TÉCNICO LISBOA

Bayesian Generalized Additive Models for Car Insurance Data

João Pedro Rodrigues Góis

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisor: Prof. Giovani Loiola da Silva

Examination Committee

Chairperson: Prof. António Manuel Pacheco Pires
Supervisor: Prof. Giovani Loiola da Silva
Members of the Committee: Prof. Carlos Daniel Mimoso Paulino
Prof. Marília Cristina de Sousa Antunes

October, 2019

Acknowledgments

I want to express my gratitude to:

My parents and sister for their encouragement and support, without whom this project would not be possible.

My supervisor, Giovani Silva, for his availability and guidance during the period of elaboration of this thesis, as well as Prof. Carlos Paulino and Prof. Marília Antunes for their valuable suggestions to improve this work.

My friends for their companionship.

Abstract

Motor vehicles (cars) are increasingly becoming more susceptible to unexpected events. So, it is convenient for vehicle owners to purchase an insurance policy, which determines the claims that are legally covered by a car insurance. This work aims to analyse how claim frequency is influenced by policy risk factors. Hence, risk profiles can be defined by the company in order to apply adequate insurance premiums for policyholders and reduce monetary losses. The methodology of choice is based on Generalized Additive Models for Location, Scale and Shape using Bayesian statistics via MCMC methods.

This study was motivated by a data set comprising policies registered in Portugal mainland from 2011 to 2013. The data set includes some particularities, namely missing values in the covariates and an excess of zeros in the response variable. As such, the imputation of missing values and zero-inflated response distributions were explored throughout the analysis. Model selection suggested both better fitting for zero-inflated models and that imputation is a valid option to reduce model dispersion. Finally, the selected models were employed for estimation of actuarial quantities.

Keywords

Structured Additive Regression; Imputation; Zero Inflation; MCMC Methods; Bayesian Statistics.

Resumo

Os veículos automóveis (carros) estão cada vez mais suscetíveis a imprevistos. Então, é do interesse do proprietário do veículo a aquisição de uma apólice de seguro, que determina o tipo de participações que serão cobertas pela seguradora automóvel. Este estudo tem como objetivo analisar de que forma o número de participações de seguro automóvel é influenciado por fatores de risco presentes nas apólices. Assim, podem-se definir perfis de risco por parte da seguradora de forma a aplicar prémios de seguro adequados e reduzir prejuízos. Para tal, recorre-se a metodologia estatística recente baseada em Modelos Aditivos Generalizados para Localização, Escala e Forma, usando estatística Bayesiana através de métodos MCMC.

Este trabalho foi motivado por um conjunto de dados com informação relativa a apólices registadas em Portugal continental no período de 2011 a 2013. O conjunto de dados possui algumas especificadades, tais como a existência de valores omissos nas covariáveis e excesso de zeros na variável resposta. Nesse sentido, a substituição de valores omissos e o uso de distribuições adequadas para inflação de zeros foram explorados ao longo da análise. A etapa de seleção de modelos sugeriu quer um melhor ajustamento ao usar distribuições que acomodam excesso de zeros quer que a imputação é uma opção válida para reduzir a dispersão dos modelos. Por fim, seguiu-se a estimação de quantidades de interesse nas aplicações actuariais com base nos modelos selecionados.

Palavras Chave

Regressão Aditiva Estruturada; Imputação; Inflação de Zeros; Métodos MCMC; Estatística Bayesiana.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Throughout the years, the statistical methods for car insurance data have become more accurate in evaluating policy's risk factors. The adequacy of statistical methods is an important issue in Actuarial Science (Denuit et al, 2007). The present chapter is divided in three parts. Firstly, Section 1.1 explains the problem that motivates this work. Section 1.2 develops the ideas and methods pointed out by several authors for the problem at hand, i.e. the state-of-the-art. Finally, Section 1.3 describes the structure of the dissertation.

1.1 Motivation

As a motor vehicle (car) is susceptible to accidents or robbery, the vehicle owner must purchase an insurance policy. The related contract determines the policyholder's claims that are legally covered by the car insurance. Once the insurance premium is paid, the policy is active. Its duration defines the risk period, in which the insurance company is vulnerable to monetary losses due to coverage.

In order to ensure their sustainability, companies need to use advanced statistical tools to analyse insurance risk. For instance, regression models can be fitted to explain how claim frequency or cost vary with age of the policyholder, driving license time, place of residence, age of the vehicle, among others. These models can be used to predict actuarial quantities and characterize the risk profile of policyholders. Then, companies are more likely to apply appropriate insurance premiums.

The data set, which will be analysed herein, belongs to a insurance company, that provided anonymous information regarding policyholders and vehicles, including their geographical location. The objective is to propose precise statistical models for claim frequency with the use of recent methodology on car insurance applications, and account for different risk profiles in this context.

1.2 State-of-the-Art

For many years, linear regression models have been widely used in statistical analysis of car insurance data. Despite its simplicity of interpretation and calculation, these models are not adequate for a variety of applications. For instance, when the responses do not depend linearly on the covariates. Furthermore, the assumption of normally distributed responses with constant variance may not be realistic. Still, car insurance data analyses were initially dominated by simple statistical models (Denuit and Lang, 2004).

The methodology in regression analysis has constantly improved since the second half of the last century. Nelder and Wedderburn [1972] proposed Generalized Linear Models (GLM) as an extension of ordinary linear models. GLM are able to accommodate response distributions that belong to the exponential family, including Gaussian distribution, and introduce some degree of nonlinearity in the model structure through a link function between the response variable mean and the linear predictor. These models were initially applied to actuarial applications in order to improve *a priori* risk classification. However, GLM still have the drawback of not being able to deal with nonlinear effects of continuous covariates. In car insurance context, claim frequency modelling usually requires nonlinear effects of the continuous covariates (Denuit and Lang, 2004; Klein et al., 2014).

Hastie and Tibshirani [1990] proposed Generalized Additive Models (GAM), which can be specified in terms of smooth functions of covariates, so-called nonparametric parts of the predictor. These nonparametric terms allow the data to present the most appropriate functional form of the covariate effects and should be used when their nonlinear dependencies are known (Ruppert and Matteson, 2015). Smooth functions are often represented by spline basis functions (see e.g. Silva and Dean, 2004, for construction of B-spline). Regression splines have appealing smoothing properties and enable more flexible models than other nonlinear techniques, such as polynomial regression. Eilers and Marx [1996] suggested the use of P-splines to overcome some difficulties of regression splines. Since GAM inherit the response distributions of GLM, in particular the Poisson distribution for count data, Poisson GAM became a primary choice to model claim frequency in car insurance scenarios (Denuit and Lang, 2004). However, for analysing risk variation by geographical area, it is still needed to include spatial information into the additive predictor, and therefore an extension to GAM should be considered.

Kammann and Wand [2003] proposed Generalized Geoadditive Models (GGAM), which are an extension to GAM, accommodating a nonlinear (correlated) spatial effect on the predictor. A further extension of these models is the Generalized Geoadditive Mixed Models (GGAMM), which also include a nonlinear (uncorrelated) spatial effect into the predictor. The latter are of particular interest to handle for model heterogeneity that cannot be taken into account only from the observed covariates. Farhmeir et al. [2004] proposed a unified representation of different model terms, known as Structured Additive Regression (STAR), which includes parametric linear effects, smooth nonlinear effects of continuous

covariates, spatial effects and also interaction terms based on varying coefficient framework.

GAM were originally developed according to a classical (frequentist) approach. Consequently, these models become computationally expensive should the number of parameters and observations be sufficiently large. An appealing alternative to classical GAM employs Bayesian inference based on Markov Chain Monte Carlo (MCMC) techniques, using Bayesian P-splines for modelling nonlinear effects of continuous covariates and Markov Random Field (MRF) priors for correlated spatial effects. MCMC simulation techniques are particularly useful to extend model formulations to more complex settings, providing easy prediction and inferential information for functions of the parameters, including credible intervals. Brezger and Lang [2006] suggested the use of iteratively weighted least squares approximations for full conditional distributions of nonparametric coefficients in MCMC simulation. For some car insurance applications, see e.g. Denuit et al. [2007] and Klein et al. [2014].

In count data, the response variable often presents an excess of zeros, where zeros can be considered either true or false. In car insurance applications, policyholders that make a small number of claims are awarded premium discounts. Sometimes policyholders do not make small claims to earn those discounts, thus leading to false zeros (Denuit et al., 2007). Since false zeros are not accounted by the Poisson distribution, an alternative is to use the zero-inflated Poisson (ZIP) distribution. It consists of a mixture of two distributions: a degenerated distribution to account for false zeros and a Poisson distribution for the count data part, where true zeros are included. Another common problem is overdispersion, which is reflected by the variance being much larger than the mean. Klein et al. [2015] suggested the use of a zero-inflated negative binomial (ZINB) distribution instead, since it has a better performance than Poisson and ZIP distributions. In particular, zero-inflated models belong to the framework of generalized additive models for location, scale and shape (GAMLSS), which can be seen as a special case of STAR where all occurring parameters are related to regression predictors and may depend on a complex covariate structure. Bayesian inference via MCMC techniques was extended to GAMLSS by Klein et al. [2014].

In real application problems, data sets often contain missing values. In particular, the current car insurance data set comprises thousands of observations. Since removing observations with missing values can bring bias into the model, an alternative to handle this problem is to perform imputation. van Buuren and Groothuis-Oudshoorn [2011] proposed multiple imputation via chained equations (MICE) for either missing at random (MAR) and missing not at random (MNAR) patterns in the data. The `R` package `mice` (van Buuren, 2011) was used, as it provides functionality to perform MICE and diagnosis on the imputed values.

Model selection can be performed on the basis of Klein et al. [2014] and Klein et al. [2015], through `R` packages `bamlss` (Umlauf et al., 2017) and `gamlss` (Stasinopoulos and Rigby, 2019). These packages implement the conceptional frameworks to perform Bayesian inference via MCMC simulation techniques.

A number of functionalities for inference and visualization of functions of parameters are available, as well as tools for model diagnosis and prediction. For each of the Poisson, ZIP and ZINB distributions, a model is selected based on the Deviance Information Criterion (DIC) and the model complexity measure pD. Quantile residuals and an estimate of model dispersion are then evaluated to choose the distribution that provides a better fitting.

1.3 Structure of the Dissertation

This dissertation is structured as follows. Chapter 1 summarily introduces the issues approached in the dissertation. Chapter 2 provides an introduction to Bayesian inference concepts and computational methods that will be essential for the data analysis throughout the text. Chapter 3 describes the methodology used herein for the problem at hand. Chapter 4 presents the results from applied methodology for two data scenarios: only complete data and data completed with imputation, and prediction of actuarial quantities. Finally, Chapter 5 states the concluding remarks.

Chapter 2

Bayesian Statistics and Computation

According to O'Hagan [1994], "The fundamental problem towards which the study of statistics is addressed is that of inference. Some data are observed and we wish to make some statements, inferences, about one or more unknown features of the physical system which gave rise to these data". In other words, after producing a descriptive analysis based on past observations, the objective of a statistician is to make inferences or predictions about new observations of the same nature, as stated by Robert [1994].

The historical development of statistical inference comprises, among others, Bayesian inference and Classical (frequentist) inference. Following their own principles, statistical investigation can be planned, performed, and meaningfully evaluated (Berger, 1984). The foundations of Bayesian inference are more easily understood when compared to classical inference (vide e.g. Paulino et al., 2018).

Bayesian Statistics was developed from Bayes Theorem, introduced in the literature by Bayes [1763]. In Bayesian models, the model parameter vector $\theta \in \Theta$ is a random non-observable quantity. Being unknown according to Bayesian philosophy, it must be quantified in terms of probability. Bayesian inference is constructed depending on prior knowledge of θ, which corresponds to past or exterior knowledge to the experience. The prior information is used jointly with data to make posterior inference on the model parameter.

Classical approach was introduced by Laplace [1774] at the same period, and later developed by Neyman, Pearson and Fisher. Classical procedures are based on the principle of repeated sampling, which states that statistical methods should be characterized by their behaviour in an undefined number of hypothetical repetitions under the same conditions. In Classical inference, parameter $\theta \in \Theta$ is an unknown vector, which is assumed to be fixed. Prior knowledge of θ is mainly criticized by classical statisticians due to lack of objectivity.

On the other hand, Bayesian statisticians argue that past knowledge about the parameter should not be neglected. However, the criticism towards Bayesian inference, as well as the difficulty of its theory

and implementation, made Bayesian inference obsolete for many decades (Paulino et al., 2018).

Recently, relevant improvements in computational statistics have been made, leading to the development of new techniques and algorithms for sampling from a probability distribution. Therefore, many statistical models have been reformulated under a Bayesian perspective, with increasing usage for a variety of complex problems.

In this chapter, the fundamental concepts of Bayesian statistics will be presented at first. In further steps, the representation of prior information and hierarchical Bayesian models will be addressed. At last, recent computational techniques for Bayesian inference will be revised.

2.1 Fundamental Concepts

In probability problems, the starting point is to define a probability space (Ω, $\mathcal{A}$, P), where

- Ω is a non-empty sample space with elements $\omega \in \Omega$, and subset A $\subset \Omega$, known as event.
- $\mathcal{A}$ is a family (algebra or σ-algebra) of probability events.
- P is a probability measure defined for all probability events, where P(A) denotes the probability of event A $\subset \Omega$, A $\in \mathcal{A}$.

Consider a finite set of events $\{A_i\}_{i=1}^m$ of Ω such that (i) $P(A_i) > 0$; (ii) $A_i \cap A_j = \varnothing$, $\forall i \neq j$; and (iii) $\cup_i A_i = \Omega$, i.e. a partition of Ω. Let B be another event of Ω. It is easy to show that $B = \cup_i (B \cap A_i)$. Consequently, $P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$.

Theorem 2.1.1 (Bayes Theorem) Consider $A_1, \ldots, A_m$ events that form a partition of the sample space Ω. If B is an event defined in Ω, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^m P(B|A_j)P(A_j)}, \quad i = 1, \ldots, m \tag{2.1}$$

Bayes theorem proposes learning with experience. The initial beliefs on the probabilities of events are modified given that another event (or set of events) has occurred. In Equation 2.1:

- $P(A_i)$, $i = 1, \ldots, m$, are called prior probabilities of A_i or their probabilities before the experiment.
- $P(A_i|B)$, $i = 1, \ldots, m$, are called posterior probabilities of A_i or their probabilities given the results of the experiment.

2.2 Bayesian Inference

Under a Bayesian approach, the model parameter θ is considered a random variable. Bayesian models can be roughly seen as an extension of classical models, whose difference lies on the definition of the parameter (Paulino et al., 2018). Suppose that x is the observed value of a random variable X defined in a sample space Ω. Let $f(x|\theta)$ be any element from a family of density functions $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$, and the prior distribution of θ given by $h(\theta)$. The Bayes Theorem for densities is given by

$$h(\theta|x) = \frac{f(x|\theta)h(\theta)}{\int_\Theta f(x|\theta)h(\theta)d\theta}, \quad \theta \in \Theta \tag{2.2}$$

where $h(\theta|x)$ is the posterior distribution of θ given x. If x is now an observed sample $(X_1 = x_1, \ldots, X_n = x_n)$ of X, (2.2) can be rewritten as

$$h(\theta|x_1, \ldots, x_n) = \frac{\prod_{i=1}^n f(x_i|\theta)h(\theta)}{\int_\Theta \prod_{i=1}^n f(x_i|\theta)h(\theta)d\theta}, \quad \theta \in \Theta \tag{2.3}$$

where $h(\theta|x_1, \ldots, x_n)$ is the posterior distribution of θ given an observed sample $(x_1, \ldots, x_n)$. Sometimes calculating the integrals in the denominator of (2.2) or (2.3) is not straightforward, especially if θ is a vector, which requires the use of numerical methods. Since those denominators are not functions of θ, it is common to represent the posterior distribution in terms of its kernel, that is

$$h(\theta|x) \propto f(x|\theta)h(\theta) \tag{2.4}$$

$$h(\theta|x_1, \ldots, x_n) \propto \prod_{i=1}^n f(x_i|\theta)h(\theta) \tag{2.5}$$

In many applications, complex model formulations result in intractable (joint) posterior distributions. For many years, that was one of the main problems in Bayesian inference. However, the development of computational methods for sampling from a probability distribution (see Section 2.5) was extremely important to overcome these limitations.

2.3 Representation of Prior Information

One of the major assets regarding Bayesian inference is the employment of information prior to the experiment. According to classical statisticians, prior distributions bring subjectivity into inference. Actually, prior specification should be handled carefully. A wrongly specified prior may have a negative impact on inference.

Prior information can be represented in a variety of ways. In particular, conjugate priors and non-informative priors will be of interest in this work. The following subsections are dedicated to these representations.

2.3.1 Conjugate prior distributions

Prior distributions are called conjugate for a sampling model (or likelihood) if the posterior distribution has the same family as the prior distribution. By using conjugate priors, posterior distributions can be obtained by updating prior distribution parameters, which is convenient in a variety of applications. However, this type of prior implies very specific prior knowledge, which is something to be considered. Table 2.1 shows some common conjugate priors in Bayesian statistics.

Table 2.1: Conjugate priors

Likelihood	Prior Distribution	Posterior Distribution
Normal (σ^2 is known)	Normal	Normal
Normal (μ is known)	Inverse Gamma	Inverse Gamma
Poisson	Gamma	Gamma
Exponential	Gamma	Gamma
Uniform	Pareto	Pareto
Binomial (n is known)	Beta	Beta

2.3.2 Non-informative prior distributions

Non-informative prior distributions are used when the parameter information is little in comparison to sample information. In that case, the influence of prior information on the posterior is minimised and consequently Bayesian inferential results based on non-informative priors can be used as reference for comparison with the results of classical inference (Paulino et al., 2018).

An argument to generate non-informative priors is the Principle of Insufficient Reason (Syversveen, 1998), which determines that any element belonging to a parameter space has the same probability. Let Θ denote a parametric space. If:

- Θ is a finite set, say $\Theta = \{\theta_1, \ldots, \theta_k\}$, the non-informative prior is a uniform discrete distribution, that is, $h(\theta) = \frac{1}{k}, \quad \theta \in \Theta$.

- Θ is a countably infinite set, using the uniform distribution results in an improper prior distribution.
- Θ is an uncountably infinite set, the prior is a continuous uniform distribution. For unbounded Θ, an improper prior is obtained.

The main objection to uniform priors is that these are not invariant to injective transformations. Alternatively, Jeffreys [1939] proposed priors that are based on Fisher Information measure (Jeffreys' priors). However, the priors generated according to this rule may be improper.

The validity of improper priors is questioned by many authors. Nevertheless, it is acceptable to use these priors when prior knowledge is weak, leading to robust posterior inferences (Paulino et al., 2018). An alternative to improper priors is by means of proper diffuse priors, which are meant to represent vague prior knowledge. For instance, a common choice for location parameter prior is to use the normal distribution with zero mean and large variance.

2.4 Hierarchical Bayesian Models

Standard Bayesian models are in fact defined by a two-level hierarchy, where the first level is the sample distribution $f(x|\theta)$, i.e. the distribution of X given θ, and the second level is the prior distribution $h(\theta)$. However, only one level for the prior distribution may not quantify all the prior information by itself. In practice, it is decomposed in two or more levels, which is represented by a hierarchical structure.

Formally, a hierarchical Bayesian model is specified by $\{f(x|\theta), h(\theta)\}$, where the prior distribution $h(\theta)$ is decomposed into conditional distributions $h_1(\theta|\alpha_1), h_2(\alpha_1|\alpha_2), \ldots, h_{l-1}(\alpha_{l-2}|\alpha_{l-1})$ and marginal distribution $h_l(\alpha_{l-1})$, such that

$$h(\theta) = \int h_1(\theta|\alpha_1)h_2(\alpha_1|\alpha_2)\ldots h_{l-1}(\alpha_{l-2}|\alpha_{l-1})h_l(\alpha_{l-1})d\alpha_1\ldots d\alpha_{l-1} \tag{2.6}$$

where α_i denotes the i^{th} level hyperparameter, $i = 1, \ldots, l-1$. Note that the integration extends over all possible values of $(\alpha_1, \alpha_2, \ldots, \alpha_{l-1})$. The hyperparameter distribution is called hyperprior. For the posterior distribution, (2.4) can be rewritten as

$$h(\theta|\alpha_1, \ldots, \alpha_{l-1}, x) \propto f(x|\theta)h_1(\theta|\alpha_1)h_2(\alpha_1|\alpha_2)\ldots h_{l-1}(\alpha_{l-2}|\alpha_{l-1})h_l(\alpha_{l-1}) \tag{2.7}$$

Hyperprior specifications have much less impact than first-level priors. As such, the impact of wrong specifications is much more relevant in the lower levels of the hierarchy. In general, the first level prior

distribution, $h_1(\theta|\alpha_1)$, is chosen to be a conjugate prior arising naturally from the sample model. This is also done to simplify computational procedures. As the hierarchy level increases, it becomes difficult to specify the distributions of hyperparameters. Then, a natural choice is to use non-informative priors for higher levels of the hierarchy (Paulino et al., 2018).

2.5 Markov Chain Monte Carlo Methods

It was noted that the posterior distribution is frequently difficult to handle with analytical methods (see Section 2.2). Although conjugate priors could be employed to circumvent this problem (see subsection 2.3.1), that may not be realistic.

Monte Carlo methods were developed as an alternative to existing numerical methods to handle integration. These methods are based on stochastic simulation for approximating posterior densities and expectations. However, these were still insufficient for sampling from more complex distributions . Computational difficulties were evident until the last decade of the past century. As an extension to the previous methods, Markov Chain Monte Carlo (MCMC) methods were brought into Bayesian inference, with important contributions by Gelfand and Smith [1990]. These methods consist of simulating a Markov chain whose stationary distribution is the posterior distribution, used for making inference on the model quantities of interest. Notice that several important assumptions are required to employ that Markov chain approach, and typically a larger number of iterations than with classical Monte Carlo methods.

In MCMC methods, the time the chain takes to converge depends on the starting point. Thus, the initial period of the chain, known as burn-in, is generally discarded in order to eliminate the dependence on starting values. A good choice of burn-in is fundamental to obtain well mixed chains. Thus, convergence should be monitored in order to detect correlated patterns and collinearity. If convergence problems persist after eliminating the values related to a reasonable burn-in, an alternative is to generate more samples after the burn-in and thinning the Markov chain to make the sample approximately independent. In addition, multiple chains can be simulated for convergence diagnosis.

Among MCMC methods, the Metropolis-Hastings algorithm and the Gibbs Sampler are two common algorithms to generate appropriate Markov Chains, i.e. with the desired properties. These are seen in more detail in this section.

2.5.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm is a simulation method proposed by Metropolis et al. [1953] and later generalized by Hastings [1970].

Let the conditional distribution $q(v|u)$ be an instrumental or proposal distribution, which is used for generating simulated values. Denote π the target distribution. The simple version of M-H algorithm is

given as follows:

1. Given $u^{(t)}$, $t = 0, 1, 2, \ldots$, generate a value of $V^{(t)} \sim q(v|u^{(t)})$.
2. Calculate $R(u^{(t)}, v^{(t)})$, where $R(u, v) = \frac{\pi(v)q(u|v)}{\pi(u)q(v|u)}$, and let $\alpha(u, v) = \min\{R(u, v), 1\}$.
3. The next value of the chain is the realization of

$$U^{(t+1)} = \begin{cases} v^{(t)}, & \text{with probability } \alpha(u^{(t)}, v^{(t)}) \\ u^{(t)}, & \text{with probability } 1 - \alpha(u^{(t)}, v^{(t)}) \end{cases}$$

   To decide whether or not to accept the transition $u \to v$ with probability $\alpha = \alpha(u^{(t)}, v^{(t)})$, the following procedure is considered:

   i) Generate a value z of $Z \sim U(0, 1)$.

   ii) If $z \leq \alpha$, accept transition $u \to v$, otherwise the chain remains in u.

In M-H algorithm, the chain convergence to the target distribution π depends on the proposal distribution. Moreover, the values of the chain are, approximately, simulated values of π from a given point onward. However, some considerations should be made. The proposal distribution should be chosen such that the simulated values cover the support of the target distribution in a number of iterations. On the one hand, a very quick convergence of the chain does not mean that the algorithm is working well. On the other hand, q should not be chosen very disperse with respect to π, which results in slow convergence. Therefore, a preliminary study should be conducted, where the parameters of q should be fitted in a number of trials in order to obtain reasonable acceptance rates, suggested to be in the interval $25\% - 50\%$.

The M-H algorithm can be specialised in a variety of ways. For instance, the Random-Walk M-H, Independence Sampler, Component-wise M-H, and Gibbs sampler that can be seen in detail in Paulino et al. [2018]. All these specialisations provide improvements to the proposal distribution and other steps in the algorithm. In particular, the Gibbs sampling will be of interest in this study.

2.5.2 Gibbs sampling

Gibbs sampling (Geman and Geman, 1984) is based on a sequential construction of a Markov chain by repeatedly sampling from the conditional distributions of one variable of the target distribution given all other components, known as full conditional distributions.

Let $U = (U_1, \ldots, U_k)$ denote a random vector with density $\pi(u)$ and $U_{-j} = (U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_k)$. The basic Gibbs sampling algorithm is as follows:

1. Start at $t = 0$. Given $u^{(t)} = (u_1^{(t)}, \ldots, u_k^{(t)})$, sample each component $v_j^{(t)}$ of the new chain vector from $V_j^{(t)} \sim \pi(v_j^{(t)}|u_{-j}^{(t)})$, $j = 1, \ldots, k$.

2. At the end of iteration k, take $u^{(t+1)} = (v_1^{(t)}, \ldots, v_k^{(t)})$ and return to step 1.

According to Paulino et al. [2018], Gibbs sampling is particularly useful for hierarchical models with sample variability and diversity of prior information. However, the simulation process can be slow since only one component of U is simulated at each iteration k in a given cycle.

A variation of the basic Gibbs sampler is the Gibbs sampler with updates. The difference lies on the updating process of U. The simulation of $U_j^{(t+1)}$ is now conditioned on $u_m^{(t+1)}$, $1 \leq m < j$, and $u_m^{(t)}$, $j + 1 \leq m \leq k$. The components of U are updated as follows

$$\begin{aligned}
U_1^{(t+1)} &\sim \pi(u_1|u_2^{(t)}, \ldots, u_k^{(t)}) \\
U_2^{(t+1)} &\sim \pi(u_2|u_1^{(t+1)}, u_3^{(t)}, \ldots, u_k^{(t)}) \\
&\vdots \\
U_{k-1}^{(t+1)} &\sim \pi(u_{k-1}|u_1^{(t+1)}, \ldots, u_{k-2}^{(t+1)}, u_k^{(t)}) \\
U_k^{(t+1)} &\sim \pi(u_k|u_1^{t+1}, \ldots, u_{k-1}^{t+1})
\end{aligned} \tag{2.8}$$

Gibbs sampling with updates is also known as the standard Gibbs sampling algorithm. The algorithm becomes more efficient and computational speed is increased. Other versions of Gibbs sampling can be seen in more detail in Paulino et al. [2018].

Gibbs sampling is a better option than M-H algorithm when the updating process of full conditional distributions leads to known distributions. This is the case when conjugate priors are selected. The simulation of full conditional distributions is easier than complex joint distributions. However, if conjugate priors are not selected, the full conditional distributions may not transform into known distributions. In such a case, it is more appropriate to use M-H algorithm.

Note that functions of the parameters are easily evaluated by these MCMC methods using the samples generated from the posterior distribution. Namely, given the usefulness of Gibbs sampling, it is easy to make predictions on future data by using the predictive distribution. Let Y be a random variable whose sample distribution depends on θ, based on observations of a random variable X with distribution $f(x|\theta)$. Let $h(\theta|x)$ comprise all the available knowledge of θ. The predictive distribution density is given by

$$p(y|x) = \int_\Theta f(y|x, \theta)h(\theta|x)d\theta \tag{2.9}$$

where $f(y|x,\theta)$ is the sample distribution of Y given θ, sometimes Y can be statistically independent of X conditional on θ. Then, for a MCMC sample $\{\theta_{(j)}, j = 1,\ldots,m\}$, (2.9) can be estimated by

$$\hat{p}(y|x) = \frac{1}{m}\sum_{j=1}^{m} f(y|\theta_{(j)}) \tag{2.10}$$

based on independent and identically distributed (i.i.d.) simulated values for $h(\theta|x)$.

Chapter 3

Modelling Claim Frequency in Car Insurance

In this chapter, Sections 3.1 and 3.2 introduce the models that will be used throughout this work under a Bayesian perspective, i.e. adding the representation of prior information for the model parameters. Section 3.3 comprises the construction of the joint posterior distribution, model selection and prediction of actuarial quantities.

3.1 Generalized Additive Model

Generalized Additive Models (GAM), proposed by Hastie and Tibshirani [1990], are Generalized Linear Models (GLM) that accommodate nonlinear effects of continuous covariates in the linear predictor. Consider a sample of the response variable Y belonging to the exponential distribution family, where Y_i, $i = 1, \ldots, n$, are assumed to be independent, and the predictor, which relates the expected value of Y_i (μ_i) to the covariates through a (smooth monotonic) link function g, is given by

$$g(\mu_i) = \eta_i = \gamma_0 + \boldsymbol{z}_i^T \boldsymbol{\gamma} + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) \tag{3.1}$$

where $\boldsymbol{z}_i^T$ is a row vector from design matrix $\boldsymbol{Z}$ for any strictly parametric model components, and their parameter vector γ, and $f_j(x_j)$, $j = 1, \ldots, p$, are unknown one-dimensional smooth functions of the metrical covariates x_j comprising nonlinear effects and constitute the nonparametric part. Typically, the effects of categorical covariates are gathered in the parametric part of the predictor (Brezger and Lang, 2006). Table 3.1 illustrates the link functions for some response distributions. For Gaussian

responses, the link function is usually the identity. This is a special case of GAM, known as Additive Models (AM). Higher dimensional smooth functions $f_j(\cdot)$ can also be included in the predictor, see e.g. Lang and Brezger [2004] for further details. Although model specification in terms of smooth functions of covariates provides more flexibility for detecting nonlinear covariate effects, new theoretical problems are introduced, such as the representation of smooth functions and model degree of smoothness.

Table 3.1: Response distributions and their usual link functions

Distribution	Link Function	Model Structure
Normal	Identity	$\eta = \mu$
Exponential	Inverse	$\eta = \frac{1}{\mu}$
Gamma	Inverse	$\eta = \frac{1}{\mu}$
Inverse Gaussian	Inverse squared	$\eta = \frac{1}{\mu^2}$
Poisson	Log	$\eta = \log \mu$
Binomial	Logit	$\eta = \log(\frac{\mu}{1-\mu})$

3.1.1 Representation of smooth functions

Let f_j, $j = 1, \ldots, p$, be univariate smooth functions of covariates. Wood [2006] suggests a basis representation for each f_j. A set of basis functions $\{b_{j,k}(x)\}_{k=1,\ldots,q_j}$, is chosen such that

$$f_j(x) = \sum_{k=1}^{q_j} b_{j,k}(x)\beta_{jk} \tag{3.2}$$

where β_{jk} are the smooth function parameters and q_j denotes the basis dimension used for representing f_j. For a model matrix representation of each smooth function f_j, let $\boldsymbol{f_j}$ be a vector such that $\boldsymbol{f_j}_i = f_j(x_{ji})$, $i = 1, \ldots, n$, and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = [\beta_{j1}, \ldots, \beta_{jq_j}]'$. Then,

$$\boldsymbol{f_j} = \tilde{\boldsymbol{X}}_{\boldsymbol{j}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}, \quad j = 1, \ldots, p \tag{3.3}$$

where $\tilde{\boldsymbol{X}}_{\boldsymbol{j},ik} = \{b_{j,k}(x_{ji}) : k = 1, \ldots, q_j, i = 1, \ldots, n\}$ is an element of $\tilde{\boldsymbol{X}}_j$. The presence of more than one smooth function in the additive predictor originates an identifiability problem. For instance, consider any two smooth functions f_k and f_l. By adding a constant to f_k while subtracting it from f_l does not change model predictions. This would lead to more than one solution for both f_k and f_l. A suitable convention to handle this problem is to constrain each smooth function to have zero mean, usually taken over the set of covariate values

$$\sum_{i=1}^{n} f_j(x_{ji}) = 0, \quad j = 1, \ldots, p \tag{3.4}$$

As for smooth functions f_j, a matrix representation of (3.4) is given by

$$\mathbf{1}^T \tilde{\boldsymbol{X}}_{\boldsymbol{j}} \tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = 0, \quad j = 1, \ldots, p \tag{3.5}$$

which is easily absorbed by re-parameterization of the model.

There are several different choices for basis functions, such as polynomial functions. Wood [2006] recommends using splines for a more global study of each f_j. A spline approach for smooth functions is appealing due to good theoretic approximation properties. Cubic splines were shown to be optimal or, at least, very good approximations in interpolation problems (De Boor, 1978). In particular, Green and Silverman [1994] have shown that natural cubic splines are the smoothest in the sense of minimizing

$$J(f_j) = \int_{x_1}^{x_n} (f_j''(x))^2 dx \tag{3.6}$$

The definition of smoothness according to (3.6) determines curves with sharp ”kinks” that have larger squares of the second derivative, as opposed to flatter curves.

For regression splines, typical approaches are either to assume evenly spaced knots x_j through the range of observed x values, or place the knots at the quantiles of the distribution of unique x values. A drawback of spline-based approaches is the choice of knot location as these approaches introduce some degree of subjectivity and possibly overfitting as well. Another drawback is related to the choice of basis dimension. Choosing a large number of basis functions leads to a wide space of possible smooth functions, in contrast with a small basis. Wood [2006] states that the choice is not critical, but should be made as close as possible, since model fitting can retain some sensitivity from it. Alternatively, thin plate regression splines (TPRS) could be used, see e.g. Wood [2003], which enable the construction of knot free bases for smooth functions of any number of predictors. Nevertheless, the TPRS basis and penalties can become computationally expensive to calculate for large data sets.

3.1.2 Degree of smoothness

In order to avoid curve fitting problems, such as overfitting, penalized least squares regression can be employed. The sensitivity related to knot selection is circumvented by placing knots at all data points and by adding roughness penalties to shrink the coefficients of the estimated smooth functions (in their

basis expansion). For AM (Gaussian responses), the following minimization problem is considered

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \arg\min_{\beta} \left[||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_1 \int_{x_1}^{x_n} f_1''(x)^2 dx + \cdots + \lambda_p \int_{x_1}^{x_n} f_p''(x)^2 dx \right] \tag{3.7}$$

where λ_j, $j = 1, \ldots, p$ are smoothing parameters that control the trade-off between residual error and local variation. For given values of λ_j, solving (3.7) leads to the best compromise between model smoothness and goodness-of-fit. For smoothing parameter estimation, common approaches are to minimize the Generalized Cross Validation (GCV) score (Craven and Whaba, 1979) or the Akaike Information Criterion (AIC) [Wood, 2006]. Notice that for large λ_j, the estimated smooth functions will display little curvature (large shrinkage), as opposed to small values of the λ_j. In general, (3.7) is not solved directly, but instead orthogonal matrix methods are preferred due to greater numerical stability.

For non-Gaussian responses, the setting is different. In order to control the degree of smoothness, GAM are set up as a penalized GLM. Since model estimation in GLM is based on Iteratively Re-Weighted Least Squares (IRLS), a penalized version of IRLS is considered, known as Penalized Iteratively Re-Weighted Least Squares (P-IRLS). For further details, see e.g. Wood [2006].

3.1.3 B-splines and P-splines

An attractive approach based on penalized regression splines was presented by Eilers and Marx [1996] in a frequentist setting. The unknown functions are approximated by a polynomial spline of degree l considering equally spaced knots over the domain of x_j, given by

$$\zeta_{j0} = x_{j,min} < \zeta_{j1} < \cdots < \zeta_{j,k_j-1} < \zeta_{jk_j} = x_{j,max}, \quad j = 1, \ldots, p$$

The spline can be written in terms of a linear combination of $M_j = k_j + l$ B-spline basis functions (see De Boor, 1978) for f_j, $j = 1, \ldots, p$. Thus, denoting the respective m^{th} basis function by B_{jm}, $m = 1, \ldots, M_j$, then

$$f_j(x_j) = \sum_{m=1}^{M_j} \beta_{jm} B_{jm}(x_j) \tag{3.8}$$

or according to a matrix representation

$$\boldsymbol{f_j} = \boldsymbol{X}_j \boldsymbol{\beta}_j \tag{3.9}$$

where $\boldsymbol{X}_{j,im} = B_{jm}(x_{ij})$. As an attempt to overcome knot selection, Eilers and Marx [1996] suggested the use of a relatively large number of knots (usually between 20 and 40) to ensure enough flexibility. Also, roughness penalties are assumed for adjacent regression coefficients in order to avoid overfitting.

Model estimation is performed by direct maximization of the penalized likelihood (Eilers and Marx, 1998), which is given by

$$\mathcal{L}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\gamma}|\boldsymbol{y}) = l(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\gamma}|\boldsymbol{y}) - \lambda_1 \sum_{l=w+1}^{M_1} (\Delta^w \beta_{1l})^2 - \cdots - \lambda_p \sum_{l=w+1}^{M_p} (\Delta^w \beta_{pl})^2 \tag{3.10}$$

with respect to the unknown regression coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j,M_j})$, $j = 1, \ldots, p$, and $\boldsymbol{\gamma}$. Here, Δ^w denotes the difference operator of order w, where first ($w = 1$) or second ($w = 2$) order differences are typically used. The penalization is accomplished by smoothing parameters, which can be estimated via GCV or AIC.

A penalized spline (P-spline) approach remedies the problems of regression splines, allowing a good deal of flexibility in model estimation and interpretation. However, it was originally developed under a frequentist perspective, which can lead to severe difficulties in smoothing parameter estimation. GCV or AIC criteria usually fail if the number of smooth functions in the predictor is large (Wood, 2006).

Bayesian inference was introduced in the context of GAM to enable model estimation in application problems with a large number of parameters. Lang and Brezger [2004] proposed Bayesian P-splines, whose parameter estimation is based on Markov Chain Monte Carlo techniques. For fixed effect parameters, diffuse priors are assumed, i.e. $h(\gamma_j) \propto$ const. For parameter vectors $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$ the difference penalties in (3.10) are replaced by their stochastic analogues, respectively first and second order random walks

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho} \tag{3.11}$$

$$\beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{3.12}$$

with gaussian errors, i.e. $u_{j\rho} \sim \mathcal{N}(0, \tau_j^2)$ and initial values $\beta_{j1} \propto const$ of (3.11) , or β_{j1} and $\beta_{j2} \propto const$ of (3.12). The priors for $\boldsymbol{\beta}_j$ can be equivalently written in the form of global smoothness priors

$$\boldsymbol{\beta}_j|\tau_j^2 \propto \frac{1}{(\tau_j^2)^{rank(\boldsymbol{K}_j)/2}} \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j \right) \tag{3.13}$$

where τ_j^2, $j = 1, \ldots, p$, are smoothing variances determining the prior confidence on smoothness and

correspond to the inverse smoothing parameters in the classical approach, the matrices $\boldsymbol{K}_j$ represent the precision matrices implementing prior assumptions about the smoothness of f_j. In particular, $\boldsymbol{K}_j = \boldsymbol{D}^T\boldsymbol{D}$, where $\boldsymbol{D}$ is a difference matrix of appropriate order. The use of global smoothness priors particularly facilitates the description and implementation of MCMC inference (Brezger and Lang, 2006).

Notice that the design matrix comprises evaluations of B-spline basis functions defined upon an equidistant grid of knots. Cubic B-splines are typically chosen, with 20 inner knots, and second order penalization (Lang and Brezger, 2004).

Under a Full Bayesian inference, additional hyperparameters for the smoothing variances (and the overall variance σ^2, for gaussian responses) are estimated simultaneously with regression coefficients. Lang and Brezger [2004] proposed highly dispersed (but proper) inverse Gamma hyperpriors, IG(a_j, b_j), for smoothing variances, with $a_j = 1$ and small value for b_j, for instance, $b_j = 0.005, 0.0005$ or 0.00005. Alternatively, Klein et al. [2014] proposed $a_j = b_j = 0.001$ in order to obtain a more data-driven amount of smoothness. In some situations, the estimated smooth functions f_j may considerably depend on the choice of a_j and b_j, e.g. for very low signal-to-noise ratios or small sample sizes. Therefore, a sensitivity analysis is recommended to assess the dependence of results.

3.1.4 Modeling interactions

In situations with several covariates, a simple additive predictor may not be appropriate due to interactions between covariates. An interaction between a categorical covariate, $x_j^{(1)}$, and a continuous covariate, $x_j^{(2)}$, is conveniently modeled within the varying coefficient framework, introduced by Hastie and Tibshirani (1993). The effect of $x_j^{(1)}$ is assumed to vary smoothly over the range of $x_j^{(2)}$, also called the effect modifier of $x_j^{(1)}$. The interaction effect is represented by $x_j^{(1)} \times f_j(x_j^{(2)})$, where $f_j(\cdot)$ is typically approached via Bayesian P-splines.

If both interacting covariates are continuous, the approach is based on two-dimensional surface fitting, using mainly two-dimensional Bayesian P-splines that are described in more detail in Lang and Brezger [2004].

3.2 Generalized Geoadditive Models

3.2.1 Spatial mixed models

Fahrmeir et al. [2004] formally proposed Structured Additive Regression (STAR), which is a class of complex models whose distributional and structural assumptions, given covariates and parameters, are based on GLM. This class comprises GAM and a wide variety of extensions.

In many applications, responses depend not only on metrical and categorical covariates, but also on

the spatial location where they have been observed. In order to take into account spatial heterogeneity, a spatial effect f_{spat} is introduced into the additive predictor. Let $s = 1, \ldots, S$ be the regions in a geographical map. Therefore, Generalized Geoadditive Models (GGAM), proposed by Kammann and Wand [2003], are given by:

$$g(\mu_i) = \gamma_0 + \boldsymbol{z}_i^T \boldsymbol{\gamma} + f_1(x_{1i}) + \cdots + f_p(x_{pi}) + f_{spat}(s_i) \tag{3.14}$$

Depending on the application, the models (3.14) may not contemplate all sorts of spatial heterogeneity. There may exist heterogeneity between units that is not observed by covariates, smooth functions of covariates and geographical information. The spatial effect, f_{spat}, is further decomposed into two components: a spatially correlated effect, also known as structured spatial effect, and a spatially uncorrelated effect, known as an unstructured spatial effect. This leads to the Generalized Geoadditive Mixed Models (GGAMM), whose additive predictor is rewritten as

$$g(\mu_i) = \gamma_0 + \boldsymbol{z}_i^T \boldsymbol{\gamma} + f_1(x_{1i}) + \cdots + f_p(x_{pi}) + f_{str}(s_i) + b_{si}, \tag{3.15}$$

where f_{str} denotes the structured spatial effects and b_{si} is a unit- or group-specific (spatial) random effect, with $b_{si} = b_s$ if unit i is in $s = 1, \ldots, S$.

Since GGAM and GGAMM will possibly comprise a larger number of parameters than GAM, it is even more convenient to use a full Bayesian inference for namely estimating the model parameters.

For these GAM, the prior assumptions introduced in subsection 3.1.3 are hold concerning the spline function parameter vector $\boldsymbol{\beta_j}$, (fixed effect) regression parameter vector $\boldsymbol{\gamma}$ and variance parameters τ_j^2, $j = 1, \ldots, p$. Especially for spatial data observed on a regular lattice, a common approach for f_{str} is based on Markov Random Field (MRF) priors for $\boldsymbol{\beta}_{str}$ (Besag et al., 1991). Let $s \in \{1, \ldots, S\}$ denote the pixels of a lattice or regions in a geographical map. Then, a usual MRF prior for $f_{str}(s) = \beta_{str,s}$ is defined by

$$\beta_{str,s} | \beta_{str,u}, u \neq s, \tau_{str}^2 \sim \mathcal{N}\Big(\sum_{u \in \partial_s} \frac{1}{N_s} \beta_{str,u}, \frac{\tau_{str}^2}{N_s} \Big) \tag{3.16}$$

where N_s is the number of adjacent regions to s, and ∂_s denotes the set of neighbours to s, and τ_{str}^2 is the spatial variance parameter. The adjacency matrix is a $n \times S$ matrix, where the entries are given by $\boldsymbol{X}_{str}(i, s) = 1$, if observation i has been observed at location s, and zero otherwise.

To implement the spatial smoothness, it is more convenient to represent the prior (3.16) in the form of a global smoothness prior, with $\boldsymbol{K}_{str}$ being a penalization matrix (Rue and Held, 2005), that is,

$$\boldsymbol{\beta}_{str}|\tau^2_{str} \propto \frac{1}{(\tau^2_{str})^{rank(\boldsymbol{K}_{str})/2}} \exp\left(-\frac{1}{2\tau^2_{str}} \boldsymbol{\beta}^T_{str} \boldsymbol{K}_{str} \boldsymbol{\beta}_{str} \right) \tag{3.17}$$

where the elements of $\boldsymbol{K}_{str}$ are given by $k_{ss} = N_s$, $k_{su} = -1$ if $u \in \partial_s$, and $k_{su} = 0$ if $u \notin \partial_s$ zero otherwise. Here, the prior assumptions for τ^2_{str} are those for τ^2_j in Section 3.2. Another approach for the structured effect, f_{str}, is to use two-dimensional surface estimators. However, Brezger and Lang [2004] state that MRF priors are often superior for model fitting.

For uncorrelated spatial effects , it is generally assumed i.i.d. Gaussian distributions given by

$$b_s|\nu^2 \sim \mathcal{N}(0, \nu^2), \quad s = 1, \ldots, S \tag{3.18}$$

where ν^2 is typically assigned a highly dispersed hyperprior.

3.2.2 Zero-inflated models

In count data, zero counts are sometimes more frequently observed than other values, which in practice reflects an excess of zeros as compared to the number of zeros expected from a count probability distribution e.g. Poisson distribution. Kuhnert et al. [2005] suggest that an excess of zeros may be due to poor experimental design (for instance, the sampling period is too short) or observation errors (such as counting zeros when the true value is not zero), among others.

The distinction between false and true zeros should be made. Note that false zeros may result from policyholders not claiming in order to benefit from premium discounts or even fill-in errors. A natural extension to the Poisson distribution is the zero-inflated Poisson (ZIP) distribution, which is decomposed into two components. The first part is a degenerated distribution accounting for the probability of false zeros, whereas the second part is a usual Poisson distribution, which includes true zeros.

A common problem related to count data regression is overdispersion, e.g. the assumption of equal expectation and variance in the Poisson distribution is unrealistic. A useful approach to deal with overdispersion in count data is to consider the negative binomial distribution instead of the Poisson distribution. That is a convenient extension to the Poisson distribution by using a second parameter determining the scale of the distribution. In the context of excess of zeros, a zero-inflated negative binomial (ZINB) model is a suitable extension to the negative binomial model, and a better option than ZIP models, if

overdispersion persists. Klein et al. [2014] suggest using ZINB models to model claim frequency in car insurance data, as an alternative to Poisson and ZIP models.

ZIP and ZINB distributions are embedded in the framework of generalized additive models for location, scale and shape (GAMLSS) [Rigby and Stasinopoulos, 2005], which is an extension of GAM-based models to accommodate more complex response distributions. That is, not only the expectation, but also other parameters of the distribution can be specified by additive predictors via suitable link functions.

For zero-inflated count data, assume the response variable Y_i, as well as covariate information ν_i, have been collected for individuals $i = 1, \ldots, n$. Denote π_i a probability mass at zero and $\mathbb{1}_{\{0\}}(\cdot)$ the indicator function of zero. The conditional distribution of Y_i given ν_i is described in terms of the probability function

$$p(Y_i = y_i|\nu_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i|\nu_i) \tag{3.19}$$

arising from the hierarchical definition of the responses $Y_i = D_i\tilde{Y}_i$, where d_i is a Bernoulli selection process $D_i \sim \text{Ber}(1 - \pi_i)$, and $\tilde{Y}_i$ follows a count data distribution, which can be Poisson, $\tilde{Y}_i \sim \text{Po}(\lambda_i)$, or negative binomial, $\tilde{Y}_i \sim \text{NB}\left(\delta_i, \frac{\delta_i}{\delta_i+\mu_i}\right)$, whose probability mass functions are respectively given by

$$\tilde{p}(\tilde{y}_i) = \frac{\lambda_i^{\tilde{y}_i} e^{-\lambda_i}}{\tilde{y}_i!} \tag{3.20}$$

$$\tilde{p}(\tilde{y}_i) = \frac{\Gamma(\tilde{y}_i + \delta_i)}{\Gamma(\tilde{y}_i + 1)\Gamma(\delta_i)} \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i}\right)^{\tilde{y}_i} \tag{3.21}$$

where λ_i is the Poisson parameter, μ_i and $\delta_i = \frac{1}{\sigma_i}$ are the negative binomial parameters, with σ also being known as the dispersion parameter. Note that (3.21) corresponds to the probability mass function of negative binomial type I. Hence, the probability mass functions can be written for ZIP and ZINB distributions. For the ZIP distribution, it is given by

$$p(Y_i = k|\nu_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\lambda_i} & k = 0 \\ (1 - \pi_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!} & k = 1, 2, \ldots \end{cases} \tag{3.22}$$

The corresponding mean and variance are given by

$$\mathsf{E}[Y_i] = (1 - \pi_i)\lambda_i$$

$$\text{Var}[Y_i] = (1-\pi_i)\lambda_i + \pi(1-\pi_i)\lambda_i^2 \tag{3.23}$$

From (3.23), the probability of excess of zeros π_i weighs the mean of the Poisson distribution. Also, the variance of the distribution exceeds the mean. This is the mechanism to account for overdispersion in ZIP models.

Analogous reasoning can be used for the negative binomial case. By (3.21) the probability mass function for the ZINB distribution is given by

$$p(Y_i = k|\nu_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i} & k = 0 \\ (1-\pi_i)\frac{\Gamma(k+\delta_i)}{\Gamma(k+1)\Gamma(\delta_i)}\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}\left(\frac{\mu_i}{\delta_i+\mu_i}\right)^{k} & k = 1, 2, \dots \end{cases} \tag{3.24}$$

The use of this distribution is of interest because, since σ_i is the dispersion parameter in GLM, it can be assessed and interpreted directly. The mean and variance are given by

$$\begin{aligned} \text{E}[Y_i] &= (1-\pi_i)\mu_i \\ \text{Var}[Y_i] &= (1-\pi_i)\mu_i\left(1 + \mu_i\left(\pi_i + \frac{1}{\delta_i}\right)\right) \end{aligned} \tag{3.25}$$

The model specifications for both the probability of excess of zeros and count data parameters can be related to regression predictors via suitable link functions.

For ZIP models, the default predictor options are the log link function for μ_i and logit link function for π_i

$$\begin{cases} \eta_i^{\pi} = \text{logit}(\pi_i) \equiv \log(\frac{\pi_i}{1-\pi_i}) \\ \eta_i^{\lambda} = \log(\lambda_i) \end{cases}$$

For ZINB models, the log link function is chosen for both μ_i and δ_i, and the logit link function for π_i

$$\begin{cases} \eta_i^{\pi} = \text{logit}(\pi_i) \\ \eta_i^{\mu} = \log(\mu_i) \\ \eta_i^{\delta} = \log(\delta_i) \end{cases}$$

For both specifications, the generation of false zeros can be described by covariate effects

$$\pi_i = \frac{e^{\theta_0+\theta_1 Z_{i1}+\cdots+\theta_q Z_{iq}}}{1+e^{\theta_0+\theta_1 Z_{i1}+\cdots+\theta_q Z_{iq}}} \tag{3.26}$$

where $Z_1, \ldots, Z_q$ are covariates used in the model specification of μ and σ, or even further covariates. A common (and easy) approach is to choose an intercept-only model for π , i.e. $\text{logit}(\pi_i) = \theta_0$.

3.3 Bayesian Inference

3.3.1 Joint posterior distribution

For simplicity, a Poisson model (3.15) is considered herein whose Bayesian model specification is completed by the following assumptions:

1. Given covariates and parameters $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{str}$, $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, b_s, $s = 1, \ldots, S$, the (response) variables y_i are independent.

2. Given τ_j^2 and τ_{str}^2, $\boldsymbol{\beta}_j|\tau_j^2$ and $\boldsymbol{\beta}_{str}|\tau_{str}^2$ are independent.

3. Priors for fixed $\boldsymbol{\gamma}$ and random $\boldsymbol{b}$ effects, and for τ_j^2, $j = 1, \ldots, p$, are mutually independent.

For the following let $\boldsymbol{\alpha}$ denote the vector of all model parameters:

$$\boldsymbol{\alpha} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{\gamma}, \boldsymbol{\beta}_{str}, \tau_{str}^2, \boldsymbol{b}) \tag{3.27}$$

By assumptions above, the joint posterior distribution can be written as

$$h(\boldsymbol{\alpha}|\boldsymbol{y}) \propto h(\boldsymbol{y}|\boldsymbol{\alpha})h(\boldsymbol{\alpha}) \tag{3.28}$$

$$h(\boldsymbol{\alpha}|\boldsymbol{y}) \propto \left(\prod_{i=1}^{n} L_i(y_i;\eta_i)\right) \times \left(\prod_{k=1}^{r} h(\gamma_k)\right) \times \left(\prod_{j=1}^{p} \{h(\boldsymbol{\beta}_j|\tau_j^2)h(\tau_j^2)\}\right) \times \left(\prod_{s=1}^{S} \{h(b_s|\nu^2)h(\nu^2)\right) \tag{3.29}$$

where $\nu^2 \sim IG(a', b')$. The generic prior distributions in (3.29) are replaced by those proposed in subsections 3.1.3 and 3.2.1. After some manipulation, the joint posterior distribution (3.28) can be written as

$$h(\boldsymbol{\alpha}|\boldsymbol{y}) \propto \left(\prod_{i=1}^{n} L_i(y_i;\eta_i)\right) \exp\left\{-\sum_{j=1}^{p} \frac{1}{2\tau_j^2}(\boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j + 2b_j) - \frac{1}{2\tau_{str}^2}(\boldsymbol{\beta}_{str}^T \boldsymbol{K}_{str} \boldsymbol{\beta}_{str} + 2b_{str}) - \frac{b'S}{\nu^2} - \sum_{s=1}^{S} \frac{b_s^2}{2\nu^2}\right\} \times$$

$$\times \left(\frac{1}{\prod_{j=1}^{p} (\tau_j^2)^{rank(\boldsymbol{K}_j)/2}} \right) \times \frac{1}{(\tau_{str}^2)^{rank(\boldsymbol{K}_{str})/2}} \left(\prod_{j=1}^{p} (\tau_j^2)^{-a_j-1} \right) (\tau_{str}^2)^{-a_{str}-1} (\nu^2)^{-S(a'+1)} \tag{3.30}$$

Depending on the response distribution, the posterior distribution will vary accordingly. For a Poisson model, (3.30) is difficult to handle, i.e. assess the full conditional posterior distributions, especially for regression coefficients arising from the basis function expansion (see Appendix A.1 for more detail). In ZIP and ZINB cases, the problem becomes analytically intractable due to complex structure of the GAMLSS models. In contrast with the regression coefficients of smooth functions, the full conditionals for the smoothing variances τ_j^2 and τ_{str}^2 can be derived in closed form (see Appendix A.1), since inverse gamma distributions, IG(a_j, b_j), were assumed. The corresponding parameters can be updated as follows

$$a_j^\star = a_j + \frac{\text{rank}(\boldsymbol{K}_j)}{2} \tag{3.31}$$

$$b_j^\star = b_j + \frac{1}{2} \boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j \tag{3.32}$$

In order to proceed for regression coefficients, density proposals to approximate the full conditional posterior distributions are suggested below.

3.3.2 Iterative Weighted Least Squares proposals

Lang and Brezger [2004] proposed density approximations based on Iteratively Weighted Least Squares (IWLS) to the full conditional distributions for AM. Brezger and Lang [2006] extended this approach for GAM with non-Gaussian responses. Briefly, IWLS proposals are used to determine quadratic approximations of full conditional distributions. It consists of Gaussian proposal densities with expectation and covariance matrices corresponding to the mode and curvature of the quadratic approximation. Then, parameters are updated by Metropolis-Hastings algorithm.

Denote $\boldsymbol{\eta}^c$ the current predictor based on the current regression coefficients $\boldsymbol{\beta}_j^c$ of smooth function f_j, $j = 1, \ldots, p$ (analogously for the structured effect). To update $\boldsymbol{\beta}_j$, the current state of the chain is used. Then, a new value $\boldsymbol{\beta}_j^\star$ is proposed by drawing a random number from the multivariate Gaussian proposal distribution $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^\star)$, with precision matrix and mean given by, respectively,

$$\boldsymbol{P}_j = \boldsymbol{X}_j^T \boldsymbol{W}(\boldsymbol{\eta}^c) \boldsymbol{X}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j$$

$$\boldsymbol{m}_j = \boldsymbol{P}_j^{-1} \frac{1}{\sigma^2} \boldsymbol{X}_j^T \boldsymbol{W}(\boldsymbol{\eta}^c)(\boldsymbol{y}(\boldsymbol{\eta}^c) - \tilde{\boldsymbol{\eta}}^c) \tag{3.33}$$

where $\tilde{\boldsymbol{\eta}}^c$ is the part of the predictor related to all the remaining effects in the model, $\boldsymbol{W}$ is a matrix of appropriate working weights and $\boldsymbol{y}$ is a vector of working observations. In general, matrices $\boldsymbol{P}_j$ and $\boldsymbol{K}_j$ are sparse.

There are two possible updating schemes: sampling scheme 1 (IWLS proposals based on current mode) and sampling scheme 2 (IWLS proposals, update $\boldsymbol{\beta}_j$ and τ_j^2), see Brezger and Lang [2006] for more details. The latter is particularly useful for updating spatial structured effects based on MRF priors, as it remedies the problem of low acceptance rates originated by high dimensional parameter vectors $\boldsymbol{\beta}_j$.

Since the full conditional distributions for smoothing variances are known distributions (Appendix A.1), a simple Gibbs sampler can be used to update the respective parameters.

Density approximation via IWLS proposals specifically for ZIP and ZINB models can be seen in more detail in Klein et al. [2014].

3.3.3 Model selection

The selected model process is quite challenging because there is no automated procedure to follow. It is recommended to start by simple model formulations, and increase model complexity step by step (Brezger and Lang, 2006). This is also pointed out by Klein et al. [2014] in order to avoid convergence issues, especially for ZINB models.

Let $D(\theta) = -2 \log p(y|\theta) + 2 \log f(y)$ denote the deviance for a likelihood $p(y|\theta)$, where y are the data, θ are unknown parameters of the model and $f(y)$ is some fully specified standardising term that is a function of the data alone, thus not affecting model comparison. An approach to perform variable selection is to look for models that significantly decrease the Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002], defined by

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D \tag{3.34}$$

$$= 2\overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}}) \tag{3.35}$$

where $\overline{D(\boldsymbol{\theta})}$ is the posterior expected deviance, measuring the goodness-of-fit, $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters, measuring model complexity, and $\overline{\boldsymbol{\theta}}$ is the posterior mean of the model parameter vector. The DIC is a generalization of the Akaike Information Criterion (AIC), being widely used for model selection in hierarchical Bayesian models, whose inference is based on MCMC simulation techniques. The DIC has been shown to provide valid guidance for covariate selection in

Klein et al. [2014]. Also, Watanabe [2010] proposed the widely applicable information criterion (WAIC), which is a measure of predictive accuracy with more Bayesian focus than the DIC, given by

$$\text{WAIC} = -2\sum_{i=1}^{n} \log \mathbb{E}_{\theta|x}[f(x_i|\theta)] + 2p_W$$

for a given likelihood $f(x|\theta)$, where p_W can be obtained using two different proposals (see e.g. Paulino et al., 2018).

In order to compare the performance of model distributions, model dispersion and quantile residuals are inspected (Zuur et al., 2012; Klein et al., 2014).

The analysis of model residuals may suggest possible outliers, which can have some impact on model fitting. In particular, atypical observations may lead to chains that are not well mixed. For instance, poor mixing is demonstrated by large spikes for large lags in Auto-Correlation Function (ACF) plots and visible trends and patterns in trace plots of parameters (see Appendix A.2). Convergence issues can also be detected by multiple chains convergence tests, such as Gelman and Rubin's diagnostic (Gelman and Rubin, 1992). In order to solve this issue, the number of iterations and burn-in period are typically increased. For convergence diagnostic methods, see e.g. Paulino et al. [2018].

3.3.4 Estimation of actuarial quantities

In car insurance applications, there are some important indicators that can be used to apply appropriate insurance premiums, given the risk structure of a policy. In particular, estimations of these indicators can be obtained in order to define risk profiles.

No-claim probablity

No-claim probability is a key actuarial indicator for underwriting an annual policy renewal. This is a measure of risk based on the conditional probability that a policyholder does not make any claim, given its risk structure. Typically, small no-claim probabilities indicate that the respective policyholders are more likely to make claims. Car insurances pay special attention to these policies, thus applying larger premiums. On the contrary, policies with large no-claim probability are applied lower premiums, and policyholders may even benefit from premium discounts. Given n observations, the no-claim probability for some count distributions is given by

$$\text{Pr}(N_i = 0) = \begin{cases} e^{-\lambda_i} & \text{Poisson} \\ \left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i} & \text{NB} \\ \pi_i + (1-\pi_i)e^{-\lambda_i} & \text{ZIP} \\ \pi_i + (1-\pi_i)\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i} & \text{ZINB} \end{cases} \tag{3.36}$$

where N_i is the number of insurance incidents for observation i and $\sigma_i = \frac{1}{\delta_i}$ accounts for the dispersion of observation i.

Expected claim frequency

The expected claim frequency is another important actuarial indicator. It indicates the expected number of claims for an insurance policy. Naturally, policies with large expected claim numbers are riskier, thus higher insurance premiums are applied to those policyholders. On the other hand, policies with small expected claim numbers are less risky. Given n observations, the expected claim frequency for some count distributions is given by

$$\text{E}(N_i) = \begin{cases} \lambda_i & \text{Poisson} \\ \mu_i & \text{NB} \\ (1-\pi_i)\lambda_i & \text{ZIP} \\ (1-\pi_i)\mu_i & \text{ZINB} \end{cases} \tag{3.37}$$

Chapter 4

Car Insurance Application

4.1 Exploratory Data Analysis

The original data set comprises policies from a Portuguese car insurance in the period $2011 - 2013$, consisting of $604,649$ observations on 21 variables. Table 4.1 provides a brief description of the used variables.

Table 4.1: Description of the used variables

Variable	Type	Description
DISTRITO†	Categorical	District in which the policy is registered (21 categories)
DECR_SEXO_PESSOA	Categorical	Policyholder's gender (3 categories)*
CATEGORIA_AGREGADA	Categorical	Aggregated category of vehicle (9 categories)
ESCALAO_CILINDRADA	Categorical	Engine displacement of vehicle (3 categories)
Idade_Veiculo†	Continuous	Age of the vehicle (years)
Anos_Carta†	Continuous	Policyholder's car license time (years)
Idade_Condutor†	Continuous	Age of the policyholder (years)
DESCR_TIPO_USO	Categorical	Vehicle usage (13 categories)
GARAGEM	Binary	Is there a garage for the vehicle? (yes/no)
Credor	Binary	Is there a creditor? (yes/no)
Marca_Conv†	Categorical	Vehicle brand (35 categories)
CS_2011.2013	Continuous	Costs with insurance incidents in $2011 - 2013$
NS_2011.2013	Integer	Claim frequency in $2011 - 2013$
Apol_2011.2013	Continuous	Policy duration or risk period

† NA is a category; * the third category represents an undefined gender.

Some modifications were made for variables Marca_Conv, CATEGORIA_AGREGADA, DESCR_TIPO_USO and DISTRITO.

- Marca_Conv - The vehicle brand comprises 34 categories, some of them with very small frequency. Also, keeping all the levels obviously lead to slower computation during model fitting. Thus, it was conveniently aggregated by brand home country. Countries whose observed proportions amounted to around 95% had their categories kept in the new variable, while the remaining ones

were aggregated in a new category called "OTHER". For observations with non-specified brand, these were aggregated in a new level called "NS" (see fig. 4.1).

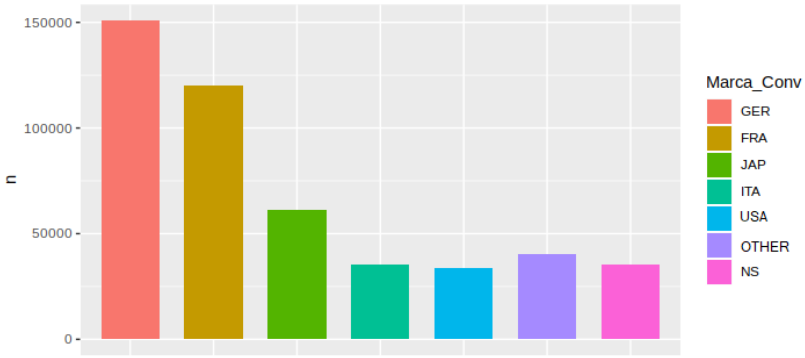

Figure 4.1: Frequency of the new covariate Marca_Conv (brand home country).

- CATEGORIA_AGREGADA - For the aggregated category of vehicle, its observed category proportions were ranked and those that sum up to, at least, 95% were kept, while the remaining ones were aggregated in a new level called "OTHER" (see fig. 4.2).
- DESCR_TIPO_USO - Most categories of vehicle usage are subcategories of "PROF". For this reason, these were suitably aggregated.

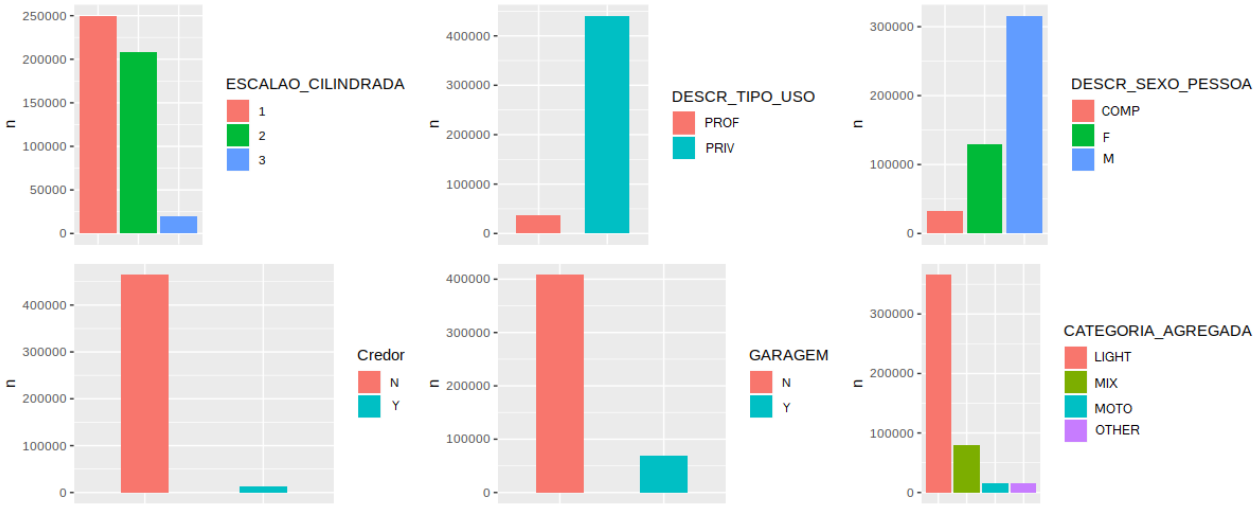

Figure 4.2: Frequency of the covariates ESCALAO_CILINDRADA, DESCR_TIPO_USO, DESCR_SEXO_PESSOA, Credor, GARAGEM and CATEGORIA_AGREGADA.

- DISTRITO - Districts MADEIRA (Autonomous Region of Madeira) and ACORES (Autonomous Region of Azores) are islands, which do not share any land borders with other Portuguese districts. The

existence of disconnected areas leads to a deficient construction of the neighborhood structure. One possibility to solve this issue could be the addition of neighbors through the neighborhood matrix, which in practice is not realistic. Therefore, only policies registered in Portugal mainland will be considered in the forthcoming data analysis (see fig. 4.3).

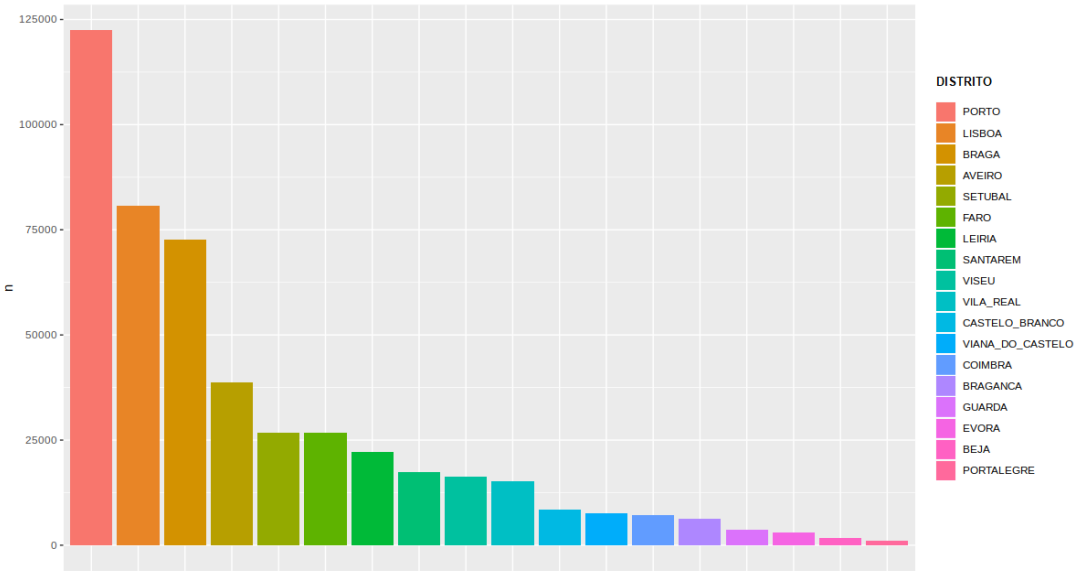

Figure 4.3: Frequency of the covariate `DISTRITO`.

After this data set reorganization, it comprises $585,256$ observations on 14 variables. The response variable `NS_2011.2013`, from now on denoted by Y, refers to claim frequency.

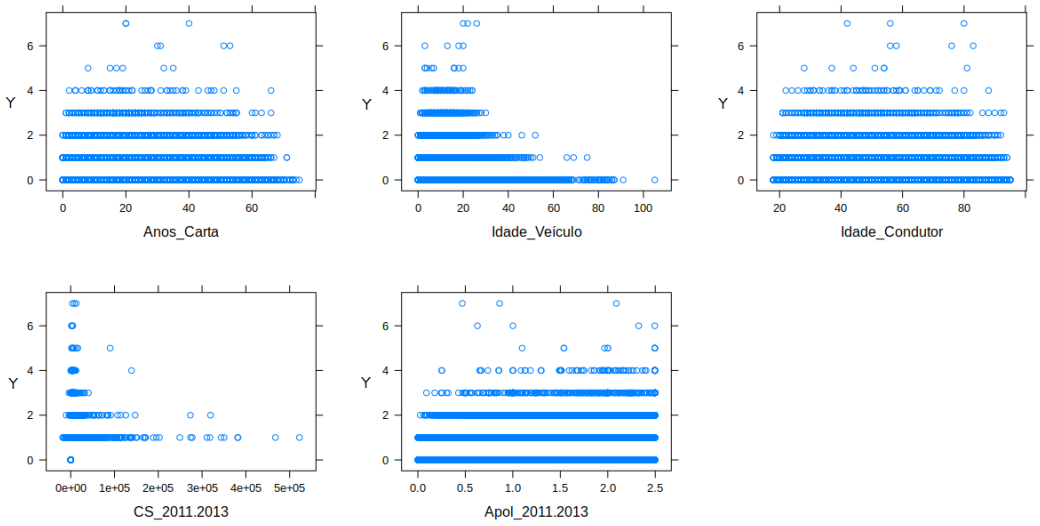

Figure 4.4: Dot plots of the quantitative covariates *versus* the response Y (or claim frequency).

Figure 4.4 shows the existence of very large values for license time (`Anos_Carta`) and age of the vehicle (`Idade_Veiculo`), which may be due to fill-in errors. There is a significant number of observations with age of the policyholder (`Idade_Condutor`) larger than 80. The range of values for the aggregated cost (`CS_20112013`) is also very wide, from -18070.7 to 522296.5. Note that a large aggregated cost does not imply a large claim frequency, and vice-versa. Thus, it is suspected that the aggregated cost cannot be a good predictor for claim frequency.

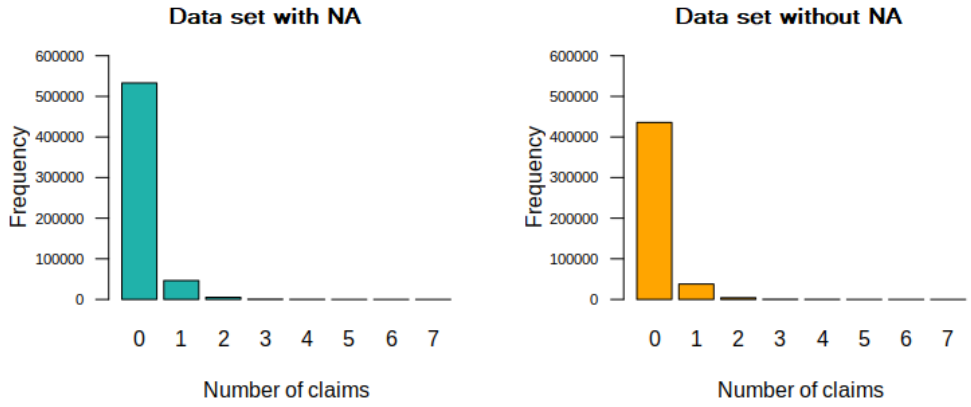

Figure 4.5: Claim frequency (`NS_20112013`) in the original data (left) and data without NA (right).

Since there are some atypical (extreme) observations, the performance of the selected model with and without those observations will be inspected. Also, figure 4.5 suggests an excess of zeros in the response variable for both original (with NA) and data set without NA.

Since the original data set contains missing values, an initial analysis was performed with the complete data, which comprises $477,997$ observations. Note that there are missing values only for the covariates. However, removing all the observations with at least one missing value may not be totally adequate. It is of interest to perform imputation and then compare the selected models in both cases.

4.2 Data Imputation

In many applications, data missing is a common problem and a threat to data analysis. If the data missing pattern resembles Missing at Random (MAR) or Missing Completely at Random (MCAR) [1], removing the observations with missing values may be acceptable. However, if data missing patterns are evident, e.g. Missing Not at Random (MNAR) here, it can bring bias into the models.

[1] If the probability of being missing is the same for all cases or only within groups defined by the observed data, then the data are said to be missing completely at random (MCAR) or missing at random (MAR), respectively. Otherwise, we have missing not at random (MNAR) i.e. the probability of being missing varies for reasons that are unknown to us (Rubin, 1976).

In the original data set, missing data typically occur for license time (`Anos_Carta`) and age of the policyholder (`Idade_Condutor`). Although many other combinations arise, most of them have negligible proportions of missing data (see figure 4.6 and table 4.2). It is not clear whether the pattern is MAR or MNAR, but the fact most of the missing data occur for continuous covariates may have some impact on the results.

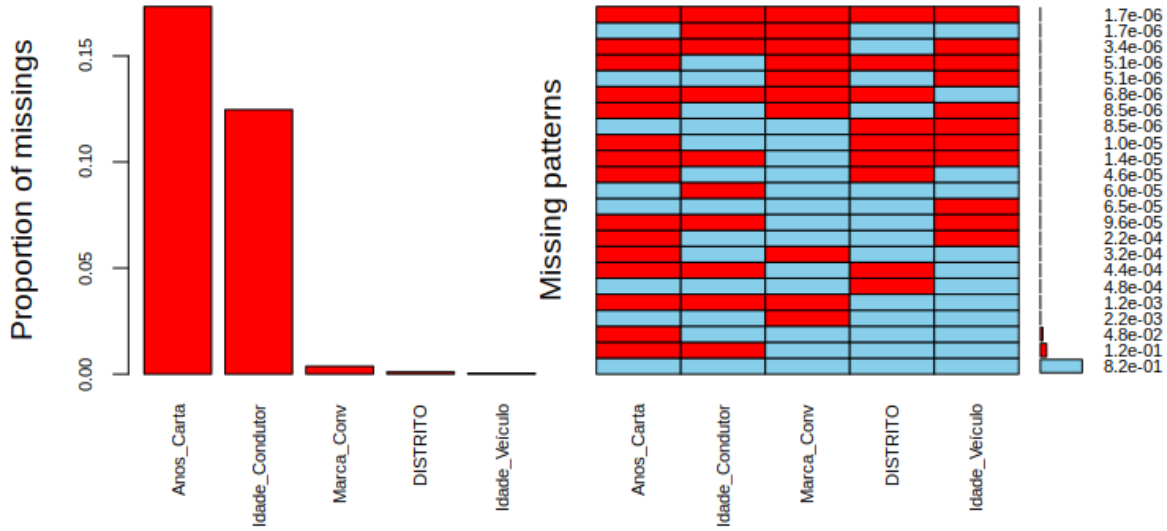

Figure 4.6: Observed missing proportions by several variables (left) and missing data patterns with corresponding proportions (right).

Multiple Imputation by Chained Equations (MICE) is a method of choice for incomplete data problems, typically used for MAR data (van Buuren, 2007). Initially, the variables with missing values are selected. Then, a univariate imputation model is specified for each of these variables, conditionally on the remaining variables. Starting from initially bootstrapped imputations, subsequent imputations are drawn by iterating over conditional densities. The `R` package `mice` provides functionality for multiple imputation, as well as diagnosis for imputed values (van Buuren and Groothuis-Oudshoorn, 2011).

Table 4.2: Percentages of missing values and imputation methods.

Variable	Missing Data (%)	Imputation Method
DISTRITO	9.84×10^{-2}	"polyreg"
CATEGORIA_AGREGADA	4.96×10^{-4}	"polyreg"
ESCALAO_CILINDRADA	2.00	"polyreg"
Idade_Veiculo	4.27×10^{-2}	"pmm"
Anos_Carta	17.14	"pmm"
Idade_Condutor	12.35	"pmm"
Marca_Conv	3.86×10^{-1}	"polyreg"

The original categories of each categorical covariate were kept for data imputation. Then, data reorganization was made following section 4.1.

Imputed data sets were generated by `mice` package. Since the data set comprises a large number of observations and covariates, `mice` is very slow, if a large number of imputed data sets and iterations is chosen. As such, these numbers would be increased only if the imputation results were very poor. van Buuren et al. [1999] suggested using `quickpred()` function to define the predictor matrix according to minimum thresholds for correlation (`mincor`) and proportion of usable cases (`minpuc`), which is useful to speed up `mice`. The default imputation methods for the variables were also considered. For categorical covariates with more than two levels and continuous covariates, polytomous regression (`method="polyreg"`) and predictive mean matching (`method="pmm"`) were selected, respectively (see table 4.2).

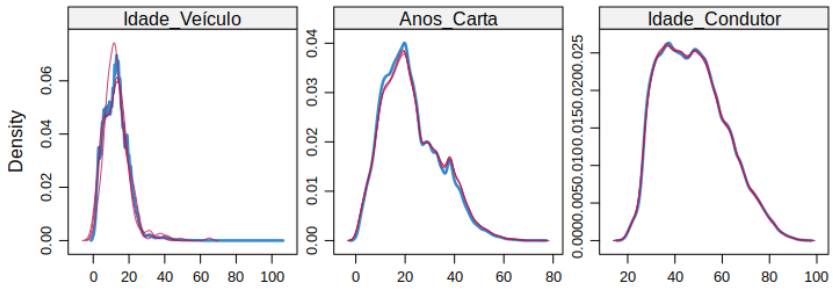

Figure 4.7: Density plots of the complete data (blue) versus imputed data (pink), using 3 imputed data sets, 5 iterations, minimum correlation 0.25 and minimum usable cases 0.25.

Figure 4.7 allows to do some diagnosis on the imputed values for continuous covariates. For `Anos_Carta` and `Idade_Condutor`, which are the covariates with largest proportions of missing values, the densities of imputed values resemble the respective densities for complete data. By increasing the number of imputed data sets and iterations, the densities with the imputed data sets did not improve significantly. Regardless of the imputed data set, the densities are similar to those of the complete data for each of these three covariates, which is a good indicator for using any of the imputed data sets.

Table 4.3: Observed proportions of categories of covariate `ESCALAO_CILINDRADA` for original and imputed data.

ESCALAO_CILINDRADA	Original data (%)	Imputed data (%)
1	51.42	50.77
2	43.65	44.36
3	4.93	4.87

For categorical covariates, the observed proportions of each category for original and imputed data

are essentially similar. For instance, table 4.3 shows those proportions for covariate ESCALAO_CILINDRADA do not differ much between the two data types.

Therefore, the imputed data set was used in order to replace the missing values in the original data set. In forthcoming analysis, the new data set is called complete data with imputation.

4.3 Model selection

In this section, model selection will be performed for both complete data (without NA) and complete data with imputation. For statistical model fitting, the first choice in terms of R packages was R2BayesX (Kneib et al., 2011), which is a R interface to execute the free software and standalone program BayesX (Belitz et al., 2015). This package provides functionality for making Bayesian mixed model inference on complex semiparametric regression models with structured additive predictor. Although this package was useful to fit Poisson GGAM or GGAMM in an early stage, it provides very limited framework on the specification and interpretation of zero-inflated GGAM or GGAMM.

Alternatively, R packages bamlss (Umlauf et al., 2017) and gamlss (Stasinopoulos and Rigby, 2019) provide more functionality to perform Bayesian inference for many more sampling distributions, in particular ZIP and ZINB distributions. However, it should be noted that these packages implement the prior distributions of smoothing variances τ_j^2, $j = 1, \ldots, p$ and τ_{str}^2 to be inverse gamma distributions with $a = b = 0.001$, thus not being possible to perform a sensitivity analysis for prior distributions. This is a limitation to be handled for future work.

4.3.1 Complete Data

According to other car insurance studies, such as Denuit and Lang [2004] and Klein et al. [2014], it was noticed that claim frequency was significantly influenced by nonlinear effects of continuous covariates. So, it was expected to obtain similar results for the current car insurance data, whose continuous covariates are Idade_Condutor (age of the policyholder), Idade_Veiculo (age of the vehicle), and Anos_Carta (license time), which were measured on policy purchase. Besides, the district in which the policy is registered (DISTRITO) provides important insights on the study of claim frequency, thus it will be used in the definition of risk profiles. Hence, basic models for each count data distribution are specified using the smooth effects of previous continuous covariates and structured spatial effects.

For model implementation, package bamlss was used with the default arguments. Thus, nonlinear covariate effects are represented via Bayesian cubic P-splines with 20 inner knots, where second order random walk priors were assigned for the corresponding splines coefficient parameters. For the spatial effects, simple Markov random field priors are used for correlated effects as in (3.16) and i.i.d. Gaussian priors for uncorrelated effects as in (3.18). For each fitted model, the fitting stage consists of running the

backfitting algorithm based on IWLS proposals (default maximum number of iterations is 400) in order to find posterior mode estimates. These are used as starting values of the sampler, with 1200 iterations and 200 iterations as the initial choices of sampling and burn-in periods, respectively.

A basic Poisson GGAM is here specified in terms of its additive predictor as follows

$$\begin{aligned}\mathsf{M}: \log \lambda_i \equiv \eta_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ + f_{\mathsf{str}}(\texttt{DISTRITO}_i)\end{aligned}$$

where $\mathsf{offset} = \log(\texttt{Apol\_20112013}/2.494536)$ is the offset term. This choice for the offset is appropriate because the risk period can vary from policy to policy, see e.g. Denuit et al [2004] for a similar approach. The value 2.494536 is the maximum risk period in this car insurance data.

Based on model M, variable selection was performed with the remaining covariates in a forward stepwise procedure, using the DIC and pD for model comparison. In an early stage, posterior mode estimates (calculated via backfitting algorithm) were chosen as initial values of the sampler. To check the performance of these estimates, package `bamlss` computes the corrected AIC (AICc), given by $\mathsf{AICc} = \mathsf{AIC} + \frac{2k^2+2k}{n-k-1}$, where $\mathsf{AIC} = -2\log(L) + 2k$ is the Akaike Information Criterion (Akaike, 1974), where L is the value of the likelihood, k the estimated number of parameters; and n is the sample size.

Table 4.4: Comparison of Poisson GGAM and GGAMM for the complete data.

Poisson Model	DIC	$\mathsf{p_D}$	AICc
M	309360.8	41.0013	309352.8
M1: $\eta = \eta_{\mathsf{M}} + \texttt{ESCALAO\_CILINDRADA}$	309014.3	43.5892	309003.5
M2: $\eta = \eta_{\mathsf{M}} + \texttt{Credor}$	309281.2	43.2569	309268.7
M3: $\eta = \eta_{\mathsf{M}} + \texttt{DESCR\_TIPO\_USO}$	309049.7	42.1491	309041.3
M4: $\eta = \eta_{\mathsf{M}} + \texttt{GARAGEM}$	309286	43.2781	309274.5
M5: $\eta = \eta_{\mathsf{M}} + \texttt{DESCR\_SEXO\_PESSOA}$	308992.8	41.896	308986.8
M6: $\eta = \eta_{\mathsf{M}} + \texttt{Marca\_Conv}$	308610.6	48.0171	308600.9
M7: $\eta = \eta_{\mathsf{M}} + \texttt{CATEGORIA\_AGREGADA}$	307014.5	41.5867	307009.7
M8: $\eta = \eta_{\mathsf{M}} + b$	309360.5	41.3844	309367

Table 4.4 shows the values of DIC, pD and AICc for the Poisson GGAM and GGAMM that result from model M adding the effects of the remaining covariates in the predictor, one at a time. In general, models M1 - M8 have a lower DIC than model M, which is more significantly reduced for models M1, M3, M5, M6 and M7. Note that the inclusion of uncorrelated spatial effects barely improves the DIC. As such, new model specifications were fitted, combining now the linear effects of `ESCALAO_CILINDRADA`, `DESCR_SEXO_PESSOA`, `DESCR_TIPO_USO`, `Marca_Conv` and `CATEGORIA_AGREGADA`. Interaction terms were also evaluated, but there were no significant improvements in these new fitting results.

The selected Poisson GGAM for the complete data is given by

$$\begin{aligned}\mathsf{M}_{sel}^{Po} : \log \lambda_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \textstyle\sum_{j=1}^{2} \beta_{1j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{2j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} \\ + \textstyle\sum_{j=1}^{6} \beta_{3j} \texttt{Marca\_Conv}_{ij} + \sum_{j=1}^{3} \beta_{4j} \texttt{CATEGORIA\_AGREGADA}_{ij}\end{aligned}$$

where DIC$= 306703$, pD$= 77.2948$. Also, the WAIC$= 310334.3$ and pW$= 89.06074$. In a final run, the number of iterations of the sampler was increased to 3000 and its burn-in to 800, but there were no significant differences.

In figure 4.8, a normal Q-Q plot of residuals of the selected Poisson GGAM after the final run are shown. Note that the sample quantiles greater than 2 are considerably large in comparison to the true (theoretical) quantiles based on the gaussian distribution. Clearly, the model performance is not very good at the tails of the data. A possible cause for this is the excess of zeros in the response.

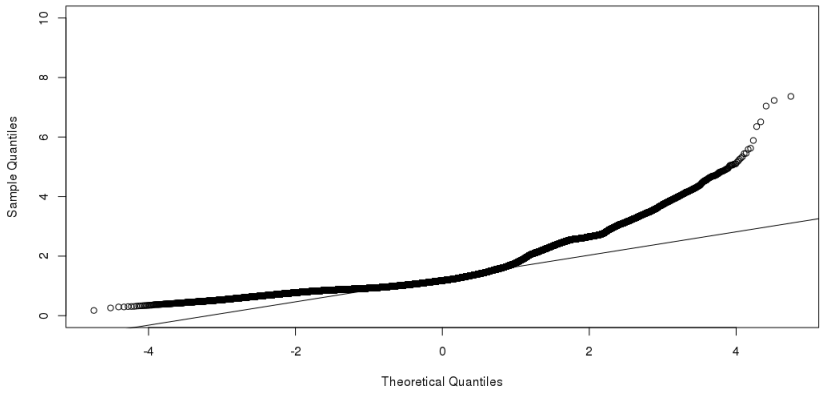

Figure 4.8: Normal Q-Q plot of residuals of the selected Poisson GGAM, M_{sel}^{Po}.

On an exploratory basis, even without a proper Bayesian interpretation, Zuur et al. [2012] proposed a dispersion measure of model. For instance, for model M_{sel}^{po} this measure can be estimated as follows

$$\frac{\mathsf{Deviance}(\mathsf{M}_{sel}^{po})}{\mathsf{Residual\ degrees\ of\ freedom}(\mathsf{M}_{sel}^{po})} = \frac{\sum_{i=1}^{n} r^2_{\mathsf{M}_{sel}^{Po},i}}{n - p_D} = \frac{976513.4}{477997 - 77.2948} \approx 2.043257$$

which is considered reasonably large and r_i denotes the crude residual for the i^{th} data point. Large model dispersion can be obtained due to unobserved risk factors, such as swiftness of reflexes, aggressiveness behind the wheel and consumption of drugs. It is also possible that the excess of zeros in claim frequency is increasing model dispersion.

The next model fitting step is to choose a response distribution that accommodates an excess of

zeros and check to what extent the model residuals and dispersion are improved. The ZIP distribution was the primary choice. As for the Poisson case, a basic ZIP GGAM was specified

$$\mathsf{M}^* : \begin{cases} \log \lambda_i \equiv \eta_i^{\lambda} = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ \qquad\qquad + f_{\mathsf{str}}(\texttt{DISTRITO}_i) \\ \mathsf{logit}(\pi_i) \equiv \eta_i^{\pi} = \gamma \end{cases}$$

For the probability of excess of zeros π, a simple intercept model was the initial choice. Model complexity was increased step by step using the DIC and pD, both for λ and π. From table 4.5 models from M1* to M8* have in general a lower DIC than model M*. Note that including the linear effects covariates `Credor` and `GARAGEM` does not lead to a relevant decrease in the DIC. The inclusion of uncorrelated spatial effects is also not relevant.

Thus, new model specifications were considered for the count predictor η^{μ}, combining the categorical covariates of models M1*, M3*, M5*, M6*, M7* and M7* in table 4.5. Interaction terms between covariates were also tested. The model with the lowest DIC (304436.1) includes covariates `ESCALAO_CILINDRADA`, `DESCR_TIPO_USO`, `Marca_Conv` and `CATEGORIA_AGREGADA` in η^{μ}. The respective model complexity pD is equal to 102.0202.

Table 4.5: Comparison of ZIP GGAM and GGAMM for the complete data.

ZIP Model	DIC	p_D	AICc
M*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307343.3	39.557	307346.3
M1*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{ESCALAO\_CILINDRADA}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307030	42.0156	307030.1
M2*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{Credor}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307269.3	39.9691	307272.3
M3*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{DESCR\_TIPO\_USO}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307062.5	38.4955	307068.7
M4*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{GARAGEM}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307278.6	41.5924	307280.8
M5*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{DESCR\_SEXO\_PESSOA}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307014	39.1487	307021.2
M6*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{Marca\_Conv}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	306664.5	46.3305	306665.3
M7*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + \texttt{CATEGORIA\_AGREGADA}$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	305178.8	39.5587	305186.2
M8*: $\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^*} + b$, $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^*}$	307347.3	41.5237	307350

Then, the model specification for π was considered. The inclusion of linear effects of covariates `Marca_Conv`, `Apol_20112013` and `CATEGORIA_AGREGADA` in the predictor of π leads to the most significant decrease in the DIC (304378.3), while model complexity remains similar (pD$= 105.0119$). Also, the

WAIC$= 306384.1$ and pW$= 119.6444$

A final run of the sampler was performed using 3000 iterations and burn-in equal to 800, with no significant changes in the fitting results. The selected ZIP GGAM for the complete data is given by

$$\mathsf{M}_{sel}^{zip}: \begin{cases} \log \lambda_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ \qquad\qquad + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{2} \beta_{1j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{2j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \qquad\qquad + \sum_{j=1}^{6} \beta_{3j} \texttt{Marca\_Conv}_{ij} + \sum_{j=1}^{3} \beta_{4j} \texttt{CATEGORIA\_AGREGADA}_{ij} \\ \mathsf{logit}(\pi_i) = \gamma_0 + \sum_{j=1}^{6} \gamma_{1j} \texttt{Marca\_Conv}_{ij} + \gamma_2 \texttt{Apol\_20112013}_i + \sum_{j=1}^{3} \gamma_{3j} \texttt{CATEGORIA\_AGREGADA}_{ij} \end{cases}$$

Although the selected ZIP GGAM has a lower DIC than the selected Poisson GGAM, there were no relevant improvements in the model residuals (see figure 4.9). Note that the model dispersion measure is still large, and possibly generated by a mechanism other than zero-inflation,

$$\frac{\sum_{i=1}^{n} r^2_{\mathsf{M}_{sel}^{zip},i}}{n - p_D} = \frac{1024564}{477997 - 105.0119} \approx 2.143924$$

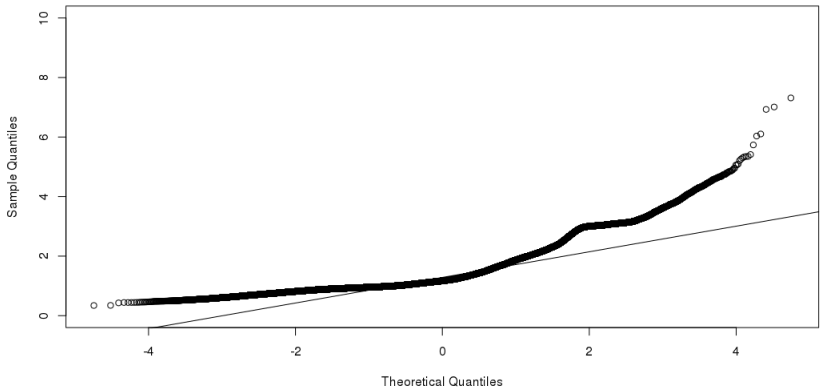

Figure 4.9: Normal Q-Q plot of residuals of the selected ZIP GGAM, M_{sel}^{zip}.

In order to attempt to reduce dispersion and improve model fitting results, the ZINB distribution was employed. Since ZINB model structure is more complex than the ZIP counterpart, the initial choice was

to consider a simple intercept model for both σ and π, given by

$$\mathsf{M}^{**}: \begin{cases} \log \mu_i \equiv \eta_i^{\mu} = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ \qquad\qquad + f_{\mathsf{str}}(\texttt{DISTRITO}_i) \\ \log \sigma_i \equiv \eta_i^{\sigma} = \alpha \\ \mathsf{logit}(\pi_i) \equiv \eta_i^{\pi} = \gamma \end{cases}$$

Again, variable selection was performed using the DIC and pD in a forward stepwise fashion to model μ. Table 4.6 provides the fitting results from model M^{**} adding the linear effects of categorical covariates in η^{μ}, while intercept models were considered for both σ and π. For most model specifications, the backfitting algorithm did not converge with the default number of iterations. In order to proceed, the selected model was fitted considering a larger number of iterations.

From table 4.6, the DIC is expressively reduced by adding the linear effects of `Marca_Conv` and `CATEGORIA_AGREGADA` in η^{μ}, much more than with the other categorical covariates. As for the Poisson and ZIP cases, the inclusion of covariates `Credor` and `GARAGEM` is not significant, as well as the uncorrelated spatial effect. The specification of μ that led to a lower DIC contains the effects of covariates `CATEGORIA_AGREGADA`, `Marca_Conv`, `DESCR_TIPO_USO`, `ESCALAO_CILINDRADA` and `DESCR_SEXO_PESSOA`. An interation term between `Idade_Condutor` and `DESCR_SEXO_PESSOA` was evaluated in η^{μ}, conducting to relevant improvements in the model fitting results. For η^{σ}, the choice was to include the same covariate effects of η^{μ}. For η^{π}, simple intercept and univariate models were fitted. In this case, the inclusion of `Apol_20112013` significantly improved the DIC.

The selected ZINB GGAM for the complete data is given by

$$\mathsf{M}_{sel}^{zinb}: \begin{cases} \log \mu_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + \sum_{j=1}^{2} \beta_{1j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} * f_2(\texttt{Idade\_Condutor}_i) \\ \quad + f_3(\texttt{Anos\_Carta}_i) + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{3} \beta_{2j} \texttt{CATEGORIA\_AGREGADA}_{ij} + \sum_{j=1}^{6} \beta_{3j} \texttt{Marca\_Conv}_{ij} \\ \quad + \sum_{j=1}^{2} \beta_{4j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{2} \beta_{5j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{6j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \log \delta_i = \mathsf{offset}_i + \alpha_0 + g_1(\texttt{Idade\_Veiculo}_i) + g_2(\texttt{Idade\_Condutor}_i) + g_3(\texttt{Anos\_Carta}_i) \\ \quad + g_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{3} \alpha_{1j} \texttt{CATEGORIA\_AGREGADA}_{ij} + \sum_{j=1}^{6} \alpha_{2j} \texttt{Marca\_Conv}_{ij} \\ \quad + \sum_{j=1}^{2} \alpha_{3j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{2} \alpha_{4j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \alpha_{5j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \mathsf{logit}(\pi_i) \equiv \eta_i^{\pi} = \gamma_0 + \gamma_1 \texttt{Apol\_20112013}_i \end{cases}$$

where DIC$= 303744.4$, pD$= 123.6081$, and WAIC$= 305543.6$, pW$= 214.6695$. In terms of DIC, this model is clearly better than the selected ZIP GGAM. The model fitting performance also seems to be improved at the tails of the data (see figure 4.10). As such, the selected ZINB GGAM is the best choice for the complete data, based on DIC and WAIC. In a final run, the number of iterations was increased, which

Table 4.6: Comparison of ZINB GGAM and GGAMM for the complete data.

ZINB Model	DIC	p_D	AICc
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}}$ M**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	307203.6	39.3307	307218.9
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{ESCALAO\_CILINDRADA}$ M1**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	306892	39.5051	306902.7
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{*}} + \texttt{Credor}$ M2**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{*}}$	307133.9	41.4714	307142.8
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{DESCR\_TIPO\_USO}$ M3**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	306936.4	41.2559	306944.7
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{GARAGEM}$ M4**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	307137.5	40.0539	307150.1
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{DESCR\_SEXO\_PESSOA}$ M5**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	306881.9	39.8941	306896.2
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{Marca\_Conv}$ M6**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	306080.5	71.7819	306089
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + \texttt{CATEGORIA\_AGREGADA}$ M7**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	305047.5	40.0193	305061.1
$\eta^{\lambda} = \eta^{\lambda}_{\mathsf{M}^{**}} + b$ M8**: $\eta^{\sigma} = \eta^{\sigma}_{\mathsf{M}^{**}}$ $\eta^{\pi} = \eta^{\pi}_{\mathsf{M}^{**}}$	307211.3	40.1864	307200.1

resulted in no relevant changes in the fitting results.

Then, the corresponding covariate effects in the predictors of μ, σ and π were inspected. In figure 4.11, the estimated nonlinear covariate functions on the mean μ are plotted together with 95% pointwise credible intervals. In figure 4.11 (a) it is clear that as the license time increases, a smaller number of accidents is expected, and it eventually appears to stabilise for large license time values.

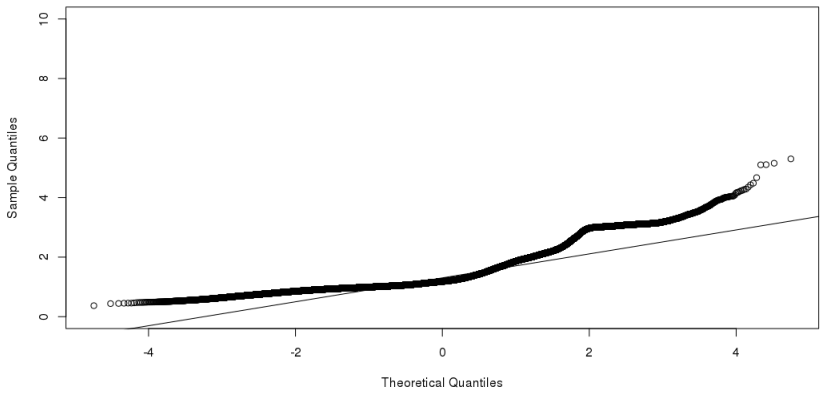

Figure 4.10: Normal Q-Q plot of residuals of the selected ZINB GGAM for the complete data, M^{zinb}_{sel}.

However, the effect increases for license time greater than 40. This is possibly originated by larger

values of license time being related to old policyholders, for whom the risk increases.

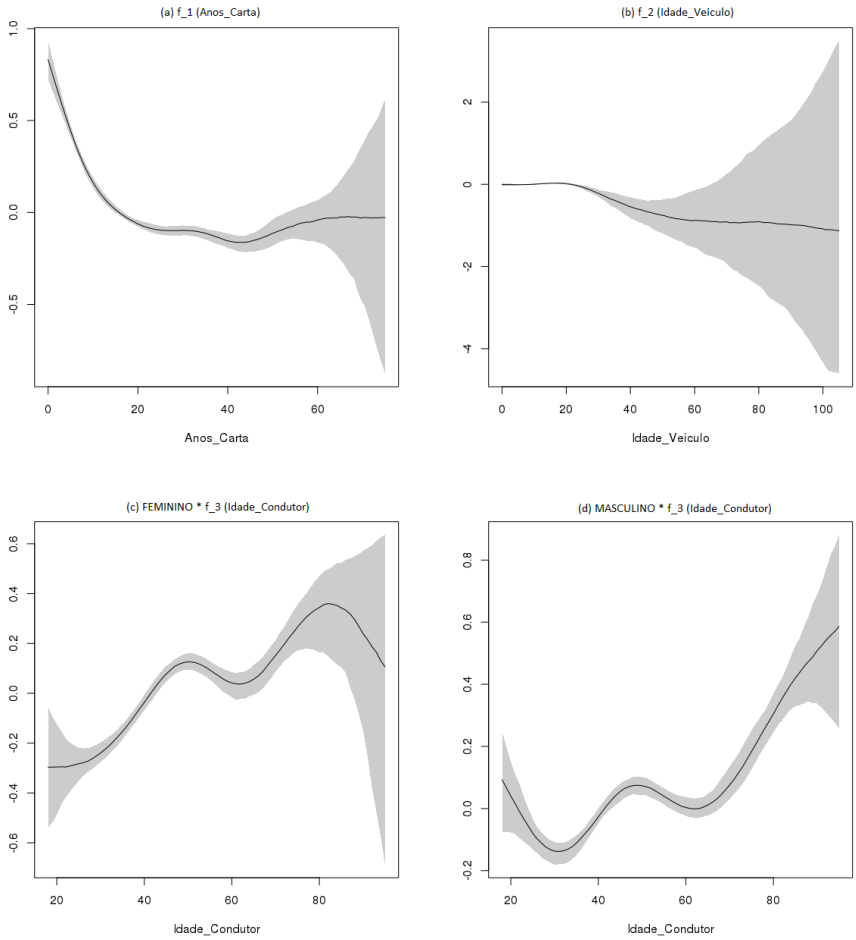

Figure 4.11: Estimated nonlinear covariate functions involved in η^{μ} of the selected model for the complete data. Together are shown the 95% pointwise credible intervals.

In figure 4.11 (b) it is suggested that vehicles aged 20 or lower have a negligible effect on the score of the predictor. From that point onward, the effect decreases gradually. This fact indicates that older cars are less risky than new cars. However, note that for larger values of age of the vehicle (namely greater than 40), the 95% pointwise credible intervals are very wide. This is possibly due to a small number of observations in that case. Figures 4.11 (c) and (d) show the interaction between gender (female and male, respectively) and age of the policyholder. It can be seen that young female policyholders have a lower effect on the predictor than young male policyholders. As the latter become more experienced, the effect gradually decreases. From the age of 30 to 50, the effect increases significantly for both female and male policyholders. A possible reason for this is that, because of extremely high premiums charged to young policyholders, they ask older relatives (for instance, parents) to purchase the policy. As expected, that effect increases again for old policyholders, being more significant for males. The 95% credible

intervals become very wide after the age of 90, which is given to a low number of observations in that case. The interaction between age of the policyholder and the undefined gender was not analysed since it may include both male and female policyholders, thus leading to unclear interpretation.

In figure 4.12 the estimated nonlinear covariate functions on σ are plotted. In this case, the effects on η^{σ} are close to linear. The smooth functions of age of the vehicle and license time are increasing, while the smooth function of age of the policyholder is approximately constant. This means that as age of the policyholder and license time increase, the contributions for dispersion also increase. Note that the 95% pointwise credible intervals are wide, especially at the upper extremes.

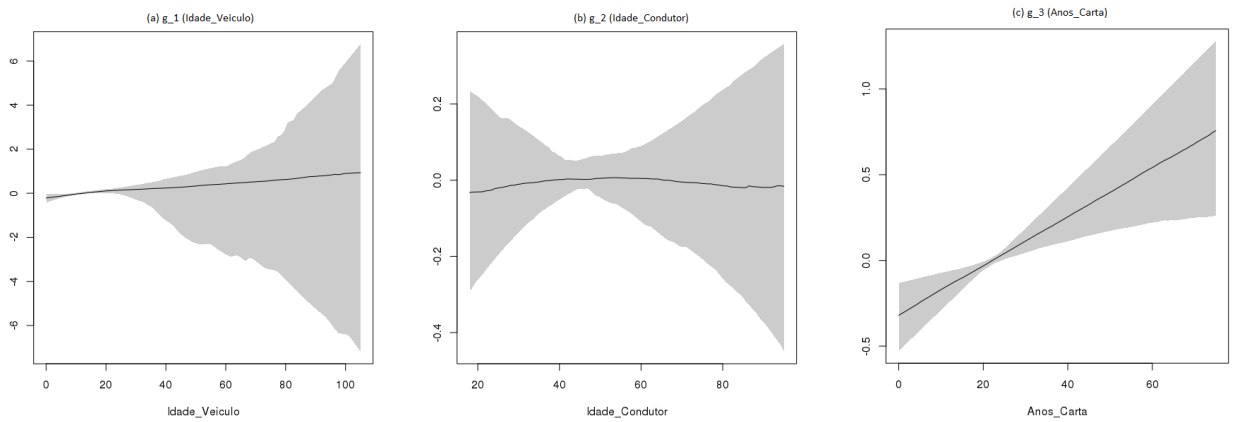

Figure 4.12: Estimated nonlinear covariate functions involved in η^{σ} of the selected model for the complete data. Together are shown the 95% pointwise credible intervals.

The selected ZINB GGAM was initially run with basis dimension for smooth functions equal to 20 both in η^{μ} and η^{σ}. By changing the basis dimension to 15 and 30, the fitting results changed significantly (see table 4.7). Since the fitting results were not improved, the default basis dimension equal to 20 was considered.

Table 4.7: Model fitting results for different basis dimensions of smooth functions.

k	DIC	pD	AICc
15	303869.8	144.0794	303891
20	303744.4	123.6081	303778.5
30	303777.1	157.7543	303810.8
35	303777.5	161.4627	303814.3

Figure 4.13 depicts the estimated correlated spatial effects $f_{str}(\texttt{DISTRITO})$ and $g_{str}(\texttt{DISTRITO})$ on the log-mean μ and σ, respectively. The districts Lisboa, Porto, Braga and Setúbal are the ones with larger spatial effects on the mean μ. That can be due to these districts being the main urban areas in Portugal mainland. On the contrary, living in the countryside reduces the correlated spatial effects, such

as in Portalegre and Bragança. In addition, it is suggested that the northern districts have lower effects on σ than the southern districts. Since σ is the dispersion parameter, it means that southern districts have larger contributions for spatial dispersion.

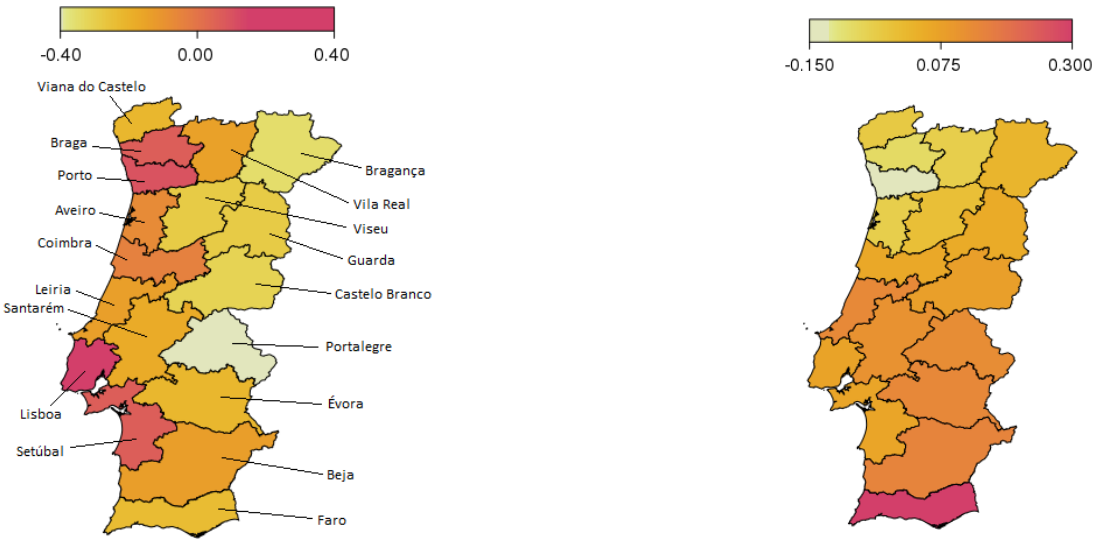

Figure 4.13: Estimated correlated spatial effect in η^{μ} (left) and η^{σ} (right) of the selected model.

Table 4.8 shows the estimated regression coefficients on the log-mean μ. It should be noted that most of the effects related to `CATEGORIA_AGREGADA` are omitted in the output. This is due to its 95% pointwise credible intervals including zero. The effects related to `Marca_Conv` are also in general small in magnitude and do not have a heavy contribution on the predictor. In addition, having a powerful car (engine displacement in class 3) increases the risk, while private use for the car has an opposite effect.

Table 4.9 shows the estimated regression coefficients on $\log \sigma$. Again, most of the effects related to `CATEGORIA_AGREGADA` are omitted and the regression coeffcients associated with `Marca_Conv` are small in magnitude. In addition, having a powerful (engine displacement class 3) car and private life usage have relevant positive contributions for dispersion.

By inspecting the range of values for covariates `Anos_Carta`, `Idade_Veiculo` and `Idade_Condutor`, there could be some atypical observations. Looking back at figure 4.11 (b), the effects on $\log \mu$ decrease for policyholders aged 90 or more, with wide 95% credible intervals for those values. Possibly, very old policyholders have paid insurance premiums, but they did not use the vehicle very often. Some attention should also be paid to policyholders with license time values greater than 70, which are associated with very old policyholders. Moreover, old cars are more prone to failures and demand more expenses in general, and hence their policyholders tend to buy a new car to replace the old one. It is possible that

Table 4.8: Estimated regression coefficients of η^{μ} in the final model (M_{sel}^{zinb}) for the complete data.

Parameter	Mean	2.5% quant.	Median	97.5% quant.
β_0	−1.4299	−1.4813	−1.4364	−1.3565
CATEGORIA_AGREGADA (MIX)	0.1724	0.1440	0.1731	0.1978
CATEGORIA_AGREGADA (MOTO)	−0.8272	−0.9074	−0.8267	−0.7353
CATEGORIA_AGREGADA (OTHER)	−1.6863	−1.8543	−1.6838	−1.5249
Marca_Conv (FRA)	0.0053	−0.0191	0.0034	0.0448
Marca_Conv (JAP)	−0.0031	−0.0410	−0.0027	0.0350
Marca_Conv (ITA)	0.0718	0.0418	0.0725	0.1031
Marca_Conv (USA)	0.0540	0.0123	0.0544	0.0870
Marca_Conv (OTHER)	0.0113	−0.0497	−0.0104	0.0241
Marca_Conv (NS)	−0.0184	−0.0776	−0.0189	0.0409
DESCR_SEXO_PESSOA (F)	0.1303	0.0378	0.1310	0.2263
DESCR_SEXO_PESSOA (M)	0.0859	−0.0015	0.0838	0.1792
ESCALAO_CILINDRADA (2)	0.0620	0.0427	0.0628	0.0793
ESCALAO_CILINDRADA (3)	0.2078	0.1486	0.2086	0.2580
DESCR_TIPO_USO (PRIV)	−0.3103	−0.3915	−0.3090	−0.2274

Table 4.9: Estimated regression coefficients of η^{σ} in the final model (M_{sel}^{zinb}) for the complete data.

Parameter	Mean	2.5% quant.	Median	97.5% quant.
β_0	−1.1284	−1.3709	−1.1178	−0.9207
CATEGORIA_AGREGADA (MIX)	−0.0858	−0.3335	−0.0893	0.2055
CATEGORIA_AGREGADA (MOTO)	0.3208	−0.43323	0.34133	0.9984
CATEGORIA_AGREGADA (OTHER)	2.3186	1.7865	2.3000	2.9549
Marca_Conv (FRA)	−0.1690	−0.3731	−0.1573	0.0129
Marca_Conv (JAP)	−0.0434	−0.2811	−0.0399	0.1629
Marca_Conv (ITA)	−0.0515	−0.3282	−0.0520	0.2145
Marca_Conv (USA)	0.1316	−0.1644	0.1439	0.3898
Marca_Conv (OTHER)	−0.1720	−0.4181	−0.1779	0.1260
Marca_Conv (NS)	0.2520	−0.1404	0.2530	0.6348
DESCR_SEXO_PESSOA (F)	−0.9826	−2.3476	−0.8758	−0.1956
DESCR_SEXO_PESSOA (M)	−0.8595	−2.2677	−0.7428	−0.1014
ESCALAO_CILINDRADA (2)	−0.0434	−0.2020	−0.0360	0.1014
ESCALAO_CILINDRADA (3)	0.3414	−0.0632	0.3585	0.6630
DESCR_TIPO_USO (PRIV)	0.7840	−0.0348	0.6375	2.3510

such large values for age of the vehicle are due to fill-in errors during the collection of the data set.

Therefore, the selected model specification was run for the complete data without possible atypical observations. Although the DIC (303437.5) and model complexity (pD$= 117.8334$) were decreased, the model residuals did not differ much from the ones of the selected model for the complete data (see figure 4.14). In section 4.4, the selected model for complete data without possible atypical observations will also be used for prediction and compared to the selected ZINB GGAM for the complete data.

4.3.2 Complete Data with Imputation

In this subsection, model selection is analogous to that of subsection 4.3.1. That is, a Poisson GGAM or GGAMM will be selected, starting with the basic Poisson model corresponds to model M used previously. Variable selection was performed in a forward stepwise fashion, using the DIC and pD.

From table 4.10, only models M1, M3, M5, M6 and M7 have the DIC significantly reduced. The

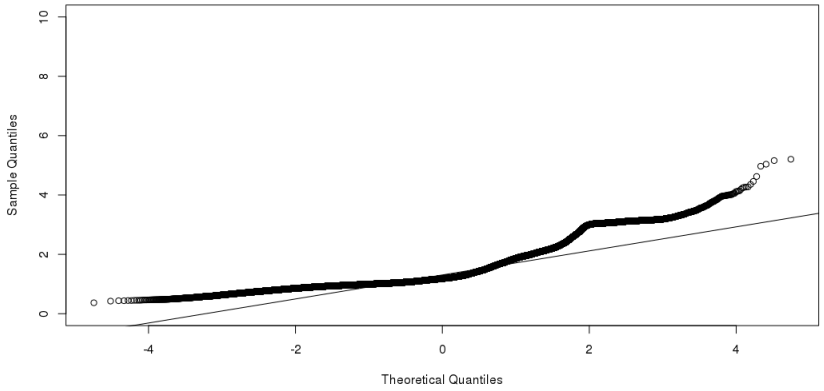

Figure 4.14: Normal Q-Q plot of residuals of the selected ZINB GGAM after removing possible outliers.

linear effects of the respective covariates were combined in the predictor, including an interaction term between covariates DESCR_SEXO_PESSOA and Idade_Condutor.

Table 4.10: Comparison of Poisson GGAM and GGAMM for the complete data with imputation.

Poisson Model	DIC	p_D	AICc
M	383431.2	32.8496	383410.4
M1: $\eta = \eta_{\mathsf{M}} +$ ESCALAO_CILINDRADA	382682.8	43.4899	382659.5
M2: $\eta = \eta_{\mathsf{M}} +$ Credor	383267.4	40.7303	383249.5
M3: $\eta = \eta_{\mathsf{M}} +$ DESCR_TIPO_USO	382993.6	46.6184	382966.8
M4: $\eta = \eta_{\mathsf{M}} +$ GARAGEM	383361.4	43.0186	383341.3
M5: $\eta = \eta_{\mathsf{M}} +$ DESCR_SEXO_PESSOA	383002.2	45.4208	382979.1
M6: $\eta = \eta_{\mathsf{M}} +$ Marca_Conv	382180.1	44.0705	382166.8
M7: $\eta = \eta_{\mathsf{M}} +$ CATEGORICA_AGREGADA	379114.9	46.9612	379107.9
M8: $\eta = \eta_{\mathsf{M}} + b$	383430.6	42.8244	383428.3

The selected Poisson GGAM for the complete data with imputation is given by

$$\begin{aligned}\mathsf{M}_{sel}^{po} : \log \lambda_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \textstyle\sum_{j=1}^{2} \beta_{1j}\texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{2j}\texttt{DESCR\_TIPO\_USO}_{ij} \\ + \textstyle\sum_{j=1}^{2} \beta_{3j}\texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{6} \beta_{4j}\texttt{Marca\_Conv}_{ij} + \sum_{j=1}^{3} \beta_{5j}\texttt{CATEGORIA\_AGREGADA}_{ij}\end{aligned}$$

where DIC$= 378702.1$ and pD$= 61.998$. Also, the WAIC$= 382935.3$ and pW$= 63.28057$. The dispersion of model M_{sel}^{po} was estimated to be

$$\frac{\sum_{i=1}^{n} r^2_{\mathsf{M}^{Po}_{sel},i}}{n - p_D} = \frac{1024690}{586256 - 61.998} \approx 1.751026$$

which is lower than the dispersion of Poisson GGAM for complete data. Nevertheless, this value is still reasonably large. In addition, figure 4.15 indicates that the model performance at the tails of the data could possibly be improved.

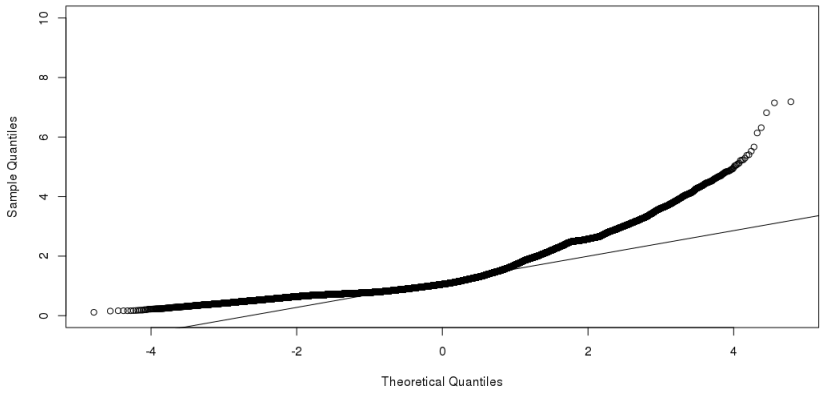

Figure 4.15: Normal Q-Q plot of residuals of the selected Poisson GGAM, M_{sel}^{Po}.

The next model fitting step is to fit a ZIP GGAM or GGAMM in order to accommodate the excess of zeros in the response (see figure 4.5), and check to what extent model dispersion can be reduced. The basic ZIP GGAM corresponds to model M^* given in subsection 4.3.1. Variable selection is performed again in a forward stepwise procedure to model both λ and π.

A simple intercept model was the first choice for π. From table 4.11, only models $M1^*$, $M3^*$, $M5^*$, $M6^*$ and $M7^*$ had the DIC significantly reduced. Thus, the effects of covariates `ESCALAO_CILINDRADA`, `DESCR_TIPO_USO`, `Marca_Conv`, `DESCR_SEXO_PESSOA` and `CATEGORIA_AGREGADA` were combined in η^{μ}. The model with lowest DIC was obtained by including the effects of all the covariates. An interaction term between `Idade_Condutor` and `DESCR_SEXO_PESSOA` was also assessed but the model fitting results were not improved.

The inclusion of covariate effects η^{π} was also evaluated. As in subsection 4.3.1, including effects of covariates `Marca_Conv`, `Apol_20112013` and `CATEGORIA_AGREGADA` significantly decreases the DIC (375554.9), while the pD (54.9107) increases but not enough. As such, the selected ZIP GGAM for the complete data with imputation is given by

Table 4.11: Comparison of ZIP GGAM and GGAMM for the complete data with imputation.

ZIP Model	DIC	p_D	AICc
M*: $\eta^{\lambda} = \eta^{\lambda}_{M^*}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380499.2	40.4522	380493.8
M1*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{ESCALAO\_CILINDRADA}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	379836.8	43.659	379824.5
M2*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{Credor}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380355.3	38.5142	380354
M3*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{DESCR\_TIPO\_USO}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380103.3	40.134	380103.3
M4*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{GARAGEM}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380437.5	40.1107	380435.1
M5*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{DESCR\_SEXO\_PESSOA}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380119.5	43.0905	380113.1
M6*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{Marca\_Conv}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	379387.7	46.1042	379386.3
M7*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + \texttt{CATEGORIA\_AGREGADA}$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	376527	46.8088	376530.6
M8*: $\eta^{\lambda} = \eta^{\lambda}_{M^*} + b$, $\eta^{\pi} = \eta^{\pi}_{M^*}$	380506.2	43.4641	380512.1

$$
\mathsf{M}^{zip}_{sel}: \begin{cases} \log \lambda_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + f_2(\texttt{Anos\_Carta}_i) + f_3(\texttt{Idade\_Condutor}_i) \\ \quad + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{2} \beta_{1j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{2j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \quad + \sum_{j=1}^{2} \beta_{3j} \texttt{Marca\_Conv}_{ij} + \sum_{j=1}^{2} \beta_{4j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{3} \beta_{5j} \texttt{CATEGORIA\_AGREGADA}_{ij} \\ \mathsf{logit}(\pi_i) = \gamma_0 + \sum_{j=1}^{6} \gamma_{1j} \texttt{Marca\_Conv}_{ij} + \gamma_2 \texttt{Apol\_20112013}_i + \sum_{j=1}^{3} \texttt{CATEGORIA\_AGREGADA}_{ij} \gamma_{3j} \end{cases}
$$

also with WAIC$= 377907.7$ and pW$= 60.03601$.

The normal Q-Q plot of residuals also shows that the model performance at the tails is still poor (see figure 4.18).

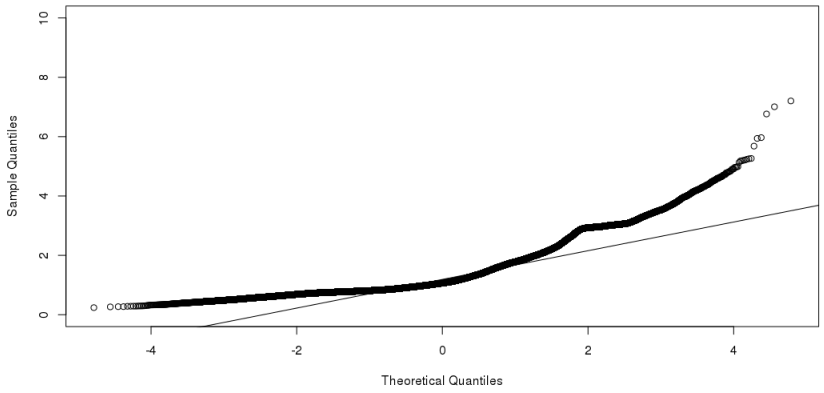

Figure 4.16: Normal Q-Q plot of residuals plot of the selected ZIP GGAM, M^{zip}_{sel}.

The dispersion of model M_{sel}^{zip} was estimated to be

$$\frac{\sum_{i=1}^{n} r^2_{\mathsf{M}_{sel}^{zip},i}}{n - p_D} = \frac{1105237}{585256 - 54.9107} \approx 1.888645$$

indicating that the model dispersion was not reduced, and is possibly driven by a mechanism other than zero-inflation. However, it should be noted that the selected Poisson and ZIP GGAM are less dispersed than the respective selected models in subsection 4.3.1. In order to attempt to reduce dispersion and improve the fitting results, the ZINB distribution was considered. The basic ZINB model corresponds to model M^{**} in the previous subsection.

Table 4.12 shows that including the effects of covariates ESCALAO_CILINDRADA, DESCR_TIPO_USO, DESCR_SEXO_PESSOA, Marca_Conv and CATEGORIA_AGREGADA in η^{μ} led to a more considerable decrease of the DIC. As in subsection 4.3.1, there were some convergence issues regarding the backfitting algorithm. The lowest DIC was obtained by including the effects of all the covariates, and also an interaction term between DESCR_SEXO_PESSOA and Idade_Condutor. For η^{σ}, the same covariate effects as in η^{μ} were included. For η^{π}, simple intercept and univariate models were evaluated, with a significant improvement of the DIC being achieved with the inclusion of linear effect of Apol_20112013.

Therefore, the selected ZINB GGAM for the complete data with imputation is given by

$$\mathsf{M}_{sel}^{zinb} : \begin{cases} \log \mu_i = \mathsf{offset}_i + \beta_0 + f_1(\texttt{Idade\_Veiculo}_i) + \sum_{j=1}^{2} \beta_{1j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} * f_2(\texttt{Idade\_Condutor}_i) \\ \quad + f_3(\texttt{Anos\_Carta}_i) + f_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{3} \beta_{2j} \texttt{CATEGORIA\_AGREGADA}_{ij} + \sum_{j=1}^{6} \beta_{3j} \texttt{Marca\_Conv}_{ij} \\ \quad + \sum_{j=1}^{2} \beta_{4j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{2} \beta_{5j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{6j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \log \delta_i = \mathsf{offset}_i + \alpha_0 + g_1(\texttt{Idade\_Veiculo}_i) + g_2(\texttt{Idade\_Condutor}_i) + g_3(\texttt{Anos\_Carta}_i) \\ \quad + g_{\mathsf{str}}(\texttt{DISTRITO}_i) + \sum_{j=1}^{3} \alpha_{1j} \texttt{CATEGORIA\_AGREGADA}_{ij} + \sum_{j=1}^{6} \alpha_{2j} \texttt{Marca\_Conv}_{ij} \\ \quad + \sum_{j=1}^{2} \alpha_{3j} \texttt{DESCR\_SEXO\_PESSOA}_{ij} + \sum_{j=1}^{2} \alpha_{4j} \texttt{ESCALAO\_CILINDRADA}_{ij} + \sum_{j=1}^{2} \beta_{5j} \texttt{DESCR\_TIPO\_USO}_{ij} \\ \mathsf{logit}(\pi_i) = \gamma_0 + \gamma_1 \texttt{Apol\_20112013}_i \end{cases}$$

where DIC$= 374746.7$ and pD$= 96.1305$. Also, the WAIC$= 377277.4$ and pW$= 202.4251$. The selected ZINB GGAM improves the DIC and WAIC in comparison to the ZIP GGAM. The model residuals are slightly improved, but not significantly (see figure 4.18). Thus, the selected ZINB GGAM is chosen.

According to figure 4.17, the estimated nonlinear covariate functions on the mean μ are similar to those of the selected ZINB GGAM for the complete data. The estimated nonlinear effects on σ are again close to linear. The major difference now is that the effect of age of the policyholder is increasing (see

Table 4.12: Comparison of ZINB GGAM and GGAMM for the imputed data set.

ZINB Model	DIC	p_D	AICc
M**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	380286.9	37.4694	380297.1
M1**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + \texttt{ESCALAO\_CILINDRADA}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	379628.5	41.9975	379628.3
M2**: $\eta^{\lambda} = \eta^{\lambda}_{M^{*}} + \texttt{Credor}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{*}}$	380148	40.9366	380158.1
M3**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + \texttt{DESCR\_TIPO\_USO}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	379903.8	40.9571	379914.9
M4**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + \texttt{GARAGEM}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	380228.1	40.1069	380236.3
M5**: $\eta^{\lambda} = \eta^{\lambda}_{M^{*}} + \texttt{DESCR\_SEXO\_PESSOA}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	379909.5	40.6314	379924
M6**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + \texttt{Marca\_Conv}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	379172	48.1627	379178.9
M7**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + \texttt{CATEGORIA\_AGREGADA}$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	376317.9	46.655	376340.9
M8**: $\eta^{\lambda} = \eta^{\lambda}_{M^{**}} + b$, $\eta^{\delta} = \eta^{\delta}_{M^{**}}$, $\eta^{\pi} = \eta^{\pi}_{M^{**}}$	380294.6	42.8132	380299.1

figure 4.19). The default basis dimension (equal to 20) to represent the smooth functions for both μ and σ was again the initial choice. From table 4.13, changing the basis dimension to 15 and 30 did not lead to significant changes in the fitting results. This seems to be the case for the complete data with imputation. Since the pD was smaller for $k = 20$, this was chosen.

Table 4.13: Model fitting results for different basis dimensions of smooth functions.

k	DIC	pD	AICc
15	374739.2	102.3919	374769
20	374746.7	96.1305	374771.9
30	374724.5	100.4538	374774.8
35	374728.8	102.4451	374777.1

From figure 4.20, Lisboa, Porto, Braga and Setúbal are the districts with higher contributions on the mean μ. The spatial effect decreases smoothly from the main urban areas to the countryside, which was also noticed for the complete data. For the dispersion σ, the northern districts have lower effects than the southern districts, i.e. the latter districts have larger contributions for dispersion.

From table 4.14, some differences are noted for the estimated linear regression coefficients on η^{μ}. In contrast with table 4.8, most of the effects in `CATEGORIA_AGREGADA` are relevant. For `Marca_Conv`, the estimated regression coefficients are again small in magnitude. In addition, having a powerful car

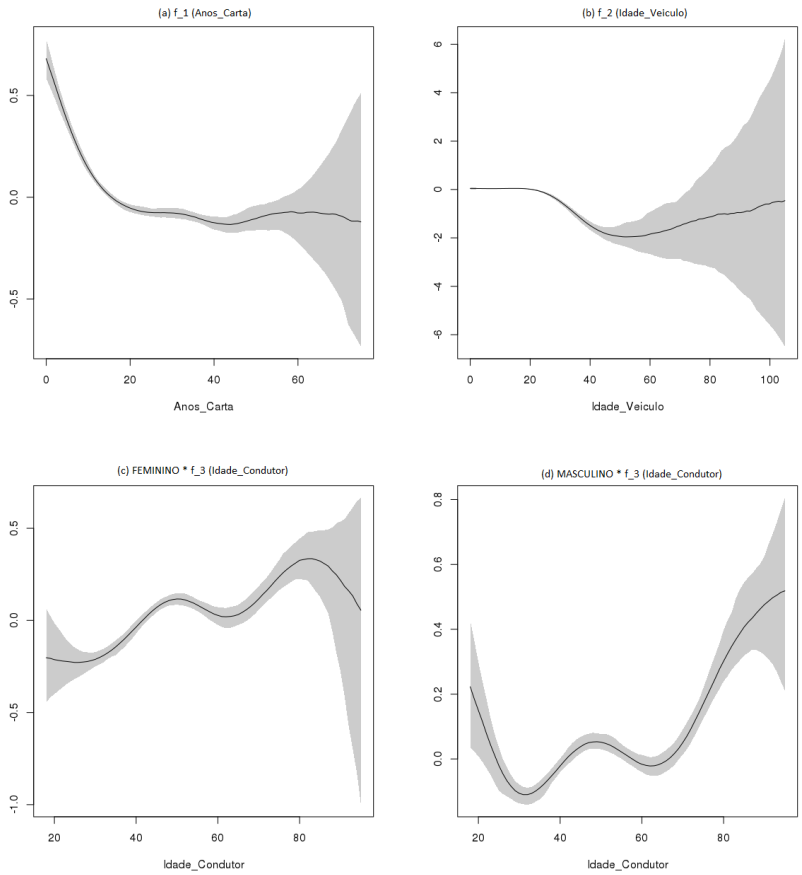

Figure 4.17: Estimated nonlinear covariate functions involved in η^{μ} of the selected model for the complete data with imputation. Together are shown 95% pointwise confidence intervals.

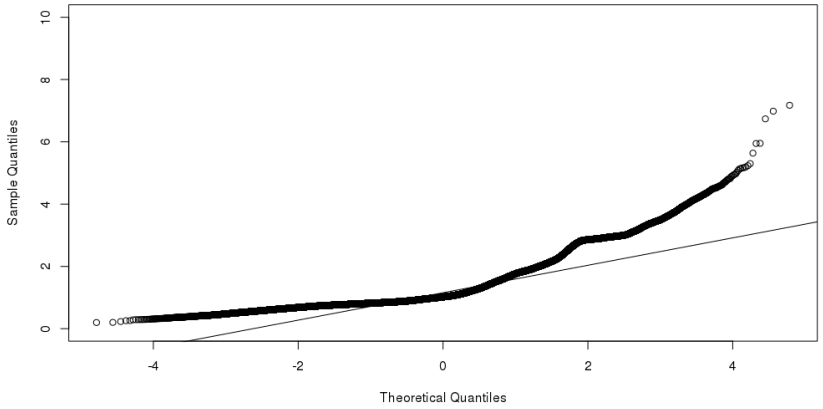

Figure 4.18: Normal Q-Q of residuals of the selected ZINB GGAM for the complete data with imputation.

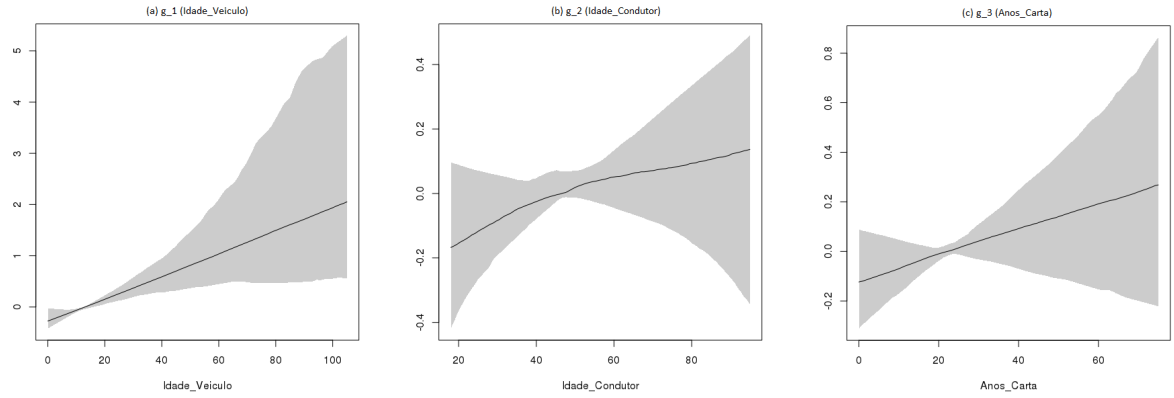

Figure 4.19: Estimated nonlinear covariate functions involved in η^{σ} of the selected model for the complete data with imputation. Together are shown the 95% pointwise confidence intervals.

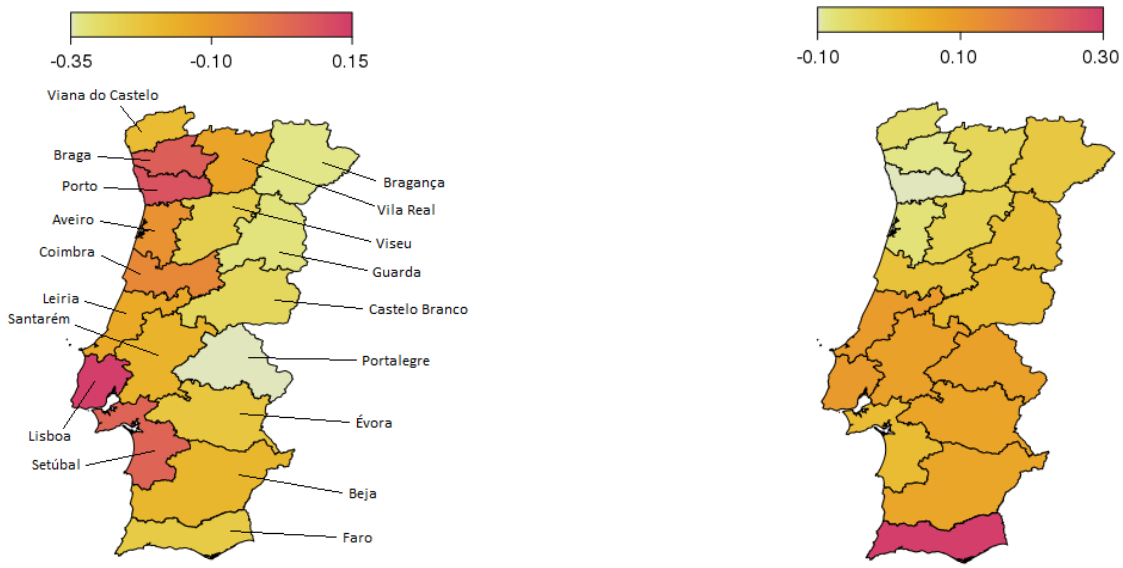

Figure 4.20: Estimated correlated spatial effect in η^{μ} (left) and η^{σ} (right) of the selected model for the complete data with imputation.

(engine displacement in class 3) increases the risk, while private life use has an opposite effect on the risk, which was also the case in the previous subsection.

From table 4.15, now most of effects of `CATEGORIA_AGREGADA` are relevant. In particular, categories `"HEAVY_PASS"` and `"TAXI"` have very small contributions for dispersion in comparison with other categories. For `Marca_Conv` the estimated regression coefficients are in general small in magnitude. In addition, being a male has a larger contribution for dispersion than being a female. In contrast with table 4.9, having a powerful car and private life usage do not have large contributions for dispersion.

A ZINB GGAM was fitted again using the imputed data set without possible atypical observations, similarly as in subsection 4.3.1. The model residuals are very similar to those of the selected model (see

Table 4.14: Estimated regression coefficients of the η^{μ} in the final model.

Parameter	Mean	2.5% quant.	Median	97.5% quant.
β_0	-1.1246	-1.3033	-1.1296	-0.9240
CATEGORIA_AGREGADA (DIVERSE)	-1.2612	-1.4671	-1.2618	-0.9402
CATEGORIA_AGREGADA (LIGHT)	-0.4165	-0.5760	-0.4158	-1.0972
CATEGORIA_AGREGADA (MIX)	-0.2430	-0.4061	-0.2382	-0.2483
CATEGORIA_AGREGADA (MOTO)	-1.2745	-1.4253	-1.2700	-0.0788
CATEGORIA_AGREGADA (HEAVY_MERC)	-0.3193	-0.4875	-0.3195	-1.0634
CATEGORIA_AGREGADA (HEAVY_PASS)	-0.6374	-1.0028	-0.6487	-0.1162
CATEGORIA_AGREGADA (TRAILER)	-4.8715	-5.5227	-4.8493	-0.249367
CATEGORIA_AGREGADA (TAXI)	0.9991	-0.2571	0.9827	2.2280
Marca_Conv (FRA)	-0.0023	-0.0274	-0.0023	0.0190
Marca_Conv (JAP)	-0.0083	-0.0358	-0.0083	0.0232
Marca_Conv (ITA)	0.0593	0.0218	0.0577	0.0955
Marca_Conv (USA)	0.0396	-0.0223	0.0408	0.0776
Marca_Conv (OTHER)	-0.0123	-0.0412	-0.0127	0.0182
Marca_Conv (NS)	-0.0737	-0.1220	-0.0745	-0.0309
DESCR_SEXO_PESSOA (F)	0.1696	0.0639	0.1719	0.2463
DESCR_SEXO_PESSOA (M)	0.1239	0.0348	0.1269	0.1941
ESCALAO_CILINDRADA (2)	0.0757	0.0511	0.0773	0.0940
ESCALAO_CILINDRADA (3)	0.2072	0.1701	0.2053	0.2462
DESCR_TIPO_USO (PART)	-0.1685	-0.2380	-0.1671	-0.1032
DESCR_TIPO_USO (PRIV)	-0.2921	-0.3587	-0.2943	-0.2026

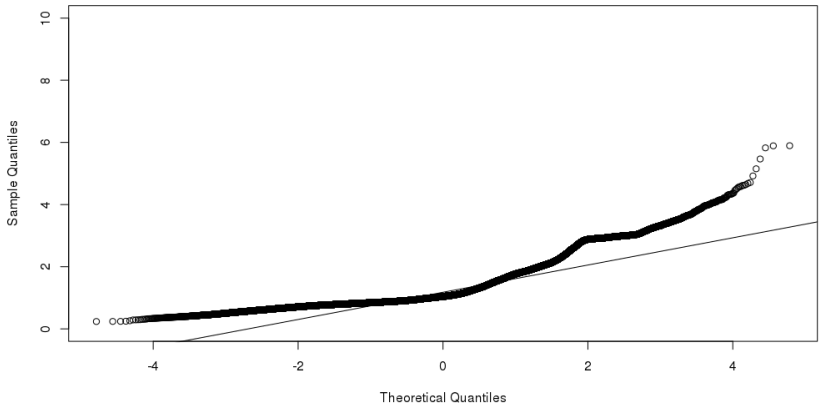

Figure 4.21: Normal Q-Q plot of residuals of the selected ZINB GGAM after removing possible outliers.

figure 4.21). The DIC decreased to 374369.7 and the model complexity increased slightly (pD$= 82.899$).

Table 4.15: Estimated regression coefficients of the η^{σ} in the final model.

Parameter	Mean	2.5% quant.	Median	97.5% quant.
β_0	-1.9842	-5.8754	-1.5957	-0.6416
CATEGORIA_AGREGADA (DIVERSE)	2.5657	1.0859	2.2326	6.2202
CATEGORIA_AGREGADA (LIGHT)	1.1716	0.0585	0.7655	4.79999
CATEGORIA_AGREGADA (MIX)	1.1035	-0.0408	0.7429	4.8567
CATEGORIA_AGREGADA (MOTO)	1.53010	0.12577	1.18389	5.32773
CATEGORIA_AGREGADA (HEAVY_MERC)	2.1573	0.94572	1.79486	5.92122
CATEGORIA_AGREGADA (HEAVY_PASS)	-293.4802	-295.7529	-293.6855	-289.3119
CATEGORIA_AGREGADA (TRAILER)	6.5063	4.6286	6.3661	9.38199
CATEGORIA_AGREGADA (TAXI)	-44.4498	-57.4435	-40.0490	-35.3789
Marca_Conv (FRA)	-0.2798	-0.4903	-0.2776	-0.0507
Marca_Conv (JAP)	-0.0460	-0.2259	-0.0459	0.12335
Marca_Conv (ITA)	-0.0415	-0.3335	-0.0283	0.1815
Marca_Conv (USA)	0.0805	-0.1118	0.0695	0.2911
Marca_Conv (OTHER)	-0.2627	-0.5371	-0.2629	0.0439
Marca_Conv (NS)	0.3805	0.1723	0.3772	0.5881
DESCR_SEXO_PESSOA (F)	-0.1491	-0.7505	-0.1470	0.3003
DESCR_SEXO_PESSOA (M)	0.1171	-0.4964	0.1224	0.5730
ESCALAO_CILINDRADA (2)	-0.1981	-0.3572	-0.1926	-0.0548
ESCALAO_CILINDRADA (3)	0.0615	-0.1178	0.0419	0.3069
DESCR_TIPO_USO (PART)	-0.0775	-0.5425	-0.1140	0.3938
DESCR_TIPO_USO (PRIV)	-0.2185	-0.6508	-0.2181	0.3117

4.4 Estimation of Actuarial Quantities of Interest

In this section, quantities of interest in actuarial applications are estimated using the selected models for both complete data and complete data with imputation. Recent advances in data analysis enable car insurances a more accurate prediction of the risk associated to a given policy. According to previous studies, some risk factors were identified, namely:

- Age of the vehicle - Older cars are more susceptible to failures, increasing the risk of an accident or even fatality.
- Age of the driver - Younger drivers are more prone to accidents due to inexperience and youthful vigour. In addition, old drivers also have relatively high accident rates due to slow reflexes and health issues, such as dementia, heart problems and Parkinson's disease.
- Gender of the driver - Studies conducted in developed countries have shown women are more patient and less reckless than men. In general, women pay lower premiums than men with a similar age, location and driving history.
- Driver's claim record - A driver who is frequently making claims is associated with riskier behaviours and the insurance company will apply a higher premium.
- Home location - Drivers living in urban areas or high-crime neighborhoods pay higher rates insurance premiums.
- Usage of the vehicle - Drivers who use their vehicles for professional purposes are commonly applied higher insurance premiums.

On the basis of previous information, it is possible design some risk profiles for Portugal mainland based on the current car insurance study. The risk profile structure would resemble the following scenarios:

- High risk - an inexperienced male driver aged between 18 and 25, living in Lisboa or Porto and driving a new powerful car.
- Medium risk - a middle-aged male driver living in an urban area and driving a $4 - 15$-year car.
- Low risk - a retired male driver living in the countryside and driving a car aged more than 15.

The chosen data for making estimations are in table 4.16. At a first glance, it is clear that some chosen values are riskier than other ones according to the previous considerations. The selected models in section 4.3 were used to estimate actuarial quantities of interest, the no-claim probability ($\mathrm{P}(\mathrm{N} = 0)$)

and the expected claim count (E(N)). The posterior means of these quantities were obtained by using the entire MCMC sample of each parameter.

Table 4.16: Chosen data for making predictions.

i	Anos_ Carta	Idade_ Veiculo	Idade_ Condutor	DISTRITO	ESCALAO_ CILINDRADA	CATEGORIA_ AGREGADA	DESCR_ SEXO_ PESSOA	Marca_ Conv	DESCR_ TIPO_ USO	Apol_ 2011 2013
1	1	1	19	LISBOA	3	LIGHT	M	USA	PRIV	2
2	0	0	18	LISBOA	3	LIGHT	F	FRA	PRIV	2
3	2	5	21	SETUBAL	2	LIGHT	F	GER	PART	2
4	1	1	20	LISBOA	2	LIGHT	M	OTHER	PART	1.5
5	2	2	23	LISBOA	2	LIGHT	M	GER	PART	1.5
6	1	2	25	LISBOA	2	LIGHT	M	GER	PART	1.5
7	0	0	18	PORTO	3	LIGHT	M	OTHER	PART	1.5
8	1	1	20	PORTO	2	LIGHT	M	GER	PART	2.2
9	1	1	19	COIMBRA	3	LIGHT	F	GER	PART	1.3
10	10	4	29	BRAGA	2	LIGHT	M	GER	PART	1.8
11	20	4	42	LISBOA	2	LIGHT	M	OTHER	PART	2
12	22	8	50	LISBOA	2	LIGHT	M	JAP	PART	1.5
13	15	6	38	LISBOA	2	LIGHT	M	ITA	PART	2
14	20	4	42	PORTO	2	LIGHT	M	FRA	PART	2
15	22	8	50	PORTO	2	LIGHT	M	GER	PART	1.5
16	15	6	38	AVEIRO	2	LIGHT	F	GER	PRIV	1.5
17	12	8	35	VISEU	2	LIGHT	M	GER	PART	2
18	1	1	19	FARO	2	LIGHT	M	FRA	PART	1
19	45	18	63	PORTALEGRE	2	LIGHT	M	FRA	PART	1.5
20	40	15	65	EVORA	2	LIGHT	M	GER	PART	2.5
21	52	23	72	BRAGANCA	2	LIGHT	M	OTHER	PART	2
22	35	12	60	BEJA	2	LIGHT	M	OTHER	PART	1.5
23	35	13	55	LISBOA	3	LIGHT	F	USA	PART	2
24	33	16	53	LISBOA	2	LIGHT	M	OTHER	PRIV	2
25	7	8	26	LISBOA	2	LIGHT	M	OTHER	PRIV	2
26	0	0	18	LISBOA	3	LIGHT	M	FRA	PRIV	2
27	0	0	18	PORTO	3	LIGHT	F	OTHER	PART	2
28	0	0	18	LISBOA	3	LIGHT	M	FRA	PRIV	0.5
29	0	0	18	PORTO	3	LIGHT	F	OTHER	PART	0.5

From table 4.17, estimations of the no-claim probability and the expected claim count based on the complete data are generally larger and smaller, respectively, than complete data with imputation. Moreover, removing possible atypical observations has a small decreasing effect of no-claim probability (thus, increasing effect of the other quantity of interest) using the complete data, but for complete data with imputation these trends are not widely visible. It can be seen that young policyholders with short license time, driving new cars and living in main urban areas (e.g. observations 1, 2, 4, 26 and 28 in table 4.16) have smaller no-claim probability and larger expected claim count. Being a young male driver makes it riskier (e.g. observation 2 vs observation 26). However, making short term policies with younger (both male and female) policyholders may be the more convenient strategy for the company (e.g. observation 26 vs observation 28 and observation 27 vs observation 29). As policyholders get older, their license time increases and they grow on experience behind the wheel, which leads to larger no-

Table 4.17: Estimations of no-claim probabilities and expected claim counts on the generated data using the select model for the complete data and complete data with imputation (imp), as well as the fitted models removing possible atypical observations.

i	P(N=0)	E[N]	P(N=0)^{out}	E[N]^{out}	P(N=0)$_{imp}$	E[N]$_{imp}$	P(N=0)$^{out}_{imp}$	E[N]$^{out}_{imp}$
1	0.6542	0.4956	0.6679	0.4881	0.5871	0.6123	0.5923	0.5955
2	0.7161	0.3759	0.7222	0.3771	0.6682	0.4469	0.6522	0.4724
3	0.7681	0.2907	0.7709	0.2938	0.7245	0.3489	0.7060	0.3777
4	0.6769	0.4399	0.6919	0.4317	0.6153	0.5319	0.6223	0.5160
5	0.7130	0.376	0.7261	0.3679	0.6614	0.4472	0.6644	0.4406
6	0.7045	0.3903	0.7167	0.3842	0.6611	0.4481	0.6618	0.4456
7	0.6524	0.4776	0.6668	0.4703	0.5845	0.5845	0.5966	0.5587
8	0.6782	0.4527	0.6899	0.4442	0.6186	0.5567	0.6238	0.5414
9	0.7995	0.2405	0.8042	0.2399	0.7617	0.2855	0.7545	0.2947
10	0.8265	0.2024	0.8311	0.1999	0.7827	0.2572	0.7806	0.2599
11	0.8263	0.2069	0.8322	0.2034	0.7841	0.2598	0.7823	0.2622
12	0.8332	0.1977	0.8378	0.1961	0.7954	0.2420	0.7935	0.2446
13	0.8256	0.2075	0.8292	0.2068	0.7835	0.2607	0.7791	0.2667
14	0.8327	0.1961	0.8377	0.1931	0.7916	0.2474	0.7896	0.2500
15	0.8390	0.1875	0.8427	0.1862	0.8024	0.2304	0.8003	0.2333
16	0.8686	0.1481	0.8711	0.1471	0.8377	0.1831	0.8355	0.1859
17	0.8765	0.1388	0.8779	0.1384	0.8421	0.1794	0.8409	0.1807
18	0.8060	0.2388	0.8149	0.2349	0.7627	0.2894	0.7674	0.2823
19	0.9101	0.1011	0.9093	0.1037	0.8827	0.1299	0.8764	0.1379
20	0.8944	0.1228	0.8951	0.1238	0.8692	0.1512	0.8651	0.1568
21	0.8850	0.1334	0.8907	0.1278	0.8648	0.1534	0.8633	0.1553
22	0.8805	0.1375	0.8860	0.1333	0.8573	0.1605	0.8575	0.1606
23	0.8219	0.2189	0.8309	0.2111	0.7793	0.2702	0.7834	0.2639
24	0.8242	0.2158	0.8302	0.2120	0.7858	0.2610	0.7844	0.2631
25	0.7814	0.2720	0.7898	0.2652	0.7315	0.3397	0.7311	0.3394
26	0.6273	0.5528	0.6432	0.5429	0.5586	0.6795	0.5681	0.6496
27	0.7250	0.3565	0.7295	0.3579	0.6781	0.4256	0.6620	0.4506
28	0.8138	0.2387	0.8268	0.2351	0.7702	0.2868	0.7790	0.2731
29	0.8681	0.1540	0.8724	0.1550	0.8431	0.1796	0.8356	0.1894

claim probability. The risk does not vary significantly from males to females in this case. Here, the main differences arise depending on the district the policyholder lives in, with main urban areas again being riskier (e.g. observations 10, 11, 16 and 23). This was more evident for old policyholders (observations 19, 20 and 21). The ones living in the countryside are associated to large no-claim probabilities, while the ones living in main urban areas have an increased risk.

Chapter 5

Conclusion

This chapter is dedicated to state the achievements obtained through this study, motivated by the current car insurance data (section 5.1) and at last some suggestions are presented for future work (section 5.2).

5.1 Achievements

The present work uses recent methodology in actuarial applications to model claim frequency of policies from a portuguese car insurance in the period 2011-2013. Model selection was performed for complete data ($477,997$ observations) and complete data with imputation ($585,256$ observations).

The application of Zero-Inflated (ZINB) models proved to be a valid alternative to Poisson models to accommodate zero-inflation and relatively large dispersion in these data. In contrast with the Poisson GGAM, the mean and other parameters related to the sampling distribution were specified in terms of various risk factors with appropriate link functions, leading to more complex models that were fitted under a Bayesian perspective via MCMC methods. Variable selection was computationally heavy for ZINB models, with frequent convergence issues in the backfitting algorithm. As such, simple model formulations were the primary choice, and the selected models were constructed in a forward stepwise procedure on the basis of DIC and pD, which proved to be important tools. It was noticed that including spatial uncorrelated effects in the count and dispersion predictors was not relevant, in contrast with spatial correlated effects. It was also clear that the nonlinear effects are more suitable for representing continuous covariate effects and the inclusion of an interaction term between gender and age of the policyholder improves the model fitting results.

For both complete data and complete data with imputation, the analysis of models residuals confirmed that a ZINB GGAM was a good alternative especially for Poisson GGAM. Since a ZINB GGAM improved model residuals and DIC in comparison with a ZIP GGAM, the former was here considered as the final selected model for the car insurance data in both scenarios. Removing possible atypical

(extreme) observations made no significant differences in the model residuals.

Finally, the selected models in both data scenarios were used to estimate actuarial quantities of interest based on a generated data set. In general, no-claim probability and expected claim count did not differ expressively in both data scenarios. Notice that zero-inflation is quite relevant, which led to relatively large no-claim probability and small expected claim count in general, even for riskier policies.

Note that using the complete data with imputation led to less dispersed models in general and the model fitting results were also less sensitive to changes in the basis dimension of nonlinear covariate functions. However, the DIC and convergence diagnostic results are not improved in comparison with the selected model for complete data. Therefore, the ZINB GGAM for complete data is the final selected model. Nevertheless, data imputation is a valid option for model comparison in car insurance problems regarding incomplete data, especially for MAR mechanism.

5.2 Future Work

As noted in exploratory data analysis (section 4.1), the analyzed car insurance data only contain policies registered in districts from Portugal mainland. An extension of this study would be to perform a separate analysis for the districts Madeira and Azores, or even a complete analysis based on all Portuguese municipalities. In particular, the prediction of actuarial quantities could lead to new insights, such as 'How does the no-claim probability of a given policy differ among all Portuguese municipalities?'. Moreover, prediction could be used for a more rigorous definition of risk profiles at a municipality level.

Despite the importance of IWLS proposals for Bayesian structured additive regression models, it is common to obtain very low acceptance rates in the estimation of spatial parameters during MCMC simulation. An alternative that could possibly improve acceptance rates is to use a multilevel framework (see e.g. Klein et at., 2015).

Moreover, hurdle Poisson and hurdle negative binomial models could be used for comparison with fitting results from the proposed ZINB models. At last, due to the natural importance of the variable claim cost, a potential modeling extension is to consider Bayesian joint models for claim frequency and claim cost.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974. DOI: 10.1109/TAC.1974.1100705.

[2] T. Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763.

[3] C. Belitz, A. Brezger, N. Klein, T. Kneib, S. Lang, and N. Umlauf. *BayesX - Software for Bayesian inference in structured additive regression models*, 2015. R package version 3.0.2.

[4] J. Berger. An overview of robust Bayesian analysis. *Test*, 3:5–124, 1994. DOI: 10.1007/BF02562676.

[5] J. Besag, J. York, and A. Mollié. Bayesian Image Restoration with Two Applications in Spatial Statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991. DOI: 10.1007/BF00116466.

[6] A. Brezger and S. Lang. Generalized structured additive regression based on bayesian P-splines. *Computational Statistics & Data Analysis*, 50:967–991, 2006. DOI: 10.1016/j.csda.2004.10.011.

[7] P. Craven and G. Whaba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[8] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978. ISBN 978-0-387-95366-3.

[9] M. Denuit and S. Lang. Non-life rate-making with bayesian GAMs. *Insurance: Mathematics and Economics*, 35:627–647, 2004. DOI: 10.1016/j.insmatheco.2004.08.001.

[10] M. Denuit, X. Maréchal, S. Pitrebois, and J. Walhin. *Actuarial Modelling of Claim Counts. Risk classification, Credibility and Bonus-Malus Systems*. Chichester: Wiley, 2007. ISBN 9780470026779.

[11] P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121, 1996. DOI: 10.1214/ss/1038425655.

[12] L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica*, 14:731–761, 2004.

[13] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods, and applications*. Springer Science & Business Media, 2013. ISBN 978-3-642-34333-9.

[14] A. Gelman and D. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7:457–472, 1992. DOI: 10.1214/ss/1177011136.

[15] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. DOI: 10.1109/TPAMI.1984.4767596.

[16] P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London, 1994. ISBN 9780412300400.

[17] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[18] T. Hastie and R. Tibshirani. Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B*, 55:757–796, 1993. DOI: 10.1111/j.2517-6161.1993.tb01939.x.

[19] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. DOI: 10.1093/biomet/57.1.97.

[20] H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

[21] E. Kammann and M. Wand. Geoadditive models. *Journal of the Royal Statistical Society*, 52:1–18, 2003. DOI: 10.1111/1467-9876.00385.

[22] N. Klein, M. Denuit, T. Kneib, and S. Lang. Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale and shape. *Insurance: Mathematics and Economics*, 55:225–249, 2014. DOI: 10.1016/j.insmatheco.2014.02.001.

[23] N. Klein, T. Kneib, and S. Lang. Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110:405–419, 2015. DOI: 10.1080/01621459.2014.912955.

[24] S. Lang and A. Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13:183–212, 2004. DOI: 10.1198/1061860043010.

[25] P. Laplace. Mémoire sur la probabilité des causes par les évènements. *Mémoires de Mathematique et de Physique, Presentés à l'Académie Royale des Sciences, Par Divers Savans Lus Dans ses Assemblées, Tome Sixième*, 6:621–656, 1774.

[26] T. Martin, B. Wintle, J. Rhodes, P. Kuhnert, S. Field, S. Low-Choy, A. Tyre, and H. Possingham. Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8:1235–1246, 2005. DOI: j.1461-0248.2005.00826.x.

[27] B. Marx and P. Eilers. Direct generalized additive modelling with penalized likelihood. *Computational Statistics Data Analysis*, 28:193–209, 1998. DOI: 10.1016/S0167-9473(98)00033-4.

[28] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953. DOI: 10.1063/1.1699114.

[29] J. Nelder and R. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370–384, 1972. DOI: 10.2307/2344614.

[30] A. O'Hagan. *Bayesian Inference*. Kendall's Advanced Theory of Statistics, vol. 2B, Arnold, London, 1994. ISBN 9780340529225.

[31] C. Paulino, M. Amaral Turkman, B. Murteira, and G. Silva. *Estatística Bayesiana*. Fundação Calouste Gulbenkian, 2nd edition, 2018. ISBN 978-972-31-1606-9.

[32] R. Rigby and D. Stasinopoulos. Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:507–554, 2005. DOI: j.1467-9876.2005.00510.x.

[33] C. Robert. *The Bayesian choice*. Springer, New York, 2nd edition, 1994. ISBN 9780387715988.

[34] D. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976. DOI: 10.1093/biomet/63.3.581.

[35] H. Rue and L. Held. *Gaussian Markov Random Field: Theory and Applications*. Chapman and Hall/CRC, 1st edition, 2005. ISBN 9781584884323.

[36] D. Ruppert and D. Matteson. *Statistics and Data Analysis for Financial Engineering*. New York, Springer, 2nd edition, 2015. ISBN 978-1-4939-2614-5.

[37] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, 64:583–639, 2002. DOI: 10.1111/1467-9868.00353.

[38] M. Stasinopoulos and B. Rigby. *gamlss: Generalized Additive Models for Location, Scale and Shape*, 2016. R package version 5.1-3.

[39] A. Syversveen. Noninformative Bayesian priors: interpretation and problems with construction and applications. *Preprint statistics*, 3:1–11, 1998.

[40] M. Templ, A. Alfons, and A. Kowarik. *VIM: Visualization of Imputation of Missing Values*, 2011. R package version 4.8.0.

[41] K. Umlauf, N. Klein, A. Zeileis, and M. Köhler. *bamlss: Bayesian Additive Models for Location, Scale and Shape (and Beyond)*, 2017. R package version 0.1-2.

[42] N. Umlauf, D. Adler, T. Kneib, S. Lang, and A. Zeileis. Structured additive regression models: An R interface to bayesx. *Journal of Statistical Software*, 63:1–46, 2015. DOI: 10.18637/jss.v063.i21.

[43] N. Umlauf, N. Klein, and A. Zeileis. BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27:612–627, 2018. DOI: 10.1080/10618600.2017.1407325.

[44] N. Umlauf and T. Kneib. A primer on bayesian distributional regression. *Statistical Modelling*, 18:219–247, 2018. DOI: 10.1177/1471082X18759140.

[45] S. van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–243, 2007. DOI: 10.1177/0962280206074463.

[46] S. van Buuren, H. Boshuizen, and D. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics and Medicine*, 18:681–694, 1999. DOI: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.

[47] S. van Buuren and K. Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. 45:1–67, 2011. DOI: 10.18637/jss.v045.i03.

[48] S. van Buuren and K. Groothuis-Oudshoorn. *mice: Multiple Imputation by Chained Equations*, 2019. R package version 3.6.0.

[49] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571—-3594, 2010.

[50] S. Wood. *Generalized Additive Models: An introduction with R*. Chapman & Hall/CRC, 1st edition, 2006. ISBN 978-1584884743.

[51] A. Zuur, A. Saveliev, and E. Ieno. *Zero Inflated Models and Generalized Linear Mixed Models with R*. Highland Statistics Ltd, Newburgh, 1st edition, 2012. ISBN 978-0957174115.

Appendix A

A.1 Full conditional posterior distributions

Denote the vector $\boldsymbol{\alpha}$ without its component j by $\boldsymbol{\alpha}_{-j}$. From the joint posterior (3.30) in subsection 3.3.1, the full conditional posterior distributions for the Poisson model, denoted by $[j|\boldsymbol{\alpha}_{-j}]$, are given by

(i) $[\beta_{jl}|\boldsymbol{\alpha}_{-\beta_{jl}}] \propto \exp\left\{\beta_{jl}\sum_{i=1}^{n} y_i b_{jl}(x_{ji}) - \sum_{i=1}^{n} e^{\eta_i} - \frac{1}{2\tau_j^2}\beta_{jl}\left[\beta_{j1}K_{l1} + \ldots + \beta_{j,l-1}K_{l,l-1} + \sum_{w=1}^{q_j}\beta_{jw}K_{wl} + \beta_{j,l+1}K_{l,l+1} + \ldots + \beta_{jq_jK_{lq_j}}\right]\right\}$, $l = 1, \ldots, q_j$, $j = 1, \ldots, p$

(ii) $[\beta_{str,l}|\boldsymbol{\alpha}_{-\beta_{str,l}}] \propto \exp\left\{\beta_{str,l}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\eta_i} - \frac{1}{2\tau_j^2}\beta_{str,l}\left[\beta_{str,1}(K_{str})_{l1} + \ldots + \beta_{str,l-1}(K_{str})_{l,l-1} + \sum_{w=1}^{q_{str}}\beta_{str,w}(K_{str})_{wl} + \beta_{str,l+1}(K_{str})_{l,l+1} + \ldots + \beta_{str,q_{str}(K_{str})_{lq_{str}}}\right]\right\}$, $l = 1, \ldots, q_{str}$

(iii) $\tau_j^2|\boldsymbol{\alpha}_{-\tau_j^2} \sim \mathsf{IG}\left(a_j + \frac{rank(\boldsymbol{K}_j)}{2}, b_j + \frac{\boldsymbol{\beta}_j^T\boldsymbol{K}_j\boldsymbol{\beta}_j}{2}\right)$, $j = 1, \ldots, p$

(iv) $\tau_{str}^2|\boldsymbol{\alpha}_{-\tau_{str}^2} \sim \mathsf{IG}\left(a_{str} + \frac{rank(\boldsymbol{K}_{str})}{2}, b_{str} + \frac{\boldsymbol{\beta}_{str}^T\boldsymbol{K}_{str}\boldsymbol{\beta}_{str}}{2}\right)$

(v) $\nu|\boldsymbol{\alpha}_{-\nu} \sim \mathsf{IG}\left(a'S + S - 1, b'S + \sum_{s=1}^{S}\frac{b_s^2}{2}\right)$

(vi) $[\gamma_0|\boldsymbol{\alpha}_{-\gamma_0}] \propto \exp\left\{\gamma_0\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\eta_i}\right\}$

(vii) $[\gamma_k|\boldsymbol{\alpha}_{-\gamma_k}] \propto \exp\left\{\gamma_k\sum_{i=1}^{n} x_{ki}y_i - \sum_{i=1}^{n} e^{\eta_i}\right\}$, $k = 1, \ldots, p^\star$

(viii) $[b_s|\boldsymbol{\alpha}_{-b_s}] \propto \exp\left\{b_s\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\eta_i} - \frac{b_s^2}{2\nu^2}\right\}$, $s = 1, \ldots, S$

where $\eta_i = \gamma_0 + \sum_{k=1}^{p^\star}\gamma_k x_{ki} + \sum_{k=1}^{q_1} b_{1,k}(z_{1i})\beta_{1k} + \ldots + \sum_{k=1}^{q_p} b_{p,k}(z_{pi})\beta_{pk} + f_{str}(s_i) + b_{s_i}$, $i = 1, \ldots, n$.

A.2 Convergence Diagnostic Results

It is necessary to make sure that the MCMC sampler explores the parameter space efficiently, i.e. it does not reject or accept too many proposals. In this case, chains are well mixed. On the other hand, if too many proposals are rejected, many simulations are needed to generate a sufficient number of parameter samples. Moreover, if too many proposals are accepted, there is not much information about the underlying distribution. This is the case for poor mixing.

A.2.1 Auto-Correlation Function Plots

One way to check for convergence is to look at auto-correlation between samples returned by the MCMC output. The lag-k auto-correlation is the auto-correlation between every sample and the sample k-steps before. This auto-correlation should become smaller as k increases, i.e. samples can be considered as independent. If, on the other hand, auto-correlation remains high for higher values of k, this indicates a high degree of correlation between the samples and slow mixing.

Figure A.1: Maximum ACF plots of samples for the final selected models for complete data (left) and complete data with imputation (right).

In figure A.1, it is shown an auto-correlation function (ACF) plot of the maximum auto-correlation of all parameter samples for the final selected models in both data scenarios. It can be seen that maximum auto-correlation decreases more expressively in the complete data case, which indicates better mixing.

A.2.2 Trace Plots of Parameter Samples

Trace plots provide an important tool for assessing the mixing of a chain. In the trace plots, we want to avoid flat bits (where the chain remains in the same state for too long) or too many consecutive steps in one direction.

In figure A.2, the trace plots of parameter samples for term $f(\texttt{Anos\_Carta})$ on η_μ for the final selected model for complete data. Notice that for the parameters associated with smooth functions, there is no visible trend in parameter results as the number of iterations increases. Moreover, the ACF plots show

a relatively quick decrease for autocorrelation after the first lags. Therefore, convergence is relatively attained. Since the number of estimated parameters is very large and the conclusions are similar to previous ones, the remaining trace plots are omitted.

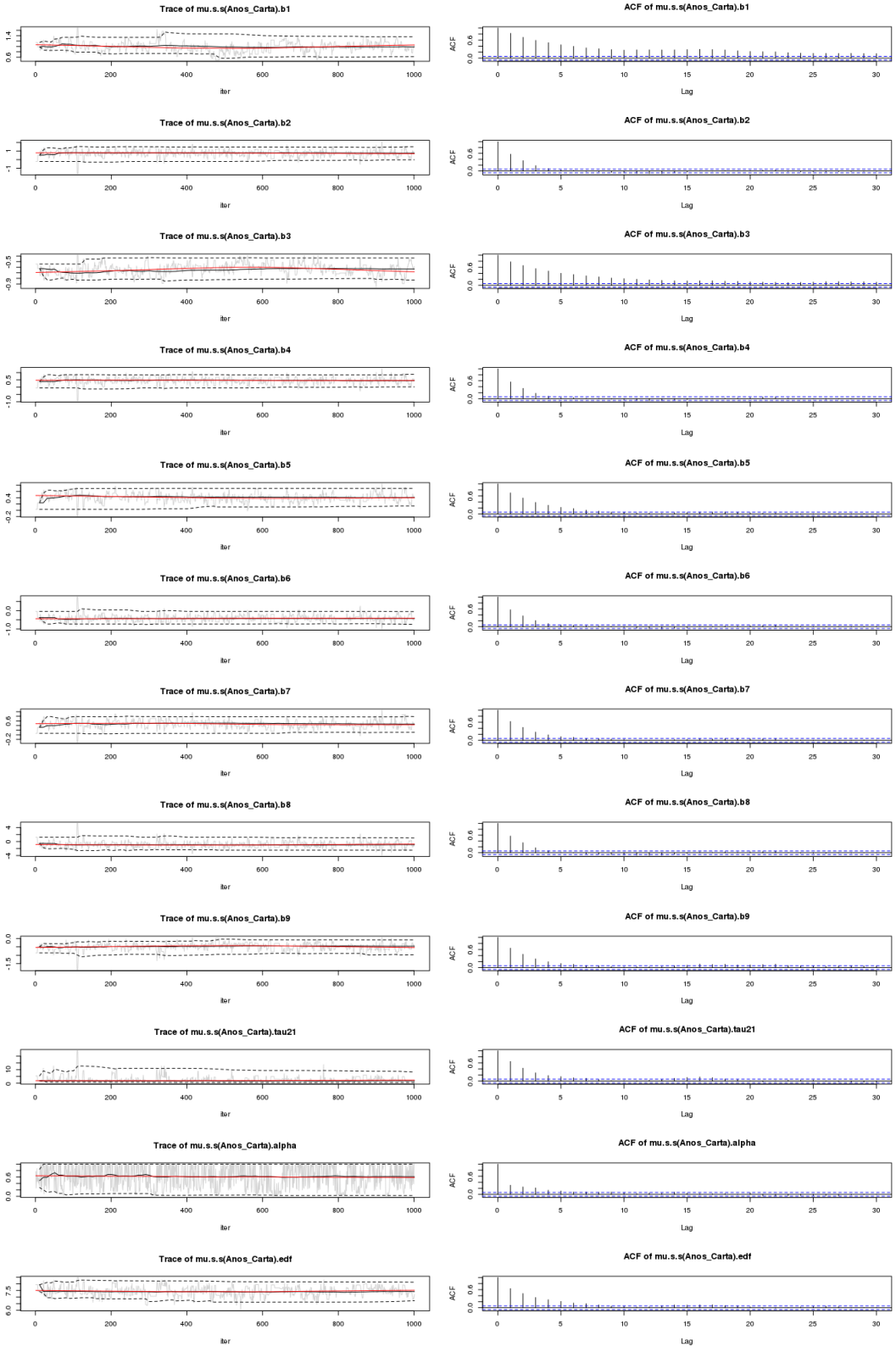

Figure A.2: Traceplots of parameter samples for term $f(\texttt{Anos\_Carta})$ on η_μ for the final selected model for complete data.

Appendix B

Code of Project

Listing B.1: R code for data preprocessing and exploratory analysis.

```
#set working directory
#read the data into R
library(openxlsx)
ins<-read.xlsx("Insurance.xlsx")

#summary
summary(ins)
#variables with missing values
sum(is.na(ins$DISTRITO))
#595
sum(is.na(ins$CONCELHO))
#41
sum(is.na(ins$ESCALAO_CILINDRADA))
#12094
sum(is.na(ins$descr_escalao_cilindrada))
#12094
sum(is.na(ins$DESCR_CATEGORIA_APOLICE))
#5
sum(is.na(ins$CATEGORIA_AGREGADA))
#3
sum(is.na(ins$Marca_Conv))
#2334

```

```r
#rename variables
colnames(ins)[11] <- "Idade_Veiculo"
colnames(ins)[21] <- "Apol_20112013"
colnames(ins)[20] <- "NS_20112013"
colnames(ins)[19] <- "CS_20112013"

#COMPLETE DATA (remove observations with at least one NA)

ins$DESCR_SEXO_PESSOA <- as.factor(ins$DESCR_SEXO_PESSOA)
ins$DESCR_TIPO_USO<-as.factor(ins$DESCR_TIPO_USO)
ins$GARAGEM<-as.factor(ins$GARAGEM)
ins$Credor<-as.factor(ins$Credor)
ins$NS_2011.2013<-as.integer(ins$NS_2011.2013)
ins$DISTRITO<-as.factor(ins$DISTRITO)
ins$CATEGORIA_AGREGADA<-as.factor(ins$CATEGORIA_AGREGADA)
ins$ESCALAO_CILINDRADA<-as.factor(ins$ESCALAO_CILINDRADA)
ins$Marca_Conv<-as.factor(ins$Marca_Conv)

dtest3_semna<-ins

library(dplyr)

#remove levels from covariate DISTRITO
dtest3_semna<-dtest3_semna[-which(dtest3_semna$DISTRITO=="MADEIRA"|
                           dtest3_semna$DISTRITO=="MADEIRA(PORTO_SANTO)"|
                           dtest3_semna$DISTRITO=="ACORES"),]

dtest3_semna$DISTRITO<-droplevels(dtest3_semna$DISTRITO)

#remove observations with at least one NA
dtest3_semna<-na.omit(dtest3_semna)

#remove variables
dtest3_semna<-dtest3_semna[,-10]
dtest3_semna<- dtest3_semna[,-5]
dtest3_semna<-dtest3_semna[,-3]
dtest3_semna<- dtest3_semna[,-3]
dtest3_semna<- dtest3_semna[,-14]
```

```
dtest3_semna<- dtest3_semna[,-1]
dtest3_semna<- dtest3_semna[,-3]

#Marca_Conv - level modifications
dtest3_semna[dtest3_semna$Marca_Conv=="MERCEDES"|
        dtest3_semna$Marca_Conv=="BMW"|
        dtest3_semna$Marca_Conv=="OPEL"|
        dtest3_semna$Marca_Conv=="SMART"|
        dtest3_semna$Marca_Conv=="AUDI"|
        dtest3_semna$Marca_Conv=="VOLKSWAGEN",]$Marca_Conv<- "GER"

dtest3_semna[dtest3_semna$Marca_Conv=="ALFA_ROMEO"|
        dtest3_semna$Marca_Conv=="FIAT"|
        dtest3_semna$Marca_Conv=="JEEP"|
        dtest3_semna$Marca_Conv=="LANCIA",]$Marca_Conv <- "ITA"

dtest3_semna[dtest3_semna$Marca_Conv=="TOYOTA"|
        dtest3_semna$Marca_Conv=="NISSAN"|
        dtest3_semna$Marca_Conv=="MAZDA"|
        dtest3_semna$Marca_Conv=="SUZUKI"|
        dtest3_semna$Marca_Conv=="HONDA"|
        dtest3_semna$Marca_Conv=="MITSUBISHI",]$Marca_Conv <- "JAP"

dtest3_semna[dtest3_semna$Marca_Conv=="RENAULT"|
        dtest3_semna$Marca_Conv=="PEUGEOT"|
        dtest3_semna$Marca_Conv=="CITROEN",]$Marca_Conv <- "FRA"

dtest3_semna[dtest3_semna$Marca_Conv=="CHEVROLET"|
        dtest3_semna$Marca_Conv=="FORD",]$Marca_Conv <- "EUA"

dtest3_semna[dtest3_semna$Marca_Conv=="LAND_ROVER"|
        dtest3_semna$Marca_Conv=="MINI"|
        dtest3_semna$Marca_Conv=="VOLVO"|
        dtest3_semna$Marca_Conv=="SKODA"|
        dtest3_semna$Marca_Conv=="SEAT"|
        dtest3_semna$Marca_Conv=="DACIA"|
        dtest3_semna$Marca_Conv=="KIA"|
        dtest3_semna$Marca_Conv=="HYUNDAI",]$Marca_Conv<-"OUTRO_PAIS"
```

```
100
101 dtest3_semna[dtest3_semna$Marca_Conv=="Z-DIVERSOS"|
102         dtest3_semna$Marca_Conv=="Z-PEQ_QUANT"|
103         dtest3_semna$Marca_Conv=="Z-LUXO"|
104         dtest3_semna$Marca_Conv=="Z-PESADOS",]$Marca_Conv<- "NS"
105
106 dtest3_semna$Marca_Conv <- droplevels(dtest3_semna$Marca_Conv)
107
108 #CATEGORIA_AGREGADA - level modifications
109 levels(dtest3_semna$CATEGORIA_AGREGADA)<-c(levels(dtest3_semna$CATEGORIA_AGREGADA),
110                                     "OUTRA")
111
112 dtest3_semna[which(dtest3_semna$CATEGORIA_AGREGADA!="LIGEIRO" &
113     dtest3_semna$CATEGORIA_AGREGADA!="MISTOS" &
114     dtest3_semna$CATEGORIA_AGREGADA!="MOTOCICLO"),]$CATEGORIA_AGREGADA <-"OUTRA"
115
116 dtest3_semna$CATEGORIA_AGREGADA<- droplevels(dtest3_semna$CATEGORIA_AGREGADA)
117
118 #DESCR_TIPO_USO - level modifications
119 dtest3_semna[which(dtest3_semna$DESCR_TIPO_USO!="VIDA_PRIVADA" &
120 dtest3_semna$DESCR_TIPO_USO!="USO_PROFISSIONAL"),]$DESCR_TIPO_USO<-"USO_PROFISSIONAL"
121
122 dtest3_semna$DESCR_TIPO_USO<- droplevels(dtest3_semna$DESCR_TIPO_USO)
123
124 # COMPLETE DATA WITH IMPUTATION
125 #level modifications are made after imputation
126 ins3<-ins[,-10]
127 ins3<- ins3[,-5]
128 ins3<-ins3[,-3]
129 ins3<- ins3[,-3]
130 ins3<- ins3[,-14]
131 ins3<- ins3[,-1]
132 ins3<- ins3[,-3]
133 #missing patterns and propotions
134 library(VIM)
135 aggr_plot <- aggr(ins3, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
136                 labels=names(data),cex.axis=0.30, cex.numbers=0.7,
137                 gap=2, ylab=c("Histogram of missing data", "Pattern"))
```

```
#multiple imputation via chained equations
library(mice)
ins3imp<-mice(ins3, m=3, pred=quickpred(ins3, mincor=0.25,  minpuc=0.25),
             nnet.MaxNWts = 3000, maxit=3, seed=123)

#imputation
ins3comp1<- complete(ins3imp,1) # complete with 1st imputed data set
ins3comp2<- complete(ins3imp,2)# complete with 2nd imputed data set
ins3comp3<- complete(ins3imp,3) #complete with 3rd imputed data set

ins3imp$meth #imputation methods

densityplot(ins3imp) # densities - complete data vs imputed data

dtest3<-ins3comp1 # choose 1st imputed data set

#level modifications for Marca_Conv, CATEGORIA_AGREGADA,
#DESCR_TIPO_USO are analogous to those for complete data

#BAR PLOTS FOR CLAIM FREQUENCY - ORIGINAL DATA vs COMPLETE DATA
#no missing values in NS_20112013
par(mfrow=c(1,2))
barplot(table(ins$NS_20112013), xlab = "Number of claims",
ylab = "Frequency", main="Data set with NA", col="lightseagreen",
las=1, cex.axis=0.7, ylim=c(0, 600000))
barplot(table(dtest3_semna$NS_20112013), xlab = "Number of claims",
ylab = "Frequency", main="Data set with NA removed", col="orange",
las=1, cex.axis=0.7, ylim=c(0, 600000))

par(mfrow=c(1,1))

# DOTPLOTS AND XYPLOTS OF COVARIATES
library(lattice)
library(gridExtra)

pp1<- dotplot(NS_20112013~Idade_Veiculo, data=dtest3)
pp2<-dotplot(NS_20112013~Idade_Condutor, data=dtest3)
grid.arrange(pp1,pp2, ncol=2)
```

```
p1<-xyplot(NS_2011.2013~Anos_Carta, data=ins3test)
p2<- xyplot(NS_2011.2013~Idade_Veiculo, data=ins3test)
p3<-xyplot(NS_2011.2013~Idade_Condutor, data=ins3test)
p4<-xyplot(NS_2011.2013~CS_2011.2013, data=ins3test)
p5<-xyplot(NS_2011.2013~Apol_2011.2013, data=ins3test)
grid.arrange(p1,p2,p3,p4,p5 nrow=2, ncol=3)

#proportions of levels of ESCALAO_CILINDRADA for imputed data and complete data
imputed<- ins3comp1test$ESCALAO_CILINDRADA[is.na(ins3test$ESCALAO_CILINDRADA)]
original<- ins3comp1test$ESCALAO_CILINDRADA[!is.na(ins3test$ESCALAO_CILINDRADA)]
table(imputed)
table(original)
#analogous for other imputed categorical covariates

#barplots for categorical covariates
aux<- dtest3_semna %>% count(ESCALAO_CILINDRADA)
aux1<- dtest3_semna %>% count(DESCR_TIPO_USO)
aux2<- dtest3_semna %>% count(CATEGORIA_AGREGADA)
aux3<- dtest3_semna %>% count(DESCR_SEXO_PESSOA)
aux4<- dtest3_semna %>% count(Credor)
aux5<- dtest3_semna %>% count(GARAGEM)
aux6<- dtest3_semna %>% count(Marca_Conv)
aux7<- dtest3_semna %>% count(DISTRITO)

p1<-ggplot(aux, aes(x=ESCALAO_CILINDRADA,y=n, fill=ESCALAO_CILINDRADA))
        + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(), axis.ticks.x=element_blank())

p2<-ggplot(aux1, aes(x=DESCR_TIPO_USO,y=n, fill=DESCR_TIPO_USO))+
        geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(), axis.ticks.x=element_blank(), width=0.65)

p3<-ggplot(aux3, aes(x=DESCR_SEXO_PESSOA,y=n, fill=DESCR_SEXO_PESSOA))+
        geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(), axis.ticks.x=element_blank())

p4<-ggplot(aux4, aes(x=Credor,y=n, fill=Credor))
        + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
```

```
214         axis.text.x=element_blank(), axis.ticks.x=element_blank(), width=0.40)
215
216 p5<-ggplot(aux5, aes(x=GARAGEM,y=n, fill=GARAGEM))
217         + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
218         axis.text.x=element_blank(), axis.ticks.x=element_blank(), width=0.45)
219
220 p6<-ggplot(aux2, aes(x=CATEGORIA_AGREGADA,y=n, fill=CATEGORIA_AGREGADA))
221         + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
222         axis.text.x=element_blank(), axis.ticks.x=element_blank())
223
224 p7<-ggplot(aux7, aes(x=reorder(DISTRITO, -n),y=n, fill=reorder(DISTRITO, -n)))
225         + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
226         axis.text.x=element_blank(), axis.ticks.x=element_blank())
227         +theme(legend.title = element_blank())
228
229 p8<-ggplot(aux6, aes(x=Marca_Conv,y=n, fill=Marca_Conv))
230         + geom_bar(stat="identity")+theme(axis.title.x=element_blank(),
231         axis.text.x=element_blank(), axis.ticks.x=element_blank())
232
233 grid.arrange(p1,p2,p3,p4,p5, p6, nrow=2, ncol=3)
234
235 # BOXPLOTS FOR CONTINUOUS COVARIATES
236 op <- par(mar = c(5, 8, 4, 2) + 0.1)
237 boxplot(dtest3[, c("Anos_Carta", "Idade_Veiculo", "Idade_Condutor")],
238 col=c("orange", "red", "green"), border="brown", notch=TRUE,
239 horizontal=TRUE, las=2, cex.axis=0.8)
240 par(op) #reset margins
241
242 #MAP
243 #read shapefile into R
244 library(rgdal)
245 shpd<- readOGR(dsn="C:/Users/joaopedro/Desktop/Dist",layer="dist")
246
247 #remove polygons for Azores and Madeira
248 shpd<-shpd[shpd$ID_1!="Azores",]
249 shpd<-shpd[shpd$ID_1!="Madeira",]
250 writeOGR(obj=shpd, dsn="C:/Users/joaopedro/Desktop/Dist",
251         layer="dist",driver="ESRI Shapefile")
```

```r
#convert to boundary format (bnd) file
library(R2BayesX)
bndd <-shp2bnd(shpname="C:/Users/joaopedro/Desktop/Dist/dist",
        regionnames="ID_1", check.is.in=FALSE)

#plot geographical map
plotmap(bndd)
```

Listing B.2: R code for model selection and prediction.

```r
library(bamlss)
library(gamlss)

#poisson models
po1<- bamlss(NS_20112013 ~ log(Apol_20112013/2.494635)
    +s(Idade_Veiculo)+s(Idade_Condutor)+s(Anos_Carta)
    +s(DISTRITO, bs = "mrf", xt=list(polys=bndd)), data = dtest3_semna,
    family="poisson", optimizer = bfit, sampler = GMCMC, maxit = 500)
#add interaction term
po2<- bamlss(NS_20112013 ~ log(Apol_20112013/2.494635)
    +DESCR_SEXO_PESSOA+DESCR_SEXO_PESSOA*s(Idade_Condutor)
    +s(Idade_Veiculo)+s(Idade_Condutor)+s(Anos_Carta)
    +s(DISTRITO, bs = "mrf", xt=list(polys=bnd)),data = dtest3_semna,
    family="poisson", optimizer = bfit, sampler = GMCMC, maxit = 500)

#modelos zip
#simple intercept model for pi
zip1<- bamlss(NS_20112013 ~ log(Apol_20112013/2.494635)
    +s(Idade_Veiculo)+s(Idade_Condutor)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)), data = dtest3_semna, family="poisson",
    optimizer = bfit, sampler = GMCMC, maxit = 500)
#introduce covariate effects into the preditor of pi
ff1<- list(
    NS_20112013 ~ log(Apol_20112013/2.494635)+CATEGORIA_AGREGADA
    +Marca_Conv+DESCR_TIPO_USO+s(Idade_Veiculo)+s(Idade_Condutor)
    +s(Anos_Carta)+s(DISTRITO, bs = "mrf", xt=list(polys=bnd)),
    nu~ 1)

```

```
zip2<- bamlss(ff1, data = dtest3_semna, family=ZIP,
    optimizer = bfit, sampler = GMCMC, maxit = 500)
#add interaction term
zip3<- bamlss(NS_20112013 ~ log(Apol_20112013/2.494635)
    +DESCR_SEXO_PESSOA+DESCR_SEXO_PESSOA*s(Idade_Condutor)
    +s(Idade_Veiculo)+s(Idade_Condutor)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)), data = dtest3_semna,
    family="poisson", optimizer = bfit, sampler = GMCMC, maxit = 500)

#modelos zinb
#simple intercept model for sigma and pi, which is denoted by nu
zinb1<- bamlss(NS_20112013 ~ log(Apol_20112013/2.494635)
    +s(Idade_Veiculo)+s(Idade_Condutor)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)), data = dtest3_semna,
    family=ZINBI, optimizer = bfit, sampler = GMCMC, maxit = 500)
#simple intercept model for pi
f1<-list(
    NS_20112013 ~ log(Apol_20112013/2.494635)+CATEGORIA_AGREGADA
    +Marca_Conv+DESCR_TIPO_USO+s(Idade_Veiculo)
    +s(Idade_Condutor)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)),
    sigma~ s(Anos_Carta),
    nu~ 1)

zinb2<- bamlss(f1, data = dtest3_semna, family=ZINBI,
optimizer = bfit, sampler = GMCMC, maxit = 500)
#covariate effects for both pi and sigma
f2<-list(
    NS_20112013 ~ log(Apol_20112013/2.494635)+CATEGORIA_AGREGADA
    +Marca_Conv+DESCR_TIPO_USO+s(Idade_Veiculo)
    +s(Idade_Condutor)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)),
    sigma~ s(Anos_Carta),
    nu~ Apol_20112013)

zinb3<- bamlss(f2, data = dtest3_semna, family=ZINBI,
optimizer = bfit, sampler = GMCMC, maxit = 500)

```

```
#add interaction terms
f3<-list(
    NS_20112013 ~ log(Apol_20112013/2.494635)+CATEGORIA_AGREGADA+
    DESCR_SEXO_PESSOA+Marca_Conv+DESCR_SEXO_PESSOA*s(Idade_Condutor)
    +s(Idade_Veiculo)+s(Anos_Carta)+
    s(DISTRITO, bs = "mrf", xt=list(polys=bnd)),
    sigma~ s(Anos_Carta),
    nu~ Apol_20112013)

zinb4<- bamlss(f3, data = dtest3_semna, family=ZINBI,
optimizer = bfit, sampler = GMCMC, maxit = 500)

#COVARIATE EFFECTS AND MODEL DIAGNOSIS: EXAMPLE
#estimated covariate effects for mu
plot(zinb1, model="mu")
plot(zinb1, model="mu", term="s(Anos_Carta)")
#estimated covariate effects for sigma
plot(zinb1, model="mu")
plot(zinb1, model="sigma", term="s(Anos_Carta)")

#map representation of spatial effects
spd<- bnd2sp(bndd)
stest<-spd
#get spatial effects and district order
zinb1$results$mu$s.effects$`s(DISTRITO)
#re-order spatial polygon (sp) object according to district order
stest@polygons<-stest@polygons[c("LISBOA", "SANTAREM", "PORTO",
"SETUBAL","VILA_REAL","AVEIRO", "BEJA", "BRAGA", "CASTELO_BRANCO",
"GUARDA", "FARO","COIMBRA", "BRAGANCA", "LEIRIA", "VIANA_DO_CASTELO",
"PORTALEGRE", "VISEU", "EVORA")]

#plot spatial effects on the map
#define scale color, range and position

#mu
plotmap(stest, x=zinb1$results$mu$s.effects$`s(DISTRITO)`$`50%`,
id=zinb1$results$mu$s.effects$`s(DISTRITO)`$DISTRITO,
pos="topleft", color=heat_hcl, lrange=c(-0.40,0.40))
```

```
105
106 #sigma
107 plotmap(stest, x=zinb1$results$sigma$s.effects$`s(DISTRITO)`$`50%`,
108 id=zinb1$results$sigma$s.effects$`s(DISTRITO)`$DISTRITO,
109 pos="topleft", color=heat_hcl, lrange=c(-0.40,0.40))
110
111 #model performance and diagnosis
112
113 summary(zinb1) #check DIC and pD
114 plot(zinb1, which="samples") #get samples of parameters
115 plot(zinb1, which="qq-resid") #get residual plot
116 plot(zinb1, which="max-acf", lag=200) #get maximum autocorrelation plot
117
118 # remove possible outliers for Anos_Carta (analogouly for other covariates)
119 outliers<- boxplot(dtest3$Anos_Carta, plot=FALSE)$out
120 dtest3[-which(dtest3$Anos_Carta %in% outliers),]
121
122 #PREDICTION
123
124 #generated data set with 29 observations
125 nd<- data.frame(
126 Anos_Carta=c(1,0,2,1,2,1,0,1,1,10,20,22,15,20,22,15,12,1,45,40,52,35,
127 35,33,7, 0,0,0,0,0),
128 Idade_Veiculo=c(1,0,5,1,2,2,0,1,1,4,4,8,6,4,8,6,8,1,18,15,23,12,13,16,
129 8,0,0,0,0,3),
130 Idade_Condutor=c(19,18,21,20,23,25,18,20,19,29,42,50,38,42,50,38,35,19,
131 63,65,72,60,55,53,26,18,18,18,18,18),
132 DISTRITO=c("LISBOA", "LISBOA", "SETUBAL", "LISBOA","LISBOA","LISBOA",
133 "PORTO", "PORTO", "COIMBRA", "BRAGA", "LISBOA","LISBOA","LISBOA",
134 "PORTO","PORTO", "AVEIRO","VISEU","FARO","PORTALEGRE", "EVORA", "BRAGANCA",
135 "BEJA", "LISBOA","LISBOA","LISBOA","LISBOA", "PORTO", "LISBOA","PORTO","SETUBAL"),
136 ESCALAO_CILINDRADA=c("3","3","2","2","2","2","3","2","3","2","2","2","2","2",
137 "2","2","2","2","2","2","2","2","3","2","2","3","3","3","3","2"),
138 CATEGORIA_AGREGADA=c("LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO",
139 "LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO",
140 "LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO",
141 "LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO","LIGEIRO",
142 "LIGEIRO","LIGEIRO","LIGEIRO","MISTOS" ),
```

```
 DESCR_SEXO_PESSOA=c("MASCULINO","FEMININO","FEMININO",
 "MASCULINO","MASCULINO","MASCULINO","MASCULINO","MASCULINO",
 "FEMININO", "MASCULINO","MASCULINO","MASCULINO","MASCULINO",
 "MASCULINO","MASCULINO","FEMININO","MASCULINO","MASCULINO",
 "MASCULINO","MASCULINO","MASCULINO","MASCULINO","FEMININO",
 "MASCULINO","MASCULINO","MASCULINO","FEMININO", "MASCULINO",
 "FEMININO","MASCULINO"),
 Marca_Conv=c("EUA","FRA","GER","OP","GER","GER","OP","GER","GER",
 "GER","OP","JAP","ITA","FRA","GER","GER","GER","FRA","FRA","GER",
 "OP","OP","EUA","OP","OP","FRA","OP","FRA","OP","OP"),
DESCR_TIPO_USO=c("VIDA_PRIVADA", "VIDA_PRIVADA","USO_PART","USO_PART",
"USO_PART","USO_PART","USO_PART","USO_PART","USO_PART","USO_PART",
"USO_PART","USO_PART","USO_PART","USO_PART","USO_PART","VIDA_PRIVADA",
"USO_PART","USO_PART","USO_PART","USO_PART","USO_PART","USO_PART",
"USO_PART", "VIDA_PRIVADA","VIDA_PRIVADA","VIDA_PRIVADA", "USO_PART",
"VIDA_PRIVADA","USO_PART","VIDA_PRIVADA"),
Apol_20112013=c(2,2,2,1.5,1.5,1.5,1.5,2.2,1.3,1.8,2,1.5,2,2,1.5
,1.5,2,1,1.5,2.5,2,1.5,2,2,2,2,2,0.5,0.5,1))

#predictions
model <- predict(zinb1, newdata=nd,type="parameter", FUN=function(x){x})

#no claim probability - posterior means
ncp<- function(model, n, m){i=1; res2=matrix(0, n, 1);
for(i in 1:n){ j=1; delta=matrix(0, m, 1);res=matrix(0, m, 1);
for (j in 1:m){delta[j]=1/(model$sigma)[i,j];
res[j,1]=(model$nu)[i,j]+
(1-(model$nu)[i,j])*(delta[j]/(delta[j]+(model$mu)[i,j]))^(delta[j])};
res2[i,1]=mean(res)};
return(res2);}


ncp(model,30,1001)

#expected claim frequency - posterior means
ec<-function(model, n, m){i=1; res2=matrix(0, n, 1);
for(i in 1:n){ j=1;res=matrix(0, m, 1);
for (j in 1:m){ res[j,1]=(1-(model$nu)[i,j])*(model$mu)[i,j]};
```

```
181 res2[i,1]=mean(res)};
182 return(res2);}
183
184 ec(model, 30, 1001)
```