# Spatial Statistical Modelling of Insurance Claim Frequency

## Using Markov Chain Monte Carlo based inference with Riemannian Langevin diffusion and continuous spatial dependence

Daniel Faller

Master's thesis
2022:E7

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Visiting address: Sölvegatan 18, 223 62 Lund. Postal address: Box 118, 221 00 Lund.

## Abstract

In this thesis a fully Bayesian hierarchical model that estimates the number of aggregated insurance claims per year for non-life insurances is constructed using Markov chain Monte Carlo based inference with Riemannian Langevin diffusion. Some versions of the model incorporate a spatial effect, viewed as the relative spatial insurance risk that originates from a policyholder's geographical location and where the relative spatial insurance risk is modelled as a continuous spatial field. It is shown that the inclusion of a spatial effect derived from a Gaussian Markov random field with Matérn covariance in a generalised linear mixed model (GLMM) has better predictive performance regarding the number of aggregated claims in an insurance portfolio compared to GLMMs that lack such a spatial effect.

**Keywords**: Insurance risk, claim frequency, Markov chain Monte Carlo (MCMC), Riemann manifold Metropolis adjusted Langevin algorithm (MMALA), spatial statistics, Gaussian Markov random field (GMRF), preconditioned Crank Nicolson Langevin algorithm (pCNL), Gibbs sampling, Bayesian hierarchical modelling, high dimensional, shrinkage prior, horseshoe prior, regularisation.

# Acknowledgements

This master's thesis in mathematical statistics was conducted during the fall semester of 2021, as the final part of a Master of Science in Engineering, Risk Management and Safety Engineering. The thesis equals 30 ECTS and was completed in January 2022. The thesis was supervised by the LTH Faculty of Engineering, Centre for Mathematical Sciences at Lund University, Sweden.

The trajectory of the last 6 months working with this project share many similarities with the trajectories that are modelled in it. These similarities are many small nudges from one state to another more meaningful state. For many nudges I thank my family and friends. For one specific nudge I thank Prof. Magnus Wiktorsson who directed me to my supervisor Prof. Johan Lindström, and to Johan; it has been a privilege having you as my supervisor, thank you for all your guidance.

Stugan, January 2022

Daniel Faller

Daniel Faller

# Modelling Insurance Risk with Continuous Spatial Dependence

**The geographic rating factor used to determine the insurance risk originating from the geographical location of a policyholder can be modelled with a continuous spatial dependence. Continuous models allow the geographic risks to vary in larger pricing areas which is not the case with constant, or discrete models. Constant geographic risks can cause the geographic risks of larger pricing areas to have a greater influence on neighbouring pricing areas than feasible.**

The purpose of insurances is to protect against financial loss. To be insured by an insurance, the policyholder has to pay a premium. The price of the premium needs to be proportionate to the size of the future and uncertain losses of the policyholder. These losses may or may not be financial, but they need to be reducible to financial terms. To determine the future and uncertain losses for a specific policyholder, an insurance company looks at the individual traits of the policyholder and compares these traits with the traits of policyholders that have incurred historical losses. The insurance company then assumes that these traits are indicative for future losses. How indicative certain traits are, can be quantified and used to estimate future losses with probabilities. These probabilities are needed to define the risk premium which is based on the number of times during a specific period a policyholder is expected to suffer a loss together with the expected sizes of these losses. With the risk premium it is possible to determine a proportionate price for an insurance policy. One of the traits an insurance company can look at is called the geographic factor, or spatial effect. The spatial effect indicates how much of a policyholder's insurance risk originates from the region which the policyholder resides in. It was shown in a case study performed by Tufvesson[1] in 2016 that the spatial effect derived from a discrete model improved claim frequency predictions, i.e. how many insurance claims will be made during a specific period. However, for the cost, or severity of the claims no spatially associated risk was found.

Based on real data[2,3], it has been shown in a case study performed by Faller[4] in 2021 that a continuous spatial model also improves claim frequency predictions.

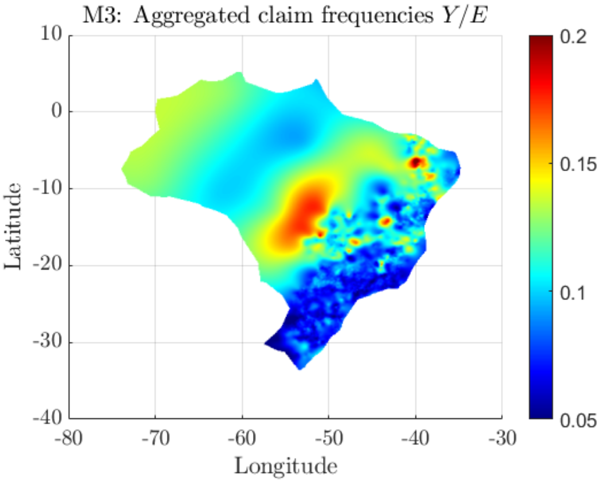


*Figure: Claim frequency predictions for vehicle damage insurances over Brazil, modelled with a continuous spatial dependence*[4]*.*

The continuous model has less requirements regarding the resolution of the geographic data used to determine the geographic rating factors compared to discrete models. Discrete models require *micro-geographical* data, e.g. instead of estimating the spatial effect for a part of Stockholm's inner city with 13,831 areas, as done in Tufvesson's[1] discrete model, the continuous model[4] estimates the spatial effect for the whole of Brazil with 3,109 areas. The relaxed requirements regarding the geographic data in the continuous model[4] enable accurately priced insurances with the use of a spatial effect, even if micro-geographical data is not available.

*Written by: Daniel Faller, 4 January 2022.*

[1] Tufvesson, O. (2016). Spatial statistical modelling of insurance risk: A spatial epidemiological approach to car insurance.

[2] SUSEP. (2015). *Autoseg - susep automobile statistics system.* Retrieved December 30, 2021, from http://www2.susep.gov.br/menuestatistica/Autoseg/menu2.aspx

[3] IBGE. (2010). *Index of Censos Censo Demografico.* Retrieved December 30, 2021, from https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/

[4] Faller, D. (2022). Spatial Statistical Modelling of Insurance Claim Frequency.

# Contents

CHAPTER 1

# Introduction

In this chapter a general background on insurance is given, the scope and purpose of this thesis is also presented in this chapter. The chapter concludes with an outline of the thesis' disposition.

## 1.1 Background

The insurance market consists of different key actors and each actor has different roles. Figure 1.1 presents an overview of the insurance market's key actors. The choice of actors is motivated by the function of the insurance market and which actors primarily ensure the function of the insurance market, namely to spread financial risks (EIOPA, 2021; EU, 2009; FI, 2020; If P&C Insurance, 2021; Regeringskansliet, 2020; Valecký et al., 2017). The topic for this thesis is linked to one of the most central key activities in the insurance market, the quantification of risk.

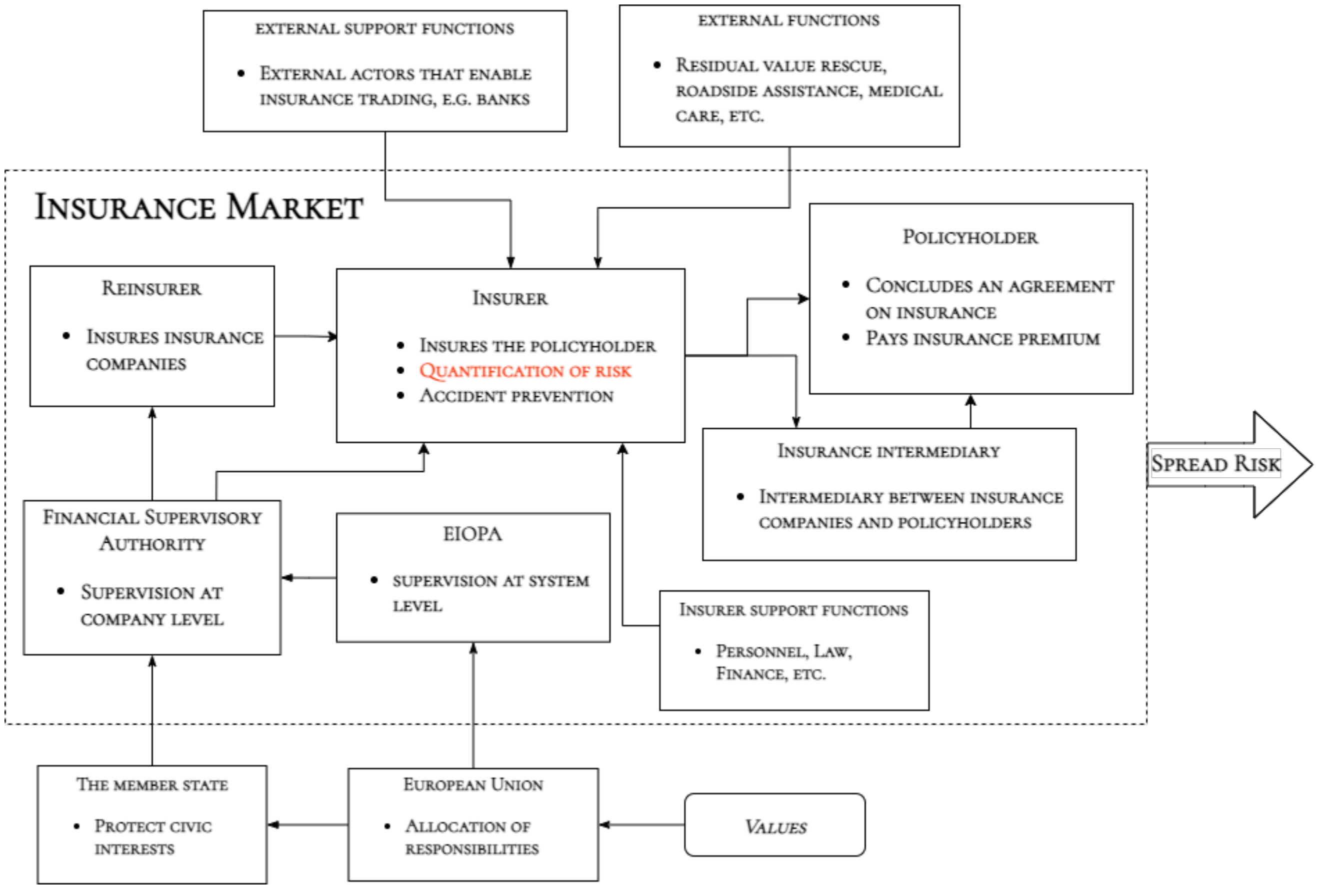

External support functions
External actors that enable insurance trading, e.g. banks
External functions
Residual value rescue, roadside assistance, medical care, etc.
Insurance Market
Reinsurer
Insures insurance companies
Insurer
Insures the policyholder
Quantification of risk
Accident prevention
Policyholder
Concludes an agreement on insurance
Pays insurance premium
Insurance intermediary
Intermediary between insurance companies and policyholders
Spread Risk
Financial Supervisory Authority
Supervision at company level
EIOPA
Supervision at system level
Insurer support functions
Personnel, Law, Finance, etc.
The member state
Protect civic interests
European Union
Allocation of responsibilities
Values

Figure 1.1: The picture above highlights the central activity "Quantification of risk" in the European insurance market.

### 1.1.1 The risk premium

To spread financial risks requires actuarially fair pricing of insurances. What this means is that the risk a policyholder is exposed to must be proportional to the risk premium that the insurer charges. Pricing is also actualised from an enterprise risk management point of view as an event that could threaten the goals of an organisation, as products that are not priced at optimal margins are business threatening (O'Donnell, 2005). Actuarial justice is thus pivotal for the insurer's function to spread risk. The risk group that a policyholder is within must be reflected in the price they pay to insure themselves. To achieve this, it is required that the risk premium is set correctly. Equation (1.1) below defines the risk premium $R$ (Tufvesson et al., 2019).

$$R = \frac{C}{E} \tag{1.1}$$

where $C$ is total claims cost and $E$ is the total duration of policies. The risk premium can also be defined as,

$$R = \frac{C}{Y} \cdot \frac{Y}{E} = S \cdot F, \tag{1.2}$$

where $Y$ is the number of claims, $S$ is severity and $F$ is claims frequency. The tariff which determines the risk premium is generally multiplicative and consists of rating factors. The expected risk premium for the j$^{th}$ policyholder is then modelled as

$$\mathrm{E}[\mathrm{R_j}] = \mathrm{E}[S_j \cdot F_j] = \underbrace{\rho_0 \cdot \prod_{h=1}^{H} \rho_{j,h}}_{\text{Expected severity}} \cdot \underbrace{\gamma_0 \cdot \prod_{k=1}^{K} \gamma_{j,k}}_{\text{Expected frequency}} \tag{1.3}$$

where E[ $\cdot$ ] denotes the expectation operator, $\rho_0$ and $\gamma_0$ are the base risk level, or offsets. $\rho_{j,k}$ and $\gamma_{j,k}$ are rating factors which can be thought of as the relative risk for the j$^{th}$ policyholder with respect to the k$^{th}$ and h$^{th}$ factor respectively (Tufvesson et al., 2019).

To set an actuarially fair price for the risk premium, it becomes central to correctly model claim frequency and claim severity. The claim frequency is often assumed to be Poisson distributed and the claim severity is often assumed to be gamma distributed. Usually, claim frequency and claim severity are modelled separately. The main reason for this is that claim frequency is most often more stable than claim severity, allowing for better predictions for the claim frequency (Ohlsson & Johansson, 2010). This thesis focuses on modelling of the claim frequency for vehicle damage insurances, as there is suitable data available for this class of event, but the method can also be generalised to other types of enumerable events.

The following predictor variables have a proven correlation with the claim frequency and severity and they are used to determine the risk premium (Styrud, 2017): age of policyholder [year], mileage per year [km/year], engine power [kW], length of car ownership [year], car age [year], time since obtaining driver's license [year], population density at the place of residence [persons/km$^2$], whether the car is imported and car brand. At the insurance company If P&C Insurance, geographical location has long been used as one of these variables (Tufvesson, 2016). The motivation for having location dependencies in a pricing model is that it is more likely to be involved in a car accident in a city with denser traffic, than in less populated areas. The estimation of this location dependence will be the focus of this thesis.

## 1.2 Situation analysis

Tufvesson (2016) provided a statistical model for assessing the relative insurance risk associated with the policyholder's geographical location. He modelled claim frequency and claim severity separately, where the Poisson distribution was assumed for claim frequency and the gamma distribution was assumed for claim severity. Basing the models on a Bayesian approach and using the INLA-package (Lindgren & Rue, 2015) in R (R Core Team, 2016) for inference, he showed that the inclusion of a spatial effect from a conditional auto regressive model with first order neighbours (CAR(1)) in an ordinary generalised linear model, improves the prediction quality for claim frequency. He used spatially referenced data of high resolution which makes the conditioning in the CAR model valid. Earlier work in spatial modelling of claim frequencies and severity also include (Gschlößl & Czado, 2007) which used Markov Chain Monte Carlo (MCMC) for inference and a CAR model for the spatial dependence. However, the CAR model might suffer from bias if the available data is not of sufficiently high spatial resolution. The reason for this is that the claim severity and claim frequency of larger pricing areas could have a greater influence on neighbouring pricing areas than feasible and that the relative risk could vary in larger pricing areas.

In (Boskov & Verrall, 1994) the authors proposed a method for premium rating by post-code area. Their method is based on spatial models in a Bayesian framework and uses the Gibbs sampler for estimation. To the extent of my knowledge, I have not encountered any other published work that provides a spatial model for assessing insurance risk.

## 1.3 Purpose of thesis

Uncertainty is sometimes classified into two categories: epistemic and aleatory uncertainty. Epistemic uncertainty can be described as the inadequate understanding of underlying processes, and aleatory uncertainty refers to the inherent uncertainty due to probabilistic variability. The goal of this thesis is to provide a spatial statistical model for estimating the spatial relative risk with respect to the claim frequency of non-life insurances. The inclusion of a spatial effect derived from a GMRF aims to reduce epistemic uncertainty in claim frequency modelling. The model also aims to relax the requirement of spatially referenced data of high resolution as implemented in (Tufvesson, 2016). This is done by implementing a fully Bayesian model on a data set with spatially referenced data of low resolution, i.e. instead of modelling a part of Stockholm's inner city with 13,831 areas as done in (Tufvesson, 2016), the proposed method will model the spatial relative risk for the whole of Brazil with 3,109 areas. The previously stated aims and goals can be framed as one research question; will a spatial effect derived from a continuous spatial dependence reduce uncertainty in claim frequency predictions for vehicle damage insurances?

## 1.4 Thesis disposition

In chapter one a general background on insurance is given as well as the scope and purpose of the thesis. Chapter two presents the data that is used in the thesis and was provided by Prof. Johan Lindström. In chapter three the generalised mixed model for the aggregated insurance claim frequency is presented. Chapter three also presents the motivations behind the construction choices that was made when developing the model. In chapter four the Markov chain Monte Carlo (MCMC) estimation methodology used for the parameter estimations in the model is outlined. Chapter five presents the results of the model in a comparative way, where comparisons with different modifications of the model is made with respect to predictive performance. Chapter six concludes the thesis by summarising the main findings and proposes recommended future research which can increase knowledge within the research field. The appendices contain the derivations made to implement the MCMC-code in MATLAB®.

# CHAPTER 2

# The data

In this chapter, the predictor variables that are used for the estimation of the aggregated claim frequency are presented as well as the geo-statistical data that is used for the spatial modelling. A principal component analysis of the predictor variables is also presented.

## 2.1 The data set

The insurance data is retrieved from the Brazilian organisation Superintendence of Private Insurance (SUSEP, 2015) and the demographic data is retrieved from the Brazilian Institute of Geography and Statistics (IBGE, 2010). The data consists of 3828 polygons representing the 5568 municipalities in Brazil (IBGE, 2020), the polygons can also be referred to as municipalities since they largely coincide. The geographical division in the insurance data is also coarser for the less inhabited municipalities in the west and north-west of Brazil.

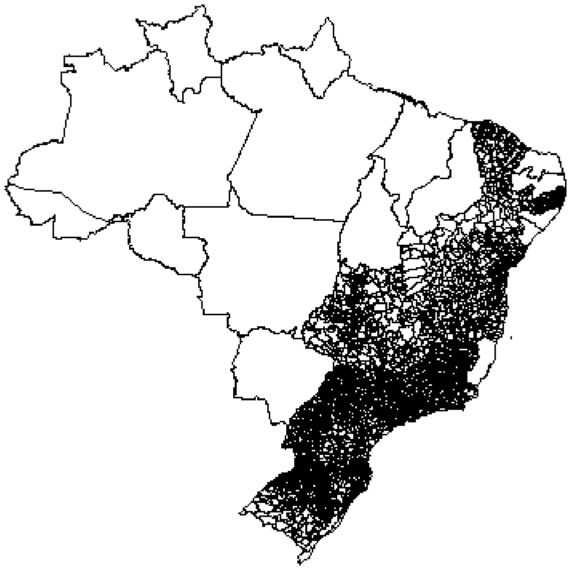


Figure 2.1: Division of Brazil into 3828 polygons each with the twelve attributes specified in table 2.1.

Each polygon has a set of predictor variables for which a model is to be fitted. The geo-statistical data consists of 3828 polygons but only 3109 polygons are used for the inference. The 719 excluded polygons had either missing predictor variables or a too small areal which leads to rows with only zeros in the integration matrix (see section 3.4.3), where no mapped grid elements for a polygon leads to numerically unstable log-likelihoods. The lack of mapping grid elements can be resolved by increasing the spatial resolution, i.e. from $0.1° \times 0.1°$ to $0.05° \times 0.05°$ [Long.×Lat.]. But since the number of operations grows with an order of $\mathcal{O}(n^2)$ per $n$ grid elements this payoff is accepted as the purpose is to model the whole of Brazil, in a reasonable time frame. If one was more concerned with the specific smaller municipalities, a finer division is possible by dividing Brazil into smaller regions which in turn include the smaller municipalities. With the current spatial resolution, municipalities with an area less than 120 km$^2$ are excluded. Each municipality is effectively treated as a policyholder resulting in a geographic rating factor $\gamma_G$ for the policyholder residing in their respective municipality. The response variable for the municipalities is the number of incurred claims for one year due to vehicle collisions. There is a large difference between municipalities regarding the number of incurred claims, e.g. São Paulo has 64,501 incurred claims and 1,264,725 insured vehicles whereas some municipalities only have four insured vehicles (SUSEP, 2015).

The predictor variables for each municipality, or polygon consist of demographic attributes. The following twelve predictor variables are included in the demographic data retrieved from IBGE (2010).

Table 2.1: Available demographic predictor variables for all polygons in Brazil (with square root of variance inflation factors), retrieved from (IBGE, 2010).

| **Available predictor variables for each municipality** | | |
|---|---|---|
| 1. pop.m | (1.19) | Ratio of registered male residents |
| 2. pop.urban | (1.98) | Ratio of residents living in urban areas |
| 3. y18.24 | (1.65) | Ratio of residents between 18-24 years |
| 4. y60 | (1.51) | Ratio of residents above 60 years |
| 5. households | (1.05) | Number of registered households |
| 6. h.owned | (1.90) | Ratio of households that are owned by their residents |
| 7. h.rent | (2.55) | Ratio of households that are rented by their residents |
| 8. h.no.el | (1.25) | Ratio of households with no electricity |
| 9. literacy | (2.21) | Literacy ratio among residents |
| 10. income | (5.96) | Average monthly income per resident |
| 11. income.urban | (5.16) | Average monthly income per resident in urban areas |
| 12. income.rural | (2.00) | Average monthly income per resident in rural areas |

The square root of the variance inflation factor indicates how much larger the standard error for an ordinary least square regression coefficient becomes compared to if its corresponding predictor variable had 0 correlation with the other predictor variables used in the regression. This means that the standard error for a regression coefficient of the predictor variable income is 5.96 times larger than if that predictor had 0 correlation with the other predictors in the data (Miles, 2014). This indicates that income or urban income need to be excluded to avoid severe multicollinearity. The pairwise correlation between income and urban income is 0.9689.

## 2.2 Principal component analysis

In this section an initial principal component analysis is performed to provide quantitative evidence regarding the potential dimensionality reduction of the predictor variables. The reduction of dimensionality aims to increase the interpretability of the final model but at the same time minimise the information loss by keeping the most informative predictors (Jolliffe & Cadima, 2016).

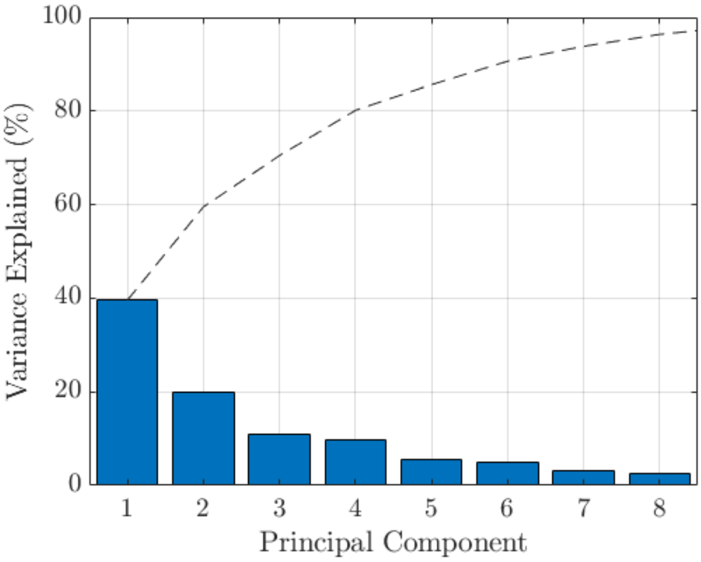


Figure 2.2: A scree plot displaying how well the variance in the attributes of the polygons can be explained by the principal components. By reducing the 12 attributes to 3 principal components it is possible to still explain around 70% of the variance in the normalised predictors.

Figure 2.3, 2.4 and 2.5 identifies the underlying relationships between the observed predictors. The axes consist of transformed coefficients of the principal components so that they are orthonormal, this is done to find a parameterisation in which each predictor has only a small number of large coefficients. That is, each predictor variable is affected by a small number of principal components, preferably only one. This can often make it easier to interpret what the factors represent. It is apparent in Figure 2.3 and 2.4 that the first principal component represents literacy and income level, whereas the second axis represents age. It is worth noting that people between 18-24 are represented as negative and people over 60 are represented as positive on the second principal component, this can be seen in Figure 2.3 and 2.5. It is also worth to note the contrast in Figure 2.5 between houses that are owned and houses that are rented where the latter has similar scores as the ratio of people living in urban areas.

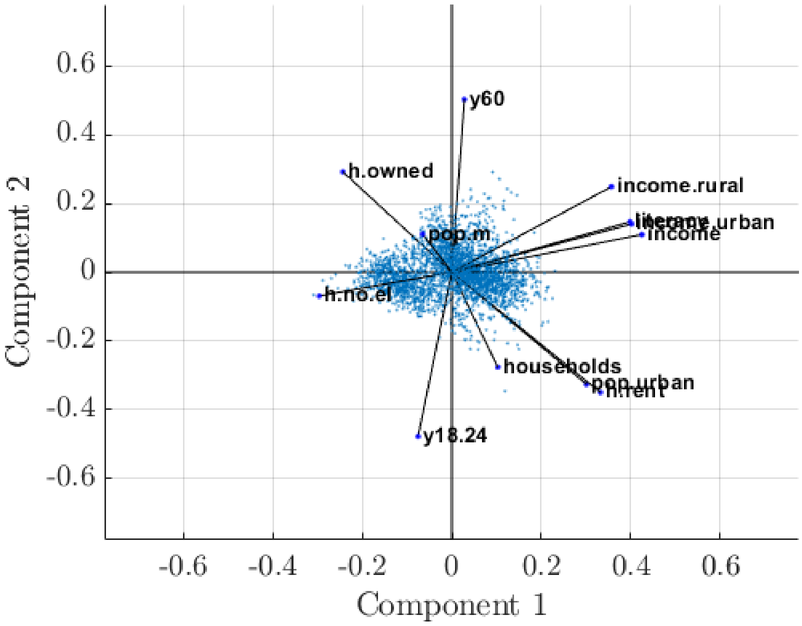


Figure 2.3: Visualisation of the two most variance explaining orthonormal principal component coefficients for each attribute and the principal component scores for each polygon in a single plot.

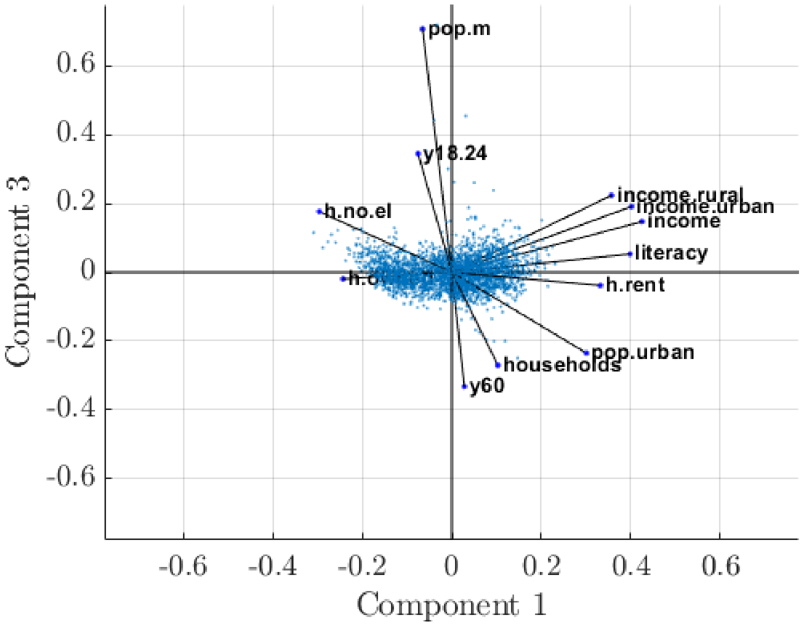


Figure 2.4: Same visualisation as Figure 2.3 but with principal component three on the y-axis.

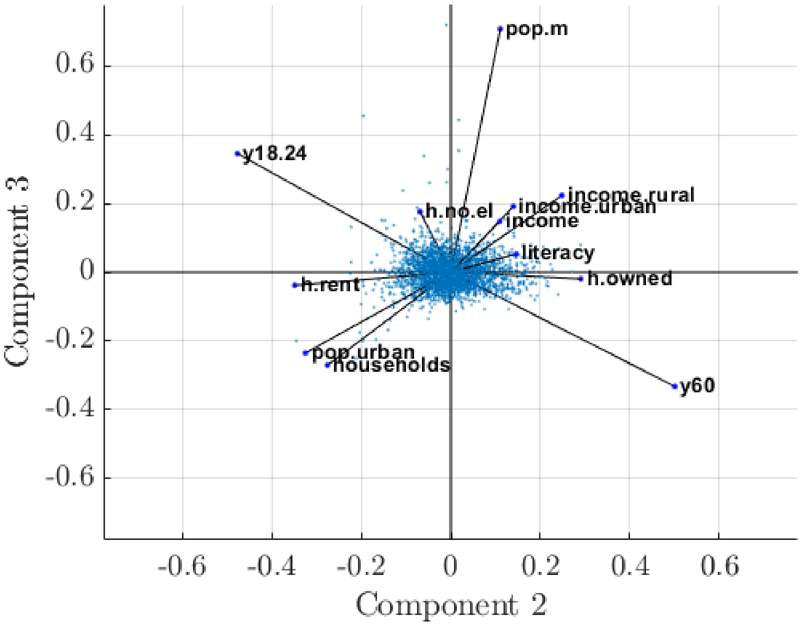


Figure 2.5: Same visualisation as Figure 2.3 but with principal component two and three.

The results from the principal component analysis imply that the initial 12 predictor variables can be reduced due to the proximity of some predictors in the 3-D hyper plane spanned by the three most variance explaining principal components. This is the motivation for the use of Horseshoe priors as a regularisation method for the regression coefficients $\beta$ in section 3.5. The eigenvalue decomposition of the covariance matrix of the predictor variables used in the principal component analysis was performed with the MATLAB® function `pca`.

# CHAPTER 3

# Model

In this chapter the generalised mixed model for the aggregated insurance claim frequency is presented. The implementation of the developed model includes a spatial effect derived from a Gaussian Markov random field (GMRF), which aims to catch underlying effects in the geography of the policyholder that affect the insurance risk. The model also includes an independent and identically distributed (i.i.d) lognormal effect that accounts for the overdispersion in the data and Horseshoe priors that are implemented for the potential regularisation of some predictor variables. In some sections it is more convenient to refer to the model as the model for the number of aggregated insurance claims rather than the model for the aggregated insurance claims frequencies. The only difference is the inclusion or exclusion of the number of insured vehicles $E$ in the linear predictor, where the exclusion yields the frequency model with response variable $Y/E$ instead of $Y$.

## 3.1 Generalised linear mixed model

The predicted number of insurance claims $Y_i$ in a region $B_i$ will be modelled with a Poisson distribution and a log link as,

$$Y_i|\eta_i \sim \mathcal{P}o(\exp(\eta_i)) \tag{3.1a}$$

$$\eta_i = \underbrace{\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta}}_{\text{Fix effect}} + \underbrace{u_i}_{\text{Spatial effect}} + \underbrace{v_i}_{\text{I.i.d effect}} + \underbrace{\log(E_i)}_{\text{Offset}}, \tag{3.1b}$$

where $\boldsymbol{\beta}$ are the regression coefficients for the predictor variables and $\boldsymbol{B}$ consists of a suitable set of predictors. The offset $\log(E_i)$ is included in the linear predictor $\eta_i$ to account for varying vehicle population sizes. In this case, the offset will be the number of insured vehicles $E_i$ that is insured in a municipality, this is equivalent to the exposure or total duration of policies in one municipality. The spatial mixture effect $u_i$ is covered in section 3.4 and the i.i.d mixture effect $v_i$ is covered in section 3.2.

## 3.2 Overdispersion

In this section a first-round test of apparent versus inherent overdispersion is performed by modelling the data using both the Poisson and negative binomial model. The overdispersion is defined as the ratio between the variance divided by the expectation of a stochastic variable, i.e. overdispersion = $V[Y]/E[Y]$.

Poisson models assume the conditional means are equal to the conditional variances, this is not the case with negative binomial models, where the conditional moments are described in equation (3.4). If the estimate of the ancillary parameter $\kappa$ is near zero, then the negative binomial model can be discarded as it equivalent to a Poisson model. Following the arguments regarding how to model overdispersed count data presented in (Ver Hoef & Boveng, 2007), Figure 3.1 is presented to determine which model is more suitable for modelling the overdispersion. The residual intervals are chosen so that the number of collisions in each of the 10 residual intervals is 1/10 of the total collisions.

$$Y \sim \mathcal{P}o(\mu) \Rightarrow E[Y] = V[Y] = \mu \tag{3.2}$$

$$Y \sim quasi\mathcal{P}o(\mu, \phi) \Rightarrow \begin{cases} E[Y] = \mu \\ V[Y] = \phi \cdot \mu \\ \text{Overdispersion} = \phi \end{cases} \tag{3.3}$$

$$Y \sim \mathcal{NB}(\mu, \kappa) \Rightarrow \begin{cases} E[Y] = \mu \\ V[Y] = \mu + \kappa\mu^2 \\ \text{Overdispersion} = 1 + \kappa\mu \end{cases} \tag{3.4}$$

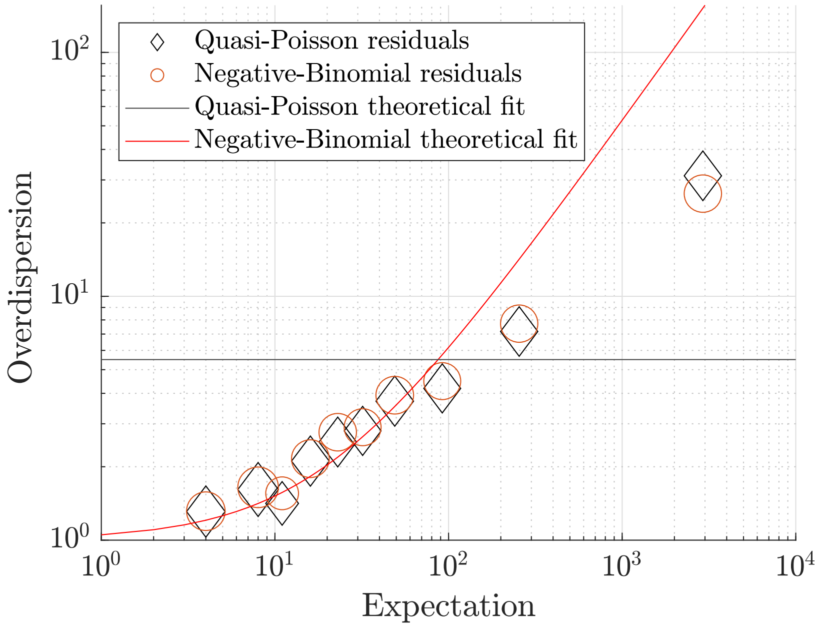


Figure 3.1: Estimated overdispersion-to-mean relationship, the axes are on the logarithmic scale. The markers are averaged squared residuals in ten intervals, where the circle markers are for quasi-Poisson with a fix overdispersion parameter estimated to $\hat{\phi} = 5.5009$ and the diamond markers are for the negative binomial model with an ancillary parameter estimated to $\hat{\kappa} = 0.0517$.

Looking at the overdispersion, a logarithmised linear trend is more prevalent than a fix trend. Hence, a negative binomial model with linearly increasing overdispersion is indicated as a more suitable choice than a quasi-Poisson model with fix overdispersion for the aggregated claim frequency model.

The first-round models were fitted using the MATLAB® toolbox `nbreg` implemented by Surojit (2013), `nbreg` uses iteratively reweighted least squares and $\chi^2$ dampening. The overdispersion parameter $\phi$ is estimated using the Pearson's $\chi^2$ statistic and the degree of freedom

$$\hat{\phi} = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{n - p}, \tag{3.5}$$

where $n$ is the number of observations and $p$ is the number of parameters including the intercept.

### 3.2.1 Negative binomial model as a Poisson-gamma mixture

The negative binomial model can be constructed with the inclusion of an unobserved effect $v_i$ in the linear predictor $\eta_i$ of the generalised linear mixed model in equation (3.1). Excluding the spatial effect and offset; the conditional Poisson mean $\mu_i$ for the claim frequency is (Hardin & Hilbe, 2007, p. 245)

$$\begin{aligned}\log(\mu_i) = \eta_i &= \boldsymbol{B}_i\boldsymbol{\beta} + v_i && (3.6)\\ &= \log(\lambda_{B_i}) + \log(\lambda_{v_i}). && (3.7)\end{aligned}$$

The claim frequency $y_i$ conditioned on the predictor variables and the unobserved effect remains Poisson distributed with the conditional mean and variance given by $\mu_i = \lambda_{B_i}\lambda_{v_i}$,

$$f(y_i|\mu_i) = \frac{e^{-\lambda_{B_i}\lambda_{v_i}}(\lambda_{B_i}\lambda_{v_i})^{y_i}}{y_i!}. \tag{3.8}$$

The conditional mean $\lambda_{B_i}\lambda_{v_i}$ and the unconditional distribution for the claim frequency is given by (Hardin & Hilbe, 2007, p. 246)

$$\begin{aligned}f(y_i, \boldsymbol{B}_i) &= \int_0^\infty \frac{e^{-\lambda_{B_i}\lambda_{v_i}}(\lambda_{B_i}\lambda_{v_i})^{y_i}}{y_i!} g(\lambda_{v_i}) d\lambda_{v_i} && (3.9)\\ &= \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\Big(\frac{1}{1+\lambda_{B_i}/\theta}\Big)^\theta\Big(1-\frac{1}{1+\lambda_{B_i}/\theta}\Big)^{y_i}, && (3.10)\end{aligned}$$

which gives the following moments for the Poisson-gamma mixture (negative binomial) distribution:

$$\begin{aligned}E[Y_i] &= \mu_i && (3.11)\\ V[Y_i] &= \mu_i + \kappa\mu_i^2 && (3.12)\end{aligned}$$

where $\mu_i = \exp(\eta_i)$ and $\kappa = \frac{1}{\theta}$. $g(\cdot)$ is a gamma distribution with mean equal to 1 for the i.i.d effect $\lambda_{v_i} = \exp(v_i)$. The $\theta$ parameter comes from using the shape-rate parametrisation of the gamma distribution which implies $\lambda_{v_i} \sim \mathcal{G}(\theta, \theta)$ to ensure $E[\lambda_{v_i}] = 1$.

The generalised linear mixture model for the number of aggregated insurance claims extended with the multiplicative spatial effect $\exp(u_i)$ (see section 3.4) and multiplicative i.i.d effect $\exp(v_i)$ becomes

$$\begin{aligned}\boldsymbol{Y}|\boldsymbol{\eta} &\sim \mathcal{P}o(\exp(\boldsymbol{\eta})) && (3.13)\\ \eta_i &= \underbrace{\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta}}_{\text{Fix effect}} + \underbrace{u_i}_{\text{Spatial effect}} + \underbrace{v_i}_{\text{I.i.d effect}} + \underbrace{\log(E_i)}_{\text{Offset}}, && (3.14)\end{aligned}$$

where the gamma i.i.d effect is replaced with a lognormal i.i.d effect which gives a similar model (Harrison, 2014)

$$v_i \sim \mathcal{N}(0, q^{-1}). \tag{3.15}$$

The use of a lognormal prior on $\lambda_{v_i} = \exp(v_i)$ instead of a gamma prior enables to fit the model more efficiently by using only one normal apriori-distribution in the MMALA block of the target distribution in section 4.1. The precision parameter $q$ is further explained in section 4.7.

## 3.3 Inhomogeneous Poisson process

The inhomogeneous spatial Poisson process is a Poisson process where the intensity parameter $\lambda_u$ for the Poisson distribution is the surface integral over some bounded region $B_i \in \mathbb{R}^2$.

$$\lambda_u = \exp(u_i) \sim \mathcal{P}o(\Lambda_{B_i}(\boldsymbol{s})), \quad \boldsymbol{s} \in B_i \tag{3.16a}$$

$$\mathrm{E}[\exp(u_i)] = \Lambda_{B_i}(\boldsymbol{s}) \tag{3.16b}$$

$$\Lambda_{B_i}(\boldsymbol{s}) = \frac{\int_{B_i} \exp(X(\boldsymbol{s}))\,\mathrm{d}\boldsymbol{s}}{\int_{B_i} 1\,\mathrm{d}\boldsymbol{s}}, \tag{3.16c}$$

where $\boldsymbol{s}$ is the coordinates of a point in the spatial domain $B_i$. The normalisation of the total area of $B_i$ in equation (3.16c) is performed because the spatial relative risk measure is chosen to be defined as indifferent to the areal size of the spatial domain $B_i$. This ensures that larger regions in the north-west of Brazil does not get an inflated spatial relative risk. The equations in (3.16) form the framework for the construction of a spatial effect derived from a specific region $B_i$ in Brazil in the aggregated claims frequency model, where $X(\boldsymbol{s})$ will be a latent Gaussian Markov random field.

## 3.4 Spatial modelling

In this section the latent spatial model is derived. The latent spatial model for the spatial relative risk is

$$\exp(u_i) = \frac{\int_{B_i} \exp(X(\boldsymbol{s}))\,\mathrm{d}\boldsymbol{s}}{\int_{B_i} 1\,\mathrm{d}\boldsymbol{s}} \approx [\boldsymbol{F}\boldsymbol{A}\exp(\boldsymbol{X})]_i \tag{3.17}$$

$$\boldsymbol{X} | \kappa_x, \tau_x \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{Q}(\kappa_x, \tau_x)^{-1}). \tag{3.18}$$

Note that the boldface $\boldsymbol{X}$ is the node weight vector used to discretise the theoretical Gaussian Markov random field denoted $X(\boldsymbol{s})$ (see section 3.4.1), the matrices $\boldsymbol{F}$ and $\boldsymbol{A}$ are covered in section 3.4.3. The precision matrix $\boldsymbol{Q}$ for the node weights is derived from the stationary solution to the stochastic partial differential equation (SPDE) (3.19) (Lindgren & Rue, 2015)

$$(\kappa_x - \Delta)^{\frac{\alpha}{2}}(\tau_x X(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega. \tag{3.19}$$

Where $\mathcal{W}(\boldsymbol{s})$ is standard Gaussian noise, $\alpha$ controls the smoothness of the field $X(\boldsymbol{s})$ and $\Omega$ is the spatial domain. $\Delta$ is the Laplacian defined as a sum of second order derivatives w.r.t the coordinates $\boldsymbol{s}_i$

$$\Delta = \sum_{i}^{d} \frac{\partial^2}{\partial \boldsymbol{s}_i^2} \tag{3.20}$$

where $d$ is the dimension of the spatial domain. $\kappa_x$ is a scaling coefficient that governs how strong the correlation decay is between two regions. $\tau_x$ is a variance parameter that governs the variation of the field (Moraga, 2019). The link between the precision matrix

$\boldsymbol{Q}$, $\kappa_x$ and $\tau_x$ comes from the Matérn covariance function (Matern et al., 1960) of the Gaussian field that is the exact solution to the aforementioned SPDE (3.19) (Blangiardo & Cameletti, 2015; Lindgren et al., 2011; Moraga, 2019; Whittle, 1954),

$$\boldsymbol{Q}_{ij}^{-1} = \boldsymbol{\Sigma}_{ij} = \operatorname{Cov}[X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)] = \frac{\sigma^2}{\Gamma(\nu)\cdot 2^{\nu-1}} \cdot (\kappa_x \|\boldsymbol{s}_i - \boldsymbol{s}_j\|)^{\nu} \cdot K_{\nu}(\kappa_x \|\boldsymbol{s}_i - \boldsymbol{s}_j\|), \tag{3.21}$$

where $K_\nu$ is the modified Bessel function of the second kind and order $\nu > 0$. $\sigma^2$ denotes the marginal variance of the spatial field and is related to the parameters as

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa_x^{2\nu}\tau_x^2}. \tag{3.22}$$

### 3.4.1 The Gaussian Markov random field

Using Neumann boundary conditions, the precision matrix $\boldsymbol{Q}$ for the node weight vector $\boldsymbol{X} = [X_1, \ldots, X_n]^T$ is given by

$$\boldsymbol{Q} = \tau_x^2(\kappa_x^4 \boldsymbol{C} + 2\kappa_x^2 \boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}). \tag{3.23}$$

The elements of the diagonal matrix $\boldsymbol{C}$ is $C_{ii} = \int a_i(\boldsymbol{s})d\boldsymbol{s}$ and the elements of the sparse matrix $\boldsymbol{G}$ is $G_{ij} = \int \nabla a_i(\boldsymbol{s})\nabla a_j(\boldsymbol{s})d\boldsymbol{s}$, where $\nabla$ denotes the gradient operator and $a_i$ is the basis function as in equation (3.24). The precision matrix $\boldsymbol{Q}$ is sparse and its elements depend on the range parameter $\kappa_x$ and the field precision parameter $\tau_x$. The sparseness of the precision matrix $\boldsymbol{Q}$ makes the node weights $\boldsymbol{X}$ a GMRF with distribution $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ (Blangiardo & Cameletti, 2015).

### 3.4.2 Finite element approximation

The solution to the SPDE (3.19) represented by the stationary and isotropic Matérn Gaussian field $X(\boldsymbol{s})$ can be approximated using the finite element method through a basis function representation defined on a triangulation of the domain consisting of Brazil.

$$X(\boldsymbol{s}) = \sum_{i=1}^{G} a_i(\boldsymbol{s})X_i, \tag{3.24}$$

where G is the total number of vertices of the triangulation, $a_i$ is the set of (deterministic) basis functions, and $X_i$ are zero mean but correlated Gaussian distributed node weights. To obtain a Markov structure, the basis functions are chosen to have local support by being piecewise linear on each triangle, i.e., $a_i$ is 1 at vertex $i$ and 0 at all other vertices (Blangiardo & Cameletti, 2015). The piecewise linear representation of the isotropic Matérn Gaussian field generates a finite element mesh as in Figure 3.3

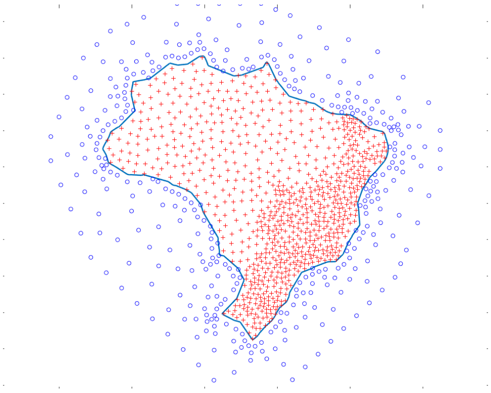


Figure 3.2: The 1204 nodes over Brazil that form the foundation of the finite element mesh.

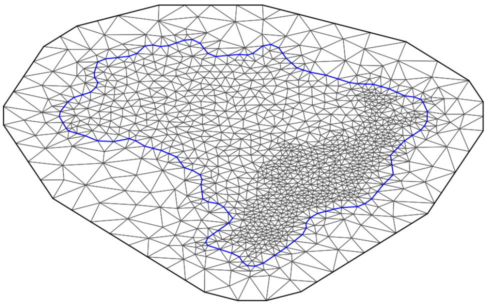


Figure 3.3: Finite element mesh over Brazil using Delaunay triangulation.

### 3.4.3 The integration matrix $\boldsymbol{F}$ and the projection matrix $\boldsymbol{A}$

To be able to integrate the spatial effect of the inhomogeneous spatial Poisson process as in equation (3.16c), an integration grid is placed over Brazil. Where each grid element (or pixel) has an area of 0.1° Long. × 0.1° Lat. ≈ 120 km$^2$ per pixel. The size of the integration matrix $\boldsymbol{F}$ becomes with the current spatial resolution $[N_{\text{regions}} \times 72129]$, i.e. 72,129 grid elements are generated to integrate the spatial effect. All the non-zero elements in the un-normalised integration matrix $\hat{\boldsymbol{F}}$ are set to 1 when the integration matrix is initially constructed in the R-INLA package (Lindgren & Rue, 2015; R Core Team, 2021), this needs to be adjusted for equation (3.16c) to be fulfilled. The adjustment is achieved when each element in the normalised integration matrix $\boldsymbol{F}_{ij}$ corresponds to

$$\boldsymbol{F}_{ij} = \frac{\hat{\boldsymbol{F}}_{ij}}{\sum\limits_{j=1}^{n_{pixels}} \hat{\boldsymbol{F}}_{ij}} \tag{3.25}$$

which gives the following approximation

$$\exp(u_i) = \frac{\int_{B_i} \exp(X(\boldsymbol{s}))\,\mathrm{d}\boldsymbol{s}}{\int_{B_i} 1\,\mathrm{d}\boldsymbol{s}} \approx [\boldsymbol{F}\boldsymbol{A}\exp(\boldsymbol{X})]_i. \tag{3.26}$$

Equation (3.26) shows the discretisation of the surface integral for the bounded i$^{th}$ region $B_i$ over the latent Gaussian field divided by the total area of $B_i$. The projection matrix $\boldsymbol{A}$ is also constructed in the R-INLA package (Lindgren & Rue, 2015; R Core Team, 2021). The elements of the projection matrix consist of the basis function representation defined on a triangulation of the domain consisting of Brazil, see Figure 3.3.

$$X(\boldsymbol{s}_j) \approx \sum_{i=1}^{3} a_i(\boldsymbol{s}_j) X_i = [\boldsymbol{A}\boldsymbol{X}]_j \tag{3.27}$$

$a_i(s)$ are basis functions that weight each j$^{th}$ pixel value to three adjacent node weights $X_i$ whose values will be estimated (see section 4.4). Each row of the projection matrix

$\boldsymbol{A}$ contains three non-zero valued elements which sum to one. These elements are the corresponding weights of the adjacent nodes for one pixel.

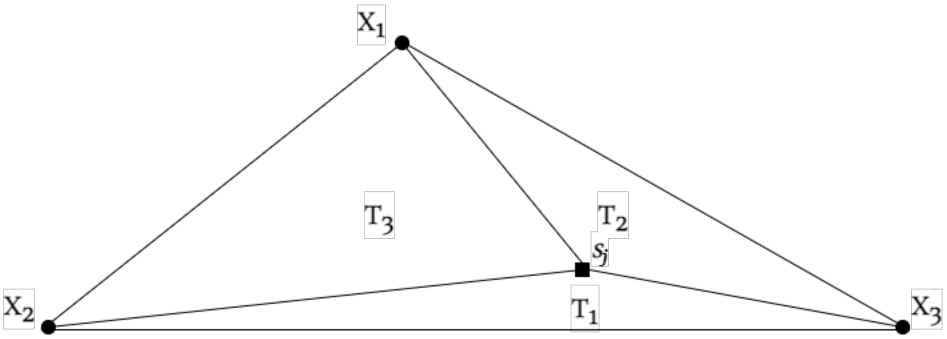


Figure 3.4: A figure that shows how the j$^{th}$ pixel at location $\boldsymbol{s}_j$ inside one triangle of the mesh is weighted w.r.t the three nearest nodes $\boldsymbol{X}_{1,2,3}$ of the discretised GMRF (Moraga, 2019).

The value of the $[\boldsymbol{AX}]_j$ element then corresponds to how one pixel is weighted w.r.t its three adjacent nodes $\boldsymbol{X}_{1,2,3}$

$$[\boldsymbol{AX}]_j = \sum_{i=1}^{3} \frac{T_i}{T} \boldsymbol{X}_i, \quad T = \sum_{k=1}^{3} T_k. \tag{3.28}$$

Lastly, the spatial effect $\boldsymbol{u}$ in equation (3.14) will have the following form

$$\boldsymbol{u} = \log(\boldsymbol{FA} \cdot \exp(\boldsymbol{X})). \tag{3.29}$$

## 3.5 Horseshoe priors

To obtain a sparse solution for the regression coefficients $\boldsymbol{\beta}$, Horseshoe priors for $\boldsymbol{\beta}$ will be used. Horseshoe priors act as an effective method to push non-significant parameters towards zero. Below follows the Horseshoe hierarchy proposed by Carvalho et al. (2009).

$$\boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{\Lambda}^{-1}\right) \tag{3.30a}$$

$$\boldsymbol{\Lambda} = \tau_\beta^2 \cdot \text{diag}(\lambda_{\beta_1}^2, ..., \lambda_{\beta_p}^2) \tag{3.30b}$$

$$\lambda_{\beta_1}, ..., \lambda_{\beta_p} \sim \mathcal{C}^+(0,1) \tag{3.30c}$$

$$\tau_\beta \sim \mathcal{C}^+(0,1) \tag{3.30d}$$

The name Horseshoe arises from the observation that, for $\tau_\beta = 1$, $\sigma = 1$ and $y_i|\boldsymbol{B}_i\beta_i \sim \mathcal{N}(\boldsymbol{B}_i\beta_i, \sigma^2\boldsymbol{I})$ (Carvalho et al., 2009)

$$E[\boldsymbol{B}_i\beta_i|y_i] = \int_0^1 (1-k_i)y_i(\kappa_i|y_i)\mathrm{d}\kappa_i = (1 - E[\kappa_i|y_i])y_i, \tag{3.31}$$

where $\kappa_i = 1/(1+\lambda_i^2)$ and $\mathrm{E}[\kappa_i|y]$ can be interpreted as the amount of shrinkage towards zero. If $\kappa \approx 0$ virtually no shrinkage is exercised, but if $\kappa \approx 1$ then near total shrinkage is exercised on the parameter $\beta_i$. The resulting probability density function for the shrinkage effect $p(\kappa)$ will be U-shaped and this gives rise to the name Horseshoe (Carvalho et al., 2009). The non-standard form of the conditional posterior distributions for the local shrinkage parameters $\mathbb{P}(\boldsymbol{\lambda}_\beta|\boldsymbol{\beta},\boldsymbol{y})$ and global shrinkage parameter $\mathbb{P}(\tau_\beta|\boldsymbol{\beta},\boldsymbol{y})$ makes Gibbs sampling (see appendix A) difficult to implement. Makalic and Schmidt (2015) proposed to use a scale mixture representation of the half-Cauchy distribution on the positive real numbers to make sampling from the conditional posterior more straight forward. The method uses auxiliary variables that lead to conjugate conditional posterior distributions for the local and global shrinkage parameters enabling efficient Gibbs sampling. Makalic and Schmidt (2015) make use of the following scale mixture representation of the half-Cauchy distribution. Let $c$ and $m$ be random variables such that

$$c^2|m \sim \mathcal{IG}\Big(\frac{1}{2},\frac{1}{m}\Big) \quad \text{and} \quad m \sim \mathcal{IG}\Big(\frac{1}{2},\frac{1}{D^2}\Big) \tag{3.32}$$

then c $\sim \mathcal{C}^+(0,D)$ where $\mathcal{IG}(\cdot,\cdot)$ is the inverse-Gamma distribution with probability density function

$$\mathbb{P}(z|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1}\exp(-\frac{\beta}{z}). \tag{3.33}$$

Using the proposed decomposition for the Horseshoe hierarchy (3.30) leads to the following revised Horseshoe hierarchy for the regression coefficients $\beta$

$$\beta_i|\lambda_{\beta_i},\tau_\beta \sim \mathcal{N}(0,\lambda_{\beta_i}^2\tau_\beta^2) \tag{3.34a}$$
$$\lambda_{\beta_i}^2|\nu_i \sim \mathcal{IG}(1/2,1/\nu_i) \tag{3.34b}$$
$$\tau_\beta|\xi \sim \mathcal{IG}(1/2,1/\xi) \tag{3.34c}$$
$$\nu_1,...,\nu_p,\xi \sim \mathcal{IG}(1/2,1), \tag{3.34d}$$

where the predictor, or covariate matrix $\boldsymbol{B}$ that contains the predictors will be normalised by subtraction of the mean and division with the standard deviation for each type of predictor to ensure equal shrinkage on all regression coefficients.

## 3.6 Hierarchical model and priors

In this section the entire Bayesian hierarchy is presented and summarised. The aggregated insurance claims are modelled with a Poisson distribution as

$$\boldsymbol{Y}|\boldsymbol{\eta} \sim \mathcal{P}o(\exp(\boldsymbol{\eta})), \tag{3.35}$$

with the linear predictor

$$\boldsymbol{\eta}|\boldsymbol{\beta},\boldsymbol{X},\boldsymbol{v} = \beta_0 + \boldsymbol{B}\boldsymbol{\beta} + \log(\boldsymbol{F}\boldsymbol{A}\cdot\exp(\boldsymbol{X})) + \boldsymbol{v} + \log(\boldsymbol{E}). \tag{3.36}$$

The Horseshoe hierarchy is

$$\beta_i | \lambda_{\beta_i}, \tau_\beta \sim \mathcal{N}(0, \lambda_{\beta_i}^2 \tau_\beta^2) \tag{3.37}$$

$$\lambda_{\beta_i}^2 | \nu_i \sim \mathcal{IG}(1/2, 1/\nu_i) \tag{3.38}$$

$$\tau_\beta | \xi \sim \mathcal{IG}(1/2, 1/\xi) \tag{3.39}$$

$$\nu_1, ..., \nu_p, \xi \sim \mathcal{IG}(1/2, 1), \tag{3.40}$$

the intercept should not be regularised hence an uninformative normal prior is used

$$\beta_0 \sim \mathcal{N}(0, \psi_1), \quad \psi_1 = 10^6. \tag{3.41}$$

The GMRF with its parameters is modelled as

$$\boldsymbol{X} | \kappa_x, \tau_x \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{Q}^{-1}) \tag{3.42}$$

and the precision matrix of the node weights as

$$\boldsymbol{Q} = \tau_x^2 (\kappa_x^4 \boldsymbol{C} + 2\kappa_x^2 \boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}). \tag{3.43}$$

The i.i.d lognormal effect is modelled as

$$v_i \sim \mathcal{N}(0, q^{-1}), \tag{3.44}$$

where the precision parameter has an uninformative gamma prior

$$q \sim \mathcal{G}(\alpha_v, m_v), \quad \alpha_v = 1.5, \quad m_v = 0.1. \tag{3.45}$$

Below the conditional dependencies in the Bayesian hierarchy for the number of aggregated insurance claims $\boldsymbol{Y}$ is presented.

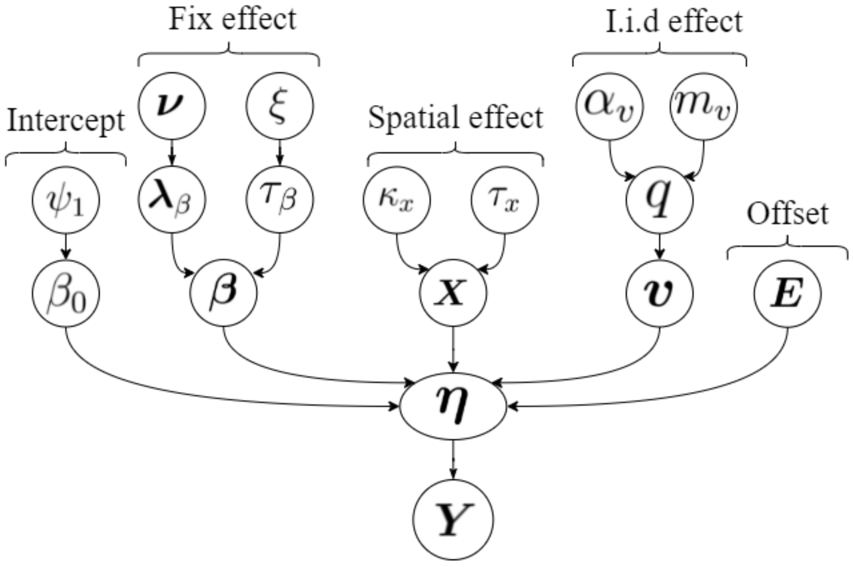


Figure 3.5: Directed acyclic graph describing the conditional dependencies in the hierarchical model.

Summarising the introduced building blocks of the entire model and how they interact, the matrix $\boldsymbol{F}$ is an integration matrix of size $[N_{\text{regions}} \times N_{\text{pixels}}]$ which maps each of the $N_{\text{pixels}}$ to their respective region. The matrix $\boldsymbol{A}$ is the projection matrix of size $[N_{\text{pixels}} \times N_{\text{nodes}}]$ and evaluates each grid element with regard to the basis functions of the mesh. $\boldsymbol{X}$ is the node weight vector of size $[N_{\text{nodes}} \times 1]$, where each element is the weight at a certain node of the triangulation mesh. $E_i$ is the number of insured vehicles in the specified i$^{th}$ region, i.e. exposure (or offset). The offset can be set to be on the natural scale $E_i \cdot \exp(\eta_i)$ or it can be included as an offset in the linear predictor, i.e. $\eta_i = \beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + \log([\boldsymbol{FA} \cdot \exp(\boldsymbol{X})]_i) + v_i + \log(E_i)$. When multiplying $\boldsymbol{F}$, $\boldsymbol{A}$ and $\exp(\boldsymbol{X})$ we get $[\boldsymbol{FA} \cdot \exp(\boldsymbol{X})]_i \approx \int_{s\in i} \exp(X(\boldsymbol{s}))d\boldsymbol{s} \,/\, \int_{s\in i} 1\, d\boldsymbol{s}$ which is a normalised surface integral for the latent Gaussian field over the i$^{th}$ region. $[\boldsymbol{FA} \cdot \exp(\boldsymbol{X})]_i$ is viewed as the incurred claim intensity in a 2-D inhomogeneous Poisson process stemming from the geo-location of the i$^{th}$ region, i.e. the spatial effect. $\boldsymbol{\beta}$ are the regression coefficients for the predictor variables, and $\boldsymbol{B}$ consists of a suitable set of predictors. To account for the inherent overdispersion in the data an i.i.d lognormal effect $v_i$ is added to the linear predictor $\eta_i$. Lastly, Horseshoe priors are used to potentially regularise some regression coefficients.

## 3.7 Approximate minimum degree permutation

To generate proposals for the regression coefficients $\boldsymbol{\beta}$ and the node weights $\boldsymbol{X}$ requires performing a Cholesky decomposition of the precision matrix $\boldsymbol{Q}$ (see equation (B.8)). The Cholesky is a decomposition of a symmetric, positive-definite matrix into the product of a triangular matrix and its transpose

$$\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{R}^T. \tag{3.46}$$

To speed up computations involving $\boldsymbol{R}$ a symmetric reordering is performed prior to the Cholesky decomposition. This reordering reduces the number of non-zero elements in the Cholesky factor $\boldsymbol{R}$ (see Figure 3.6) . The matrix $\hat{\boldsymbol{G}} = \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$ from equation (3.23) has the most structure, hence, the approximate minimum degree (AMD) permutation vector will be generated for this matrix using the MATLAB® function `amd`.

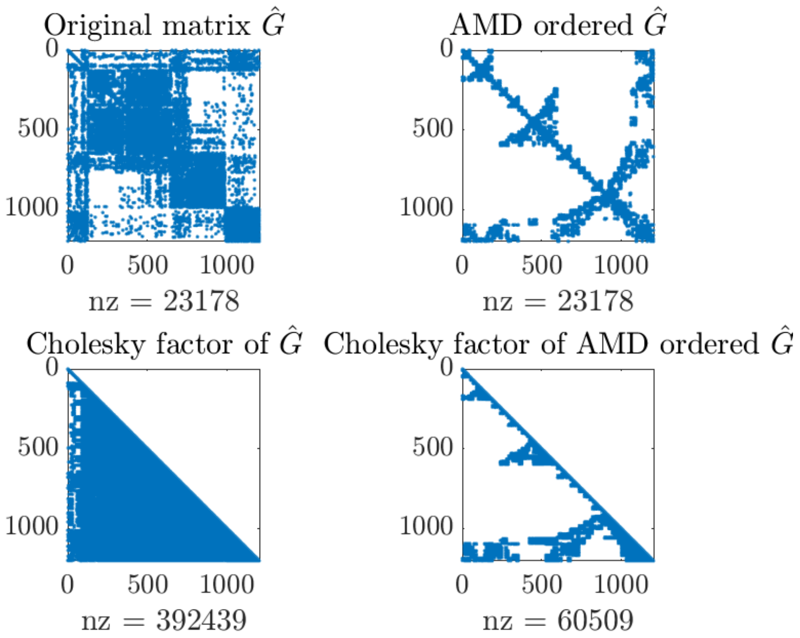


Figure 3.6: Plot of the sparsity patterns for the $\hat{\boldsymbol{G}} = \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$ matrix. The Cholesky factor obtained from the AMD-permuted matrix is considerably sparser compared to the factor of the matrix in its original ordering.

CHAPTER 4

# Estimation using MCMC

This chapter outlines the Markov chain Monte Carlo based sampling methods implemented by the author in MATLAB® and are used for the parameter estimations of the hierarchical model presented in section 3.6.

## 4.1 The target density for the Markov chain

The conditional posterior for the latent model $\mathbb{P}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\lambda}_\beta, \tau_\beta, \kappa_x, \tau_x, \boldsymbol{\nu}, \xi, \boldsymbol{v}, q, \alpha_v, m_v, \psi_1 | \boldsymbol{Y})$ given the observations $\boldsymbol{Y}$ follows below, see appendix D for derivation.

$$\begin{gathered}\mathbb{P}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\lambda}_\beta, \tau_\beta, \kappa_x, \tau_x, \boldsymbol{\nu}, \xi, \boldsymbol{v}, q, \alpha_v, m_v, \psi_1 | \boldsymbol{Y}) \\ \propto \mathbb{P}(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\lambda}_\beta, \tau_\beta, \kappa_x, \tau_x, \boldsymbol{\nu}, \xi, \boldsymbol{v}, q, \alpha_v, m_v, \psi_1) \\ \cdot\mathbb{P}(\boldsymbol{X} | \kappa_x, \tau_x) \cdot \mathbb{P}(\boldsymbol{\beta}, \boldsymbol{v} | q, \alpha_v, m_v, \boldsymbol{\lambda}_\beta, \tau_\beta, \boldsymbol{\nu}, \xi, \psi_1) \\ \cdot\mathbb{P}(\boldsymbol{\lambda}_\beta | \boldsymbol{\nu}) \cdot \mathbb{P}(\tau_\beta | \xi) \cdot \mathbb{P}(q | \alpha_v, m_v) \cdot \mathbb{P}(\kappa_x, \tau_x) \cdot \mathbb{P}(\boldsymbol{\nu}) \cdot \mathbb{P}(\xi).\end{gathered} \tag{4.1}$$

Applying a Metropolis within Gibbs algorithm (see appendix A), the target density can be divided into four main blocks.

1. $\pi(\hat{\boldsymbol{\beta}}|\cdot) \propto \pi(\boldsymbol{Y}|\boldsymbol{\eta}) \cdot \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}_\beta, \tau_\beta) \cdot \pi(\boldsymbol{v}|q) \cdot \pi(\beta_0|\psi_1)$, $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{v} & \beta_0 & \boldsymbol{\beta} \end{bmatrix}^T$
2. $\pi(\boldsymbol{X}|\cdot) \propto \pi(\boldsymbol{Y}|\boldsymbol{\eta}) \cdot \pi(\boldsymbol{X}|\kappa_x, \tau_x)$
3. (a) $\pi(\kappa_x, \tau_x|\cdot) \propto \pi(\kappa_x, \tau_x|\boldsymbol{X})$
   (b) $\pi(\boldsymbol{\nu}, \xi|\cdot) \propto \pi(\boldsymbol{\nu}, \xi|\,\boldsymbol{\lambda}_\beta, \tau_\beta)$
   (c) $\pi(q|\cdot) \propto \pi(q|\boldsymbol{v}, \alpha_v, m_v)$
4. $\pi(\boldsymbol{\lambda}_\beta, \tau_\beta|\cdot) \propto \pi(\boldsymbol{\lambda}_\beta, \tau_\beta|\boldsymbol{\beta}, \boldsymbol{\nu}, \xi)$,

where $\cdot$ denotes the conditioning on all other variables. Block 1, 2 and 3(a) will be updated using the Metropolis Hastings algorithm (due to intractable posteriors) but with different proposals. For block 3(b) and 4 the conditional posteriors exist in a tractable form due to the scale mixture representation in equation (3.32). The conditional posterior also exists in a tractable form for block 3(c), hence block 3(b), 3(c) and 4 will be updated using the Gibbs algorithm.

## 4.2 Metropolis Hastings algorithm

The predominant methodology to sample from un-normalised probability densities $\tilde{p}(\beta)$ is Markov chain Monte Carlo (MCMC) sampling. The most general algorithm that defines a Markov process is the Metropolis Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). The Metropolis Hastings algorithm proposes transitions $\beta \mapsto \beta^*$ with the candidate transition kernel $q(\beta^*|\beta)$, the proposals $\beta^*$ are then rejected or accepted with the following probability (Girolami & Calderhead, 2011)

$$\alpha(\beta, \beta^*) = \min\left\{1, \frac{\tilde{p}(\beta^*)q(\beta|\beta^*)}{\tilde{p}(\beta)q(\beta^*|\beta)}\right\}. \tag{4.2}$$

This acceptance and rejection methodology ensures that the Markov chain is reversible with respect to the stationary target density $\tilde{p}(\beta)$ and satisfies the detailed balance criterion (Robert & Casella, 2013)

$$\alpha(\beta_{i+1}, \beta_i)q(\beta_{i+1}, \beta_i)\tilde{p}(\beta_i) = \alpha(\beta_i, \beta_{i+1})q(\beta_i, \beta_{i+1})\tilde{p}(\beta_{i+1}), \tag{4.3}$$

using $\beta_i = \beta$ and $\beta_{i+1} = \beta^*$.

## 4.3 Updating $\boldsymbol{\beta}$ with MMALA

The first block $\pi(\hat{\boldsymbol{\beta}})$ in section 4.1 (referred to as $\pi(\boldsymbol{\beta})$ for notational simplicity) will be updated using the Metropolis adjusted Langevin algorithm (MALA), where the Langevin diffusion process is preconditioned with the observed negative Fisher information. The ordinary MALA algorithm uses a proposal derived from a discretised Langevin diffusion with a drift term that pushes towards maximising the likelihood for the parameter estimates. The drift term is based on gradient information of the target density for the Markov chain (Girolami & Calderhead, 2011). The form of the candidate density $q(\beta^*|\beta)$ is derived from the Langevin diffusion $\boldsymbol{L}_t$ which is constructed to converge to the target distribution $\pi(\beta)$ under suitable regularity conditions (Roberts & Tweedie, 1996). Roberts and Rosenthal (1998) formally define the reversible Langevin diffusion $\boldsymbol{L}_t$ for the n-dimensional density $\pi$ with variance $\sigma^2$, as the diffusion process $\{\boldsymbol{L}_t\}$ which satisfies the n-dimensional stochastic differential equation

$$d\boldsymbol{L}(t) = \frac{\sigma^2 \nabla \log \pi(\boldsymbol{L}(t))}{2} dt + \sigma d\mathcal{W}(t), \quad \nabla = (\frac{\partial \log \pi}{\partial L_1(t)}, ..., \frac{\partial \log \pi}{\partial L_n(t)})^T. \tag{4.4}$$

It can be shown that $\boldsymbol{L}_t$ has $\pi$ as a stationary distribution, see (Roberts & Tweedie, 1996). Applying equation (4.4) on the regression coefficients $\boldsymbol{\beta}$ yields the stochastic differential equation which defines the Langevin diffusion for $\boldsymbol{\beta}$

$$d\boldsymbol{\beta}(t) = \frac{\sigma^2 \nabla \log \pi\Big(\boldsymbol{\beta}(t)\Big)}{2} dt + \sigma d\mathcal{W}(t). \tag{4.5}$$

Using a forward Euler step as an approximation of the LHS of equation 4.5 gives the following discrete approximation of a preconditioned Langevin diffusion for the regression coefficients $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}(t+\delta) \approx \frac{\mathcal{K}}{2} \nabla \log \pi(\boldsymbol{\beta}(t))\, \delta + \boldsymbol{\beta}(t) + \sqrt{\mathcal{K}}\Big(\mathcal{W}(t+\delta) - \mathcal{W}(t)\Big), \tag{4.6}$$

$$\tag{4.7}$$

using the following notations

$$df(t) \approx f(t+\delta) - f(t), \quad \sigma^2 = \mathcal{K}, \quad dt = \delta. \tag{4.8}$$

Taking the expectation and variance gives

$$E[\boldsymbol{\beta}(t+\delta)] = \frac{\mathcal{K}}{2} \nabla \log \pi(\beta(t))\, \delta + \boldsymbol{\beta}(t) \tag{4.9}$$

$$V[\boldsymbol{\beta}(t+\delta)] = \mathcal{K}\, \delta, \tag{4.10}$$

where $\mathcal{K}$ is a preconditioning matrix that is positive definite. The preconditioning defines the Langevin diffusion on a Riemann manifold with metric tensor $\mathcal{K}$ (Girolami & Calderhead, 2011). The variance in equation (4.10) is given by the increments of a Wiener process (Lindström et al., 2018, p.121). The motivation for the preconditioning is that standard Langevin dynamics gives an isotropic proposal distribution which leads to slow mixing of the MCMC chain if the components of $\boldsymbol{\beta}$ have very different scales or are highly correlated, preconditioning can help with this and lead to better mixing. The preconditioner $\mathcal{K}$ is a user chosen matrix which allows for local adaptation (Girolami & Calderhead, 2011; Patterson & Teh, 2013).

Here the preconditioning matrix for $\boldsymbol{\beta}$ will be the observed negative Fisher information, see appendix B.2 for derivation. Since the discretisation in equation (4.6) introduces a discretisation error, the proposals will be accepted or rejected with a Metropolis Hastings correction step, see appendix B for the derivation of the acceptance rate. This leads to the MMALA transition kernel for the regression coefficients $\boldsymbol{\beta}$ (technically referred to as simplified MMALA by Girolami and Calderhead, 2011),

$$\boldsymbol{\beta}_{i+1}|\boldsymbol{\beta}_i \sim \mathcal{MVN}\Big(\frac{\mathcal{K}}{2}\nabla \log \pi(\boldsymbol{\beta}_i)\,\delta + \boldsymbol{\beta}_i,\, \mathcal{K}\delta\Big). \tag{4.11}$$

## 4.4 Updating $\boldsymbol{X}$ with pCNL

The second block $\pi(\boldsymbol{X})$ in section 4.1 containing the node weights $\boldsymbol{X}$ of the discretised GMRF will be updated with the preconditioned Crank Nicolson Langevin algorithm. $\boldsymbol{X}$ has a Gaussian latent field prior $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ of dimensionality $D = 1204$. This motivates the choice of the pCNL algorithm when updating $\boldsymbol{X}$ since the convergence properties of pCNL are independent of the dimensionality of the target distribution (Hairer et al., 2014). The pCNL discretises the same SDE as MALA but uses a central difference, or Crank Nicolson step for the linear Gaussian part of the gradient from the conditional posterior. The preconditioned Langevin diffusion process for the node weights is

$$\begin{aligned} d\boldsymbol{X}(t) &= \frac{\mathcal{K}}{2}\nabla \log \pi(\boldsymbol{X}(t))dt + \sqrt{\mathcal{K}}d\mathcal{W}(t) && (4.12)\\ &= \frac{-\mathcal{K}}{2}(\nabla\Phi_{\text{GMRF}}(\boldsymbol{X}(t)) + \nabla\Phi_{\text{Po}}(\boldsymbol{X}(t)))dt + \sqrt{\mathcal{K}}d\mathcal{W}(t), && (4.13)\end{aligned}$$

where the gradient of the negative log posterior (NLP) w.r.t $\boldsymbol{X}$ for the Poisson part is denoted by $\nabla\Phi_{\text{Po}}$ and the gradient of the NLP for the GMRF part is $\nabla\Phi_{\text{GMRF}} = \boldsymbol{QX}$, see appendix C.1 for derivations. The pCNL implies sampling from the transition kernel (see appendix C for derivation),

$$\boldsymbol{X}_{i+1} \sim \mathcal{MVN}\left(\frac{1}{4+\delta}\left(-2\boldsymbol{Q}^{-1}\nabla\Phi_{\text{Po}}(\boldsymbol{X}_i)\delta + (4-\delta)\boldsymbol{X}_i\right), 16\boldsymbol{Q}^{-1}\delta\right). \tag{4.14}$$

The proposals generated from the pCNL transition kernel (4.14) will also be accepted or rejected with a Metropolis Hastings correction step to account for any discretisation errors introduced in equation (C.2). When deriving the acceptance probability $\alpha(\boldsymbol{X}_{i+1}, \boldsymbol{X}_i)$ the determinants of the precision matrix cancel due to the symmetry in the transition kernel (4.14) (see equation C.11), this is also a motivation for using pCNL instead of MMALA. The observed negative Fisher information that is used as a preconditioner for the regression coefficients $\boldsymbol{\beta}$ would be computationally heavy to obtain for $\boldsymbol{X}$ due to its high dimensionality, this is also a reason for the use of the prior covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ as a preconditioner for the node weights $\boldsymbol{X}$.

## 4.5 Updating $\kappa_x$ and $\tau_x$ with MHRW

Block 3(a) $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = [\kappa_x, \tau_x]^T$ in section 4.1 containing the field range parameter $\kappa_x$ and the field precision parameter $\tau_x$ will be updated using the Metropolis-Hastings algorithm with random walk proposals (MHRW) and adaptive step size. The update rule will be of the following transformed form to ensure positive values for $\kappa_x$ and $\tau_x$

$$\boldsymbol{\theta}_{i+1} = \exp\Big(\log(\boldsymbol{\theta}_i) + \boldsymbol{\epsilon} \cdot h_i\Big) \quad \text{where} \quad \epsilon \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\theta). \tag{4.15}$$

The candidate $\boldsymbol{\theta}_{i+1}$ will either be accepted or rejected with a Metropolis-Hastings step as in equation (4.2) and the step size $h_{i+1}$ will be updated according to equation (4.24). The covariance matrix $\boldsymbol{\Sigma}_\theta$ for the field parameters is estimated from samples in an initial pilot run using an identity matrix as initial covariance matrix. This implies sampling from the following bivariate lognormal distribution

$$\log(\boldsymbol{\theta}_{i+1}) = \mathcal{MVN}\Big(\log(\boldsymbol{\theta}_i), h_i^2\, \boldsymbol{\Sigma}_\theta\Big). \tag{4.16}$$

The resulting proposal process will not have symmetric proposals due to the log transform and the transition probability becomes:

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = f_{\mathcal{N}(g^{-1}(\boldsymbol{\theta}), h^2\,\boldsymbol{\Sigma}_\theta)}\Big(g^{-1}(\boldsymbol{\theta}^*)\Big) \cdot \left|\frac{\partial g^{-1}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}\right|, \quad g(\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}), \tag{4.17}$$

which then gives the correction factor

$$\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} = \frac{\det\begin{bmatrix} 1/\kappa_x & 0 \\ 0 & 1/\tau_x \end{bmatrix}}{\det\begin{bmatrix} 1/\kappa_x^* & 0 \\ 0 & 1/\tau_x^* \end{bmatrix}} = \frac{\kappa_x^* \tau_x^*}{\kappa_x \tau_x}, \tag{4.18}$$

where the normal probabilities cancel due to symmetry.

## 4.6 Updating $\boldsymbol{\lambda}_\beta$ and $\tau_\beta$ with auxiliary variables

Block 3(b) $\pi(\boldsymbol{\lambda}_\beta, \tau_\beta)$ and the fourth block $\pi(\boldsymbol{\nu}, \xi)$ in section 4.1 containing the local shrinkage parameters $\boldsymbol{\lambda}_\beta$, the global shrinkage parameter $\tau_\beta$, and the auxiliary variables $\boldsymbol{\nu}$ and $\xi$ will be updated with the Gibbs algorithm. The scale mixture representation with auxiliary variables of the half-Cauchy distribution on the positive real (3.32) proposed by Makalic and Schmidt (2015) results in the following conditional posterior distributions required for the update of $\pi(\boldsymbol{\lambda}_\beta, \tau_\beta)$ and $\pi(\boldsymbol{\nu}, \xi)$

$$\lambda_j^2|\nu_j, \beta_j, \tau_\beta \sim \mathcal{IG}\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau_\beta^2}\right), \quad (j = 1, 2, \ldots, N_\beta) \tag{4.19}$$

$$\tau_\beta^2|\xi, \boldsymbol{\lambda}_\beta \sim \mathcal{IG}\left(\frac{N_\beta + 1}{2}, \frac{1}{\xi} + \frac{1}{2}\sum_{j=1}^{N_\beta}\frac{\beta_j^2}{\lambda_j^2}\right) \tag{4.20}$$

$$\nu_j|\lambda_j \sim \mathcal{IG}\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad (j = 1, 2, \ldots, N_\beta) \tag{4.21}$$

$$\xi|\tau_\beta \sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau_\beta^2}\right). \tag{4.22}$$

## 4.7 Updating $q$ with the conditional posterior

Block 3(c) $\pi(q|\cdot) \propto \pi(q|\boldsymbol{v}, \alpha_v, m_v)$ in section 4.1 containing the precision parameter $q$ for the i.i.d lognormal effect $\boldsymbol{v}$ will be updated with the Gibbs algorithm since the conditional posterior can be directly sampled from. Below follows the conditional posterior of $q$

$$q|\boldsymbol{v}, \alpha_v, m_v \sim \mathcal{G}\left(\alpha_v + \frac{n}{2}, \frac{1}{m_v + \frac{1}{2}\sum_{i=1}^{n} v_i^2}\right). \tag{4.23}$$

## 4.8 Adaptive step size

The Metropolis Hastings implementations used to realise the hierarchical model in section 3.6 has a step size $h$ that determines how far from the current point in the parameter space that new parameter candidates will be proposed. An optimum acceptance rate for random walk proposals in high dimensional parameter spaces is around 0.234 (Gelman et al., 1997). An optimum acceptance rate ensures sufficient exploration of the parameter space but at the same time not too many rejected proposals. When considering other proposals than random walks for the Metropolis Hastings algorithm, the optimal proposal scaling can be increased. With Langevin diffusion in the proposal, the optimal asymptotic acceptance rate is 0.57 due to the added gradient information in the proposal process (Roberts & Rosenthal, 1998; Svensson, 2019).

Below follows the update rule for the step size $h$ that is used to generate proposed realisations of the MCMC chain.

$$\log(h_{i+1}) = \log(h_i) + \frac{1}{i^{-0.51}}(\alpha_{acc} - \alpha^\star), \tag{4.24}$$

where $\alpha^\star$ is the optimum acceptance rate and $i$ specifies which realisation of the MCMC chain that is simulated (Givens & Hoeting, 2012, p.248). The adaptation will adjust the step size to obtain optimal acceptance rates.

## 4.9 Implementation of the model

In this section a more practical overview is presented regarding the implementation of the aggregated claim frequency model and its validation. The model is implemented with the following steps

1. Create approximate solution to the SPDE (3.19) using the R-INLA package by Lindgren and Rue (2015), this yields the matrices $\boldsymbol{G}, \boldsymbol{C}, \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}, \boldsymbol{A}$ and $\boldsymbol{F}$.
2. Normalise $\boldsymbol{F}$ for equation (3.16c) to be fulfilled.
3. Randomly partition the data into ten sets with equal number of observations in each set, then exclude one set and train the model on the remaining sets.
4. Normalise the predictor variables in the covariate matrix $\boldsymbol{B}$ so that each attribute of the municipalities has mean zero and standard deviation one.
5. Initialise model parameters, where the regression coefficients $\boldsymbol{\beta}$ are initialised with ridge regression.
6. Generate the approximate minimum permutation vector from $\boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$, then permute the matrices $\boldsymbol{G}, \boldsymbol{C}, \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$ and $\boldsymbol{A}$.
7. Start the Metropolis within Gibbs sampling with N iterations, where N is set to $10^4$ iterations.
   (a) Sample the node weights $\boldsymbol{X}$ with pCNL and update constants w.r.t the regression coefficients $c_\beta$ (see appendix B.1).
   (b) Sample the regression coefficients $\boldsymbol{\beta}$ and the i.i.d lognormal effect $\boldsymbol{v}$ with MMALA and update constants w.r.t the node weights $c_x$ (see appendix C.1).
   (c) Sample the local shrinkage parameters $\boldsymbol{\lambda}_\beta$ and the global shrinkage parameter $\tau_\beta$ with auxiliary variables.
   (d) Sample the field parameters $\kappa_x$ and $\tau_x$ with MHRW and update the precision matrix $\boldsymbol{Q}$ for the node weights $\boldsymbol{X}$ according to equation (3.23).
   (e) Sample the i.i.d lognormal precision parameter $q$ from its conditional posterior in equation (4.23).
8. Un-normalise the predictor variables $\boldsymbol{B}$.
9. Re-scale the regression coefficients $\boldsymbol{\beta}$, this needs to be done since the model was trained with normalised predictor variables.
10. Compute the sample mean as an estimate for the model parameters from the sampling in step 7, with a burnin of $2.5 \cdot 10^3$ iterations.
11. Permutate back the matrices in step 6.
12. Evaluate and save results.
13. Perform a 10-fold cross validation by repeating steps 1-12.

# CHAPTER 5

# Results

The MCMC diagnostics and results presented in this chapter indicate how well the MCMC based inference performed on the Brazilian insurance data. Five models are presented, whereof model one is to be viewed as a benchmark standard for high-dimensional Bayesian regularised count regression. Model 2-5 uses MCMC implementations written by the author in MATLAB® for the estimation of the parameters in the models.

- M1: Only includes a fix effect in the linear predictor, the parameter estimation is based on Hamiltonian Monte Carlo (HMC) combined with the 'No-U-Turn Sampler' (NUTS), the parameter estimation is performed by the MATLAB® toolbox `bayesreg` implemented by Makalic and Schmidt (2016); the toolbox was updated 2020-11-30 to include Poisson regression. The toolbox `bayesreg` is also available in R (R Core Team, 2021).
- M2: Only includes a fix effect in the linear predictor and uses MMALA for inference.
- M3: Is the same as model two but has a modified linear predictor that includes a spatial effect $u_i$ which is estimated using pCNL.
- M4: Is the same as model two but has an i.i.d lognormal effect $v_i$ in the linear predictor that is motivated by the overdispersion in the Brazilian count data, the i.i.d effect is sampled with MMALA in the same block as the regression coefficients $\boldsymbol{\beta}$, the variance $q^{-1}$ of $\boldsymbol{v}$ has a known conditional posterior and is estimated through Gibbs sampling.
- M5: Uses the estimated $\boldsymbol{\beta}$ and $\boldsymbol{v}$ from model four and an intercept that is estimated from an initial ridge regression in the linear predictor but includes a spatial effect that is estimated using pCNL. The reason not to estimate all parameters exclusively in model five is because the i.i.d lognormal effect $\boldsymbol{v}$ fits an individual $v_i$ to each observation point which fully fits the data and pushes the GMRF towards zero.

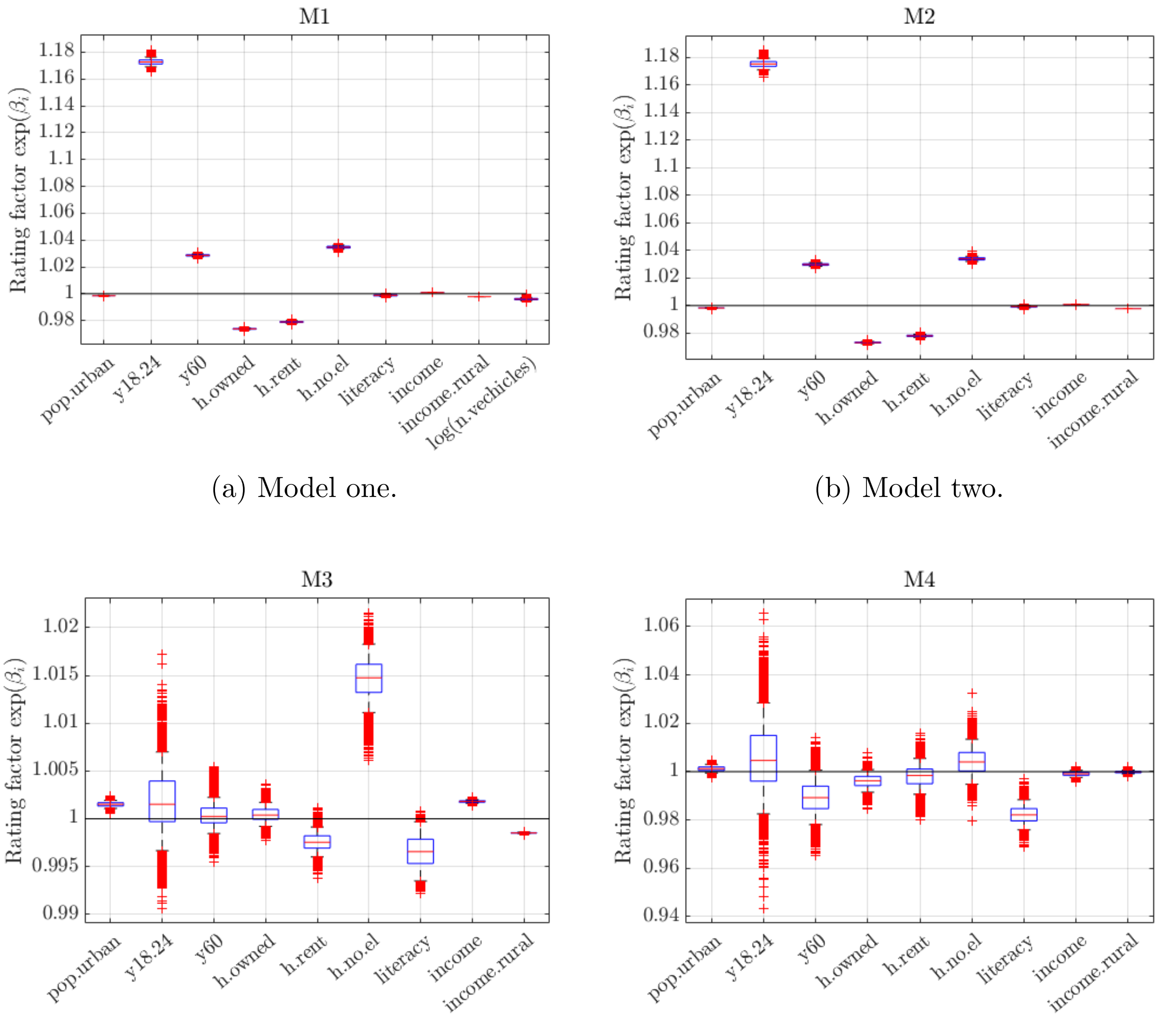
The linear predictors for the 5 different aggregated claim models are summarised in Table 5.1.

Table 5.1: Linear predictors $\eta_i$ for the five models.

| Models | $\eta_i$ |
|---|---|
| M1: Fix HMC NUTS | $\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + \log(E_i)$ |
| M2: Fix MMALA | $\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + \log(E_i)$ |
| M3: Fix MMALA + spatial pCNL | $\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + u_i + \log(E_i)$ |
| M4: Fix + i.i.d MMALA | $\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + v_i + \log(E_i)$ |
| M5: Fix + i.i.d MMALA + spatial pCNL | $\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + u_i + v_i + \log(E_i)$ |

## 5.1 Regression coefficient estimates

Figure 5.1 shows that model one and model two have identical regression coefficient estimates (rating factors) except that the `bayesreg` toolbox does not allow for offset specifications, so the offset is included as a predictor variable in model one. Model three and four show more conservative regression coefficient estimates, model three exhibits especially conservative estimates. It is worth to note that the added variance in model four renders estimates insignificant at a higher significance level compared to model three. The estimates are based on the same training set with the same seed for the pseudorandom number generators used in the MATLAB® code. A first round estimation was performed where the Horseshoe priors pushed: ratio of male residents, average monthly income per residents in urban areas, and number of registered households towards zero, hence these predictor variables are not included in Figure 5.1.



(a) Model one.

(b) Model two.

(c) Model three.

(d) Model four.

Figure 5.1: Boxplot for the regression coefficients, income rural is pushed towards zero by the Horseshoe priors in all models. Model five is excluded as its regression coefficient estimates are identical to those of model four.

The regression coefficient estimates in Figure 5.1 need to be interpreted with caution from an inferential point of view, due to multicollinearity among the predictor variables. A comparative GLM was fitted to the principal components presented in section 2.2 using the MATLAB® function `fitglm`. The use of uncorrelated principal components gives a more stable estimation of causality among the demographic attributes of the regions and their claim frequencies. From Figure 2.4 it can be seen that positive PC1 scores represents literacy and income levels which reduces the claim frequency estimates for a region, with a p-value of 0.01 (presented in Table 5.2). PC2 represents ratio of houses that are owned and ratio of residents above 60 years with positive scores, while ratio of houses that are rented and ratio of residents between 18-24 years with negative scores. The possible increase in claim frequencies caused by higher PC2 scores has a p-value of 0.071.

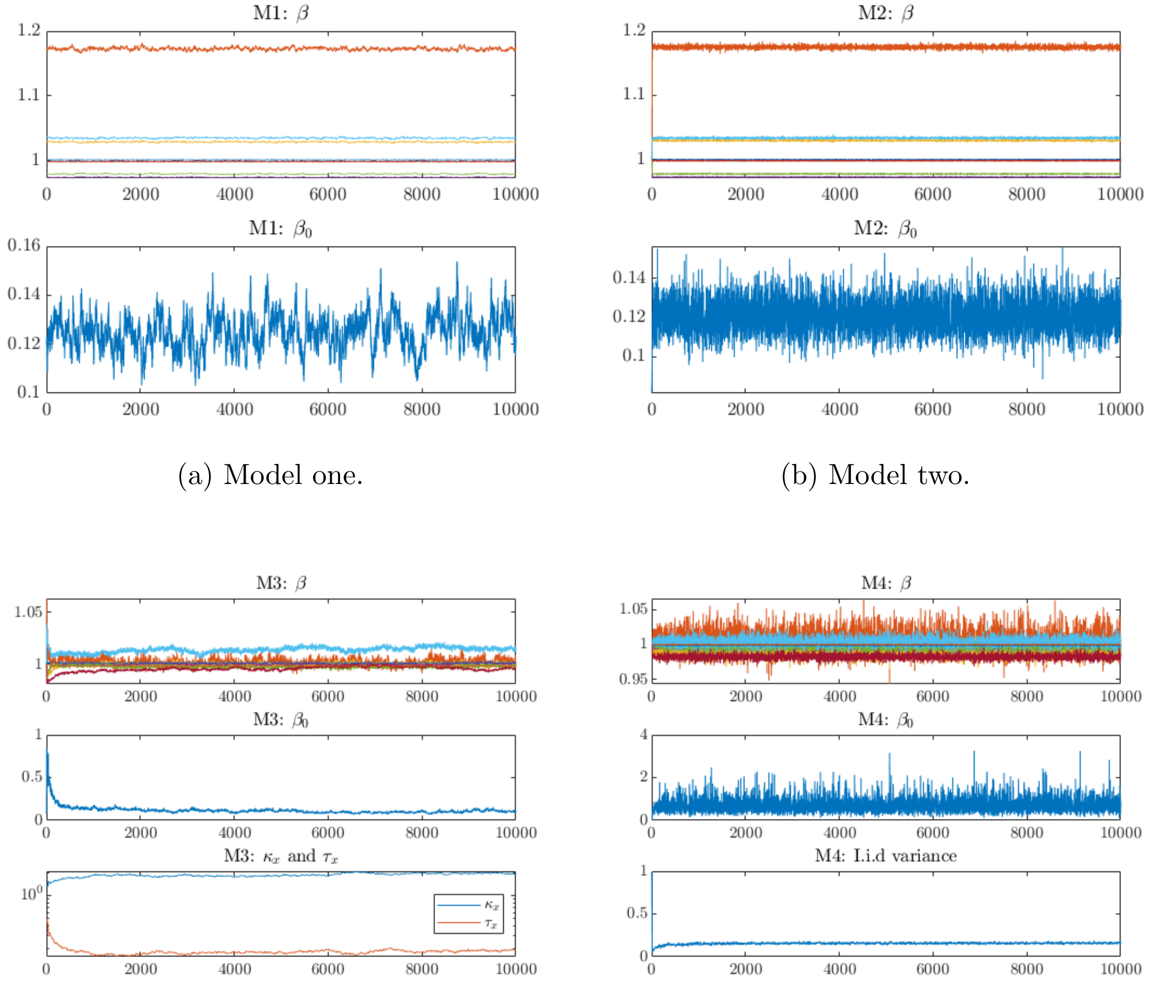
Table 5.2: Principal component regression.

| **Predictor** | $\exp(\beta_i)$ | p-value |
|---|---|---|
| Intercept | 0.0849 | 0.0000 |
| PC1 | 0.9315 | 0.0100 |
| PC2 | 1.0790 | 0.0710 |
| PC3 | 1.0121 | 0.8290 |

The p-values of 0.0100, 0.0710, and 0.8290 indicate that only the coefficient of PC1 is statistically significant at the 95% significance level.

## 5.2 Trace plots

Figure 5.2 shows the mixing of the MCMC chains. Most MCMC chains in all models exhibit rapid mixing, the difference in the MCMC chains for the intercept $\beta_0$ in model one and two is because model one uses a five-iteration thinning interval. In addition to $\boldsymbol{\beta}$ and $\beta_0$, model three has two additional field parameters and model four has one additional i.i.d lognormal effect variance parameter that is estimated. A sample of the 1204 possible trace plots for the node weights $\boldsymbol{X}$ is presented in Figure 5.5.



(a) Model one.

(b) Model two.

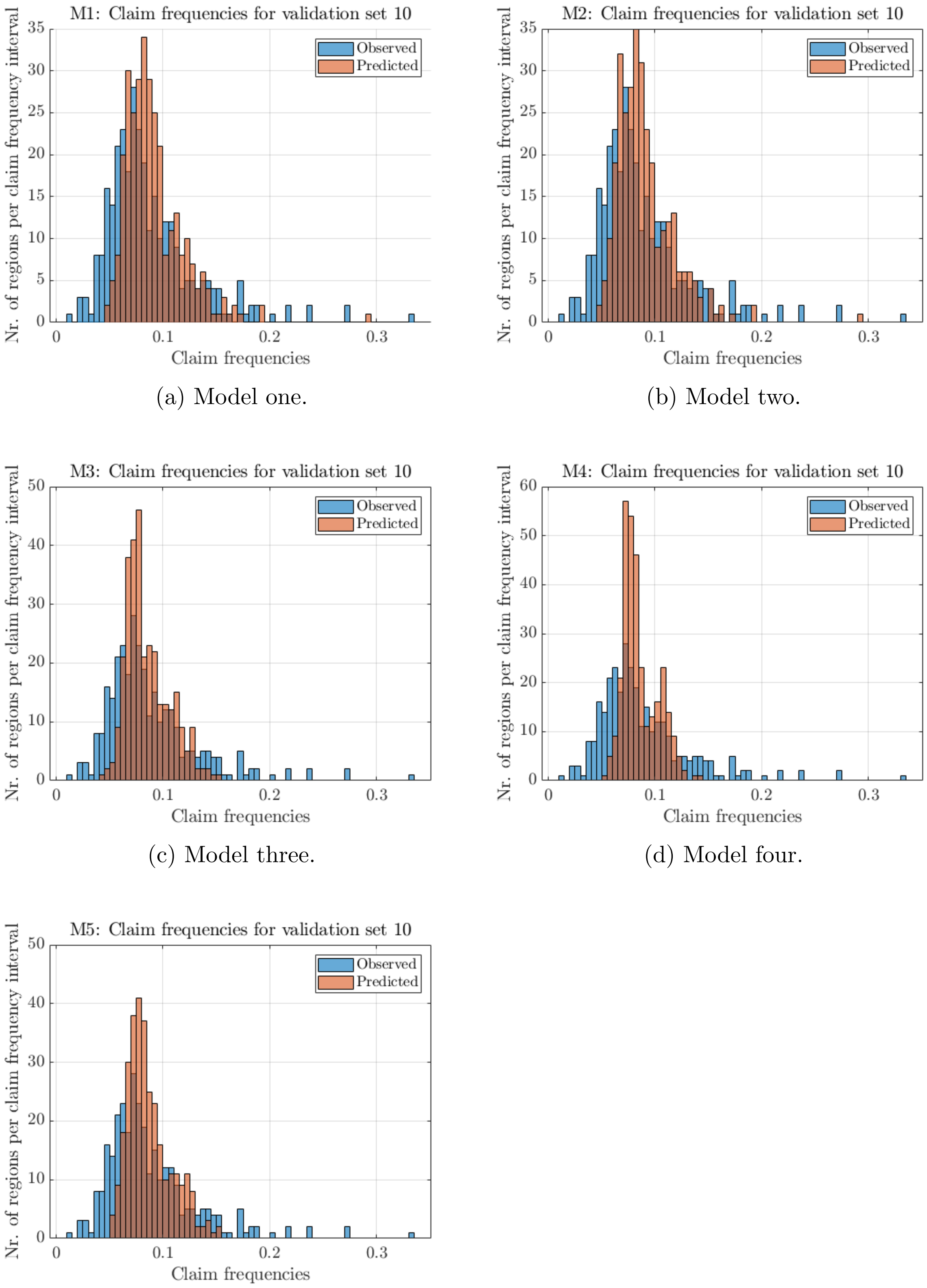
(c) Model three.

(d) Model four.

Figure 5.2: Trace plots for the model parameters. Model five is excluded as its regression coefficients estimates and i.i.d lognormal effect variance (i.i.d variance) is identical to those of model four.

## 5.3 Claim frequency predictions

Figure 5.3 shows the predicted mean claim frequencies and the observed claim frequencies for the same 310 municipalities in the $10^{th}$ validation set. Model 1-3 models the mean and variance as the Poisson distribution, i.e. $E[Y] = V[Y]$. Model four models the variance of the insurance claim frequency with an extra lognormal i.i.d effect. The variance of the lognormal i.i.d effect in model four is estimated to be $\hat{q}^{-1} = 0.1581$, i.e. $\exp(v_i) \sim log\mathcal{N}(0, 0.1581)$.



(a) Model one.

(b) Model two.

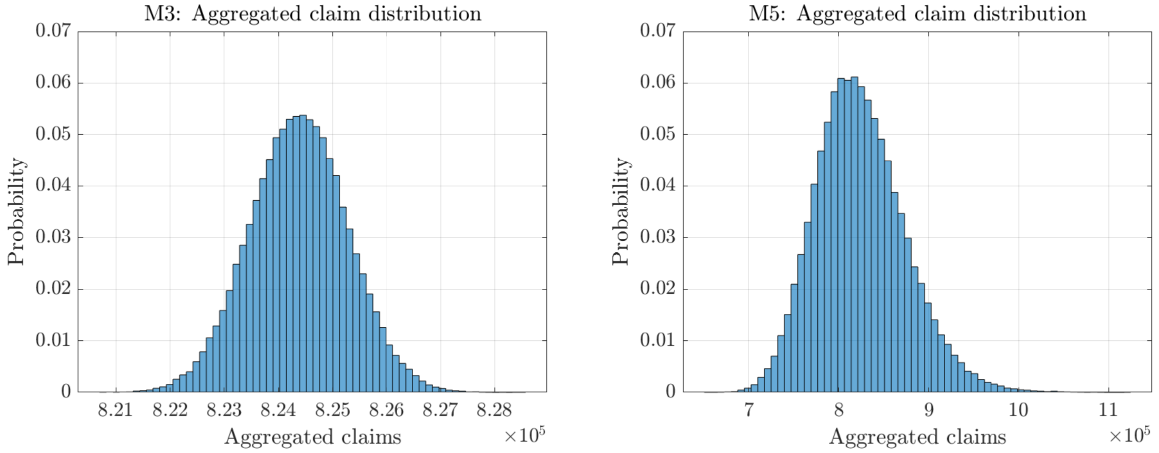
(c) Model three.

(d) Model four.

(e) Model five.

Figure 5.3: Histogram over the predicted and observed insurance claim frequencies for the same 310 validation regions, shown for model 1-5.

Figure 5.4 shows the aggregated claim distribution for the entire insurance portfolio using model three and five for inference. The distributions were created through a bias corrected parametric bootstrap based on 100,000 realisations of the portfolio. All the parameters used in the simulation were re-drawn at each realisation from a normal distribution with mean equal to the parameter estimates and variance scaled by the variance of the parameter estimates. It is evident that the lognormal i.i.d effect in the linear predictor of model five results in a more widespread distribution with right skewness and lepto kurtosis (fat tails). The observed number of claims in the entire portfolio is 824,379 claims.



(a) Model three.

(b) Model five.

Figure 5.4: Aggregated claim distribution for model three and five.

## 5.4 Validation

The validation consists of a 10-fold cross-validation for each model. The cross-validation measure ($\mathrm{CV}_{\mathrm{error}}$) is the sum of the normalised absolute prediction errors,

$$\mathrm{CV}_{\mathrm{error}} = \sum_{i=1}^{n} \frac{|y_i - \eta_i|}{\sqrt{\eta_i}}. \tag{5.1}$$

The motivation for the division with the square root of the expected number of incurred claims $\eta_i$ is that the CV measure should not be dominated by municipalities with larger vehicle populations. The square root of $\eta_i$ corresponds to the variance of the prediction due to the Poisson property of the model, hence the $\mathrm{CV}_{\mathrm{error}}$ can be viewed as the sum of normalised absolute errors. The time to fit one training set takes approximately 30 seconds for model five, 6 minutes for model four, 2 minutes for model three, 3 seconds for model two and 5 seconds for model one, on a personal computer (Intel® Core™ i7-11800H CPU (2021) with 16 GB memory).

Table 5.3: Sum of normalised absolute errors (and standard deviation) from a 10-fold cross-validation for each model (best value in bold).

| **Models** | $\mathrm{CV}_{\mathrm{error}}$ | (sd) |
|---|---|---|
| M1: Fix HMC NUTS | 688.78 | (113.33) |
| M2: Fix MMALA | 674.77 | (105.18) |
| M3: Fix MMALA + spatial pCNL | **529.46** | (**67.98**) |
| M4: Fix + i.i.d MMALA | 609.26 | (125.56) |
| M5: Fix + i.i.d MMALA + spatial pCNL | 531.11 | (75.46) |

A two-sample t-test is also performed to see if there is a significant difference between the $\mathrm{CV}_{\mathrm{error}}$ from M3 and M2. The performed t-test returns a test decision for the null hypothesis that the $\mathrm{CV}_{\mathrm{error}}$ from M3 and M2 comes from independent random samples from normal distributions with equal means without assuming that the cross-validation errors also have equal variances. The alternative hypothesis is that the $\mathrm{CV}_{\mathrm{error}}$ from M3 and M2 comes from populations with unequal means. The result $h$ is true if the test rejects the null hypothesis at the 5% significance level, and false otherwise. The result of the t-test is that $h$ is true, indicating that the t-test rejects the null hypothesis at the 5% significance level, even if equal variances are not assumed. The p-value of the t-test is $p_{value} = 0.0022$ and the 95% confidence interval for the difference between the $\mathrm{CV}_{\mathrm{error}}$ for model three and two is: $CI_{0.95}\ [CV_{error}^{M3} - CV_{error}^{M2}] = [-229.5321; -61.0878]$.

### Model three: Fix MMALA + spatial pCNL

In this section model three with its spatial component is further evaluated. Table 5.4 contains 9 of the 1204 node weights that discretise the GMRF from which the spatial effect is derived from. The nine node weights are selected from three types of location categories: High risk for regions with high relative risk, low risk for regions with low relative risk, and sparse for regions with few neighbouring municipalities. Figure 5.6 displays the location of the nine node weights on a map where the Z component is the value of the multiplicative geographic rating factor $\gamma_{G_i}$. Figure 5.5 shows the trace plots of the nine selected node weights in Table 5.4, where some nodes seem to have converged and some seem not to have converged. Figure 5.7a shows the adaptive step lengths for the parameters in model three and Figure 5.7b shows the acceptance rates for the Metropolis-Hastings updated parameters which converged to the optimal acceptance rates, 0.57 for Langevin proposals and 0.243 for random walk proposals.

Table 5.4: Nine node weights from three types of location categories.

| **Node weights $X_i$** | **Long., Lat.** | **Location** |
|---|---|---|
| $X_{594}$ | -46.1906, -16.4869 | High risk |
| $X_{404}$ | -51.1696, -15.6677 | High risk |
| $X_{942}$ | -39.9113, -6.25333 | High risk |
| $X_{723}$ | -54.1208, -27.3682 | Low risk |
| $X_{687}$ | -47.1449, -22.7482 | Low risk |
| $X_{433}$ | -38.4012, -12.7691 | Low risk |
| $X_{496}$ | -70.4735, -9.21299 | Sparse |
| $X_{920}$ | -64.6530, -4.15386 | Sparse |
| $X_{567}$ | -53.0643, -3.97494 | Sparse |

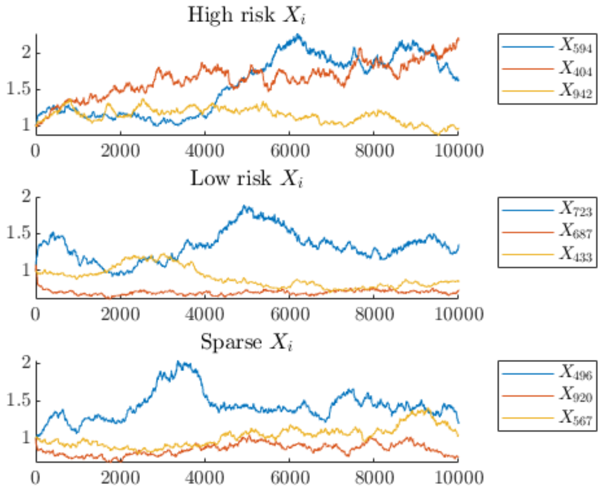


Figure 5.5: Trace plots of the nine node weights in Table 5.4

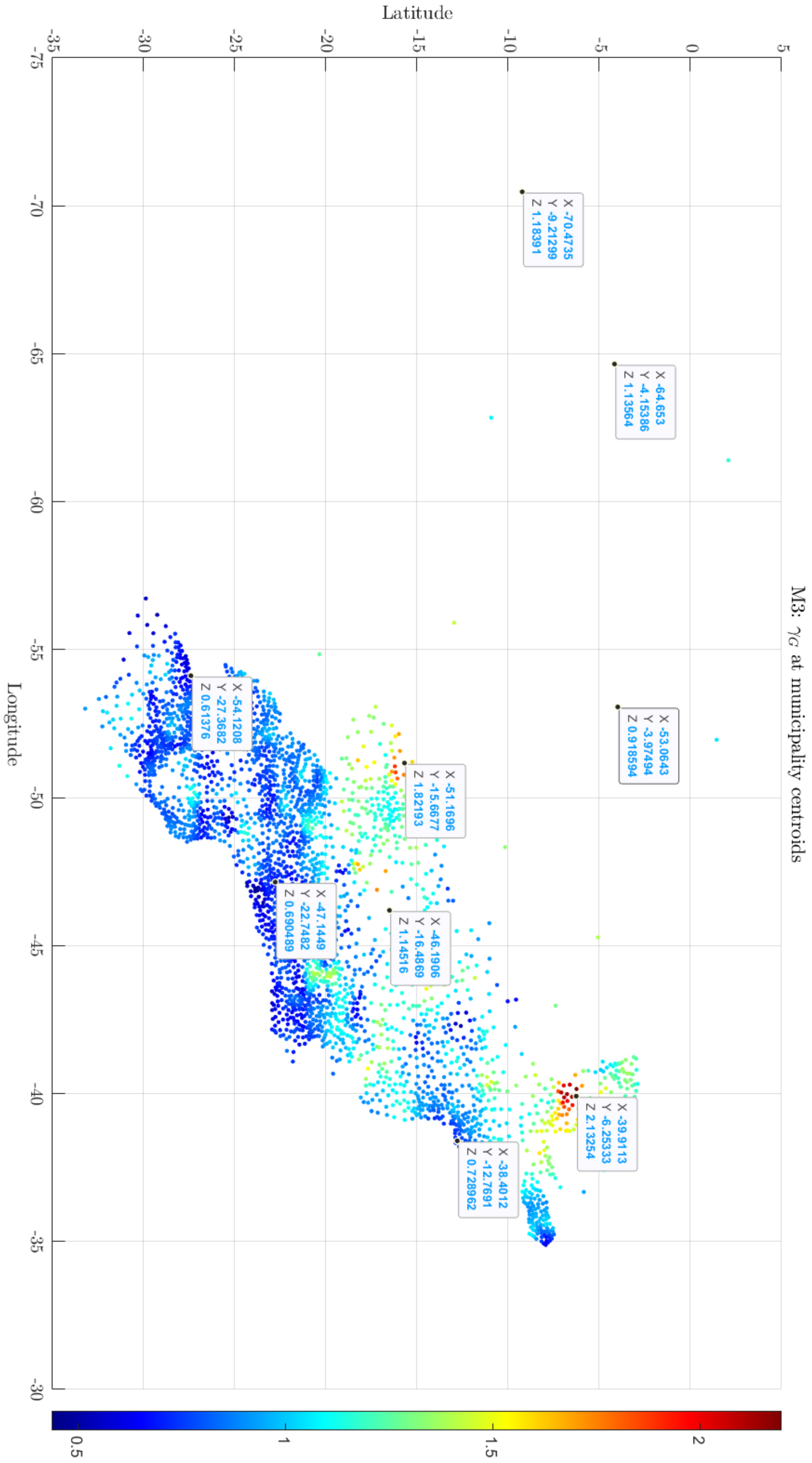
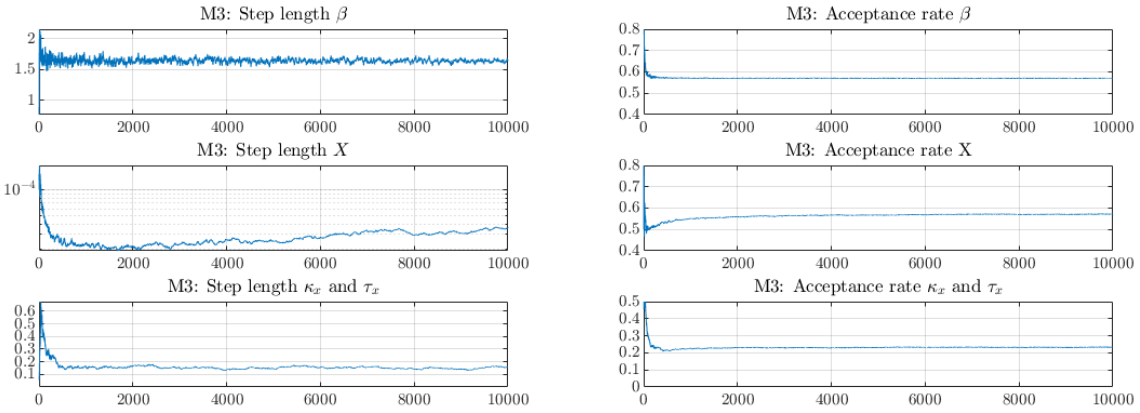


Figure 5.6: The nine node weights from Table 5.4 at each highlighted municipality centroid, X=Long. Y=Lat. Z = $\gamma_{G_i} = \exp(\beta_0 + \boldsymbol{B}_i\boldsymbol{\beta} + u_i)$.



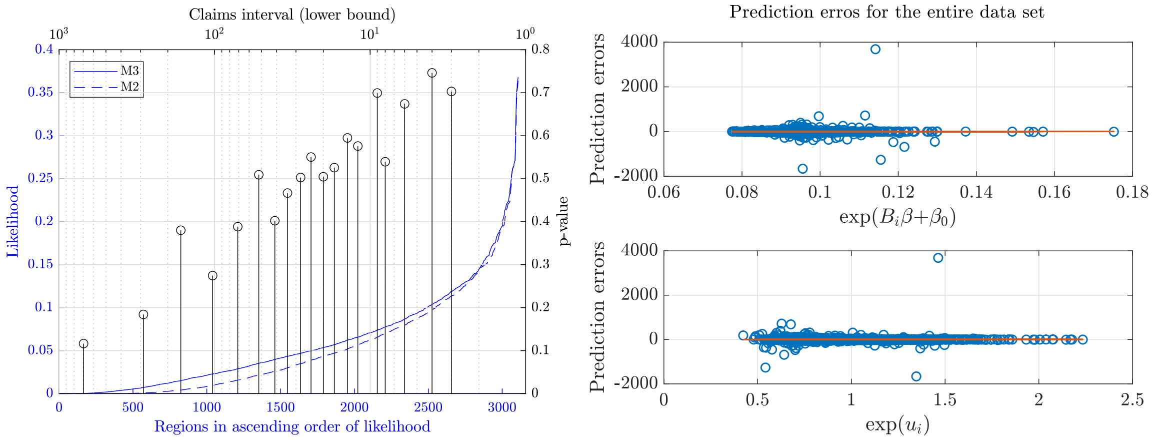
(a) M3: Step lengths.

(b) M3: Acceptance rates.

Figure 5.7: Step lengths for the regression coefficients, node weights and field parameters (a). Acceptance rates for the regression coefficients, node weights and field parameters (b).

Figure 5.8a displays the Poisson likelihoods (in blue) on the left y-axis for the number of aggregated insurance claims for each municipality in the entire data set, given the linear predictor from M2 (dashed blue line) or M3 (continuous blue line), i.e. $\mathbb{P}(y_i|\eta_i^{M2})$ (dashed blue line) or $\mathbb{P}(y_i|\eta_i^{M3})$ (continuous blue line). It is apparent that the inclusion of a spatial effect yields higher likelihoods for regions with a larger number of aggregated insurance claims, mainly because the potential added predictive performance from using the spatial effect does not show for aggregated claims less than approximately 700 claims, this is based on a likelihood ratio test that was performed for 20 claim intervals, whose lower bounds are based on every fifth quantile of the claims data, and whose p-values are displayed on the right y-axis of Figure 5.8a. Lower p-values mean that the null hypothesis can be rejected with a higher significance. The null hypothesis is that the unrestricted model (M3) does not explain the data better. The p-value for regions with claims between 700 and 64,500 (based on 155 observations) is $p_{value} = 0.116$, this can be seen in Figure 5.8a by looking at the leftmost black stem.

Figure 5.8b shows a residual error analysis, it is evident that the prediction errors are independent of the elements in the linear predictor $\eta_i$, the prediction errors were plotted against $E[\exp(u_i)]$ and $E[\exp(B_i\beta + \beta_0)]$ in Figure 5.8b (bottom) and Figure 5.8b (top) respectively.

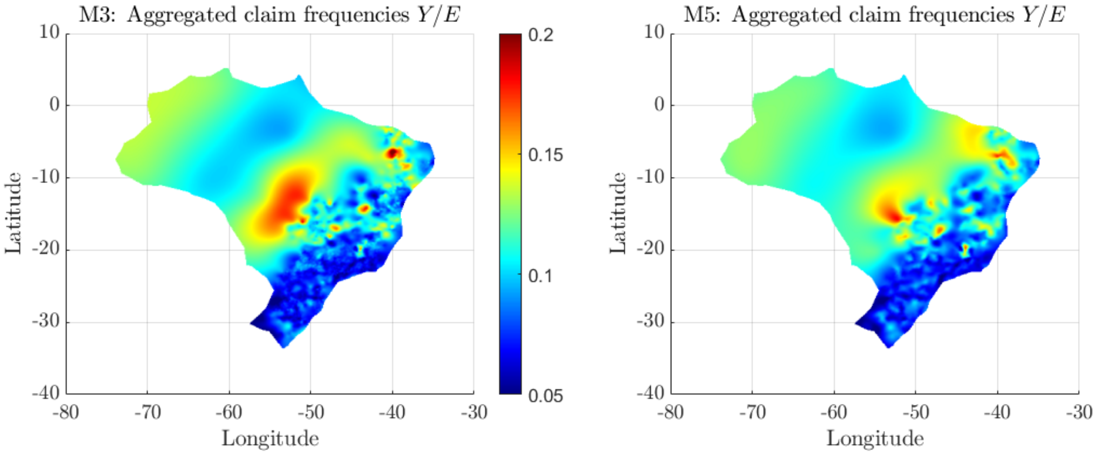


(a) Likelihood comparison between M2 and M3.

(b) M3: Prediction errors.

Figure 5.8: Likelihood comparison between model two and model three (5.8a). Prediction errors plotted against E[exp($\beta_0$ + $\boldsymbol{B\beta}$)] (5.8b top) and E[exp($u_i$)] (5.8b bottom) with a trend line in red.

Figure 5.9 shows the continuously modelled aggregated claim frequencies. Both model three and model five indicate that the regions west of the capital Brasilia and the regions between Boa Vista and São Bento are high risk regions from a vehicle damage insurance perspective. In Figure 5.10 the spatial resolution was increased from 0.1° × 0.1° [Long.×Lat.] (hereafter abbreviated as 0.1°) to 0.08°. A resolution of 0.1° generates 72,129 grid elements, results in the exclusion of 154 municipalities, and takes approximately 110 seconds for one training set to be fitted. A resolution of 0.08° results in 112,402 grid elements, excludes 58 municipalities, and takes approximately 140 seconds to fit. A resolution of 0.05° was attempted, generating 238,678 grid elements × 3,259 regions = 777,851,602 elements in the integration matrix, and excludes only 4 municipalities, but more than 16.1 gigabytes are needed to manipulate the matrix causing memory shortage. The 154 municipalities that are not included in model three, which was used in the 10-fold cross-validation, is deemed not to have affected the estimate of the spatial effect, as can be seen in Figure 5.10, where the field is estimated with 0.08° but does not show too much different spatial structure compared to the fields in Figure 5.9 with a resolution of 0.1°.


M3: Aggregated claim frequencies $Y/E$
M5: Aggregated claim frequencies $Y/E$
Latitude
Longitude
10
0
-10
-20
-30
-40
-80
-70
-60
-50
-40
-30
0.2
0.15
0.1
0.05

(a) Model three.

(b) Model five.

Figure 5.9: Comparison between model three and five for the aggregated claim frequencies.

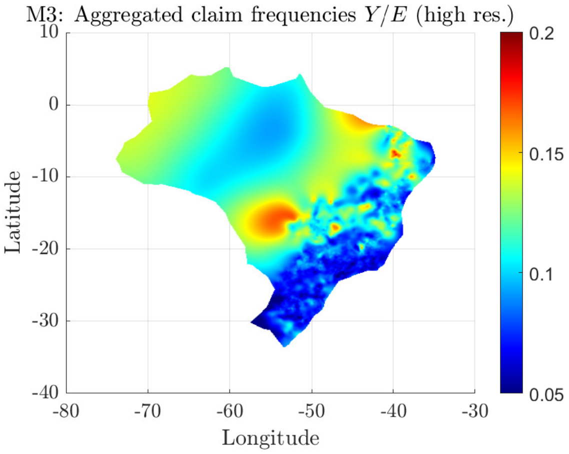
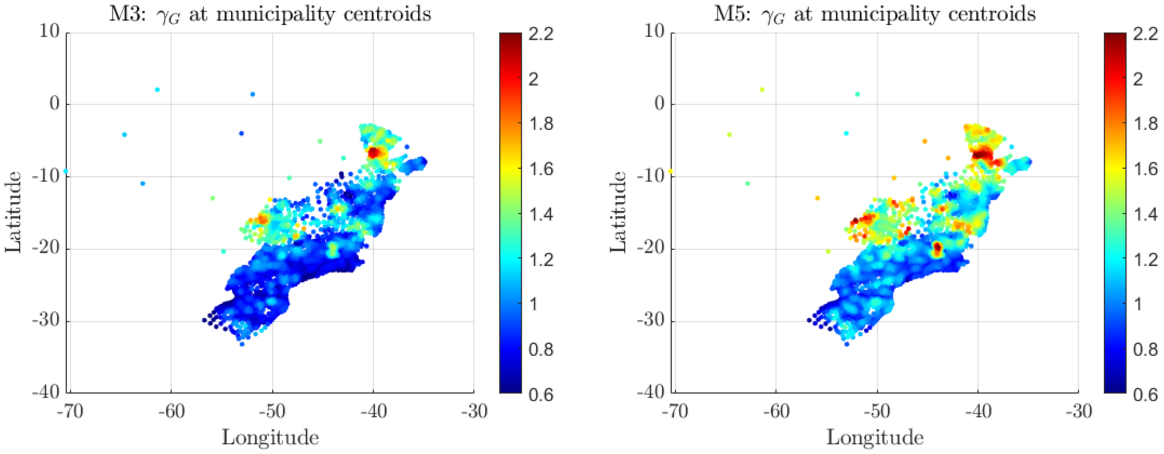


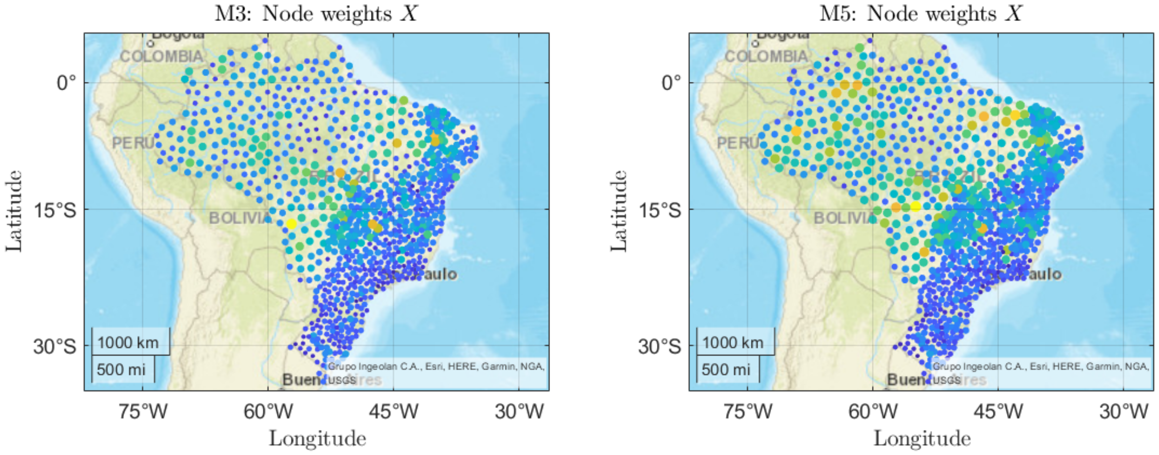
Figure 5.10: Aggregated claim frequencies over Brazil with 55.83% more grid elements used to discretise the GMRF in model three.



(a) Model three. (b) Model five.

Figure 5.11: Comparison between model three and five for the geographic rating factors $\gamma_G$ evaluated at the municipality centroids.

Figure 5.11 displays the geographic rating factors $\gamma_G$ evaluated at the municipality centroids, and Figure 5.12 shows the node weights from model three and model five that discretise the GMRF over a map of Brazil. It is apparent that the node weights $\boldsymbol{X}$ and geographic rating factors $\gamma_G$ are estimated similarly in both models.



(a) Model three. (b) Model five.

Figure 5.12: Comparison between model three and five for the node weights of the discretised GMRF.

# CHAPTER 6

# Discussion

In this chapter the methodology used to implement the GLMM that models the aggregated insurance claims for vehicle damages in Brazil is discussed, and the thesis is concluded by summarising the main findings.

In this thesis a Bayesian MCMC sampling methodology has been used to estimate the parameters of a GLMM that models aggregated insurance claims. The GLMM includes a mixture of three effects which were sampled in three blocks. The MMALA implementation used in the second block is technically referred to as simplified MMALA by Girolami and Calderhead, 2011. The unsimplified Riemann manifold Metropolis adjusted Langevin algorithm (MMALA) requires an additional third order derivative term of the log-likelihood w.r.t the parameters of the block to fully define the Brownian motion of the Langevin diffusion on a Riemann manifold with a metric tensor defined by the observed negative Fisher information. The third order term relates to changes in local curvature of the manifold and reduces to zero if the curvature of the log-likelihood is constant everywhere (Girolami & Calderhead, 2011). The simplified MMALA still generates a proposal process with local adaptation for the Brownian motion without the third order term, by scaling the size of the Brownian motion w.r.t the local curvature of the log-likelihood, this can be seen by looking at the variance of the transition kernel in equation (4.11). The simplified MMALA in conjunction with the acceptance probability still define a correct MCMC method that converges to the target distribution even if the curvature of the manifold is not constant (Girolami & Calderhead, 2011).

The lack of mapping grid elements caused by the specified resolution of the integration grid which excludes regions with an area smaller than 120 km$^2$, could potentially be resolved by using a Dirichlet tessellation (Voronoi diagram). The Dirichlet tessellation partitions the plane consisting of Brazil with its region centroids into convex polygons such that each polygon contains exactly one centroid. The convex polygons for regions smaller than the specified resolution could then be used as pixels to integrate the spatial effect. This partition would then generate convex polygons for an additional 154 regions in the data, and the total observed claims would increase to 875,864 claims from the current

824,379 claims. The Dirichlet tessellation could also lead to a reduced time complexity for the integration scheme.

The development of a spatial effect used in a GLMM derived from a GMRF is an attempt to improve on existing prediction models for insurance claims. The geographic rating factor derived from a GMRF has more relaxed requirements for the spatial data used to infer the geographic rating factor than the requirements of the method presented in (Tufvesson, 2016). An i.i.d lognormal effect was included to minimise overdispersion and thus recover less biased regression coefficient estimates in the Poisson mixed effects model. Failing to account for overdispersion in model one and two indicates to have inflated the estimates of the proportion of variance explained by the fixed effects, relative to when the i.i.d lognormal effect was included, which is in line with the conclusions of (Harrison, 2014). Ignoring overdispersion leads to reduced standard errors for the parameter estimates and narrower prediction intervals for the aggregated claims. Reduced standard errors for the parameter estimates could be dangerous when estimating capital requirements for an insurer, i.e. not accounting for excess kurtosis in the distribution of the aggregated claims when determining the value at risk could lead to an undervaluation of the value at risk. The need to account for excess kurtosis is nothing new and well established, e.g. in the established collective risk model used to model aggregate claims, the number of aggregated claims is given by a compound process, where: the number of claims is Poisson distributed with a structure variable q that represents short-term fluctuations which is assumed gamma distributed with equal parameters (Nino & Clemente, 2012), resulting in a NB2 model as demonstrated in section 3.2. In premium differentiation, the application of individual risk models is important for pure premium calculation in which the annual premium is increasingly determined according to the relevant individual characteristics of a policyholder (rating factors) (Valecký et al., 2017), the developed geographic rating factor in this thesis can be used for this purpose.

Future research could look at extending the model to include discontinuities in the terrain like water or other natural obstacles. This is a limitation of the current model since it does not account for natural obstacles. Implementing the proposed model on other data sets with different regions and different spatial scales will help to determine if the spatial effect gives consistently better claim frequency predictions in all types of geographical settings. A further extension of the model is to include temporal effects like seasonality. Seasonality caused by various factors like weather or holidays could possibly generate predictable patterns in the claim frequency levels. The GMRF used to model the claim frequencies is assumed to be stationary with fix field parameters $\kappa_x$ and $\tau_x$. The stationary assumption means that the spatial correlation is assumed to be the same throughout the domain. It is possible to consider a non-stationary GMRF by specifying spatially varying field parameters $\kappa_x(\boldsymbol{s})$ and $\tau_x(\boldsymbol{s})$. The non-stationary GMRF would account for topographical variables (e.g., altitude, river, lakes, etc.) that might influence the spatial dependence structure (Blangiardo & Cameletti, 2015, p. 197).

Summarising the findings; it is shown that the inclusion of a spatial effect yields better predictions and the likelihood ratio test presented in section 5.4 shows that the potential gain becomes more significant for larger Poisson observations. For the predictor variable selection, the Horseshoe priors shrunk rural and urban income levels, ratio of male residents, and number of households in the regions from which the geographic rating factor was inferred from. The focus of the predictor variable selection was on optimal forecasting to see if the spatial effect gave better predictions, hence the regression coefficient estimates presented in Figure 5.1 need to interpreted with caution from an inferential point of view. If two variables are highly correlated increases in one may be offset by decreases in another so the combined effect is to negate each other. This can cause an important predictor to become insignificant if it has a collinear relationship with other predictors. The results presented in Table 5.2 give a more stable estimation for the possible causation of changes in the claim frequencies caused by different principal component scores. The term possible causation is used as correlation does not imply causation. Especially model three exhibit conservative fix effect estimates which could be caused by the spatial effect being a stronger optimiser of the latent model-likelihood causing the fix effect to be reduced.

In conclusion of the thesis; modelling reality is complex and involves a lot of uncertainties. When there is uncertainty, risk can arise since risk does not exist by itself. To manage the uncertainties when choosing predictor variables for what has been observed, Horseshoe priors can be used. The inclusion of an i.i.d lognormal effect in the mixed effects model, infers a greater and more correct aleatory uncertainty which refers to the inherent uncertainty due to the probabilistic variability of the claim frequencies. And lastly, the inclusion of a spatial effect derived from a GMRF in a Bayesian framework has been shown to reduce epistemic uncertainty in claim frequency modelling.

# Bibliography

Amestoy, P. R., Davis, T. A., & Duff, I. S. (1996). An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications*, *17*(4), 886–905.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53), 370–418.

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r-inla*. John Wiley & Sons.

Boskov, M., & Verrall, R. (1994). Premium rating by geographic area using spatial models. *ASTIN Bulletin: The Journal of the IAA*, *24*(1), 131–143.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. *Artificial Intelligence and Statistics*, 73–80.

EIOPA. (2021). *Mission and tasks*. Retrieved October 15, 2021, from https://www.eiopa.europa.eu/about/mission-and-tasks_en

EU. (2009). *Directive 2009/138/ec of the european parliament and of the council of 25 november 2009 on the taking-up and pursuit of the business of insurance and reinsurance (solvency ii)*. Retrieved October 15, 2021, from https://eur-lex.europa.eu/legal-content/SV/TXT/?uri=celex:32009L0138

FI. (2020). *Insurance distribution*. Retrieved October 15, 2021, from https://www.fi.se/en/insurance/apply-for-authorisation/insurance-distribution/

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

Gelman, A., Gilks, W. R., & Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, *7*(1), 110–120.

Girolami, M., & Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(2), 123–214.

Givens, G. H., & Hoeting, J. A. (2012). *Computational statistics* (Vol. 703). John Wiley & Sons.

Gschlößl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, *2007*(3), 202–225.

Hairer, M., Stuart, A. M., & Vollmer, S. J. (2014). Spectral gaps for a metropolis–hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, *24*(6), 2455–2490.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. Stata press.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

IBGE. (2010). *Index of censos censo demografico 2010*. Retrieved October 29, 2021, from https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/tabelas_pdf/

IBGE. (2020). *Municipal mesh*. Retrieved October 29, 2021, from https://www.ibge.gov.br/en/geosciences/territorial-organization/territorial-meshes/18890-municipal-mesh.html?=&t=sobre

If P&C Insurance. (2021). *If, annual report 2020*. Retrieved October 15, 2021, from https://www.sampo.com/investors/financial-information/annual-reports/?s=if-pc-insurance

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of statistical software*, *63*(1), 1–25.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Lindström, E., Madsen, H., & Nielsen, J. N. (2018). *Statistics for finance*. CRC Press.

Makalic, E., & Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, *23*(1), 179–182.

Makalic, E., & Schmidt, D. F. (2016). High-dimensional bayesian regularised regression with the bayesreg package. *arXiv preprint arXiv:1611.06649*.

Matern, B. et al. (1960). Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran Statens Skogsforskningsinstitut*, *49*(5).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Miles, J. (2014). Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online*.

Moraga, P. (2019). *Geospatial health data: Modeling and visualization with r-inla and shiny*. Chapman; Hall/CRC.

Nino, S., & Clemente, G. P. (2012). Modelling aggregate non life underwriting risk : Standard formula vs internal model. *Giornale dell'Istituto Italiano degli attuari*, 301–339.

O'Donnell, E. (2005). Enterprise risk management: A systems-thinking framework for the event identification phase. *International Journal of Accounting Information Systems*, *6*(3), 177–195.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 174). Springer.

Patterson, S., & Teh, Y. W. (2013). Stochastic gradient riemannian langevin dynamics on the probability simplex. *NIPS*, 3102–3110.

Petersen, K. B., Pedersen, M. S. et al. (2008). The matrix cookbook. *Technical University of Denmark*, *7*(15), 510.

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Regeringskansliet. (2020). *How sweden is governed.* Retrieved October 15, 2021, from https://www.regeringen.se/other-languages/english---how-sweden-is-governed/

Robert, C., & Casella, G. (2013). *Monte carlo statistical methods.* Springer Science & Business Media.

Roberts, G. O., & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 255–268.

Roberts, G. O., & Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 341–363.

Styrud, L. (2017). Risk premium prediction of car damage insurance using artificial neural networks and generalized linear models. *Master's Thesis in Mathematical Sciences.*

Surojit, B. (2013). Negative binomial regression. https://www.mathworks.com/matlabcentral/fileexchange/40642-negative-binomial-regression

SUSEP. (2015). *Autoseg - susep automobile statistics system.* Retrieved October 29, 2021, from http://www2.susep.gov.br/menuestatistica/Autoseg/menu2.aspx

Svensson, L. (2019). Reconstruction of past european land cover from pollen data: Using spatial statistics and crank-nicolson monte carlo. *Master's Thesis in Mathematical Sciences.*

Tufvesson, O. (2016). Spatial statistical modelling of insurance risk: A spatial epidemiological approach to car insurance. *Master's Thesis in Mathematical Sciences.*

Tufvesson, O., Lindström, J., & Lindström, E. (2019). Spatial statistical modelling of insurance risk: A spatial epidemiological approach to car insurance. *Scandinavian Actuarial Journal*, *2019*(6), 508–522.

Valecký, J. et al. (2017). Calculation of solvency capital requirements for non-life underwriting risk using generalized linear models. *Prague Economic Papers*, *26*(4), 450–466.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 434–449.

# APPENDIX A

## The Metropolis within Gibbs algorithm

In this section the Metropolis Hastings algorithm is first presented, then the Gibbs algorithm is presented, and last the Metropolis within Gibbs algorithm is presented as a combination of the two.

Let there be a probability distribution for the parameter $\beta$ which is proportional to the un-normalised probability distribution of $\beta$, i.e $p(\beta) \propto \hat{p}(\beta)$, then the Metropolis Hastings algorithm follows as

**Algorithm 1:** Metropolis Hastings algorithm

```
Result: A sample of the r.v $\beta$ with size M from the un-normalised probability
        distribution $\hat{p}(\beta)$ that is representative of the probability density
        function $p(\beta)$.
M=N+burn in
$\beta(1) = \beta_0$ is some reasonable initial value
$\delta = \delta_{opt}$ is chosen such that $\alpha \approx 0.24$ or $\alpha \approx 0.57$ if Langevin diffusion is used
for $i = 1$ to $(M - 1)$ do
    $\beta^* \sim q(\beta^*|\beta)$
    $\alpha = \min\left\{1, \frac{\tilde{p}(\beta^*)q(\beta(i)|\beta^*)}{\tilde{p}(\beta(i))q(\beta^*|\beta(i))}\right\}$ if $u \leq \alpha$ where $u \sim \mathcal{U}(0,1)$ then
        $\beta(i+1) = \beta^*$
    else
        $\beta(i+1) = \beta(i)$
    end
end
```

For the Gibbs algorithm assume that there are two random variables A and B for which the joint distribution $\mathbb{P}(A, B)$ is unknown but the conditional distributions $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are known. The Gibbs algorithm then starts by sampling $(A_0, B_0) \sim \pi(\cdot)$ from some probability distribution, which has support over the allowed values of A and B. The Gibbs algorithm then iterates for M iterations and does the following: first it samples a value $A_{i+1}$ from the conditional probability $A_{i+1} \sim \mathbb{P}(A|B_i)$, then it uses this value of $A_{i+1}$ to define a conditional probability distribution of B and samples $B_{i+1}$, $B_{i+1} \sim \mathbb{P}(B|A_{i+1})$.

**Algorithm 2:** Gibbs algorithm

```
Result: A sample of size M from the conditional probability densities $\mathbb{P}(A|B)$
        and $\mathbb{P}(B|A)$ that represents the joint probability density $\mathbb{P}(A,B)$, i.e.
        $f_{A,B}(a,b)$.
M=N+burn in
$A(1) = A_0$
$B(1) = B_0$
for $i = 1$ to $(M - 1)$ do
    $A_{i+1} \sim \mathbb{P}(A|B_i)$
    $B_{i+1} \sim \mathbb{P}(B|A_{i+1})$
end
```

The Gibbs algorithm could be viewed as a Metropolis Hastings algorithm with a 100 percent acceptance rate.

It is often convenient to consider combinations of the Gibbs and the Metropolis Hastings algorithm when some blocks of a target distribution has a tractable conditional posterior whereas for other blocks the conditional posterior is intractable. This combination yields the **Metropolis within Gibbs algorithm** (Hybrid MCMC) which is performed by the following steps:

- Divide the target distribution into blocks and aim for Gibbs sampling.
- If the conditional distribution of a block is known, update according to Gibbs.
- If there are blocks for which we cannot find a closed form of the conditional distribution, insert a MH-step instead.

# APPENDIX B

# The acceptance rate for MMALA

From equation (4.11) we get the MMALA transition kernel for the proposals $\boldsymbol{\beta}^*$ given the current state $\boldsymbol{\beta}$ as

$$q(\boldsymbol{\beta}^*|\boldsymbol{\beta}) = \mathcal{MVN}\Big(\frac{\mathcal{K}}{2}\nabla \log \pi(\boldsymbol{\beta})\delta + \boldsymbol{\beta},\, \mathcal{K}\delta\Big), \tag{B.1}$$

where the preconditioner is set to the observed negative Fisher information $\mathcal{K}^{-1}(\boldsymbol{\beta}) = -\Delta \log\pi(\boldsymbol{\beta})$ derived in appendix B.1.

The probability density function for a multivariate normal distribution parametrised with the precision matrix $\boldsymbol{Q}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}$ is

$$\mathcal{MVN}(\boldsymbol{\beta}^*; \boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{Q}(\boldsymbol{\beta})) = \frac{\sqrt{|\boldsymbol{Q}(\boldsymbol{\beta})|}}{\sqrt{(2\pi)^d}} \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}^* - \boldsymbol{\mu}(\boldsymbol{\beta})\big)^T \boldsymbol{Q}(\boldsymbol{\beta})\big(\boldsymbol{\beta}^* - \boldsymbol{\mu}(\boldsymbol{\beta})\big)\Big). \tag{B.2}$$

For the MMALA the resulting mean and precision for the proposal $\boldsymbol{\beta}^*$ thus depend on the current state $\boldsymbol{\beta}$ as

$$\boldsymbol{\mu}(\boldsymbol{\beta}^*) = \boldsymbol{\beta} + \frac{\delta}{2}\mathcal{K}(\boldsymbol{\beta})^{-1}\nabla \log \pi(\boldsymbol{\beta}) \tag{B.3}$$

$$\boldsymbol{Q}(\boldsymbol{\beta}^*) = \frac{1}{\delta}\mathcal{K}(\boldsymbol{\beta}). \tag{B.4}$$

The correction factor then becomes

$$\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^*|\boldsymbol{\beta})} = \frac{\sqrt{|\boldsymbol{P}(\boldsymbol{\beta}^*)|}\, \exp\Big(-\frac{1}{2\delta}\big(\boldsymbol{\beta} - \boldsymbol{\mu}(\boldsymbol{\beta}^*)\big)^T \boldsymbol{P}(\boldsymbol{\beta}^*)\big(\boldsymbol{\beta} - \boldsymbol{\mu}(\boldsymbol{\beta}^*)\big)\Big)}{\sqrt{|\boldsymbol{P}(\boldsymbol{\beta})|}\, \exp\Big(-\frac{1}{2\delta}\big(\boldsymbol{\beta}^* - \boldsymbol{\mu}(\boldsymbol{\beta})\big)^T \boldsymbol{P}(\boldsymbol{\beta})\big(\boldsymbol{\beta}^* - \boldsymbol{\mu}(\boldsymbol{\beta})\big)\Big)} \tag{B.5}$$

The quota of the probabilities from the stationary distributions from the first block in section 4.1 multiplied with the correction factor (B.5) then gives the acceptance rate for the MMALA

$$\alpha_{acc}(\boldsymbol{\beta}^*, \boldsymbol{\beta}) = \frac{\pi\big(\boldsymbol{\beta}^*\big)q\big(\boldsymbol{\beta}|\boldsymbol{\beta}^*\big)}{\pi\big(\boldsymbol{\beta}\big)q\big(\boldsymbol{\beta}^*|\boldsymbol{\beta}\big)}, \tag{B.6}$$

where the proposals are drawn using the Cholesky factorisation of (B.4)

$$\begin{aligned}\boldsymbol{\beta}^* &= \boldsymbol{\mu}(\boldsymbol{\beta}^*) + \boldsymbol{R}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{I}) && \text{(B.7)}\\ \boldsymbol{R}(\boldsymbol{\beta}^*) &= \text{chol}(\boldsymbol{Q}(\boldsymbol{\beta}^*)). && \text{(B.8)}\end{aligned}$$

## B.1 Derivation of $\nabla\mathcal{L}(\beta)$

In this section the gradient of the log-posterior w.r.t the regression coefficents is derived. Its purpose it to propose new parameters that maximise the likelihood in an efficient manner in the MMALA algorithm.

Let

$$\mathcal{L}(\boldsymbol{\beta}) :\propto \log\Big\{\mathbb{P}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\lambda_\beta}, \tau_\beta)\Big\}. \tag{B.9}$$

$$\mathcal{L}(\boldsymbol{\beta}) = \log\Bigg(\Bigg(\prod_{i=1}^{N_{obs}} \eta_i^{y_i} \cdot \exp(-\eta_i)\Bigg) \cdot \exp(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta})\Bigg). \tag{B.10}$$

$$= \Big(\underbrace{\sum_{i=1}^{N_{obs}} y_i \cdot \log(\eta_i) - \eta_i}_{\propto \text{ Poisson part}} \quad \underbrace{-\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta})}_{\propto \text{ HS regularisation part}} \quad \Big). \tag{B.11}$$

$$\boldsymbol{\eta} = \exp\Big(\boldsymbol{B\beta} + \log\big(\boldsymbol{FA} \cdot \exp(\boldsymbol{X})\big) + \log(\boldsymbol{E})\Big). \tag{B.12}$$

Define a constant w.r.t $\boldsymbol{\beta}$, consisting of the spatial field and the offset as

$$\boldsymbol{c}_{\beta[\text{Nobs}\times 1]} := \exp\Big(\log\big(\boldsymbol{FA} \cdot \exp(\boldsymbol{X})\big) + \log(\boldsymbol{E})\Big) \tag{B.13}$$

Then $\boldsymbol{\eta} = \boldsymbol{c_\beta} \odot \exp(\boldsymbol{B\beta})$, where $\odot$ is the Hadamard (elementwise) product.

Define inner function $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{B\beta}$ and apply the chain rule to the Poisson part.

$$\begin{aligned} \frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\beta}}_{[1\times \text{N}\beta]} &= \frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\eta}}_{[1\text{xNobs}]} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{g}}_{[\text{Nobs}\times\text{Nobs}]} \cdot \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}}_{[\text{Nobs}\times\text{N}\beta]} \\ \frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\eta}} &= \frac{\partial}{\partial \boldsymbol{\eta}}\Bigg(\sum_{i=1}^{N_{obs}} y_i \cdot \log(\eta_i) - \eta_i\Bigg) = \sum_{i=1}^{N_{obs}} \Big(\boldsymbol{Y} \oslash \boldsymbol{\eta} - \mathbb{1}\Big)_i, \end{aligned} \tag{B.14}$$

where $\oslash$ is the Hadamard (elementwise) division.

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{g}} = \frac{\partial}{\partial \boldsymbol{g}}\Bigg(\boldsymbol{c_\beta} \odot \exp(\boldsymbol{g})\Bigg) = \boldsymbol{c_\beta} \odot \exp(\boldsymbol{g}). \tag{B.15}$$

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\Big(\boldsymbol{B\beta}\Big) = \boldsymbol{B}. \tag{B.16}$$

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{g}} \odot \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}} = \boldsymbol{c_\beta} \odot \exp(\boldsymbol{g}) \odot \boldsymbol{B}. \tag{B.17}$$

$$\frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\beta}} = \frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \Big(\boldsymbol{Y} \oslash \boldsymbol{\eta} - \mathbb{1}\Big)^T \cdot \boldsymbol{c_\beta} \odot \exp(\boldsymbol{g}) \odot \boldsymbol{B}. \tag{B.18}$$

Equation (97) in (Petersen, Pedersen, et al., 2008) gives the Jacobian for the HS regularisation part.

$$\frac{\partial l_{\text{HS}}}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\Bigg(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}\Bigg) = -\frac{1}{2}(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T)\boldsymbol{\beta} = -\boldsymbol{\Lambda}\boldsymbol{\beta}. \tag{B.19}$$

with the last step due to $\mathbf{\Lambda}$ being a diagonal matrix. Combining the Poisson with the HS regularisation part gives,

$$\nabla l(\boldsymbol{\beta}) = \frac{\partial l_{\text{Po}}}{\partial \boldsymbol{\beta}} + \frac{\partial l_{\text{HS}}}{\partial \boldsymbol{\beta}} = \left(\left(\boldsymbol{Y} \oslash \boldsymbol{\eta} - \mathbb{1}\right)^T \cdot \boldsymbol{c}_{\beta} \odot \exp(\boldsymbol{\beta}^T)\right)^T - \mathbf{\Lambda}\boldsymbol{\beta}. \tag{B.20}$$

## B.2 Derivation of $\Delta\mathcal{L}(\beta)$

In this section the observed negative Fisher information $-\Delta\mathcal{L}(\beta)$ of the log-posterior w.r.t the regression coefficents is derived. Its purpose is to define the Langevin diffusion on a Riemann manifold which allows for local adaptation of the proposal in the negative-log-posterior space.

Let

$$\Delta\mathcal{L}(\boldsymbol{\beta}) :\propto \frac{\partial^2}{\partial\beta_i\beta_j}\log\left\{\mathbb{P}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\lambda}_{\boldsymbol{\beta}}, \tau_\beta)\right\}. \tag{B.21}$$

We earlier derived that,

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) = \underbrace{\left(\left(\boldsymbol{Y} \oslash \boldsymbol{\eta} - \mathbb{1}\right)^T \cdot \boldsymbol{c}_{\boldsymbol{\beta}} \odot \exp(\boldsymbol{\beta}^T)\right)^T}_{\text{Poisson part}} - \underbrace{\mathbf{\Lambda}\boldsymbol{\beta}}_{\text{HS regularisation part}} \tag{B.22}$$

Expanding the Poisson part gives

$$\frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_i\partial\beta_j} = \begin{bmatrix} \frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_1\partial\beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_n\partial\beta_1} & \cdots & \frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_n\partial\beta_n} \end{bmatrix}. \tag{B.23}$$

And expanding one second order partial derivative.

$$\begin{gathered} \frac{\partial^2\mathcal{L}_{\text{Po}}}{\partial\beta_i\partial\beta_j} = \frac{\partial}{\partial\beta_i}\underbrace{\Big((1 - c_1\exp(g_1))B_{1,j} + \ldots + (1 - c_n\exp(g_n))B_{n,j}\Big)}_{\partial\mathcal{L}_{\text{Po}}/\partial\beta_j} = \\ -c_1\exp(g_1)B_{1,j}B_{1,i} - \ldots - c_n\exp(g_n)B_{n,j}B_{n,i}. \end{gathered} \tag{B.24}$$

The elements of the Hessian for the Poisson part thus become

$$\frac{\partial^2\mathcal{L}_{Po}}{\partial\beta_i\partial\beta_j} = -\sum_{k=1}^{N_{obs}}\Big([\boldsymbol{c}_\beta \odot \exp(\boldsymbol{B}\boldsymbol{\beta})]_k \odot \boldsymbol{B}_{kj} \odot \boldsymbol{B}_{kj}\Big), \tag{B.25}$$

and the HS regularisation part becomes

$$\frac{\partial^2\mathcal{L}_{\text{HS}}}{\partial\beta_i\partial\beta_j} = \frac{\partial}{\partial\beta_i}\Big([-\mathbf{\Lambda}\boldsymbol{\beta}]_j\Big) = -\Lambda_{ij}. \tag{B.26}$$

Finally, combining the Poisson with the HS regularisation part gives,

$$\Delta\mathcal{L}(\boldsymbol{\beta})_{ij} = \frac{\partial^2\mathcal{L}_{\mathrm{Po}}}{\partial\beta_i\partial\beta_j} + \frac{\partial^2\mathcal{L}_{\mathrm{HS}}}{\partial\beta_i\partial\beta_j} = -\sum_{k=1}^{N_{obs}}\Big([\boldsymbol{c}_\beta \odot \exp(\boldsymbol{B}\boldsymbol{\beta})]_k \odot \boldsymbol{B}_{kj} \odot \boldsymbol{B}_{ki}\Big) - \boldsymbol{\Lambda}_{ij}, \tag{B.27}$$

or equivalently put

$$-\Delta\mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{B}^T \cdot \Big(\boldsymbol{c}_\beta \odot \exp(\boldsymbol{B}\boldsymbol{\beta}) \odot \boldsymbol{B}\Big) + \boldsymbol{\Lambda}. \tag{B.28}$$

# APPENDIX C

## The acceptance rate for pCNL

The pCNL discretises the linear part of the NLP $\nabla\Phi_{\mathrm{GMRF}}(\boldsymbol{X}(t))$ (4.13) in the following way

$$\nabla\Phi_{\mathrm{GMRF}}(\boldsymbol{X}(t)) \approx \frac{1}{2}\Big(\underbrace{\nabla\Phi_{\mathrm{GMRF}}(\boldsymbol{X}(t+\delta)) + \nabla\Phi_{\mathrm{GMRF}}(\boldsymbol{X}(t))}_{\text{Central difference}}\Big) \tag{C.1}$$

This yields the following discretisation of the SDE (4.12)

$$\begin{gathered}\boldsymbol{X}(t+\delta) - \boldsymbol{X}(t) = \\ \frac{-\mathcal{K}}{2}\left(\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t)) + \frac{1}{2}\Big(\underbrace{\boldsymbol{Q}\boldsymbol{X}(t+\delta) + \boldsymbol{Q}\boldsymbol{X}(t)}_{\text{Central difference}}\Big)\right)\delta + \sqrt{\mathcal{K}}(\mathcal{W}(t+\delta) - \mathcal{W}(t))\end{gathered} \tag{C.2}$$

Solving for $\boldsymbol{X}(t+\delta)$ yields

$$\Big(\boldsymbol{I} + \frac{\delta\mathcal{K}\boldsymbol{Q}}{4}\Big)\boldsymbol{X}(t+\delta) = -\frac{\mathcal{K}}{2}\left(\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t)) + \frac{\boldsymbol{Q}\boldsymbol{X}(t)}{2}\right)\delta + \boldsymbol{X}(t) + \sqrt{\mathcal{K}}(\mathcal{W}(t+\delta) - \mathcal{W}(t)) \tag{C.3}$$

$$\boldsymbol{X}(t+\delta) = \Big(\boldsymbol{I} + \frac{\delta\mathcal{K}\boldsymbol{Q}}{4}\Big)^{-1}\left[-\frac{\mathcal{K}}{2}\left(\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t)) + \frac{\boldsymbol{Q}\boldsymbol{X}(t)}{2}\right)\delta + \boldsymbol{X}(t) + \sqrt{\mathcal{K}}(\mathcal{W}(t+\delta) - \mathcal{W}(t))\right] \tag{C.4}$$

Now using $\mathcal{K} = \boldsymbol{Q}^{-1}$

$$\boldsymbol{X}(t+\delta) = \frac{4}{4+\delta}\left[ - \frac{\boldsymbol{Q}^{-1}}{2}\left(\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t)) + \frac{\boldsymbol{Q}\boldsymbol{X}(t)}{2}\right)\delta + \boldsymbol{X}(t) + \sqrt{\boldsymbol{Q}^{-1}}\Big(\mathcal{W}(t+\delta) - \mathcal{W}(t)\Big)\right] \tag{C.5}$$

$$= \frac{1}{4+\delta}\left( - 2\boldsymbol{Q}^{-1}\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t))\delta - \boldsymbol{X}(t)\delta + 4\boldsymbol{X}(t) + 4\sqrt{\boldsymbol{Q}^{-1}}\Big(\mathcal{W}(t+\delta) - \mathcal{W}(t)\Big)\right) \tag{C.6}$$

$$E[\boldsymbol{X}(t+\delta)] = \frac{1}{4+\delta}\left( - 2\boldsymbol{Q}^{-1}\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}(t))\delta + (4-\delta)\boldsymbol{X}(t)\right) \tag{C.7}$$

$$V[\boldsymbol{X}(t+\delta)] = 16\boldsymbol{Q}^{-1}\delta. \tag{C.8}$$

which implies sampling from

$$\boldsymbol{X}_{i+1} \sim \mathcal{MVN}\left(\frac{1}{4+\delta}\left( - 2\boldsymbol{Q}^{-1}\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X}_i)\delta + (4-\delta)\boldsymbol{X}_i\right), 16\boldsymbol{Q}^{-1}\delta\right). \tag{C.9}$$

For the pCNL only the resulting mean for the proposal $\boldsymbol{X}^*$ thus depend on the current state $\boldsymbol{X}$ as

$$\boldsymbol{\mu}(\boldsymbol{X}^*) = \frac{1}{4+\delta}\left( - 2\boldsymbol{Q}^{-1}\nabla\Phi_{\mathrm{Po}}(\boldsymbol{X})\delta + (4-\delta)\boldsymbol{X}\right) \tag{C.10}$$

which gives the correction factor

$$\frac{q(\boldsymbol{X}|\boldsymbol{X}^*)}{q(\boldsymbol{X}^*|\boldsymbol{X})} = \frac{\exp\Big( - \frac{1}{32\delta}\big(\boldsymbol{X} - \boldsymbol{\mu}(\boldsymbol{X}^*)\big)^T\boldsymbol{Q}\big(\boldsymbol{X} - \boldsymbol{\mu}(\boldsymbol{X}^*)\big)\Big)}{\exp\Big( - \frac{1}{32\delta}\big(\boldsymbol{X}^* - \boldsymbol{\mu}(\boldsymbol{X})\big)^T\boldsymbol{Q}\big(\boldsymbol{X}^* - \boldsymbol{\mu}(\boldsymbol{X})\big)\Big)}. \tag{C.11}$$

The quota of the probabilities from the stationary distributions from the second block in section 4.1 multiplied with the correction factor (C.11) then gives the acceptance rate for the pCNL

$$\alpha_{acc}(\boldsymbol{X}^*, \boldsymbol{X}) = \frac{\pi\big(\boldsymbol{X}^*\big)q\big(\boldsymbol{X}|\boldsymbol{X}^*\big)}{\pi\big(\boldsymbol{X}\big)q\big(\boldsymbol{X}^*|\boldsymbol{X}\big)}, \tag{C.12}$$

where the proposals are drawn using the Cholesky factorisation

$$\boldsymbol{X}^* = \boldsymbol{\mu}(\boldsymbol{X}^*) + \boldsymbol{R}(\boldsymbol{X}^*)^{-1}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{I}) \tag{C.13}$$

$$\boldsymbol{R}(\boldsymbol{X}^*) = \mathrm{chol}(\boldsymbol{Q}). \tag{C.14}$$

## C.1 Derivation of $\nabla\Phi(X)$

The $\Phi(X)$ function goes by many names; the function maps the negative-log-posterior space from the parameters of the model. In a Hamiltonian setting it is referred to as the potential and in machine learning it goes under the name "loss function". In this section it will be referred to as the negative-log-posterior (NLP). Below the gradient of the NLP w.r.t the node weights, $\boldsymbol{X}$ is derived. Its purpose is to give new proposed parameters that maximise the likelihood for our observations and is used in the pCNL algorithm.

Let

$$\Phi(\boldsymbol{X}) :\propto -\log\Big\{\mathbb{P}(\boldsymbol{X}|\boldsymbol{Y}, \kappa_x, \tau_x)\Big\}. \tag{C.15}$$

$$\Phi(\boldsymbol{X}) = -\log\Bigg(\Bigg(\prod_{i=1}^{N_{obs}} \eta_i^{y_i} \cdot \exp(-\eta_i)\Bigg) \cdot \exp(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X})\Bigg). \tag{C.16}$$

$$= -\Big( \underbrace{\sum_{i=1}^{N_{obs}} y_i \cdot \log(\eta_i) - \eta_i}_{\propto \text{ Poisson part}} \quad \underbrace{-\frac{1}{2}(\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X})}_{\propto \text{ Latent GMRF part}} \Big). \tag{C.17}$$

$$\boldsymbol{\eta} = \exp\Big(\boldsymbol{B\beta} + \log\big(\boldsymbol{FA} \cdot \exp(\boldsymbol{X})\big) + \log(\boldsymbol{E})\Big). \tag{C.18}$$

Define constant w.r.t $\boldsymbol{X}$, consisting of the fix effect, offset and pixel mapping as

$$\boldsymbol{c}_{x[\text{Nobs}\times\text{Nx}]} := \exp\big(\boldsymbol{B\beta} + \log(\boldsymbol{E})\big) \odot \boldsymbol{FA}, \tag{C.19}$$

where $\odot$ is the Hadamard (elementwise) product.

Then then linear predictor can be written as

$$\boldsymbol{\eta} = \boldsymbol{c_x} \cdot \exp(\boldsymbol{X}). \tag{C.20}$$

Apply the chain rule to the Poisson part.

$$\frac{\partial\Phi_{\text{Po}}}{\partial\boldsymbol{X}}_{[1\times\text{Nx}]} = \frac{\partial\Phi_{\text{Po}}}{\partial\boldsymbol{\eta}}_{[1\times\text{Nx}]} \cdot \frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{X}}_{[\text{Nobs}\times\text{Nx}]} \tag{C.21}$$

$$\frac{\partial\Phi_{\text{Po}}}{\partial\boldsymbol{\eta}} = \frac{\partial}{\partial\boldsymbol{\eta}}\Bigg(\sum_{i=1}^{N_{obs}} \eta_i - y_i \cdot \log(\eta_i)\Bigg) = \sum_{i=1}^{N_{obs}} \Big(\mathbb{1} - \boldsymbol{Y} \oslash \boldsymbol{\eta}\Big)_i, \tag{C.22}$$

where $\oslash$ is the Hadamard (elementwise) division.

$$\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{X}} = \frac{\partial}{\partial\boldsymbol{X}}\Bigg(\boldsymbol{c_x} \cdot \exp(\boldsymbol{X})\Bigg) = \boldsymbol{c_x} \odot \exp(\boldsymbol{X}^T). \tag{C.23}$$

$$\frac{\partial\Phi_{\text{Po}}}{\partial\boldsymbol{X}} = \frac{\partial\Phi_{\text{Po}}}{\partial\boldsymbol{\eta}} \cdot \frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{X}} = \Big(\mathbb{1} - \boldsymbol{Y} \oslash \boldsymbol{\eta}\Big)^T \cdot \boldsymbol{c_x} \odot \exp(\boldsymbol{X}^T). \tag{C.24}$$

Using equation (97) in (Petersen, Pedersen, et al., 2008) gives the Jacobian for the latent GMRF part.

$$\frac{\partial \Phi_{\text{GMRF}}}{\partial \boldsymbol{X}} = \frac{\partial}{\partial \boldsymbol{X}}\left(\frac{1}{2}\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X}\right) = \frac{1}{2}(\boldsymbol{Q} + \boldsymbol{Q}^T)\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{X}. \tag{C.25}$$

The last step follows since the precision matrix $\boldsymbol{Q}$ is symmetric. Finally, combining the Poisson with the GMRF part gives,

$$\nabla\Phi(\boldsymbol{X}) = \frac{\partial \Phi_{\text{Po}}}{\partial \boldsymbol{X}} + \frac{\partial \Phi_{\text{GMRF}}}{\partial \boldsymbol{X}} = \left(\left(\mathbb{1} - \boldsymbol{Y} \oslash \boldsymbol{\eta}\right)^T \cdot \boldsymbol{c}_{\boldsymbol{x}} \odot \exp(\boldsymbol{X}^T)\right)^T + \boldsymbol{Q}\boldsymbol{X}. \tag{C.26}$$

# APPENDIX D

## The conditional posterior for the latent model

The derivation of the conditional posterior for the latent model $\mathbb{P}(\cdot|\boldsymbol{Y})$ given the observations $\boldsymbol{Y}$ follows below, where $\cdot$ denotes the conditioning on all other parameters of the model .

Using the property of conditional probability; the prior distribution $\mathbb{P}(\cdot)$ and the sampling distribution $\mathbb{P}(\boldsymbol{Y}|\cdot)$ describes the joint probability probabilities as

$$\mathbb{P}(\cdot,\boldsymbol{Y}) = \mathbb{P}(\cdot)\mathbb{P}(\boldsymbol{Y}|\cdot) \tag{D.1}$$

yielding Bayes theorem (Bayes, 1763; Gelman et al., 1995)

$$\mathbb{P}(\cdot|\boldsymbol{Y}) = \frac{\mathbb{P}(\cdot,\boldsymbol{Y})}{\mathbb{P}(\boldsymbol{Y})} = \frac{\mathbb{P}(\boldsymbol{Y}|\cdot)\mathbb{P}(\cdot)}{\mathbb{P}(\boldsymbol{Y})} \tag{D.2}$$

using the properties form above; a hierarchical breakdown of the model introduced in section 3.6 gives the conditional posterior as

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1|\boldsymbol{Y})\\
&= \frac{\mathbb{P}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1,\boldsymbol{Y})}{\mathbb{P}(\boldsymbol{Y})}\\
&= \frac{\mathbb{P}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1)\cdot\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1)}{\mathbb{P}(\boldsymbol{Y})}
\end{aligned} \tag{D.3}$$

Where some conditional dependencies and joint distributions cancel due to the specified hierarchy in figure 3.5 yielding the conditional posterior for the latent model up to proportionality

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\boldsymbol{\nu},\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1|\boldsymbol{Y})\\
&\propto \mathbb{P}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\lambda}_\beta,\tau_\beta,\kappa_x,\tau_x,\boldsymbol{\nu},\xi,\boldsymbol{v},q,\alpha_v,m_v,\psi_1)\\
&\cdot\mathbb{P}(\boldsymbol{X}|\kappa_x,\tau_x)\cdot\mathbb{P}(\boldsymbol{\beta},\boldsymbol{v}|q,\alpha_v,m_v,\boldsymbol{\lambda}_\beta,\tau_\beta,\boldsymbol{\nu},\xi,\psi_1)\\
&\cdot\mathbb{P}(\boldsymbol{\lambda}_\beta|\boldsymbol{\nu})\cdot\mathbb{P}(\tau_\beta|\xi)\cdot\mathbb{P}(q|\alpha_v,m_v)\cdot\mathbb{P}(\kappa_x,\tau_x)\cdot\mathbb{P}(\boldsymbol{\nu})\cdot\mathbb{P}(\xi)\\
&\propto \prod_{i=1}^{n}\eta_i^{y_i}\exp(-\eta_i)\cdot\exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}})\cdot\exp(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X})\cdot\prod_{i=1}^{p}\nu_i^{-1/2}\lambda_{\beta_i}^{-3/2}\exp(-\frac{1}{\nu_i\cdot\lambda_{\beta_i}})\\
&\cdot\xi^{-1/2}\tau_\beta^{-3/2}\exp(-\frac{1}{\xi\cdot\tau_\beta})\cdot\frac{1}{\hat{b}^{\hat{a}}\Gamma(\hat{a})}q^{\hat{a}-1}\exp(\frac{-q}{\hat{b}}),
\end{aligned}\tag{D.4}$$

where

$$\hat{a}=\alpha_v+\frac{n}{2},\ \hat{b}=\frac{1}{m_v+\frac{1}{2}\sum\limits_{i=1}^{n}v_i^2},\ \hat{\boldsymbol{\beta}}=\begin{bmatrix}\boldsymbol{v}\\ \beta_0\\ \boldsymbol{\beta}\end{bmatrix},\ \text{and}\ \hat{\boldsymbol{\Lambda}}=\begin{bmatrix}\boldsymbol{I}\odot q & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \psi_1^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}\end{bmatrix}.\tag{D.5}$$

Master's Theses in Mathematical Sciences 2022:E7
ISSN 1404-6342

LUTFMS-3437-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/