
&= \mathbb{E}\left[\theta^2(M+1)^2 + \theta(M+1)\right] - \theta^2(\lambda+1)^2 \\
&= \theta^2\mathbb{E}\left[M^2 + 2M + 1\right] + \theta(\lambda+1) - \theta^2(\lambda+1)^2 \\
&= \theta^2\left(\lambda^2 + \lambda + 2\lambda + 1\right) + \theta(\lambda+1) - \theta^2(\lambda+1)^2 \\
&= \theta^2\lambda + \theta(\lambda+1).
\end{aligned} \tag{7.2.3}$$

Note, the variance of the defined compound Poisson distribution is greater than the mean unless $\lambda = 0$. The transformed parameters ϕ_1 and ϕ_2 are then derived from θ and λ, based on (7.2.2) and (7.2.3), via the transformation

$$(\theta, \lambda) \longmapsto \left(\ \theta(\lambda+1),\ \theta^2\lambda\ \right) = (\phi_1, \phi_2)\,. \tag{7.2.4}$$

Hence, θ and λ are conversely defined in terms of ϕ_1 and ϕ_2 as

$$\begin{aligned}
\theta &= \frac{\phi_1}{2} - \sqrt{\frac{\phi_1^2}{4} - \phi_2} \\
\lambda &= \phi_1 \left[\ \frac{\phi_1}{2} - \sqrt{\frac{\phi_1^2}{4} - \phi_2}\ \right]^{-1} - 1.
\end{aligned} \tag{7.2.5}$$

The estimates of ϕ_1 and ϕ_2 for the simulated data are provided in Figure 7.2.2. Results indicate an improved mixing and less dependence of the parameter estimates.

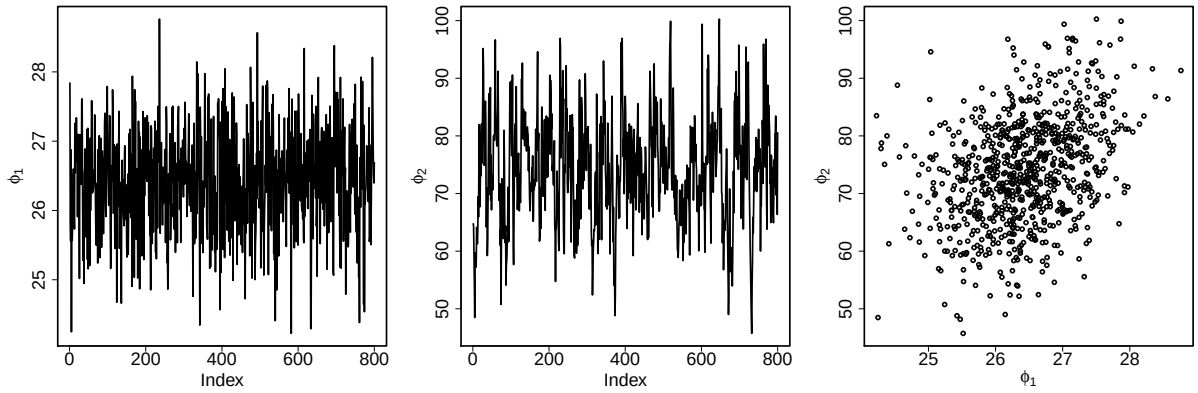


Figure 7.2.2: Sampled Markov chains for the reparametrized parameters ϕ_1 (Plot 1) and ϕ_2 (Plot 2) of the compound Poisson distribution. Plot 3 illustrates the sampled pairs of parameters ϕ_1 and ϕ_2.

In conclusion, the proposed transformed has improved the efficiency of the MCMC algorithm. The distribution of the number of claims N can then be defined via the compound Poisson (7.2.1). Further work is, however, required to define the covariate structure of the model parameters. Since both ϕ_1 and ϕ_2 are positive, a log-linear model could be specified.

7.2.3 Effect Study of Climate

The improved claim models developed using the approaches in this thesis should finally be applied to assess the impact of climate change. Similarly to Haug et al. (2011), it is of interest to predict the claim frequency for a scenario period, such as, 2071-2100. To obtain these claim frequencies, the covariates introduced in Chapter 6 have to be derived based upon general circulation models (GCM). This requires both downscaling and calibration in order to achieve future weather data from these climate models. Some research on future rainfall events in Norway has already been undertaken (Orskaug et al., 2011). However, the new covariates proposed here are defined over time periods which may correspond to more than one severe weather event and this may make the derivation of future covariate values more difficult.

Conditional on future weather covariates being derivable, the models in this thesis can then be applied in combination with a claim size distribution as formulated by Haug et al. (2011). In particular, the monetary losses over each period are assumed to be Gamma distributed as the sparse data will lead to a high uncertainty in the estimates for more complex statistical models. The results by Haug et al. (2011) indicate an increase in both the claim frequency and claim size in the future. However, their claim model does not allow differentiate between the emissions

models A2 and B2 as the corresponding approximate 95% confidence intervals overlap. This result is quite unsatisfactory as politics and economy should have more certainty in how far the climate change affects the society. Further, politics and science are interested in the differences between different emission models such that they have additional information for future decision making. Therefore, the improved claim models in this thesis may provide a better understanding by applying them in combination with climate modelling.

Appendix A

Supplementary Material Chapter 1

A.1 Temporal Variation in the Rainfall Levels

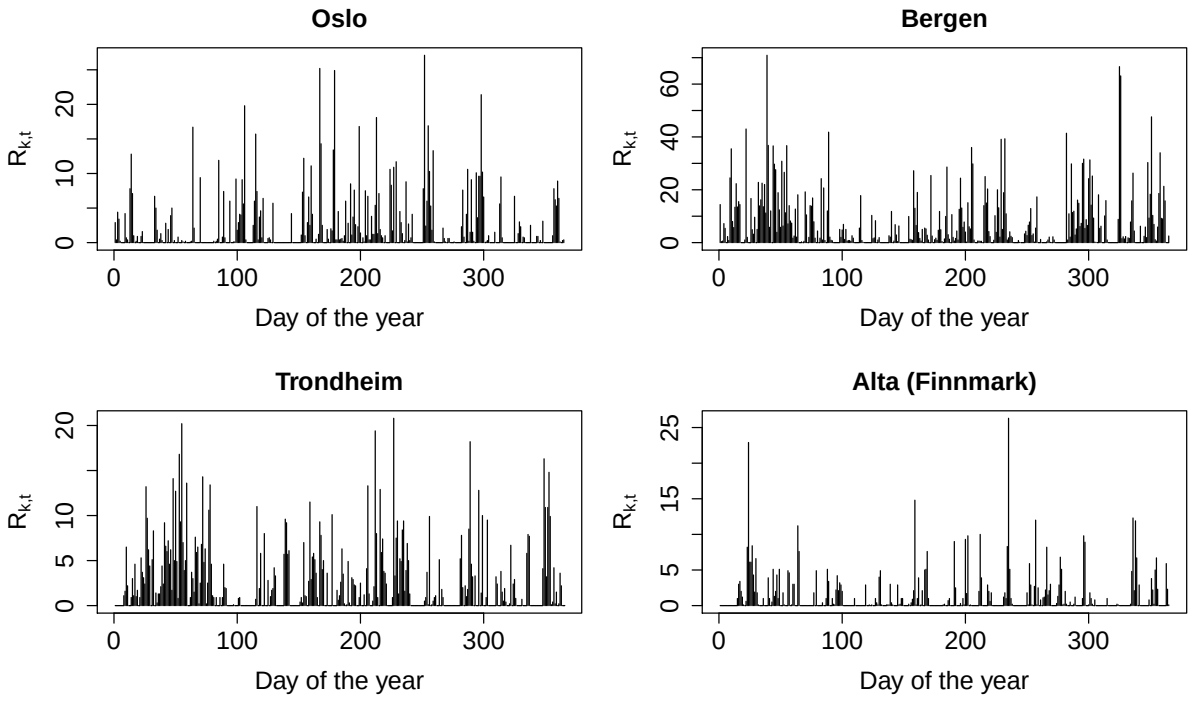


Figure A.1.1: Observations for $R_{k,t}$ for four municipalities across Norway for 1998.

A.2 Spatial Variation of the Difference in Snow-water Equivalent

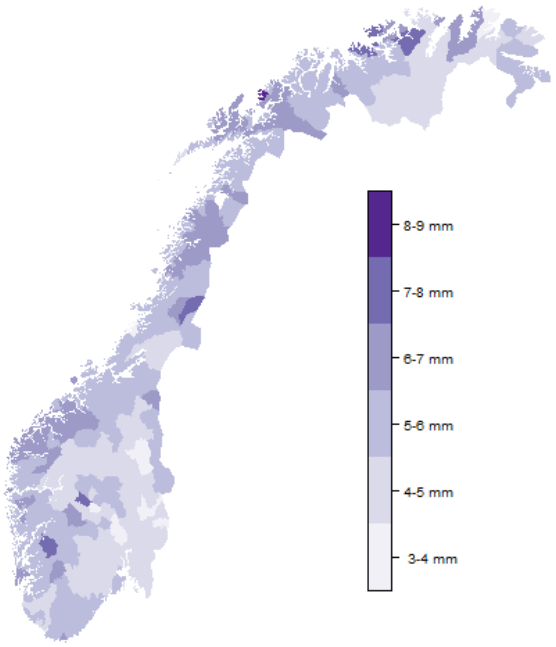


Figure A.2.1: Average positive difference in snow-water equivalent $\Delta S_{k,t} | \Delta S_{k,t} > 0$ between 1997 and 2006.

A.3 Correlation between Claims and Rainfall

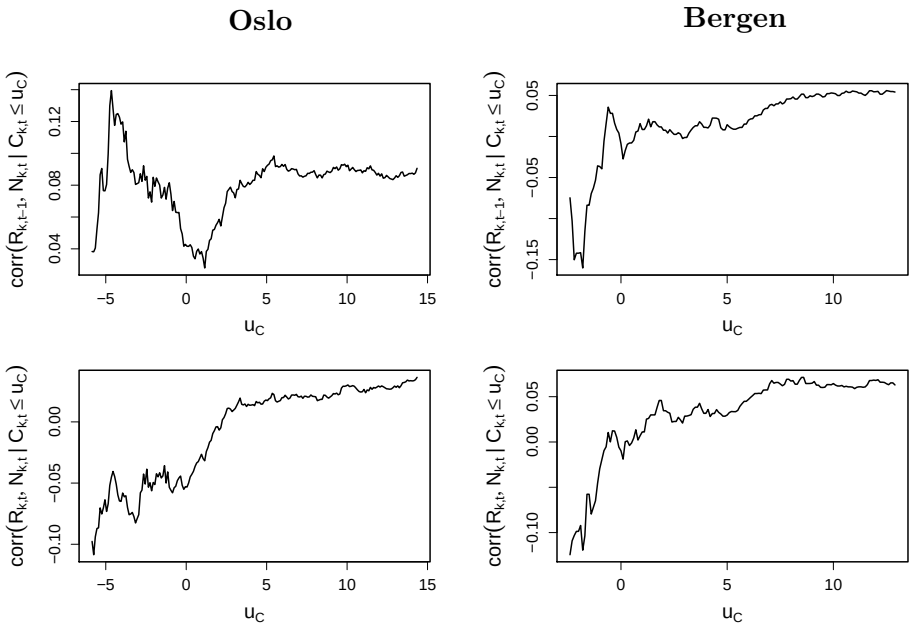


Figure A.3.1: Kendall's correlation coefficient between number of claims $N_{k,t}$ and amounts of precipitation $R_{k,t-1}$ and $R_{k,t}$ in dependency on $C_{k,t}$. The functional level corresponds to the correlation between $N_{k,t}$ and $R_{k,t-1}$ (Row 1), and $N_{k,t}$ and $R_{k,t}$ (Row 2) for Oslo (Column 1) and Bergen (Column 2) conditional on $C_{k,t}$ being smaller or equal u_C.

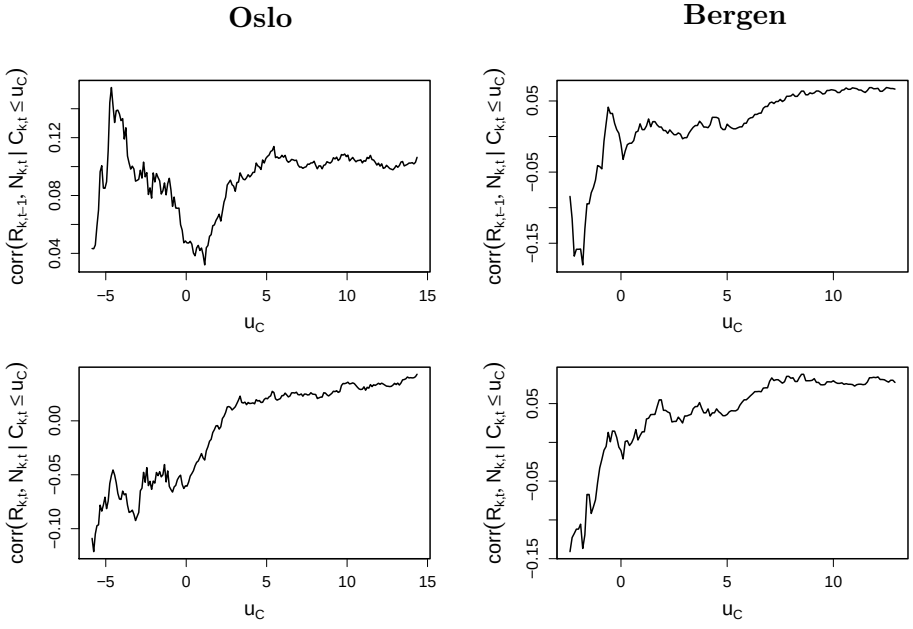


Figure A.3.2: Spearman's correlation coefficient between number of claims $N_{k,t}$ and amounts of precipitation $R_{k,t-1}$ and $R_{k,t}$ in dependency on $C_{k,t}$. The functional level corresponds to the correlation between $N_{k,t}$ and $R_{k,t-1}$ (Row 1), and $N_{k,t}$ and $R_{k,t}$ (Row 2) for Oslo (Column 1) and Bergen (Column 2) conditional on $C_{k,t}$ being smaller or equal u_C.

Appendix B

Supplementary Material Chapter 3

B.1 Trace plots for the sampled intercepts and covariate effects for Oslo and Bergen

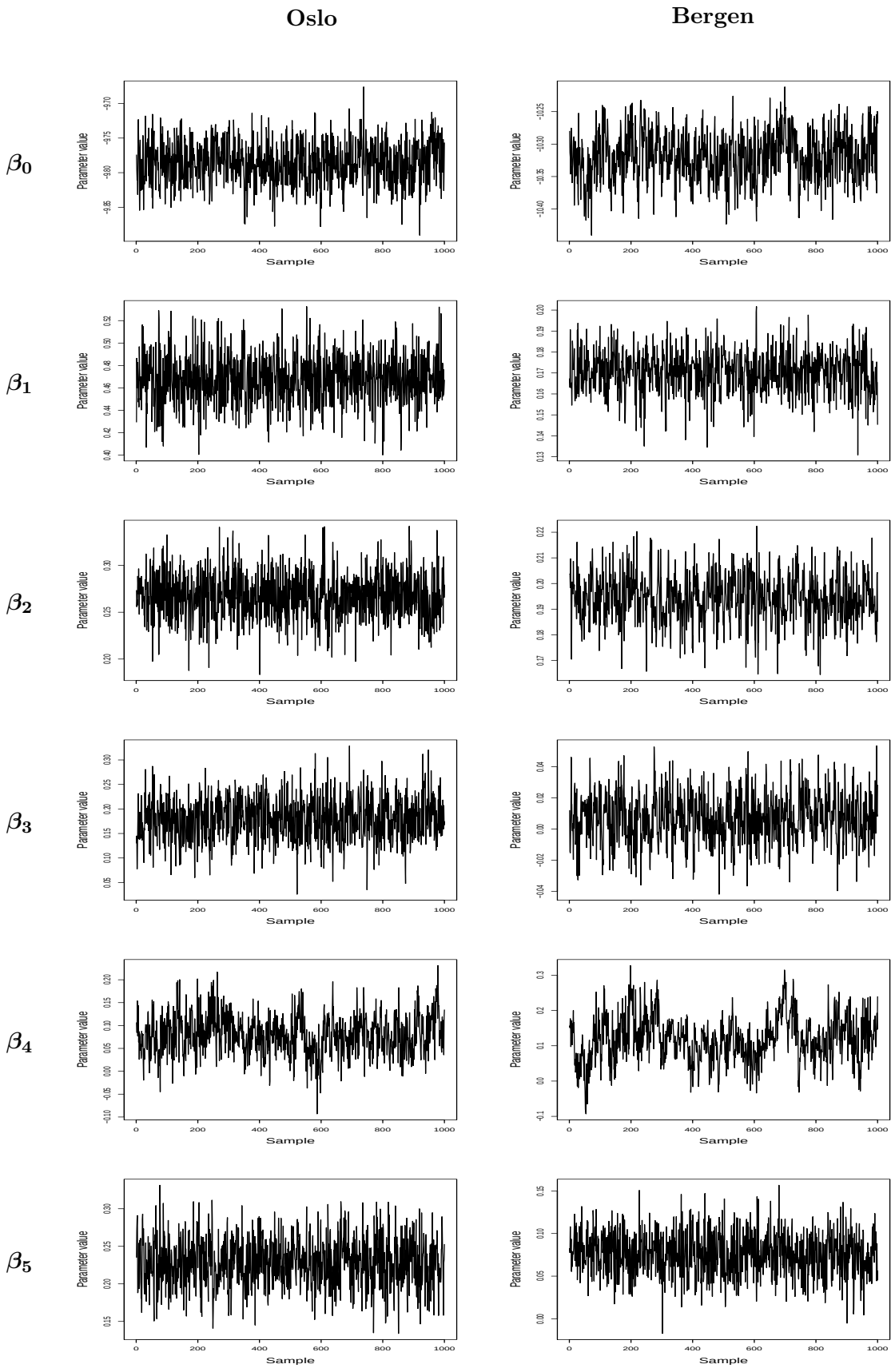


Figure B.1.1: Trace plots of the sampled realizations from the posterior distribution for the Binomial model and the original covariates for Oslo and Bergen.

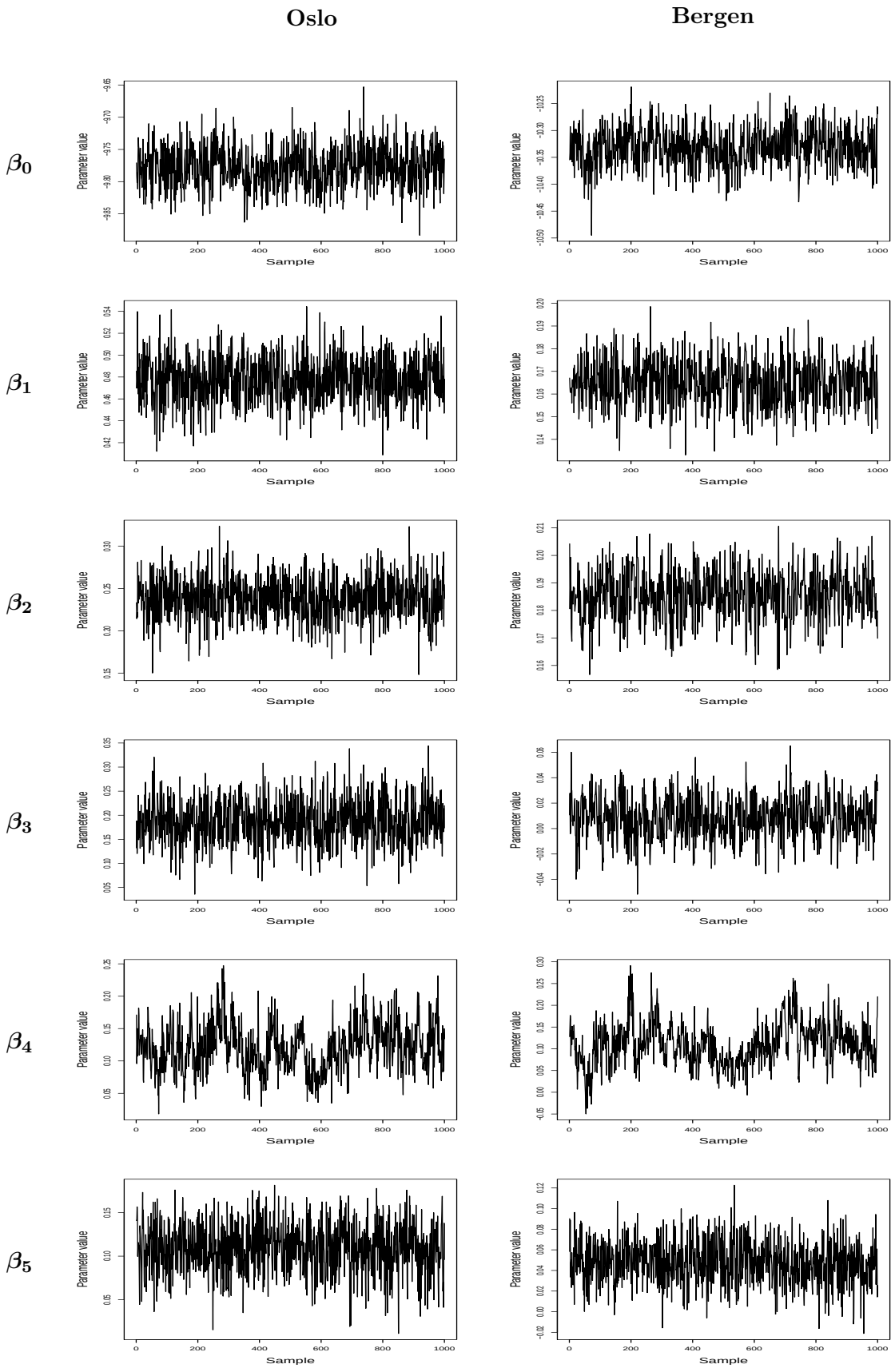


Figure B.1.2: Trace plots of the sampled realizations from the posterior distribution for the Binomial model and the proposed covariates for Oslo and Bergen.

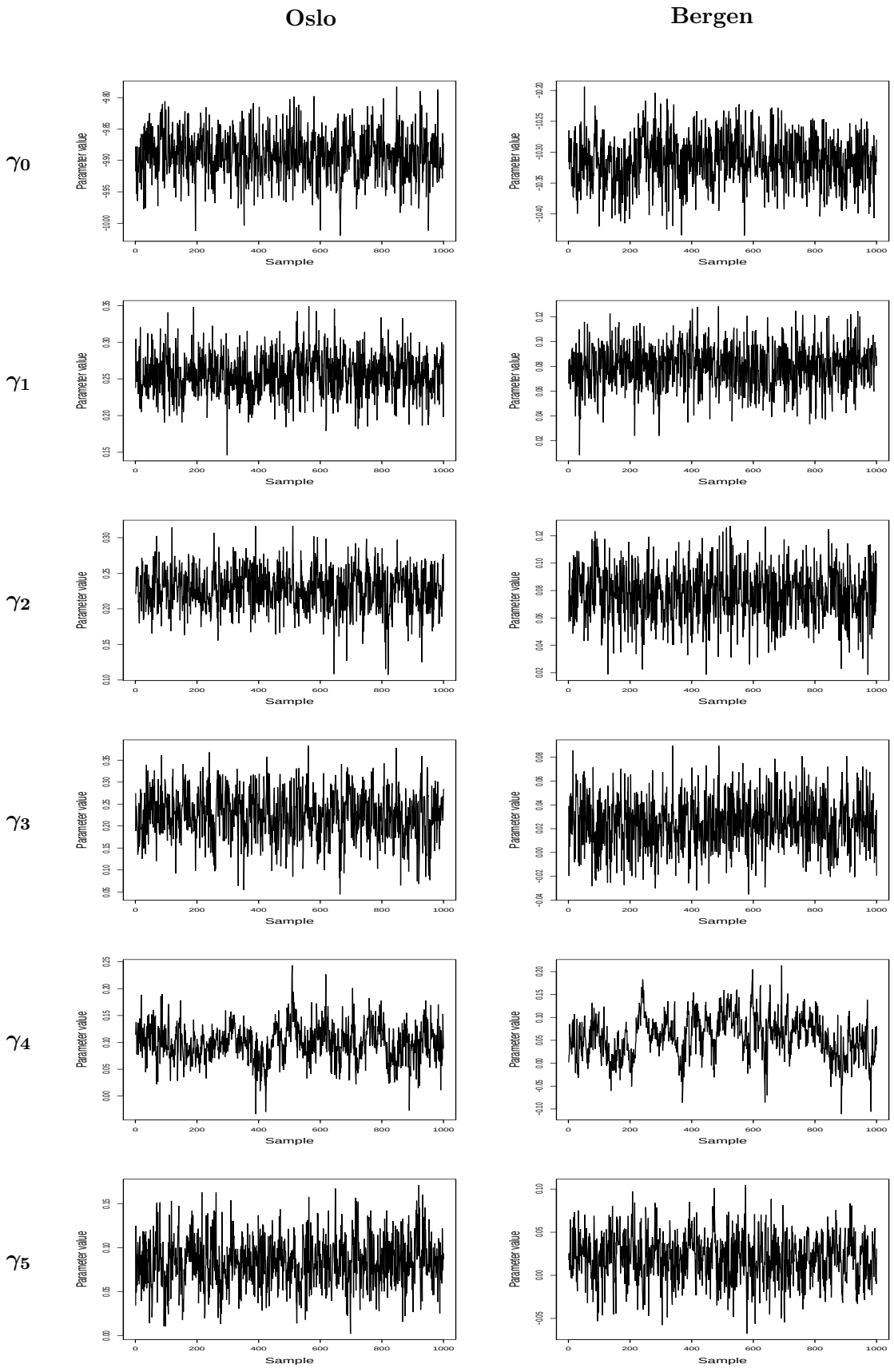


Figure B.1.3: Trace plots of the sampled realizations from the posterior distribution for the Hurdle component of the Poisson-Hurdle model and the proposed covariates for Oslo and Bergen.

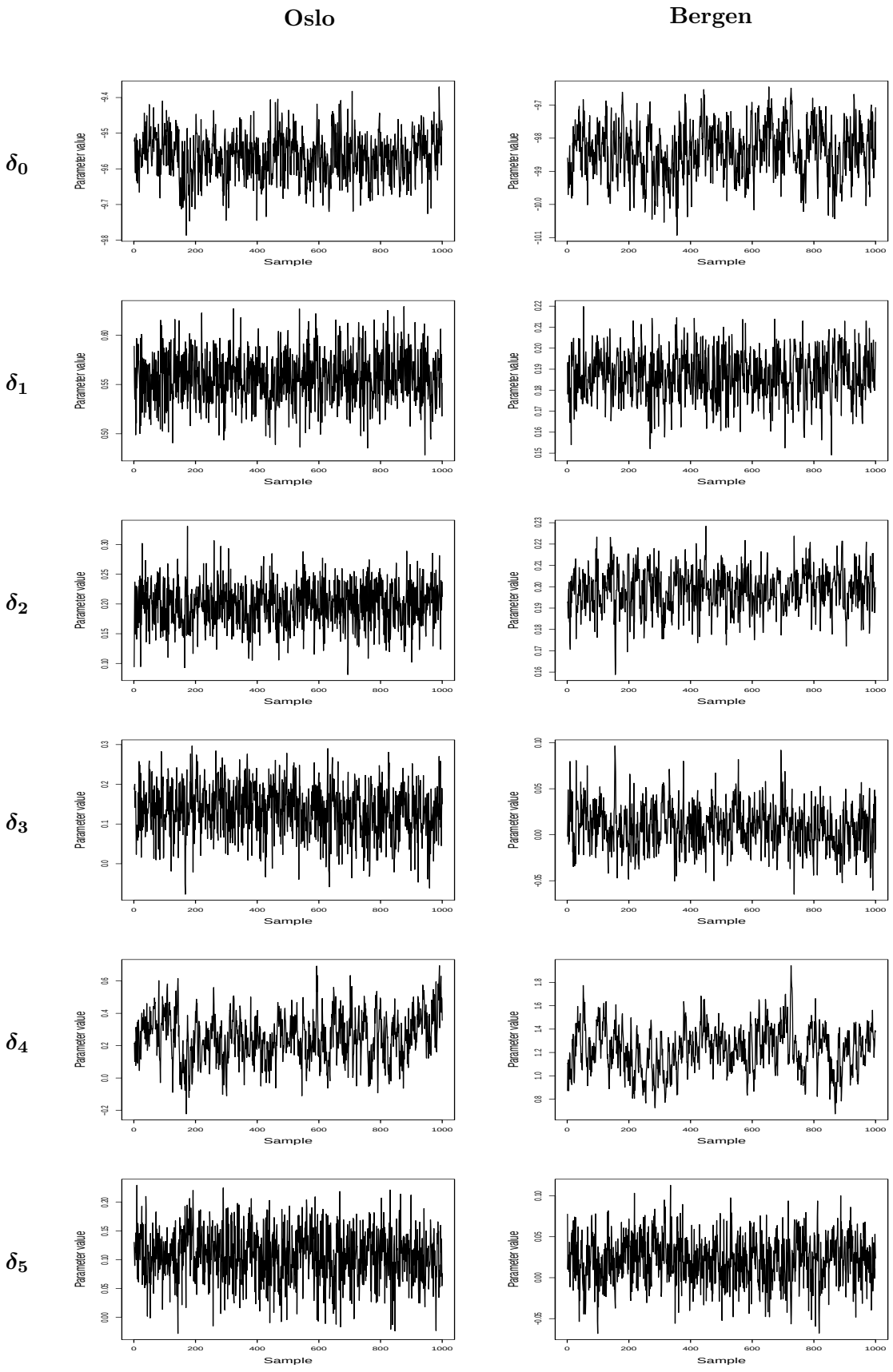


Figure B.1.4: Trace plots of the sampled realizations from the posterior distribution for the Poisson component of the Poisson-Hurdle model and the proposed covariates for Oslo and Bergen.

B.2 Posterior Mean estimates for the Binomial model with proposed covariates

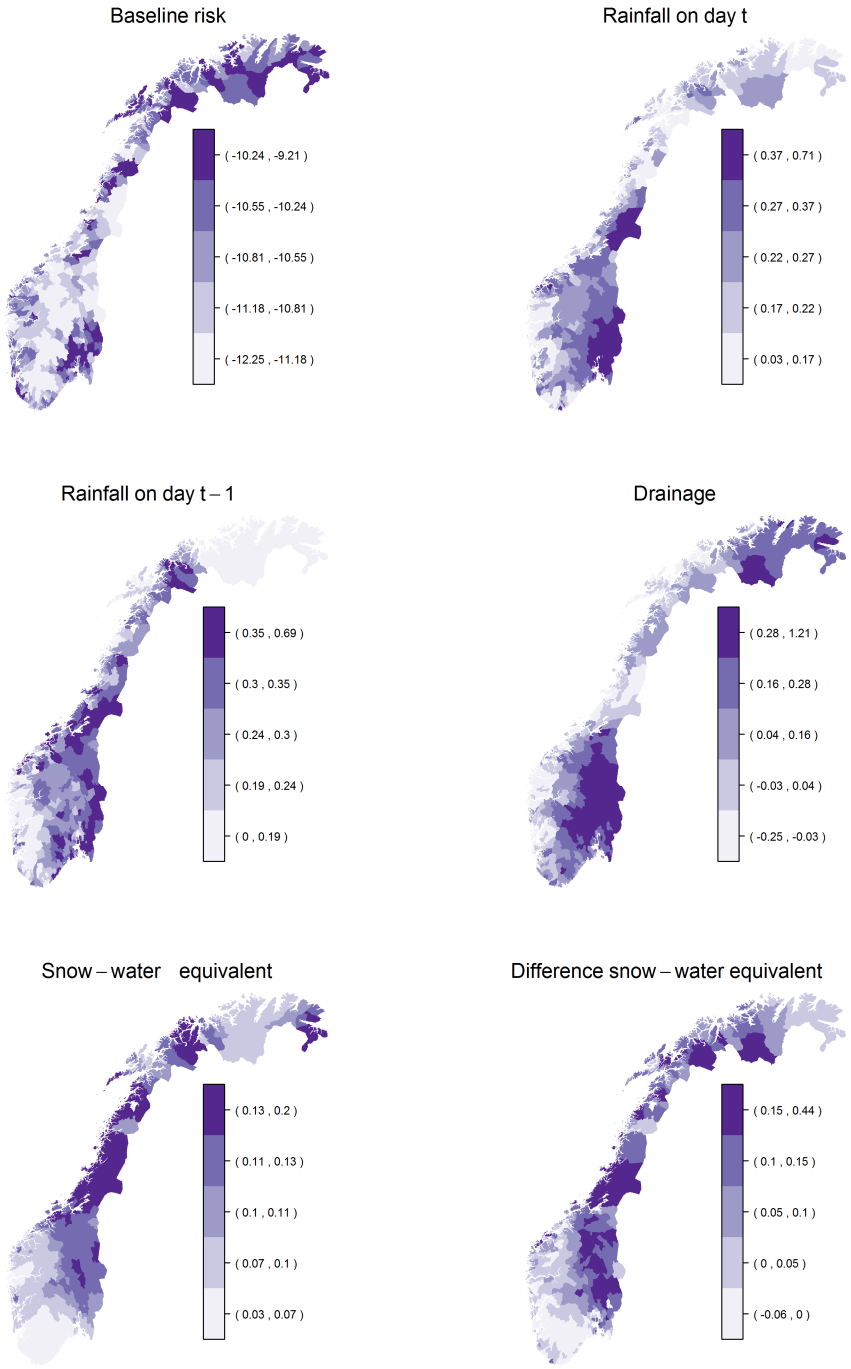


Figure B.2.1: Posterior mean estimates of the baseline risk and the covariate effects obtained for the Binomial model with the proposed data.

B.3 Predicting the weekly average number of claims

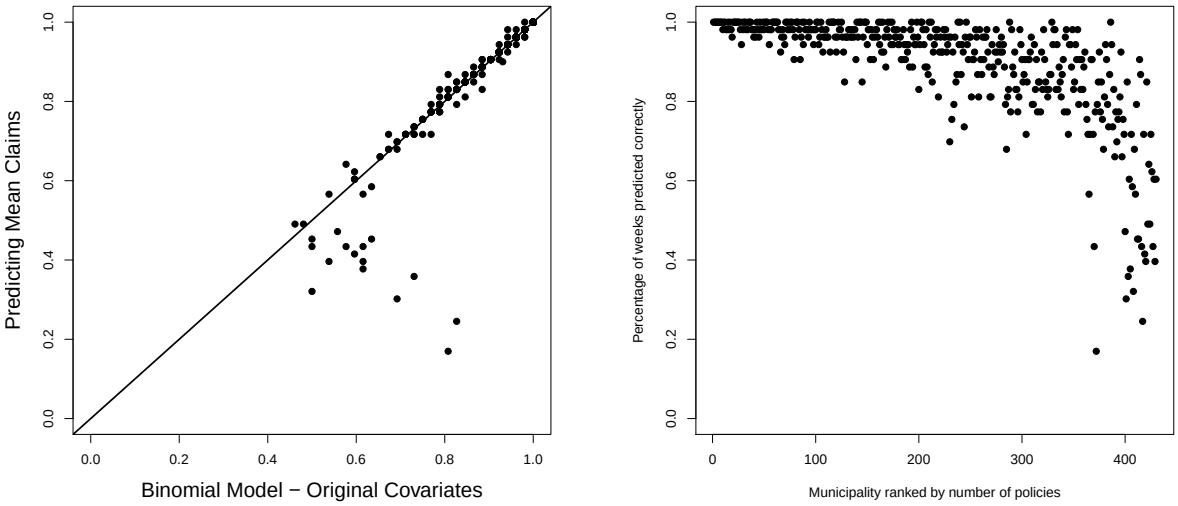


Figure B.3.1: Comparison of predictive performance for the Binomial model with original covariates and a prediction of the average weekly number of claims over the test period (left panel). The right panel illustrates the performance of predicting the average number of weekly claims with respect to the rank of the municipality in terms of number of policies.

Appendix C

Supplementary Material Chapter 4

C.1 Derivation of the Prior Density $\phi(\boldsymbol{\Delta}_k \mid \eta)$ in Section 4.2.3

We derive the prior density $\phi(\boldsymbol{\Delta}_k \mid \eta)$ in expression (4.2.8) in Section 4.2.3. For notational simplicity, the index k is dropped in the following. The number of points in $\boldsymbol{\Delta}$, corresponding to the number of jumps of λ, is defined as geometrically distributed with probability $\eta^{-1},\ \eta > 1$. Hence, its probability mass function is formally given as

$$\mathbb{P}(n(\boldsymbol{\Delta}) = n \mid \eta) = \frac{1}{\eta}\left(1 - \frac{1}{\eta}\right)^n.$$

Conditional on $n(\boldsymbol{\Delta})$, the distribution of the number of points in the subprocesses $\Delta_1, \ldots, \Delta_I$ is specified as being uniform over the set of possibilities to allocate $n(\boldsymbol{\Delta})$ points to I processes. Thus, the distribution $n(\Delta_1), \ldots, n(\Delta_I) \;\Big|\; n(\boldsymbol{\Delta})$ has probability mass function

$$\mathbb{P}\left[n(\Delta_1) = n_1, \ldots, n(\Delta_I) = n_I \;\Big|\; n(\boldsymbol{\Delta}) := \sum_{i=1}^{I} n(\Delta_i) = n\right] = \binom{n + I - 1}{n}^{-1}.$$

For subprocess $\Delta_i,\ i = 1, \ldots, I$, the location $\boldsymbol{\xi}_{i,j},\ j = 1, \ldots, n(\Delta_i)$, is uniformly distributed on the subspace X_i with volume $|X_i|$. The density for the vector $\underline{\boldsymbol{\xi}_i} = \left(\boldsymbol{\xi}_{i,1}, \ldots, \boldsymbol{\xi}_{i,n(\Delta_i)}\right)$, given $n(\Delta_i)$, is thus

$$\pi\left[\ \underline{\boldsymbol{\xi}_i} \mid n(\Delta_i)\ \right] = \left[\frac{1}{|X_i|}\right]^{n(\Delta_i)},\ i = 1, \ldots, I.$$

Given the $n(\boldsymbol{\Delta})$ locations, the marks are defined to be uniformly distributed on the prespecified interval $[\delta_{\min}, \delta_{\max}]$, subject to them satisfying the monotonic constraints. Formally, the density

yields to

$$\pi\left(\underline{\delta_1},\ldots,\underline{\delta_I} \mid \underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right) = \frac{n\left(\boldsymbol{\Delta}\right)!}{Z\left(\underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right)}\left(\frac{1}{\delta_{\max}-\delta_{\min}}\right)^{n(\boldsymbol{\Delta})},$$

where $n\left(\boldsymbol{\Delta}\right)!$ is the total number of permutations and $Z\left(\underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right)$ denotes the number of permutations which satisfy the monotonic constraints. For instance, if the covariate space is one-dimensional, $m = 1$, then the monotonic constraint imposes a total ordering and, hence, $Z\left(\underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right) = 1$.

Combining the individual densities and by application of the chain rule, $\phi\left(\boldsymbol{\Delta}_k \mid \eta\right)$ results in

$$\begin{aligned}
&\pi\left(\underline{\delta_1},\ldots,\underline{\delta_I} \mid \underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right) \times \prod_{i=1}^{I} \pi\left[\underline{\boldsymbol{\xi}_i} \mid n\left(\Delta_i\right)\right] \times \pi\left[n(\Delta_1),\ldots,n(\Delta_I) \mid n(\boldsymbol{\Delta})\right] \times \pi\left[n(\boldsymbol{\Delta}) \mid \eta\right] \\
=&\frac{n(\boldsymbol{\Delta})!}{Z\left(\underline{\boldsymbol{\xi}_1},\ldots,\underline{\boldsymbol{\xi}_I}\right)}\left(\frac{1}{\delta_{\max}-\delta_{\min}}\right)^{n(\boldsymbol{\Delta})} \times \prod_{i=1}^{I}\left[\frac{1}{|X_i|}\right]^{n(\Delta_i)} \times \binom{n(\boldsymbol{\Delta})+I-1}{n(\boldsymbol{\Delta})}^{-1} \times \frac{1}{\eta}\left(1-\frac{1}{\eta}\right)^{n(\boldsymbol{\Delta})}.
\end{aligned}$$

C.2 Details of the RJMCMC Algorithm

The following calculations derive the acceptance probabilities for the defined moves *Birth*, *Death* and *Shift*. For notational clarity, we first derive the acceptance probabilities in case $\omega = 0$, that is, the prior for $\boldsymbol{\Delta}_k,\ k = 1,\ldots,K$, yields to $\phi\left(\boldsymbol{\Delta}_k \mid \eta\right)$. As $\lambda_1,\ldots,\lambda_K$ are independent in this case, we consider $\lambda_k,\ k = 1,\ldots,K$, individually and drop the index k in the following. The proposed moves and associated acceptance probabilities are then similar to Saarela and Arjas (2011).

Prior to sampling the proposed move, one of the I processes $\Delta_1,\ldots,\Delta_I$ is randomly selected with equal probability. Let Δ_i denote the process which is to be updated. Next, one of the three defined moves *Birth*, *Death* and *Shift* is randomly selected with probability p_{Birth}, p_{Death} and $1 - p_{Birth} - p_{Death}$, respectively. If Δ_i contains no point, a proposed *Death* or *Shift* is rejected immediately. As *Birth* and *Death* lead to an increase and decrease, respectively, in the dimension of the parameter space, their acceptance probability has to be derived according to Green (1995).

In case of *Birth*, we propose to add a point $(\boldsymbol{\xi}^*, \delta^*)$ to the current process Δ_i. The proposal $(\boldsymbol{\xi}^*, \delta^*)$ is generated by first sampling the proposed location $\boldsymbol{\xi}^*$ uniformly on X_i. Next, the associated mark δ^* is sampled uniformly over the interval of values which satisfy the monotonic constraint, that is, the proposal distribution for δ^* depends on both $\boldsymbol{\xi}^*$ and the current set of

marked point processes $\boldsymbol{\Delta}_k$. The mapping between the parameter spaces then corresponds to the identity function and, thus, the determinant of the Jacobian is equal to 1. The reverse move *Death* selects one of the points in Δ_i with equal probability and proposes to remove it.

The acceptance probability for a *Birth* is then of the form

$$\min\left\{1, \prod_{t=1}^{T} \frac{f\left(y_t \mid \lambda^*\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)}{f\left(y_t \mid \lambda\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)} \times \frac{\phi\left(\boldsymbol{\Delta}^* \mid \eta\right)}{\phi\left(\boldsymbol{\Delta} \mid \eta\right)} \times \frac{1}{q\left(\boldsymbol{\xi}^*, \delta^* \mid \boldsymbol{\Delta}\right)} \times \frac{p_{Death}}{p_{Birth}}\right\}.$$

By using the expression for $\phi\left(\boldsymbol{\Delta} \mid \eta\right)$ derived in the previous Section C.1, the right-hand term in the acceptance probability can be written as

$$\prod_{t=1}^{T} \frac{f\left(y_t \mid \lambda^*\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)}{f\left(y_t \mid \lambda\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)} \times \frac{\pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i^*, \ldots \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i^* \ldots, \boldsymbol{\xi}_I\right)}{\pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i, \ldots \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i \ldots, \boldsymbol{\xi}_I\right)} \times \frac{\pi\left[\boldsymbol{\xi}_i^* \mid n\left(\Delta_i\right)+1\right]}{\pi\left[\boldsymbol{\xi}_i \mid n\left(\Delta_i\right)\right]} \times$$
$$\frac{\pi\left[n\left(\Delta_1\right), \ldots, n\left(\Delta_i\right)+1, \ldots, n\left(\Delta_I\right) \mid n\left(\boldsymbol{\Delta}\right)+1\right]}{\pi\left[n\left(\Delta_1\right), \ldots, n\left(\Delta_i\right), \ldots, n\left(\Delta_I\right) \mid n\left(\boldsymbol{\Delta}\right)\right]} \times \frac{\pi\left[n\left(\boldsymbol{\Delta}\right)+1 \mid \eta\right]}{\pi\left[n\left(\boldsymbol{\Delta}\right) \mid \eta\right]} \times \frac{1}{q\left(\boldsymbol{\xi}^*, \delta^* \mid \boldsymbol{\Delta}\right)} \times \frac{p_{Death}}{p_{Birth}}.$$

where $\boldsymbol{\xi}_i^* = \{\boldsymbol{\xi}_{i,j} \ : \ j = 1, \ldots, n\left(\Delta_i\right)\} \cup \boldsymbol{\xi}^*$ and $\boldsymbol{\delta}_i^* = \{\delta_{i,j} \ : \ j = 1, \ldots, n\left(\Delta_i\right)\} \cup \delta^*$. By using conditional probabilities and evaluating the prior densities, this term can be simplied to

$$\prod_{t=1}^{T} \frac{f\left(y_t \mid \lambda^*\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)}{f\left(y_t \mid \lambda\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)} \times \frac{\pi\left(\delta^* \mid \boldsymbol{\Delta}, \xi^*\right) \times \pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i^*, \ldots, \boldsymbol{\xi}_I\right)}{\pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_I\right)} \times \frac{1}{|X_i|} \times$$
$$\frac{\binom{n(\boldsymbol{\Delta})+I-1}{n(\boldsymbol{\Delta})}}{\binom{n(\boldsymbol{\Delta})+I}{n(\boldsymbol{\Delta})+1}} \times \left(1 - \frac{1}{\eta}\right) \times \frac{1}{q\left(\delta^* \mid \boldsymbol{\Delta}, \xi^*\right) \times q\left(\xi^*\right)} \times \frac{p_{Death}}{p_{Birth}}.$$

This equation is simplified further as follows: Firstly, $\pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i^*, \ldots, \boldsymbol{\xi}_I\right)$ in the numerator is independent of the proposed location $\boldsymbol{\xi}^*$ and hence cancels with the term $\pi\left(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i, \ldots, \boldsymbol{\delta}_I \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_I\right)$ in the denominator. Secondly, the proposal density $q\left(\boldsymbol{\xi}^*\right)$ is equal to $1/\left|X_i\right|$ and hence cancels with the second term of the prior ratio. Finally, $\pi\left(\delta^* \mid \boldsymbol{\Delta}, \xi^*\right)$ is a uniform density on the interval of marks which satisfy the monotonic constraints imposed by the locations in $\boldsymbol{\Delta}^*$. As this is identical to the proposal density, $\pi\left(\delta^* \mid \boldsymbol{\Delta}, \xi^*\right)$ cancels with $q\left(\delta^* \mid \boldsymbol{\Delta}, \xi^*\right)$. Consequently, the acceptance probability for a *Birth* yields

$$\min\left\{1, \prod_{t=1}^{T} \frac{f\left(y_t \mid \lambda^*\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)}{f\left(y_t \mid \lambda\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)} \times \frac{n\left(\boldsymbol{\Delta}\right)+1}{n\left(\boldsymbol{\Delta}\right)+I} \times \left(1 - \frac{1}{\eta}\right) \times \frac{p_{Death}}{p_{Birth}}\right\}.$$

Diametrically, in case the proposed move is *Death*, the proposed set $\boldsymbol{\Delta}^*$ of marked point

processes contains $n(\boldsymbol{\Delta}) - 1$ points. Hence, the acceptance probability for *Death* results in

$$\min\left\{1, \prod_{t=1}^{T} \frac{f\left(y_t \mid \lambda^*\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)}{f\left(y_t \mid \lambda\left(\mathbf{x}_t\right), \boldsymbol{\theta}\right)} \times \frac{n\left(\boldsymbol{\Delta}\right) + I - 1}{n\left(\boldsymbol{\Delta}\right)} \times \left(1 - \frac{1}{\eta}\right)^{-1} \times \frac{p_{Birth}}{p_{Death}}\right\}.$$

Finally, a *Shift* proposes to shift both the location and level of an existing support point while preserving the current partial ordering of the support points. The point $(\boldsymbol{\xi}_{i,j}, \delta_{i,j}) \in \Delta_i$ to be shifted is selected with equal probability. A new location $\boldsymbol{\xi}^*_{i,j}$ is then sampled uniformly with the lower and upper bounds in each covariate being given by the next higher and lower covariate values; see Saarela and Arjas (2011) for details. The proposed mark δ^*_{i,j^*} is then sampled uniformly on the set of possible values which preserve the monotonic constraint. This approach implies that the current and proposed set of marked point processes, $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}^*$, respectively, have the same prior and proposal density. Hence, the acceptance probability is equal to the likelihood ratio.

After deriving the acceptance probabilities for $\omega = 0$, the case $\omega > 0$ is considered. As the general prior $\pi\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K \mid \omega, \eta\right)$ is the product of $\phi\left(\boldsymbol{\Delta}_k \mid \eta\right)$ and the defined dependence model $\pi\left(\lambda_1, \ldots, \lambda_K \mid \omega\right)$, the acceptance probability has to be extended by the ratio in the dependence components, but remains the same otherwise. For instance, in case a *Birth* is proposed for process $\Delta_{k,i}$, the acceptance probability is given as

$$\begin{aligned}
&\min\left\{1, \prod_{t=1}^{T_k} \frac{f\left(y_{k,t} \mid \lambda^*_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)}{f\left(y_{k,t} \mid \lambda_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)} \times \frac{\pi\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}^*_k, \ldots, \boldsymbol{\Delta}_K \mid \omega, \eta\right)}{\pi\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_k, \ldots, \boldsymbol{\Delta}_K \mid \omega, \eta\right)} \times \frac{1}{q\left(\boldsymbol{\xi}^*, \delta^* \mid \boldsymbol{\Delta}\right)} \times \frac{p_{Death}}{p_{Birth}}\right\} \\
&= \min\left\{1, \prod_{t=1}^{T_k} \frac{f\left(y_{k,t} \mid \lambda^*_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)}{f\left(y_{k,t} \mid \lambda_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)} \times \frac{\pi\left(\lambda_1, \ldots, \lambda^*_k, \ldots, \lambda_K \mid \omega\right)}{\pi\left(\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K \mid \omega\right)} \times \frac{\phi\left(\boldsymbol{\Delta}^*_k \mid \eta\right)}{\phi\left(\boldsymbol{\Delta}_k \mid \eta\right)} \times \right. \\
&\qquad \left. \frac{1}{q\left(\boldsymbol{\xi}^*, \delta^* \mid \boldsymbol{\Delta}\right)} \times \frac{p_{Death}}{p_{Birth}}\right\}. \\
&= \min\left\{1, \prod_{t=1}^{T_k} \frac{f\left(y_{k,t} \mid \lambda^*_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)}{f\left(y_{k,t} \mid \lambda_k\left(\mathbf{x}_{k,t}\right), \boldsymbol{\theta}_k\right)} \times \prod_{\substack{k'=1 \\ k' \neq k}}^{K} \frac{\exp\left[-\omega \cdot d_{k,k'} \cdot D_{p,q}\left(\lambda^*_k, \lambda_{k'}\right)\right]}{\exp\left[-\omega \cdot d_{k,k'} \cdot D_{p,q}\left(\lambda_k, \lambda_{k'}\right)\right]} \times \left(1 - \frac{1}{\eta}\right) \times \right. \\
&\qquad \left. \frac{n\left(\boldsymbol{\Delta}_k\right) + 1}{n\left(\boldsymbol{\Delta}_k\right) + I} \times \frac{p_{Death}}{p_{Birth}}\right\}.
\end{aligned}$$

as only the pairs in $\pi\left(\lambda_1, \ldots, \lambda_K \mid \omega\right)$ involving λ_k have to be evaluated.

C.3 Detection of Discontinuities via Sampled Point Processes

Interest lies in the detection of discontinuities in λ_k based on the samples obtained via the RJMCMC algorithm. For notational simplicity, the index k is dropped in the following as the outlined procedure considers the K functions independently. To detect discontinuities, sampled points have to be distinguished into those representing a jump and those approximating a continuous shape. In general, discontinuities are expected to occur in most of the samples, i.e. they are removed with low probability and a shift is only likely to be accepted if it changes the point marginally in both location and mark. Assume that a functional change in λ is defined as a discontinuity if the change in the functional level exceeds a threshold ρ.

Based on these considerations, each sampled point is classified as follows: Consider the $r+1$th sample, $\boldsymbol{\Delta}^{(r+1)}$. Further, let $\boldsymbol{\Psi}^{(r)} = \left\{\boldsymbol{\psi}_1^{(r)}, \ldots, \boldsymbol{\psi}_{n(\boldsymbol{\Psi}^{(r)})}^{(r)}\right\}$ denote the set of potential discontinuities after iteration $r = 1, \ldots, R$. A point $\left(\boldsymbol{\xi}_j^{(r+1)}, \delta_j^{(r+1)}\right),\ j = 1, \ldots, n\left(\boldsymbol{\Delta}^{(r+1)}\right)$, in $\boldsymbol{\Delta}^{(r+1)}$, is then examined as follows:

1. If the functional level difference of $\lambda^{(r+1)}$ at $\boldsymbol{\xi}_j^{(r+1)}$ is smaller than ρ, the point is not classified as potential discontinuity. Formally, it is checked whether

$$\lambda_k^{(r+1)}\left(\boldsymbol{\xi}_j^{(r+1)}\right) = \delta_j^{(r+1)} < \lambda_k^{(r+1)}\left(\boldsymbol{\xi}_j^{(r+1)} - \boldsymbol{\epsilon}\right) + \rho,$$

   where $\boldsymbol{\xi}_j^{(r+1)} - \boldsymbol{\epsilon}$ refers to close point on which $\boldsymbol{\xi}_j^{(r+1)}$ puts a monotonic constraint.

2. Given that the functional level difference exceeds ρ, we examine whether it coincides with one of the potential discontinuities in $\boldsymbol{\Psi}^{(r)}$. Consider one potential discontinuity $\boldsymbol{\psi}_h^{(r)} = \left(\widetilde{\boldsymbol{\xi}}_{\psi_h}^{(r)}, \widetilde{\delta}_{\psi_h}^{(r)}, \widetilde{n}_{\psi_h}^{(r)}\right),\ h = 1, \ldots, n(\Psi^{(r)})$, where the third entry denotes the number of occurrences of this discontinuity in the first r samples. Then, $\left(\boldsymbol{\xi}_j^{(r+1)}, \delta_j^{(r+1)}\right)$ is considered as the same discontinuity as $\boldsymbol{\psi}_h^{(r)}$ if is close to it in both its location and mark. Formally, we check whether

   (a) $\left\|\boldsymbol{\xi}_j^{(r+1)} - \widetilde{\boldsymbol{\xi}}_{\psi_h}^{(r)}\right\|_2^2 \leq \tau$, i.e the points are close in the covariate space and,

   (b) $\left|\delta_j - \widetilde{\delta}_{\psi_h}\right| \leq \upsilon$, i.e. the points have similar functional level,

   where τ and υ are prespecified constants. If no point in $\boldsymbol{\Psi}$ fulfills these two properties, $\left(\boldsymbol{\xi}_j^{(r+1)}, \delta_j^{(r+1)}\right)$ is added to the set. Otherwise, the current discontinuity ψ_h is updated

via

$$\begin{aligned}
\widetilde{\boldsymbol{\xi}}_{\psi_h}^{(r+1)} &= \frac{\widetilde{n}_{\psi_h}^{(r)}}{\widetilde{n}_{\psi_h}^{(r)}+1}\widetilde{\boldsymbol{\xi}}_{\psi_h}^{(r)} + \frac{1}{\widetilde{n}_{\psi_h}^{(r)}+1}\boldsymbol{\xi}_j \qquad \text{component-wise,} \\
\widetilde{\delta}_{\psi_h}^{(r+1)} &= \frac{\widetilde{n}_{\psi_h}^{(r)}}{\widetilde{n}_{\psi_h}^{(r)}+1}\widetilde{\delta}_{\psi_h}^{(r)} + \frac{1}{\widetilde{n}_{\psi_h}^{(r)}+1}\delta_j, \\
\widetilde{n}_{\psi_h}^{(r+1)} &= \widetilde{n}_{\psi_h}^{(r)}+1
\end{aligned}$$

Based on this classification, potential discontinuities are listed and their empirical occurrence rate across the R samples is derived.

C.4 Posterior Mean Plots for Sensitivity Analysis on η

In the following, the posterior mean plots for regions 1 and 2 for the four settings and five studies in Table 4.3.1 on page 90 are presented. While the posterior mean plots for the settings with $\omega \neq 0$ are more or less the same, some differences are found, when we compare them to the case $\omega = 0$. The plots show that, due to the lack of information for extrapolation, the functional levels are different from the truth at the border of the sample space. Hence, the performance was only evaluated based on the convex hull of the observations for the region itself. In other words, the poor fit in the lower left and upper right corner does not affect the results in Table 4.3.1. This section further presents some trace plots which were used to confirm convergence and to assess the mixing of the sampled Markov chains. All trace plots indicate convergence and also a moderate to good mixing, in particular, for region 2. The quality of mixing also correlates negatively with the number of points as a higher number means that the functional level at an arbitrary point changes less frequently. For instance in Study 1 for $\eta = \widehat{\eta}$, the average number of points for region 1 is 280 while it is about 60 for region 2.

Study 1

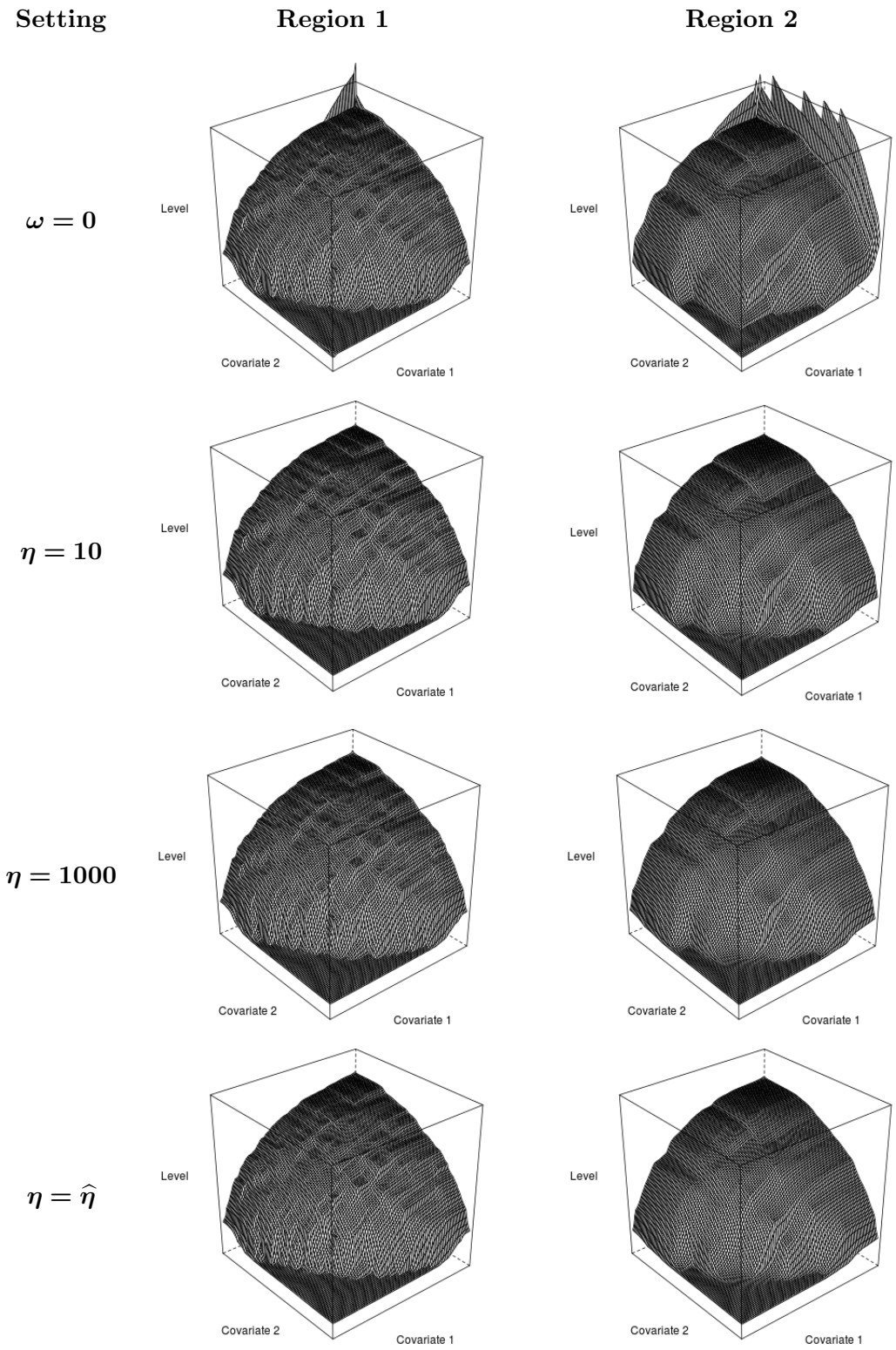


Figure C.4.1: Posterior mean plots for region 1 (left column) and region 2 (right column) for different settings of η and $\omega = 0$.

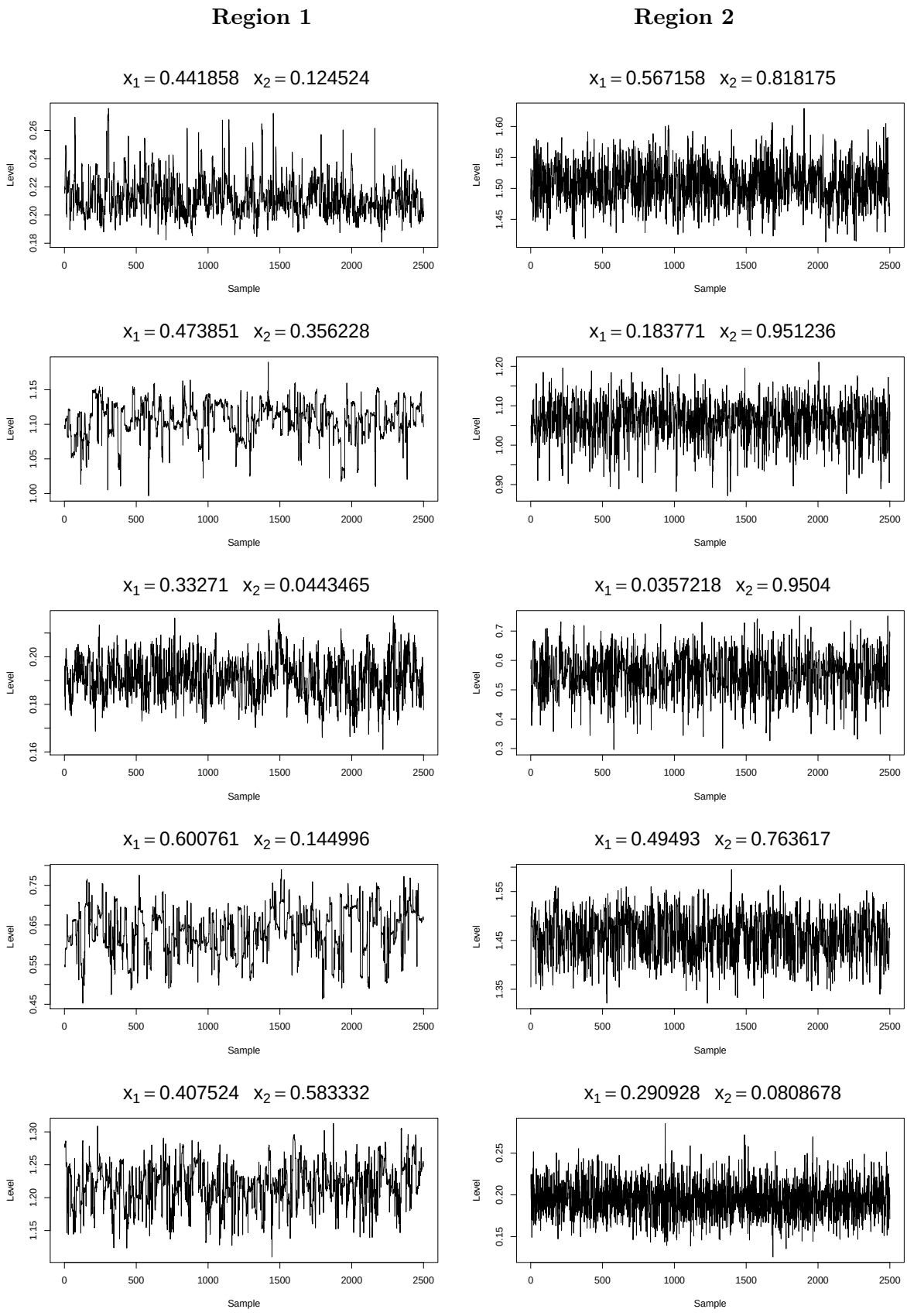


Figure C.4.2: Trace plots of the functional level at five random points for region 1 (left column) and region 2 (right column) for $\eta = \widehat{\eta}$.

Study 2

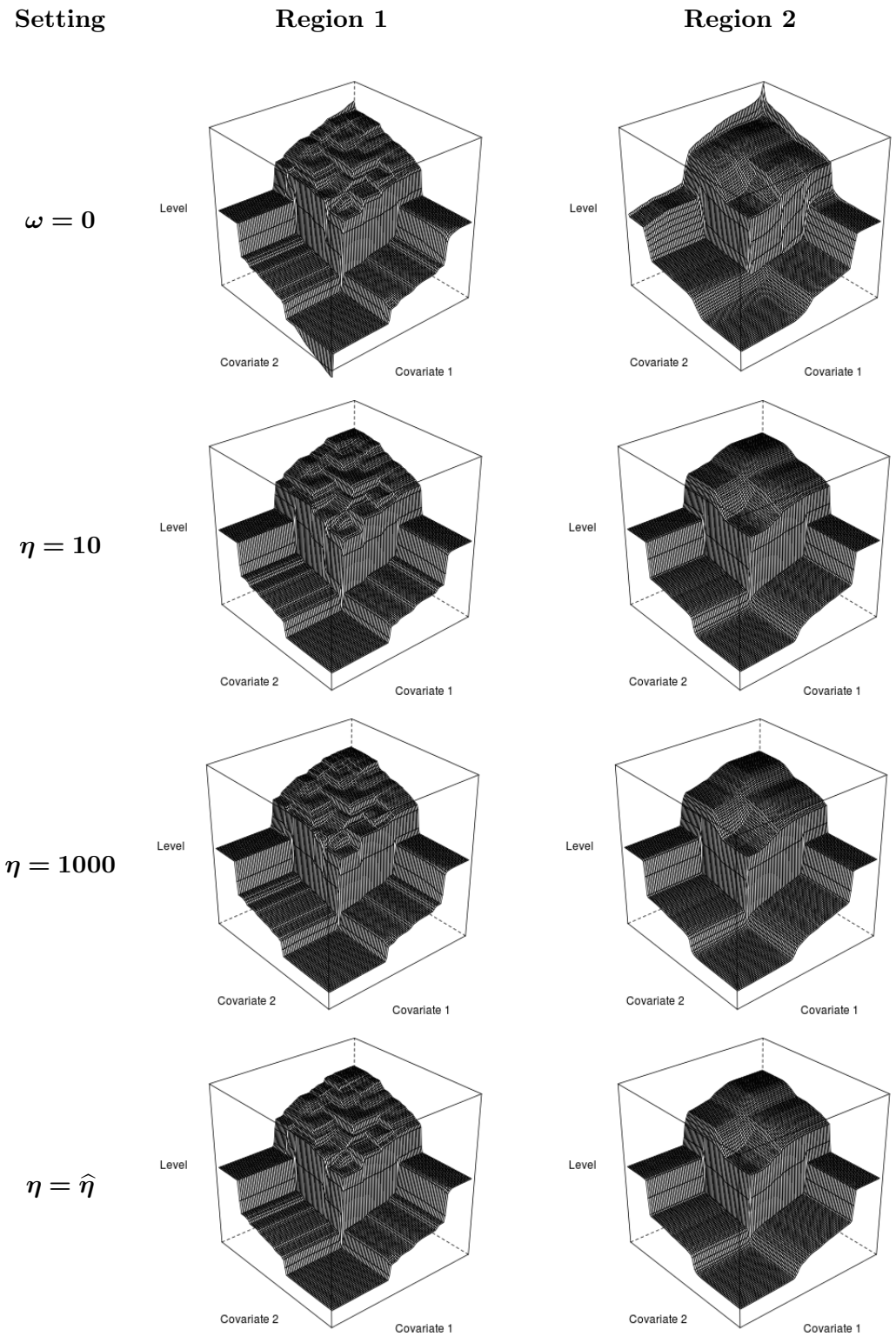


Figure C.4.3: Posterior mean plots for region 1 (left column) and region 2 (right column) for different settings of η and $\omega = 0$.

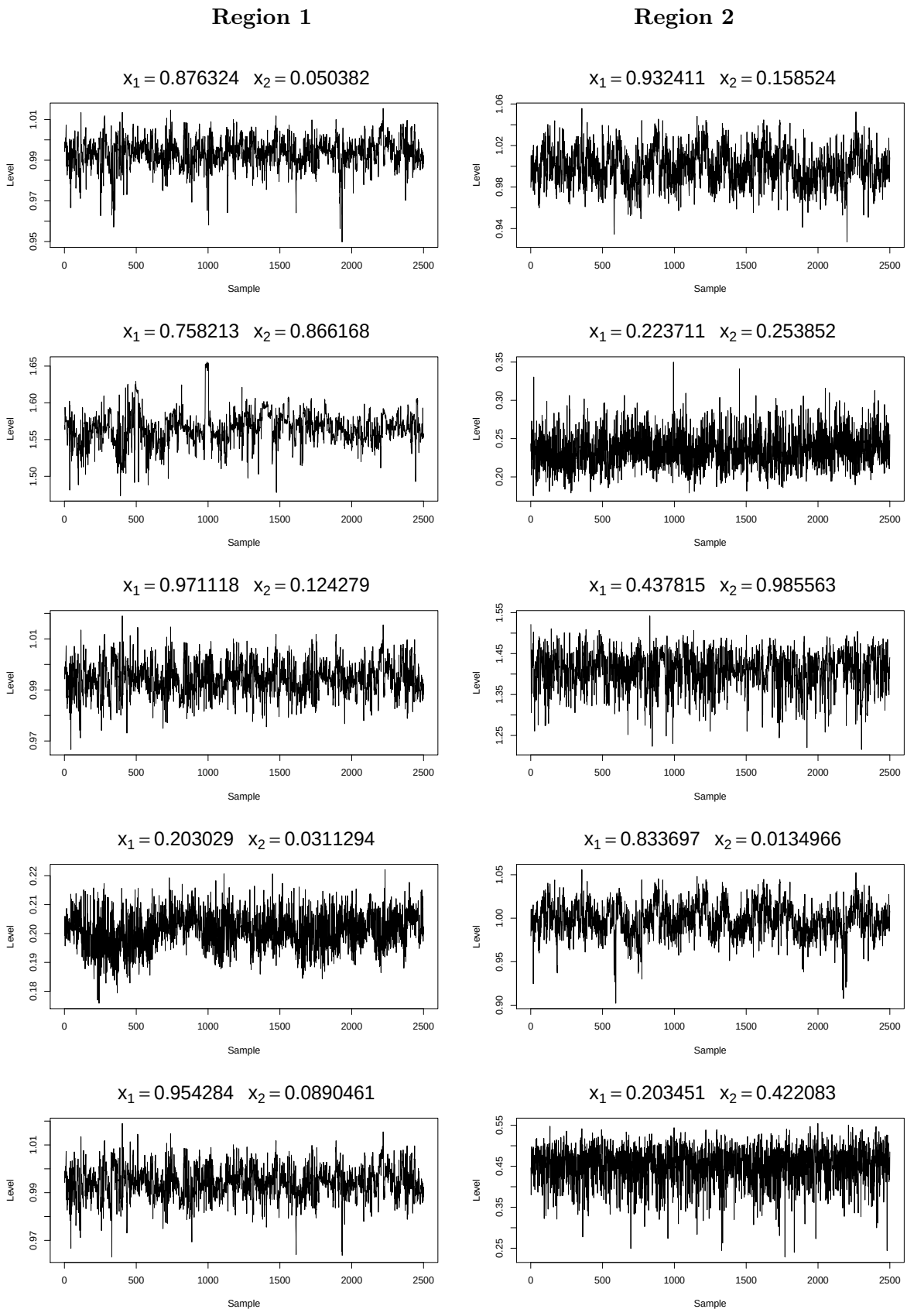


Figure C.4.4: Trace plots of the functional level at five random points for region 1 (left column) and region 2 (right column) for $\eta = \widehat{\eta}$.

Study 3

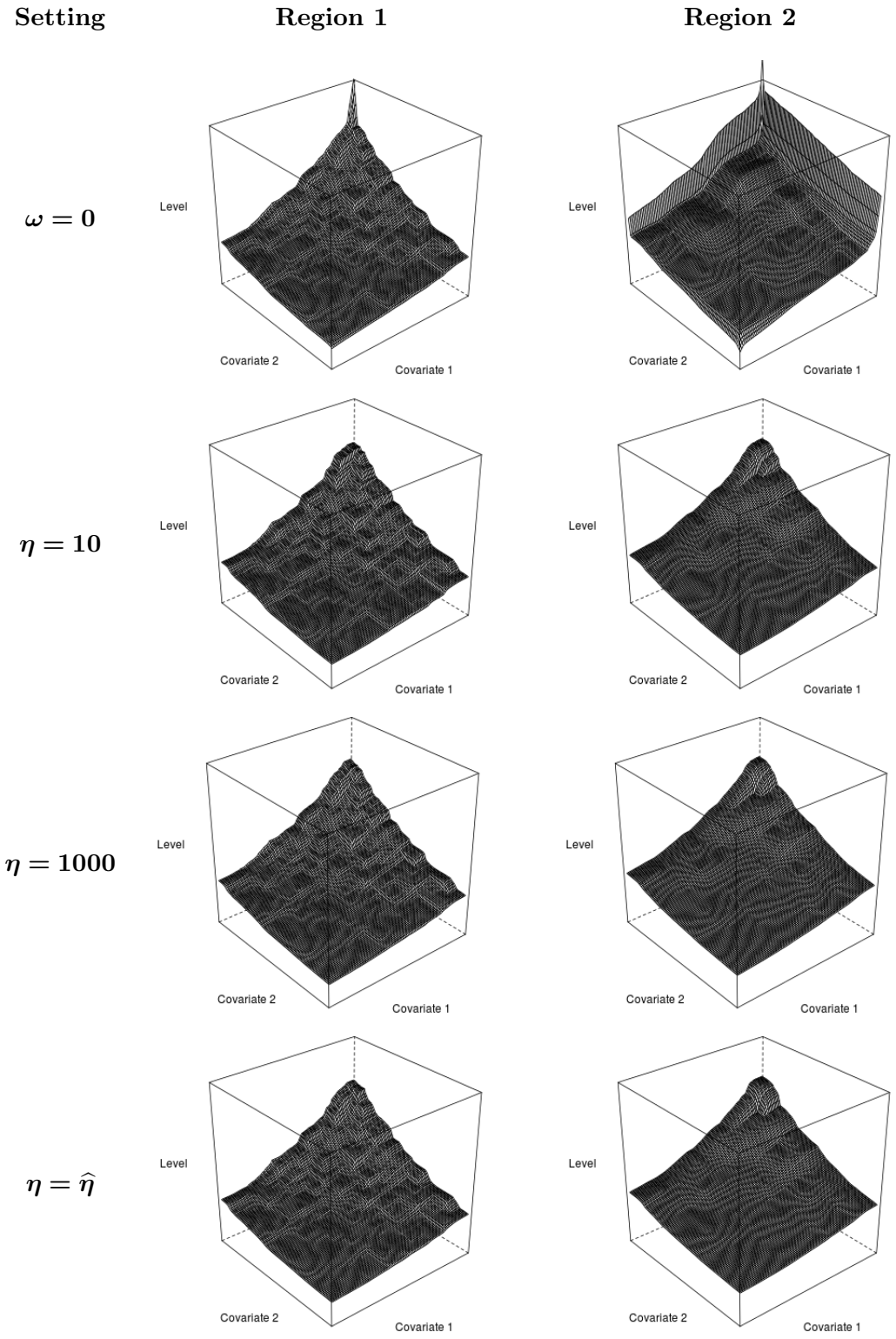


Figure C.4.5: Posterior mean plots for region 1 (left column) and region 2 (right column) for different settings of η and $\omega = 0$.

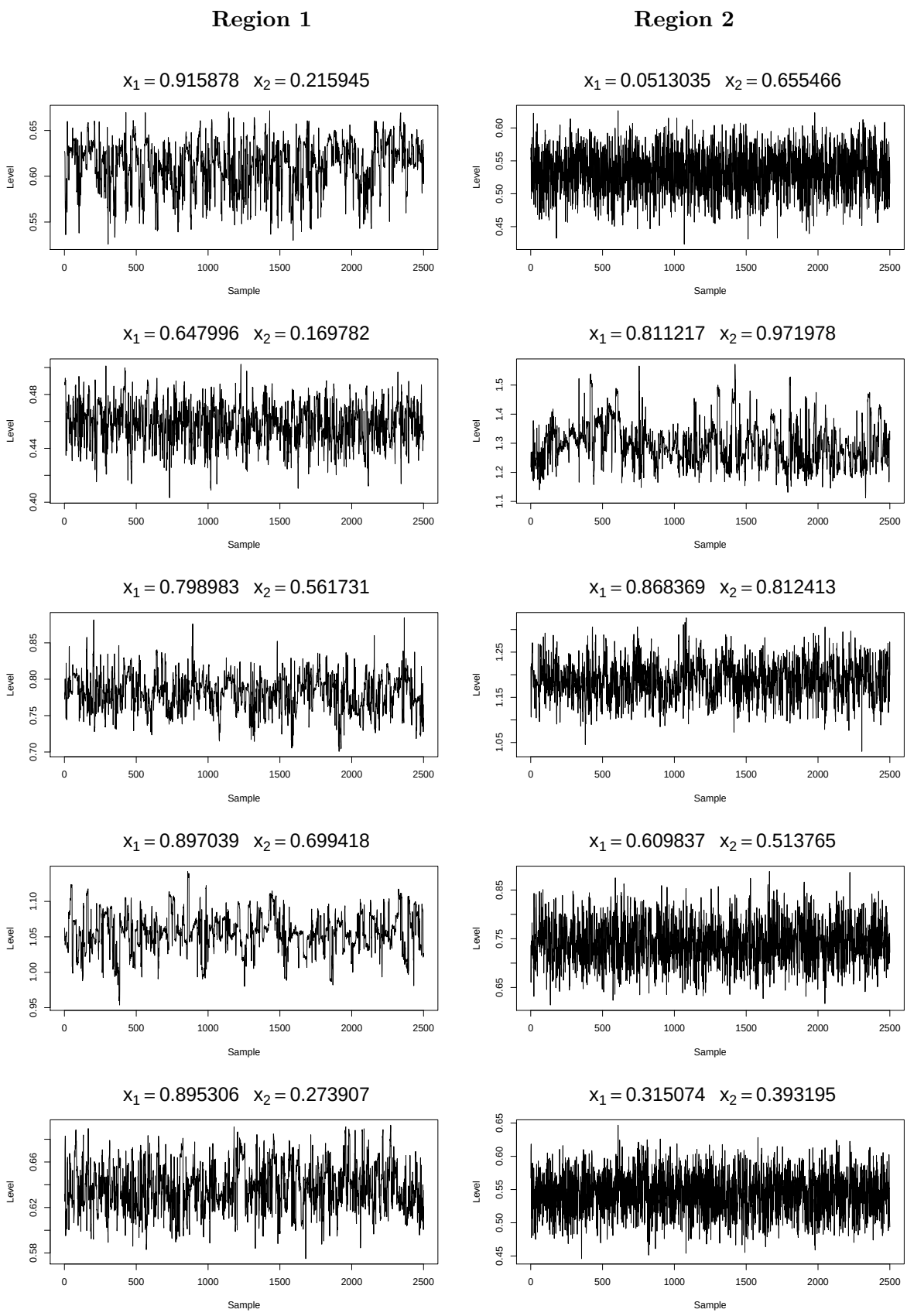


Figure C.4.6: Trace plots of the functional level at five random points for region 1 (left column) and region 2 (right column) for $\eta = \widehat{\eta}$.

Study 4

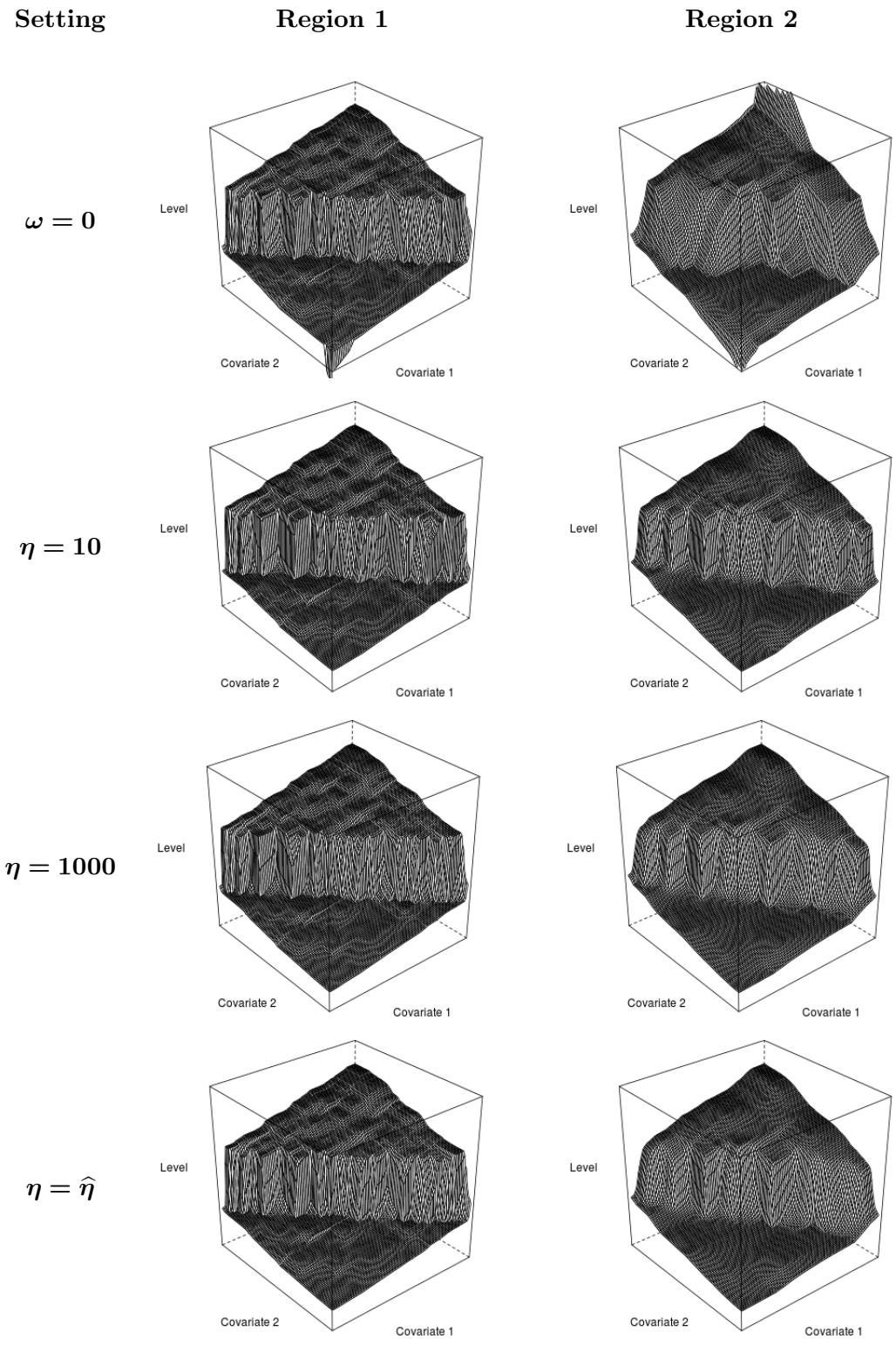


Figure C.4.7: Posterior mean plots for region 1 (left column) and region 2 (right column) for different settings of η and $\omega = 0$.

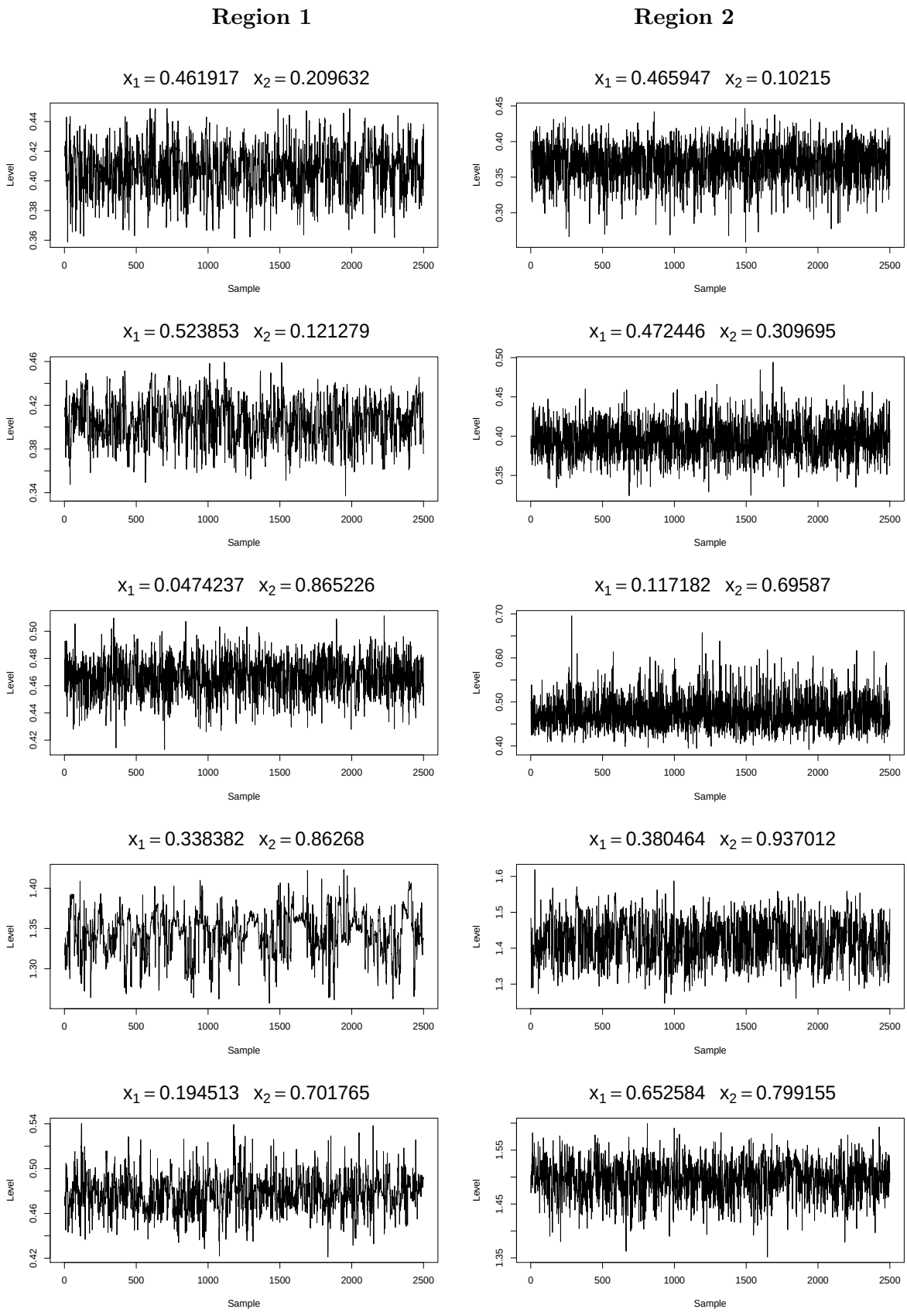


Figure C.4.8: Trace plots of the functional level at five random points for region 1 (left column) and region 2 (right column) for $\eta = \widehat{\eta}$.

Study 5

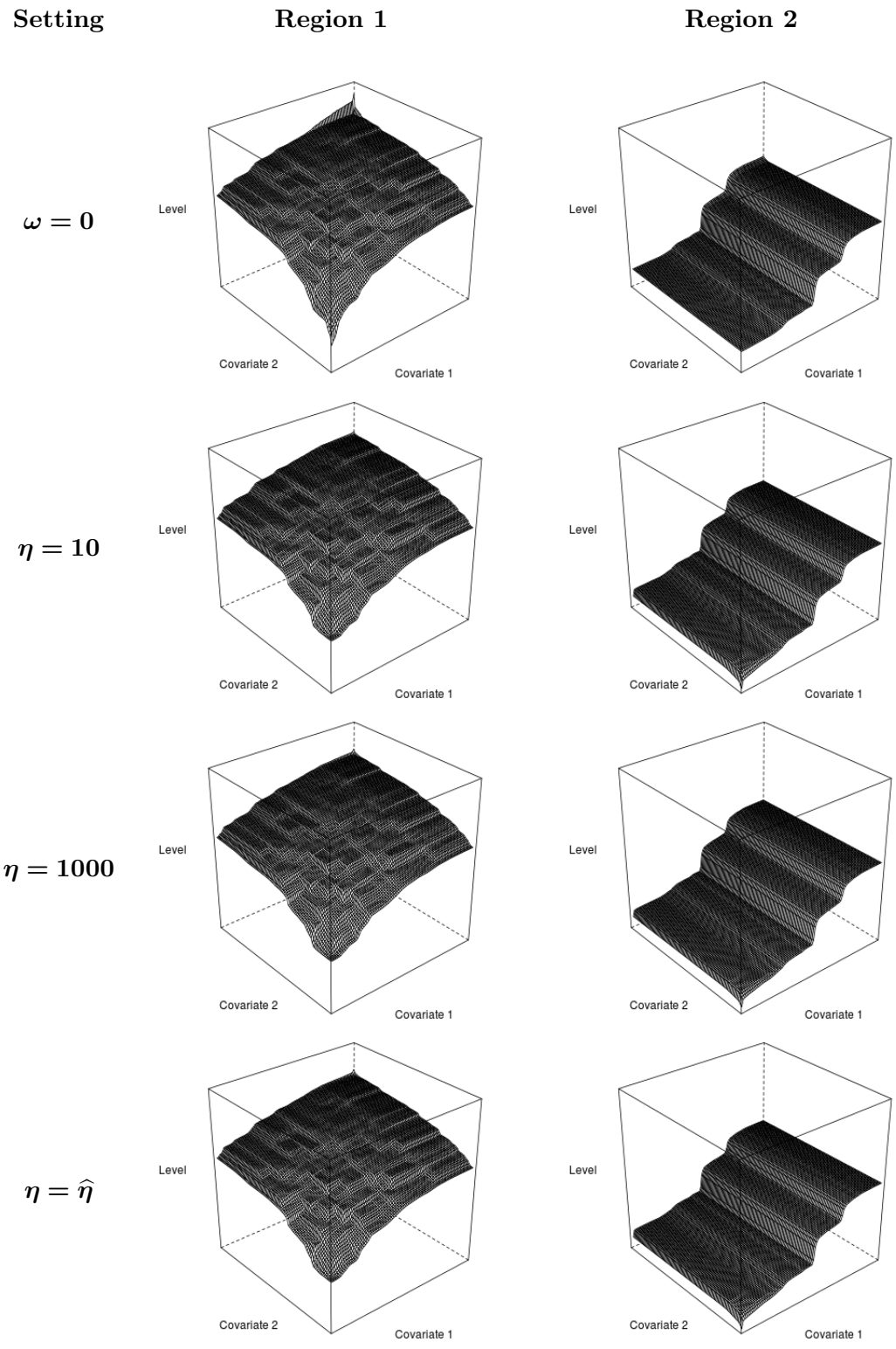


Figure C.4.9: Posterior mean plots for region 1 (left column) and region 2 (right column) for different settings of η and $\omega = 0$.

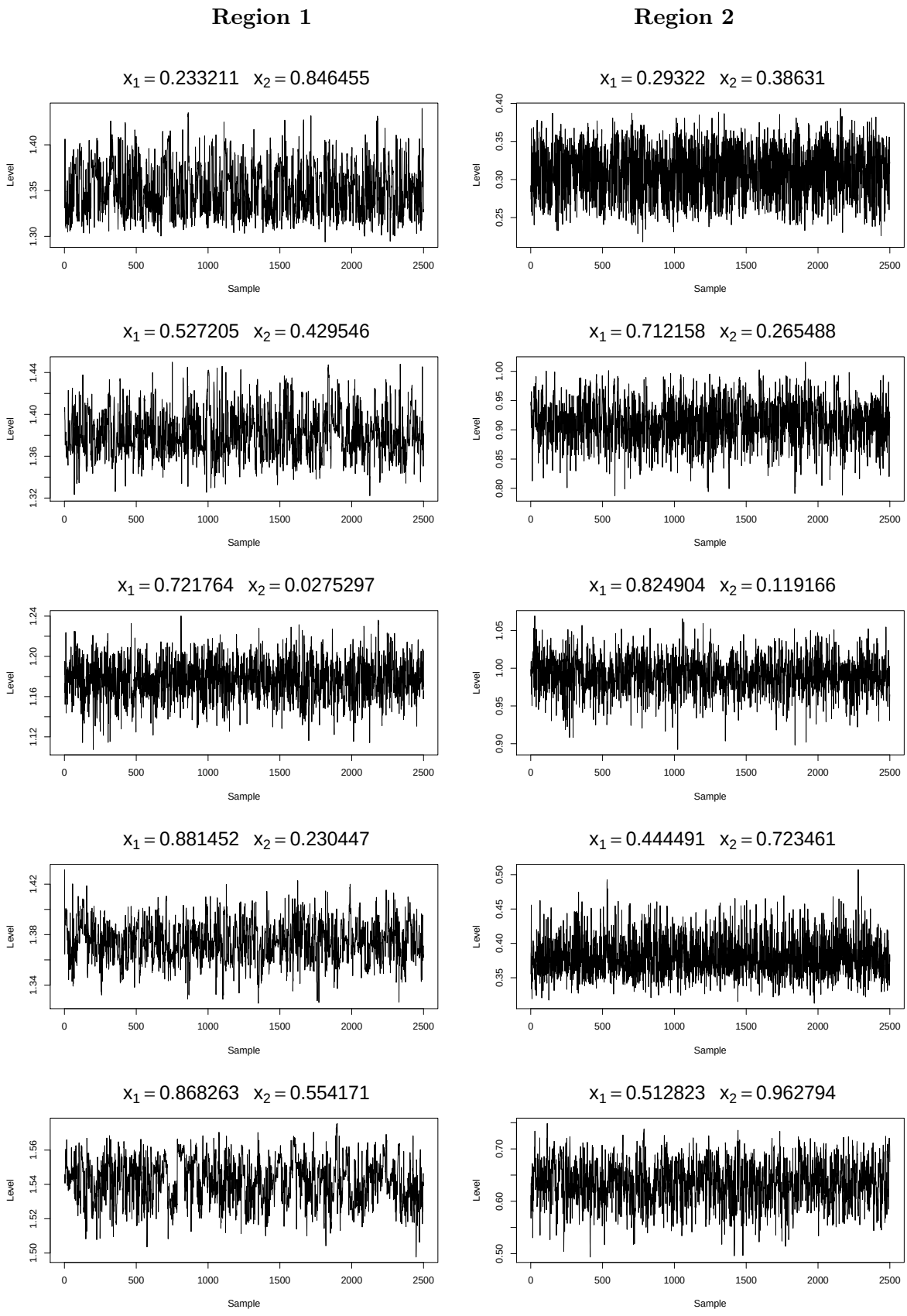


Figure C.4.10: Trace plots of the functional level at five random points for region 1 (left column) and region 2 (right column) for $\eta = \widehat{\eta}$.

C.5 Posterior Mean Plots for Sensitivity Analysis on p and q

Study 1 - Region 1

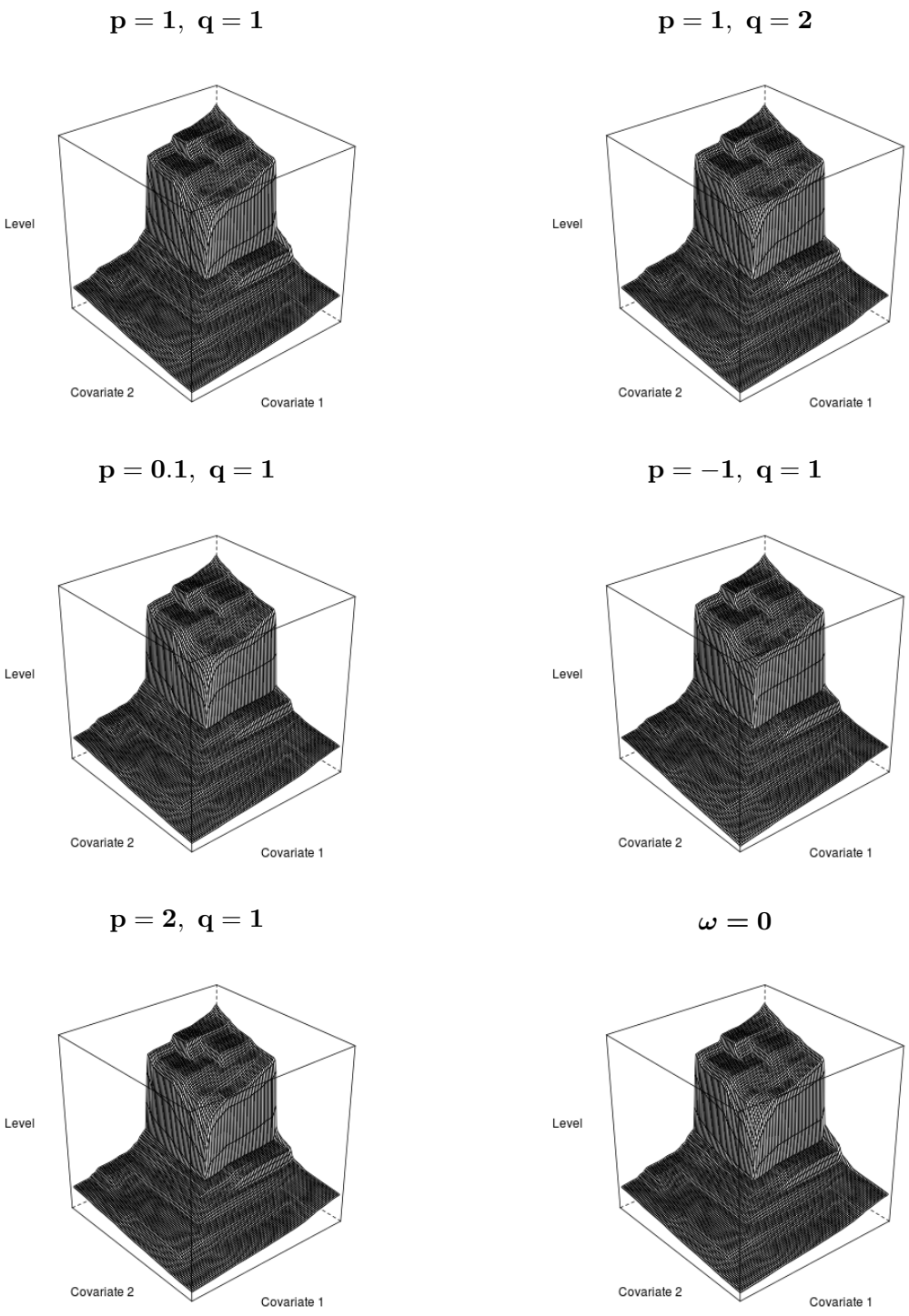


Figure C.5.1: Posterior mean plots for Region 1 for different settings of p and q, and for $\omega = 0$.

Study 1 - Region 2

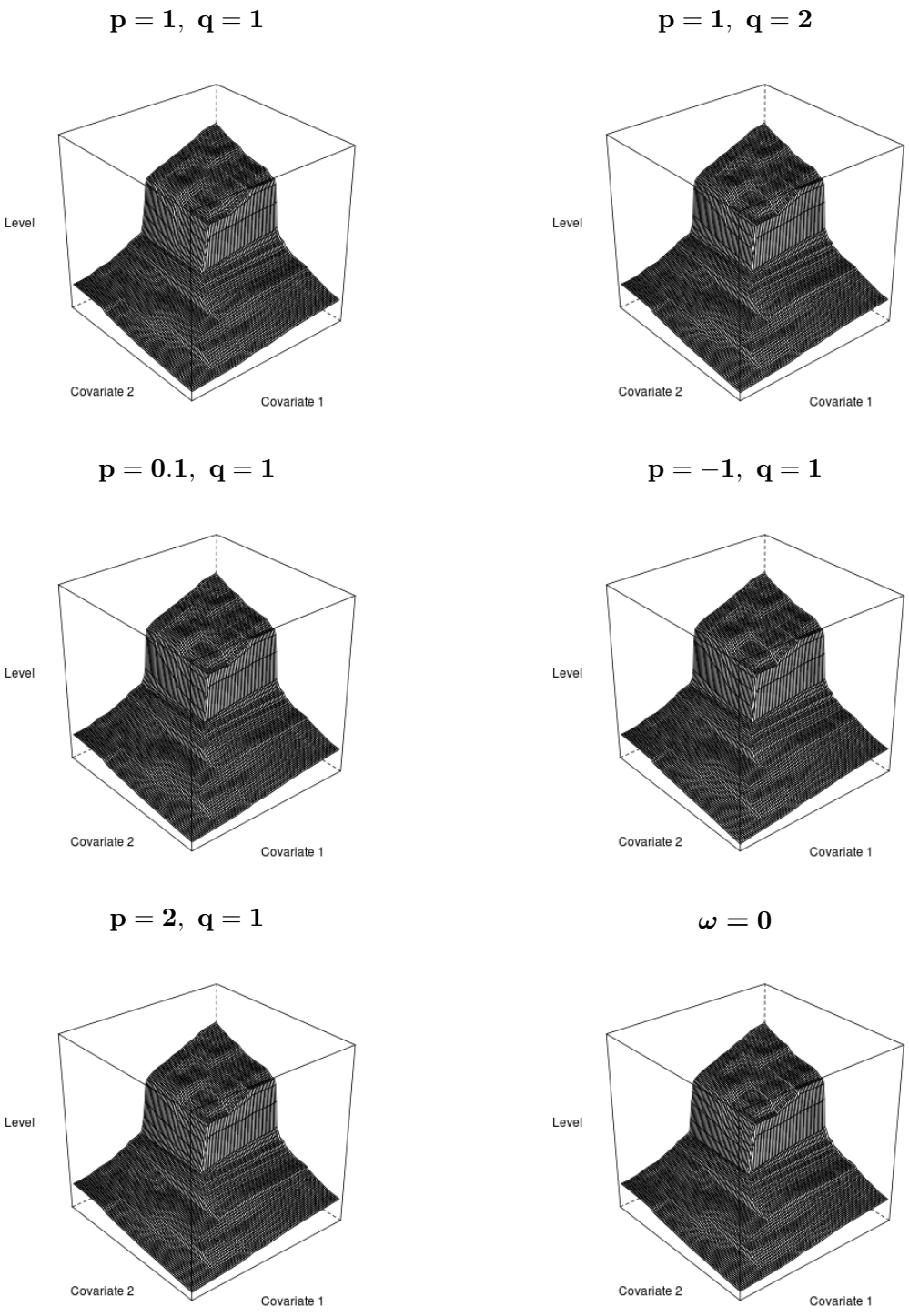


Figure C.5.2: Posterior mean plots for Region 2 in Study 1 for different settings of p and q, and for $\omega = 0$.

Study 2 - Region 1

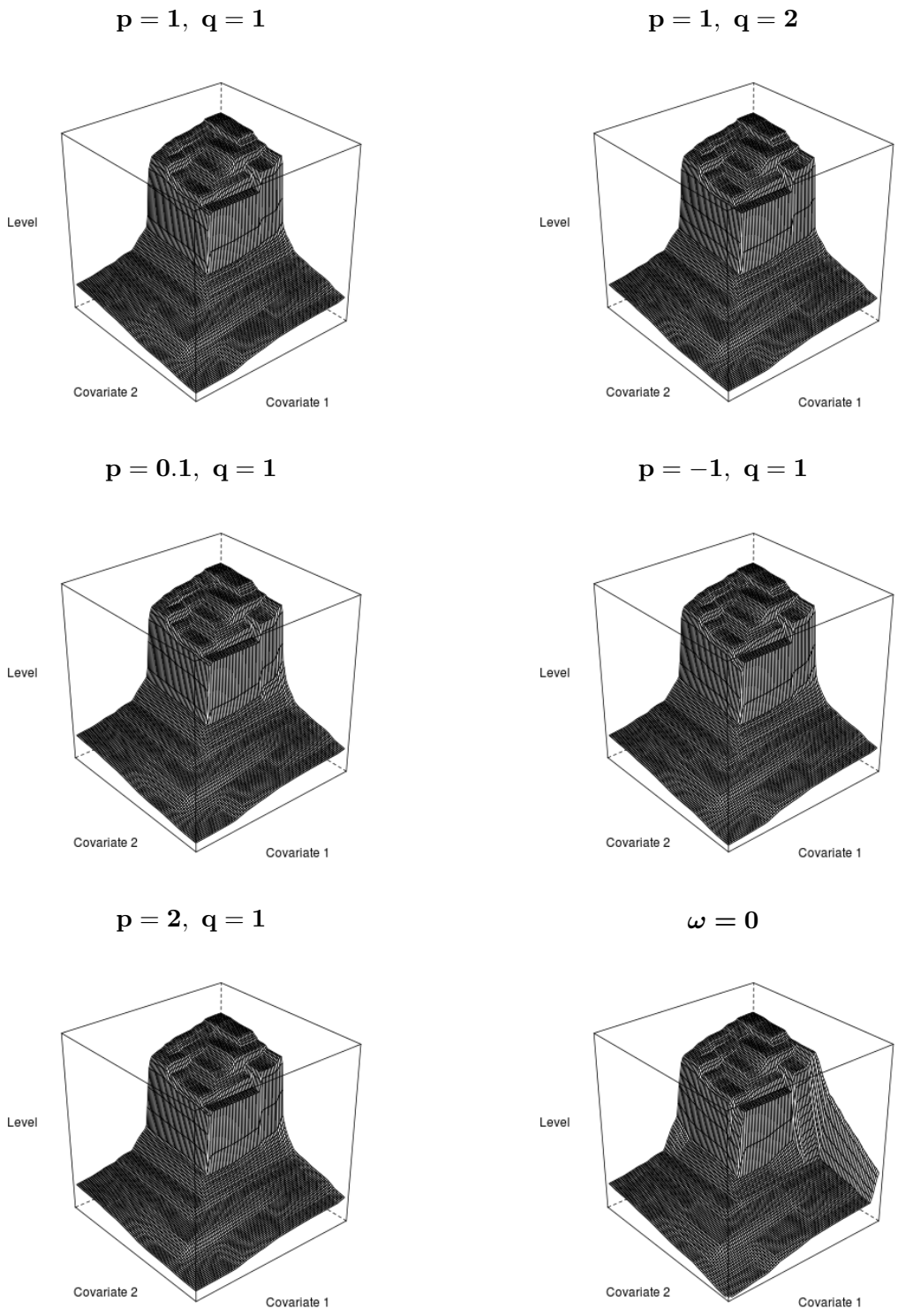


Figure C.5.3: Posterior mean plots for Region 1 in Study 2 for different settings of p and q, and for $\omega = 0$.

Study 2 - Region 2

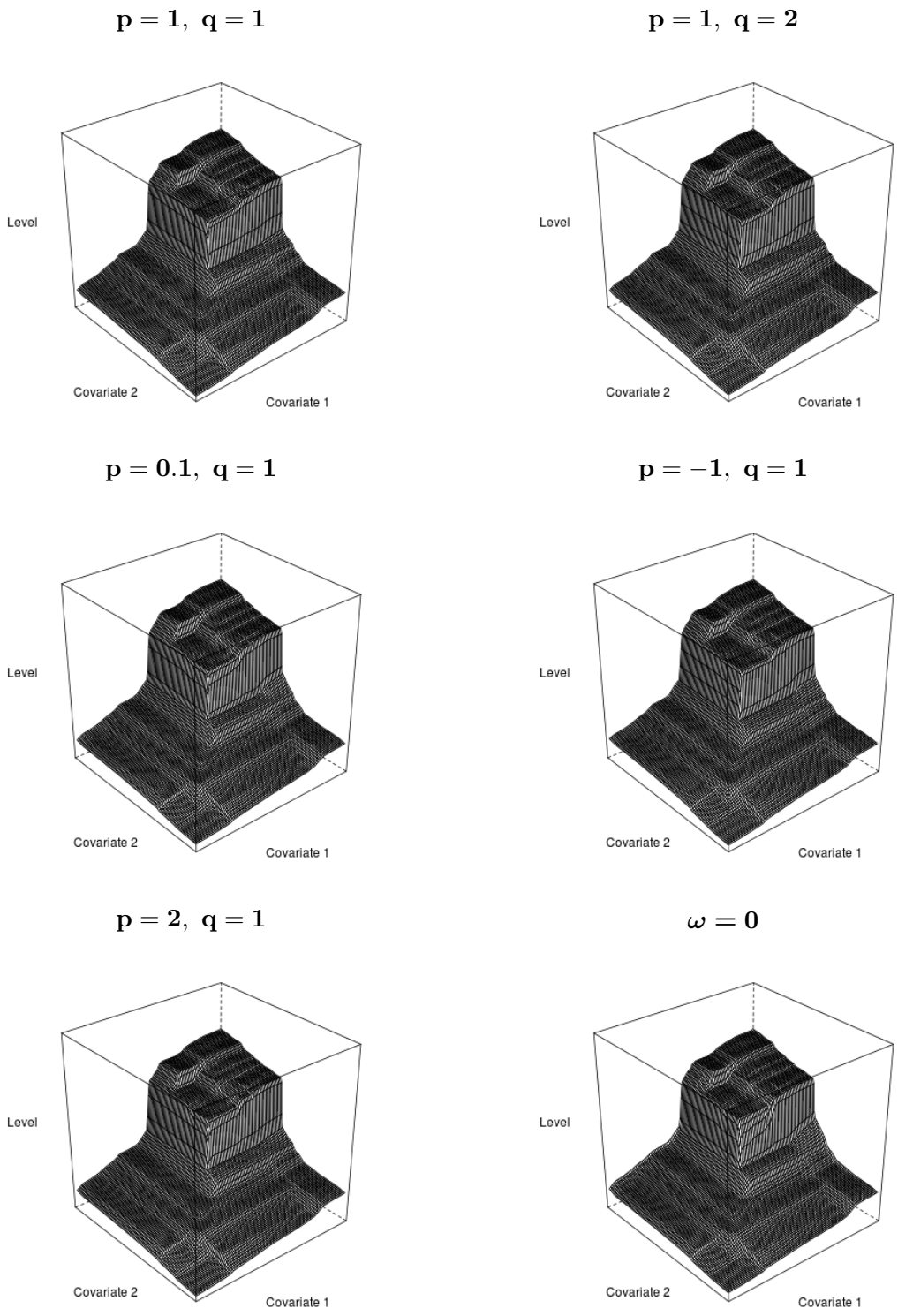


Figure C.5.4: Posterior mean plots for Region 2 in Study 2 for different settings of p and q, and for $\omega = 0$.

Study 3 - Region 1

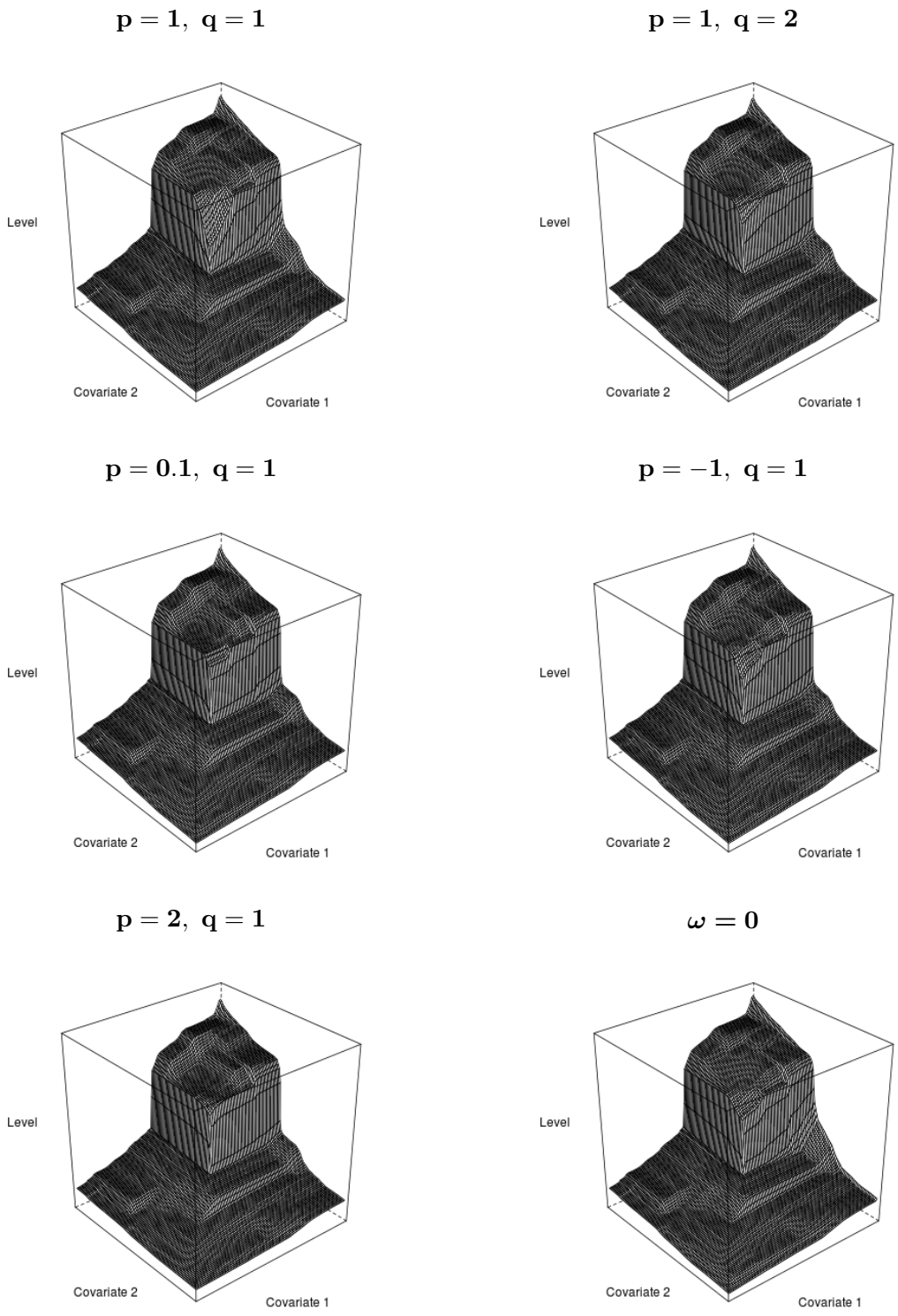


Figure C.5.5: Posterior mean plots for Region 1 in Study 3 for different settings of p and q, and for $\omega = 0$.

Study 3 - Region 2

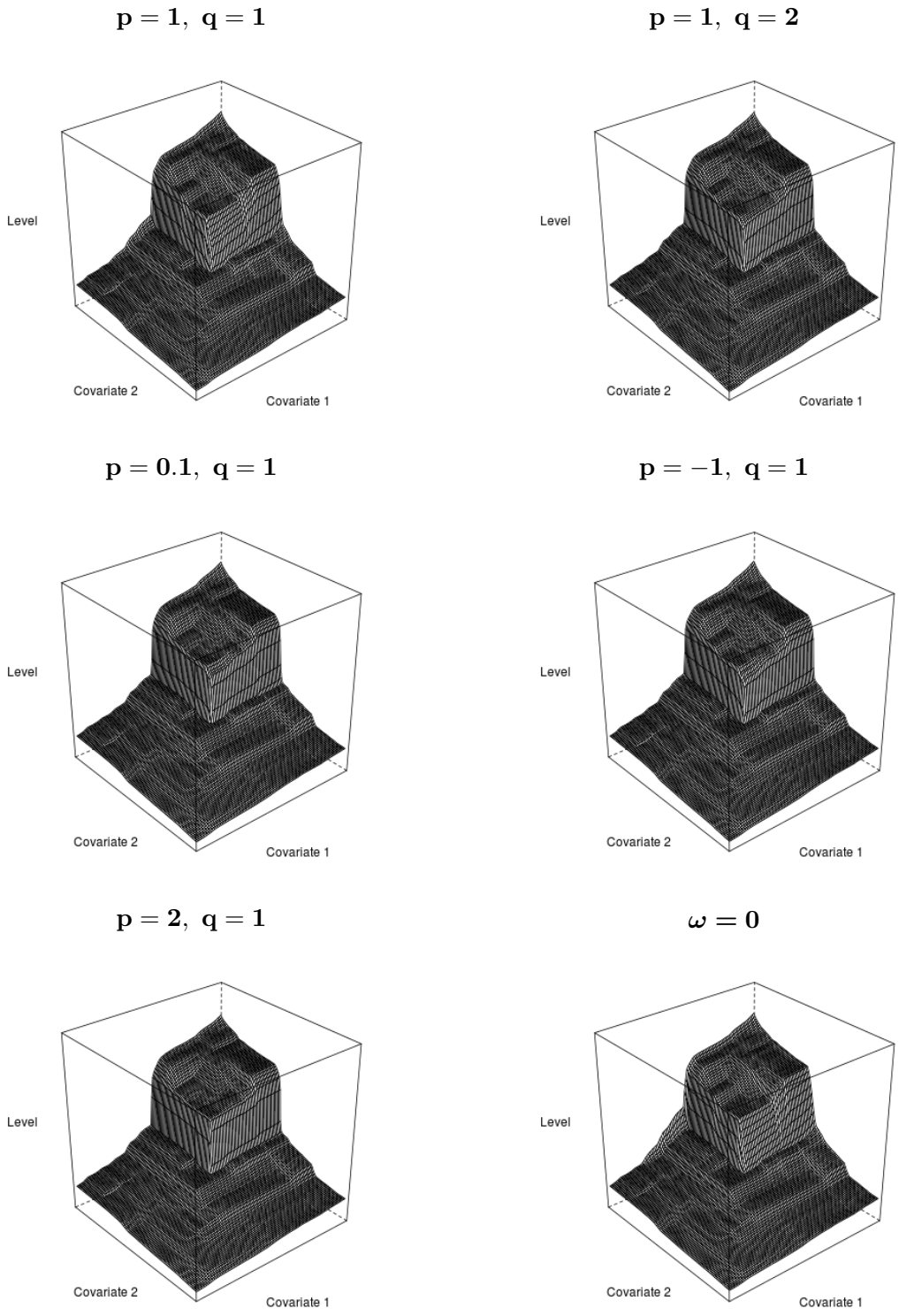


Figure C.5.6: Posterior mean plots for Region 2 in Study 3 for different settings of p and q, and for $\omega = 0$.

C.6 Posterior mean plots for Case Study

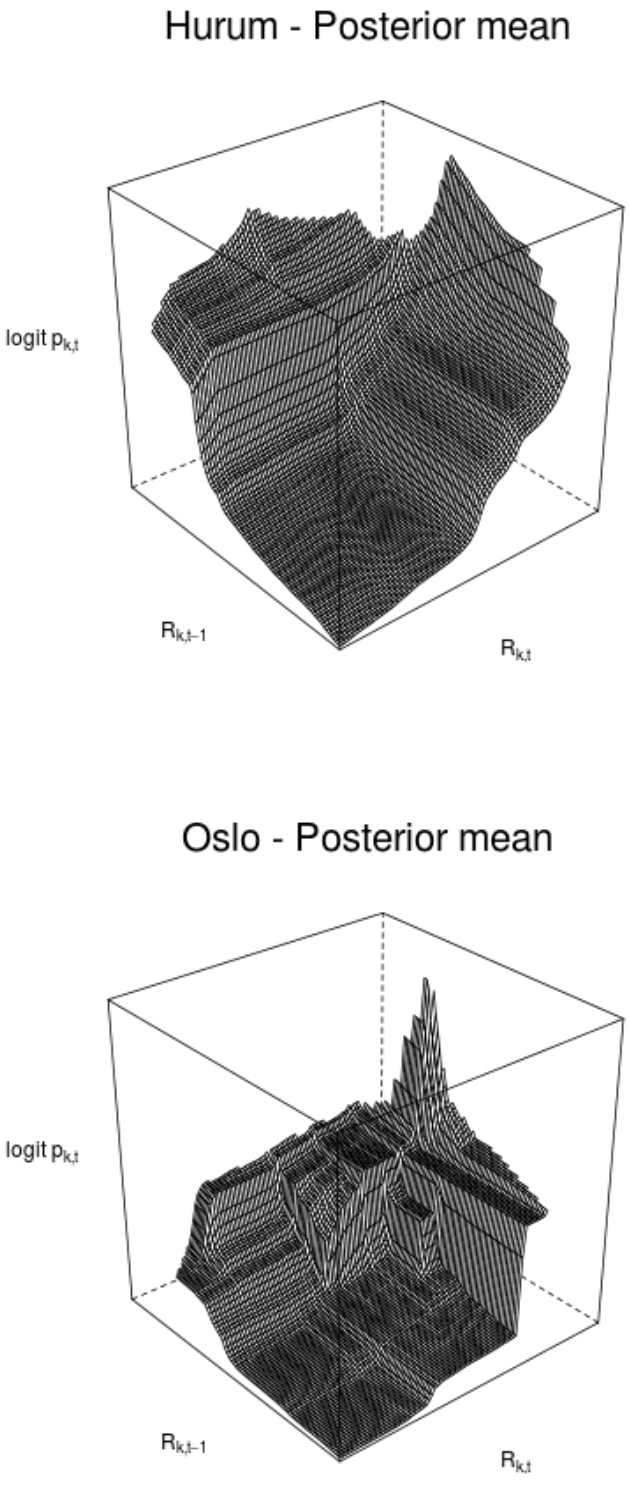


Figure C.6.1: Posterior mean plots for $\omega = \omega_{opt}$ for the Norwegian municipalities of Hurum (top) and Oslo (bottom) considered in Section 4.4.

Appendix D

Supplementary Material Chapter 6

D.1 Threshold-stability of the IGPD

Lemma 1. *Let N be an integer-valued random variable with $N \mid N > u \sim IGPD(\sigma_u, \xi, u)$, $u \in \mathbb{R}$. Then for any threshold $v > u$, the distribution $N \mid N > v$ corresponds to a IGPD with scale $\sigma_u + \xi(\lfloor v \rfloor - \lfloor u \rfloor)$ and shape ξ.*

Proof. We prove the lemma via the survival function $\mathbb{P}(N > n \mid N > v)$, where n is integer. By applying conditional probabilities, $\mathbb{P}(N > n \mid N > v)$ can be expressed by

$$\begin{aligned}\mathbb{P}(N > n \mid N > v) &= \frac{\mathbb{P}(N > n \mid N > u)}{\mathbb{P}(N > v \mid N > u)} \\ &= \frac{\mathbb{P}(H > n - \lfloor u \rfloor)}{\mathbb{P}(H > \lfloor v \rfloor - \lfloor u \rfloor)}\end{aligned}$$

where H is GPD with parameters $\mu = 0$, σ_u and ξ. Next, the two cases $\xi = 0$ and $\xi \neq 0$ are considered separately. For $\xi = 0$, $\mathbb{P}(N > n \mid N > v)$ simplifies to

$$\begin{aligned}\mathbb{P}(N > n \mid N > v) &= \frac{1 - \left[1 - \exp\left(-\frac{n - \lfloor u \rfloor}{\sigma_u}\right)\right]}{1 - \left[1 - \exp\left(-\frac{\lfloor v \rfloor - \lfloor u \rfloor}{\sigma_u}\right)\right]} \\ &= \frac{\exp\left(-\frac{n - \lfloor u \rfloor}{\sigma_u}\right)}{\exp\left(-\frac{\lfloor v \rfloor - \lfloor u \rfloor}{\sigma_u}\right)} \\ &= \exp\left(-\frac{n - \lfloor v \rfloor}{\sigma_u}\right).\end{aligned}$$

This corresponds to the survival function of a IGPD for threshold v with scale σ_u and shape

$\xi = 0$. Next, for $\xi \neq 0$, $\mathbb{P}(N > n \mid N > v)$ can be expressed as

$$\begin{aligned}
\mathbb{P}(N > n \mid N > v) &= \frac{\left[1 + \frac{\xi(\ n\ - \lfloor u \rfloor)}{\sigma_u}\right]_+^{-\frac{1}{\xi}}}{\left[1 + \frac{\xi(\lfloor v \rfloor - \lfloor u \rfloor)}{\sigma_u}\right]_+^{-\frac{1}{\xi}}} \\
&= \left[\frac{\sigma_u + \xi\,(\ n\ - \lfloor u \rfloor)}{\sigma_u + \xi\,(\lfloor v \rfloor - \lfloor u \rfloor)}\right]_+^{-\frac{1}{\xi}} \\
&= \left[\frac{\sigma_u + \xi\,(\ n\ - \lfloor v \rfloor + \lfloor v \rfloor - \lfloor u \rfloor)}{\sigma_u + \xi\,(\lfloor v \rfloor - \lfloor u \rfloor)}\right]_+^{-\frac{1}{\xi}} \\
&= \left[1 + \frac{\xi\,(\ n\ - \lfloor v \rfloor)}{\sigma_u + \xi\,(\lfloor v \rfloor - \lfloor u \rfloor)}\right]_+^{-\frac{1}{\xi}},
\end{aligned}$$

which is the survival function of a IGPD above threshold v with scale parameter $\sigma_u + \xi\,(\lfloor v \rfloor - \lfloor u \rfloor)$ and shape parameter ξ. Consequently, the threshold stability is proven for each threshold $v > u$ and any pair of parameter values $\sigma_u > 0$ and ξ. □

D.2 Threshold-stability of the mixture tail

Lemma 2. *Let N be an integer-valued random variable with $N \mid N > u$ having distribution function*

$$\mathbb{P}(N = n \mid N > u) = p\ \mathbb{P}(Y = n) + (1 - p)\ \mathbb{P}(Z = n)$$

where $Y \sim IGPD(\sigma_u, \xi, u)$ and Z being a truncated Poisson above threshold u with parameter κ. Then for any $v > u$, the random variable $N \mid N > v$, is distributed according to a mixture of an $IGPD(\sigma_u + \xi\,(\lfloor v \rfloor - \lfloor u \rfloor), \xi, v)$ and a truncated Poisson above v with rate parameter κ and mixture probability

$$p_v = \frac{p\ \mathbb{P}(Y > v)}{p\ \mathbb{P}(Y > v) + (1 - p)\ \mathbb{P}(Z > v)}.$$

Proof. Consider any combination $n > v > u$. Then, based on conditional probabilities,

$$\begin{aligned}
&\mathbb{P}(N > n \mid N > v) \\
&\qquad = \frac{\mathbb{P}(N > n \mid N > u)}{\mathbb{P}(N > v \mid N > u)} \\
&\qquad = \frac{p\ \mathbb{P}(Y > n) + (1 - p)\ \mathbb{P}(Z > n)}{p\ \mathbb{P}(Y > v) + (1 - p)\ \mathbb{P}(Z > v)} \\
&\qquad = \frac{p\ \mathbb{P}(Y > n \mid Y > v)\ \mathbb{P}(Y > v) + (1 - p)\ \mathbb{P}(Z > n | Z > v)\ \mathbb{P}(Z > v)}{p\ \mathbb{P}(Y > v) + (1 - p)\ \mathbb{P}(Z > v)}
\end{aligned}$$

By defining

$$p_v = \frac{p\ \mathbb{P}(Y > v)}{p\ \mathbb{P}(Y > v) + (1-p)\ \mathbb{P}(Z > v)},$$

we obtain

$$\mathbb{P}(N > n \mid N > v) = p_v\ \mathbb{P}(Y > n \mid Y > v) + (1 - p_v)\ \mathbb{P}(Z > n \mid Z > v).$$

Based on the threshold-stability in Appendix D.1, $Y \mid Y > v \sim \text{IGPD}(\sigma_u + \xi(\lfloor v \rfloor - \lfloor u \rfloor), \xi, v)$. Further, $Z \mid Z > v$ is a truncated Poisson above v with rate κ. Hence, $N \mid N > v$ is distributed according to a mixture of an IGPD and a truncated Poisson. □

D.3 Details of the MCMC algorithm

Let $\mathcal{D} = \{(\widetilde{n}_i, \widetilde{\mathbf{x}}_i)\,,\ i = 1, \ldots, m\}$ denote the set of observed claim numbers and covariates effects. Further, a latent binary variable v_i is introduced for each observation $\widetilde{n}_i$ which is defined by

$$v_i = \begin{cases} 1 & \text{if } \widetilde{n}_i \text{ is a realization from the distribution } \widetilde{Y} \\ 0 & \text{otherwise.} \end{cases}$$

We set a $Beta(1,1)$ prior on the mixing probability p and an improper prior on the remaining parameters, $\pi(\boldsymbol{\beta}, \boldsymbol{\delta}, \xi, \kappa) \propto 1$. Hence, the posterior distribution $\pi(p, \boldsymbol{\beta}, \xi, \boldsymbol{\delta}, \kappa, v_1, \ldots, v_I|\mathcal{D})$ is proportional to

$$\prod_{i=1}^{m} \left\{ \left[p\ \mathbb{P}\left(\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}, \xi, \boldsymbol{\delta}, \widetilde{\mathbf{x}}\right)\right]^{v_i} \ \left[(1-p)\ \mathbb{P}\left(\widetilde{Z} = \widetilde{n}_i \mid \kappa\right)\right]^{1-v_i} \right\} \ \pi(p)$$

Realizations from this posterior distribution are sampled by a Metropolis-within-Gibbs algorithm which runs for a fixed number of iterations J. Let $p^{(0)}, \boldsymbol{\beta}^{(0)}, \xi^{(0)}, \boldsymbol{\delta}^{(0)}$ and $\lambda^{(0)}$ denote the initial parameter values. The update procedure for all parameters within one iteration step $j = 1, \ldots, J$ is as follows:

At the start of iteration step j, the latent variables $v_1^{(j)}, \ldots, v_m^{(j)}$ are sampled from a Bernoulli distribution

$$v_i^{(j)} \sim \text{Bernoulli}\left[w_i^{(j)}\right].$$

The probability of observation $\widetilde{n}_i$ being sampled from the covariate-driven component $\widetilde{Y}$, $w_i^{(j)}$,

is given by

$$w_i^{(j)} = \frac{p^{(j-1)}\ \mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j-1)}, \xi^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}{p^{(j-1)}\ \mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j-1)}, \xi^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right] + \left[1 - p^{(j-1)}\right]\ \mathbb{P}\left[\widetilde{Z} = \widetilde{n}_i\right]}.$$

Since we placed a conjugate Beta prior on p, the parameter value is updated by sampling from the full-conditional Beta posterior

$$p^{(j)} \sim \text{Beta}\left(\sum_{i=1}^{I} v_i^{(j)} + 1,\ I - \sum_{i=1}^{I} v_i^{(j)} + 1\right).$$

The model parameters $\boldsymbol{\beta}$, ξ and $\boldsymbol{\delta}$ are updated separately via Random-Walk-Metropolis with Gaussian proposal. For the covariate effects $\boldsymbol{\beta}$, the proposal $\boldsymbol{\beta}^*$ is accepted with probability

$$\min\left\{1, \prod_{v_i^{(j)}=1,\ \widetilde{n}_i > u} \frac{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^*, \xi^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j-1)}, \xi^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}\right\},$$

whilst the proposal ξ^* has acceptance probability

$$\min\left\{1, \prod_{v_i^{(j)}=1,\ \widetilde{n}_i > u} \frac{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j)}, \xi^*, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j)}, \xi^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}\right\}.$$

Note, the likelihood needs only to be evaluated for the observations with latent variable $v_i^{(j)} = 1$ and the number of observations $\widetilde{n}_i$ greater than the threshold. Next, the covariate effects for the rate parameter κ are updated. Here, the likelihood has to be evaluated for all observations with $v_i^{(j)} = 1$ as $\boldsymbol{\delta}$ effects the threshold exceedance model. The acceptance ratio is thus given by

$$\min\left\{1, \prod_{v_i^{(j)}=1} \frac{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j)}, \xi^{(j)}, \boldsymbol{\delta}^*, \widetilde{\mathbf{x}}_i\right]}{\mathbb{P}\left[\widetilde{Y} = \widetilde{n}_i \mid \boldsymbol{\beta}^{(j)}, \xi^{(j)}, \boldsymbol{\delta}^{(j-1)}, \widetilde{\mathbf{x}}_i\right]}\right\}.$$

Finally, the rate parameter κ is updated via an independence sampler with uniform proposal distribution. The acceptance probability then yields to

$$\min\left\{1, \prod_{v_i^{(j)}=0} \frac{\mathbb{P}\left[\widetilde{Z} = \widetilde{n}_i \mid \kappa^*\right]}{\mathbb{P}\left[\widetilde{Z} = \widetilde{n}_i \mid \kappa^{(j-1)}\right]}\right\}.$$

D.4 Estimates for Leaving Highest Observation Out

Table D.4.1: Posterior mean estimates, lower 5% quantile ($q_{0.05}$) and upper 5% quantile ($q_{0.95}$) of the model parameters for the municipalities of Oslo, Bærum and Bergen with thresholds $u_k = 4, 2$ and 4, respectively, when leaving the highest claim observation out.

City	Statistic	p	β_0	β_1	β_2	β_3	ξ	δ_0	δ_1	δ_2	δ_3	κ
Oslo	Mean	0.89	0.24	0.07	0.22	0.75	-0.30	-0.17	0.42	0.32	0.71	2.20
	$q_{0.05}$	0.83	-0.37	-0.12	0.10	0.43	-0.69	-0.31	0.29	0.22	0.57	1.66
	$q_{0.95}$	0.95	0.91	0.25	0.33	1.04	0.10	-0.04	0.56	0.42	0.86	2.89
Bærum	Mean	0.81	-1.19	0.25	0.26	0.81	0.11	-0.94	0.49	0.35	0.97	1.06
	$q_{0.05}$	0.61	-2.20	0.08	0.10	0.31	-0.30	-1.65	0.18	0.21	0.57	0.67
	$q_{0.95}$	0.95	-0.37	0.44	0.44	1.32	0.54	-0.57	1.05	0.58	1.68	1.65
Bergen	Mean	0.88	-0.46	0.01	0.19	0.26	0.58	-0.53	0.14	0.14	0.41	1.20
	$q_{0.05}$	0.79	-1.51	-0.15	0.04	-0.04	0.14	-0.73	0.09	0.07	0.34	0.66
	$q_{0.95}$	0.95	0.48	0.17	0.34	0.55	1.23	-0.36	0.20	0.21	0.49	1.89

Bibliography

Anderson, C. W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7(1):99–113.

Anderson, C. W. (1980). Local limit theorems for the maxima of discrete random variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 88, pages 161–165. Cambridge Univ Press.

Anderson, C. W., Coles, S. G., and Hüsler, J. (1997). Maxima of Poisson-like variables and related triangular arrays. *The Annals of Applied Probability*, 7(4):953–971.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050.

Assunção, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14(5):453–473.

Athreya, J. S. and Sethuraman, S. (2001). On the asymptotics of discrete order statistics. *Statistics & Probability Letters*, 54(3):243–249.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647.

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association*, 84(405):289–294.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Bank, M. and Wiesner, R. (2011). Determinants of weather derivatives usage in the Austrian winter tourism industry. *Tourism Management*, 32(1):62–68.

Barlow, R. E. and Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147.

Barnett, B. J. and Mahul, O. (2007). Weather index insurance for agriculture and rural areas in lower-income countries. *American Journal of Agricultural Economics*, 89(5):1241–1247.

Barthel, F. and Neumayer, E. (2012). A trend analysis of normalized insured damage from natural disasters. *Climatic Change*, 113(2):215–237.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. *JAMA*, 292(19):2372–2378.

Benth, F. E. and Šaltytė Benth, J. (2011). Weather derivatives and stochastic modelling of temperature. *International Journal of Stochastic Analysis*, 2011.

Bergersen, L. C., Tharmaratnam, K., and Glad, I. K. (2014). Monotone splines lasso. *Computational Statistics & Data Analysis*, 77:336 – 351.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Best, M. J. and Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439.

Best, M. J., Chakravarti, N., and Ubhaya, V. A. (2000). Minimizing separable convex functions subject to simple chain constraints. *SIAM Journal on Optimization*, 10(3):658–672.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics*, 65(1):198–205.

Bornkamp, B., Ickstadt, K., and Dunson, D. B. (2010). Stochastically ordered multiple regression. *Biostatistics*, 11(3):419–431.

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6(1):25–47.

Botzen, W. J. and van den Bergh, J. C. (2012). Risk attitudes to low-probability climate change risks: WTP for flood insurance. *Journal of Economic Behavior & Organization*, 82(1):151–166.

Botzen, W. W. and Van Den Bergh, J. C. (2008). Insurance against climate change and flooding in the Netherlands: present, future, and comparison with other countries. *Risk Analysis*, 28(2):413–426.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations*. Oxford University Press.

Bowman, A., Jones, M., and Gijbels, I. (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7(4):489–500.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):607–616.

Brunk, H. D., Ewing, G. M., and Utz, W. R. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics*, 7(1):833–847.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression-modelling spatial non-stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(3):431–443.

Bühlmann, H. (1980). An economic premium principle. *ASTIN Bulletin*, 11:52–60.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*. Springer.

Cahill, M. and Mulligan, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review*, 25(2):174–193.

Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines. *Journal of the American Statistical Association*, 102(480):1158–1171.

Carreau, J. and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76.

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2016). An extreme value approach for modeling operational risk losses depending on covariates. *The Journal of Risk and Insurance*, 83(3):735–776.

Chen, Y. and Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):729–754.

Cheng, G. (2009). Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference*, 139(6):1980 – 1991.

Cheng, G., Zhao, Y., and Li, B. (2012). Empirical likelihood inferences for the semiparametric additive isotonic regression. *Journal of Multivariate Analysis*, 112:172 – 182.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

Coles, S. G. and Pauli, F. (2001). Extremal limit laws for a class of bivariate poisson vectors. *Statistics & Probability Letters*, 54(4):373–379.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):377–392.

Congdon, P. (2003). Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes. *Journal of Geographical Systems*, 5(2):161–184.

Congdon, P. (2006). A model for non-parametric spatially varying regression effects. *Computational Statistics & Data Analysis*, 50(2):422 – 445.

Costain, D. A. (2009). Bayesian partitioning for modeling and mapping spatial casecontrol data. *Biometrics*, 65(4):1123–1132.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, N. A. C. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Cunningham, J. P. (1982). Multiple monotone regression. *Psychological Bulletin*, 92(3):791 – 800.

Curry, H. B. and Schoenberg, I. J. (1966). On Pólya frequency functions IV: the fundamental spline functions and their limits. *Journal d'analyse mathématique*, 17(1):71–107.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–442.

De Boer, W. J., Den Besten, P. J., and Ter Braak, C. F. (2002). Statistical analysis of sediment toxicity by additive monotone regression splines. *Ecotoxicology*, 11(6):435–450.

De Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5).

De Melo Mendes, B. V. and Lopes, H. F. (2004). Data driven estimates for mixtures. *Computational Statistics & Data Analysis*, 47(3):583–598.

Dickson, D. C. M. (2005). *Insurance Risk and Ruin*. Cambridge University Press.

Dickson, G. C. A. and Steele, J. T. (1986). *Introduction to Insurance*. Pitman, London, 2nd edition.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-temporal Point Patterns*. CRC Press.

Dkengne, P. S., Eckert, N., and Naveau, P. (2016). A limiting distribution for maxima of discrete stationary triangular arrays with an application to risk due to avalanches. *Extremes*, 19(1):25–40.

Dorfleitner, G. and Wimmer, M. (2010). The pricing of temperature futures at the Chicago Mercantile Exchange. *Journal of Banking & Finance*, 34(6):1360–1370.

Downton, M. W. and Pielke, R. A. J. (2005). How accurate are disaster loss data? The case of U.S. flood damage. *Natural Hazards*, 35(2):211–228.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825.

Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470):618–627.

Economou, T., Stephenson, D. B., and Ferro, C. A. (2014). Spatio-temporal modelling of extreme storms. *The Annals of Applied Statistics*, 8(4):2223–2246.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling Extremal Events: for Insurance and Finance*, volume 33 of *Stochastic Modelling and Applied Probability*. Springer.

Eshita, N. (1977). An estimation of claims distribution. *ASTIN Bulletin*, 9(1-2):111–118.

Fang, Z. and Meinshausen, N. (2012). LASSO isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, 21(1):72–91.

Farah, M., Kottas, A., and Morris, R. D. (2013). An application of semiparametric Bayesian isotonic regression to the study of radiation effects in spaceborne microelectronics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):3–24.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235.

Fuchs, A. and Wolff, H. (2011). Concept and unintended consequences of weather index insurance: the case of Mexico. *American Journal of Agricultural Economics*, 93(2):505–511.

Furman, E. and Zitikis, R. (2008). Weighted premium calculation principles. *Insurance: Mathematics and Economics*, 42(1):459–465.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.

Gelfand, A. E. and Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika*, 78(3):657–666.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

Gerber, H. U. (1974). On additive premium calculation principles. *Astin Bulletin*, 7(3):215–222.

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *The Annals of Statistics*, 28(4):1054–1082.

Goovaerts, M. J. and Haezendonck, J. (1984). *Insurance Premiums: Theory and Applications*. North-Holland.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Guttorp, P. (1991). Spatial statistics in ecology. In National Research Council, editor, *Spatial Statistics and Digital Image Analysis*, pages 129–146. The National Academy Press, Washington, DC.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89(426):368–378.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Haug, O., Dimakos, X. K., Vårdal, J. F., Aldrin, M., and Meze-Hausken, E. (2011). Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, 2011(1):1–20.

Hawn, M. T., Graham, L. A., Richman, J. S., Itani, K. F., Henderson, W. G., and Maddox, T. M. (2013). Risk of major adverse cardiac events following noncardiac surgery in patients with coronary stents. *JAMA*, 310(14):1462–1472.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93(442):643–650.

Heikkinen, J. (2003). Trans-dimensional Bayesian non-parametrics with spatial point processes. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 203–207. Oxford University Press.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25(3):435–450.

Heimfarth, L. E. and Musshoff, O. (2011). Weather index-based insurances for farmers in the North China Plain: An analysis of risk reduction potential and basis risk. *Agricultural Finance Review*, 71(2):218–239.

Hermanussen, M., Thiel, C., von Büren, E., de Lama, M., Romero, A., Ruiz, C., Burmeister, J., and Tresguerres, J. (1998). Micro and macro perspectives in auxology: findings and considerations upon the variability of short term and individual growth and the stability of population derived parameters. *Annals of Human Biology*, 25(4):359–385.

Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59(2):317–322.

Holmes, C. C. and Heard, N. A. (2003). Generalized monotonic regression using random change points. *Statistics in Medicine*, 22(4):623–638.

Holmes, J. and Moriarty, W. (1999). Application of the generalized Pareto distribution to extreme value analysis in wind engineering. *Journal of Wind Engineering and Industrial Aerodynamics*, 83(1):1–10.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942 – 4956.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

Jenkins, G. J., Perry, M. C., and Prior, M. J. (2008). The climate of the United Kingdom and recent trends. Met Office Hadley Centre, Exeter, UK.

Jewson, S., Brix, A., and Ziehmann, C. (2005). *Weather Derivative Valuation: The Meteorological, Statistical and Mathematical Foundations*. Cambridge University Press.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Kaluszka, M., Laeven, R. J. A., and Okolewski, A. (2012). A note on weighted premium calculations priniciples. *Insurance: Mathematics and Economics*, 51(2):379–381.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46(4):1071–1085.

Keshvari, A. and Kuosmanen, T. (2013). Stochastic non-convex envelopment of data: Applying isotonic regression to frontier estimation. *European Journal of Operational Research*, 231(2):481 – 491.

Khattree, R., Schmidt, D., and Schochetman, I. (1999). An isotonic regression problem for infinite-dimensional parameters. *Journal of Optimization Theory and Applications*, 103(2):359–384.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893.

Knabb, R. D., Rhome, J. R., and Brown, D. P. (2005). Tropical Cyclone Report: Hurricane Katrina: 23-30 August 2005. Technical report, National Hurricane Center.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17):2555–2567.

Knorr-Held, L. (2003). Some remarks on Gaussian Markov random field models for disease mapping. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 260–264. Oxford University Press.

Kottas, A., Branco, M. D., and Gelfand, A. E. (2002). A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics*, 58(3):593–600.

Kubilay, A., Derome, D., Blocken, B., and Carmeliet, J. (2013). CFD simulation and validation

of wind-driven rain on a building facade with an Eulerian multiphase model. *Building and Environment*, 61:69–81.

Kyng, R., Rao, A., and Sachdeva, S. (2015). Fast, provable algorithms for isotonic regression in all L_p-norms. In *Advances in Neural Information Processing Systems*, pages 2719–2727.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lavine, M. and Mockus, A. (1995). A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference*, 46(2):235 –248.

Lawson, A. B. (2013). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. CRC press.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89.

Lee, D. (2013). CARBayes: An `R` package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13).

Lee, D. and Lawson, A. (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *Annals of Applied Statistics*, 10(3):1427–1446.

Leitenstorfer, F. and Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, 8(3):654–673.

Leroux, B. G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–192. Springer.

Li, Y., Cai, W., and Campbell, E. (2005). Statistical modeling of extreme rainfall in southwest western Australia. *Journal of Climate*, 18(6):852–863.

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17(3):909–927.

Litterman, R. (2011). Pricing climate change risk appropriately. *Financial Analysts Journal*, 67(5):4–10.

Liu, M.-H. and Ubhaya, V. A. (1997). Integer isotone optimization. *SIAM Journal on Optimization*, 7(4):1152–1159.

Loayza, N. V., Olaberría, E., Rigolini, J., and Christiaensen, L. (2012). Natural disasters and growth: Going beyond the averages. *World Development*, 40(7):1317–1336.

Luss, R., Rosset, R., and Shahar, S. (2012). Efficient regularized isotonic regression with application to gene-gene interaction search. *The Annals of Applied Statistics*, 6(1):253–283.

Luss, R. and Rosset, S. (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics*, 23(1):192–210.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

Malinovski, V. K. (2008). Adaptive control strategies and dependence of finite time ruin on the premium loading. *Insurance: Mathematics and Economics*, 42(1):81–94.

Maxwell, W. L. and Muckstadt, J. A. (1985). Establishing consistent and realistic reorder intervals in production-distribution systems. *Operations Research*, 33(6):1316–1341.

McBride, S. J., Williams, R. W., and Creason, J. (2007). Bayesian hierarchical modeling of personal exposure to particulate matter. *Atmospheric Environment*, 41(29):6143–6155.

McCormick, W. P. and Park, Y. S. (1992). Asymptotic analysis of extremes from autoregressive negative binomial processes. *Journal of Applied Probability*, 29(4):904–920.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapmann and Hall/CRC, 2nd edition.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015). Space–time smoothing of complex survey data: Small area estimation for child mortality. *The Annals of Applied Statistics*, 9(4):1889–1905.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miles, R. E. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 46(3/4):317–327.

Miller, S., Muir-Wood, R., and Boissonade, A. (2008). An exploration of trends in normalized weather-related catastrophe losses. In Diaz, H. F. and Murnane, R. J., editors, *Climate Extremes and Society*, pages 225–247. Cambridge University Press.

Mills, E. (2005). Insurance industry in a climate of change. *Science*, 309(5737):1040–1044.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Moore, K. S. and Young, V. R. (2003). Pricing equity-linked pure endowments via the principle of equivalent utility. *Insurance Mathematics and Economics*, 33(3):497–516.

Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H. O., and Binks, K. (2000). Additive isotonic regression models in epidemiology. *Statistics in Medicine*, 19(6):849–859.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Nadarajah, S. and Mitov, K. (2002). Asymptotics of maxima of discrete random variables. *Extremes*, 5(3):287–294.

Nadarajah, S. and Mitov, K. (2004). Extremal limit laws for discrete random variables. *Journal of Mathematical Sciences*, 122(4):3404–3415.

Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406.

Neelon, B., Ghosh, P., and Loebs, P. F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):389–413.

Neumayer, E. and Barthel, F. (2011). Normalizing economic loss from natural disasters: a global analysis. *Global Environmental Change*, 21(1):13–24.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

Obozinski, G., Lanckriet, G., Grant, C., Jordan, M. I., and Noble, W. S. (2008). Consistent probabilistic outputs for protein function prediction. *Genome Biology*, 9(1).

Orskaug, E., Scheel, I., Frigessi, A., Guttorp, P., Haugen, J. E., Tveito, O. E., and Haug, O. (2011). Evaluation of a dynamic downscaling of precipitation over the Norwegian mainland. *Tellus A*, 63(4):746–756.

Ozenne, B., Subtil, F., Østergaard, L., and Maucort-Boulch, D. (2015). Spatially regularized mixture model for lesion segmentation with application to stroke patients. *Biostatistics*, 16(3):580–595.

Paiva, T., Chakraborty, A., Reiter, J., and Gelfand, A. E. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine*, 33(11):1928–1945.

Paulson, N. D., Hart, C. E., and Hayes, D. J. (2010). A spatial Bayesian approach to weather derivatives. *Agricultural Finance Review*, 70(1):79–96.

Penttinen, A., Stoyan, D., and Henttonen, H. M. (1992). Marked point processes in forest statistics. *Forest Science*, 38(4):806–824.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Pielke, R. A. J. and Landsea, C. W. (1998). Normalized hurrican damages in the United States: 1925-95. *Weather and Forecasting*, 13(3):621–631.

Prahl, B. F., Rybiski, D., Kropp, J. P., Burghoff, O., and Held, H. (2012). Applying stochastic small-scale damage functions to German winter storm. *Geophsyical Research Letters*, 39(6).

Prieto, F., Gómez-Déniz, E., and Sarabia, J. M. (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention*, 71:38–49.

Qian, S. (1992). Minimum lower sets algorithms for isotonic regression. *Statistics & Probability Letters*, 15(1):31 – 35.

Qian, S. S., Craig, J. K., Baustian, M. M., and Rabalais, N. N. (2009). A Bayesian hierarchical modeling approach for analyzing observational data from marine ecological studies. *Marine Pollution Bulletin*, 58(12):1916–1921.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer series in statistics. Springer, New York, 2nd edition.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer.

Richards, T. J., Manfredo, M. R., and Sanders, D. R. (2004). Pricing weather derivatives. *American Journal of Agricultural Economics*, 86(4):1005–1017.

Ricker-Gilbert, J., Jayne, T. S., and Chirwa, E. (2011). Subsidies and crowding out: A double-hurdle model of fertilizer demand in Malawi. *American Journal of Agricultural Economics*.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *AISTATS*, volume 9, pages 645–652.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Robertson, T. and Wright, F. T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *The Annals of Statistics*, pages 645–651.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York, 1st edition.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim: Two `R` packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1).

Royston, P. (2000). A useful monotonic non-linear model with applications in medicine and epidemiology. *Statistics in Medicine*, 19(15):2053–2066.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

Rueda, C. (2013). Degrees of freedom and model selection in semiparametric additive monotone regression. *Journal of Multivariate Analysis*, 117:88 – 99.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Russell, B. T., Cooley, D., Porter, W. C., Reich, B. J., and Heald, C. L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events. *Annals of Applied Statistics*, 10(3):1673–1698.

Saarela, O. and Arjas, E. (2011). A method for Bayesian monotonic multiple regression. *Scandinavian Journal of Statistics*, 38(3):499–513.

Sanders, C. H. and Phillipson, M. C. (2003). UK adaption strategy and technical measures: the impacts of climate change on buildings. *Building Research & Information*, 31(3–4):210–221.

Sasabuchi, S., Inutsuka, M., and Kulatunga, D. D. S. (1992). An algorithm for computing multivariate isotonic regression. *Hiroshima Mathematical Journal*, 22(3):551–560.

Sasabuchi, S., M., I., and Kulatunga, D. D. S. (1983). A multivariate version of isotonic regression. *Biometrika*, 70(2):465–472.

Schabenberger, O. and Gotway, C. A. (2004). *Statistical Methods for Spatial Data Analysis*. CRC press.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2011). A Bayesian hierarchical model with spatial variable selection: the effect of weather

on insurance claims. Derivation of distributions and MCMC sampling schemes. Statistical Research Report ISSN 0806-3842 no. 2, Department of Mathematics, University of Oslo. available from `http://urn.nb.no/URN:NBN:no-28752`.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2013). A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100.

Schell, M. J. and Singh, B. (1997). The reduced monotonic regression method. *Journal of the American Statistical Association*, 92(437):128–135.

Schuster, S. S., Blong, R. J., and McAneney, K. J. (2006). Relationship between radar-derived hail kinetic energy and damage to insured buildings for severe hailstorms in Eastern Australia. *Atmospheric Research*, 81(3):215–235.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scott, J. G., Shively, T. S., and Walker, S. G. (2015). Nonparametric Bayesian testing for monotonicity. *Biometrika*, 102(3):617–630.

Shimura, T. (2012). Discretization of distributions in the maximum domain of attraction. *Extremes*, 15(3):299–317.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

Shively, T. S., Walker, S. G., and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161(2):166 – 181.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4(3):283–291.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431.

Smith, R. L. and Goodman, D. J. (2000). Bayesian risk analysis. *Extremes and Integrated Risk Management*, pages 235–251.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spouge, J., Wan, H., and Wilbur, W. (2003). Least squares isotonic regression in two dimensions. *Journal of Optimization Theory and Applications*, 117(3):585–605.

Stephenson, A. G. and Tawn, J. A. (2013). Determining the best track performances of all time using a conceptual population model for athletics records. *Journal of Quantitative Analysis in Sports*, 9(1):67–76.

Stoppa, A. and Hess, U. (2003). Design and use of weather derivatives in agricultural policies: the case of rainfall index insurances in Morocco. In *Internationale Conference: Agricultural policy reform and the WTO: Where are we heading?*

Stout, Q. F. (2012). Strict L_∞ isotonic regression. *Journal of Optimization Theory and Applications*, 152(1):121–135.

Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica*, 71(2):450–470.

Taib, C. M. I. C. and Benth, F. E. (2012). Pricing of temperature index insurance. *Review of Development Finance*, 2(1):22–31.

Tancredi, A., Anderson, C. W., and O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9(2):87–106.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, 53(1):54–61.

Tol, R. S. J. (1998). Climate change and insurance: a critical appraisal. *Energy Policy*, 26(3):257–262.

Turvey, C. G., Weersink, A., and Chang, C. S. (2006). Pricing weather insurance with a random strike price: the Ontario ice-wine harvest. *American Journal of Agricultural Economics*, 88(3):696–709.

Tutz, G. and Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics*, 16(1):165–188.

Špička, J. (2011). Weather derivative design in agriculture - a case study of barley in the Southern Moravia Region. *AGRIS on-line Papers in Economics and Informatics*, 3(3):53–59.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wadsworth, J. L. and Tawn, J. A. (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):543–567.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Waller, L. A. and Carlin, B. (2010). Disease mapping. In Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, pages 217–243. Chapman & Hall.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons.

Waller, L. A., Zhu, L., Gotway, C. A., Gorman, D. M., and Gruenewald, P. J. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21(5):573–588.

Wang, L. and Dunson, D. B. (2011). Bayesian isotonic density regression. *Biometrika*, 98(3):537–551.

Wang, L. and Xue, L. (2015). Constrained polynomial spline estimation of monotone additive models. *Journal of Statistical Planning and Inference*, 167:27 – 40.

Wang, S. S. (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin*, 26(1):71–92.

Wang, W. and Small, D. S. (2015). Monotone B-Spline smoothing for a generalized linear model response. *The American Statistician*, 69(1):28–33.

Williams, C. A. J., Smith, M. L., and Young, P. C. (1995). *Risk Management and Insurance*. McGraw-Hill, New York, 7th edition.

Wilson, A., Reif, D. M., and Reich, B. J. (2014). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics*, 70(1):237–246.

Wood, S. N. (2006). *Generalized Additive Models : An Introduction with R*. Chapman & Hall/CRC, 1st edition.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Xiao, Y. (2011). Pricing a contract of linking home reversion plan and long-term care insurance via the principle of equivalent utility. *Quality & Quantity*, 45(2):465–475.

Yeganova, L. and Wilbur, W. J. (2009). Isotonic regression under Lipschitz constraint. *Journal of Optimization Theory and Applications*, 141(2):429–443.

Young, V. R. (2006). Premium principles. In Teugels, J. and Sundt, B., editors, *Encyclopedia of Actuarial Science*. Wiley.

Yu, K. (2014). On partial linear additive isotonic regression. *Journal of the Korean Statistical Society*, 43(1):11 – 17.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques - Essays in Honour of Bruno de Finetti*, pages 233–243. Elsevier.

Zhang, L. and Shi, H. (2004). Local modeling of tree growth by geographically weighted regression. *Forest Science*, 50(2):225–244.

Zhou, M., Chen, H., Paisley, J., Ren, L., Li, L., Xing, Z., Dunson, D., Sapiro, G., and Carin, L. (2012). Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21(1):130–144.