Universiteit van Amsterdam
Amsterdam School of Economics

Master's thesis Actuarial Science and Mathematical Finance

SMALL-AREA STATISTICAL ANALYSES OF CLAIM FREQUENCY IN MOTOR INSURANCE

Martin Haringa (10845666)

supervised by

Dr. K. antonio

Amsterdam, June 2016

Martin Haringa: *Small-Area Statistical Analyses of Claim Frequency in Motor Insurance* , Master's thesis , Master of Science (MSc.), © June 2016

SUPERVISORS:
Dr. K. Antonio
Dr. U. Can

LOCATION:
Amsterdam

TIME FRAME:
June 2016

ABSTRACT

In this thesis the claim frequency in a Belgian motor thirdparty liability (MTPL) insurance portfolio on the municipality level is discussed. In particular, the claim frequency in the MTPL insurance portfolio on the municipality level is analyzed in an hierarchical Bayesian framework using integrated nested Laplace approximations (INLA).

We show that an hierarchical Bayesian framework fit the claim frequency in the MTPL insurance portfolio on the municipality level better than the usual *generalized additive model* (GAM) framework. Moreover, INLA results are obtained in seconds, whereas the computation time to perform a GAM grows rapidly when the number of observations increases.

KEYWORDS: Bayesian hierarchical spatial regression models, integrated nested Laplace approximations (INLA), visualizing spatial data, motor insurance, leave-one-out-cross-validation, spatial econometrics.

All places are related but nearby places are more related than distant places.

— Waldo Tobler, *1970 - First law of geography*

ACKNOWLEDGMENTS

This thesis is made as a completion of the master's degree program in Actuarial Sciences and Mathematical Finance at the University of Amsterdam.

Several persons have contributed academically, practically and with support to this master's thesis. I would therefore firstly like to thank my supervisor Katrien Antonio for her time and support throughout the entire thesis' period.

Finally, I would like to thank my family and friends for being helpful and supportive during my time studying at the University of Amsterdam.

Haarlem, June 2016

CONTENTS

1

INTRODUCTION

In the European motor thirdparty liability (MPTL) insurance industry, an increasing competition among insurers has been encouraged as a mechanism to resort to increasingly advanced statististical methods to analyze the risk profile of the policyholders. Actuaries help formulate a fair and reasonable tariff associated with these risks. A correct premium will be charged for old drivers relative to young drivers, for males relative to females, etc. Insurance portfolios are thus partitioned into classes with all policyholders belonging to the same class paying the same a priori premium (Denuit and Lang, 2004).

The classification variables are called *a priori* variables (as their values can be determined before the policyholder starts to drive). In motor insurance, they include the age and gender of the policyholder, the type and use of their car and the geographical zone where the policyholder lives (urban/nonurban for instance, or a finer geographical resolution according to postal codes or municipalities). It is common to assume that claim characteristics at geographical areas near each other tend to be similar (after other factors have been accounted for). The idea of geographic rating models is to exploit this spatial smoothness by borrowing information over neighboring areas.

The concepts used in Boskov and Verrall (1994) and Brouhns et al. (2002) for premium rating on a small-area level are closely related to those applied in spatial epidemiology to the study of the distribution of disease. Boskov and Verrall (1994) made use of a Markov chain Monte Carlo (MCMC) algorithm to implement a Bayesian approach to the observation in each area. The model is based on the work by Besag, York, and Mollié (1991) who proposed an appropriate way of fitting spatial models using MCMC methods. The model splitted the variability on a small-area level as the sum of a spatially correlated variable (which depends on the values of its neighbors) plus an area-independent effect (which reflects local heterogeneity). Taylor (2001) indicated that the main advantage of the Bayesian framework is that it recognizes the magnitudes of sampling error and incorporates the concept of smoothing over neighboring areas.

An introduction to MCMC and its main applications, including disease mapping, can be found in Gilks, Richardson, and Spiegelhalter (1996).

Although in theory always possible to implement, MCMC algorithms applied to latent Gaussian models come not without problems, both in terms of convergence and computational time. The integrated nested Laplace approximation (INLA hereafter) substitutes MCMC simulations with accurate, deterministic approximations to

posterior marginal distributions (Rue, Martino, and Chopin, 2009). Unlike MCMC methods that sample from the posterior distribution of parameters, INLA uses the mode as the mean and use the derivative in the area of the mode to approximate the variance. The closer the target posterior distribution is to Gaussian, the better the approximation (DiMaggio, 2015). Laplace approximations have been known for some time, but only recently developed sufficiently to be accurate enough for practical application. The main benefit of these approximations is computational: where MCMC algorithms need hours and days to run, INLA provide more precise estimates in seconds.

A detailed description of the INLA method and a thorough comparison with MCMC results can be found in Rue, Martino, and Chopin (2009).

In this thesis, we analyze the claim frequency in a MTPL insurance portfolio on the municipality level with Bayesian hierarchical spatial models using INLA. Bayesian hierarchical models make an appropriate framework for the development of spatially structured models. The model is specified in different layers, so that each one accounts for different sources of variation. For example, the model can cope with socio-demographic factors and traffic-related factors at a small-area level at the same time as borrowing strength from neighbors to improve the quality of estimates. We show that Bayesian hierarchical spatial regression models fit the claim frequency in a MTPL insurance portfolio on the municipality level better than the usual *generalized additive model* (GAM) framework. Moreover, INLA results are obtained in seconds, whereas the computation time to perform a GAM grows rapidly when the number of observations increases. An application to a comprehensive data set from a Belgian motor insurance company studied in Denuit and Lang (2004) is given.

This thesis adds to the existing research in modeling the claim frequency on a small-area level in a motor insurance portfolio in two ways. First, Bayesian hierarchical spatial models using INLA have previously not been used for modeling the claim frequency on a small-area level in a non-life insurance portfolio. And second, the thesis extends the well-recognized association of claim frequency in motor insurance with income and traffic-related factors, which may offer additional insights into the complex interplay of economic and traffic-related environment in vehicle claim risk.

THE REMAINDER OF THIS THESIS proceeds as follows. Chapter 2 presents the data set used in this thesis, concerning a Belgian motor thirdparty liability (MTPL) insurance portfolio. In Chapter 3 a data-driven method to find categorizations for the continuous risk factors in the MTPL insurance portfolio is applied. In Part ii GLM and GAM approaches are compared to the INLA framework to model the claim frequency in the MTPL insurance portfolio on the municipality level.

Chapter 8 gives a final claim frequency model on the policyholder's level (based on the results in Part ii). Chapter 9 concludes.

Part I

DATA AND METHODS

2

DESCRIPTION OF THE DATA SET

In Section 2.1 the data used on the level of the policyholder is described. Section 2.2 describes the data used on the municipality level, and Section 2.3 describes how choropleth maps can be used to obtain better insights in the spatial patterns of the data on the municipality level.

2.1 DATA ON THE POLICYHOLDER LEVEL

The data set at hand consists of 163,660 records concerning a Belgian motor thirdparty liability (MTPL hereafter) insurance portfolio. Third party coverage provides protection in the event the vehicle owner causes harm to another party, who recovers their cost from the policyholder (Denuit et al., 2007). All the observations relate to the year 1997. It is the portfolio Brouhns et al. (2002), Denuit and Lang (2004) and Klein et al. (2014) based their case studies on.

Table 2.1 shows the covariates available in the MTPL insurance portfolio. It should be stressed that the bonus-malus scale level is not a covariate of the same type as the others. It is a function of the number of claims reported in previous years and, as such, should not be incorporated in a priori risk classification (Denuit and Lang, 2004). In addition to these covariates, the number of claims reported during the year 1997 is known for each policyholder (Table 2.2) as well as the municipality where the policyholder lives.

Figure 2.1 gives histograms for the continuous covariates in the MTPL insurance portfolio. The correlations between these continuous covariates are given in Figure 2.2.

2.2 DATA ON THE MUNICIPALITY LEVEL

To obtain insights in the well-recognized association of claim frequency in motor insurance with economic and traffic-related factors, covariates for income, road density and traffic density were added to the data set. These covariates were supplied by the national statistical agency Belgium and are available for each Belgian municipality.

1 In the British Isles they are sometimes called *multi-purpose vehicle* (MPV) or people carriers.

Variable	Description
AGEPH	Age of the policyholder (in years)
AGEC	Age of the car (in years)
POWER	Engine power (in kilowatts)
BM	Level occupied in the 23-level bonus-malus scale (the higher the level occupied, the worse the claim history)
DURATION	Number of days the policy was in force during 1997
COVERAGE	1 = TPL only (58.3%)
	2 = TPL + limited material damage and theft (28.2%)
	3 = TPL + comprehensive damage (13.5%)
FLEET	1 = vehicle does not belong to a fleet (3.2%)
	2 = vehicle belongs to a fleet (96.8%)
FOR	0 = normal vehicle (99.7%)
	1 = 4×4 vehicle (0.3%)
FUEL	1 = gasoline (69.0%)
	2 = diesel (30.8%)
	3 = LPG (0.2%)
	4 = other (0.0%)
MONOVOL	0 = normal vehicle (98.9%)
	1 = minivan[1] (1.1%)
SEX	1 = female policyholder (26.4%)
	2 = male policyholder (73.5%)
	3 = company (0.1%)
SPORT	1 = sports vehicle (0.9%)
	2 = normal vehicle (99.1%)
USE	1 = private use (95.1%)
	2 = professional (4.9%)

Table 2.1: Covariates available in the MTPL insurance portfolio.

The income variable is the median household income per tax return for the year 2014 in increments of thousand euros per municipality (Belgian Federal Government, 2013a). The traffic-related measures are the road density (road length in kilometers per square kilometer) per municipality and the average traffic density (vehicle kilometers traveled per day per square kilometer) per municipality (Belgian Federal Government, 2013b). The traffic density measure is transformed

k	n_k
0	145312
1	16602
2	1562
3	162
4	17
5	2

Table 2.2: Number of claims per policyholder in the MTPL insurance portfolio.

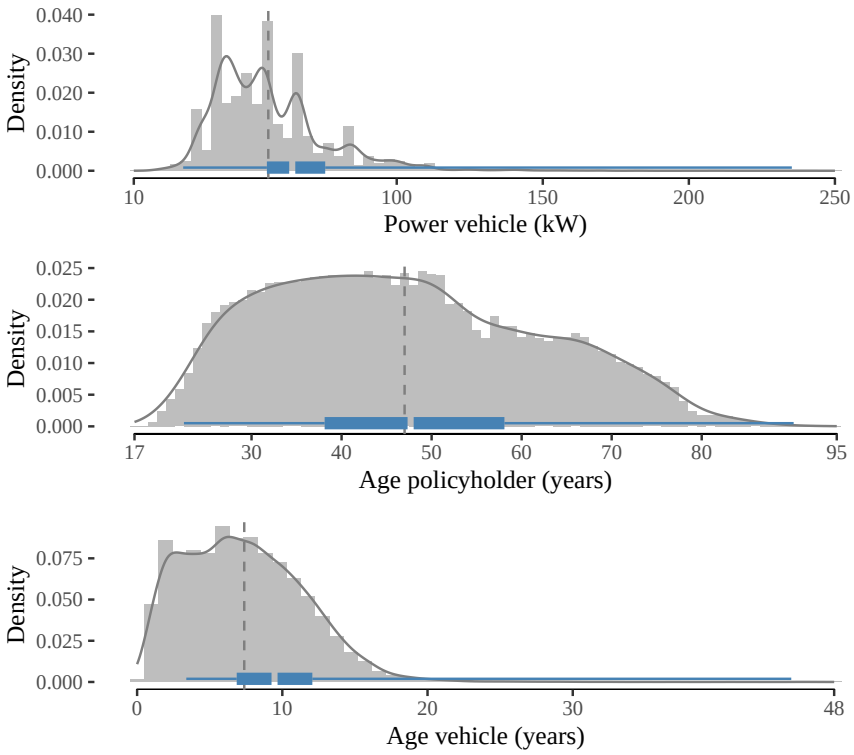


Figure 2.1: Histograms and Tufte's boxplot for the covariates on the policyholder level: power of the vehicle, age of the policyholder and age of the vehicle. The dashed line represents the mean.

into standardized units. Figure 2.3 gives histograms for the median income, road density and traffic density.

According to the data sets provided by the national statistical agency Belgium there are 589 municipalities in Belgium. This number is in line with the number of municipalities in our spatial object. Actually the MTPL insurance portfolio consists of 583 municipalities. To

The spatial object on the municipality level is obtained from the Global Administrative Areas (GADM) database (`http://www.gadm.org`).

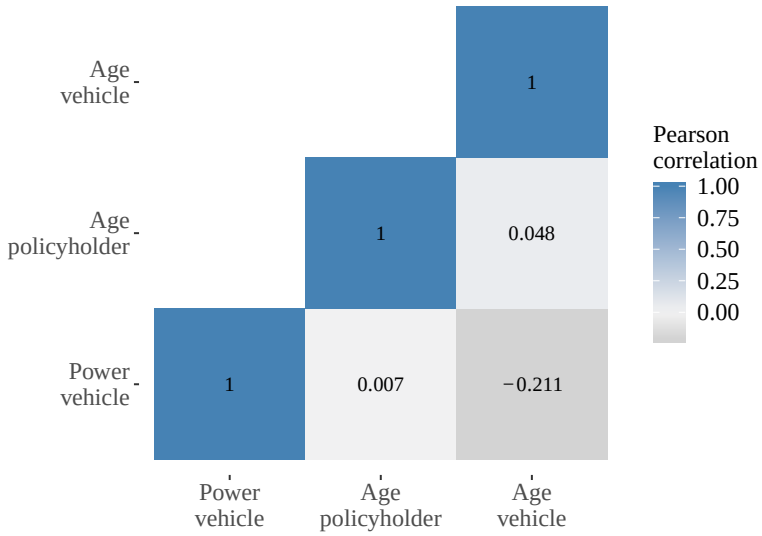


Figure 2.2: Correlation matrix for the covariates on the policyholder level: power of the vehicle, age of the policyholder and the power of the vehicle.

deal with this discrepancy, we set the number of claims and the exposure for the missing municipalities equal to their neighbor means, corrected for population size. This is done by dividing the number of claims and the exposure of the neighbors of the missing municipality by their population size, taking averages of these rates for both the number of claims and the exposure of the neighbors of the missing municipality, and then multiplying these averages by the population size of the missing municipality.

Figure 2.4 gives the correlations between the variables supplied by the national statistical agency Belgium. The high correlation between the traffic density and the number of claims is in line with what is expected.

2.3 CHOROPLETH MAPS

To obtain better insights in the overall geographic patterns of the claim frequency in Belgium, choropleth maps are used in this thesis. A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurements of the statistical variable being displayed on the map.

Two issues need to be addressed in the context of creating choropleth maps: class intervals and color palettes. The first issue, classification, or data partioning, can take many forms including equal

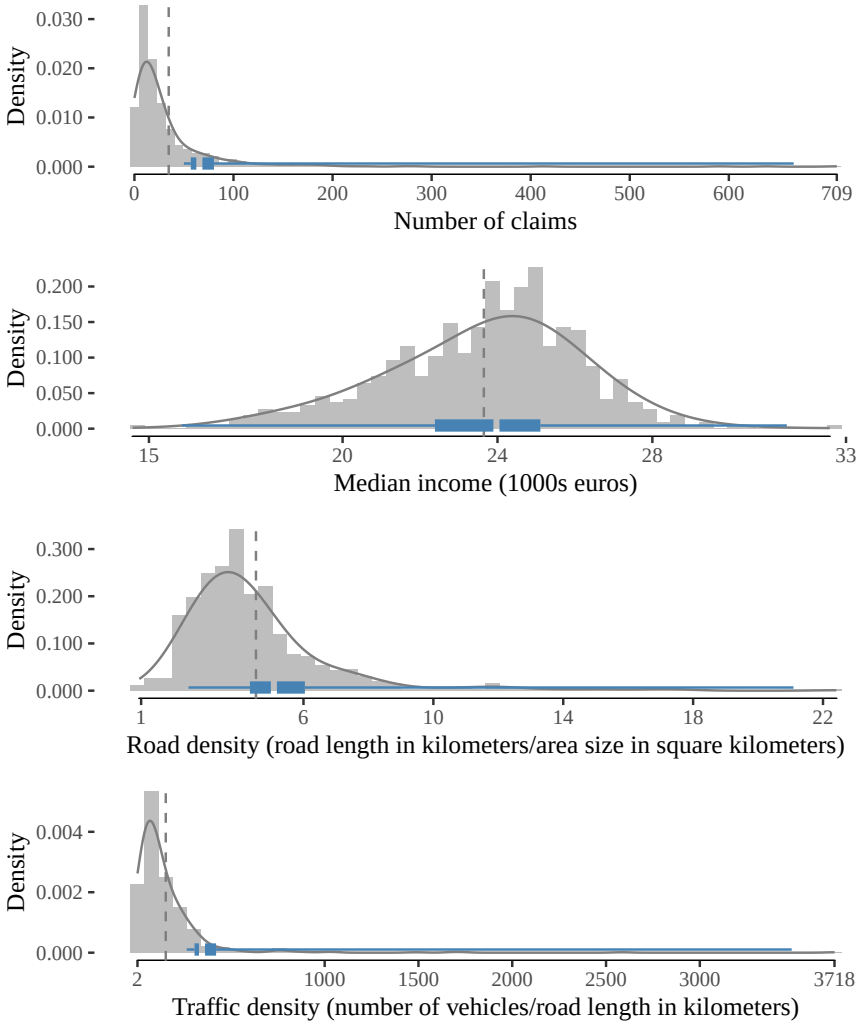


Figure 2.3: Histograms and Tufte's boxplot for the covariates on the municipality level: number of claims, median income, road density and traffic density. The dashed line represents the mean.

interval, frequency, or standard deviations from a mean (Brewer and Pickle, 2002). We follow Bivand, Pebesma, and Gómez-Rubio (2008) and try the quantile method and the Fisher-Jenks method. The empirical cumulative distribution function suggests that using quantiles is not necessarily a good idea (Figure 2.5a). While of course the number of sites in each class is equal by definition, the observed values are far from uniformly distributed. Therefore, the Fisher-Jenks method is used in this thesis (Figure 2.5b). The Fisher-Jenks natural breaks classification method is a classification scheme that finds class breaks

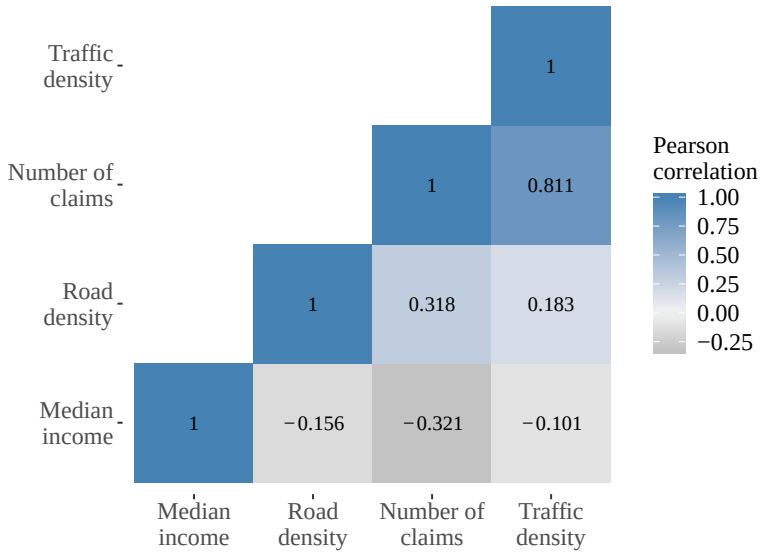


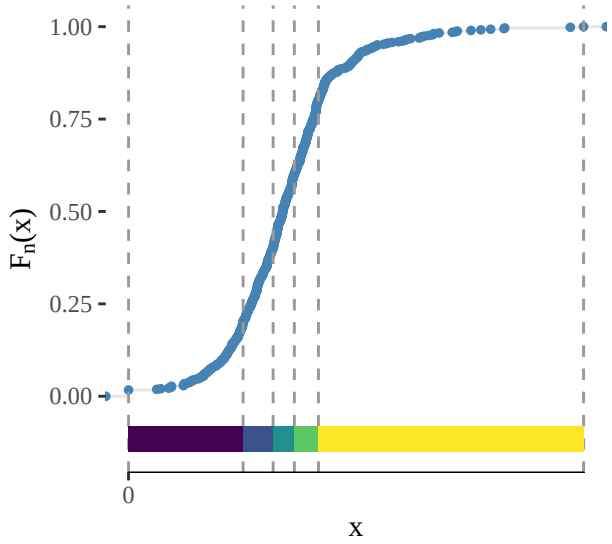
Figure 2.4: Correlation matrix for the covariates on the municipality level: median income, road density, number of claims and traffic density.

that will minimize within-class variance and maximize between-class differences.

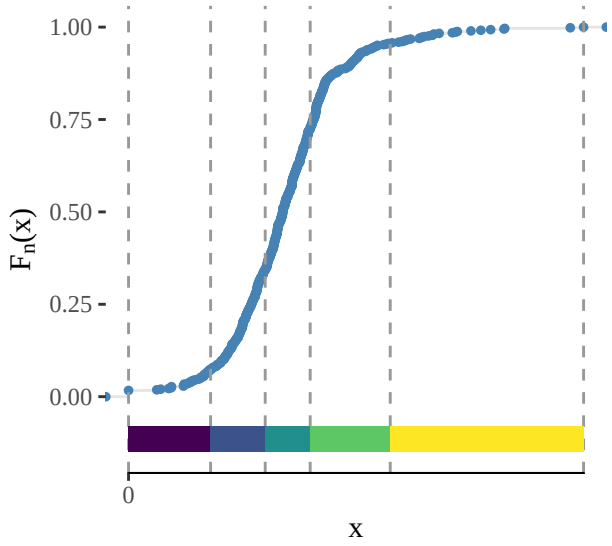
Both methods have been collected in the `classInt` package (Bivand, 2015b) in R.

And second, a specific color progression should be used to depict the data properly. In this thesis the `viridis` palette is used, this palette is easier to read by those with colorblindness, and print well in grey scale. This palette is provided by the `viridis` package (Garnier, 2018) in R. Figure 2.6 projects the ratio between the total number of claims and the total exposure per municipality on a map of Belgium. It is observed that in the Brussels-Capital Region many municipalities exist with higher ratios compared to the nation-wide average. Choropleth maps of the total number of claims per municipality and the total exposure per municipality show that the highest number of claims and highest exposures are observed in the municipalities Antwerp, Ghent, and Charleroi (Figure B.1 in the Appendix).

Is a region in Belgium comprising nineteen municipalities, including the City of Brussels which is the capital of Belgium. The region has a population of 1.2 million.



(a) Quantile classification.



(b) Fisher's natural breaks classification.

Figure 2.5: Empirical cumulation distribution functions for the quantile method and Fisher-Jenks classification method for the ratio between the total number of claims and the total exposure on the municipality level.

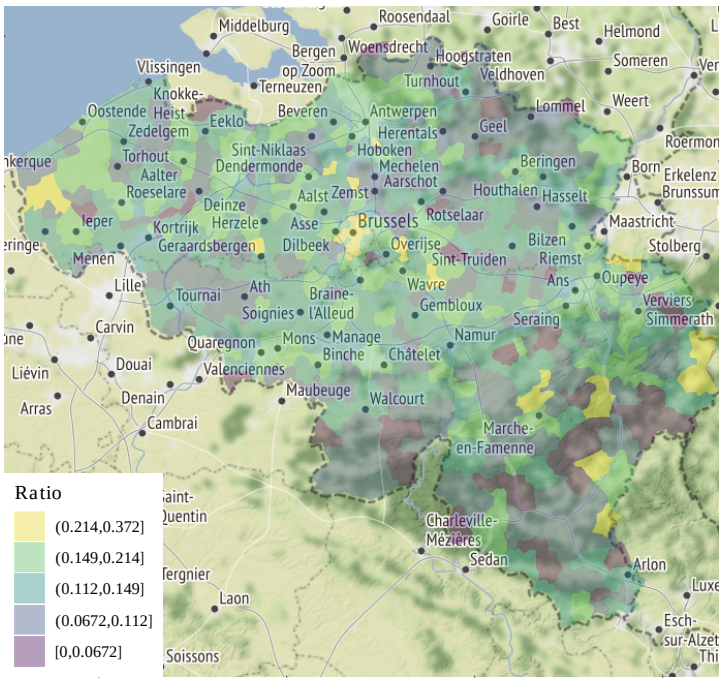


Figure 2.6: Projection of the ratio between the total number of claims per municipality and the total exposure per municipality on the map of Belgium.

3 BINNING CONTINUOUS VARIABLES

As industry practice requires, this thesis uses categorizations for the continuous risk factors. However, in practice the categories are often binned by expert judgement, we follow Clijsters (2015) and utilize a more advanced and data-driven method to find categories that are homogeneous with respect to their response variable. In particular, we fit *generalized additive models* (GAMs) for the continuous covariates in the MTPL insurance portfolio (Section 3.1), and use the resulting nonparametric estimates as input for regression trees, which on their turn will bin the data in risk homogeneous classes (Section 3.2).

3.1 GAM FOR CONTINUOUS VARIABLES

GAMs are very similar to GLMs, but they also allow for including non-linear terms in the linear predictor term (Hastie and Tibshirani, 1990; Wood, 2006). An interesting feature of GAMs is that they provide insight to the specification of meaningful categories for the continuous risk factors (Antonio and Valdez, 2012).

The fitted univariate model for every continuous variable in our data set on the policyholder level has the following structure

$$\log\left(\mathrm{E}(N_i)\right) = f(\mathrm{x}_i), \tag{3.1}$$

where N_i, the observed number of claims, follow a Poisson distribution, x_i is the continuous variable, and f is a smooth function. The degree of smoothness (within certain limits) of f is estimated by minimizing the *unbiased risk estimator* (UBRE) function see Wood, 2006, for more details.

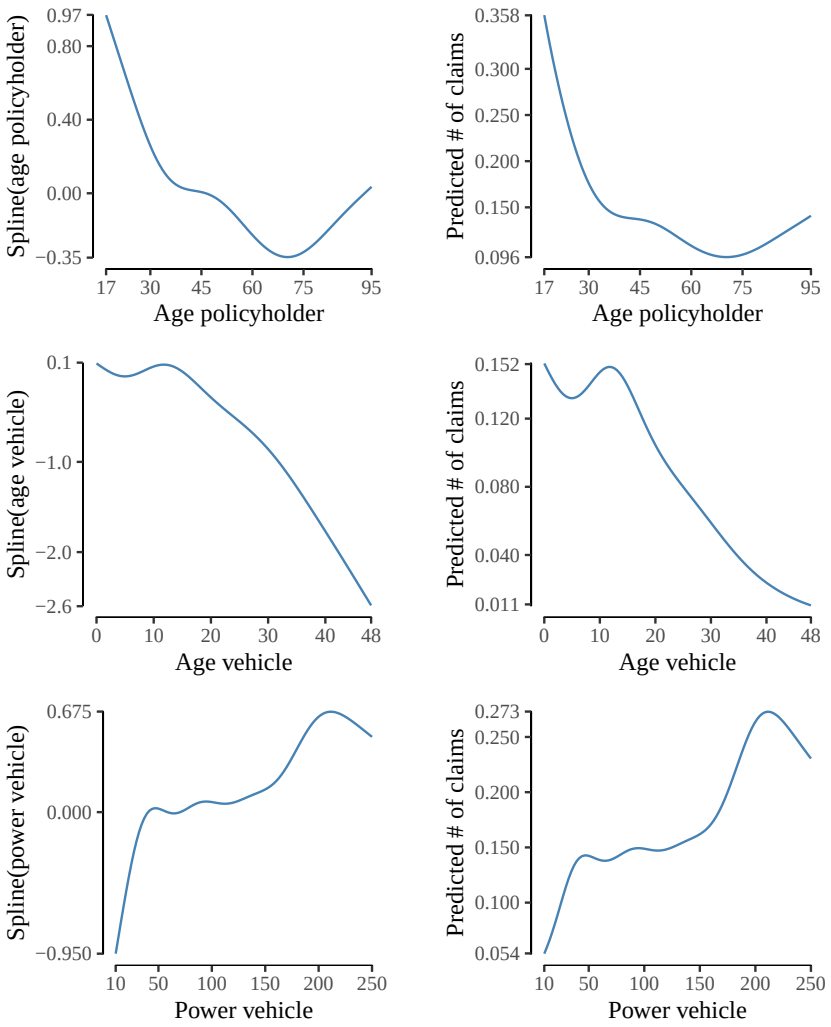
The GAM estimate for Equation 3.1 is given by

$$\log\left(\widehat{N}_i^{pred}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{s}\,(x_i)$$

where $\widehat{\beta}_1$ is the estimate of the linear effect of the continuous variable x_i on the observed number of claims and $\widehat{s}\,(x_i)$ the nonparametric estimate. The `mgcv` package (Wood, 2000) in `R` is used to fit the GAMs.

Figure 3.1a visualizes the nonparametric effect $\widehat{s}\,(\cdot)$ for the age of the policyholder, the age of the vehicle and the power of the vehicle, respectively. Figure 3.1b shows the predicted number of claims for the continuous explanatory variables. To obtain the predicted number of

claims the `predict.gam` function from the `mgcv` package (Wood, 2000) in `R` is used.



(a) Nonparametric effect. (b) Predicted number of claims.

Figure 3.1: Nonparametric effect and the predicted number of claims for the age of the policyholder, the age of the vehicle, and the power of the vehicle.

When we take a closer look at Figure 3.1b, a higher reported number of claims is observed for young policyholders until the age of 35, as well as for drivers above the age of 70. In between, we notice a peak around the age of 45. This peak can be attributed to the fact that, because of the extremely high premiums charged to young policyholders, it has become common practice in Belgium to ask older relatives to purchase the policy. The peak around 45 years could perhaps be attributed to accidents caused by children behind the wheel. New ve-

hicles appear more dangerous the first three years. Again this can be explained by Belgian characteristics. All cars over four years old must undergo annual inspections organized by the State. So, drivers with high annual mileage often keep their car only for three years and then buy a new one. And for vehicles older than twelve years the number of claims decreases. This can perhaps be attributed to the fact that the vehicle is still insured, but no longer used as the main vehicle and thus no longer fully exposed to risk. Finally, we notice the number of claims reported is almost stable for the effect of the power of the vehicle for a power of 50 kilowatt uptil 150 kilowatt. Afterwards, the effect of the power of the vehicle is more or less linear. However, the decreasing effect for extremely high values (above 200 kilowatt) is surprising, but can perhaps also be attributed to the fact that those vehicles are hardly exposed to risk (Denuit and Lang, 2004).

In Belgium, vehicle inspections are known as autokeuring (Dutch) or contrôle technique (French).

3.2 BINNING BY MEANS OF REGRESSION TREES

The next step is to bin the resulting nonparametric GAM estimates (Figure 3.1b) into risk homogeneous categories. From Figure 3.1b it can be observed that categories can be broader when the risk is stable (i.e. small slope), whereas more categories are necessary when the risk varies more (i.e. steep slope). In order to determine splits for creating risk homogeneneous categories *classification and regression trees* (CARTs) are used (Breiman et al., 1984).

The CART methodology is incorporated in the `rpart` *package (Therneau, Atkinson, and Ripley, 2015) in* R.

We consider a regression tree with only one independent variable and one dependent variable, which are the explanatory variable and the corresponding nonparametric GAM estimate for the number of claims, respectively. This means that we are in fact using artificial data sets, containing only one observation for each value of the explanatory variable. For example, the data set used to grow the regression tree for the age of the policyholder consists of 79 observations, namely the ages 17 to 95 and their corresponding GAM estimates (Clijsters, 2015).

Figure 3.1b shows that the number of claims decreases strongly for vehicles with ages above twenty years. For that reason, growing a regression tree would result in a large number of categories for ages above twenty. However, Figure 2.1 shows that ages above twenty only account for a small part of the total exposure and therefore are not of great interest to the insurance company. Consequently, different ages should be weighted in terms of their exposure, such that the regression tree chooses its splits not only based on the values of the nonparametric GAM estimates but also on the relative importance of each of the age groups.

The trees for the continuous variables age of the vehicle and power of the vehicle are given in Figure A.1 (in the Appendix).

The tree obtained for the age of the policyholder variable is given in Figure 3.2. Above each node the split condition is displayed, which is the condition evaluated by the tree to bin the variable policyholder's age based upon its nonparametric GAM estimate. Note that if the split condition is true, the left-hand side path is taken, whereas the path on the right is taken if the condition is false. In Figure 3.3 the splits for the categories are indicated by the vertical dashed lines. Table 3.1 summarizes the obtained categories.

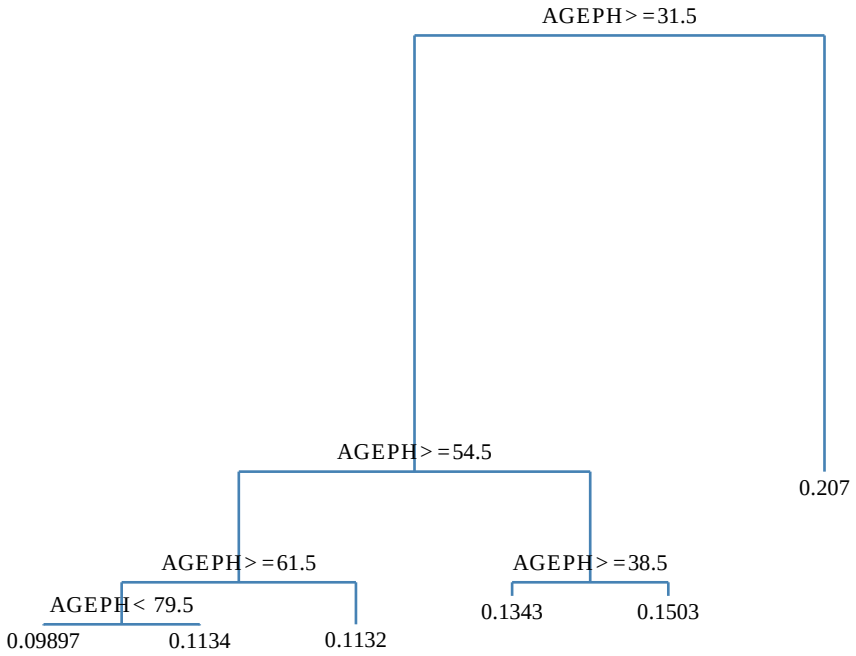


Figure 3.2: Dendrogram for the age of the policyholder with a complexity factor of $c_p = 0$.

It is important to note that the number of categories for the continuous risk factors strongly relies on the choice of the cost-complexity parameter c_p. Smaller values for c_p tend to result in a larger number of categories, whereas larger values for c_p may give a smaller number of categories. As our goal is to find regression trees that generate a *reasonable* number of homogeneous risk categories we evaluate the regression tree for different values of c_p. It turns out that a complexity factor of $c_p = 0$ gives a *reasonable* number of categories for both the age of the policyholder and the age of the vehicle. For the power of the vehicle a complexity factor of $c_p = 0$ results in 17 categories, therefore a complexity factor of $c_p = 0.01$ is choosen, which gives a more *reasonable* number of categories.

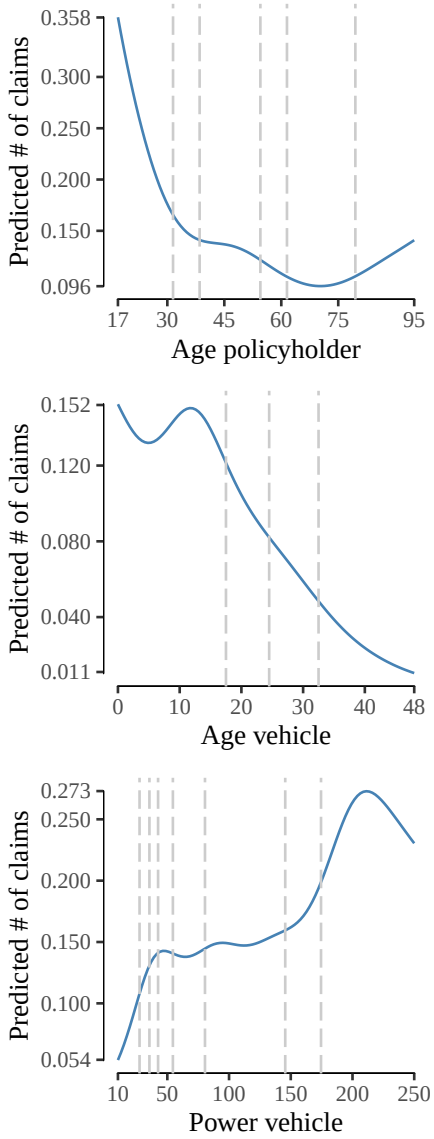


Figure 3.3: Categories added to the predicted number of claims for continuous risk factors.

Variable	Categories							
Age policyholder (year)	17-31	32-37	38-54	55-61	62-79	≥ 81		
Age vehicle (year)	0-17	18-24	25-32	≥ 33				
Power vehicle (kW)	0-27	28-34	35-41	42-55	56-78	79-157	158-174	≥ 175

Table 3.1: Categorizations for the continuous risk factors.

Part II

MODELING APPROACHES

In this chapter the GLM and GAM approaches are compared to the more sophisticated INLA methodology to model the claim frequency in the MTPL insurance portfolio on the municipality level. The GLM approach is presented in Chapter 4. The GAM approach is given in Chapter 5, and in Chapter 6 Bayesian hierarchical spatial regression models using INLA are applied to our data set. Chapter 7 compares the results from the three modeling approaches.

4

GLM

Within the context of *a priori* ratemaking, it is now common practice since McCullagh and Nelder (1989) to use *generalized linear models* (GLMs). GLMs can be used to predict the claim frequency on the policyholder level and ultimately the claim frequency on the municipality level.

The paper by Haberman and Renshaw (1996) gives a comprehensive review of its applications to actuarial problems.

Recall that the continuous variables age of the policyholder, age of the vehicle, and power of the vehicle are binned based on the outcome of the GAM and the corresponding regression trees, as summarized in Table 3.1. The mean $E(Y_i)$ of the independent Poisson distributed claim count Y_i is modeled as an exponential function of the linear predictor η_i, i.e. $E(Y_i) = \exp(\eta_i)$, with the predictor equal to

$$\begin{aligned}\eta_i^{\text{freq}} =& \gamma_0 + \mathit{offset}_i + \sum_{k=1}^{6} \gamma_1 \left(\mathit{age\ policyholder}_{ik}\right) + \\ & \sum_{k=1}^{4} \gamma_2 \left(\mathit{age\ vehicle}_{ik}\right) + \sum_{k=1}^{8} \gamma_3 \left(\mathit{power\ vehicle}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_4 \left(\mathit{coverage}_{ik}\right) + \sum_{k=1}^{2} \gamma_5 \left(\mathit{fleet}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_6 \left(\mathit{for}_{ik}\right) + \sum_{k=1}^{4} \gamma_7 \left(\mathit{fuel}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_8 \left(\mathit{monovol}_{ik}\right) + \sum_{k=1}^{3} \gamma_9 \left(\mathit{sex}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_{10} \left(\mathit{sport}_{ik}\right) + \sum_{k=1}^{2} \gamma_{11} \left(\mathit{use}_{ik}\right) + \\ & \gamma_{12} \left(\mathit{median\ income}_i\right) + \gamma_{13} \left(\mathit{traffic\ density}_i\right) + \\ & \gamma_{14} \left(\mathit{road\ density}_i\right), \end{aligned} \tag{4.1}$$

where the offset_i is defined as the logarithm of the vehicle exposure (i.e. $\mathit{offset}_i = \log\left(\mathit{duration}_i / 365\right)$). Adding an *offset* seems reasonable since the number of claims per policyholder is directly proportional to the vehicle exposure per policyholder The *offset* regression coefficient is known to be 1.

We use the `stepAIC` function from the `MASS` package (Venables and Ripley, 2002) in `R` to perform stepwise model selection by AIC on Model 4.1. The model with the lowest AIC score becomes

$$\begin{aligned}\eta_i^{\text{freq}} =& \gamma_0 + \mathit{offset}_i + \sum_{k=1}^{6} \gamma_1 \left(\mathit{age\ policyholder}_{ik}\right) + \\ & \sum_{k=1}^{4} \gamma_2 \left(\mathit{age\ vehicle}_{ik}\right) + \sum_{k=1}^{8} \gamma_3 \left(\mathit{power\ vehicle}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_4 \left(\mathit{coverage}_{ik}\right) + \sum_{k=1}^{2} \gamma_5 \left(\mathit{fleet}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_6 \left(\mathit{for}_{ik}\right) + \sum_{k=1}^{4} \gamma_7 \left(\mathit{fuel}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_9 \left(\mathit{sex}_{ik}\right) + \gamma_{12} \left(\mathit{median\ income}_i\right) + \\ & \gamma_{13} \left(\mathit{traffic\ density}_i\right) + \gamma_{14} \left(\mathit{road\ density}_i\right), \end{aligned} \tag{4.2}$$

where all elements are defined as in Equation 4.1. A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as *overdispersion* and indicates that the model is not appropriate. In our case the inter-vehicle variability is (very) small (overdispersion taken to be 1.0312). Consequently, a (basic) Poisson GLM seems appropriate.

The predicted number of claims according to Equation 4.2 on the policyholder level is aggregated on the municipality level. At this point, a basic estimate of the vehicle claim risk in a given municipality can be computed as the observed number of claims per municipality divided by the predicted number of claims per municipality. This ratio is known as the *standardized incidence ratio* (SIR) (Bivand, Pebesma, and Gómez-Rubio, 2008). SIR values greater than 1.0 indicate more claims observed in a given municipality than expected. The resulting SIRs according to the GLM approach are presented in Figure 4.1.

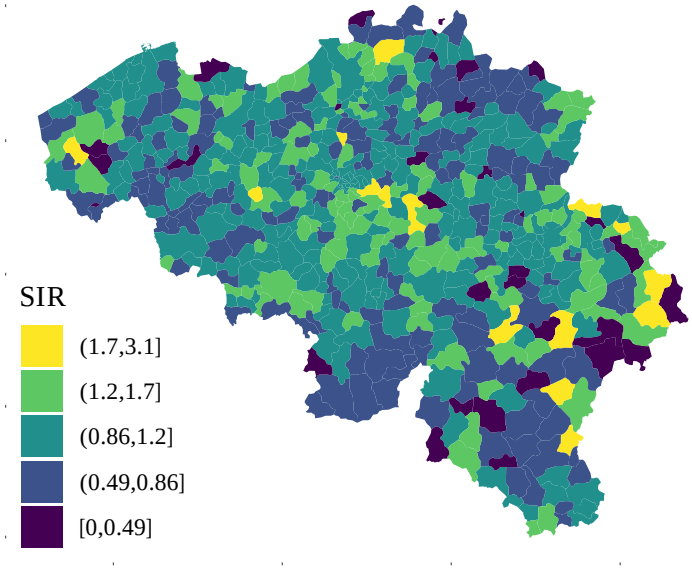


Figure 4.1: *Standardized incidence ratio* (SIR) based on the GLM approach per municipality.

5

GAM

Generalized additive models (GAMs) are very similar to GLMs, but they also allow for including non-linear terms in the linear predictor term (Wood, 2006).

In our GAM, a second order spline on the sphere (based on Wendelberger, 1981) is used to add smooth spatial structure from the residuals to the GLM approach (as defined in Model 4.2). This spline on the sphere is the analogue of a second order thin plate spline in a two-dimensional space. The two arguments to such a smooth are taken to be latitude (in degrees) and longitude (in degrees). These pairs are determined as the centroids of the municipalities.

The function `gCentroid` in the `rgeos` package (Bivand and Rundel, 2016) in `R` is used to calculate the centroids of the municipalities. The `mgcv` package in `R` is used to fit the GAMs.

The predicted number of claims on the level of the policyholder is aggregated on the municipality level. The resulting SIRs are presented in Figure 5.1. The SIRs according to the GAM approach show a similar spatial pattern to the SIRs according to the GLM approach (as calculated in Chapter 4).

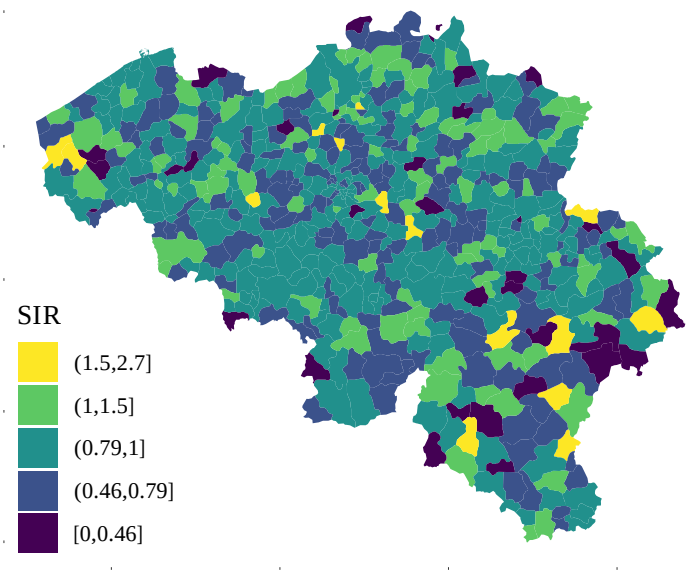


Figure 5.1: *Standardized incidence ratio* (SIR) based on the GAM approach per municipality.

6 INLA

In Section 6.2 different Bayesian hierarchical spatial regression models using INLA are applied to our data set. In Section 6.1 two methods to determine the expected number of claims per municipality are compared. The method where the ratios between the observed number of claims per municipality and the expected number of claims per municipality are closest to one are used in the Bayesian hierarchical spatial regression models in Section 6.2.

6.1 Modeling the expected number of claims per municipality

In this section the expected number of claims per municipality e_i is determined using the *indirect standardization* method and using a GLM. The results from the approach giving the expected number of claims per municipality closest to the observed number of claims per municipality will be used in the *offset* variable in the Bayesian hierarchical spatial regression models in Section 6.2.

6.1.1 *Indirect standardization*

Indirect standardization seeks to answer the question: What would be the number of claims expected if policyholders within a certain municipality contracted claims at the same rate as policyholders within Belgium? (Mausner, Kramer, and Bahn, 1985).

Since our data set includes a variable indicating the municipality, it is easy to sum the number of observed claims and exposure per municipality, which is denoted by Y_i^{obs}, and

$$EXP_i = duration_i/365,$$

for municipality i, where *duration* is aggregated on the municipality level. Summing again over all municipalities give the totals which are denoted by Y_+^{obs} and EXP_+. The expected number of claims in municipality i is calculated as

$$e_i = EXP_i \cdot r_+,$$

where r_+ is the overall incidence ratio

$$r_+ = \frac{Y_+^{obs}}{EXP_+}.$$

At this point, a basic estimate of the vehicle claim risk in a given municipality can be computed as

$$\text{SIR}_i = \frac{Y_i^{obs}}{e_i}$$

(Bivand, Pebesma, and Gómez-Rubio, 2008). In Figure 6.1 the SIRs based on the indirect standardization approach are given.

It is important to note that standardized rates represent *summaries* across population strata experiencing differing risks. As summaries, standardized rates may mask important differences occuring at the stratum level (Bivand, Pebesma, and Gómez-Rubio, 2008). Therefore, a GLM is used to determine the expected number of claims per municipality. In the GLM the covariates given in our MTPL insurance portfolio are taken into consideration.

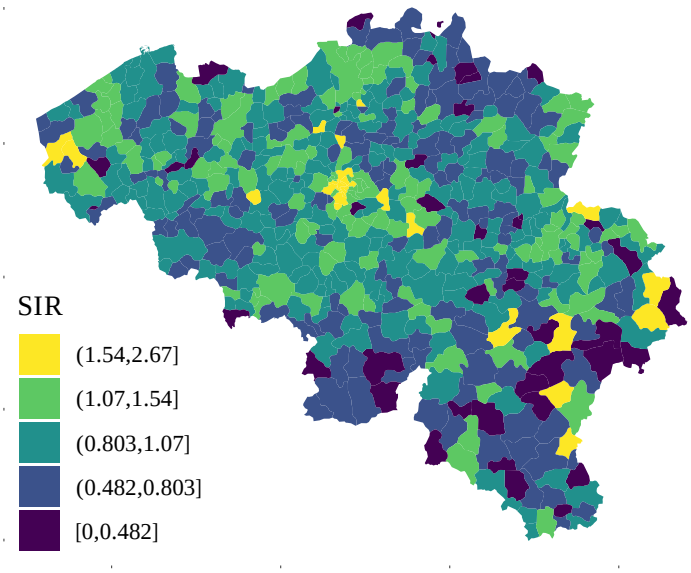


Figure 6.1: Observed number of claims per municipality divided by the predicted number of claims based on the indirect standardization method per municipality.

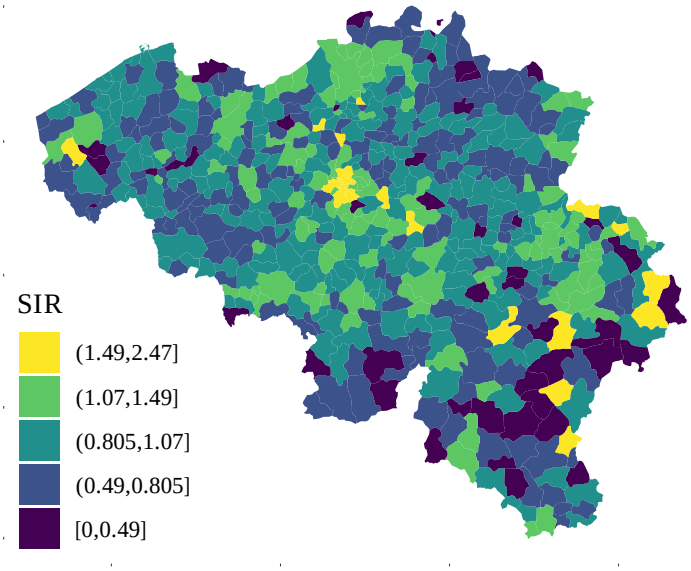


Figure 6.2: Observed number of claims per municipality divided by the predicted number of claims based on the GAM approach per municipality.

6.1.2 *Generalized linear models (GLMs)*

In this section we use a similar approach to the GLM approach described in Chapter 4. Recall that the continuous variables age of the policyholder, age of the vehicle, and power of the vehicle are binned based on the outcome of the GAM and the corresponding regression trees, as summarized in Table 3.1. The mean $E(e_i)$ of the independent Poisson distributed claim count e_i is modeled as an exponential func-

tion of the linear predictor η_i, i.e. $E(e_i) = \exp(\eta_i)$, with the predictor equal to

$$\begin{aligned}\eta_i^{\text{freq}} =& \gamma_0 + \mathit{offset}_i + \sum_{k=1}^{6} \gamma_1 \left(\mathit{age\ policyholder}_{ik}\right) + \\ & \sum_{k=1}^{4} \gamma_2 \left(\mathit{age\ vehicle}_{ik}\right) + \sum_{k=1}^{8} \gamma_3 \left(\mathit{power\ vehicle}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_4 \left(\mathit{coverage}_{ik}\right) + \sum_{k=1}^{2} \gamma_5 \left(\mathit{fleet}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_6 \left(\mathit{for}_{ik}\right) + \sum_{k=1}^{4} \gamma_7 \left(\mathit{fuel}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_8 \left(\mathit{monovol}_{ik}\right) + \sum_{k=1}^{3} \gamma_9 \left(\mathit{sex}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_{10} \left(\mathit{sport}_{ik}\right) + \sum_{k=1}^{2} \gamma_{11} \left(\mathit{use}_{ik}\right), \end{aligned} \tag{6.1}$$

where the offset_i is defined as the logarithm of the vehicle exposure (i.e. $\mathit{offset}_i = \log\left(\mathit{duration}_i / 365\right)$).

We again use the `stepAIC` function from the `MASS` package in `R` to perform stepwise model selection by AIC (Venables and Ripley, 2002). The model with the lowest AIC score becomes

$$\begin{aligned}\eta_i^{\text{freq}} =& \gamma_0 + \mathit{offset}_i + \sum_{k=1}^{6} \gamma_1 \left(\mathit{age\ policyholder}_{ik}\right) + \\ & \sum_{k=1}^{4} \gamma_2 \left(\mathit{age\ vehicle}_{ik}\right) + \sum_{k=1}^{8} \gamma_3 \left(\mathit{power\ vehicle}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_4 \left(\mathit{coverage}_{ik}\right) + \sum_{k=1}^{2} \gamma_5 \left(\mathit{fleet}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_6 \left(\mathit{for}_{ik}\right) + \sum_{k=1}^{4} \gamma_7 \left(\mathit{fuel}_{ik}\right) + \sum_{k=1}^{3} \gamma_9 \left(\mathit{sex}_{ik}\right). \end{aligned} \tag{6.2}$$

The predicted number of claims on the level of the policyholder is aggregated on the municipality level. Figure 6.2 shows the resulting SIRs. The spatial pattern of the SIRs based on the GLM approach (Figure 6.2) is very similar to the spatial pattern of the SIRs based on the indirect standardization approach (Figure 6.1).

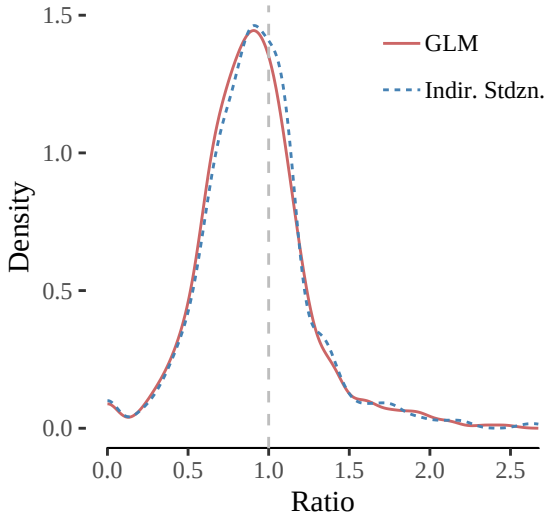


Figure 6.3: Density plot of the SIRs according to the indirect standardization method and the GLM approach.

6.1.3 *Conclusion*

In Figure 6.3 we compare the densities of the SIRs of the GLM approach and the *indirect standardization* approach. It is observed that the ratios from the *indirect standardization* approach and the GLM approach yield very similar results. It seems that the ratios obtained from the indirect standardization method are closest to one. Therefore, the expected number of claims per municipality according to the indirect standardization approach will be used in the *offset* variable in the Bayesian hierarchical spatial regression models in Section 6.2.

6.2 MODEL SPECIFICATIONS

In this section different Bayesian hierarchical spatial regression models are applied to our data set. In particular, a baseline model that consisted solely of an intercept term with unstructured and spatially structured random effect terms is extended to include for income and traffic-related covariates Moreover, the association of these covariates to the vehicle claim frequency at the municipality level is quantified and interpreted.

6.2.1 *Random Effects Model*

The model that consisted solely of an intercept term and a spatially unstructured random effects component is used as a baseline against which to measure more informative models. The model is defined as

$$Y_i \sim \text{Poisson}\left(\lambda_i = e_i\theta_i\right) \tag{6.3}$$

$$\log(\theta_i) = \eta_i + v_i \tag{6.4}$$

$$v \sim \mathcal{N}\left(0, \tau_v\right) \tag{6.5}$$

where,

(6.3) the number of claims Y_i is Poisson distributed, in municipality $i = 1, \ldots, 589$, and where the mean $\lambda_i = e_i\theta_i$ is defined in terms of a claim rate θ_i and the expected number of claims e_i (as determined in Section 6.1.1) as offset,

(6.4) a logarithmic transformation, $\log\left(\lambda_i\right)$, allows

$$\eta_i^{\text{freq}} = \beta_0,$$

where β_0 is an *intercept*, interpreted as the average nation-wide risk on the log scale, along with

(6.5) a spatially *unstructured* random effects component $\left(v_i\right)$ that is independent and identically normally distributed with mean zero.

The R package and documentation are available on http://www.r-inla.org/.

In order to fit the model, we use the R package INLA (Rue, Martino, and Chopin, 2009). A call to R-INLA consists of two parts: a formula statement, followed by a model fitting statement.

Listing 6.1 gives the implementation of the random effects model in R-INLA. In this statement the offset e is the expected number of claims per municipality e_i according to the *indirect standardization* method presented in Section 6.1.1.

Listing 6.1: Implementation of random effects model in R-INLA.

```
formula.re <- Y ~ f(municipality, model = "iid")
inla.re <- inla(formula.re, family = "poisson",
                data = claim.data, E = e)
```

The fixed effects consist only of the intercept β_0, which has a mean of 0.9864 (95% credible interval = 0.9683 - 1.0043), and translates to a yearly rate of 9.86% claims. The fixed effects are stored in inla.re-$summary.fixed. The mean random effect term for each municipality

is the random variation, on the logarithmic scale, around the mean or intercept value of the number of claims in a municipality. The random effects are stored in `inla.re$summary.random$municipality[,2]`.

Figure 6.4 shows that the spatially unstructured heterogeneity random effect term appears reasonably normally distributed. Figure 6.5 shows that the spatially unstructured heterogeneity random effect term appears spatially random.

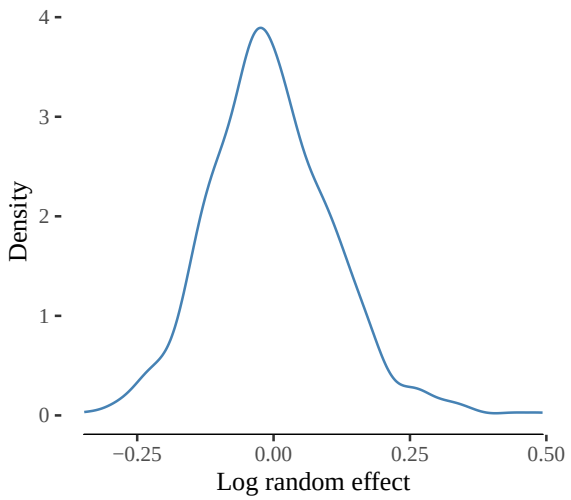


Figure 6.4: Random effects term for random effects model.

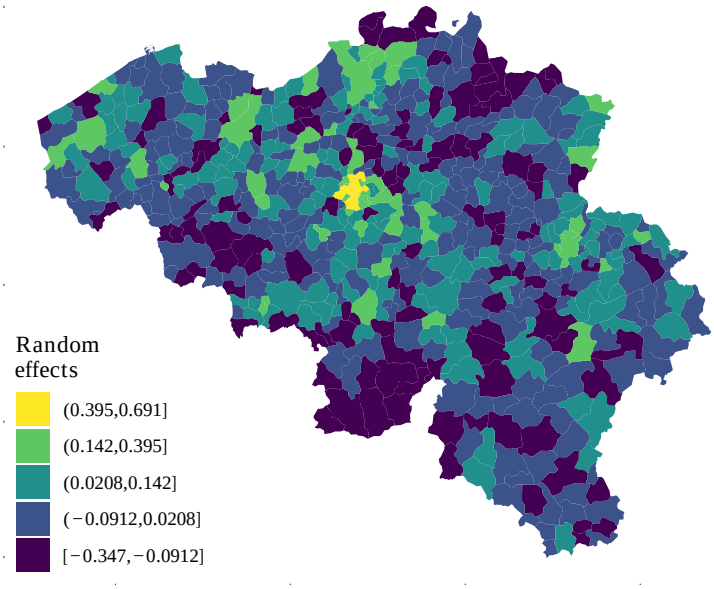


Figure 6.5: Unstructured heterogeneity for random effects model.

6.2.2 *Besag-York-Mollié (BYM) Model*

The so-called Besag-York-Mollié (BYM henceforth) model (Besag, York, and Mollié, 1991) adds a spatially-structured conditional autoregression term (ν) to the spatially-unstructured heterogeneity random effect term (υ) of the random effects model defined in the previous paragraph.

We restrict ourselves to only the incorporation of the BYM model in our vehicle claim frequency framework. For a general introduction to the BYM model and the use in Bayesian hierarchical disease mapping see Banerjee, Carlin, and Gelfand (2014), DiMaggio (2015), Haining (2003), Schabenberger and Gotway (2004), and Waller and Gotway (2004).

The BYM model is defined as

$$Y_i \sim \text{Poisson}\left(\lambda_i = e_i\theta_i\right) \tag{6.6}$$

$$\log(\theta_i) = \eta_i + \upsilon_i + \nu_i \tag{6.7}$$

$$\upsilon \sim \mathcal{N}\left(0, \tau_\upsilon\right) \tag{6.8}$$

$$\nu \sim \mathcal{N}\left(\bar{\nu}_\delta, \tau_\nu / \nu_\delta\right) \tag{6.9}$$

where,

(6.6) the number of claims Y_i is Poisson distributed, in municipality $i = 1, \ldots, 589$, and where the mean $\lambda_i = e_i\theta_i$ is defined in terms of a claim rate θ_i and the expected number of claims e_i (as determined in Section 6.1.1) as offset,

(6.7) a logarithmic transformation, $\log\left(\lambda_i\right)$, allows

$$\eta_i^{\text{freq}} = \beta_0,$$

where β_0 is an *intercept*, interpreted as the average nation-wide risk on the log scale, along with

(6.8) a spatially *unstructured* random effects component (υ_i) that is independent and identically normally distributed with mean zero, and

(6.9) a conditional autoregressive spatially structured component, in which a *neighborhood* consisting of spatially adjacent municipalities is characterized by the normally distributed mean of the spatially structured random effect terms for the municipalities that make up the neighborhood $\bar{\nu}_\delta$, and the standard deviation of that mean divided by the number of municipalities in the neighborhood τ_ν / ν_δ.

`"bym"` refers to Besag, York, Mollié to whom this model is attributed.

Listing 6.2 gives the implementation of the BYM model in `R-INLA`. In `R-INLA` the model can be specified using the single `"bym"`. The `"bym"` specification contains both the `"iid"` and standard `"besag"` model specifications, and where `"belgium.graph"` is an adjacency matrix, saying which municipalities are neighbors to each other. We created a neighbors object for the municipality contiguities using the

`poly2nb` function in the `spdep` package (Bivand and Piras, 2015) in `R`. It takes an object extending the `SpatialPolygonsDataFrame` class as its first argument, and using heuristics identifies polygons sharing boundary points as neighbors. By default, the contiguity condition is met when at least one point of its neighbor. This relationship is given by the argument `queen=TRUE` by analogy with movements on a chessboard (Figure 6.6).

Listing 6.2: Implementation of BYM model in R-INLA.

```
formula.car <- Y ~ f(municipality, model = "bym",
                     graph = "belgium.graph")
inla.car <- inla(formula.car, family = "poisson",
                   data = claim.data, E = e)
```

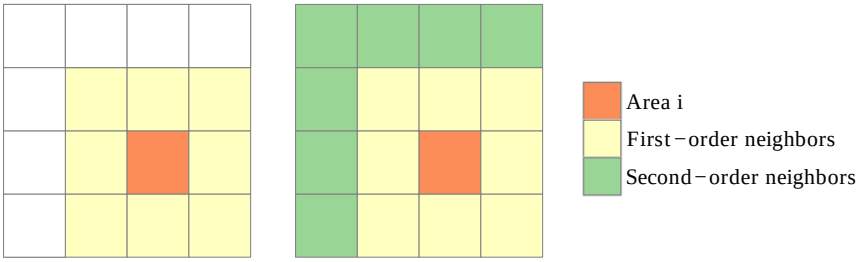


Figure 6.6: Neighboring structure according to the queen contiguity for area i: first-order neighbors (left), first- and second-order neighbors (right).

Figure 6.7 gives the queen contiguities for Belgium on the municipality level. The neighbors object obtained using the `poly2nb` function is then formatted through the `nb2INLA` function from the `inla` package in `R` to get the neighbors object into the correct format for `R-INLA` (Bivand, 2015a).

The intercept β_0 is stored in `inla.car$summary.fixed`, and can be interpreted as the average claim risk across all municipalities, which is a yearly rate of 9.86% claims (95% credible interval = 9.67 - 10.05).

The results for each component of the model are concatenated in a single `summary.random$municipality` result, with the random effects terms listed first, followed by the conditional autoregressive terms. In Figure 6.8 it is observed that the density plot of the random effects terms of the BYM model looks similar to the random effects terms of the random effects model (Figure 6.4). Figure 6.9 shows that the

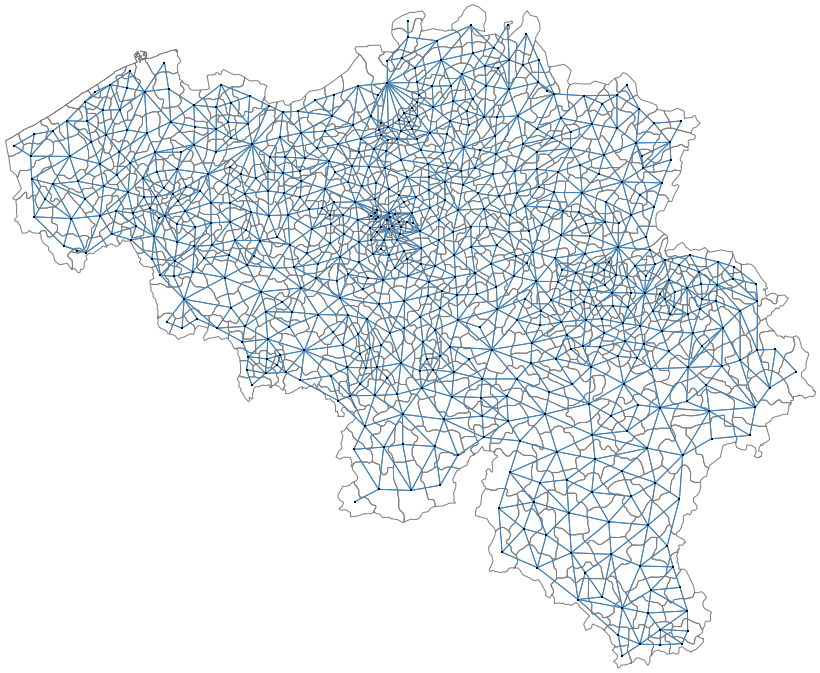


Figure 6.7: Belgian municipality queen contiguities. A line is shown when municipalities touch.

density plot of the conditional autoregressive term of the BYM model appears reasonably normally distributed and symmetric about zero.

Figure 6.10 shows a spatial structuring of vehicle claim risk, with nearby municipalities demonstrating similar risk. Figure 6.11 presents the *spatial risk* ($\zeta = v + \nu$) term at the municipality level. The computation of the posterior mean for the random effects ($\zeta = v + \nu$) is performed in two steps as we have more than one parameter. First, the marginal posterior distribution for each element of the random effects is extracted and then the exponential transformation is applied to calculate the posterior mean for each of them. Spatial risk is interpreted as the residual relative claim risk for each municipality (compared to the national average of Belgium), identifying municipalities of increased claim risk.

Finally, it could be interesting to evaluate the proportion of variance explained by the structured spatial component. The quantity σ_ν^2 is the variance of the conditional autoregressive specification, while σ_v^2 is the variance of the marginal unstructured component. Thus, the two are not directly comparable. Nevertheless, it is possible to ob-

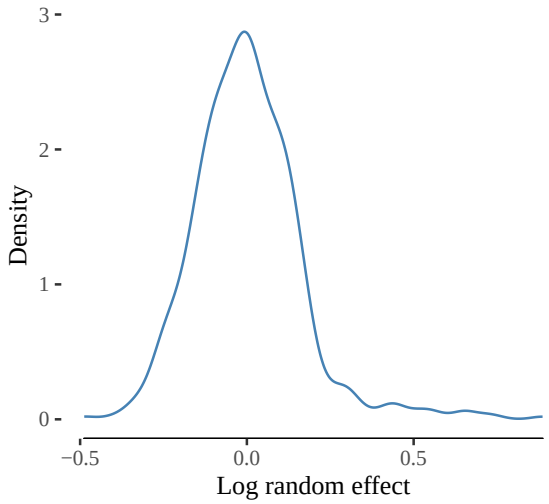


Figure 6.8: Density plot for random effects in the BYM model.

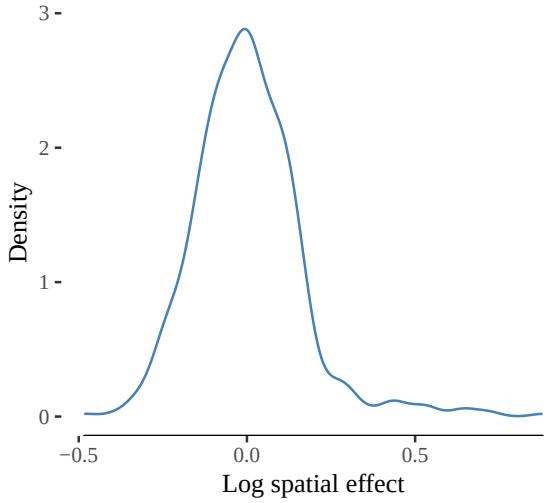


Figure 6.9: Density plot for conditional autoregressive term in the BYM model.

tain empirically an estimate of the posterior marginal variance for the structured effect through

$$s_v^2 = \frac{\sum_{i=1}^{n} (v_i - \bar{v})^2}{n-1},$$

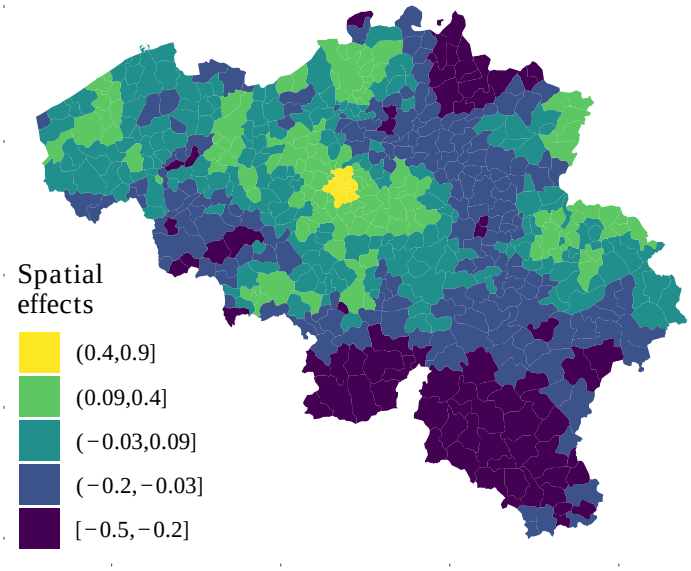


Figure 6.10: Spatially structured heterogeneity (v) in BYM model.

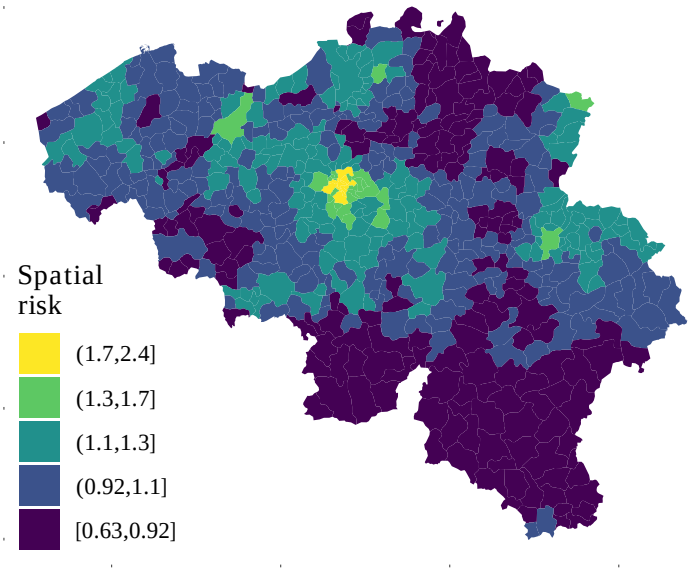


Figure 6.11: Spatially structured risk estimates $(\zeta = v + \nu)$ in BYM model.

where $\bar{v}$ is the average of $\nu_1, \ldots, \nu_{589}$, and then compare it to the posterior marginal variance for the unstructured effect, provided by σ_v^2, as

$$\text{frac}_{\text{spatial}} = \frac{s_\nu^2}{s_\nu^2 + \sigma_v^2} \tag{6.10}$$

(Blangiardo et al., 2013).

The proportion of spatial variance is about 97.5% suggesting that the variability is almost completely explained by spatial structure.

6.2.3 *BYM Model with Covariates*

We add the *median income*, *traffic density* and *road density* covariates to the BYM model given the previous paragraph. The model becomes

$$Y_i \sim \text{Poisson}\left(\lambda_i = e_i\theta_i\right) \tag{6.11}$$

$$\log(\theta_i) = \eta_i + v_i + \nu_i \tag{6.12}$$

$$v \sim \mathcal{N}\left(0, \tau_v\right) \tag{6.13}$$

$$\nu \sim \mathcal{N}\left(\bar{v_\delta}, \tau_\nu / \nu_\delta\right) \tag{6.14}$$

where,

(6.11) the number of claims Y_i is Poisson distributed, in municipality $i = 1, \ldots, 589$, and where the mean $\lambda_i = e_i\theta_i$ is defined in terms of a claim rate θ_i and the expected number of claims e_i (as determined in Section 6.1.1) as offset,

(6.12) a logarithmic transformation, $\log\left(\lambda_i\right)$, allows a linear, additive model of regression terms

$$\begin{aligned}\eta_i^{\text{freq}} =& x_i'\beta \\ =& \beta_0 + \beta_1(\textit{income})_i + \beta_2(\textit{road density})_i + \\ & \beta_3(\textit{traffic density})_i\end{aligned}$$

where β_0 is an *intercept*, interpreted as the average nation-wide risk on the log scale adjusted for the covariates and random effects, β_1, β_2 and β_3 are *regression parameters*, and income_i, road density_i and traffic density_i are *explanatory variables* on the municipality level, along with

(6.13) a spatially *unstructured* random effects component (v_i) that is independent and identically normally distributed with mean zero. This unstructured heterogeneity represents, essentially, *noise* that arises from the data that we cannot capture with our covariates, and

(6.14) a conditional autoregressive spatially structured component, in which a *neighborhood* consisting of spatially adjacent municipalities is characterized by the normally distributed mean of the spatially structured random effect terms for the municipalities that make up the neighborhood $\bar{v}_\delta$, and the standard deviation

of that mean divided by the number of municipalities in the neighborhood τ_ν / ν_δ.

Listing 6.3 gives the implementation of this model in R-INLA.

Listing 6.3: Implementation of BYM model with covariates in R-INLA.

```
formula.cov <- Y ~ f(municipality, model = "bym",
                     graph = "belgie.graph") +
                     median.income + traffic.density +
                     road.density
inla.cov <- inla(formula.cov, family = "poisson",
                 data = claim.data, E = e)
```

The results for the fixed effects of the covariates for the final covariate model are presented in Figure 6.12. Holding all the other covariates to zero, for every thousand euros increase in the median income per municipality, there is a 1.7% decrease in vehicle claim risk (95% credible interval = 0.9730 - 0.9930), every single unit increase in the traffic density is associated with 0.8% increase in vehicle claim risk (95% CrI = 0.999 - 1.016), and every one single unit increase in the standardized road density is associated with 3.1% increase in vehicle claim risk (95% CrI = 1.020 - 1.042).

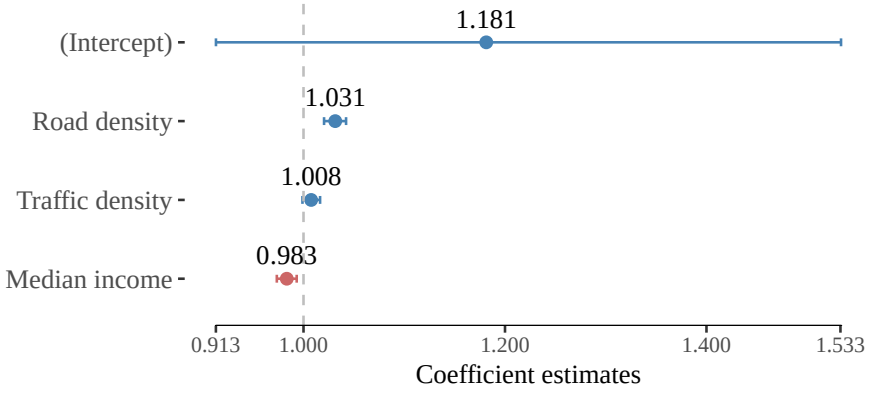


Figure 6.12: Summary statistics: posterior mean and posterior 95% credibility interval for the fixed effect of the final covariate model. Positive effects (> 1) in blue.

Figure 6.13 and Figure 6.14 show that the density plots for the random effect term and the conditional autoregressive term for this model specification are normally distributed. Figure 6.15 shows that the conditional autoregressive term displays more spatial structure compared to the BYM model without the income and traffic-related

covariates. Figure 6.17 shows the spatial risk ($\zeta = v + \nu$). The spatial risk is interpreted as the residual relative risk for each municipality (compared to the nation-wide average) after income level, road density and traffic density are taken into account, identifying municipalities of increased claim risk.

The uncertainty associated with the posterior means can also be mapped and provides useful information (Richardson et al., 2004). The resulting map of their posterior probability of exceeding 1 is presented in Figure 6.18. Table C.1 (in the Appendix) lists the municipalities with the highest spatial risks and corresponding posterior probabilities for an municipality's spatial risk estimate exceeding 1.

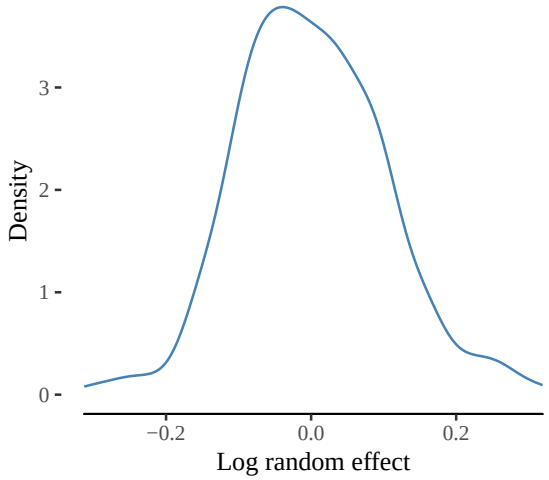


Figure 6.13: Density plot for random effects in covariate model.

In Figure 6.16 it is observed that the highest fitted values ($\theta = \alpha + v + \nu$) are found in the Brussel-Capital Region, Liège and Ghent. The fitted values are stored in `inla.cov$summary.fitted.values[,1]`. The fitted values multiplied by the expected number of claims (determined by the method of *indirect standardization* in Section 6.1) are the predicted number of claims per municipality. The highest fitted values can be seen as the municipalities where the conditional autoregressive term, and the income and traffic-related covariates have the highest influence on the expected number of claims per municipality (as determined in Section 6.1).

6.3 MODEL SELECTION

In this section we compare the different model specifications (Section 6.2) based on model fit criteria. The most commonly used measure of model fit based on the deviance for Bayesian models is the

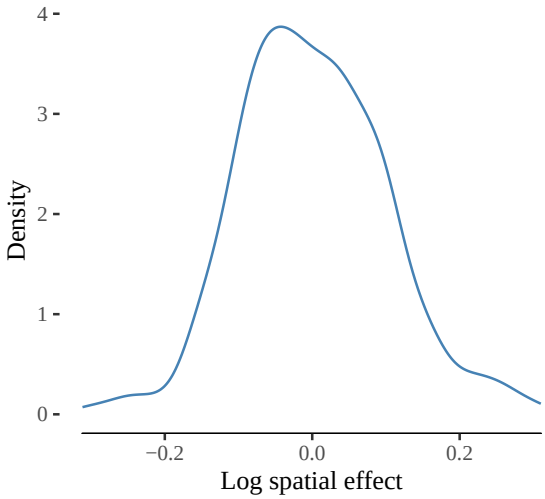


Figure 6.14: Density plot for conditional autoregressive term in covariate model.

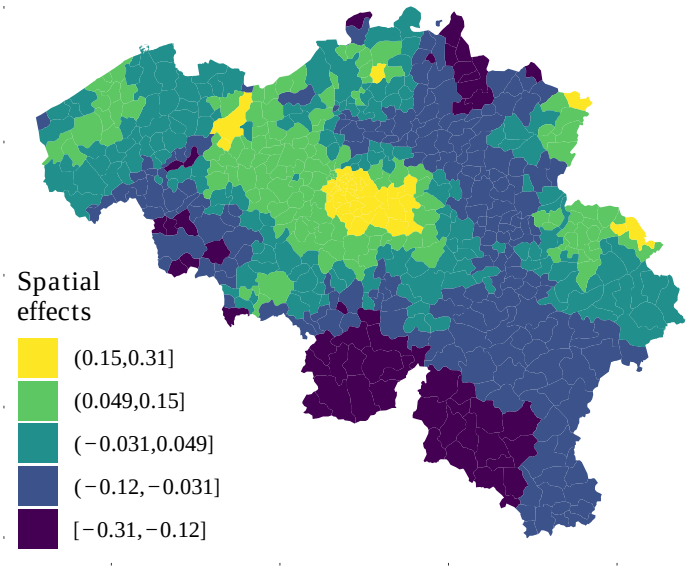


Figure 6.15: Spatially structured heterogeneity for covariate model.

deviance information criterion (DIC), proposed by Spiegelhalter et al. (2002). It is a generalization of the Akaike information criterion (AIC), developed especially for Bayesian model comparison and it is the sum of two components, one for quantifying the model fit and the other for evaluating the complexity of the model. The first com-

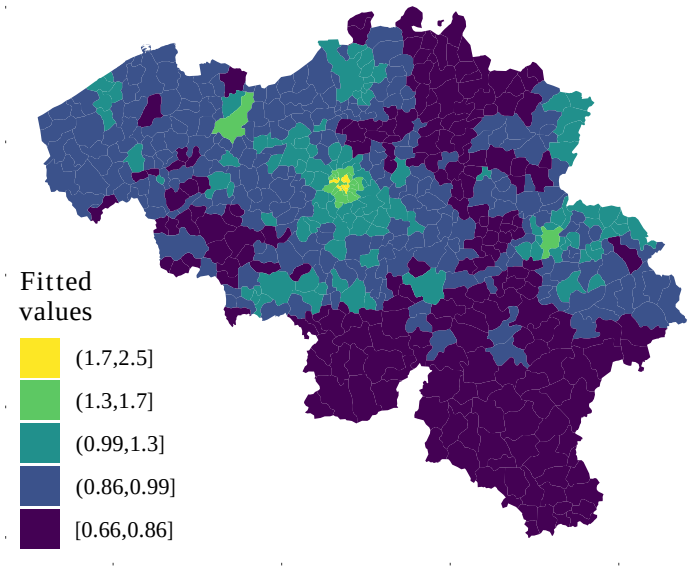


Figure 6.16: Fitted model values for covariate model.

ponent is measured through the posterior expectation of the deviance, while the model complexity is measured through the *effective number of parameters*. Analogously to the AIC, models with smaller DIC are better supported by the data.

However, the DIC may underpenalize complex models with many random effects (Plummer, 2008; Riebler and Held, 2009). Therefore, the cross-validated logarithmic score for model choice is used in this thesis (Gneiting and Raftery, 2007). Analogously to the DIC, a smaller value of the logarithmic score indicates a better prediction quality of the model.

The cross-validated logarithmic score is based on the *conditional predictive ordinate* (CPO) (Geisser, 1993; Pettit, 1990):

$$\text{CPO}_i = \pi\left(y_i^{obs} | y_{-i}\right),$$

where y_{-i} denotes the observations y with the i-th component omitted. The so-called *leave-one-out-cross validation* (LOOCV) is used in the determination of the cross-validated logarithmic score. This more sophisticated version of training/test sets is a more efficient use of the available data, that predicts a municipality value using all the data except the value for that municipality. A small CPO indicates an observation that is unlikely under the model fit without the observation

One way to measure the predictive ability of a model is to test it on a set of data not used in estimation. Data miners call this a "test set" and the data used for estimation is the "training set", originating the so-called cross-validation. However, there is often not enough data to allow some of it to be kept back for testing.

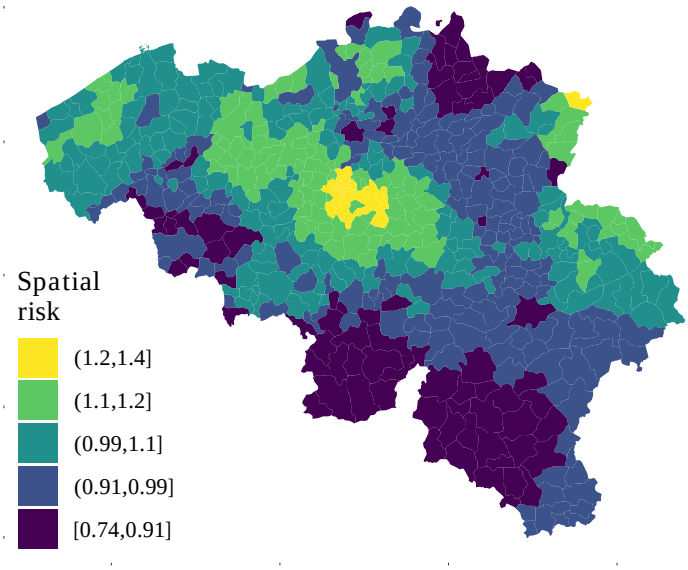


Figure 6.17: Spatially structured risk estimates $(\zeta = v + \nu)$ for covariate model.

in question. CPO measures are discussed among others by Congdon (2003), Gelfand, Dey, and Chang (1992), and Gilks, Richardson, and Spiegelhalter (1996).

CPO values are computed in `R-INLA` routinely without rerunning the model (Rue, Martino, and Chopin, 2009). As highlighted by Held, Schrödle, and Rue (2010), numerical problems may occur when the CPO indexes are computed. To this regards, `R-INLA` provides automatically a `failure` vector which contains a 0 or 1 value for each observation. In particular, a value equal to 1 indicates that for the corresponding observation the predictive measures are not reliable due to some problems in the calculation. In our case two failures were detected. Therefore, these two cases were recomputed manually by removing Y_i from the data set, fit the model again and predict Y_i. In the updated model failures were no longer detected, and consequently the assumptions were satisfied.

Table 6.1 shows measures of model fit for the model specifications given in Section 6.2. It is shown that based on the logarithmic score the final (or 'winning') model (further referred to as *final covariate model*) is the model with both unstructured (random effect) and structured (conditional autoregression) spatial terms, and covariates for median income, traffic density and road density. This means that the

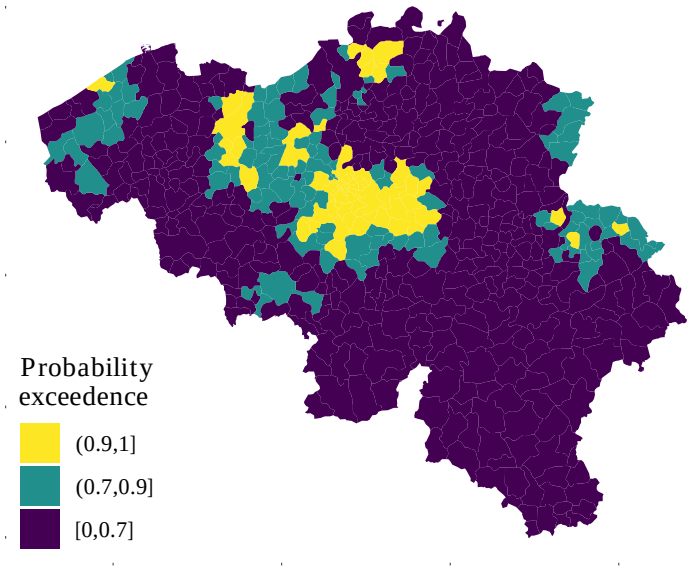


Figure 6.18: Posterior probability $\Pr(\zeta_i > 1 | y)$ for covariate model.

variables on median income, road density and traffic density explain part of the variability in the claim risk. This is also confirmed if we look at the proportion of variance explained by the spatially structured component, which goes down to 95.6% from 97.3%, registered for the BYM model without covariates (using Expression 6.10), suggesting that some of the spatial patterns in the risk of claims can be explained by the added covariates in the regression.

6.4 MODEL ASSESSMENT

In this section the *final covariate model* is assessed. First, the challenge in Bayesian statistics of specifying the prior is discussed. And second, the *probability integral transform* (PIT) is used for calibration checks.

6.4.0.1 *Prior Specification*

Even though Bayesian hierarchical models are a coherent framework for the development of spatially structured models, implementation of BYM models involves solving the challenge in Bayesian statistics of specifying the prior. The most common approach in the literature is assigning a so-called *noninformative* prior, particularly for applied researchers who often lack information on the parameters in

Model No.	Model Specification	DIC	No. of Effective Parameters	Log Score
i	random effects model	3577.93	258.88	3.1668
ii	BYM model	3476.94	184.06	2.9973
iii	(model ii) + median income	3459.12	155.06	2.9660
iv	(model ii) + traffic density	3467.70	171.88	2.9809
v	(model ii) + road density	3464.43	131.57	2.9690
vi	(model iii) + traffic density	3457.29	153.13	2.9655
vii	(model iii) + road density	3453.59	124.88	2.9543
viii	(model iv) + road density	3459.58	128.56	2.9647
ix	**(model vi) + road density**	**3452.63**	**124.40**	**2.9401**

Table 6.1: DIC, number of effective parameters, and logarithmic score for different model specifications. The final model is marked in bold face.

their studies and want the data to *speak for themselves* (Blangiardo and Cameletti, 2015). Therefore, we employ minimally informative priors specified on: (i) the log of the unstructured effect precision $\log(\tau_v) \sim \text{logGamma}(1, 0.0005)$, and (ii) the log of the structured effect precision $\log(\tau_v) \sim \text{logGamma}(1, 0.0005)$. This specification is also the default in the `"bym"` model specication in `R-INLA`. In the model fitting statement in `R-INLA` priors are set using `hyper=` in the call to the formula statement `f()`.

To assess the dependence of results on the hyperparameters a_j and b_j of variance components τ_j^2 we estimated the model with three different choices for logGamma(a_j, b_j),

- $a_j = 1, b_j = 0.0005$;
- $a_j = b_j = 0.001$;
- $a_j = b_j = 0.0001$.

For the present model, the differences between results were little, so that we present results only for our standard choice $a_j = 1$, $b_j = 0.0005$ (and skip a carefully conducted sensitivity analysis, as this is not in the scope of this thesis).

More detailed discussions on prior structures in spatial model exercises can be found in Lesaffre and Lawson (2012).

6.4.1 *Calibration Checks*

Dawid (1984) proposed the use of the *probability integral transform* (PIT) for calibration checks. Calibration refers to the statistical con-

sistency between the probabilistic forecasts and the observations, and is a joint property of the predictive distributions and the events or values that materialize. The PIT is simply the value that the predictive cumulation function attains at the value that materializes.

As described by Rue, Martino, and Chopin (2009) and Held, Schrödle, and Rue (2010), PIT values are computed in `R-INLA` routinely without rerunning the model. However, in the case of count data, the predictive distribution is discrete, and therefore, an adapted version should be used. Czado, Gneiting, and Held (2009, Section 2.1) adapted the methods proposed by Gneiting, Balabdaoui, and Raftery (2007) to the case of count data and proposed a *non-randomized* yet uniform version of the PIT histogram.

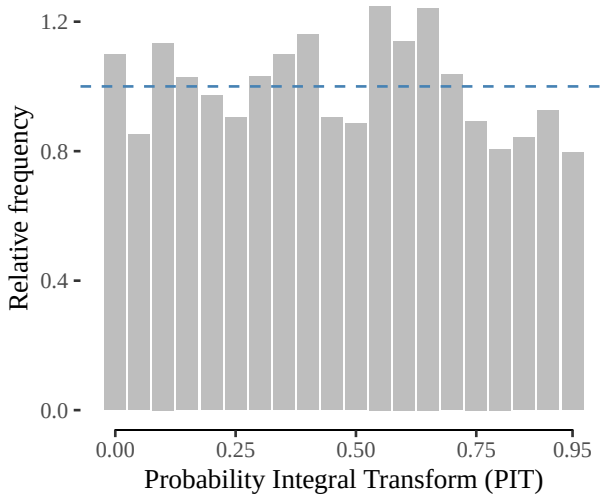


Figure 6.19: Histogram of the adapted PIT.

The histogram of this adapted PIT measure is depicted in Figure 6.19, and shows a distribution which tends to uniformity. Deviations from uniformity hint at reasons for forecast failures and model deficiencies. U-shaped histograms indicate underdispersed distributions, hump or inverse-U shaped histograms point at overdispersion, and skewed histograms occur when central tendencies are biased (Czado, Gneiting, and Held, 2009). For the present model, the PIT histogram does not show any sign on wrong calibration. Therefore, we conclude that the predictive distribution is coherent with our data.

Figure 6.20 shows the scatterplot of the posterior means for the predictive distributions versus the observed values, while the posterior p-value is represented in Figure 6.21. It is clear that on average the prediction is very close to the observed values. Looking at the posterior predictive p-value it seems that most of the municipalities have

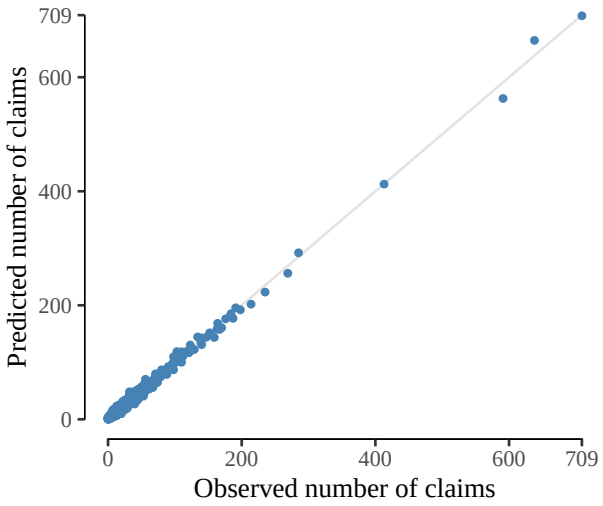


Figure 6.20: Observed and predicted number of claims per municipality.

a p-value close to one, suggesting that the model fits the data (very) well.

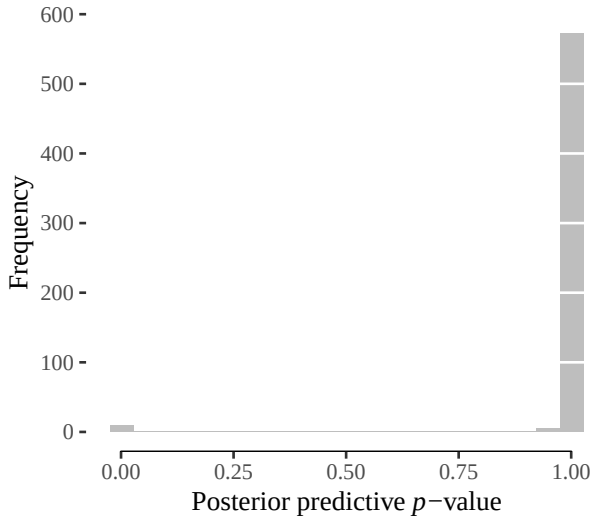


Figure 6.21: Histogram of the posterior predictive p-value.

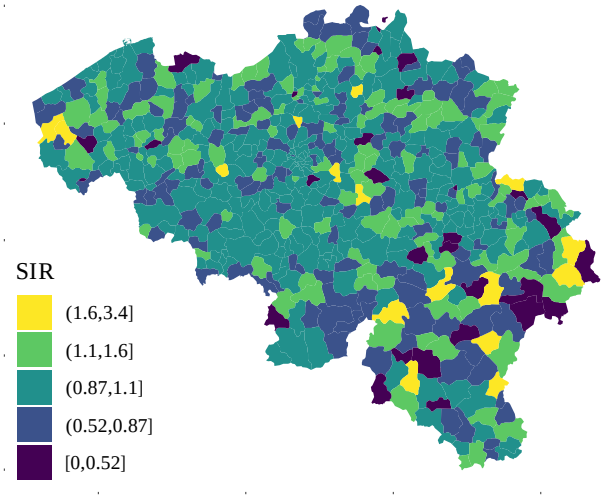


Figure 6.22: Standardized incidence ratios obtained from the INLA approach.

7

COMPARISON OF MODELING APPROACHES

This chapter concludes with a comparison of the fit of the claim frequency according to the GLM (Chapter 4), GAM (Chapter 5), and INLA approaches (Chapter 6).

Recall that the SIR is defined as the observed number of claims per municipality divided by the predicted number of claims per municipality.

Figure 7.1 shows the SIRs according to the GLM, GAM, and INLA modeling approaches. It is observed that the SIRs obtained from the INLA approach are closest to one. The spatial structure of the SIRs obtained from the INLA approach are given in Figure 6.22. Therefore, we conclude that our INLA modeling approach with a Besag-York-Mollié model with both unstructured (random effect) and structured (conditional autoregression) spatial terms, and explanatory covariates for median income, traffic density, and road density fits the expected average claim frequency for policyholders within municipalities better than the GLM and GAM modeling approaches. Moreover, the INLA results are obtained in seconds, whereas the computation time to perform GAM grows rapidly when the number of observations increases.

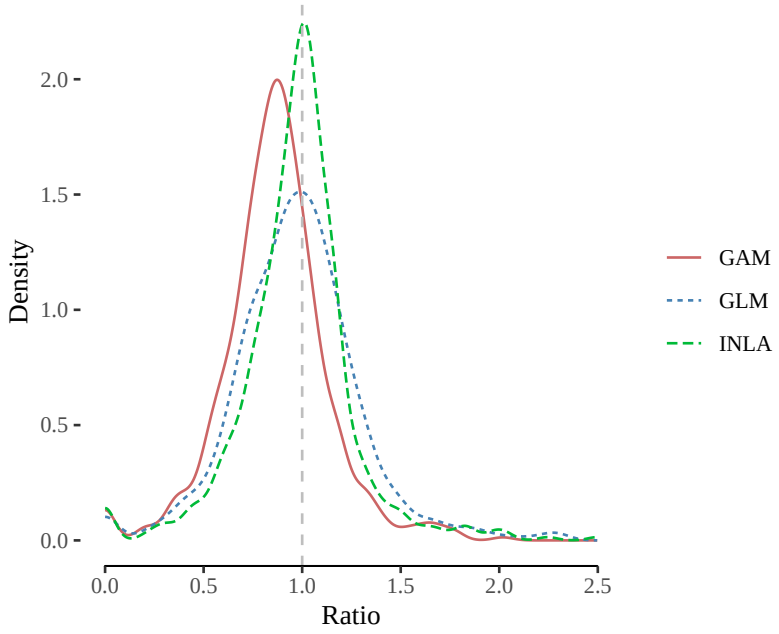


Figure 7.1: Density plots for the observed number of claims per municipality divided by the predicted number of claims per municipality according to the GAM, GLM, and INLA approaches.

8 FINAL CLAIM FREQUENCY MODEL

In this final chapter first the municipalities are divided in zones of municipalities with similar spatial claim risk (Section 8.1). These zones are then used to define a categorical covariate. These categorical covariate is used to create a final claim frequency model on the level of the policyholder (Section 8.2).

8.1 MUNICIPALITIES WITH SIMILAR SPATIAL CLAIM RISK

In Part ii it was shown that our INLA modeling approach with a Besag-York-Mollié model with both unstructured (random effect) and structured (conditional autoregression) spatial terms, and explanatory covariates for median income, traffic density, and road density fits the expected average claim frequency for policyholders within municipalities better than the GLM and GAM modeling approaches. This model specification is used to define municipalities with similar spatial claim risk.

Figure 8.1 shows the spatially structured risk estimates ($\zeta = v + \nu$) categorized into seven zones. Recall that the spatial risk is interpreted as the residual relative risk for each municipality (compared to the nation-wide average), identifying municipalities with a similar claim risk. These spatial risk zones are used to define a categorical variable *geoclass* in our final claim frequency model. The categorical variable *geoclass* takes the policyholder's municipality into account, where zone 1 groups municipalities with the lowest claim risk (dark green in Figure 8.1), whereas zone 7 combines the municipalities that face the highest claim risk (red in Figure 8.1). The categorical variable *geoclass* is defined as

$$\text{GEOCLASS} = \begin{cases} 1 = \text{if the policyholder's municipality belongs to zone 1} \\ 2 = \text{if the policyholder's municipality belongs to zone 2} \\ 3 = \text{if the policyholder's municipality belongs to zone 3} \\ 4 = \text{if the policyholder's municipality belongs to zone 4} \\ 5 = \text{if the policyholder's municipality belongs to zone 5} \\ 6 = \text{if the policyholder's municipality belongs to zone 6} \\ 7 = \text{if the policyholder's municipality belongs to zone 7} \end{cases}$$

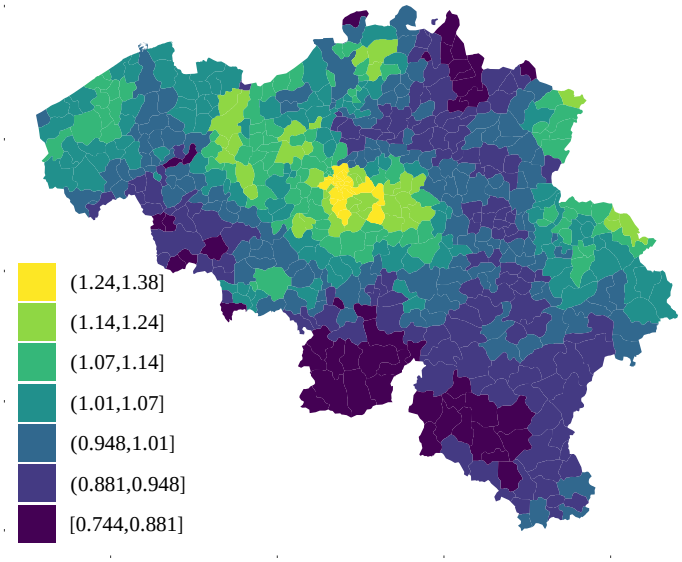


Figure 8.1: Spatially structured risk estimates ($\zeta = v + \nu$) for covariate model.

8.2 FINAL CLAIM FREQUENCY MODEL

In this section we create our final (*a priori*) claim frequency model on the level of the policyholder. We add the new categorical variable *geoclass* to the loglinear Poisson GLM defined earlier in Equation 6.2. The mean $\mathrm{E}(Y_i)$ of the independent Poisson distributed claim count Y_i is modeled as an exponential function of the linear predictor η_i, i.e. $\mathrm{E}(Y_i) = \exp(\eta_i)$, with the predictor equal to

$$\begin{aligned}\eta_i^{\text{freq}} =& \gamma_0 + \mathit{offset}_i + \sum_{k=1}^{6} \gamma_{1k} \left(\mathit{age\ policyholder}_{ik}\right) + \\ & \sum_{k=1}^{4} \gamma_{2k} \left(\mathit{age\ vehicle}_{ik}\right) + \sum_{k=1}^{8} \gamma_{3k} \left(\mathit{power\ vehicle}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_{4k} \left(\mathit{coverage}_{ik}\right) + \sum_{k=1}^{2} \gamma_{5k} \left(\mathit{fleet}_{ik}\right) + \\ & \sum_{k=1}^{2} \gamma_{6k} \left(\mathit{for}_{ik}\right) + \sum_{k=1}^{4} \gamma_{7k} \left(\mathit{fuel}_{ik}\right) + \\ & \sum_{k=1}^{3} \gamma_{8k} \left(\mathit{sex}_{ik}\right) + \sum_{k=1}^{7} \gamma_{9k} \left(\mathit{geoclass}_{ik}\right), \end{aligned} \tag{8.1}$$

where the offset_i is defined as the logarithm of the vehicle exposure (i.e. $\mathit{offset}_i = \log\left(\mathit{duration}_i / 365\right)$).

Table C.2 (in the Appendix) provides summary statistics of the parameter estimates for all explanatory variables in Equation 8.1. The coefficients given are relative to the standard class, which are indicated as reference group. For these classes, the coefficients are taken to be zero. Positive coefficients indicate a higher risk compared to the reference class, whereas, negative values indicate a lower risk than the reference class. All seven levels of the categorical variable *geoclass* turn out to be significant.

From the estimates in Table C.2, it is easy to determine how many claims on average a driver with the worst factor level produces, compared with someone having the best combination. In view of the sign of the coefficients, the best class is the one with a policyholder age between 61 and 79 years, vehicle age between 32 and 48 years, vehicle power between 0 and 27 kilowatt, TPL + limited material damage and theft, vehicle does not belong to a fleet, not a 4×4 vehicle, gasoline engine, male policyholder and the policyholder lives in a municipality with the lowest risk profile. The corresponding average number of claims equals 0.0016, that is, one claim each 347 years on average. The worst class is the one with a policyholder age between 17 and 31 years, vehicle age between 0 and 17 years, vehicle power between 174 and 250 kilowatt, TPL only, vehicle belongs to a fleet, 4×4 vehicle, diesel engine, female policyholder and the policyholder lives in a municipality with the highest risk profile. The corresponding average number of claims equals 1.2616, that is, one claim each 290 days.

Part III

CONCLUSION

9

CONCLUSIONS AND FURTHER RESEARCH

Our main conclusion is that Bayesian hierarchical spatial regression models using integrated nested Laplace approximations fit the claim frequency in a MTPL insurance portfolio on the municipality level better than the usual GAM framework. Moreover, INLA results are obtained in seconds, whereas the computation time to perform a GAM grows rapidly when the number of observations increases. The final Bayesian hierarchical spatial regression model includes variables for median household income, average vehicle speed, and traffic density, as well as a spatially unstructured random effect term, and a spatially structured conditional autoregression term.

Secondary conclusions are the associations of the economic and traffic-related measures with the vehicle claim risk on the municipality level. Every one thousand euro increase in the median income per municipality is associated with a 1.7% decrease in vehicle claim risk, every one standardized unit increase in traffic density per municipality is associated with a 0.7% increase in vehicle claim risk, and every one standardized unit increase in road density per municipality is associated with a 3.1% increase in vehicle claim risk.

Two of the statistical approaches used in this thesis come with caveats. First, exceedance probabilities, which have been proposed as a Bayesian approach to hotspot identification, can be sensitive to model specifications. And second, the proportion of variance explained by the spatially structured conditional autoregression term is not, strictly speaking, a variance partition coefficient because the structured and unstructured spatial terms may not be directly comparable. It is, though, an indication of the relative contribution of each of the spatial components (Blangiardo et al., 2013).

Approaching claim frequency data from a small-area spatial perspective poses challenges. The way the municipalities are defined, such as their shapes and boundaries, is arbitrary and may be to large and undifferentiated to the spatial distribution of the claim frequency in motor insurance. Therefore, municipalities may more properly be thought of as convenient 'bins' into which the number of claims are gathered, rather than as objects of interest in themselves. It may be interesting to apply our methods to claim frequency data indexed at finer geographical resolutions (DiMaggio, 2015).

While the model represents an acceptable trade-off between goodness of fit and complexity, efforts should include external validation

of the results. This could be statistical (e.g., comparing predictions based on the model to actual data from years beyond 1997) or could involve statistical associations with external sources of covariate data. Ideally, external validation would include the onsite evaluation of high-risk high-probability municipalities (DiMaggio, 2015).

THERE IS SCOPE FOR EXTENDING THIS LINE OF RESEARCH FURTHER. It may be interesting to extend the spatial model to include temporal characteristics for a space-time model of claim frequencies in municipalities in fixed-time periods. The Besag-York-Mollié model was extended by Bernardinelli et al. (1995) to include a linear term for space-time interaction, and by Knorr-Held and Raßer (2000) to include a non-parametric spatio-temporal time trend. The space-time interaction term is essentially a random-effect added to the linear model in the way the unstructured heterogeneity term and spatially-structured conditional autoregression (CAR) heterogeneity terms are added to spatial convolution models. In claim frequency models for motor insurance, space-time interaction may represent periods of heavy rainfall or snowfall, like short-term clusters.

BIBLIOGRAPHY

Antonio, Katrien and Emiliano A Valdez (2012). "Statistical concepts of a priori and a posteriori risk classification in insurance." In: *AStA Advances in Statistical Analysis* 96.2, pp. 187–224.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Belgian Federal Government (2013a). *Fiscale inkomens*. Retrieved from: http://statbel.fgov.be/nl/statistieken/cijfers/arbeid_leven/fisc/.

– (2013b). *Verkeer en vervoer*. Retrieved from: http://statbel.fgov.be/nl/modules/publications/statistiques/verkeer_vervoer/distances_routieres_parcourues_par_type_de_route_et_par_commune.jsp.

Bernardinelli, L, D Clayton, C Pascutto, C Montomoli, M Ghislandi, and M Songini (1995). "Bayesian analysis of space-time variation in disease risk." In: *Statistics in medicine* 14.21-22, pp. 2433–2443.

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics." In: *Annals of the institute of statistical mathematics* 43.1, pp. 1–20.

Bivand, Roger S (2015a). *Creating Neighbours*. URL: https://cran.r-project.org/web/packages/spdep/vignettes/nb.pdf.

– (2015b). *classInt: Choose Univariate Class Intervals*. R package version 0.1-23. URL: http://CRAN.R-project.org/package=classInt.

Bivand, Roger S, Edzer J Pebesma, and Virgilio Gómez-Rubio (2008). *Applied spatial data analysis with R*. Springer.

Bivand, Roger S and Gianfranco Piras (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." In: *Journal of Statistical Software* 63.18, 1–36.

Bivand, Roger S and Colin Rundel (2016). *rgeos: Interface to Geometry Engine - Open Source (GEOS)*. R package version 0.3-19. URL: https://CRAN.R-project.org/package=rgeos.

Blangiardo, Marta and Michela Cameletti (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blangiardo, Marta, Michela Cameletti, Gianluca Baio, and Håvard Rue (2013). "Spatial and spatio-temporal models with R-INLA." In: *Spatial and spatio-temporal epidemiology* 7, pp. 39–55.

Boskov, M and RJ Verrall (1994). "Premium rating by geographic area using spatial models." In: *Astin Bulletin* 24.01, pp. 131–143.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen (1984). *Classification and regression trees*. CRC press.

Brewer, Cynthia A and Linda Pickle (2002). "Evaluation of methods for classifying epidemiological data on choropleth maps in series." In: *Annals of the Association of American Geographers* 92.4, pp. 662–681.

Brouhns, Natacha, Michel Denuit, Bernard Masuy, Richard Verrall, et al. (2002). *Ratemaking by geographical area in the Boskov and Verrall model: a case study using belgian car insurance data*. Tech. rep. UCL.

Clijsters, Maxime (2015). "Dealing with continuous variables and geographical information in non-life insurance ratemaking." MA thesis. KU Leuven.

Congdon, Peter (2003). "The basis for, and advantages of, Bayesian model estimation via repeated sampling." In: *Applied Bayesian Modelling*, pp. 1–30.

Czado, Claudia, Tilmann Gneiting, and Leonhard Held (2009). "Predictive model assessment for count data." In: *Biometrics* 65.4, pp. 1254–1261.

Dawid, A Philip (1984). "Present position and potential developments: Some personal views: Statistical theory: The prequential approach." In: *Journal of the Royal Statistical Society. Series A (General)*, pp. 278–292.

Denuit, M., X. Marechal, S. Pitrebois, and J.F. Walhin (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley. ISBN: 9780470517413. URL: https://books.google.nl/books?id=xJ9CNm0i6mcC.

Denuit, Michel and Stefan Lang (2004). "Non-life rate-making with Bayesian GAMs." In: *Insurance: Mathematics and Economics* 35.3, pp. 627–647.

DiMaggio, Charles (2015). "Small-area spatiotemporal analysis of pedestrian and bicyclist injuries in New York City." In: *Epidemiology* 26.2, pp. 247–254.

Garnier, Simon (2018). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.5.1. URL: https://CRAN.R-project.org/package=viridis.

Geisser, Seymour (1993). *Predictive inference*. Vol. 55. CRC Press.

Gelfand, Alan E, Dipak K Dey, and Hong Chang (1992). *Model determination using predictive distributions with implementation via sampling-based methods*. Tech. rep. DTIC Document.

Gilks, Walter R, Sylvia Richardson, and David J Spiegelhalter (1996). "Introducing Markov chain Monte Carlo." In: *Markov chain Monte Carlo in practice* 1, p. 19.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E Raftery (2007). "Probabilistic forecasts, calibration and sharpness." In: *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology) 69.2, pp. 243–268.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation." In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Haberman, Steven and Arthur E Renshaw (1996). "Generalized linear models and actuarial science." In: *The Statistician*, pp. 407–436.

Haining, Robert P (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Hastie, Trevor J and Robert J Tibshirani (1990). *Generalized additive models*. Vol. 43. CRC Press.

Held, Leonhard, Birgit Schrödle, and Håvard Rue (2010). "Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA." In: *Statistical modelling and regression structures*. Springer, pp. 91–110.

Klein, Nadja, Michel Denuit, Stefan Lang, and Thomas Kneib (2014). "Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape." In: *Insurance: Mathematics and Economics* 55, pp. 225 –249.

Knorr-Held, Leonhard and Günter Raßer (2000). "Bayesian detection of clusters and discontinuities in disease maps." In: *Biometrics* 56.1, pp. 13–21.

Lesaffre, Emmanuel and Andrew B Lawson (2012). *Bayesian biostatistics*. John Wiley & Sons.

Mausner, J.S., S. Kramer, and A.K. Bahn (1985). *Epidemiology: An Introductory Text*. Saunders. ISBN: 9780721661810. URL: https://books.google.nl/books?id=qWBrAAAAMAAJ.

McCullagh, Peter and John A Nelder (1989). *Generalized linear models*. Vol. 37. CRC press.

Pettit, LI (1990). "The conditional predictive ordinate for the normal distribution." In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 175–184.

Plummer, Martyn (2008). "Penalized loss functions for Bayesian model comparison." In: *Biostatistics* 9.3, pp. 523–539.

Richardson, Sylvia, Andrew Thomson, Nicky Best, and Paul Elliott (2004). "Interpreting posterior relative risk estimates in disease-mapping studies." In: *Environmental Health Perspectives*, pp. 1016–1025.

Riebler, Andrea and Leonhard Held (2009). "The analysis of heterogeneous time trends in multivariate age–period–cohort models." In: *Biostatistics*, kxp037.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using in-

tegrated nested Laplace approximations." In: *Journal of the Royal Statistical Society: Series B (statistical methodology)* 71.2, pp. 319–392.

Schabenberger, Oliver and Carol A Gotway (2004). *Statistical methods for spatial data analysis*. CRC press.

Spiegelhalter, David J, Nicola G Best, Bradley P Carlin, and Angelika Van der Linde (2002). "Bayesian measures of model complexity and fit." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639.

Taylor, Greg C (2001). "Geographic premium rating by Whittaker spatial smoothing." In: *Astin Bulletin* 31.01, pp. 147–160.

Therneau, Terry, Beth Atkinson, and Brian Ripley (2015). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10. URL: https://CRAN.R-project.org/package=rpart.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Waller, Lance A and Carol A Gotway (2004). *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons.

Wendelberger, James G (1981). *The Computation of Laplacian Smoothing Splines with Examples.* Tech. rep. DTIC Document.

Wood, S. N. (2000). "Modelling and smoothing parameter estimation with multiple quadratic penalties." In: *Journal of the Royal Statistical Society (B)* 62.2, pp. 413–428.

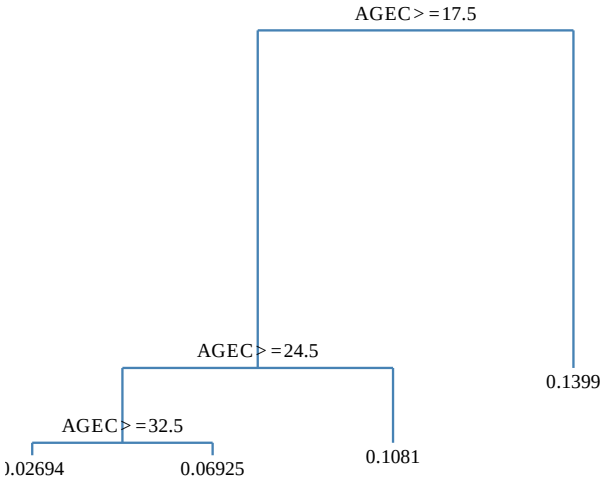
– (2006). *Generalized additive models: An introduction with R*. CRC press.

Part IV

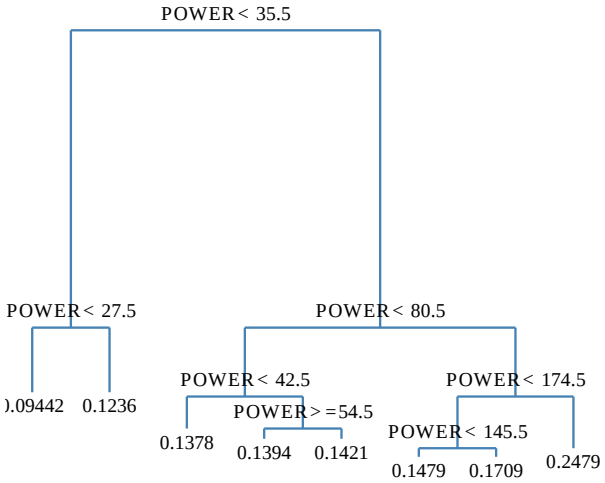
APPENDIX

A

DENDROGRAMS



(a) Age of the vehicle with a complexity factor of $c_p = 0$.

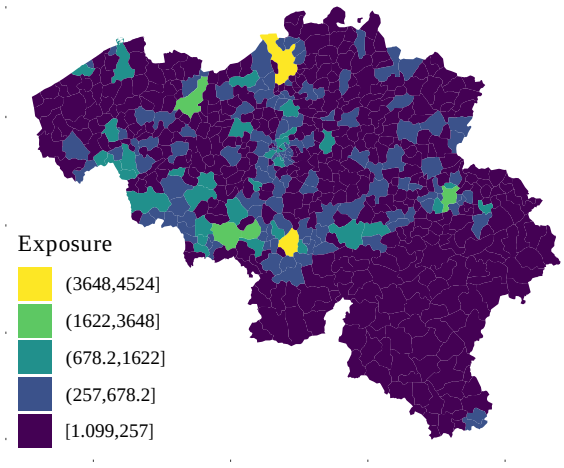


(b) Power of the vehicle with a complexity factor of $c_p = 0.005$.

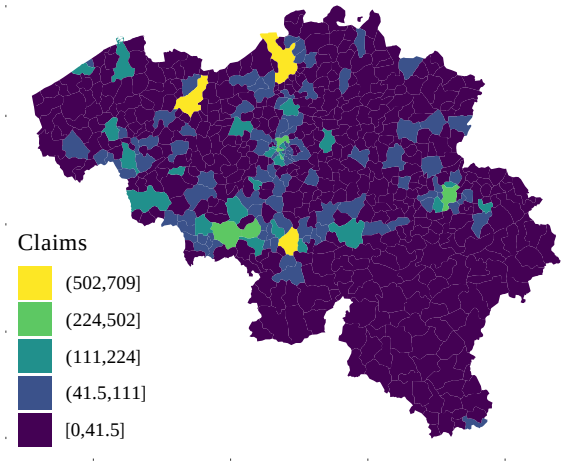
Figure A.1: Dendograms for the age of the vehicle in years (above), and for the power of the vehicle in kilowatt (below).

B

EXPOSURE AND NUMBER OF CLAIMS ON THE MUNICIPALITY LEVEL



(a) Exposure per municipality.



(b) Claim frequency per municipality.

Figure B.1: Total exposure (left) and total number of claims (right) on the municipality level.

C

TABLES

C.1 MUNICIPALITIES WITH HIGHEST SPATIAL RISKS

Municipality	Region	Spatial Risk	Y^{obs}	$\widehat{Y}^{inla}$	$\widehat{Y}^{stdzn}$
Etterbeek	Brussels-Capital	1.38	110	100.10	49.80
Ukkel	Brussels-Capital	1.38	159	143.70	96.00
Sint-Joost-ten-Node	Brussels-Capital	1.36	42	39.20	15.70
Elsene	Brussels-Capital	1.33	139	135.70	73.10
Linkebeek	Brussels periphery	1.32	15	12.90	8.40
Anderlecht	Brussels-Capital	1.31	214	202.00	125.40
Watermaal-Bosvoorde	Brussels-Capital	1.31	57	52.00	37.70
Sint-Gillis	Brussels-Capital	1.31	81	79.20	39.10
Oudergem	Brussels-Capital	1.30	56	54.80	38.30
Huldenberg	Brussels periphery	1.30	20	9.80	9.40
Sint-Lambrechts-Woluwe	Brussels-Capital	1.29	108	104.30	67.20
Schaarbeek	Brussels-Capital	1.28	184	185.30	102.90
Tervuren	Brussels periphery	1.27	44	33.40	31.70
Brussels	Brussels-Capital	1.26	235	223.20	135.20
Sint-Jans-Molenbeek	Brussels-Capital	1.26	167	157.60	87.00
Sint-Genesius-Rode	Brussels periphery	1.26	33	31.50	27.90
Sint-Pieters-Woluwe	Brussels-Capital	1.25	64	66.80	47.30
Wavre	Brussels periphery	1.25	72	65.10	53.50
Waterloo	Brussels periphery	1.25	75	70.50	57.00
Evere	Brussels-Capital	1.24	76	76.60	50.70

Table C.1: The municipalities in the highest spatial risk zone. Column 3 denotes the spatial risk (ζ). Column 4 gives the observed number of claims. Column 5 is the predicted number of claims according to the final covariate model. Column 6 denotes the predicted number of claims based on the indirect standardization method.

C.2 GLM OUTPUT FOR FINAL CLAIM FREQUENCY MODEL

	Estimate	Std. Error	z-value	Pr($>\|z\|$)
Intercept	-2.5662	0.1078	-23.81	<.001
Age Policyholder				
$17 <$ years ≤ 31	ref. group			
$31 <$ years ≤ 38	-0.3273	0.0223	-14.70	<.001
$38 <$ years ≤ 54	-0.4204	0.0186	-22.59	<.001
$54 <$ years ≤ 61	-0.5659	0.0275	-20.55	<.001
$61 <$ years ≤ 79	-0.6827	0.0242	-28.26	<.001
$79 <$ years ≤ 95	-0.4277	0.0735	-5.82	<.001
Age Vehicle				
$0 <$ years ≤ 17	ref. group			
$17 <$ years ≤ 24	-0.2940	0.0846	-3.48	<.001
$24 <$ years ≤ 32	-0.4577	0.2361	-1.94	0.053
$32 <$ years ≤ 48	-2.4121	0.9999	-2.41	0.016
Power Vehicle				
$0 <$ kilowatt ≤ 27	ref. group			
$27 <$ kilowatt ≤ 35	0.1700	0.0909	1.87	0.061
$35 <$ kilowatt ≤ 42	0.2751	0.0884	3.11	0.002
$42 <$ kilowatt ≤ 54	0.3238	0.0879	3.68	<.001
$54 <$ kilowatt ≤ 80	0.3135	0.0876	3.58	<.001
$80 <$ kilowatt ≤ 145	0.3977	0.0893	4.45	<.001
$145 <$ kilowatt ≤ 174	0.5211	0.1787	2.92	0.004
$174 <$ kilowatt ≤ 250	1.1007	0.2259	4.87	<.001
Coverage				
TPL only	ref. group			
TPL + limited material damage and theft	-0.1393	0.0166	-8.39	<.001
TPL + comprehensive damage	-0.1434	0.0219	-6.54	<.001
Vehicle Belongs to a Fleet				
No	ref. group			
Yes	0.1635	0.0431	3.79	<.001
4×4 vehicle				
No	ref. group			
Yes	0.2677	0.1086	2.47	0.014
Fuel				
Gasoline	ref. group			
Diesel	0.1700	0.0159	10.68	<.001
LPG	0.0542	0.1694	0.32	0.749
Other	-0.5475	1.0001	-0.55	0.584
Gender				
Female	ref. group			
Male	-0.0453	0.0163	-2.78	0.005
Company	0.1438	0.2367	0.61	0.543
Geoclass				
Geoclass 1	ref. group			
Geoclass 2	0.2920	0.0495	5.90	<.001
Geoclass 3	0.5125	0.0478	10.72	<.001
Geoclass 4	0.5558	0.0487	11.40	<.001
Geoclass 5	0.6092	0.0498	12.23	<.001
Geoclass 6	0.7179	0.0510	14.07	<.001
Geoclass 7	1.0967	0.0514	21.32	<.001

Table C.2: Parameter estimates for the final claim frequency model.