



Network for Studies on Pensions, Aging and Retirement

A comparative analysis of statistical models for the pricing of health insurance

Miguel Degeneffe

MSc 07/2020-012

NETSPAR ACADEMIC SERIES

A comparative analysis of statistical models for the pricing of health insurance.

Predicting pure premiums with the use of LM, GLM and GAM applied on a Belgian hospitalization product.

Maastricht University
Master thesis in Econometrics & Operations Research
Name: Miguel Degeneffe
Student ID: i6106629
Supervisor: Prof. Dr. Antoon Pelsser
Date submitted: July 10, 2020

Abstract

Actuaries frequently classify policyholders in terms of their risk potential in order to price insurance contracts, with the goal of balancing fair tariffs with guarding against expected losses. Statistical models enable actuaries to determine pure premiums by applying the frequency-severity approach. Accordingly, this thesis investigates the use of standard linear models (LM), generalized linear models (GLM), and generalized additive models (GAM) for the pricing of insurance contracts. Specifically, this paper explores the properties of these models, their purpose, and then studies their ability to deal with continuous and spatial variables. Methodologically, we construct a framework to demonstrate how actuaries can take these variables into account in the different models and how one can compare the statistical models and their associated premiums. We construct a case study with data from a Belgian hospitalization insurance and apply the corresponding models to predict the severity and frequency of the claims, and obtain pure premiums. We conclude that a log-normal LM for claim severity and a negative binomial GLM for claim frequency lead to the preferred premium structure in terms of predictive performance and fair pricing. Furthermore, we notice that the use of GAM points to a similar premium structure.

Acknowledgement

At the end of every achievement there comes the important moment of thanking everyone who was part of it and made it happen. In my case, the achievement is this master thesis which forms the culminating point of my studies at the University of Maastricht. Although it really is a cliché, it is nevertheless true that I couldn't have done this without every one of these people.

First of all, I would like to thank my supervisor, prof. dr. Antoon Pelsser. It is thanks to his feedback that I was able to lift my thesis to a higher level.

Further, I render thanks to KMPG Belgium for giving me the opportunity to combine my thesis with an internship. Special thanks to Daphne, my coach, for helping me find an interesting research topic and to Jeroen and Cédric for providing and introducing me to the data.

Last but not least I am grateful for all the love, time and support I received from my parents and girlfriend Elise. Not only during this year but throughout my whole time at Maastricht University, they were an endless support, which deserves a very big thank you.

Contents

1	Introduction	5
2	Introduction to Health insurance	7
3	Concepts and models for insurance pricing	8
3.1	Fair valuation of insurance assets and liabilities	8
3.2	Premium principles	9
3.3	Pure Premium	10
3.4	Assumptions	11
3.5	Risk factors	12
3.6	Probability distributions	13
3.6.1	Claim severity	13
3.6.2	Claim frequency	14
4	Statistical Models	15
4.1	Standard linear models	15
4.2	Generalized Linear Models	16
4.3	Generalized Additive Models	17
4.3.1	Smoothing splines	18
4.3.2	Smoothing parameter	19
4.4	Trade-offs: dealing with non-linear effects of continuous and spatial variables	20
5	Approach	21
5.1	Model specification	22
5.1.1	Probability distribution	22
5.1.2	Variable selection	23
5.2	Clustering continuous and spatial variables	23
5.3	Model comparison	25
5.3.1	Measures of fit	25
5.3.2	Prediction accuracy	26
5.3.3	Measures of lift: the ordered Lorenz curve and Gini index	27
6	Database & Descriptive statistics	28
7	Claim Severity analysis	32
7.1	Base model	32
7.2	GAM	34
7.3	LM with clustered spatial effects	35

8	Claim Frequency analysis	36
8.1	GLM	37
8.2	GAM	37
9	Comparative analysis	38
9.1	Claim severity and frequency models	38
9.2	Premium models: Inflow and predictive accuracy	39
9.3	Premium models: fair tariffs	40
10	Conclusion	41
11	Appendix	46
11.1	Figures	46
11.2	Tables	48

1 Introduction

Every year, a large number of Belgians require hospitalization services. As per a report by the Federal Public Service of Health (2019), in 2017 alone, 6,214,325 hospital visits and stays took place. The expenses of even a brief stay at the hospital may add up for a patient. Fortunately, a portion of these costs are reimbursed by accredited Health Insurance Funds or by the Auxiliary Illness and Disability Insurance Fund (Hulpkas voor Ziekte- en Invaliditeitsverzekering). Health insurance is part of the social security system in Belgium; every citizen is obligated to have health insurance and, therefore, must choose one of the aforementioned funds. Nevertheless, these funds seldom cover all expenses. The supplemental fees can be surprisingly high if you, for example, stay in a private room. Additional hospitalization insurance can intervene here, limiting the amount that a patient ultimately might have to pay on their own. Nearly 80% of Belgians have this type of hospitalization insurance (Assuralia, 2019). In 2019, there were 24 private health insurers active in Belgium (Nationale Bank van België, 2019).

If these insurance companies use flat premiums across the entire portfolio, the resulting tariffs encourage groups with lower than average health risk levels to terminate their contracts and accept a better offer at a competing company. This results in a phenomenon known as adverse selection in which the insurer is left with high-risk groups which pay too low of a premium. To avoid such a situation, insurance companies use risk factors to group policyholders with similar risk profiles into classes (such that the corresponding premiums reflect the risk level of each class as closely as possible) (Henckaerts et al., 2017). In this way, insurers are able to minimize the information asymmetry within the portfolio in a way that allows for them to successfully price insurance plans according to the risk of the individuals; as a result, they can attract low-risk individuals by offering them insurance at lower premiums.

A well-known approach to calculate the pure premium, which is equal to the expected loss, is to combine the expected claim frequency with the expected claim severity, estimated by the observed risk factors. In order to measure the effects of a set of risk factors and set corresponding tariffs for different risk classes, actuaries are in need of statistical models. The use of standard linear regressions is a well-known technique favored by many econometricians when measuring the impact of risk factors on a ‘phenomenon of interest’. However, over the last few decades, generalized linear models (GLMs) have been the standard for risk classification in the insurance setting and have been the subject of substantial amounts of research in the field of pricing non-life insurance (e.g., Denuit et al. (2007); Antonio and Valdez (2011)). Researchers often prefer the utilization of these models (GLMs), mainly due to the fact that the response variable is allowed to have a probability distribution different from the normal, which is often necessary for the data being dealt with in insurance. Furthermore, generalized additive models (GAMs), which are an extension of GLMs, have been shown to be useful for insurance pricing (Denuit et al., 2019). These models allow for the modeling of continuous variables in a non-parametric way and are, therefore, a helpful tool in identifying non-linear

effects of continuous risk factors (e.g., age).

Important risk drivers that influence the frequency and costs of hospital visits are, for example, the age, medical background, and place of residence of the insured. Many of these variables can be seen as categorical and can be modeled by the use of binary variables. However, the age and geographical information of the policyholders, which are expected to be crucial determinants that influence the pure premium, are classified as continuous variables. These variables can be interpreted as categorical factors with a large number of levels (Ohlsson and Johansson, 2010). However, including them as categorical factors in statistical models may lead to unreliable results in cases where sparse data are available for some of those levels. More optimal options to incorporate continuous variables would be either (1) to model them in a GAM, or (2) to implement a classification method that can transform them into categorical variables (see Denuit et al. (2019) for the first and Henckaerts et al. (2017) for the latter approach).

This Master's thesis will analyze the use of the introduced statistical models (LM, GLM, & GAM) with a focus on pricing of a hospitalization insurance policy with the aim of finding a pricing structure that would enable fair tariffs sufficient to cover the total loss. This leads to the following research question:

What are the differences between a parametric (LM or GLM) and a non-parametric (GAM) modeling approach in light of the hospitalization product being investigated, i.e., which premium model leads to the most accurate results?

For the purpose of answering this research question, the following sub-questions were formulated: (1) Which probability distributions result in the best fit for the frequency and severity data? (2) Does the predictive performance of the models increase by including spatial risk factors? (3) How can we compare the different models, and their corresponding premiums, from both a statistical and an actuarial perspective?

This paper seeks to answer these questions by providing the theoretical background of LMs, GLMs, and GAMs, and the concepts and use of these models in tariff setting in non-life insurance (i.e., the pure premium and frequency-severity approach). Furthermore, we thoroughly examine the possible options for incorporating the effects of continuous risk factors into the models and investigate how one can compare the results of these models. Afterwards, several models are applied to the data set: (1) a simple parametric model with the available categorical risk factors (which can be seen a starting point), (2) a GAM, and (3) a LM or GLM that incorporates the effects of the continuous and spatial variables (by clustering them into a set of categorical variables). After the creation of different models, the models and their results are compared with the aim of answering the research question of this thesis.

In terms of organizational structure, this thesis assumes the following form: In Chapter 2, a brief introduction about hospitalization and its background in health insurance is provided to give a general overview about the insurance cover being investigated. Then, in Chapter 3,

concepts and models central in insurance pricing are introduced and explained. An introduction to the predictive models is provided in Chapter 4. This chapter focuses on the models studied in this thesis: standard LM, GLM and GAM. Afterwards, in Chapter 5, the research methodology used in this study is presented. Next, in Chapter 6, will be a brief overview of the data of the insurance product used in this thesis. Having visualized the data, in the following two sections (Chapter 7 & 8), the thesis constructs and analyzes the predictive models we would like to compare for both claim severity and frequency. An evaluation of the models and a statistical comparison is presented in Chapter 9. Finally, a conclusion, the limitations of the study and alternative approaches are discussed with regard to future research within the field.

2 Introduction to Health insurance

This paper illustrates the methodology and comparison of the statistical models in the context of a hospitalization insurance. Therefore, before we consider the relevant theory for this thesis, we will briefly provide the necessary background information about health insurance, and more specifically, hospitalization insurances.

Broadly speaking, health insurance refers to a large class of insurance products that provide benefits in case a need arises due to either illness or accident, and leads to either partial or total loss of income which can be permanent or non-permanent. It also covers expenses such as those made for hospitalization, surgery, medicine, nursery, rehabilitation, etc. (Pitacco, 2014).

Pricing health insurance relies, to a large extent, on both non-life and life actuarial fields, depending on the type of coverage. First of all, health insurance has characteristics of a non-life insurance product, such as the fact that the premium calculations for all types of health insurances depend on the claim frequencies. Furthermore, in case of expense reimbursements, they also depend on the claim costs, and even on the degree of disability for cases where the benefits are graded (Pitacco, 2014). Life-insurance pricing facets come into play when pricing multi-year or lifelong health products. In these cases, mortality assumptions are crucial for setting premiums; life tables and mortality laws can be used to allow for the unknown future mortality trend, which implies the presence of aggregate longevity risk (i.e., the risk that policyholders live longer or shorter than expected). Furthermore, life-insurance pricing takes into account the time-value of money (i.e., an interest assumption), which comes back for rating multi-year health products.

The insurance, which is discussed from Chapter 6 onwards, mainly covers hospital visits and medicines that are not reimbursed by the basic health insurance funds and consist of one-year coverages. Premium calculations for pricing one-year coverage have features typically associated with non-life insurance, and therefore, relevant concepts for non-life insurance come back in this paper.

3 Concepts and models for insurance pricing

“Pricing is never a disconnected bunch of techniques, but more-or-less a formalized process that starts with collecting information about a particular policyholder or risk and ends with a commercially informed rate,” Pietro Parodi (2015) once said. Through this statement, he elucidated that pricing “is a complex endeavor, which is best conceptualized as a process.” Pricing of insurance utilizes ideas from different areas of mathematical finance and actuarial science. This thesis focuses on a statistical modeling approach to rate premiums based on the policyholder’s risk characteristics, which is just one small step in the overall process.

This chapter provides a brief background concerning the fair valuation of insurance assets. Subsequently, we focus on premium principles and introduce the concept of the pure premium that we link to the claim frequency-severity approach; a risk-based pricing method with which the expected claim amount and count for each policyholder is estimated. Concepts key to this approach are provided: we discuss the necessary assumptions; investigate the type of risk factors, which are used to predict claim severity and frequency; and lastly, discuss commonly used probability distributions for the response variables.

3.1 Fair valuation of insurance assets and liabilities

For actuaries, the subject of how to assess the fair valuation of insurance assets and liabilities, is becoming more pertinent day by day. A lot of modern solvency regulations, such as Solvency II, call for a fair valuation of assets and liabilities. A well-known definition of the fair value is the “amount at which an asset or a liability could be exchanged between knowledgeable and willing parties in an arm’s length transaction.”¹

Dhaene et al. (2017) gave us a more specific definition by defining it as an actuarial and market-consistent valuation; it is: (1) market-consistent meaning that any hedgeable part of a claim is valued at the price of its hedge, and (2) actuarial in the sense that claims with payoffs that are independent of the evolution of asset prices are valued taking into account actuarial judgment. Hence, a fair valuation combines the two important elements: the financial approach of a market-consistent valuation and the actuarial approach of an actuarial valuation.

The valuation of hedgeable claims under a market-consistent valuation is consistent with risk-neutral pricing. In this case, risk neutral pricing is based on an $E[\text{equivalent}] M[\text{artingale}] M[\text{easure}] \mathbb{Q}$. On the other hand, an actuarial valuation is generally performed with an actuarial premium principle (based on a true probability measure \mathbb{P}). The problem the actuary is solving in such a setting, is to value the claim in a way that the observed claim amount at the end of the period will be able to be paid by the insurer, ignoring the existence of a financial market (Dhaene et al., 2017).

¹The arm’s length principle is the condition that all parties engaged in a transaction are perceived to be acting independently and on an equal footing.

Several fair valuation techniques exist; Dhaene et al. (2017) discusses the class of convex hedge-based (CHB) valuations, where the sum of the financial market price of the hedge and the actuarial value of the remaining claim represents the fair value in a one-period setting. Another approach, proposed by Pelsser and Stadje (2014), is the use of two-step valuations, which can be used to apply a market and time consistent evaluation. In the following section, we provide more information about the concept of premium principles, used in the actuarial valuation, and more specifically introduce the pure premium principle, which forms the basis of this thesis.

3.2 Premium principles

Kaas et al. (2008) define premium principles as a group of mathematical methods used to calculate the premium from the probability distributions of the claims. Under the conditions of validity of the law of large numbers, which proposes that variability around the mean decreases when the number of observations increases, the total premium income should be at least equal to the expected total claim amount. Moreover, a loading, which is used for building reserves, should be added to compensate the insurance company for being in a ‘less safe’ position and to avoid ruin (Kaas et al., 2008). Furthermore, next to the capital which is needed to ensure solvency, the loading consists of the amount required to cover for other expenses (e.g., administrative costs) and the profit margin, allowing for a profit. Premium principles, which are used for finding a minimum premium covering the claim costs to which the loading is added, and the assessment of the proper degree of capital required to guarantee solvency frames the actuarial aspect of insurance pricing.

Various well-known premium principles exist, such as the equivalence principle, expected value principle, the standard deviation principle, exponential principle, etc. (for more information about these principles, see for instance Kaas et al. (2008)). For the purposes of this master’s thesis, we aim to calculate the individual rates for each policyholder. As discussed in the introduction, it is desirable to charge premiums based on the policyholder’s risk factors (also called risk classification) as the use of flat premiums across the entire portfolio would encourage those with profiles that pay over-charged premiums to terminate their contracts. Therefore, we work with the pure premium principle. The pure premium of an insurance product refers to the portion of the total insurance rate that is needed to cover the losses. To acquire from this premium the gross rate, the rate that is actually charged to the insured, the loading is added to the pure premium. The paper further endeavours to precisely predict this pure premium without looking deeper into the actuarial concepts of loadings. So, in the next section, we will take a closer look at the pure premium concept.

3.3 Pure Premium

Following from the previous section, the pure premium accounts for the economic risk that is transferred from the policyholder to the insurer. Based on the the law of large numbers, the total loss of the insurance company, which is the sum of a large number of comparatively small independent losses, should be better to predict than the value of an individual loss (Ohlsson and Johansson, 2010). This means that the actual loss should not be excessively far from its expected value. Therefore, actuaries define the pure premium as the expected cost of all claims that policyholders will file during the coverage period (Denuit et al., 2007). Before specifying the pure premium, we will introduce claim frequency (F_i) and claim severity (S_i) for each policyholder i . Claim frequency stands for the ratio of the total number of claims N_i reported during a policy period to the total exposure t_i :

$$F_i = \frac{N_i}{t_i} \quad (1)$$

where t_i refers to the fraction of the policy period for which the policyholder is insured (e.g., a t_i of 1 represents a complete year when N_i is given for an entire 1 year contract). Claim severity (S_i) represents the average claim amount, expressed as the ratio of the total loss (L_i) for each policyholder to the corresponding number of claims causing this total loss (N_i):

$$S_i = \frac{L_i}{N_i}. \quad (2)$$

Now, the pure premium (p_i) for policyholder i can be defined as the multiplication of the claim frequency (F_i) times the claim severity (S_i):

$$p_i = \frac{L_i}{t_i} = \frac{N_i}{t_i} \frac{L_i}{N_i} = F_i S_i \quad (3)$$

Assuming independence between claim frequency and claim severity, the expected value principle for p_i can be used to show that:

$$\pi_i = E(p_i) = E(F_i)E(S_i) \quad (4)$$

Hence, multiplying the estimates of both response variables, obtained by statistical models, would underlie the pure premium to be charged for each risk class of policyholders. Therefore, regression techniques are vital for actuaries to estimate both the expected frequency and severity from historical data.

Another possible way to determine pure premiums, instead of building two separate models for claim severity and claim frequency, is by modeling the expected pure premium ($\frac{L_i}{t_i}$) directly. We do not opt for this approach in this paper as focusing on a frequency-severity approach has several advantages (Brockman and Wright (1992), Goldburd et al. (2019)):

- (1) Greater insight into the underlying risk factors contributing to claim severity and frequency is provided. When modeling pure premiums directly, some interesting effects may disappear due to counteracting effects on its components: for instance, a variable that is severely negatively affecting frequency but is equally positively impacting severity would go totally unnoticed.
- (2) Claim severity and frequency usually follow different distributions.
- (3) It will be easier to observe the effects and make preliminary conclusions from the data.

3.4 Assumptions

In this subsection, we briefly delve into the assumptions made in the previous chapter as it is necessary to discuss the possible concerns specific to the current insurance environment.

Assumption 1. *Let n be the number of policyholders and X_i denote the response for policyholder i . Then X_1, \dots, X_n are independent.*

Assumption 2. *Let t denote the number of policy periods and X_j denote the response for policy period j . Then X_1, \dots, X_t are independent.*

For applying the law of large numbers, we assumed independence between the policyholders. Furthermore, independence assumptions come back in Sections 3.6 and 4. Assumption 1 states that there exists independence between individual policyholders. This assumption in non-life insurance, and more specifically in health, is not always valid as dependencies among policy holders may exist. In the present era, it is relevant to consider the spread of a disease (e.g., COVID-19), which creates a dependence as many policyholders are affected. Ohlsson and Johansson (2010) state that one can exclude these extreme possibilities as these circumstances call for other types of methods than those used for risk-classification pricing, such as the acquisition of reinsurance, to reduce their impact.

Additionally, Assumption 2, which requires independence over time periods for a given policyholder, is typically not perfectly fulfilled; for instance, dependence may exist when a policyholder suffers from a long-term illness and therefore returns to the hospital multiple times over subsequent years. For this thesis, this assumption can be dropped as our research focuses on cross-sectional data of one policy year. Notwithstanding, researchers working with panel data have to keep this assumption in mind.

Assumption 3. *Let n be the number of policyholders and S_i and F_i denote the claim severity and frequency respectively for policyholder i . Then S_i and F_i are independent of each other for each policyholder i .*

Lastly, Assumption 3, assumed in equation (4) for frequency-severity modeling, states that the number of claims does not influence the claim amounts of a policyholder i , and vice versa. For instance, upon observing many claims, they generally do not contain any information as to whether these cases are of smaller or larger size.

This assumption may in some cases be violated. See Gschlossl and Czado (2007) and Shi et al. (2015) for different models where one relaxes the independence assumption between claim severity and frequency.

3.5 Risk factors

The response variables for each policyholder are estimated by including risk factors as explanatory variables in the models. Risk factors can be either categorical or continuous. A categorical variable can be seen as a variable that takes one of the options between a limited number of possibilities (e.g., gender with 2 possible outcomes). Continuous variables, on the other hand, can take any value in a consecutive interval. Within the group of continuous variables, we define a subgroup of spatial variables, which are represented by the postal codes and their corresponding coordinates.

Furthermore, several variables can be implemented in the models as both categorical and continuous; for instance, age can be taken into account as a categorical variable when several age groups are clustered or as a continuous variable. This paper will investigate the use and effect of these continuous and categorical variables in the coming chapters.

The most common risk factors used in determining health insurance tariffs are:

- Age
- Additional information about the policyholder: marital status, BMI, tobacco use, pre-existing illness
- Information about the geographic region in which the policyholder resides (spatial risks)
- Gender

Although used in the past, setting different prices based on gender is no longer allowed.² However, displaying the gender-based effects in the models remains valuable. Therefore, actuaries often make a distinction between risk factors, which are expected to have effect on claim frequency and severity, and rating factors that are actually used for determining the tariffs. In this thesis, the main focus lies on risk factors, nevertheless, in practice, for actual pricing the strict legislation must be taken into consideration.

The risk factors concerning information about (1) the policyholders, (2) the specific type of coverage and (3) the geographic region where the policyholders reside of the hospitalization insurance which is being studied in this thesis, are discussed in Chapter 6.

²On 1 March 2011, the European Court of Justice concluded that any discrimination on grounds of sex is prohibited and declared that, since 21 December 2012, equality between men and women has to be guaranteed in the European Union (EU) (European Union, 2017).

3.6 Probability distributions

When the pure premium is calculated by the multiplication of the expected claim severity and frequency, both variables must be estimated individually. To this end, it is enlightening to discuss their properties, with the goal of modeling them.

3.6.1 Claim severity

Data from healthcare claim costs generally follow a non-negative and right-skewed distribution. This follows from the fact that costs have a lower-bound ($cost > 0$), but ranges widely on the upper end. Therefore, the gamma distribution, the inverse Gaussian and the log-normal distributions often do well when modeling claim severity. The most widely used distribution to model claim severity in practice is the gamma distribution (Murphy et al., 2000).

The probability density function (pdf) of a gamma distribution is given by:

$$f(x_i) = \frac{\beta_i^{\alpha_i} x_i^{\alpha_i-1} e^{-\beta_i x_i}}{\Gamma(\alpha_i)} \quad x_i > 0, \alpha > 0, \beta > 0 \quad (5)$$

where α is the index parameter and β the scale parameter. The distribution function (5) can be utilized when modeling a single claim cost X_i . Moreover, sums of independent gamma distributions with equal β are gamma distributed with the same scale parameter β and an index parameter, which is the sum of the individual α (Ohlsson and Johansson, 2010). Therefore, the total claim amount, L_i , which is the sum of N_i independent gamma distributed single claim costs, X_i , is gamma distributed $L_i \sim Gam(N_i \alpha_i, \beta_i)$. Hence, the pdf for claim severity $S_i = \frac{L_i}{N_i}$ is given by:

$$f_{S_i}(s_i) = N_i f_{L_i}(N_i s_i) = \frac{N_i \beta_i^{N_i \alpha_i} s_i^{N_i \alpha_i - 1} e^{-N_i \beta_i s_i}}{\Gamma(N_i \alpha_i)} \quad (6)$$

with $S_i \sim Gam(N_i \alpha_i, N_i \beta_i)$.

Furthermore, this paper looks at the log-normal distribution. When X is normally distributed, it follows that $Y = e^X$ has a log-normal distribution with the pdf given by:

$$f_Y(y) = \frac{f_X(\ln y)}{y} = \frac{1}{y\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{\ln y - \mu}{\sigma} \right)^2 \right\}; \quad y > 0, \lambda > 0 \quad (7)$$

with $Y \sim Lognormal(\mu, \sigma^2)$.

Lastly, the pdf of an inverse-Gaussian distribution is given by:

$$f_Y(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp \left[-\frac{\lambda(y - \mu)^2}{2\mu^2 y} \right]; \quad y > 0, \lambda > 0, \mu > 0 \quad (8)$$

with $Y \sim (\mu, \lambda)$.

3.6.2 Claim frequency

We now turn to claim frequency, the ratio of the total claim count over the total exposure. Let N_i stands for the number of claims reported by a policyholder i . It is possible to model claim counts by assuming that the exposure of each policyholder is the same ($= 1$), which stands for the complete policy period of one year. However, it is hardly realistic in practice; as an example, two different policyholders with different lengths of insurance coverage (e.g., 1 month and 12 months) could have recorded the same number of hospitalizations. As the expected number of claim counts would be proportional to the length of coverage, actuaries should not treat these two policyholders' loss experiences identically in the modeling process. Therefore, it is crucial to model claim frequency, which takes into account the total exposure (t_i).

When looking at probability distributions for count data, several possibilities are available for modeling N_i and F_i . The most commonly used distribution for response variables that are counts is the Poisson distribution $N_i \sim POI(t_i\mu_i)$. When $t_i = 1$, then $E(N_i) = t_i\mu_i$ and the probability distribution function can be expressed as:

$$f_{N_i}(n_i) = \frac{e^{-t_i\mu_i} t_i\mu_i^{n_i}}{n_i!} \quad (9)$$

Claim frequency can be expressed as $F_i = N_i/t_i$. Then the probability distribution function of F_i is given as follows:

$$f_{F_i}(f_i) = \frac{e^{-t_i\mu_i} t_i\mu_i^{t_i f_i}}{t_i f_i!} \quad (10)$$

A Poisson distribution is a suitable choice when one is able to assume that the claims of the policyholders occur independently and at a constant rate. However, in practice, the assumptions may not hold as the expected claim frequency of a policyholder may vary over time and claims might not always be independent of each other. This phenomenon when the variance is larger than the mean is called overdispersion.

A more feasible option in practice to avoid overdispersion might be the use of mixed distributions, where the mean of the Poisson distribution itself is treated as random variable modeled from another distribution. An often choice of probability distribution for the mean, is the gamma distribution (with $y \sim POI(\mu)$ and $\mu \sim gamma$). This is equivalent to a negative binomial distribution for y (Goldburd et al., 2019).

4 Statistical Models

Statistical models are vital ‘tools’ for actuaries that enable them to determine pure premiums by applying the frequency-severity approach. This section contextualizes and describes statistical models that are relevant for risk-based pricing in non-life insurance contracts.

Linear models – such as the standard linear regression model, or ordinary linear model (LM) and the generalized linear model (GLM) (which is a generalization of the ordinary linear regression model that provides extra flexibility) – involve a response variable and a set of predictor variables. Due to their relative simplicity researchers often use linear models to solve multiple regression problems. Essentially, the predictor effects are modeled in the form of a linear predictor. While they are advantageous for their efficiency in dealing with categorical risk factors, the models may fail for continuous risk factors with non-linear effects. To that end, generalized additive models (GAM) are suitable alternatives. They enable a more flexible relationship, as the effects that the continuous predictors have on the response can be captured through smooth functions.

To explore the utility and differences of each modeling technique in determining future premiums, we will examine and assess the different models (i.e., LM, GLM, and GAM) in the following subsections.

4.1 Standard linear models

The ordinary linear regression model is a means of modeling the relationship between a variable whose outcome we wish to predict (i.e., claim severity and frequency) and one or more independent explanatory variables (i.e., the risk factors). The linear relationship between the dependent variable y_i and the independent variables x_1, \dots, x_p can be written as:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon_i; \quad i = 1, 2, \dots, n \quad (11)$$

where β_0 stands for the intercept. ϵ_i represents the error term, which can be interpreted as a random noise accounting for all underlying non-systematic effects that contribute to the measurement error of the model. In matrix notation, this relationship is given as:

$$Y = X\beta + \epsilon \quad (12)$$

where $X\beta$ is called the linear predictor. Using $E(\epsilon) = 0$, it follows that $E(Y|X = x) = X\beta$. Note that the linearity constraint for linear regression only applies to the coefficients β , which means any non-linear relation among explanatory and response variables may still be modeled. Polynomial regression, which is a form of the linear regression, fits a nonlinear relationship between the value of x and the corresponding conditional mean of y .

It is assumed that the errors of the response variables are uncorrelated with each other, which can occasionally be a problem in practice (see section 3.4). Furthermore, one of the

issues with the standard linear model is that it requires the response variable to follow a normal distribution. Response variables in insurance are often right-skewed; a solution could be to find a transformation that normalizes the data to allow analysis using the normal linear model (e.g., a log transformation of claim severity data usually results in a distribution that is more symmetric and closer to normality) (De Jong and Heller, 2008).

4.2 Generalized Linear Models

In cases where the standard linear model offers insufficient flexibility, even with a log-transformation of the data, it may be beneficial to utilize a methodology that has been further developed to enable modeling of non-normal data and more complex interactions between predictor variables.

Generalized linear models (GLMs), as their name suggests, are a generalization of the standard linear regression models (11). These models are based on the normal distribution extended to the exponential class of distributions (e.g., the normal, Poisson, binomial and gamma distributions). The class of GLMs was introduced by Nelder and Wedderburn (1972) as a general framework for handling a range of common statistical models for normal and non-normal data and can be seen as the industry standard for modeling the relation between the response and predictor variables. GLMs in the insurance setting are thoroughly reviewed in Kaas et al. (2008), De Jong and Heller (2008), Ohlsson and Johansson (2010) and Denuit et al. (2019).

A GLM is defined by three components: a random component, that specifies the probability distribution within the exponential family for the response variable Y ; a systematic component, that relates the parameter η to the predictors X , which is called the linear predictor $\eta = \beta X$; and a link function $g(\cdot)$, which must be monotone and differentiable, that connects the random and systematic components (Tibshirani, 2014). The relationship between the dependent variable and its predictors can be represented as follows:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^z \beta_j x_{ij}; \quad i = 1, 2, \dots, n \quad (13)$$

where the intercept is denoted by β_0 , the explanatory variables by x_{i1}, \dots, x_{iz} and its coefficients by $\beta_{i1}, \dots, \beta_{iz}$.

Hence, the mean of the response variable can be written as:

$$\mu_i = E(Y_i) = g^{-1}(\eta_i) \quad (14)$$

Therefore, the two principle distinctions between the generalized linear model (13) and the ordinary linear regression (11) are:

- In the linear model, the mean is a linear function of the explanatory variables, while in GLMs “some monotone transformation of the mean, which is represented by the

link function $g(\cdot)$, is a linear function of the explanatory variables with linear and multiplicative models as special cases” (Ohlsson and Johansson, 2010).

- In GLMs, the response variable can follow any distribution that belongs to the exponential class of distributions, while in the linear model the distribution of the response variable is restricted to the normal distribution.

Within insurance settings, the generalized linear model is often preferred over the ordinary linear regression for several reasons. First, the assumption of having a normal distribution in linear regression models is frequently not fulfilled, as the response variable (severity and frequency) tends to have other distributions. Second, for insurance pricing, it is often more reasonable to use multiplicative models (Murphy et al., 2000).

That being said, in practice the ordinary linear regression often performs well for claim severity data. Moreover, when using GLMs, it is possible to use the standard linear model as a special case, namely when defining an identity link $g(\mu_i) = \mu_i$ with a normal distribution. In general, with the assumption that the response variables y_1, \dots, y_n are independent, it may be assumed that Y has a probability density function, or a probability mass function when the distribution is discrete, of the form:

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right) \quad (15)$$

where ϕ is the dispersion parameter and is fixed for all i . θ_i stands for the natural parameter, which can depend on i . The cumulant function $b(\theta_i)$ is assumed to be twice continuously differentiable for which the second derivative is invertible. Each choice of this function leads to a distribution that is part of the exponential family of probability distributions (Ohlsson and Johansson, 2010). The probability distributions introduced in Chapter 3.6, with the exception of the log transformation of a normal distribution, belong to this exponential family of probability distributions.

When modeling claim severity and frequency, the link function $g(\cdot)$ is usually specified by the natural log function ($g(x) = \ln(x)$), as it allows a multiplicative relationship between the dependent and independent variables. When a log-normal distribution is specified for claim severity, the identity link function is recommended (i.e., the ordinary linear model).

4.3 Generalized Additive Models

In the 1980’s, the issue of how to examine the effect of continuous factors was the object of much research and prompted an augmentation of GLMs known as generalized additive models (GAMs), introduced by Hastie and Tibshirani (1986). In their model, properties of GLMs were combined with those of additive models. As with the GLM discussed in the previous section, a GAM depends on a link function to relate the mean of the response to

the explanatory variables, and the probability distribution of the response also belongs to the exponential family. However, in the ordinary and generalized linear models, a linear or parametric relationship between the response variable and explanatory variables is captured.³ While using a GAM, it is possible to model non-linear effects of the explanatory variables on the response variable as the linear predictor of the form $\beta_i x_i$ is replaced by an additive predictor of the form $f_i x_i$. These estimated functions f with an unspecified, non-parametric, functional form may reveal possible non-linearities in the effect of the predictors if present (Hastie and Tibshirani, 1986). This makes the GAM especially useful in the insurance setting, as uncertainty often exists about the effects of continuous and spatial variables (e.g., ages and postal codes, respectively).

In the GAM, the relationship between the response variable and its predictors is represented as follows:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^z \beta_j x_{ij} + \sum_{j=1}^q f_j y_{ij} + \sum_{j=1}^r f_j(v_{ij} w_{ij}) \quad (16)$$

where the intercept is denoted by β_0 and the categorical variables by x_{i1}, \dots, x_{iz} , which are modeled in the same way as in the GLM. New compared to the GLM are the univariate $(f_1(y_{i1}), \dots, f_q(y_{iq}))$ and bivariate $(f_1(v_{i1} w_{i1}), \dots, f_r(v_{ir} w_{ir}))$ smooth functions which capture the effect of the continuous and spatial predictors non-parametrically. Bivariate smooth functions $f(v_i w_i)$ are used to model interaction effects and spatial effects (i.e., latitude and longitude coordinates).

To summarize, the model described in (16) can be called semi-parametric as the predictors x_{i1}, \dots, x_{iz} are modeled linearly, while the predictors y_{i1}, \dots, y_{iq} non-parametrically.

4.3.1 Smoothing splines

Splines are used to estimate the unknown functions for modeling the smooth effects. The methodology behind splines involves taking a number of polynomials, specified on a set of different intervals (instead of a single polynomial over the complete domain), and placing continuity constraints at the points, named knots, where the intervals adjoin.

In this thesis we make use of the *mgcv* package in R introduced by Wood (2019). For the univariate cases, we work with cubic smoothing splines. While, for modeling multi-dimensional relations, such as the combined effects of continuous variables (i.e., latitude, longitude) on the response, we utilize thin-plate regression splines based on thin-plate smoothing splines.

We can write the cubic smoothing spline as a linear combination of the basis functions:

$$f(x) = \sum_{j=0}^3 \beta_j x^j + \sum_{k=1}^K \theta_k (x - \xi_k)_+^3 = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K \theta_k (x - \xi_k)_+^3. \quad (17)$$

³Note that the polynomial regression, as mentioned earlier, is a form of linear regression that can be used to model non-linear relationships.

ξ_1, \dots, ξ_K represent the K knots. $1, x, x^2, x^3, (x - \xi_1)_+^3, \dots, (x - \xi_K)_+^3$ stand for the truncated cubic basis functions. The coefficients are represented by $\beta_0, \beta_1, \beta_2, \beta_3, \theta_1, \dots, \theta_K$. Furthermore, the following conditions for $f(x)$ hold:

- a cubic polynomial exists between every pair of knots $(\xi_i, \xi_i + 1)$.
- the polynomial pieces fit together at the knots s.t. f itself, its first and second derivatives are continuous at each knot.

The spline coefficients are estimated such that the cubic smoothing spline function minimizes the penalized sum of squares over the class of all twice differentiable functions.

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x_i)^2 dx \quad (18)$$

where λ is the fixed smoothing parameter. The first and second term stand for the deviance and regularization, respectively. The deviance measures the goodness of fit, while the regularization penalizes curvature in the function (Hastie et al., 2009). This penalty will be large if the function $f(x)$ has a large variation. Conversely, in cases of limited flexibility, this value will be little. λ controls the smoothness of the splines, as it establishes a trade-off between both terms.

Note that cubic smoothing splines can be seen as a special case (i.e., one-dimensional case) of the thin-plate splines, which were introduced by Duchon (1977). As mentioned earlier, this generalization allows us to model interactions between continuous variables. In a similar manner to the cubic smoothing spline, the coefficients are estimated such that the unknown function minimizes the bivariate version of the penalized least square criterion (for a detailed analysis on thin plate splines we refer to Duchon (1977) and Wood (2003)). Wood (2003) discusses that thin-plate smoothing splines suffer from a high rank – “its computation requires estimation of as many parameters as there are data”, leading to high computational cost. Therefore, Wood (2003) introduced the thin-plate regression splines in the *mgcv* package, by providing low rank approximations to thin-plate smoothing splines; the low rank smoother is obtained through performing an eigendecomposition (for an elaborate analysis, see Wood (2003)).

4.3.2 Smoothing parameter

A wrong choice for the value of λ could result in under- or over-fitting of the data; therefore it is crucial to estimate the ‘optimal’ smoothing parameter. Different techniques exist to estimate λ such as generalized cross validation (GCV), generalized approximate cross validation (GACV), unbiased risk estimator (UBRE), akaike information criterion (AIC), and restricted maximum likelihood (REML). This papers focuses on GCV and UBRE used in

the *mgcv* package. The generalized cross validation (GCV) criterion introduced by Craven and Wahba (1978) can be used when the scale parameter is unknown. This method selects the smoothing parameter such that (19) is minimized:

$$nD/(n - DoF)^2 \tag{19}$$

When the scale parameter is known, another possibility for estimating the smoothing parameter is the unbiased risk estimator (UBRE) criterion which is effectively the AIC rescaled:

$$D/n + 2sDoF/n - s \tag{20}$$

where D is the deviance, n is the number of observations, DoF is the effective degrees of freedom of the model and s is the scale parameter (Wood, 2017).

4.4 Trade-offs: dealing with non-linear effects of continuous and spatial variables

The group of linear models explored throughout this thesis can be used by researchers and actuaries to model the severity and frequency of policyholders' claims. Nevertheless, the problem emerges when we would like to include continuous and spatial variables such as age and postal codes in the model. In more detail, the ordinary LM and GLM assume that there is a linear or parametric relationship between the predictors and response variables. Yet, within an insurance setting, actuaries anticipate and require models that respond to non-linear effects.

For example, the effect of age on claim count and severity is expected to fluctuate, as different age groups are more likely to end up in a hospital than others. Hence, entering continuous variables into a (generalized) LM may lead to erroneous conclusions about the effects of the continuous variables with broader-scale implications. This could result in a situation in which the insurer fails to offer fair tariffs based on the policyholder's risk factors.

As previously noted, polynomial regressions are a form of linear regression in which the relationship between the response and predictor variable is modeled as an n th degree polynomial (x, x^2, \dots, x^n). Accordingly, polynomial regressions can be helpful to measure the possibly non-linear effects. That being said, it can be challenging to choose the degree of the polynomial. Whereas a low-degree polynomial may result in a poor fit, a high degree often leads to unstable estimates (Denuit et al., 2019). Furthermore, the presence of extreme values among the predictor variables may seriously skew the overall results, as this method is highly sensitive to outliers. Given these considerations, while polynomial regressions may suffice at times, the issue is that non-linear relationships are typically not easily specified.

Instead of treating age as a continuous predictor in the model, continuous variables can be seen as categorical, with a high number of levels (defined as "multi-level factors" by Ohlsson and Johansson (2010)). Including these categorical variables in ordinary LMs or GLMs

would lead to many risk classes wherein some groups are made up of relatively few policyholders. Ultimately, this could result in statistically insignificant results and the possibility of overfitting. Hence, a solution is to group continuous and spatial risk factors in a limited set of categorical variables.

An obvious disadvantage of grouping continuous risk factors is that the premium for two policies with different (but close) risk factor values may have substantially different premiums, if the values happen to belong to different intervals (Ohlsson and Johansson, 2010). Hence, this approach leads to a loss of information owing to the fact that the effects of a continuous variable are captured by a piecewise constant function (Denuit et al., 2019).

It is worth pointing out that insurers often prefer a model that contains categorical variables, which are easier to understand (and work with) (Henckaerts et al., 2017) (Klein et al., 2014). Appreciating the contextual constraints, when modeling continuous risk factors by categorical variables, it is important to have clusters that reflect the effect of the continuous variables as closely as possible. In the chapter that follows, we will discuss how we can transform these continuous variables into categorical variables, with a limited number of levels while keeping the loss of information to a minimum.

As shown, with the use of smooth functions, GAMs capture the possible non-linear effects of continuous and spatial variables. In contrast to a parametric approach (such as the polynomial regression) there is no need to specify a certain function in advance, or as stated by Hastie and Tibshirani (1986), “the GAM has the advantage of being completely automatic, i.e., no detective work is needed on the part of the statistician.” GAMs are much more flexible when using continuous variables than linear models (Denuit et al., 2019). From a statistical modeling point of view, this is highly appealing. However, as the GAM focuses on data in a semi- or non-parametrical fashion, its estimated functions are more difficult to interpret than the estimates of a GLM.

5 Approach

In the previous section, the theoretical backgrounds of several statistical models for predicting pure premiums using a frequency-severity approach were discussed. Furthermore, we identified the potential issues with the use of continuous and spatial risk factors and looked into the viable means of implementing these variables in the models; two main approaches were identified: (1) a GAM that offers a flexible way to model the non-linear effects, and (2) a generalized or ordinary LM, depending on the specification of the probability distribution, in which the continuous and spatial variables are transformed into categorical variables. This latter approach is often seen as the preferred method by insurers (see previous section); however, it results in a loss of information. Consequently it is crucial to find a clustering technique to group the continuous risk factors in a manner that keeps the clusters sufficiently small to avoid too much variable fluctuation within a group, and yet sufficiently large

to accomplish good estimate accuracy.

Given the trade-offs of both approaches, it remains an open question whether the differences between these approaches are significant when pricing a hospitalization product using a real-life data set. This chapter aims to explain the methodology needed to answer our research question; we illustrate the process of specifying the models and how we attempt to compare them. First, we discuss the criteria used to select the appropriate probability distribution (resulting in a LM or GLM) and explanatory variables. Second, we investigate how we can model a LM/GLM where the continuous and spatial variables are transformed into categorical variables with limited levels; for this, we follow the methodology of Henckaerts et al. (2017). Lastly, several possible ways to measure the differences between the resulting models are discussed with the purpose of answering the research questions laid out in Chapter 1.

5.1 Model specification

5.1.1 Probability distribution

The first objective when setting up statistical models is to select the appropriate probability distribution. As mentioned in Chapter 3, several probability distributions (or transformations of the response to normalize the data) exist that often form a reasonable fit for modeling claim severity and frequency data.

Before selecting the final probability distribution to work with, it is recommended to compare the possibilities with each other. An evaluation concerning the goodness of fit of the models should be performed to select the distribution that best fits the data.

When examining continuous distributions, one possibility is to analyze the deviance residuals of the models. Intuitively, the deviance residuals can be seen as the residuals “adapted for the shape of the assumed distribution”, such that the specified distribution will be appropriate if the deviance residuals are approximately normally distributed (Goldburd et al., 2019). To determine in which model the deviance residuals follow a normal distribution (and are random, such that they do not show a trend), histograms of the deviance residuals and normal quantile-quantile (Q-Q) plots can be analyzed. The histograms and normal Q-Q plots show us whether the deviance residuals are normal, and hence, which models have a good fit.

When analyzing claim frequency distributions, which are discrete, the deviance residuals are not expected to follow a normal distribution. According to Goldburd et al. (2019), randomized quantile residuals can be used instead as they have properties similar to those of deviance residuals, “but add random jitter to the discrete points so that they wind up more smoothly spread over the distribution.” Furthermore, the dispersion statistic can show if a Poisson distribution is suitable, or if an alternative such as the negative binomial distribution is preferred when overdispersion is present.

As a final remark, besides the graphical methods introduced in this section, the performance and fitting criteria discussed in Chapter 5.3 can be used for the purpose of choosing an adequate probability distribution.

5.1.2 Variable selection

The following step aims to refine the model by selecting the optimal number of explanatory variables. Including risk factors that do not have a genuine relation with the response variable, may only add noise to the model. This could result in overfitting as some variables might be falsely found to be statistically significant. The “optimal” number of predictors can be selected by referring to the information criteria and comparing their values for different numbers of combinations of explanatory variables. This allows us to select the number of risk factors with the lowest information criteria (IC) of choice. We decide to focus on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The AIC (21) and BIC (22) equations are given by:

$$AIC(k) = 2 \cdot k - 2 \cdot \text{loglikelihood} \quad (21)$$

$$BIC(k) = \log(n) \cdot k - 2 \cdot \text{loglikelihood} \quad (22)$$

where k , which stands for the degrees of freedom, is the number of parameters in the model and n is the total number of data points. The goodness of fit of both IC are measured by $-2 \cdot \text{loglikelihood}$; this term declines as the model fit to the data improves (and the *loglikelihood* increases). For AIC, the penalty term is represented by $2 \cdot k$; this term increases the AIC as a “penalty” for each added parameter (Akaike, 1974). The BIC favors less complex models as its penalty term $\log(n) \cdot k$ is more severe than that of the AIC (Schwarz, 1978).

5.2 Clustering continuous and spatial variables

This chapter discusses the methodology we use to transform continuous and spatial variables into categorical variables, which subsequently can be implemented in a parametric approach. When clustering continuous variables into categorical variables, it is crucial that the categorical variables present the effects of the spatial and continuous variables as accurately as possible. Clijsters (2015) and Henckaerts et al. (2017) introduced a strategy to cluster continuous and spatial variables. Their research explains that one method is to start with a GAM where the continuous and spatial variables are modeled non-parametrically. In this way, we are able to identify possible non-linear effects. Thereafter, the smooth effects of the variables, with comparable riskiness, can be clustered. This results in categorical variables that can be represented by a set of separate binary variables and can, in this way, be efficiently included in a parametric model. This approach of Clijsters (2015) and Henckaerts

et al. (2017) tries to best represent the nonlinear effects in the resulting categorical variables with a limited number of levels and is therefore applied in this thesis.

Spatial variables, including postal codes and their corresponding coordinates, and continuous variables, such as policyholders' ages, require different clustering approaches for the purpose of tariff setting. The goal of clustering for both types of variables is to create groups with similar risk levels; however, for clustering ages, an additional condition is added requiring that only consecutive values are classified together. This limitation is put in place for practical purposes, which require a pricing scheme that is easy to understand.

The objective of clustering spatial riskiness is to group municipalities with comparable geographical riskiness together. This 'riskiness' can be presented by the smooth bivariate effects of the longitudes and latitudes, which represent the effects of postal codes and are obtained from a modeled semi-parametric GAM that includes all the other chosen categorical variables. The *ClassInt* package introduced by Bivand et al. (2015) is comprised of commonly used methods for grouping different municipalities. Contingent upon the data and the motivation behind the classification technique, some of these methods are more fitting than others.

For this thesis, we focus on the Fisher-Jenk's natural breaks algorithm introduced by Fisher (1958). This algorithm attempts to classify close values into groups with maximal distance to other groups according to breaks that naturally exist between the groups of close values. To obtain these groups, the algorithm uses an iterative way to find the best possible classes of close values by minimizing the variation from the groups' means, while likewise maximizing the variance between the groups (Slocum et al., 2008).

Choosing the optimal clustering algorithm is somewhat subjective; however the results of alternate methods will be examined to verify the "correctness" of the chosen technique. We refer you to Bivand et al. (2013) for more information concerning widely used techniques for creating univariate class intervals for spatial data, and to Henckaerts et al. (2017) for a comprehensive analysis of several of these classification methods applied in insurance pricing of an MTPL portfolio.

To select the optimal number of levels, the AIC and BIC of the parametric models should be evaluated using different numbers of levels within the constructed categorical variables. As one requires consecutive intervals for clustering continuous variables (e.g., age), Henckaerts et al. (2017) propose the use of regression trees which produce intuitive splits, and evolutionary trees, which combine the framework of regression trees with genetic algorithms. A complete survey of the theory behind regression and evolutionary trees is beyond the scope of this paper (see Henckaerts et al. (2017) for an overview).

5.3 Model comparison

Once the probability distributions are specified and the GAM and (generalized or ordinary) LMs with and without grouped risk factors are estimated, the main questions of this thesis arises: Is the larger model with added spatial categorical variables better than the smaller one where we do not take spatial risk into account? How can we compare the statistical models, and their resulting pure premiums? To answer these questions, statistical and insurance-related approaches are presented to assess the models. The different approaches for comparing the frequency, severity and corresponding pricing models can be divided into three subcategories that are discussed in this section. This chapter introduces methods that measure the fit, the predictive accuracy, and the lift of the models.

5.3.1 Measures of fit

When comparing models with each other, from a statistical point of view, a major objective is to select the model that achieves the right balance between goodness of fit and parsimony. To review, the goodness of fit expresses how well the model fits the data, while the concept of parsimony prefers simplicity. When working with a parsimonious model, there is the risk that it fails to capture the true signal, resulting in an overly simplistic model and underfitting of the data. On the other hand, using complex models, which generally have a good fit, may lead to overfitting. Several statistical instruments are available that can help us examine the fit and complexity of models. In this way, we can find the ‘preferred’ model that adequately fits the data, while being able to discriminate the signal from the noise.

This thesis uses information criteria (AIC and BIC), which are introduced in Chapter 5.1, to compare the performance of the models. Furthermore, when comparing the nested models (i.e., when comparing the larger model with added categorical variables representing the spatial risk with the model without spatial risk), we may also perform the F-test. As deviance always reduces when predictor variables are added to a model, the F-test tries to explain whether the added variables indeed lead to an improvement of the model. The F-test is represented by:

$$F = \frac{UnscaledDeviance_{small} - UnscaledDeviance_{big}}{NewParameters * Dispersion_{Big}} \quad (23)$$

As the F-statistic follows an F-distribution, one may compare the obtained F-statistic to the appropriate F-distribution value; in this way one can conclude if the parameters are significant, and hence conclude whether adding a specific variable improves the model sufficiently to outweigh “the cost” of adding additional parameters.

5.3.2 Prediction accuracy

The measurements discussed in the previous section are helpful to understand the model's ability to properly balance between the goodness of fit and complexity on the in-sample data. However, to set pure premiums, it is crucial to examine the predictive accuracy of the models on new data. Prediction errors are used to measure the predictive performance of models, and help us to compare distinctively different models based on their predictive accuracy; a model that performs excellently on the in-sample data, but performs poorly on new data is useless and should not be preferred.

A prediction error for a variable y_i is simply the difference between its actual value and predicted value ($\hat{e}_i = y_i - \hat{y}_i$). Many well known metrics exist that incorporate prediction errors such as mean absolute error (MAE), average error, mean absolute percentage error (MAPE), mean squared error (MSE), and root mean squared error (RMSE) (see for example Shmueli et al. (2017) for an overview). In this thesis, we use the MSE as measurements to assess prediction performance. The MSE is defined by:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (24)$$

By calculating the MSE on an out-of-sample data set for each model, we are able to compare their degree of performance accuracy.

The fact that these metrics can be calculated not only for the pure premiums that are obtained, but also for the underlying severity and frequency models allows for a more intuitive understanding of the individual models. It is important to note that since the MSE depends on the unit of the examined response variable, it will result in different scales based on the dependent variables of interest (e.g., pure premium, claim severity, or claim frequency).

Furthermore, it should be noticed that MSE values are sensitive towards outliers (i.e., very large claims), which are difficult for the models to reliably predict (Dugas et al., 2003). To visualize the outlier influence, Shmueli et al. (2017) suggests that simple histograms of the distribution of the prediction errors are in fact quite valuable and can contribute to the interpretation of the metrics.

Next to comparing the predictive performance of the different models on a holdout data set, it is important to pay attention to the stability of the models itself. It bears asking: What if the models have approximately the same MSE values? Is the model with the lowest MSE the most optimal in terms of predictive performance? We involve model stability in this context by studying the differences between the values of the error metrics in the in-sample and out-of-sample data.

In this way we are able to spot complex models that have a great fit on the in-sample data, but perform poorly on new data. Generally speaking, the greater the difference, the worse the model is in terms of its predictive accuracy (Aleandri, 2019).

Lastly, in terms of process, we want to emphasize that we perform a simple data splitting resulting in an in-sample (80%) and out-of-sample (20%) data set. After fitting the models on the in-sample data set, we work with the out-of-sample data set to compare the predictive performance accuracy. To this end, Hastie et al. (2009) accentuate that, in a “data-rich” situation, analyzing the performance of different models in order to choose the best one (model selection) should be performed on a different data set than the data, which is new and not used for model selection but rather to assess the prediction error of the final chosen model. Hence, to test the chosen model on new data, ideally, it is most optimal to work with three randomly divided data sets (i.e., a training, validation, and test set).

5.3.3 Measures of lift: the ordered Lorenz curve and Gini index

Along with the statistical criteria that assess the fit and predictive performance of the models, actuarial tools are needed to gain insight into the ‘economic value’ of the pure premiums estimated by the different models. Examining lift metrics helps researchers acquire such insight.

Lift plots endeavor to visualize and evaluate the economic value of a model, which in the insurance setting can be seen as the ability of the model to prevent adverse selection (i.e., the ability that the model leads to fair actuarial prices based on the risks of the policyholders) (Goldburd et al., 2019). Therefore, the use of lift plots makes it possible to measure the comparative advantages of the models, with regard to avoidance of adverse selection.⁴ Simple quantile plots, double lift charts and the Lorenz curve are several options the researcher has to measure lift (see Goldburd et al. (2019)).

In this paper, we opt to measure the lift of the models with the help of an extension of the Lorenz curve, a well-known concept in economics initially developed by Lorenz (1905) for representing inequality in the distribution of wealth. Frees et al. (2011) introduced the ordered Lorenz curve and the corresponding Gini index, developed by Gini (1912), to insurance modeling. Furthermore, Lorenz curves have been examined and applied in several studies for model validation by, for example, Frees et al. (2013), Shi et al. (2015) and Frees et al. (2016).

The aim of developing these curves is to compare the distribution of losses to the distribution of predicted premiums for a portfolio of policies. This can be done by creating an ordered form of the Lorenz curve and utilizing the Gini index as a statistical measure through which the different statistical models can be compared. Finally, an analysis of the Gini indices will indicate which model is best able to avoid adverse selection.

Following Frees et al. (2013), we define for a policy i , L_i as its loss and $\pi_i(x)$ as the corresponding premium based on the policyholder’s risk factors x . Furthermore, the losses and

⁴When measuring a model’s lift, it is necessary to test the model on a holdout sample of the data to prevent overfitting.

premiums got ordered based on a random variable $R = \frac{\pi(x)^{new}}{\pi(x)^{base}}$, called the relativity, which is defined as the predicted premiums over the base premiums used as a benchmark. The premium ($\hat{F}_\pi(s)$) and loss ($\hat{F}_L(s)$) distributions can then be given as:

$$\left(\hat{F}_\pi(s) = \frac{\sum_{i=1}^n \pi_i \mathbf{I}(R_i \leq s)}{\sum_{i=1}^n \pi_i}, \hat{F}_L(s) = \frac{\sum_{i=1}^n L_i \mathbf{I}(R_i \leq s)}{\sum_{i=1}^n L_i} \right) \quad (25)$$

where $\hat{F}_\pi(s)$ ($\hat{F}_L(s)$) stands for the proportion of premiums (losses) for the policyholders with relativities (R_i) less than or equal to s .

The Gini index can be used to summarize the ordered Lorenz curve; which represents double the area between the line of equality and the ordered Lorenz curve. This index is used as a direct economic interpretation and as a measure to represent the ability of a model to avoid adverse selection.

Frees et al. (2013) showed that the average profit of an insurer is approximately equal to the Gini index divided by two. Furthermore, for comparing the model's ability leading to actuarial fair tariffs, Frees et al. (2013) introduced the mini-max approach, which looks for the base model with the minimal maximum Gini index; put another way, for each model, the predicted premiums are specified as the baseline premiums, and the other model(s) are used as the scores. Afterwards, from the results presented in a matrix, we select the model with the smallest of the maximum Gini indices, indicating that the selected model gives rise to a tariff structure with a low probability of suffering from the phenomenon of adverse selection.

6 Database & Descriptive statistics

To enable a comparative analysis, we consider a data set concerning a Belgian hospitalization insurance. The data, provided by KPMG Belgium, contains information about the claim losses and policyholders' between 2012 and 2019 (1,864,151 observations). The paper focuses on a single policy year, 2019, wherein 449,073 contracts were established. This comprehensive data set comprises out of all contracts with and without claims, which is used to build a predictive model for claim frequency. From this data set, we construct a subsequent one that consists of all contracts for which at least one claim is recorded (29,073 observations). Note, the latter is used to develop a predictive model for claim severity.

A considerable amount of time is spent on exploring and preparing the data, namely setting the exposures and ages of policyholders (through the availability of contract and birth dates), correcting erroneous cells, analyzing the quality of the data, and ensuring the databases can be merged. In the original data, a small number of claim costs with negative values were identified – representing reimbursements to the insurance company. Our simplified model does not analyze these numbers, as the severity distributions that were introduced require non-negative values.

Following these adjustments, we obtained a data set with 13 variables for each observation. Each row in the data file represents an insurance contract of an individual during the policy period of 2019. Table 1 lists the variables used in this project. Recall, in Chapter 3 there were several useful risk factors given for pricing health insurance – including information about BMI, tobacco use, marital status, and pre-existing illness. However, since the insurer only kept track of the standard information for the current case study, the variables mentioned in Table 1 will be our focus.

Table 1: Variables used in the analysis.

Variable	Explanation
NSI	Unique identification number of the insured.
Year Hospi	The policy period.
Exposure	Fraction of the policy period for which the policyholder is insured.
Age	Age of the policyholder at the beginning of the policy period.
Age category	The age category, set by the insurance company, in which the insured is.
Postal code	Postal code of the policy holder’s place of residence.
Lat Long	Latitude and longitude of the municipality in which the insured is established.
Sex	Gender of the policyholder: 1 = male, 2 = female.
Type of product	Type of coverage provided by the insurance policy: Hospimut: offers a financial intervention in the invoices of hospitalization in double rooms, as well as a fixed price intended to cover the costs pre and post hospitalization. Optio: offers financial assistance in the invoices for hospitalization in private rooms, as well as a package intended to cover pre and post hospitalization costs. Within this class there are three subclasses with different conditions.
Type of recipient	Code designating the type of standard beneficiary in the insurance policy.
Total amount	The total cost for the policyholder in Euros.
Average claim amount	The total amount for the policyholder divided by its number of claims.
Number of claims	The number of claims recorded for the policyholder. In case the policyholder did not file a claim, this number is 0.

By way of a preview, we will outline a few univariate statistics below. Table 8 and 9 (contained in the Appendix) represent some of the data set’s descriptive statistics.

To offer a clearer presentation of the data in context, we start the analysis by plotting histograms of the observed claim severities and claim counts. The histograms, presented in Figure 2, show us that the data for both claim severity and claim count is skewed to the right and is non-negative. Additionally, the skewness coefficient of 20.75 and 6.99, respectively for claim severity and frequency, confirms this visualization. During the examined period,

6.47% of the policyholders filed at least one claim, resulting in a total loss of EUR 4.33M. The average claim severity is EUR 115.26. The smallest amount observed was EUR 1.01, while the largest was EUR 12,500, which is the maximum yearly reimbursement per contract, set by the insurance company to limit losses. In total, 99.12% of the claims amounted to less than EUR 1000. In addition, only 10 claims were observed that exceeded EUR 10,000.

The left panel of Figure 3 shows the average claim amounts for men and women in similar age groups. On average, there is a small difference between the average claim amount for men (EUR 109) and women (EUR 122). In younger age groups (< 19), the average hospitalization costs for males is higher. This outcome was not unexpected, since younger males are more reckless and therefore more likely to be involved in accidents that require more expensive hospitalization. With respect to older age classes, the average costs for women are 11% higher. For men and women aged 20-29, we observe a significant difference (65%) in the average claim amounts. This is consistent with the findings of a study issued by the Belgian Federal Public Service of Health (2019), which reported a higher rate of hospital care among women aged 20-40 that is related to childbirth. On average, people aged 25-29 had the highest claim costs – almost 3 times higher than the lowest average claim severity for the age group 15-17.

The right panel of Figure 3 demonstrates significant differences between the average claim frequencies, with respect to the age of the policyholders. The overall claim frequency was the highest for those aged 90+ (0.22) and the lowest for the age group 10-14 (0.02).

Moreover, Table 9 shows that the hospital costs for those who own the sub-product Optio 200 are typically 1.46 times higher than for those who are Hospimut holders. This makes sense considering Optio 200 is the most costly option given its coverage of hospital bills for private rooms, while Hospimut only covers double rooms.

In order to gain some insight about the average claim amounts between the various cities, we plot the mean claim severity by postal code. Figure 1 shows that the average claim amounts in several urban areas are significantly higher than other municipalities. In other words, one's place of residence might influence the hospitalization costs. A recent investigation of the Christelijke Mutualiteit (2019) indicated that some medical interventions are up to 27 times more expensive between different Belgian hospitals. The ethics of this pricing scheme is beyond the scope of this paper, but could be a potential indication that hospitalization costs are influenced by the place of residence of the insured. What's more, postal codes serve as a proxy for socioeconomic characteristics that represent the neighborhood where a policyholder resides (Henckaerts et al., 2017).

Moreover, by seeing this chart one can obviously observe that the explored Insurance product is mainly active in Wallonia and Brussels (i.e., the insurance is active in 664 of the 1152 Belgian municipalities).

It is worth emphasizing that the descriptive analysis may not establish the true dependency of the response on the discussed variables, but rather indicates that the risk factors are likely

to play an important role in the statistical analyses of the expected severity and frequency of claims.

Figure 1: Map of Belgium representing the average claim severity by municipality obtained from the Belgian hospitalization data set (2019).

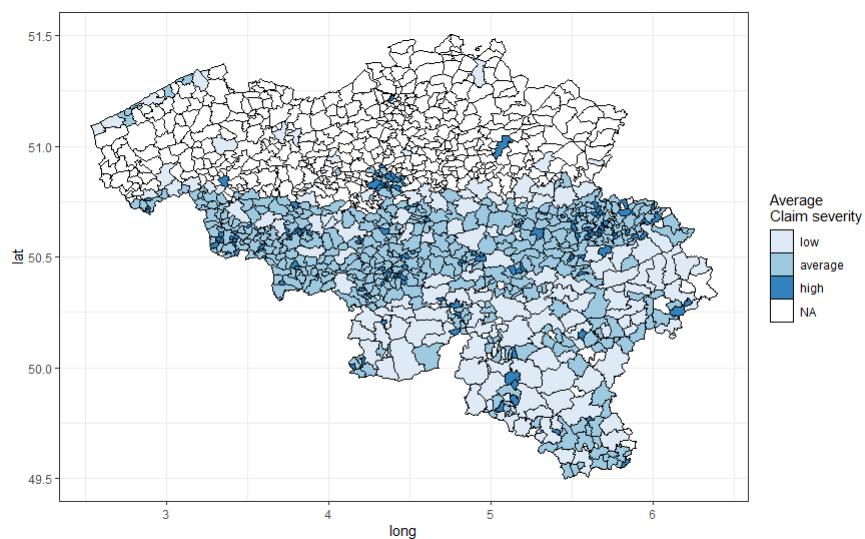


Figure 2: Histograms of the observed claim severities and claim counts obtained from the Belgian hospitalization data set (2019).

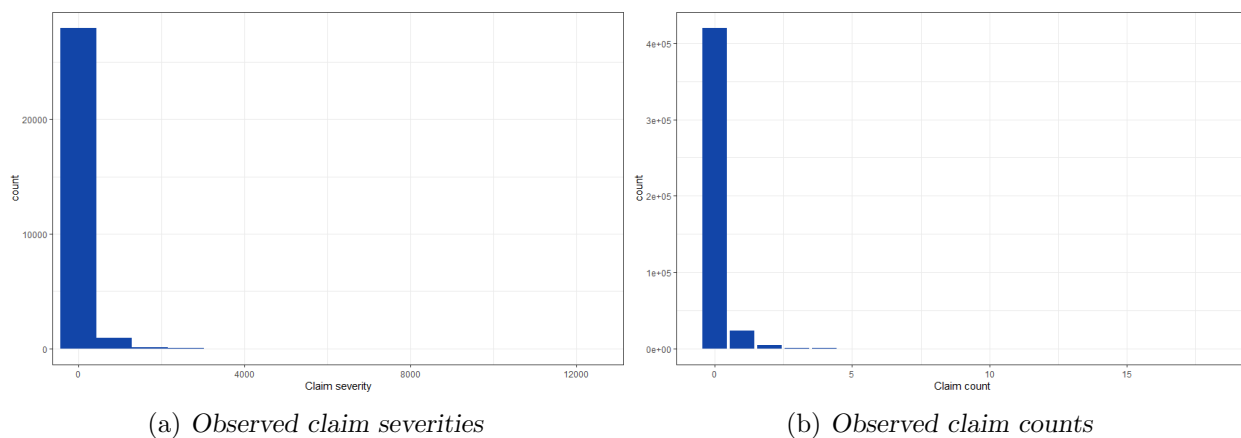
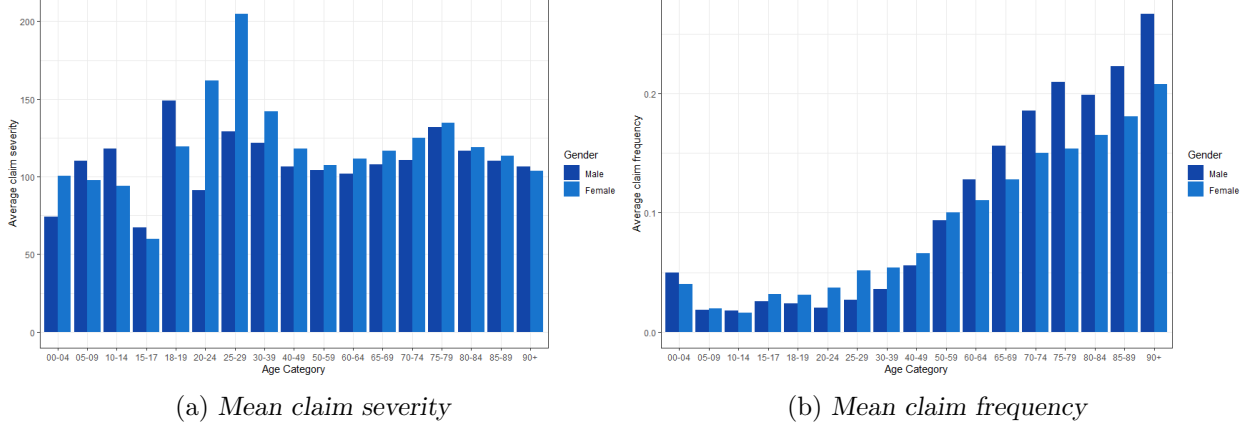


Figure 3: Histograms showing the average claims severity and frequency by gender and age groups obtained from the Belgian hospitalization data set (2019).



7 Claim Severity analysis

This chapter focuses on modeling claim severity, while the next chapter presents the specified models for claim frequency. For $E(S_i)$, we work with the data set that exclusively contains policyholders who registered at least one hospitalization during the examined period ($NumClaims > 0$). The response variable, claim severity, is defined as $AverageClaimAmount = TotalPayments/NumClaims$ where $NumClaims$ is taken into account as a weight.

7.1 Base model

Our first objective when creating statistical models for claim severity is to specify the appropriate probability distribution. In Chapter 3, we introduced three probability distributions that often form a reasonable fit for continuous, non-negative and right-skewed data (i.e., a simple normal distribution with log transformation, gamma distribution, and inverse-Gaussian distribution). As our data exhibit these properties, we model the distributions in three different models to identify which probability distribution fits our data most closely. When we specify a gamma or inverse-Gaussian distribution, the GLM for claim severity (S_i) is presented as follows:

$$\log(E(S_i)) = \beta_0 + \beta_1 Sex_i + \sum_{j=1}^{J-1} \beta_{2j} TypeProd_{ij} + \sum_{k=1}^{K-1} \beta_{3k} AgeCat_{ik} + \sum_{l=1}^{L-1} \beta_{4l} TypeClient_{il} \quad (26)$$

where Sex_i , $TypeProd_{i1}, \dots, TypeProd_{iJ-1}$, $AgeCat_{i1}, \dots, AgeCat_{iK-1}$ and $TypeClient_{i1}, \dots, TypeClient_{iL-1}$ represent the explanatory variables. $\beta_1, \dots, \beta_{4L-1}$ stand for their correspond-

ing coefficients, and the intercept is denoted by β_0 . As the categorical variables, represented by dummy variables, can take z (J , K and L respectively) different possibilities, we define only $z - 1$ binary variables in the model.⁵ Furthermore, a natural log link function ($g(x) = \ln(x)$) is specified.

The third possible way to model claim severity is by assuming a normal distribution with a log transformation of the response variable (i.e., in this case we log transform the observed values and do not take the log of the expectation):

$$E(\log(S_i)) = \beta_0 + \beta_1 Sex_i + \sum_{j=1}^{J-1} \beta_{2j} TypeProd_{ij} + \sum_{k=1}^{K-1} \beta_{3k} AgeCat_{ik} + \sum_{l=1}^{L-1} \beta_{4l} TypeClient_{il} \quad (27)$$

Unlike (26), in this GLM, an identity link (i.e., no link), which is completely equivalent to a normal LM with log transformation of the response, is assumed.

To select the distribution that best fits the data, an evaluation concerning the goodness of fit of the three models is performed. We do this by examining whether the deviance residuals of each model (approximately) follow a normal distribution. Normal Q-Q plots (Figure 7) and histograms (Figure 8) of the deviance residuals are presented in the Appendix. The Normal Q-Q plots show us that none of the examined distributions fit the data perfectly, which was expected as it is very unlikely to be the case for real-life insurance data. However, at first glance, the distributions approximately fit the data, but a poor fit is visible in the lower and upper quantiles, (corresponding to small and large claim amounts, respectively). At these edges, the quantiles lie below and above the reference line indicating that the observations are more skewed than the chosen distribution or transformation. Furthermore, when examining the histograms, it becomes clear that the deviance residuals of the log normal distribution most adequately approximate a normal distribution. Furthermore, the information criteria (AIC & BIC), shown in Table 2, substantiate that the Gaussian distribution with a log-transformed response may be more appropriate in the current situation. In the following sections, we continue to build on this model.

Lastly, based on both IC, the risk factor presenting the type of client was dropped from the base model. Most of the variables are statistically significant (p-value < 0.001), although the coefficients of the variables representing gender and the four youngest age groups are statistically insignificant (see Table 11 in the Appendix).

The chosen model, a normal LM with only categorical variables, can be seen as the starting point of this severity analysis. In the next two sections, as the probability distribution and categorical explanatory variables have been specified, the paper analyzes how to include the continuous variables.

⁵A z th dummy variable is unnecessary; it reveals no new information and leads to multicollinearity, also known as the dummy trap. In the current situation, reference levels for each group of dummy variables are represented by: $Sex(men)$, $TypeProd(Hospimut)$, $AgeCat(00 - 04)$ and $TypeClient(1)$.

Table 2: Information criteria: claim severity distributions. The value between brackets stands for the corrected version of the log-normal IC such that comparison is possible.

	gamma distribution log link	inverse-Gaussian log link	Gaussian with log-transformed response identity link
AIC	345 419	341 673	77 702 (261 790)
BIC	345 596	341 850	77 878 (261 967)

7.2 GAM

When examining Figure 1, it appears that spatial information may be a significant factor for tariff-setting as a relationship among policyholders' residences and the claim costs, costs of hospitalizations, is expected. Furthermore, rather than utilizing the predetermined age classifications, the non-linear effects of age can be modeled as a continuous variable in the GAM.

The GAM, which allows for spatial and continuous variables, can be formulated as follows:

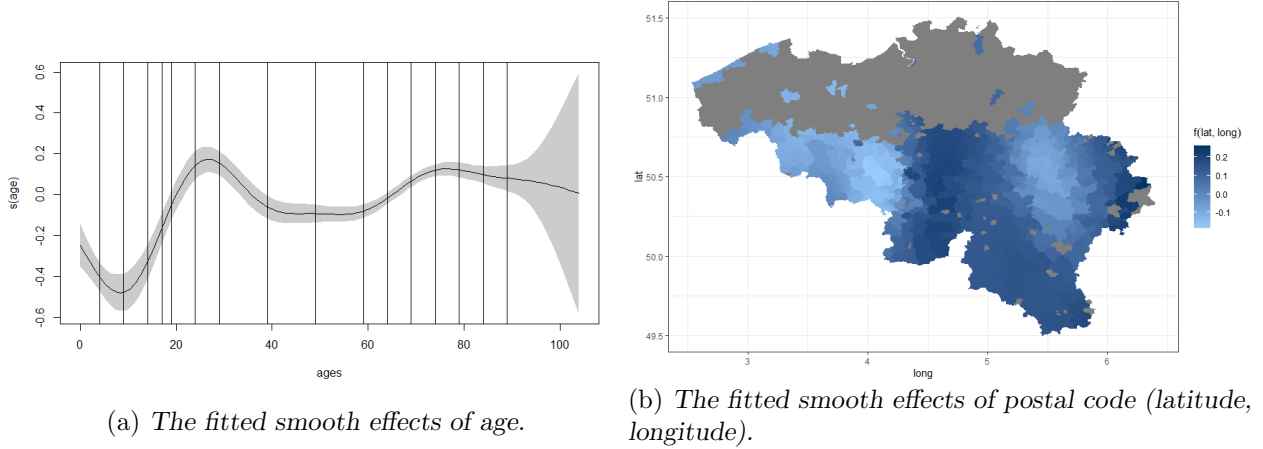
$$E(\log(S_i)) = \beta_0 + \beta_1 Sex_i + \sum_{j=1}^{J-1} \beta_{2j} TypeProd_{ij} + f_1(age_i) + f_2(latitude_i, longitude_i) \quad (28)$$

where the categorical risk factors and their coefficients are presented parametrically in a manner similar to those of the log-normal LM (27). This model extends the parametric model by including the effects of the continuous risk factors captured by the univariate f_1 and bivariate f_2 smooth effects for age and postal codes, respectively.

The left-hand panel of Figure 4 visualizes the fitted smooth effects of age. The 95% confidence interval is shown by the dashed lines. These intervals are larger for policyholders over 90 as limited data is available for this age group. Examining this plot clearly confirms the nonlinear effects of a policyholder's age on claim severity. The graph shows that on average the claim costs are the highest for persons in their late twenties and elderly people. Furthermore, the age groups set by the insurer are represented by the vertical lines that, at first glance, appear to show the effects fairly accurately.

The smooth bivariate effects of latitude and longitude, representing the geographical effect of postal codes, are visualized in the right-hand panel of Figure 4. This map provides evidence that there exists a significant spatial effect on the average claim cost. Furthermore, Table 12, provided in the Appendix, shows that both the smooth terms and the parametric coefficients are statistically significant (p-value < 0.001 and p-value < 0.10 for the coefficient of gender).

Figure 4: Continuous and spatial effects on claim severity.



7.3 LM with clustered spatial effects

In this section, we aim to represent the non-linear effects visualized in Figure 4 using categorical variables, which can be implemented in a parametric model. Following from Figure 4, we already recognize that the original age classes represent the age effects quite well. This comes as no surprise as the classes were specified based on the judgment of actuarial experts. We mentioned in Chapter 5.2 that regression trees can be used to create consecutive age intervals. For the current case, we opt not to apply this method because we choose to limit ourselves to the insurer’s clusters, as we assume these were made with good reason. Furthermore, we execute Fisher’s natural breaks algorithm with the objective of grouping the municipalities with comparable geographical riskiness together. To avoid over-fitting, we follow the approach of Henckaerts et al. (2017), discussed in Chapter 5, by clustering the fitted smooth effects obtained from the GAM (28). Hence, our last model, using a log-normal distribution, can be presented as:

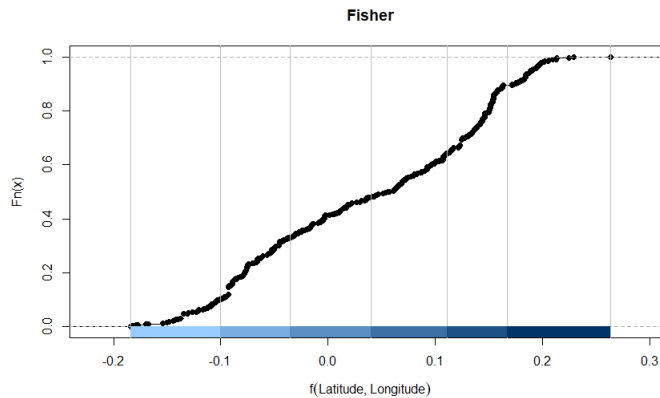
$$E(\log(S_i)) = \beta_0 + \beta_1 Sex_i + \sum_{j=1}^{J-1} \beta_{2j} TypeProd_{ij} + \sum_{k=1}^{K-1} \beta_{3k} AgeCat_{ik} + \sum_{m=1}^3 \beta_{5l} Location_{im} \quad (29)$$

where the first categorical variables are the same as in our base model (27). New in this model are the binary variables representing the clustered postal codes. To select the optimal number of clusters, we examine the AIC and BIC of a set of GAMs with different numbers of clusters representing the spatial effect. Based on Table 10 in the Appendix, we opt for 6 classes; this results in the lowest AIC and 3rd lowest BIC. Figure 4 shows the clusters obtained through the Fisher-Jenks’ natural breaks algorithm; the vertical lines represent the

breaks between the classes and the empirical cumulative distribution function of the fitted smooth spatial effects is represented by the curve (Bivand et al., 2013).⁶

As shown by Table 13 in the Appendix, this model results in statistically significant coefficients for most of the risk factors (p-value < 0.001), and for few coefficients of younger age classes (p-value < 0.1), except for the age group 15-17 for which the resulting coefficient are statistically insignificant.

Figure 5: Spatial clusters specified by the Fisher-Jenk’s natural breaks algorithm.



8 Claim Frequency analysis

For modeling the expected claim frequency $E(F_i)$, we utilize the data set containing all policyholder information available in the portfolio. The number of claims, $NumClaims(N_i)$, is used as the response variable. Furthermore, the frequency models include an offset represented by the logarithm of the exposure, $log(t_i)$. As inconsistent data on the policyholders’ residences are available for the policyholders who did not file a claim, we are not able to measure the spatial effects on claim frequency. Therefore, the analysis of frequency modeling in this thesis is limited to single parametric (GLM) and semi-parametric (GAM) approaches. Furthermore, as the methodology used, based on Chapter 5, aligns with the analysis of the severity model, we only briefly describe the aspects of frequency analysis that differs from what was previously described.

⁶The Fisher-Jenk’s natural breaks algorithm results in 6 different classes with 67, 148, 99, 105, 164 and 69 numbers of municipalities in each class, respectively.

8.1 GLM

As discussed in Chapter 3, the Poisson and negative binomial distributions are widely used probability distributions for modeling claim counts. Similarly to the severity analysis, we analyze a parametric model for both cases.

First, we examine the Poisson distribution with the equidispersion restriction, meaning that the mean and variance of the claim count should be equal. However, testing for overdispersion shows that we are dealing with data that is overdispersed (*variance* > *mean*), which is a common concern when modeling frequency data. Therefore, as an alternative, we analyze the negative binomial distribution. Working with this distribution leads to a lower AIC and BIC (see Table 3), and because of its ability to handle overdispersion, we, consequently, select the negative binomial distribution for modeling claim frequency.

Furthermore, the explanatory variable representing the type of product is dropped from the model; unlike for modeling claim severity, where it makes sense that the type of coverage influences the claim cost, it does not affect the claim frequency. The estimates of the coefficients are presented in Table 14 in the Appendix. Apart from two age groups, the remaining coefficients of the variables are statistically significant (p-value < 0.001).

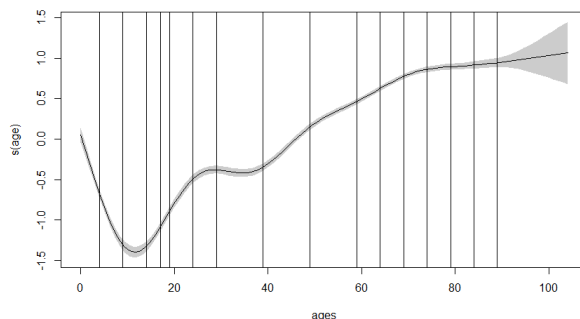
Table 3: Information criteria: claim frequency distributions.

	Poisson	negative binomial
AIC	211 951	197 578
BIC	212 145	197 783

8.2 GAM

Next, we analyze a GAM where we define age as a continuous variable instead of using the age classes. The non-linear effects of age, shown in Figure 6, clearly demonstrate that the expected claim frequency increases from the age of 12 years old onwards; the lowest number of expected hospitalizations is observed for younger people, while the curve peaks for older policyholders. Lastly, the coefficients of the smooth terms (p-value < 0.001) and the parametric coefficient of gender (p-value < 0.10) are statistically significant (shown in the Appendix in Table 15).

Figure 6: The fitted smooth effects of age on claim frequency.



9 Comparative analysis

9.1 Claim severity and frequency models

Before combining the frequency and severity models to compare the pure premiums, we will focus on them individually to observe the differences between the underlying models. Specifically, we begin our comparative analysis by employing the statistical techniques introduced in Chapter 5 to assess the goodness of fit and predictive performance of the different models.

To assess the models in terms of complexity and goodness of fit, we examine the information criteria for the severity and frequency models (shown in Table 4). With respect to the claim severity model, when either the GAM or LM account for the spatial effects, it leads to a lower AIC and BIC than those of the base LM. This suggests that including spatial information in the model is beneficial, yielding more accurate results. Furthermore, the LM with clustered spatial effects attains the lowest BIC, while the GAM has the lowest AIC. This reveals the trade-offs between a simpler LM, which is preferred by the BIC and the more flexible GAM preferred by the AIC. When examining the two frequency models, it is noticeable that using the GLM leads to a lower AIC and BIC.

Table 5 compares the predictive performance among the severity and frequency models based on the mean square error (MSE). For the claim severity models, the results suggest that the LM, without spatial effects, leads to the worst MSE. Furthermore, the GAM is preferable to the LM with spatial clusters since it has the lowest MSE. Even so, the results are similar and the difference between the in-sample and out-of-sample MSE is the smallest for the LM with spatial clusters, which could be an indication of a more stable model. For the frequency models, it is apparent that the GLM leads to the lowest MSE.

Table 4: Information criteria of the severity models with log normal distributions and frequency models with negative binomial distributions.

	AIC	BIC
Severity LM	77 702	77 878
Severity GAM	77 483	77 742
Severity LM with clustered spatial effects	77 489	77 706
Frequency GLM	197 578	197 783
Frequency GAM	197 832	197 961

Table 5: Prediction accuracy among the claim severity and frequency models based on MSE.

	MSE (in-sample)	MSE (out-of-sample)
Severity LM	84 790	91 359
Severity GAM	83 912	90 755
Severity LM with spatial clusters	83 937	90 766
Frequency GLM	0.1397	0.1318
Frequency GAM	0.1399	0.1319

9.2 Premium models: Inflow and predictive accuracy

In a next step, we multiply the expected claim frequency and severity to assess the total premium inflow and the predictive accuracy of the pure premium models. For this, we work with a holdout sample of 89,115 observations. The total loss for this sample consists of EUR 848,773. Table 6 demonstrates the three premium structures that we obtained, the corresponding total premium inflows, and MSE values.

Each premium model is sufficient to cover the total loss of the portfolio. Premium model 1, consisting of the base models without spatial effects, leads to the highest total premium inflow. The lowest total inflow is calculated for Premium model 2, which uses the GAMs. Premium model 3, based on the severity LM taking into account spatial effects and the frequency GLM, leads to a slightly higher but roughly the same total inflow.

Furthermore, the lowest MSE is observed in the Premium Model 3. Using the GAMs, the model's MSE is much the same. Alternatively, the MSE is the highest for Premium Model 1. As mentioned in Chapter 5, MSE values are highly affected by large claims, which no model could reliably predict. Therefore, we observe that the MSE values are similar across

the models.

These findings demonstrate that using spatial information can help actuaries make more accurate predictions on premiums. Furthermore, the results suggest that Premium Model 3, resulting from the frequency GLM and severity LM with spatial effects, is the most optimal in terms of predictive accuracy.

However, the differences in MSE and premium inflow between the parametric and semi-parametric models (accounting for the spatial effects) may be negligible. Examining their ability in fair pricing could make a significant difference in selecting the preferred model. To identify which model offers the fairest rates in terms of adverse selection, in the next section we will investigate the Gini indices and ordered Lorenz curves.

Table 6: Total premium inflows and MSE among the different premium models.

	Severity model	Frequency model	Total pure premium inflow	MSE out
Premium 1	Baseline LM	GLM	1 095 222	9 934
Premium 2	GAM	GAM	886 587	9 920
Premium 3	GLM with spatial clusters	GLM	886 659	9 918

9.3 Premium models: fair tariffs

While measuring the MSE showed us which model(s) perform better in terms of their predictive performance, our next step is crucial to assess which models can best withstand adverse selection. To measure the economic value of the obtained tariffs (i.e., the model’s capacity to generate fair tariffs), we analyze the ordered Lorenz curves and Gini indices for the three pure premium models. For this lift measure, we worked with the holdout sample introduced in the previous section.

To illustrate, Figure 9 in the Appendix shows us the ordered Lorenz curves for the three models. From this visualization, we note that the ordered Lorenz curves for both Pure premium models 2 and 3, which account for spatial effects, have a slightly convex shape when the base model is used as a benchmark, indicating that both models are able to spot tariff discrepancies in the benchmark model. Furthermore, when the model based on the GAMs is used as a benchmark, the curve for the Premium model 1 and 3 fall slightly below the line of equality. As it is hard to distinguish the exact differences between these models through the visualization of their Lorenz curves, the Gini indices are summarized in such a way that we are able to compare the models.

Table 7 denotes the matrix of Gini indices for the three pure premium models. The rows indicate the benchmark models used as the denominator in the relativity equation (R), while the columns indicate the models generating alternative tariffs, which represent the numerator

of the relativity (R). Hence, for the mini-max analysis introduced by Frees et al. (2013), the maximum value of each row is selected first; Pricing model 3, consisting of the parametric models with spatial effects, has the highest index (6.80) when the baseline model (Premium model 1) is specified as a benchmark. Furthermore, a slightly lower Gini index of 5.94 is retrieved for Premium model 2 (GAMs). When Premium model 2 is used as benchmark, the highest index is again achieved by Premium model 3 (4.91). Lastly, when Pricing model 3 is used as benchmark, the simple base model gives the highest index of 0.92. Of these three maximum indices, the benchmark model with the lowest value is selected: the GLM with clustered spatial effects (Premium model 3). As per the mini-max approach, this model can be seen as the most optimal choice, resulting in the fairest rates in terms of adverse selection. In other words, this model is the most capable of distinguishing good risks from bad risks and pricing them accordingly.

Table 7: Matrix of Gini indices of the competing pricing models. The benchmark models are represented by the different rows, while the alternative models are shown in the columns.

	Premium 1	Premium 2	Premium 3
Premium 1 (Base model)	0	5.94	6.80
Premium 2 (GAM with spatial effects)	2.98	0	4.91
Premium 3 (GLM with spatial effects)	0.92	0.34	0

10 Conclusion

In order to measure the effects of risk factors and set corresponding tariffs for different risk classes, the actuary can choose among several predictive models, including the ordinary LM, GLM, and GAM. Based on these models and the claim frequency-severity approach, this thesis constructed a framework that actuaries can use to predict pure premiums. Specifically, we conducted a hospitalization insurance case study to facilitate a comparative analysis of the models and corresponding tariffs. To perform this analysis, we studied the different statistical models and discussed methods to assess the models.

Since hospitalization data for claim costs and counts typically does not follow a normal probability distribution, we specified the adequate probability distribution for each model. Following an analysis of the deviance residuals and information criteria (AIC and BIC), we showed that the best fit involved a normal distribution with log transformation for claim severity and a negative binomial distribution for claim frequency.

Given that continuous (e.g., age) and spatial variables are crucial predictors for health insurance purposes, we continued by modeling a GAM that presents the non-linear effects of

these risk factors in a non-parametric way. As a second option, we considered a parametric LM or GLM, where we replaced the continuous variables with categorical variables. We then used the pre-defined age classes and clustered the smooth spatial effects of the GAM by using the Fisher-Jenk's natural breaks algorithm.

To identify whether it is possible to accurately predict pure premiums for the given insurance product, we combined the frequency and severity models and then used several measures to assesses the different models.

Based on out-of-sample data, each pricing model was able to cover the total losses. Ultimately, the preferred choice in terms of predictive performance was the premium model consisting out of the LM (with clustered spatial effects) for claim severity and the GLM for claim frequency. The measured MSE and total premium inflow of the premium model consisting out of the GAMs were similar. However, after analyzing the ordered Lorenz curve and Gini index, it was apparent that the pure premiums resulting from the parametric models (with spatial effects) suffered the least from adverse selection. Hence for the investigated insurance product, compared to a semi-parametric GAM, working with (generalized) linear models that include information about the policy holder's residence results in more optimal premiums.

It is worth noting that this analysis focused on the data of a Belgian insurer, for which limited risk factors were available. Hence, the foundational methodology contained in this thesis can enable researchers to build upon these findings – drawing conclusions for other insurance providers with more available risk factors, as per the models examined herein. Future research can also go beyond these statistical models by investigating machine learning methods (such as random forests, gradient boosting trees, regression trees) that are changing the world of insurance pricing by offering new methods supporting actuaries to obtain more accurate predictive pricing models.

As a final remark, focusing on other risk factors (such as tobacco use, pre-existing illness, BMI), acting on different underlying assumptions, or using additional models may give rise to different conclusions. Be that as it may, the method employed in this thesis extends a purely statistical comparison by also focusing on insurance related measurements. This enables researchers to assess how different predictive models can be helpful in drawing conclusions on insurance pricing that are not only valuable in theory but also in practice.

“Yet there is more work to do. Having a pure premium model is just one component of an overall pricing exercise.”

— Frees et al. (2014)

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Aleandri, M. (2019). *Data Science in Insurance*. PhD thesis, Institute and Faculty of Actuaries.
- Antonio, K. and Valdez, E. A. (2011). Statistical concepts of a priori and a posteriori risk classification in insurance. *ASTA Advances in Statistical Analysis*, 96(2):187–224.
- Assuralia (2019). kerncijfers per tak. <https://www.assuralia.be/nl/studies-en-cijfers/kerncijfers-per-tak/>.
- Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Springer New York.
- Brockman, M. J. and Wright, T. S. (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries*, 119(3):457–543.
- Christelijke Mutualiteit (2019). 15e cm ziekenhuisbarometer. <https://www.cm.be/professioneel/pers/persberichten-2019/ziekenhuisbarometer/>.
- Clijsters, M. (2015). Dealing with continuous variables and geographical information in non-life insurance ratemaking. Master’s thesis, KU Leuven, Belgium.
- Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.
- De Jong, P. and Heller, G. (2008). *Generalized linear models for insurance data*. Cambridge University Press, Cambridge New York.
- Denuit, M., Hainaut, D., and Trufin, J. (2019). *Effective statistical learning methods for actuaries I : GLMs and extensions*. Springer Publishing, Cham, Switzerland.
- Denuit, M., Marchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts*. John Wiley & Sons, Ltd.
- Dhaene, J., Stassen, B., Barigou, K., Linders, D., and Chen, Z. (2017). Fair valuation of insurance liabilities: Merging actuarial judgement and market-consistency. *Insurance: Mathematics and Economics*, (76):14–27.
- Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer Berlin Heidelberg.

- Dugas, C., Bengio, Y., Chapados, N., Vincent, P., Denoncourt, G., and Fournier, C. (2003). Statistical learning algorithms applied to automobile insurance ratemaking.
- European Union (2017). Gender equal access to goods and services directive 2004/113/ec. <https://www.europarl.europa.eu/RegData/etudes/STUD/2017/>.
- Federal Public Service of Health (2019). Key data in healthcare. <https://www.health.belgium.be/en/keydata-healthcare>.
- Frees, E., Lee, G., and Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1):4.
- Frees, E. W., Meyers, G., and Cummings, A. D. (2011). Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, 106(495):1085–1098.
- Frees, E. W., Meyers, G., and Derrig, R. A. (2014). *Predictive modeling applications in actuarial science*. Cambridge University Press, New York.
- Frees, E. W. J., Meyers, G., and Cummings, A. D. (2013). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 81(2):335–366.
- Goldburd, M., Khare, A., Tevet, D., and Gullder, D. (2019). *Generalized linear models for insurance rating*. Casualty Actuarial Society, 2nd edition.
- Gschlossl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.
- Henckaerts, R., Antonio, K., Clijsters, M., and Verbelen, R. (2017). A data driven binning strategy for the construction of insurance tariff classes. *SSRN Electronic Journal*.
- Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008). *Modern Actuarial Risk Theory*. Springer Publishing, 2nd edition.
- Klein, N., Denuit, M., Lang, S., and Kneib, T. (2014). Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55(C):225–249.
- Murphy, K. P., Brockman, M., and Lee, P. K. W. (2000). Using generalized linear models to build dynamic pricing systems.

- Nationale Bank van België (2019). Toezichtsdomeinen: verzekeringsondernemingen. <https://www.nbb.be/>.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, (135):370–384.
- Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer, Berlin London.
- Parodi, P. (2015). *Pricing in general insurance*. CRC Press, Boca Raton, FL.
- Pelsser, A. and Stadje, M. (2014). Time-consistent and market-consistent evaluations. *Mathematical Finance*, (24):25–65.
- Pitacco, E. (2014). *Health insurance : basic actuarial models*. Springer, Cham.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417–428.
- Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., and Lichtndahl, K. C. (2017). *Data mining for business analytics : concepts, techniques, and applications in R*. Wiley, Hoboken, New Jersey.
- Slocum, T. A., McMaster, R. M., Kessler, F. C., Howard, H. H., and Mc Master, R. B. (2008). *Thematic Cartography and Geographic Visualization*. Prentice Hall, 3rd edition.
- Tibshirani, R. (2014). Generalized linear models [powerpoint presentation]. <https://www.stat.cmu.edu/~ryantibs/advmethods/notes/glm.pdf>.
- Wood, S. (2017). Generalized additive models with integrated smoothness estimation. <https://www.rdocumentation.org/packages/mgcv/versions/1.8-31/topics/gam>.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 65(1):95–114.

11 Appendix

11.1 Figures

Figure 7: Normal Q-Q plots: gamma, inverse Gaussian, log normal distributions.

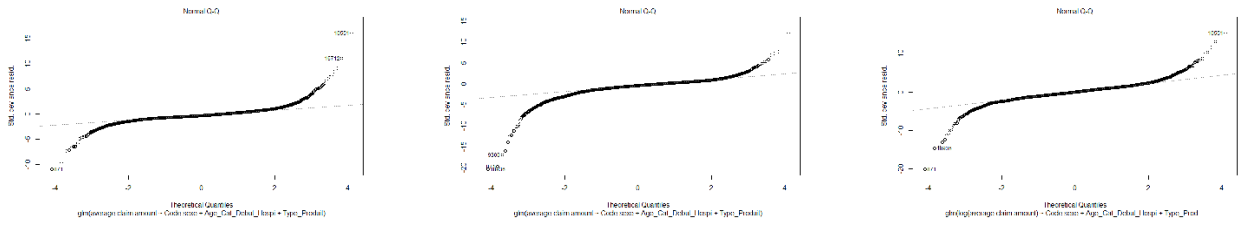


Figure 8: Histogram deviance residuals: gamma, inverse Gaussian, log normal distributions.

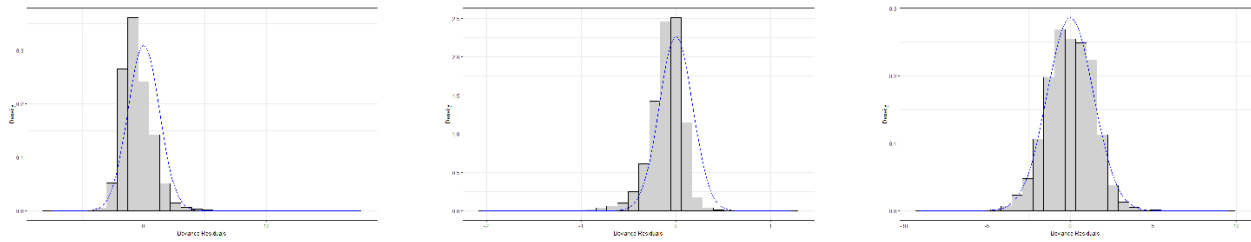
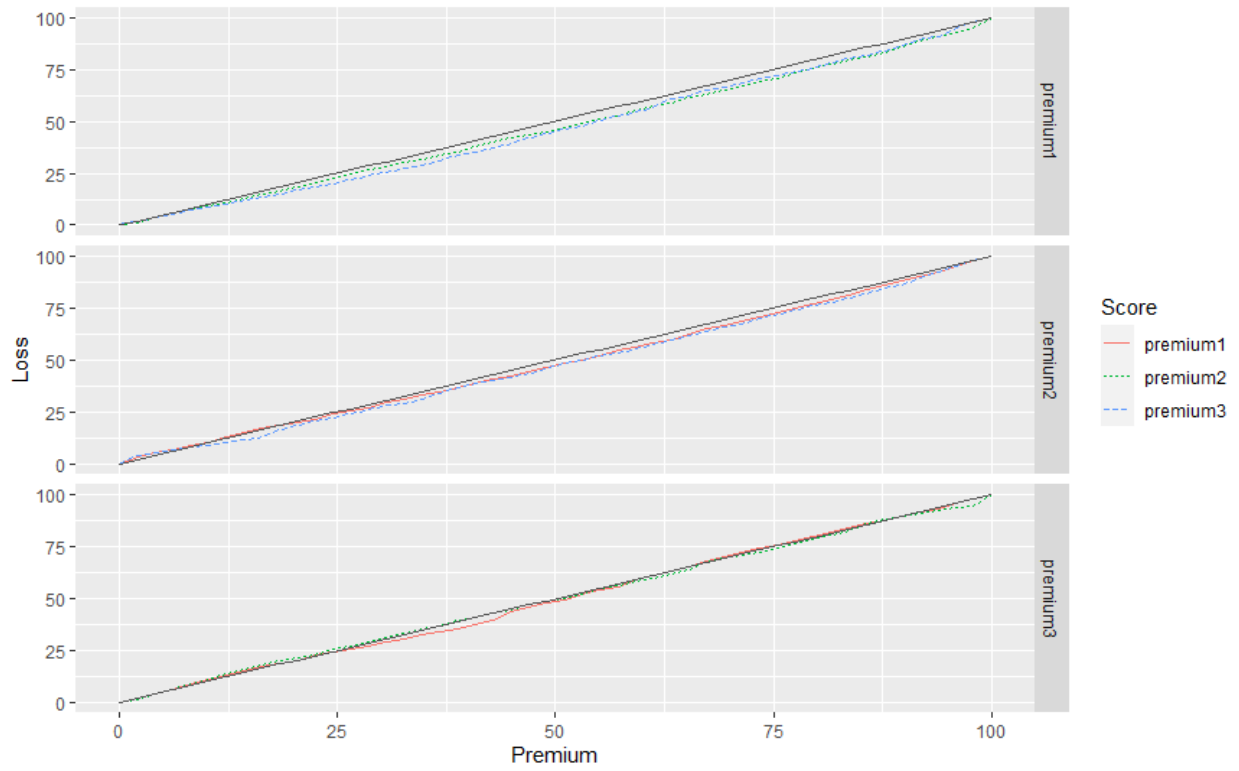


Figure 9: Ordered Lorenz curves



11.2 Tables

Table 8: Descriptive statistics: claim severity

	<i>Mean</i>	<i>Min.</i>	<i>Max.</i>
Number of claims	1.29	1	18
Average claim cost	115.26	1.01	12 500.00
	<i>Unique</i>	<i>Min (mean)</i>	<i>Max (mean)</i>
Gender	2	male (109)	female (122)
Postal code	664	8700 (1.44)	3290 (1 618.06)
Age categories	17	15-17 (63.1)	25-29 (186)
Type product	4	Optio 150 (185)	Hospimut (97.3)
Type recipient	3	N (85.5)	B (125)

Table 9: Descriptive statistics: claim frequency

	<i>Mean</i>	<i>Min.</i>	<i>Max.</i>
Number of claims	1.29	1	18
	<i>Unique</i>	<i>Min (mean)</i>	<i>Max (mean)</i>
Gender	2	male (0.085)	female (0.086)
Age categories	17	10-14 (0.017)	90+ (0.221)
Type product	4	Optio 100 (0.070)	Hospimut (0.091)
Type recipient	3	B (0.050)	S (0.100)

Table 10: Information criteria: used for the selection of spatial clusters.

nr of classes	AIC	BIC
3	77 475	77 613
4	77 469	77 611
5	77 470	77 620
6	77 462	77 620
7	77 462	77 629
8	77 463	77 635

Table 11: Estimates of the log-normal GLM for Claim severity.

Coefficients:											
	Estimate	Std. error	t value	Pr(> t)							
(Intercept)	3.58467	0.05213	68.771	< 2e-16	***						
Code.sexe	0.01233	0.01639	0.752	0.452							
Age_Cat_Debut_Hospi05-09	-0.13730	0.0875	-1.569	0.117							
Age_Cat_Debut_Hospi10-14	0.11125	0.08964	1.241	0.215							
Age_Cat_Debut_Hospi15-17	-0.05065	0.09144	-0.554	0.580							
Age_Cat_Debut_Hospi18-19	0.15611	0.10744	1.453	0.146							
Age_Cat_Debut_Hospi20-24	0.42809	0.07566	5.658	1.55e-08	***						
Age_Cat_Debut_Hospi25-29	0.50273	0.06676	7.531	5.24e-08	***						
Age_Cat_Debut_Hospi30-39	0.32099	0.05711	5.621	1.92e-08	***						
Age_Cat_Debut_Hospi40-49	0.27012	0.05392	5.010	5.49e-07	***						
Age_Cat_Debut_Hospi50-59	0.21265	0.05051	4.210	2.56e-05	***						
Age_Cat_Debut_Hospi60-64	0.28075	0.05256	5.341	9.31e-08	***						
Age_Cat_Debut_Hospi65-69	0.36336	0.05122	7.094	1.34e-12	***						
Age_Cat_Debut_Hospi70-74	0.42106	0.05091	8.271	< 2e-16	***						
Age_Cat_Debut_Hospi75-79	0.43279	0.05312	8.147	3.92e-16	***						
Age_Cat_Debut_Hospi80-84	0.45108	0.05482	8.229	< 2e-16	***						
Age_Cat_Debut_Hospi85-89	0.41232	0.05894	6.996	2.70e-12	***						
Age_Cat_Debut_Hospi90+	0.39376	0.07189	5.477	4.37e-08	***						
Type_Produitopio 100	0.18220	0.04005	4.550	5.40e-06	***						
Type_Produitopio 150	0.33133	0.04023	8.237	< 2e-16	***						
Type_Produitopio 200	0.11548	0.02255	5.122	3.05e-07	***						
- - -											
Signif. Codes:	0	***	0.001	**	0.01	*	0.05	.	0.1	'	1

Table 12: Estimates of the log-normal GAM for Claim severity.

Parametric coefficients:							
	Estimate	Std. Error	t value	Pr(> t)			
(Intercept)	3.85148	0.02823	136.442	<2e-16	***		
Code.sexe	0.03193	0.01679	1.902	0.0571	.		
Type_ProduitOptio 100	0.21275	0.04065	5.233	1.68e-07	***		
Type_ProduitOptio 150	0.33008	0.04179	7.898	2.96e-15	***		
Type_ProduitOptio 200	0.10794	0.02317	4.658	3.21e-06	***		
- - -							
Signif. Codes:	0	****	0.001	***	0.01 **	0.5 .	0.1 ' ' 1
Approximate significance of smooth terms:							
	edf	Ref.df	F	p-value			
s(AgeFirstClaim)	8.653	8.954	25.68	< 2e-06	***		
s(Latitude.Longitude)	17.796	22.231	10.17	< 2e-06	***		
- - -							
Signif. Codes:	0	****	0.001	***	0.01 **	0.5 .	0.1 ' ' 1

Table 13: Estimates of the log-normal GLM with clusters for Claim severity.

Coefficients:											
	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	3.37246	0.05347	63.070	< 2e-06	***						
Age_Cat_Debut_Hospi05-09	-0.15273	0.08369	-1.825	0.068006	.						
Age_Cat_Debut_Hospi10-14	0.17225	0.08705	1.979	0.047846	*						
Age_Cat_Debut_Hospi15-17	-0.04688	0.08752	-0.536	0.592213							
Age_Cat_Debut_Hospi18-19	0.17587	0.10397	1.691	0.090755	.						
Age_Cat_Debut_Hospi20-24	0.44667	0.07311	6.110	1.01e-09	***						
Age_Cat_Debut_Hospi25-29	0.59014	0.06470	9.122	< 2e-06	***						
Age_Cat_Debut_Hospi30-39	0.38460	0.05540	6.942	3.97e-12	***						
Age_Cat_Debut_Hospi40-49	0.30925	0.05258	5.881	4.13e-09	***						
Age_Cat_Debut_Hospi50-59	0.23282	0.04930	4.723	2.34e-06	***						
Age_Cat_Debut_Hospi60-64	0.26720	0.05162	5.176	2.29e-07	***						
Age_Cat_Debut_Hospi65-69	0.33873	0.05022	6.746	1.56e-11	***						
Age_Cat_Debut_Hospi70-74	0.4153	0.05000	8.251	< 2e-06	***						
Age_Cat_Debut_Hospi75-79	0.40026	0.05239	7.639	2.27e-14	***						
Age_Cat_Debut_Hospi80-84	0.40938	0.05394	7.589	3.35e-14	***						
Age_Cat_Debut_Hospi85-89	0.34961	0.05814	6.013	1.84e-09	***						
Age_Cat_Debut_Hospi90+	0.37252	0.71320	5.223	1.78e-07	***						
Code.sexe	0.03179	0.01678	1.894	5.82e-02	.						
Clusters[-0.097,-0.028)	0.07927	0.02282	3.474	5.15e-04	***						
Clusters[-0.028,-0.055)	0.21406	0.03157	6.780	1.23e-11	***						
Clusters[0.055,0.12)	0.26619	0.03301	8.064	7.75e-16	***						
Clusters[0.12,0.17)	0.31821	0.02736	11.630	< 2e-06	***						
Clusters[0.17,0.25]	0.34691	0.03130	11.084	< 2e-06	***						
Type.ProduitOptio 100	0.21880	0.04031	5.428	5.76e-08	***						
Type.ProduitOptio 150	0.32439	0.04124	7.865	3.84e-15	***						
Type.ProduitOptio 200	0.10590	0.02289	4.626	3.74e-06	***						
- - -					***						
Signif. Codes:	0	'***'	0.001	'**'	0.01	'*'	0.5	'.'	0.1	' '	1

Table 14: Estimates of the negative binomial GLM for claim frequency.

Coefficients	Estimate	Std. Error	z value	Pr (> z)								
(Intercept)	-2.9448759	0.0421845	-69.809	<2e-16	***							
Code.sexe	-0.0285924	0.0143283	-1.996	0.046	*							
Age_Cat_Debut_Hospi05-09	-0.8965429	0.0666252	-13.457	<2e-16	***							
Age_Cat_Debut_Hospi10-14	-1.0339654	0.0676775	-15.278	<2e-16	***							
Age_Cat_Debut_Hospi15-17	-0.5348420	0.0702454	-7.614	2.66e-14	***							
Age_Cat_Debut_Hospi18-19	-0.4711434	0.0814129	-5.787	7.16e-09	***							
Age_Cat_Debut_Hospi20-24	-0.3968652	0.0585983	-6.773	1.26e-11	***							
Age_Cat_Debut_Hospi25-29	-0.0834314	0.0525481	-1.588	0.112								
Age_Cat_Debut_Hospi30-39	0.0006341	0.0451810	0.014	0.989								
Age_Cat_Debut_Hospi40-49	0.2580767	0.0429567	6.008	1.88e-09	***							
Age_Cat_Debut_Hospi50-59	0.6915066	0.0406275	17.021	<2e-16	***							
Age_Cat_Debut_Hospi60-64	0.8919083	0.0427891	20.844	<2e-16	***							
Age_Cat_Debut_Hospi65-69	1.0599984	0.0418857	25.307	<2e-16	***							
Age_Cat_Debut_Hospi70-74	1.2157083	0.0418348	29.060	<2e-16	***							
Age_Cat_Debut_Hospi75-79	1.2772898	0.0443156	28.823	<2e-16	***							
Age_Cat_Debut_Hospi80-84	1.2727560	0.0461291	27.591	<2e-16	***							
Age_Cat_Debut_Hospi85-89	1.3696301	0.0511806	26.761	<2e-16	***							
Age_Cat_Debut_Hospi90+	1.4724061	0.0665413	22.143	<2e-16	***							

Signif. Codes:	0	****	0.001	***	0.01	**	0.5	.	0.1	'	'	1

Table 15: Estimates of the negative binomial GAM for claim frequency.

Parametric coefficients:												
	Estimate	Std. Error	z value	Pr(> z)								
(Intercept)	-2.60120	0.02393	-108.702	< 2e-16								
Code.sexe	-0.02386	0.01434	-1.664	0.0961								

Signif. Codes:	0	****	0.001	***	0.01	**	0.5	.	0.1	'	'	1
Approximate significance of smooth terms:												
	edf	Ref.df	Chi.sq	p-value								
s(age)	8.916	8.998	7025	<2e-16								

Signif. Codes:	0	****	0.001	***	0.01	**	0.5	.	0.1	'	'	1