SPATIAL STATISTICAL MODELING OF INSURANCE RISK

AN EPIDEMIOLOGIST APPROACH TO IMPROVED CAR INSURANCE PREMIUMS

OSKAR TUFVESSON

Master’s thesis
2017:E02

LUND UNIVERSITY
Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics
Abstract

The aim of this thesis is to provide a statistical method for assessing the relative insurance risk associated with the policyholders geographical location. Number of claims and claims cost are modeled separately, where the Poisson distribution is assumed for the former and the Gamma distribution is assumed for the latter. The models are based on a Bayesian approach, and inference is made using Integrated Nested Laplace Approximation (INLA). It is shown that an ordinary generalized linear model is sufficient for the Gamma distributed claims cost - while the model for number of claims can be improved by combining the ordinary generalized linear model with a spatial component in a conditional auto regressive model. In this study car insurance data from If P&C Insurance was used together with spatial referenced data of high resolution, provided by Insightone.
Acknowledgements

The thesis was conducted at If P&C Insurance, where I also worked during the time of writing. It is fair to say that the thesis would not have been what it is today without the valuable feedback and great support from all colleagues.

Also a special thanks should be directed to Johan Lindström who supervised the work, and modified the original idea into an interesting and fruitful thesis project.

It should be pointed out that after working on a subject intensely for five months, it is doubtlessly an intricate task to compress the ideas into one page that is understandable and interesting at the same time. Therefore I would like to thank Robin Vetter for his efforts in helping me translating the ideas and result of this thesis into a popular science summary.

Lastly, I want to thank friends and family for their support, not only during the last five months, but during my entire period of studies.

Stöten, February 2017
An epidemiologist approach to improved car insurance premiums

When deciding the premium for a specific car insurance, insurers take a plethora of factors, such as age and area of residence, into consideration. But a new approach is bringing added resolution to the picture.

To price their insurance contracts in an optimal way, the insurance company needs to do their utmost in accurately predicting the future cost of claims from the insured policyholder. For this purpose they use different explanatory variables. At If, the geo-location of the policyholder has been one of these variables for a long time. As an example one can imagine that a car gets more frequently damaged in a city, than in a more sparsely populated area, due to the denser traffic. But it is possible to detect differences in risk even on a geographically very detailed level.

As the areas become smaller, the amount of observed claim data in each area gets scarcer. Previously at If, the geographic assessment of risk was based on demographic and socio-economic variables from each location to bypass the issue of data scarcity. Using a new approach to the assumed model, If looks to improve insurance pricing by allowing neighbouring areas to borrow information from each other.

Borrows from epidemiology
The model tested in a case study utilizes a geographical division provided by the analysis company Insightone. The areas are of varying size and densely populated areas can sometimes be as small as only an apartment block. The method relies upon a model that was developed for image restoration, and has been applied in epidemiology for estimating the risk of a disease given the number of occurrences at each location.

By using the variables from each location and at the same time allowing the risk in an area to be dependent on its neighbours, the prediction can be improved if a spatial pattern is recognized.

Tested on real data
In the case study on vehicle hull damage insurance from If, it is shown that the new approach works well for predicting the number of claims, but that information from neighbours do not improve predictions of average claim cost. The result also demonstrates that the insurance company can improve their pricing by accounting for neighbourhood structure, when estimating number of accidents.

This way of assessing risk associated with geographical location could very well be applied to other insurance products, and the method could prove efficient in cases where good demographic or socio-economic variables do not exist. So, if you know what happened around a policy - why don’t you peek around the corner to improve prediction?

Lunds Tekniska Högskola, Master’s Thesis
Title: Spatial Statistical Modeling of Insurance Risk, Author: Oskar Tufvesson
# Contents

1 Introduction 1
  1.1 Insurance 1
  1.2 Risk differentiation 1
  1.3 The tariff 1
  1.4 Modeling claims frequency and severity separately 1
  1.5 Geographical micropricing 3
  1.6 Available data 3
  1.7 Situation analysis 4
  1.8 Other geographical pricing models in insurance 5
  1.9 Purpose of thesis 5

2 Model 7
  2.1 Spatial referenced data 7
  2.2 Poisson distribution for aggregated claim frequency 8
  2.3 Gamma distribution for aggregated claim severity 9
  2.4 Modeling the spatial dependence 10
    2.4.1 Conditional auto regression 10
    2.4.2 Model specification 11

3 Inference 13
  3.1 Bayesian Inference 13
  3.2 Integrated Nested Laplace Approximation 14
    3.2.1 Laplace approximation 14
    3.2.2 Nested laplace approximation 15
    3.2.3 Integrated nested Laplace approximation 15
    3.2.4 The R-INLA package 16
  3.3 Choice of hyper-prior distribution 16
  3.4 Process of choosing co-variates 16
    3.4.1 Penalized maximum likelihood via the elastic net 16
    3.4.2 Correlation between co-variates 17
    3.4.3 Elastic net in the case of Gamma distributed observations 17
  3.5 Model selection and validation 18
    3.5.1 Prediction errors 18
    3.5.2 Deviance information criteria 18
Introduction

Some general background on insurance and insurance pricing is provided in Sections 1.1 - 1.5. In Sections 1.6 - 1.8 the current situation, the problem background and related published articles are discussed. The main purpose of the thesis is given in Section 1.9.

1.1 Insurance

An insurance is in essence a contract which protects against uncertain financial loss. It is formed as a contract between the provider of the insurance and the insured, often called the policyholder. The fundamental principle of an insurance business is to spread the risk among a large enough number of policies, and hence be able to cover financial losses that the individual policyholder cannot cover or are not willing to.

Non-life insurance companies are often involved in many types of different insurance contracts, e.g. the homeowners insurance or the vehicle hull damage insurance which covers damages on your car (but not liability, theft, fire etc.). Clearly different types of insurance contracts involve different risks and consequently differ in price.

1.2 Risk differentiation

The price of an insurance contract does not only vary according to what is covered by the insurance but it also depends on what object is insured and who the policyholder is. The main principle is that if the insurance company accurately can predict the risk associated with each policy, then a higher price can be given to the more risky policies and a lower price to the less risky policies.

1.3 The tariff

The tariff is the model that governs estimation of risk associated with each insurance contract. The aim is to model the so called risk premium. This is defined as the total claims cost divided by the total exposure during a specified period of time. Here,
the exposure during a period of time is measured by the total duration of all active policies during that window in time.

\[
\text{Risk Premium} = \frac{\text{Total Claims Cost}}{\text{Total Duration of policies}}. \tag{1.1}
\]

In the end this amount should cover the insurer’s expected cost for a policy. As mentioned above the expected risk premium is different depending on specific attributes of the object insured or the policyholder. Generally in insurance the tariff is multiplicative and consists of rating factors. An example of such a factor in a hull damage tariff could be car owners age. In Table 1.1 an example of how the rating factor might look is given. The rating factors can be thought of as relative risk in comparison to some base group. For example if a group of policyholders have a rating factor larger than one then this group is expected be riskier than the base group (having rating factor 1).

Table 1.1. An example of a rating factor in the multiplicative tariff.

<table>
<thead>
<tr>
<th>Car owners age</th>
<th>Level (j)</th>
<th>Rating factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-25</td>
<td>1</td>
<td>(\gamma_{i,1})</td>
</tr>
<tr>
<td>26-50</td>
<td>2</td>
<td>(\gamma_{i,2})</td>
</tr>
<tr>
<td>51-</td>
<td>3</td>
<td>(\gamma_{i,3})</td>
</tr>
</tbody>
</table>

Assume that \(K\) such variables are included \(\gamma_{1,j_1}, \ldots, \gamma_{K,j_K}\). The second index, \(j\), specifies which tariff cell the specific contract belongs to. A tariff cell is simply a unique combination of indices \(j_1, \ldots, j_K\). The expected risk premium is modeled as

\[
\mathbb{E}[\text{Risk Premium}_{j_1, \ldots, j_K}] = \gamma_0 \cdot \gamma_{1,j_1} \cdots \gamma_{K,j_K}, \tag{1.2}
\]

where \(\mathbb{E}[]\) denotes the expectation operator. Here \(\gamma_0\) is referred to as the base level. The multiplicative model is the standard model in insurance pricing, as discussed by Ohlsson and Johansson (2010).

### 1.4 Modeling claims frequency and severity separately

In Equation (1.1) it can be seen that the number of claims did not occur anywhere in the definition, even though this is a tractable quantity to model. There is however a possibility to split the model into two parts - modeling claims frequency and severity separately. This is a quite common choice in practice. Introducing some notation:

\[
\begin{align*}
RP &:= \text{Risk Premium} \\
C &:= \text{Total Claims Cost} \\
Y &:= \text{Number of Claims} \\
E &:= \text{Duration (exposure)} \\
F &:= \frac{Y}{E} = \text{Frequency} \\
S &:= \frac{C}{Y} = \text{Severity},
\end{align*}
\tag{1.3}
\]
Equation (1.1) can be written as

\[ RP = \frac{C}{E} = \frac{C}{Y} \times \frac{Y}{E} = S \times F \]  

(1.4)

As seen, the frequency represents the probability of a claim to a given policy and the severity represents the average claims cost. In this setting, different multiplicative models can be assigned to the claims frequency \( F \) and the claim severity \( S \) respectively. As discussed in Ohlsson and Johansson (2010, p.34) the standard tariff analysis is to do separate analyses for claim frequency and claim severity - and then the relativities for risk premium are found by multiplying the results. One main reason for this separation is that claim frequency is usually much more stable than claim severity, whence the relativities for frequency can be estimated with greater accuracy. A more detailed discussion of the model specification will be presented in Sections 2.2 and 2.3.

1.5 Geographical micropricing

Geography has been a part of If’s pricing models for a long time. Already in the late 60’s prices were differentiated according to whether the policyholder lived in a large town or not. However it is not until quite recently that micro-geography, i.e. geographical location on a small scale, was taken into account in pricing.

1.6 Available data

At the moment If has access to a geographical division of Stockholm provided by a company called Insightone, who specializes in data driven customer knowledge. The division is depicted in Figure 1.1. One possible advantage of Insightone’s division is that it does account for ”natural” obstacles, such as water etc. This is incorporated in the neighbourhood specification. Due to confidentiality Insightone has decided not to provide If with exact area borders, but rather with area centroid coordinates and neighbouring structure. It is however enough for the analysis in this thesis.

In the division each of the areas are indexed - and associated to each area there is a generous amount of underlying data such as gender distribution, age distribution and average income. In addition to this external data, If has internal claims and insurance data from the years 2011-2015. For the purpose of this thesis, claims and insurance data on the vehicle hull damage policy has been used. This choice is motivated by the large amount of data and the relatively nice distribution of claims cost arising from this type of policies. The idea is that if it turns out to be too hard to implement a new model using this data, then there is probably no use in continuing trying the model on other types of insurance data. The analysis in this thesis will be limited to a bounded geographical area, called the model window for future reference. The reason for this limitation is mainly computational - but also interpreting the results will be easier. The window can easily be extended, as long as there is computational power available. The model window was chosen as a rectangle covering Stockholm’s inner city and suburbs - in total yielding a division containing 13 831 areas.
Figure 1.1. A zoomed in picture of the division of Stockholm. The circles represents the center of each area, but the exact borders are not disclosed. Clearly Insightone’s division is very granular in central Stockholm, but will get sparser in the outskirts of Stockholm.

1.7 Situation analysis

Currently If uses another (courser) geographical division. Let $s = \{s_1, \ldots, s_N\}$ denote the $N$ areas in the division and let $\zeta_i$ be the relative risk associated with area $i$. Using the notation introduced in Section 1.3, the expected risk premium for a policy in tariff cell $j_1, \ldots, j_K$ and area $i$ is modeled as

$$E[RP_{j_1, \ldots, j_K}(s_i)] = E[RP_{j_1, \ldots, j_K}] \times E[\zeta_i]. \quad (1.5)$$

In line with the ordinary pricing models in insurance, discussed in Section 1.3, the expected relative risk $\zeta_i$ in each area is modeled as multiplicative. As underlying variables to the rating factors, demographic and socio-economic variables of the form discussed in Section 1.6 are used. Letting $\alpha$ denote the rating factors in this model, it will be on the form

$$E[\zeta_i] = \alpha_{1,i} \cdots \alpha_{n,i}. \quad (1.6)$$

This is in principle the same approach as in the ordinary tariff, only aggregated on area-level. In principle this translates to assuming a generalized linear model (GLM) on the relative risk factor $\zeta_i$ - letting the areas act as policyholders. Once $E[\zeta_i]$ has been estimated a spatial smoothing is applied to avoid giving drastically different prices to near neighbouring areas. This is rather a business side decision than any statistical measure. The smoothing in the current setting causes a less efficient risk prediction.
1.8 Other geographical pricing models in insurance

In Boskov and Verall (1994) the authors examines a spatial model for assessing the geographical risk on post code area level. Their approach is based on observed loss ratios, which is calculated as claims cost divided by paid premium. They apply their model directly onto the loss ratios and use Markov Chain Monte Carlo (MCMC) methods for Bayesian inference. The model specification is of the same structure as will be proposed in this thesis, and the authors advocates the potential for these methods in insurance pricing.

Models for claim frequency and severity in non-life insurance that allows for spatial dependence are considered in Gschlößl and Czado (2007). Gschlößl and Czado follows a fully Bayesian approach, where model parameters are estimated using MCMC methods. Moreover, an application is given on a data set from a German car insurance provider, where they show that the inclusion of spatial effect significantly improves the models for both claim frequency and severity. The data set in the application consisted of 440 geographical areas.

To the extent of my knowledge, I have not encountered any other published article discussing the possibility of implementing a spatial model for assessing insurance risk.

1.9 Purpose of thesis

The purpose of this thesis is to implement a spatial statistical model for estimating the spatial relative insurance risk. The model should be a generalization of the geographical pricing model used at If, which did not account for spatially structured dependence. In detail, the aim is to evaluate the model adequacy and performance given the observed car insurance data and also discuss the possibilities to use the generalized model in practice.
Model

The problem is articulated in the spatial setting in Section 2.1. Continuing, in Sections 2.2-2.3 it is justified that number of claims and claims cost on aggregated level should be modeled as Poisson and Gamma distributed respectively. Some further theory on how to incorporate spatial dependence is discussed in Section 2.4 and lastly the full model is specified in Section 2.4.2.

2.1 Spatial referenced data

The insurance claims data available in combination with the division provided by Insightone yields spatial referenced claims data. That is, to each of the $N$ areas in the analysis window the number of claims and the claims cost has been mapped. This spatial data can be viewed as realizations of a (possibly multivariate) stochastic field

$$y := \{y(s), s \in \mathcal{D}\}$$

where $\mathcal{D}$ is a fixed countable subset of $\mathbb{R}^2$. To specify, $s$ is the coordinates giving the location of each area, and $\mathcal{D}$ will be the division provided by Insightone. To ease notation $y(s_i)$ is shortened to $y_i$. When constructing spatial models using areal data the spatial dependency is modeled through the neighbourhood structure of $\mathcal{D}$. The neighbours can be defined through $\mathcal{N}_i$ as the set of all of areas which share borders to area $i$, which is illustrated in Figure 2.1. These are called the first order neighbours. It would also be possible to consider $\mathcal{N}_i$ as the set of second order neighbours, i.e the all the areas that share border with it (first order) plus the areas which share borders with the first order neighbours (second order).

At the moment, the stochastic process $y$ could be either the number of claims or the total claims cost in each of the $N$ areas. If $y$ is a multivariate process, then it could be specified as a two dimensional process where number of claims is the first component and claims cost the second. In any case, a reasonable model assumption is a multiplicative model for $y_i$, with an additional rating factor corresponding to areal affiliation. Denote this factor $\eta_i$, where the index corresponds to the area index. This is so far building on the same model assumptions as the previous geographical pricing model at If - compare with Equation (1.5).
Chapter 2. Model

Second order neighbours
First order neighbours
s(i)

Figure 2.1. Illustration of neighbourhood structure. If $\mathcal{N}_i$ is the set of first order neighbours then it contains all the medium gray areas.

$$E[y_i|\eta_i] = E_i \exp(\eta_i)$$ (2.2)

In Equation (2.2), $E_i$ is the exposure in area $i$. It reflects the effect on the outcome, accounting for the total duration or number of claims in an area, depending on if number of claims or claims cost is modeled. Naturally, when the analysis window is divided the areas will likely not be homogeneous with respect to demography and socioeconomic status. While this is preferable for geographical analysis of the data, a certain level of the variation in risk between areas can be explained by the ordinary tariff alone. This, however, can be taken into account in the measure of exposure. For example when modeling claim counts in area $i$, a weighted exposure $E_i^*$ in area $i$ is used rather than the raw exposure $E_i$. The weighting is computed according to the composition of policyholders in each area. This method of obtaining a weighted exposure $E_i$ for expected counts compares to the weighted population used in models for disease mapping, where the distribution of e.g. age and sex in each area is accounted for in the weighted population, as discussed in Blangiardo and Cameletti (2015, p.179) and Papoila et al. (2014, p.1).

2.2 Poisson distribution for aggregated claim frequency

A standard model for claims frequency is to assume that the number of claims for a single policy follows a Poisson distribution (Ohlsson and Johansson, 2010, p.18). By assuming that policies are independent the number of claims on the aggregate level
(i.e. areas) will also follow a Poisson distribution, which is shown in Isham (2010). From Equation (2.2) it follows that

\[ Y_i \sim \text{Po} (E_i \exp(\eta_i)) , \]  

(2.3)

where \( Y_i \) is the number of claims in area \( i \), and \( \text{Po}(\lambda) \) denotes the Poisson distribution with intensity \( \lambda \). For the rest of this thesis the number of claims will be considered as response variable, but one often seeks to estimate the claims frequency. The expected frequency can however always be found as

\[ E[F_i|\eta_i] = \frac{1}{E_i} E[Y_i|\eta_i] = \frac{1}{E_i} E_i \exp(\eta_i) = \exp(\eta_i), \]  

(2.4)

using the definition from Equation (1.3).

### 2.3 Gamma distribution for aggregated claim severity

For this analysis, the claims cost \( C \) is modeled conditioned on the number of claims. That is, number of claims, \( Y \), acts as the non-random measure of exposure in this model (compare with \( E \) in the frequency model). Again, the model will be specified with the total claims cost in each area, \( C_i \), as response variable rather than the severity. However, for the moment the area index \( i \) will be dropped. For claims cost the Gamma distribution has become more or less the standard option in modeling (Ohlsson and Johansson 2010, p.20). First the cost, \( C \), of one claim is assumed to be Gamma distributed. This is the case \( Y = 1 \). The claims cost \( C \) then has the following probability density function

\[ \pi_{C|Y=1}(c) = \frac{b^a}{\Gamma(a)} c^{a-1} \exp(-bc), \quad a > 0, \quad b > 0, \quad c > 0, \]  

(2.5)

where \( E[C|Y = 1] = a/b = \mu \) and \( V[C|Y = 1] = a/b^2 \). Moreover, \( \Gamma(\cdot) \) denotes the Gamma function and \( V[\cdot] \) the variance of its argument. This parametrization of the Gamma distribution is denoted \( \Gamma(a,b) \). Considering an area with \( Y = y > 0 \) claims, the total claims cost \( C \) can be viewed as the sum of \( y \) independent identically distributed claims. Now, sums of independent Gamma distributions with the same scale parameter, \( b \), are also Gamma distributed. Specifically we have

\[ C|Y = y \sim \Gamma(ya,b), \]  

(2.6)

where \( E[C|Y = y] = ya/b = y\mu \). Hence, the total claims cost \( C \) in each area with a positive number of claims should also be modeled as Gamma distributed. If one is interested in the severity rather than the total claims cost, the scaling properties of the Gamma distribution yields

\[ C \bigg| \frac{Y}{Y} \sim \Gamma(ya,yb) \]  

(2.7)

which implies that \( E[C/Y|Y = y] = (ya)/(yb) = \mu \). For the rest of this thesis the model specification in Equation (2.6) will be used. Letting \( \log(a/b) = \log(\mu) = \eta \)
and re-introducing the area index \(i\), it is clear that this transforms to the model specification in Equation (2.2).

\[
E[C_i|Y_i = y_i, \eta_i] = y_i \exp(\eta_i) \tag{2.8}
\]

### 2.4 Modeling the spatial dependence

Modeling geographically referenced data is a well explored field of science in epidemiological applications, environmental studies and disease mapping [Besag et al., 1991]. When working with spatial data it is important to account for a possible spatial trend in the model to avoid biases in the estimates. In this setting, the Bayesian approach for inference is particularly effective [Blangiardo and Cameletti, 2015 Ch.6], given that we can solve the inference problem in a feasible amount of time. One of the main challenges in Bayesian inference for spatial models is computational given the added complexity due to the spatial structure.

For both the frequency model and the severity model, a latent Gaussian model is assumed. That is, the mean \(\mu_i\) of the observational variable \(y_i\) is linked to the structured additive predictor \(\eta_i\). In this thesis, as seen above, the link function is \(\log \mu_i = \eta_i\). The linear additive predictor will be on the following form

\[
\eta_i = \beta_0 + z_i \beta + u_i + v_i \tag{2.9}
\]

Here, \(z_i\) is a (suitable) collection of underlying co-variates in each area, discussed in Section 1.6. The Gaussian part in a latent Gaussian model stems from that Gaussian priors are assigned to the vector of (unobservable) parameters in the predictor \(\eta_i\). For latent Gaussian models a fast and accurate algorithm for Bayesian inference is provided by [Rue et al. 2009].

#### 2.4.1 Conditional auto regression

To begin, consider first the vector of all linear predictors \(\eta = [\eta_1 \ldots \eta_N]^T\) as a random vector, which indeed is the case in the Bayesian framework. If each component were located at a time point rather than a spatial point - under a Markov assumption - the joint density of \(\eta\) could be decomposed as

\[
\pi(\eta) = \pi(\eta_1) \times \pi(\eta_2 | \eta_1) \times \cdots \times \pi(\eta_N | \eta_{N-1}) \tag{2.10}
\]

where \(\pi(\cdot)\) generically denotes the probability density function of its argument. An intuitive model specification would be to specify the conditional marginal distributions \(\pi(\eta_i | \eta_{i-1})\) for \(i = 1, \ldots, N\). Since the assumption is first order markov, this compares to a AR(1) model in the standard time series analysis [Waller and Carlin 2010]. However, this is not a useful specification in the spatial setting. Nevertheless, using the intuition of specifying the conditional distributions it would be natural to specify the conditional distribution \(\pi(\eta_i | \eta_{-i})\), where \(\eta_{-i}\) denotes all elements of \(\eta\) except \(\eta_i\). The Markov assumption in the spatial setting refers to the property that \(\eta_i\) should only be dependent on a few components of \(\eta_{-i}\), namely the set of
near neighbours $\mathcal{N}_i$. If the model is specified as conditional on only the set of first order neighbours it is referred to as a conditional auto regressive model of order one (CAR(1)). One could also consider a CAR(2) model, which extends the Markovian property to the second order neighbours. These model specifications are thoroughly discussed in Waller and Carlin (2010).

The Markov property together with the Gaussian assumption gives $\eta$ as a Gaussian Markov Random Field (GMRF). An important property of the GMRF is that its precision matrix (inverse covariance matrix) is sparse, which gives great computational benefits in the process of inference. General theory on GMRFs and the associated computational benefits is given in Rue (2001) and Rue and Held (2005). Another assumption required by R-INLA, Section 3.2.4, is that all observations $y_i$ should be independent conditional on the latent field $\eta$.

### 2.4.2 Model specification

As discussed above either a Poisson distribution or Gamma distribution will be used for the observations. At the moment, however, the observation likelihood is generic but belongs to the exponential family. Let $\pi(y_i|\eta_i)$ denote the observation likelihood for the $i$th observation conditioned on the linear predictor $\eta_i$. Recall, $\mathcal{N}_i$ denotes the set of neighbours to area $i$ and let $N(\mu, \sigma^2)$ denote the normal distribution with mean $\mu$ and variance $\sigma^2$. The assumed model will be on the following form.

$$
\begin{align*}
y_i|\eta_i & \sim \pi(y_i|\eta_i) \\
\eta_i|\psi & = \beta_0 + z_i\beta + u_i + v_i \\
u_i|\{u_j : j \in \mathcal{N}_i\}, \psi & \sim N \left( \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} u_j, \frac{1}{|\mathcal{N}_i|\psi_1} \right) \\
v_i|\psi & \sim N(0, \psi^{-1}_2) \\
\psi & \sim \pi(\psi)
\end{align*}
\tag{2.11}
$$

Here, $\mathbf{y} = [y_1 \ldots y_N]^T$ denotes the random vector of observations, either number of claims or claims cost. As previously indicated, the spatial effect (if any) is modeled through $\mathbf{u} = [u_1 \ldots u_N]^T$, and the spatial dependence is constructed as the simple average of all neighbours. Other options for the spatial dependency are not considered in this project. The unstructured effects $\mathbf{v} = [v_1 \ldots v_N]^T$ models additional random variation, unexplained by geographical location. It would be possible to exclude $\mathbf{v}$ from the model - in this setting $\mathbf{v}$ in some sense controls the smoothness of the field. The larger the effect of $\mathbf{v}$ compared to $\mathbf{u}$ is in the model, the less exchange of information between areas is allowed. As seen, both $u_i$ and $v_i$ are specified as Gaussian, with variance determined by the hyperparameters $\psi_1$ and $\psi_2$ respectively. These are distributed according to some prior distribution $\pi(\cdot)$. A common choice of prior is the Gamma distribution (Blangiardo and Cameletti, 2015, ch.6), and this will be the choice in this thesis.

Denote by $\mathbf{x}$ all the (Gaussian) parameters that constructs the latent spatial field $\eta$. That is

$$
\mathbf{x} = [\beta_0 \beta^T \mathbf{u}^T \mathbf{v}^T]^T,
\tag{2.12}
$$
with \( \text{dim}(\mathbf{x}) = n \). The density \( \pi(\mathbf{x}|\psi) \) is multivariate Gaussian with covariance matrix \( Q^{-1}(\psi) \). The mean is assumed to be \( 0 \). The posterior distribution of \( \mathbf{x}, \psi|\mathbf{y} \) can be expressed as

\[
\pi(\mathbf{x}, \psi|\mathbf{y}) \propto \pi(\psi) \times \pi(\mathbf{x}|\psi) \times \pi(\mathbf{y}|\mathbf{x}, \psi)
\]

\[
\propto \pi(\psi) \times \pi(\mathbf{x}|\psi) \times \prod_{i=1}^{N} \pi(y_i|\mathbf{x}, \psi)
\]

\[
\propto \pi(\psi) \times |Q(\psi)|^{1/2} \exp \left( -\frac{1}{2} \mathbf{x}^T Q(\psi) \mathbf{x} \right) \times \prod_{i=1}^{N} \exp(\log(\pi(y_i|\mathbf{x}, \psi)))
\]

\[
\propto \pi(\psi) \times |Q(\psi)|^{1/2} \exp \left( -\frac{1}{2} \mathbf{x}^T Q(\psi) \mathbf{x} + \sum_{i=1}^{N} \log(\pi(y_i|\mathbf{x}, \psi)) \right)
\]

(2.13)

The main aim is now to approximate the posterior marginal distributions

\[
\pi(x_i|\mathbf{y}), \ \pi(\eta_i|\mathbf{y}), \ \text{and} \ \pi(\psi_i|\mathbf{y})
\]

(2.14)

allowing us to estimate parameters and latent fields.
First a very brief introduction to Bayesian inference is given in Section 3.1. Then the method for Bayesian inference using Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) is described in Section 3.2. In Section 3.3 the choice of hyper-prior distribution is addressed, while methods for solving the model selection and validation problem is presented in Sections 3.4-3.5. In this entire chapter the observation process is denoted $y$ and could be either number of claims or claims cost.

### 3.1 Bayesian Inference

Let generically $\pi(\cdot | \cdot)$ denote the conditional density function of its arguments, for example if $y$ is a random variable then $\pi(y)$ denotes the density of $y$, and $\pi(y|x)$ denotes the density of $y$ conditional on $x$. Let $y$ be the observed variable with a distribution governed by a latent (unobservable) random variable $x$. In the spatial case, the latent variable $x$ often is multidimensional, and the terminology used is latent field rather than latent variable. A Bayesian hierarchical model is formulated by giving components on the following form (Blangiardo and Cameletti, 2015, ch.5):

- Likelihood for observations: $\pi(y|x, \psi)$
- Latent field: $\pi(x|\psi)$
- Prior: $\pi(\psi)$
- Posterior: $\pi(x, \psi|y)$.

The specification in Equation (2.11) is exactly on this form. Here, an additional quantity appeared, namely $\psi$. This random variable constitutes the hyperparameter(s) for the latent field $x$ and the observation process $y$. The main aim is to make inference about the unobservable $x$ using the posterior marginal distribution $\pi(x|y)$. By Bayes theorem the data likelihood, the latent field and the posterior marginal distributions $\pi(\psi|y)$ and $\pi(x|y)$ are related through

\[
\pi(\psi|y) = \int \pi(\psi, x|y) \, dx \propto \int \pi(y|x, \psi) \pi(x|\psi) \, dx. \tag{3.1}
\]

\[
\pi(x|y) = \int \pi(x, \psi|y) \, d\psi = \int \pi(x|y, \psi) \pi(\psi|y) \, d\psi \tag{3.2}
\]
A key ingredient in Bayesian inference is that the parameter of interest is itself modeled as a random variable.

### 3.2 Integrated Nested Laplace Approximation

A common approach to approximate the posterior distributions in expression (2.14) is to use Markov chain Monte Carlo (MCMC) methods. However, in the spatial setting (with a supposively large spatial field $u$, and possibly many parameters in $\beta$) the MCMC-methods cannot provide a sufficiently small approximation error within a reasonable computing time (Rue et al., 2009). Rue et al. (2009) suggest a deterministic algorithm that can produce approximations with an acceptable error using a considerably smaller number of computations, yielding a sufficiently fast approximation in the case of GMRFs.

The INLA algorithm is based on nested Gaussian approximations of some cleverly chosen distributions involved in the calculation of the final posteriors. The following relation between the distributions in expression (2.14) is the key to the INLA procedure.

$$
\pi(\psi | y) = \int \pi(\psi | y) \, d\psi \approx \int \tilde{\pi}(\psi | y) \, d\psi \quad (3.3)
$$

$$
\pi(x_i | y) = \int \pi(x_i | \psi, y) \pi(\psi | y) \, d\psi \approx \int \tilde{\pi}(x_i | \psi, y) \tilde{\pi}(\psi | y) \, d\psi \quad (3.4)
$$

The expressions in equations (3.3) and (3.4) suggests that the algorithm needs to proceed in two steps. First a step where $\pi(\psi | y)$ is approximated and then a second where $\pi(x_i | \psi, y)$ is approximated. Then the hyperparameter $\psi$ can be integrated out numerically to find $\pi(x_i | y)$. The integration is feasible if the dimension of $\psi$ is small. However, as will be evident later, it is necessary to find an approximation of the full density of $x | y, \psi$.

#### 3.2.1 Laplace approximation

In this section an approximate expression for the density $\pi(x | y, \psi)$ is sought. Since

$$
\pi(x | \psi, y) \propto \pi(x, \psi, y) = \pi(y | x, \psi) \pi(x | \psi) \pi(\psi) \quad (3.5)
$$

it follows that

$$
\log \pi(x | \psi, y) = \log \pi(y | x, \psi) + \log \pi(x | \psi) + \text{constant}. \quad (3.6)
$$

Now, since $x | \psi$ is already Gaussian by definition, taking the logarithm yields a second order polynomial in $x$. Furthermore, using a second order Taylor expansion (which is also a polynomial of degree two) of $\log \pi(y | x, \psi)$, it appears that $\log \pi(x | \psi, y)$ can be approximated by a second order polynomial. If $\log \pi(x | y, \psi)$ is a second order polynomial, then $x | y, \psi$ is Gaussian. This is referred to as the Laplace approximation. Let $N(x; \mu, \sigma^2)$ denote the Gaussian density with mean $\mu$ and variance
σ at configuration x. Also, let \( \pi_{LA}(x|y, \psi) \) denote the Laplace approximation of \( \pi(x|y, \psi) \). Then specifically,

\[
\pi(x|y, \psi) \approx \pi_{LA}(x|y, \psi) = \mathcal{N}(x; x^*, -\left[ \frac{\partial^2 \log \pi(x^*|y)}{\partial x^2} \right]^{-1}), \tag{3.7}
\]
due to the uniqueness of the Taylor expansion. Here, \( x^* \) denotes the mode of \( \log \pi(x|\psi, y) \) or equivalently \( x^* = \arg \max_x \pi(x|\psi, y) \). As usual the approximation is most accurate at the mode.

### 3.2.2 Nested laplace approximation

Recall, to accomplish the main target (namely approximating the final posteriors in Equation (2.14)), the approximation of the full marginal posterior \( \pi(\psi|y) \) was needed. It will be obtained as follows. First, it is possible to write

\[
\pi(\psi|y) \propto \pi(x, \psi, y) = \frac{\pi(y|x, \psi)\pi(x|\psi)\pi(\psi)}{\pi(x|\psi, y)} \tag{3.8}
\]

This expression is tractable in terms of known parts, recalling Equation (2.13) and Equation (3.7). Looking at the left hand side of Equation (3.8) which contains no \( x \), it is clear that also the right hand side must be constant with respect to \( x \). Thus the choice of \( x \) for evaluation is arbitrary. Hence, for a given \( \psi \), it is possible to obtain an approximation by replacing the denominator with its Laplace approximation from Equation (3.7), and evaluate it at its most accurate point, namely the mode \( x^*(\psi) = \arg \max_x \pi(x|\psi, y) \). This gives the approximation

\[
\pi(\psi|y) \approx \tilde{\pi}(\psi|y) = \frac{\pi(y|x, \psi)\pi(x|\psi)\pi(\psi)}{\pi_{LA}(x|\psi, y)} \bigg|_{x=x^*(\psi)}. \tag{3.9}
\]

For the marginal posteriors of the latent field, \( \pi(x_i|y, \psi) \), it is possible to directly derive an approximation using numerical integration of the full posterior normal approximation \( \pi_{LA}(x|y, \psi) \), where the Cholesky decomposition is used for calculations involving the precision matrix. This alternative is very fast and in some cases sufficiently accurate (Blangiardo and Cameletti, 2015, p.113). However, it is not generally accurate enough. This issue was resolved by Rue et al. (2009), where the authors suggest the simplified Laplace approximation. In principle, this strategy utilizes a Taylor expansion of \( \pi(x|\psi, y) \), of higher degree than suggested in Section 3.2.1. Given that the accuracy of the simplified Laplace approximation has been proved to be sufficient, and that the computational time is short this is the standard option in practice.

### 3.2.3 Integrated nested Laplace approximation

Once \( \tilde{\pi}(x_i|y, \psi) \) and \( \tilde{\pi}(\psi|y) \) are obtained the approximations of the relevant marginal posteriors \( \pi(\psi_i|y) \) and \( \pi(x_i|y) \) follows directly from equations (3.3) and (3.4) respectively. The integrals can be solved numerically through a finite weighted sum, for example

\[
\tilde{\pi}(x_i|y) \approx \sum_j \tilde{\pi}(x_i|y, \psi^{(j)})\tilde{\pi}(\psi^{(j)}|y)\Delta_j \tag{3.10}
\]
using some relevant integration points $\psi^{(j)}$ with corresponding weights $\Delta_j$. Good integration points $\psi^{(j)}$ are decided by exploring the joint posterior $\tilde{\pi}(\psi | y)$ in a non-parametric way; where the details are thoroughly discussed in Rue et al. (2009).

### 3.2.4 The R-INLA package

R-INLA is the R package that implements approximate Bayesian inference using integrated nested Laplace approximation (Lindgren and Rue, 2015). It is a rich package, which enables a lot of models to be specified. It is available at http://www.r-inla.org.

### 3.3 Choice of hyper-prior distribution

The choice of hyper-prior distribution is an important choice as it represents the assumptions and information available for the parameters in the model, before any data is seen. Since there is no such prior information, the natural choice is to choose a quite flat prior - enabling the data to speak for themselves. This also happens to be the default choice in the R-INLA package. Clearly, before settling on any conclusion from Bayesian inference, one should be aware of what prior or hyper-prior distribution one has chosen and what effect it has on the results. However, this shall not be in focus for this project but is rather left for later investigation.

### 3.4 Process of choosing co-variates

As seen in the model specification in expression (2.11), there is a possibility to include a set of co-variates, $z_i$, in the model. From the available data described in Section 1.6 there was roughly 140 such variables to choose among, suggesting that some algorithmic or at least structured approach needs to be implemented in variable choice process. There are two main reasons for not wanting to include all possible co-variates in the model. First, this will most likely over-fit the model and thus lead to inferior predictive performance. Second, this makes model interpretation harder. Additionally, introducing too many parameters in the model can run into computational issues later when the model has a spatial component. To get a first set of co-variates to include in the model the idea is to first only consider frequency and severity as pure GLMs, i.e. excluding all spatial effect and unstructured effect in the model specification (2.11), and choosing as few parameters as possible while retaining as much predictive quality as possible. Completing this in an algorithmic fashion is desirable.

### 3.4.1 Penalized maximum likelihood via the elastic net

An effective algorithm for this purpose is the elastic net (Zou and Hastie, 2005). A ready-to-use implementation of the algorithm is available in the R-package glmnet (Hastie and Qian, 2014). Consider a GLM, and let $l(Y_i, \beta_0 + \beta z_i)$ denote the negative log-likelihood for observation $i$. The elastic net then solves the following problem.
Chapter 3. Inference

\[
\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} l(Y_i, \beta_0 + z_i \beta) + \lambda \left[ (1 - \alpha) \| \beta \|_2^2 + \alpha \| \beta \|_1 \right],
\]

(3.11)

over a grid of values of \( \lambda \), see Hastie and Qian (2014), where \( \| \cdot \|_p \) denotes the \( L_p \)-norm. Adding the penalty term involving the magnitude of the regression coefficients forces some of the coefficients to become small in the minimization. The intuition is that a variable really has to be more predictive than others to belong in the final model. This is often referred to as shrinkage. In the version of the package \texttt{glmnet} used in this thesis, all \( z \)-variables are standardized before coefficients are estimated. The tuning parameter \( \lambda \) controls the strength of the penalty. To obtain an optimal choice of \( \lambda \) one commonly adopts \( K \)-fold cross-validation (CV), where one successively holds out one group of the data when estimating the parameters and then calculates a validation error for the held out group. The \( K \) validation errors are weighted to a final validation error. The value of \( \lambda \) that gives minimal such validation error is the optimal one.

3.4.2 Correlation between co-variates

In the case of 140 demographic and socioeconomic variables it is likely that some or many of them will be correlated in groups. If it is not feasible to ignore these dependencies in the variable selection, this can be handled by choosing \( \alpha \) in Equation (3.11) in a proper way. The parameter \( 0 \leq \alpha \leq 1 \) controls the proportion of L1- and L2-norm to use in the penalty. It is known that \( \alpha = 0 \) shrinks the coefficients of correlated coefficients towards each other while \( \alpha = 1 \) tends to pick one of them and discard the others (Hastie and Qian, 2014). The idea of the elastic net is to mix these properties, i.e. choosing \( 0 < \alpha < 1 \). If predictors are correlated in groups, an \( \alpha \) around 0.5 tends to select the groups in or out together.

3.4.3 Elastic net in the case of Gamma distributed observations

Unfortunately the \texttt{glmnet}-package does not support Gamma distributed observations, obstructing the direct implementation in the case of the claims cost model. An alternative approach is to instead model claims cost as log-normal, since \texttt{glmnet} supports normal observations. The idea is that the algorithm should still (roughly) give the same suggestions of best parameters to include, even if another observational distribution is assumed. Using the log-normal distribution as a substitute for the Gamma distribution can be justified using two arguments. First, it shares similar qualitative properties, e.g. the skewness and positiveness. Second, indeed in practice they are often used to model the same phenomena. However, the two models differ in tail-behaviour where the log-normal distribution has a heavier tail. Thus the models will not be equally sensitive for outliers.
Chapter 3. Inference

3.5 Model selection and validation

3.5.1 Prediction errors

A common way of evaluating a models performance is to consider the prediction performance. Before any estimation is made, a certain portion of the data is set aside as a prediction set. Then the remaining data is used for estimation. Let $Y^p_i$ denote an observation in the prediction set, and define the prediction error as $\epsilon_i = Y^p_i - E[Y^p_i]$. It might also be useful to inspect the normalized prediction errors, defined as $\epsilon^\text{norm}_i = \epsilon_i / E_i$ where again $E_i$ denotes the generic exposure (could be duration or number of claims depending on the model in hand). The reason for introducing this quantity is that the variance of the data increases with increased exposure. Nevertheless, also the absolute prediction error might be useful to inspect. The idea is that one can compare models in terms of prediction quality, for example using the MSE, defined as

$$\text{MSE} = \frac{1}{N_p} \sum_i \epsilon_i^2, \quad (3.12)$$

and the normalized MSE is analogously defined. Moreover, the prediction errors should exhibit no spatial structure - meaning that all spatial structure was successfully modeled.

3.5.2 Deviance information criteria

If it is interesting to compare different models in terms of performance the deviance can be used (Blangiardo and Cameletti, 2015, p.171). The deviance is defined as

$$D(x) = -2 \log \pi(y|x), \quad (3.13)$$

where as usual $x$ identifies the parameters of the likelihood. In the Bayesian framework $x$ is a random variable, and hence also $D(x)$. Typically the posterior mean deviance,

$$\bar{D} = E[D(x)] \quad (3.14)$$

is used to quantify the deviance as a measure of fit. However, with increasing model size (i.e. number of parameters) the fit will be better and hence $\bar{D}$ smaller. Therefore it is necessary to introduce a trade-off against $\bar{D}$ which reflects the model complexity. Such a measure was proposed by Spiegelhalter et al. (2002) as the Deviance information criteria (DIC). It is a generalization of the well known Akaike information criteria. The DIC is a sum of two components, first the deviance $\bar{D}$ which measures the model fit, and second the effective number of parameters. This quantity is calculated as

$$p_D = E[D(x)] - D(E[x]) = \bar{D} - D(E[x]). \quad (3.15)$$

Finally the DIC is defined as

$$\text{DIC} = \bar{D} + p_D = 2\bar{D} - D(E[x]). \quad (3.16)$$
Here $\hat{D}$ decreases with better model fit, and $p_D$ increases with added complexity. The idea is simply that a lower value of the DIC indicates that the model is better supported by the data in hand.
Chapter 3. Inference
Results

In Section \([4.1]\) the different choices of linear predictors are specified. Then the result from the process of choosing co-variates is presented in Section \([4.2]\). In Section \([4.3]\) and Section \([4.4]\) the results from fitting claims frequency and severity are presented respectively, and the chapter ends with Section \([4.5]\) where the best models are chosen and validated in an alternative way.

### 4.1 The linear predictor

There are five models fitted for claims frequency and claims severity respectively, where all models share the form specified in Equation \((2.11)\). The difference in the models lies in whether the co-variate effect, \(z_i \beta\), the spatial effect \(u_i\) and the unstructured effect \(v_i\) is included or not. The prior distribution of the precision parameters is specified as the default from \texttt{R-INLA}. In Table \([4.1]\) the models are presented, and named for future reference.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Linear predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT</td>
<td>(\eta_i = \beta_0)</td>
</tr>
<tr>
<td>GLM</td>
<td>(\eta_i = \beta_0 + z_i \beta)</td>
</tr>
<tr>
<td>IID</td>
<td>(\eta_i = \beta_0 + z_i \beta + v_i)</td>
</tr>
<tr>
<td>BESAG</td>
<td>(\eta_i = \beta_0 + z_i \beta + u_i)</td>
</tr>
<tr>
<td>BYM</td>
<td>(\eta_i = \beta_0 + z_i \beta + u_i + v_i)</td>
</tr>
</tbody>
</table>

The data from the Insightone division consists of 13,831 areas. A prediction set of 10\% is randomly selected, so the models are fitted on the remaining 0.90 \cdot 13,831 observations. In the following modeling stages, this operation is accomplished by assigning those values \(y_i = NA\) in \texttt{R-INLA} and thus not letting them contribute to the observation likelihood. Still, \texttt{R-INLA} will predict at those points which provides an easy way of doing prediction.
4.2 Choice of co-variates

To choose which co-variates that are suitable to include in the model, the GLM specification from Table 4.1 is chosen as linear predictor - including all possible co-variates in the vector $z_i$.

4.2.1 Elastic net

The result from applying the elastic net algorithm with $\alpha = 0.5$ on all data and all available covariates is displayed in Figure 4.1. It becomes clear that the procedure of determining an optimal value of the penalty strength $\lambda$ should be individual depending on whether it is frequency or severity that is considered. In the case of frequency, the confidence intervals of the estimated validation error are reasonably small, as seen in Figure 4.1a. Hence one can argue that models with validation error within one standard deviation of the optimal model cannot be statistically distinguished. Therefore the model size can be reduced further - with the threshold indicated by the right dashed line in Figure 4.1a. This procedure would not work well in the case of severity, suggested by Figure 4.1b. The issue is the width of the confidence intervals for the validation error. Instead, the reasonable choice here is to simply accept the $\lambda$ that minimizes the validation error, and consequently accept that model size.

In Table 4.2 the number of chosen co-variates are reported. The idea is that these numbers can be reduced further when fitting the model using no penalization term, since it is possible that not all co-variates will turn out to be statistically significant.

![Figure 4.1](image)

(a) Frequency  (b) Severity

**Figure 4.1.** Validation error with 95% confidence bounds from a 10-fold CV using the elastic net algorithm in (a) and (b).

**Table 4.2.** Number of covariates suggested by the elastic net algorithm using $\alpha = 0.5$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of co-variates to include</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>16</td>
</tr>
<tr>
<td>Severity</td>
<td>31</td>
</tr>
</tbody>
</table>
4.2.2 Parameter significance

To assess the parameter significance, the model from Equation (2.11) is fitted using the R-INLA package, with the GLM specification from Table 4.1 on the linear predictor $\eta_i$. Using the estimated posterior distributions of each $\beta$-coefficient it is possible to derive (approximate) credibility intervals and $p$-values, and hence decide whether all co-variates really are significant on the 95% level. If a $\beta$-coefficient turns out to be insignificant, then it is excluded and the model is re-fitted without the co-variate. This iterative process is repeated until all co-variates show a significant effect. The order in which co-variates were removed was decided by the ratio of the mean and standard deviation. Alternatively, the estimated $p$-value could have been used. The intuition remains the same - if an estimate shows a small effect and a high degree of uncertainty it can be removed from the model. If one wants to apply a more rigorous approach, there is a rich literature on variable selection in linear regression, for example Gelman and Hill (2007). The same procedure is applied to both the frequency model and the severity separately. The final number of co-variates to include in the model are shown in Table 4.3. Due to confidentiality it is not possible to fully disclose which variables are considered.

Table 4.3. Number of significant covariates after checking the significance on the 95% level.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of significant co-variates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>9</td>
</tr>
<tr>
<td>Severity</td>
<td>8</td>
</tr>
</tbody>
</table>

4.3 Frequency models

Recapitulating, the number of claims $Y_i$ in each area $i$ conditioned on $\eta_i$ is modeled as Poisson-distributed with mean

$$E[Y_i|\eta_i] = E_i \exp(\eta_i),$$

where the linear predictor $\eta_i$ is specified according to Table 4.1. In Table 4.4 the DIC and the MSE are reported. Again, whenever adding an effect ($u_i$ or $v_i$ or both) it is a good idea to check whether all $\beta$-coefficients still are significant. If they are not, then they are removed in the order outlined in Section 4.2.2 and the model is re-fitted. The result from this procedure is indicated in the column $\dim(\beta)$ in Table 4.4.

It becomes clear that the models including the unstructured effect (IID and BYM) is favoured in terms of the DIC. However, it seems that the BESAG model performs better than the GLM model, and the BYM model performs better than the IID model - supporting the existence of a spatial effect $u_i$. Thus the best specification for the linear predictor is the BYM specification, namely $\eta_i = \beta_0 + z_i \beta + u_i + v_i$. From Table 4.4 one can also see that it is possible to reduce model complexity by removing some of the covariates. This leads to a slightly worse absolute MSE, but a marginally
### Table 4.4. Summary statistics from fitting the frequency model.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>MSE (Normalized)</th>
<th>dim(β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT</td>
<td>37 677</td>
<td>2.754 (0.03568)</td>
<td>-</td>
</tr>
<tr>
<td>GLM</td>
<td>37 404</td>
<td>2.692 (0.03545)</td>
<td>9</td>
</tr>
<tr>
<td>IID</td>
<td>36 551</td>
<td>2.699 (0.03546)</td>
<td>9</td>
</tr>
<tr>
<td>BESAG</td>
<td>36 810</td>
<td>2.646 (0.03543)</td>
<td>9</td>
</tr>
<tr>
<td>BESAG</td>
<td>36 807</td>
<td>2.648 (0.03543)</td>
<td>7</td>
</tr>
<tr>
<td>BYM</td>
<td>36 496</td>
<td>2.637 (0.03542)</td>
<td>6</td>
</tr>
<tr>
<td>BYM</td>
<td>36 493</td>
<td>2.642 (0.03542)</td>
<td>6</td>
</tr>
</tbody>
</table>

better DIC. Since the difference between the absolute MSE in the BYM models is relatively small (compared with the absolute MSE from the FLAT model), the BYM model with dim(β) = 6 will be considered as the best fit for the data in hand.

#### 4.3.1 Residual analysis for the BYM-model for frequency

Having the posterior marginals of the linear predictor $\eta_i$ in each area $i$ (including the prediction set) it is possible to obtain estimates of the expected value of $e^{\eta_i}$. In Figure 4.2 the prediction errors using the BYM model are displayed from three perspectives. Figure 4.2a shows the prediction errors plotted according to their geographical location, revealing that no spatial trend seems to remain in the data. To check that the prediction errors are independent of the elements of the linear predictor $\eta_i$, the prediction errors were plotted against $E[\exp(u_i)]$ and $E[\exp(z_i\beta)]$ in Figure 4.2b and Figure 4.2c respectively. There seems to be no structure remaining in the prediction errors, a conclusion that is further supported by the trend lines in Figure 4.2b and 4.2c. For example the zero line are within the 95% confidence bounds in both figures.

#### 4.3.2 Fitted values for the BYM-model for frequency

If the estimates are plotted according to their geographical position the spatial trend is clearly revealed. This is displayed in Figure 4.3a and Figure 4.3b respectively, where the dots are centroids of the areas and the colour indicates the estimated expected value. The spatial field $\exp(u)$ in Figure 4.3b is of course smooth in general, but there are some seemingly sharp edges. These are explained by “natural” obstacles such as water etc. Areas separated by such an obstacle are not considered to be neighbours in the graph specification as discussed earlier.
(a) Prediction errors plotted according to their coordinates.

(b) Prediction errors plotted against $E[\exp(z_i\beta)]$, with a trend line in red. 

(c) Prediction errors plotted against $E[\exp(u_i)]$, with a trend line in red.

Figure 4.2. Prediction analysis on the BYM-model for frequency. The prediction is made on the 10% prediction set.
Chapter 4. Results

(a) Estimates of the GLM-part, $e^{z_i \beta}$, in the BYM model for frequency.

(b) Estimates of the spatial part, $e^{u_i}$, in the BYM model for frequency.

Figure 4.3. Map of the model window with estimated expected values of $\exp(z_i \beta)$ and $\exp(u_i)$ respectively.
4.4 Severity models

The claims cost in each area \( i \), conditioned on the number of claims is modeled as Gamma distributed with mean

\[
E[C_i|Y_i = y_i] = y_i \exp(\eta_i), \quad y_i > 0,
\]

and again the linear predictor is specified according to Table 4.1. Repeating the same procedure as in the frequency case, summary statistics from the fitted severity models are presented in Table 4.5.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>( \sqrt{\text{MSE}} ) (Normalized)</th>
<th>( \text{dim}(\beta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT</td>
<td>183235</td>
<td>42278 (10845)</td>
<td>-</td>
</tr>
<tr>
<td>GLM</td>
<td>183055</td>
<td>39920 (10730)</td>
<td>8</td>
</tr>
<tr>
<td>IID</td>
<td>182913</td>
<td>40004 (10728)</td>
<td>8</td>
</tr>
<tr>
<td>BESAG</td>
<td>183021</td>
<td>39988 (10733)</td>
<td>8</td>
</tr>
<tr>
<td>BYM</td>
<td>182896</td>
<td>40000 (10728)</td>
<td>8</td>
</tr>
</tbody>
</table>

Comparing the performance of the GLM model with the FLAT model in Table 4.5, it becomes clear that the added co-variates enhances the model performance and fit, since both the DIC is substantially smaller and the MSE is smaller. However, comparing the IID model against the GLM there is only a slight improvement. The DIC is smaller, but in terms of prediction error they are similar. Here, the choice is not obvious. But since the GLM model in a sense is the benchmark model and a simpler structure is preferred, the GLM model will be preferred in this case. It is however hinted that the added complexity through the spatial effect, \( u_i \), does not provide any additional performance. This is because the GLM model seems to perform similar to the BESAG model, and the IID model similar to the BYM. Confirming this suspicion, the estimated expected values of \( \exp(u_i) \) in the BYM model were all in the interval [0.996, 1.002]. Hence, the spatial effect can be rejected. In total, the best choice is the ordinary GLM specification for the linear predictor, namely \( \eta_i = \beta_0 + z_i \beta \).

4.4.1 Residual analysis for the GLM-model for severity

Proceeding in a similar fashion as in the frequency case, it is possible to obtain estimates of the expected value of \( e^\eta \) and calculate prediction errors. Figure 4.4a shows the prediction errors plotted according to their geographical location, revealing that all spatial trend in the data seems to have vanished. Again, to check that the prediction errors are independent of the elements of the linear predictor \( \eta \), the prediction errors were plotted against \( E[\exp(z_i \beta)] \) in Figure 4.4b. No structure seems to be remaining, and the zero line are inside the 95% confidence bounds which further supports this conclusion.
(a) Normalized prediction errors in the severity model plotted according to their coordinates.

(b) Prediction errors plotted against $E[\exp(z_i\beta)]$, with a trend line in red.

**Figure 4.4.** Normalized prediction analysis on the GLM-model for severity. The prediction is made on the 10% prediction set, excluding all areas where there were no claim.
4.4.2 Fitted values for the severity model

Plotting the estimates of $E[\exp(z_i\beta)]$ from the GLM model for severity according to their geographical coordinates in Figure 4.5 reveals a geographical pattern. Actually, this pattern is similar to the co-variate effect for claims frequency depicted in Figure 4.3a. This indicates that the two models pick up the same underlying phenomena.

![Figure 4.5. Estimates of $E[\exp(z_i\beta)]$ in the GLM model for severity.](image)

4.5 Performance of the geographical pricing model

Another type of model validation and performance testing is to fit the model on data from all areas, but leaving out a validation set of policies already before aggregating the data on area level. A validation set consisting of 15% of the available data is set aside, and the BYM model for frequency and the GLM model for severity is fitted on the remaining data respectively. In Table 4.6, the estimated total frequency and severity are displayed, together with the observed values. The total duration of all the validation-policies inside the model window (Section 1.6) is 40 430 years, which makes for reasonably reliable estimates.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Estimate</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>9.03%</td>
<td>8.99%</td>
</tr>
<tr>
<td>Severity</td>
<td>12 430</td>
<td>11 930</td>
</tr>
</tbody>
</table>

Table 4.6. Estimated frequency and severity on the validation set of policies. The total duration of the policies inside the model window in the validation set is 40 430 years. The frequency estimate is very accurate, while the severity estimate is high.
To furthermore analyze the predictive quality of the models one could consider checking the estimates versus the observations grouped in some informative way. A good benchmark for the geographically associated risk is the old geographical pricing factor, described in Section 1.7. To get some significance the old geographical factors can be grouped into bins. The idea is to group them in a way such that the middle groups are largest and the lower and upper extreme groups are smaller. In this way the model quality in benchmark normal, very low risk and very high risk areas can be evaluated respectively. The results from this procedure are displayed in Figure 4.6. These results indicate that the frequency model is accurate in general, and captures the trend suggested by the benchmark risk. Now, the accuracy of the severity model is questionable but it at least tends to follow the outlined benchmark risk proposed by the old pricing factors.

Remembering the definitions from Section 1.4 it is possible to estimate the final risk premium having estimates of both frequency and severity. In the end, the estimated risk premium is the key ratio of primary interest seen from the insurers point of view. How does the risk premium estimates compare with the estimates of risk premium obtained by the old geographical pricing model? Or in other words, how does the new model stand against the old model? In Figure 4.7 the new and old estimates of risk premium are compared from two perspectives. In Figure 4.7a they are again grouped according to the old pricing factors. This way of grouping is advantageous for the old model, since the comparison is made on its "home turf". From the figure it seems that the old model performs better in this grouping, albeit the difference is not huge in general. In Figure 4.7b the estimates are instead grouped using the main new component in the new geographical risk model - namely the spatial field, $u$, from the BYM model for frequency. Again, the values are grouped in a way such that the duration is smaller (but not too small) in the extreme groups. In Figure 4.7b a trend is revealed. The estimated increased risk from the spatial field $u$ is indeed observed, and it seems like the old model cannot fully capture this phenomena. This validates that there are actually something new, possibly useful, picked up by the added model complexity.

Finally, one could address the question about model performance in a useful way by analyzing the observed loss ratios in the validation set. The loss ratio is defined as the sum of all claims cost divided by the total income of insurance premium. It should be noted that since the old geographical risk model has been part of Ifs pricing procedure for a long time now, the loss ratios already accounts for the old estimate of relative geographical risk. From the new model, a measure of total estimated relative risk can be obtained from

$$E[\exp(z_i^{\text{Severity}} \beta^{\text{Severity}})] \times E[\exp(z_i^{\text{Frequency}} \beta^{\text{Frequency}} + u_i)].$$

(4.1)

To analyze the actual loss ratio the change in estimated geographical risk gives a holistic perspective on the difference between the new and the old risk model. In Table 4.7 it is indeed the case that in the areas where the estimated risk was lower also the observed loss ratio was lower. Some of this difference in loss ratios is captured by the spatial effect, $u$. 


Chapter 4. Results

(a) Estimated frequency (light green) and observed frequency (dark green).

(b) Estimated severity (light blue) and observed severity (dark blue).

Figure 4.6. Estimates and observations plotted versus the old geographical pricing factors (grouped). The yellow bars is the total duration in each group.
Figure 4.7. The observed risk premium (dark blue) is shown together with estimates from the new (light blue) and the old (gray) geographical pricing model. The yellow bars show the total duration in each group.

(a) Estimates grouped by the old geographical pricing factors.

(b) Estimates grouped by the estimated spatial effect, $E[\exp(u_i)]$. 
Table 4.7. Actual observed loss ratios, grouped by the change in estimated geographical risk. All loss ratios are relative to the grand total loss ratio of the validation set.

<table>
<thead>
<tr>
<th>Decreased or increased</th>
<th>Duration</th>
<th>Old loss ratio</th>
<th>Loss ratio adjusted for $\exp(u_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>21 176</td>
<td>106%</td>
<td>102%</td>
</tr>
<tr>
<td>Decreased</td>
<td>19 283</td>
<td>94%</td>
<td>97%</td>
</tr>
</tbody>
</table>
Discussion and conclusions

5.1 Review

Statistical models for number of claims, \( Y \), and claims cost, \( C \), that allows for spatial dependence through spatial auto regression was given in Chapter 2. The models were presented in a fully Bayesian framework, and the method for inference using INLA were detailed in Chapter 3. Also the issue of choosing co-variates and general model selection were discussed in Chapter 3, where the elastic net algorithm was presented as a possible solution to the former issue and the MSE and DIC were used to resolve the latter issue.

In Chapter 4 the models were compared, where it was found that the frequency model can be improved by adding both the unstructured effect (\( v \)) and the spatial effect (\( u \)) to the ordinary GLM model, yielding the BYM model. However for the severity model no significant improvement was identified by the added complexity in model structure. Thus the already popular ordinary GLM model was suggested for severity.

In Section 4.5, it was shown that the models for both number of claims and claims cost yielded results in compliance with the benchmark measure of geographical risk. The predictive performance were encouraging for the frequency model, but unfortunately not so accurate for the severity model. As an ultimate test of the risk model’s value the predicted risk premium was compared with predictions from the old geographical risk model. Figure 4.7 showed that the estimated spatial effect, \( u \), from the frequency model really was present in the validation set in a way that the old model could not fully account for. In Table 4.7 it was shown that the spatial component could also provide value in practice.

5.2 Model structure

Specifying the model as done in Equation (2.11) leads to a simultaneous estimation of the co-variate effect (\( z_i \beta \)), the spatial effect (\( u_i \)) and the random effect (\( v_i \)). The allocation of strength in each component will then be data driven. Another approach is to first conduct a pure GLM analysis, and then use this estimate (possibly smoothed, see Section 5.5) as an offset when estimating \( u_i \) and \( v_i \). Then one
allocates as much explanatory degree as possible into the co-variate component, and only adds the spatial effect as a "layer on top". This could be preferable from the insurers perspective, since it is a more conservative approach of implementing the new estimation approach. However, it is then statistically more questionable.

5.3 Model selection

For the process of choosing co-variates to use in the model, the elastic net algorithm proved to be a very convenient tool. However, as with all plug-and-play algorithms, the output needs to be interpreted carefully. Since this problem was not the main focus in this thesis, the results from the algorithm were basically accepted as came with only some minor sanity checks. However, some of the variables will likely not be best modeled as a linear effect but rather as a spline effect. Alternatively, some variables possibly should to be interpreted as categorical rather than numerical, and there might also be interactions between variables. Considering this could improve model accuracy. The fact that the regression part of the modeling is likely not optimally conducted, is a possible explanation for the poorer prediction seen in Figure 4.7a. Improving the regression part should, however, not make any difference for the main conclusion of this thesis. It is unlikely that improving the co-variate analysis would eliminate the spatial effect entirely. With this said, in this stage of modeling improvements could have been made.

Another aspect of model selection is how to choose the hyperpriors. This issue was not addressed in this thesis, and thus the effect of a different choice of hyperpriors is not known. In practice one would want to at least roughly address this question. In any case, for the vehicle hull damage data considered in this thesis, there are no specific prior beliefs on the hyperparameters in Equation (2.11). Hence, one can argue that the chosen flat prior was at least fair.

5.4 Model extensions

5.4.1 Extending the model to a larger area

Clearly, the division used in this thesis are on a very fine scale. In the Stockholm area, there were 13 831 areas, which compares to the approximately 800 areas in the same region, using the previous division. This is, to the extent of my knowledge, a spatial model for insurance risk on a much more detailed level than in related published papers. This is what makes the inference problem such a computational intensive one, and forces the use of deterministic inference methods such as INLA. The MCMC approach proposed in Glössl and Czado (2007) would be very costly for estimation and probably not feasible considering the large data set here. As long as the number of areas remains within the magnitude of $10^5$ (Rue et al., 2009), the problem is possible to solve in feasible time using INLA. Since the granularity of the division used in this thesis decreases outside of Stockholm, the number of areas should be in the order of $10^5$. Thus a geographical extension could easily be obtained. The burden of computational complexity is somewhat relieved if one considers the procedure outlined in Section 5.2 i.e. first estimating the co-variate effect and only
afterwards estimating the spatial effect. Another way of reducing complexity is to fit
the model on each metropolitan region separately.

5.4.2 Other types of insurance

As the case study, car insurance data was considered in this thesis. However, the
outlined approach is applicable to other products in non-life insurance as well. The
introduction of a pure spatial dependence might be an alternative solution in situa-
tions where it is hard to find suitable explanatory variables.

5.5 Conclusion

It has been seen that the spatial effect, $u$, should only be considered in the frequency
model and not the severity model. This was actually the expected result beforehand.
Inspecting Figure 4.3b it seems like the spatially associated risk is higher in densely
populated areas. This is an intuitive result since dense areas very well could exhibit
higher risk for vehicle hull damage. The strong predictive performance of the BYM
model for frequency, together with the added identification of risk in Figure 4.7b,
suggest that the spatial dependence should be considered in future modeling of claims
frequency vehicle hull damage.

That there were no spatial trend to claims cost is also in line with beforehand intu-
tion. There is no specific reason to believe that the cost for a vehicle hull damage
claim should be spatially dependent. Moreover, strong predictors, such as car brand
and area specific average income, is already accounted for. However, a contingent spa-
tial effect for severity should not be ruled out entirely since it might be distributed
differently outside the metropolitan regions. It might also be the case that the spatial
model for severity is more suitable for other products mirroring other risks, such as
theft or third part liability.

As mentioned in Section 1.7, when having a geographical division of such granularity,
it is necessary from a business side perspective that the final pricing factors are
spatially smoothed. In the old geographical pricing model one simply smoothed
the factors after estimation. This approach could be applied to the estimates of
$E[\exp(z_i\beta)]$ again. By definition, the estimates of $E[\exp(u_i)]$ are already smoothed,
which is positive since no accuracy needs to be lost due to extra smoothing.

To wrap things up, accounting for spatial dependence when modeling insurance risk
yields better model fit and the results are positive from a business perspective.
Bibliography


