Geographical risk in the Dutch car insurance

A data-driven approach to measure regional effects on the claim frequency

Diederik de Bont

2020-2022 Master Industrial Engineering

Thesis Assignment, December 2022

Faculty of Behavioural, Management and Social sciences

UTwente

P.O. Box 217, 7500AE Enschede

The Netherlands

Supervisors: Dr. B. Roorda

Dr. R.A.M.G. Joosten

B.O. van Grevenhof, MSc AAG

V. Verwijmeren, MSc

UNIVERSITY OF TWENTE.

Acknowledgements

This master’s thesis in actuarial science was completed in December 2022, as the final part of a Master of Science in Financial Engineering and Management. It also represents the culmination of my time as a student. With all the unexpected occurrences, new acquaintances, and (nearly) limitless freedom, these years have been fantastic. I thoroughly enjoyed them, and I can now see why people say that ”your student years are the best years of your life”.

I would like to thank Dutch insurance company for giving me the opportunity to conduct my master thesis at the Schade Particulier department. They made it possible to conduct my research and also experience some of the aspects of the typical daily activities at the insurance firm. Everyone was welcoming and willing to answer any sort of questions. Special thanks go out to Boris van Grevenhof. His master classes made sure I was quickly up to speed and understood all important concepts. It was a privilege having you as my supervisor, thank you for all your guidance. Furthermore, I would like to thank Vera Verwijmeren for sharing her knowledge on data science with me. Her critical view helped me improve my model and my overall research significantly.

Support did not only come from within the company. Berend Roorda was my supportive supervisor at the University and it was his accurate reading, thought-provoking conversations, and in-depth feedback that allowed me to strengthen my research and adjust my course where needed. Second opinions are priceless, and I want to thank Reinoud Joosten for that. He helped me submit a thesis that met the standards demanded of a Financial Engineering student by taking the time to read it and provide invaluable input.

While my thesis was my main focus in the past months and my studies were the focus for the past years, the external activities happening outside of my academic development were the most important to me. I want to thank my parents, my brother, all my friends and everyone who helped me, for all the everlasting support, sincere advise, good talks, and the great adventures. Those are the things that make your student days the best days of your life.

Finally, I thank you, the reader, for taking the time to read this report.

Diederik de Bont
Utrecht, December 2022

Management summary

It is important that car insurance companies can set an accurate premium. In order to do this, policyholders need to be classified in different risk levels based on their risk profile. For this, the frequency-severity method is most often used, in which the claim frequency and claim severity are modeled separately. To model the claim frequency, generalized linear models (GLMs) are the current industry standard, as they are easily interpretable and explainable. However, one of the major issues with using GLMs is the implementation of numerical risk factors, as often, these cannot be inserted into the model directly. This is why, we aim to improve upon the risk classification of the claim frequency model of the car insurance product, taking the geographical area (longitude and latitude of a policyholder's postal code) as the bivariate numerical risk factor to be incorporated into the model, while maintaining the interpretability and explainability of the frequency model.

In order to achieve this goal, we use a dataset containing records from over 1.7 million policyholders from one specific car insurance policy. To deal with missing values, we pre-process this dataset using the MissForest data imputation algorithm. Subsequently, we take two different approaches for incorporating the postal codes in the frequency model. Firstly, we incorporate the spatial variable into a generalized additive model (GAM) using smooth functions. Secondly, we incorporate the spatial variable as a categorical risk factor into a GLM by clustering the postal codes using Jenks natural breaks algorithm. The latter algorithm has been chosen after an in-depth comparative analysis between several clustering methods. Using the AIC, BIC and the Goodness of Variance Fit, 28 clusters have been chosen for this specific dataset. We refer to the two models developed here as the GAM and the improved GLM, respectively. Next, we compare these models to the current industry-standard GLM, and assess them on their ability to predict accurately claim frequency using K-fold cross-validation.

When comparing the two models (GLM and GAM) that include the new spatial variable to the industry-standard GLM, we observed that the two models developed here resulted in more accurate predictions on the claim frequency (the mean squared error (MSE) and mean absolute error (MSA) decreased with 1.65% and 1.60% respectively in the out-of-sample test, respectively, for both models). These may seem relatively small improvements in the predictive performance of the models. However, based on a portfolio of 1.7 million policies, this can have a huge impact on the premiums. Between the two proposed approaches, no notable difference in predictive accuracy could be observed. However, even though the GAM allows for the flexible modeling of numerical bivariates without imposing any nonlinear assumptions, the improved GLM has a practical advantage over the GAM, since it is easier to interpret and explain to stakeholders as it is formulated within the well-known framework of generalized linear models. The approach developed here

has two shortcomings. First, determining the optimal number of clusters using these metrics did not lead to a clear-cut optimal number of clusters. Second, feature selection was not performed, even though analysis of the parameters of the improved GLM model showed that, after adding the spatial variable, not all remaining parameters remain statistically significant. Additionally, the proposed approach significantly increased computing time for the model. Despite this, our approach has led to many new insights on the origin of the effect of the spatial variable on claim frequency due to the visualizations produced in this research. Additionally, with the clustering method developed here, the spatial effect is distributed more gradually throughout the Netherlands, which leads to fairer premiums.

In conclusion, we show that both of the models developed here allowed for the implementation of a numerical spatial variable in the frequency model, and that both result in an improved risk classification. We recommend to apply the improved GLM including the spatial variable and use the produced dashboards to review shortcomings of the current industry-standard GLM. This approach can be extended for other numerical variables or interactions between covariates but also for claim severity models to determine the effect on premiums eventually. New experiments could be conducted for different scenarios to analyze specific effects.

Looking into the future, we recommend looking beyond the regression models presented here, by focusing on more flexible machine learning methods, since these methods might prove beneficial in order to predict the pricing models more accurately. Knowing that these machine learning methods also have their own limitations, we propose to focus future research on a hybrid system in which actuaries obtain part of the data from machine learning models as input for the statistical models. This seems to be the most efficient way to combine the quality benefits of an machine learning model with the standard-industry regression models.

Contents

List of Acronyms

AIC	Akaike Information Criterion
AVG	Algemene Verordening Gegevensbescherming
BIC	Bayesian Information Criterion
BM	Bonus-Malus
CSV	Comma separated-value
CDF	cumulative distribution function
GAM	Generalized additive models
GDPR	General data protection regulation
GLM	Generalised linear models
GVF	Goodness of variance fit
IDD	Independent and identically distributed
KDD	Knowledge discovery in databases
kNN	k-nearest neighbor
LM	Linear model
MAE	Mean absolute error
MAR	Missing values at random
MCAR	Missing completely at random
MF	Miss forest
MSE	Mean squared error
NRMSE	Normalized Root Mean Squared Error
PFC	Proportion of falsely classified
R2	Root squared
RF	Random forest
RMSE	Root mean squared error
TAI	Tabular accuracy index
TPL	Third party liability
WA	Wettelijke aansprakelijkheid
WAM	Wet aansprakelijkheidsverzekering motorrijtuigen

Chapter 1

Introduction

1.1 Introduction

The main activity of an insurance company is to provide financial protection against a certain type of risk. The insurance policy covers the filed claims of policyholders by charging them an insurance premium, which is the amount of money paid for the insurance policy. Within an insurance portfolio, the risk profiles of the policyholders differ. In order to represent the heterogeneity of risks within the portfolios, insurance companies differentiate premiums based on the policyholders' risk profiles (i.e. where high-risk policyholders pay a higher premium compared to low-risk policyholders) [1]. If the insurance company were to charge an equal premium across all risks profiles, the low-risk profiles would be more likely to accept an offer with a lower corresponding premium elsewhere, and the company would be left with only high-risk profiles. This phenomenon is called adverse selection. Consequently, these high-risk profiles would pay a premium that is lower than the expectation of the insured loss and the company would be expected to lose money over time.

In order to overcome this problem, insurance companies rank policyholders in different risk levels based on their risk profile. Within a risk level, policyholders pay an equal premium that is meant to reflect their inherent risk. The construction of these risk levels is known as risk classification, see e.g., Kaas et al. [2]. Insurance companies use this method to offer competitive pricing and lower the costs of insurances. This gives additional motivation for insurance companies to differentiate between policyholders further.

When modeling risk levels, the industry standard is to model two components separately, namely the claim frequency and the claim severity [3]. The claim frequency is the number of claims per unit of time for which premium has been paid, and claim severity refers to the average claim cost. The unit of time for which premium has been paid referred to as the exposure. Since claim frequency and severity are assumed to be independent, the modeled risk of a policyholder is simply given by the product of both models. So, for a given year, the estimated risk - the total predicted loss - of a specific policyholder is given by the estimated claim frequency times the estimated claim severity for that year.

The existing heterogeneity within an insurance portfolio, as well as the growing competition among insurances are motivations for an insurance company to continue to improve the analysis of the risk profile of their policyholders [4]. Due to the extensive availability of data within the car insurance portfolio, more advanced models and algorithms enable insurers to charge a premium corresponding to the risk of a particular policyholder. Claim frequency can be determined with better accuracy than claim severity, as it is usually far more stable than claim severity [5]. Therefore, we focuses on the risk classification of the claim frequency for car insurances. Alternative techniques may provide a better prediction than the current claim frequency model or provide an addition to this model.

1.2 Problem identification

The aim of risk classification is to group policyholders in risk levels so that, for each policyholder, the premium they pay reflects their inherent risk as well as possible. This premium is called the pure premium, and represents the total predicted loss of an individual policyholder. An insurance company has large amounts of data, containing historical claims as well as the corresponding characteristics of the policyholders. These data are used to construct the different risk levels in a data-driven way, in order to assign each policyholder to a risk level accurately.

For risk classification, insurance companies make use of regression techniques based on classification variables. Generalized linear models (GLMs) are commonly used models for classifying risk within the insurance sector [3]. The success of GLMs in an insurance setting is due to their mathematical and computational tractability, as well as their stability. Furthermore, when utilizing GLMs, it is relatively straightforward to construct a practical, flexible and interpretable premium that can be explained easily to stakeholders [6]. Moving away from this highly embedded standard is nearly impossible in practice, at least in the short term.

Classification variables can be divided into categorical or numerical variables. Categorical variables can take on a discrete number of possible outcomes or levels that are not in any specific order. Examples of a categorical risk factor for car insurance are the gender of the policyholder or the brand of a car. Numerical variables can be subdivided into discrete and continuous, where discrete variables (e.g., the age in years of a policyholder) are finite countable values and continuous variables (e.g., the value of the car) take on an infinite number of possible values within a given range. Within the group of discrete variables, we define a subgroup of spatial variables (i.e., the postal codes and their corresponding longitude and latitude).

Constructing a GLM is rather straightforward when all risk factors are categorical. One of the major issues when using GLMs arises with the implementation of numerical risk factors. First of all, there are usually too many values to include in the statistical model directly [7]. It is inefficient to take all these values into account separately, since it results in too many risk levels each containing very few policyholders. Risk levels need to have sufficient exposure in each group for the effects to be statistically significant. Besides, it take a lot of computation time when many factors are included. Currently within the Dutch insurance company, for implementing numerical variables in the industry-standard GLM is clustering them into different risk classes. This clustering is based on expert judgment, as experts manually determine the different clusters for each numerical risk factor. For example, the variable driver age contains many levels, one for each age in years. As assumed, two adjacent years are often not significantly different from each other, therefore, age classes containing consecutive years can be formed, which then result in significantly different effects. Naturally, this process is rather subjective and very time-consuming.

This research focuses on the geographical area (based on postal codes) where a policyholder resides. Including the geographical area into the model is important as it can have a significant effect one the claim frequency. For example, one can imagine that a car in a city is damaged more often due to the busier traffic than in a sparsely populated area. The Netherlands has 458,114 different postal codes. The postal code cannot be included in the GLM model directly, because all these 458,114 levels separately do not have enough exposure to be statistically significant and additionally, it will take a lot of computational time. In order to implement the geographical effect in the GLM, the Dutch insurance company cluster the postal codes based on geographic variables provided by an external party, Statistics Netherlands (in Dutch, Centraal Bureau voor de Statistiek (CBS)), such as the urbanization grade or the province. These geographical variables have not been provided on an individual basis but instead indicate the averages in a particular postal code. This clustering method might cause problems when, for example, two neighbors have to pay different premiums simply because they are separated by a province border [8]. Additionally, the use of these geographic variables causes a loss of information, since a step-wise function intends to capture the effect of a numerical risk factor [9]. When transforming numerical risk factors by clustering them into categorical variables, it is crucial to have clustered intervals that represent the effect of the numerical variables as closely as possible [8]. This is why we want to improve upon risk classification by researching how we can cluster the spatial variable as a risk factor in a data-driven way that reflects the risk effect of the variable as closely as possible.

1.3 Motivation

The interest of a Dutch insurance company is to implement an alternative approach to incorporating the spatial dimension in the GLM analysis of the rating algorithm. By developing a data-driven method to cluster the geographical locations of the policyholders' residence, a categorical risk factor can be obtained that reflects the claim frequency of the geographical location. This strategy might prove beneficial in predicting the claim probability and improving the pricing of the car insurance product. Additionally, the goal is to asses the adequacy and performance of the alternative approach, given the observed data, and to examine the potential applications of this approach.

As previously mentioned, just like most insurers, the Dutch insurance company's premium pricing models are built up out of GLMs. The main reasons for using GLMs is that the pricing models should be explainable to the different stakeholders, simple to program, and adaptable to marketing demands and benchmark studies where competitors are analyzed. This is why more advanced and high-dimensional models are not yet implemented in the industry. An important challenge for the alternative approach is to maintain those advantages and to be an addition to the current pricing model used within the Dutch insurance company.

We translate these motivations in a clear research objective as follows, *Improve upon the risk classification of the claim frequency model of the car insurance product by implementing the geographical area as a risk factor, while maintaining the model easily interpretable and explainable.*

With this, we try to explain an additional part of the residuals that are left unexplained by the original GLM by implementing the geographical area as a risk factor. Achieving this goal will lead to an improved accuracy of the claim frequency model for the car insurance product.

1.4 Research questions

Achieving these research objectives, motioned in the previous section, leads to an improved pricing model for the car insurance product. To accomplish these goals, we set the following research question,

How can we implement the geographic locations of policyholders residences as risk factors in the claim frequency model, while maintaining the interpretability and explainability of the model?

We formulate the following sub-questions to divide the research question into different parts.

1. How can we handle missing data in the dataset?
2. Which modeling techniques could be used to handle numerical risk factors and what are their differences??
3. How can we cluster the geographical location in such way that it reflects the effect it has on the claim frequency?
4. How can the optimal number of clusters be determined that best reflect the distributional differences across the variable dimensions?
5. How can the statistical models best be validated, and their results best be compared?

1.5 Research method

We base our research framework on the *Knowledge Discovery in databases* (KDD) methodology. KKD is an end-to-end methodology of extracting knowledge out of extensive amounts of data in the context of large databases by creating and evaluating innovative artifacts [10]. The knowledge obtained through this methodology may become additional data that can be used for further usage and discovery. This methodology is often applied in the field of research within statistical modeling, machine learning, and data visualization [11]. In this research framework, the KDD methodology leads to the following chronological and repetitive steps:

1. **Developing a deeper understanding of the goal**
   In the first phase, we start to develop a deeper understanding on the context of the car insurance product and the associated pricing model to understand the application domain involved and the knowledge that is required. This first phase is outlined in Chapter 2.

2. **Understanding and pre-processing the dataset**
   The second phase is outlined in Chapter 3. In this phase we visualize and analyze the selected dataset by performing simple statistics and we plot some figures in order to create a deeper understanding of the data. Furthermore, we perform literature research relating to strategies for handling missing data, removal of noise or outliers and models used for car insurance datasets. This provides support for the pre-processing and cleaning of the dataset design process and helps to justify the research. In addition, this phase incorporates all activities necessary to improve the reliability of the data and process the initial dataset into the final dataset used in the modeling. The tasks we conduct in this phase are handling the missing data.

3. **Analyzing the regression models**
   In the third phase we discuss different modeling techniques that are relevant for risk-based pricing in the insurance industry. Furthermore, we investigate the effectiveness of and differences between the modeling techniques in determining the claim frequency and focus on how these techniques handle numerical risk factors. This phase is outlined in Chapter 4.

4. **Selecting the clustering method**
   In the fourth phase we compare different clustering algorithms. Caruana & Mizil [12], and Han & Kamber [13] prove that the decision on the selection of the clustering technique is crucial for the outcome. Therefore, it is important to understand the situation under which an algorithm is most suitable. Furthermore, a performance metric measures how well the clustering algorithm performs and to select the optimal number of bins. Therefore, we select an appropriate performance measure. This research is outlined in Chapter 5.

5. **Interpretation of the results**
   In fifth phase we analyze and interpret the obtained outcomes of the frequency models. We identify important features and their relation to the predictions. This phase is outlined in Chapter 6.

6. **Model comparison and validation**
   In sixth phase we evaluate the new model on the selected performance metrics and compare the results of the different models. This phase is outlined in Chapter 7.

7. **Define Conclusions and recommendations**
   We need to organize and present the obtained knowledge of this research in such way that it is useful for the Dutch insurance company. Therefore, in Chapter 8 we discuss the research problem, answer the research questions defined in Section 1.4, and provide some final recommendations.

1.6 Ethical framework

Personalized premiums can ensure 'fairer' premiums, since the costs of individual risks are better estimated and priced. At the same time, individualized contributions can undermine the solidarity within the Dutch insurance sector, as well as enhance indirect discrimination. When consciously responding to weaknesses of certain target groups there is discrimination, which can lead to ethical concerns, such as the loss of privacy, perception of unfair prices, power shift, but also a reduction in wealth [14]. Therefore, the use of personalized premiums is limited by law, social acceptance as well as the moral framework of insurers. For example, it is legally prohibited to differentiate by race or sex. This is why the Dutch insurance company distinguishes between risk factors, which have an effect on the claim frequency or severity and are legal for determining the tariffs, and factors, which are not legal for determining the tariffs, but are still used to analyze the risks. We focus only on the postal code of the residence of an policyholder, which is not prohibited by law. This means that we can use this factor directly in the model. Nevertheless, in practice, for actual pricing, the strict legislation around the risk factors one is evaluating must always be taken into consideration.

Collecting, processing, and observing personal data of their customers on a large scale is the main activity of insurers' work in order to set premiums and estimate risks. To protect the data of the customer, the new European privacy legislation, General Data Protection Regulation (GDPR in English) or the Algemene Verordening Gegevensbescherming (AVG in Dutch), has been in force since May 25th, 2018 [15]. Insurance companies must meet strict requirements in order to comply with the privacy standards of this legislation. The research conducted here incorporates sensitive data associated with private individuals. Therefore, data protection has been of utmost importance. To ensure data protection during this research we perform several concrete actions:

1. The name of the company for which the study is being conducted is anonymized within this report and shall be named as ”the Dutch insurance company”.
2. All documentation that is made public must be approved by an agreed person from the Dutch insurance company, to prevent personal and confidential information and from being published.
3. All data shown in the report contains noise and therefore does not represent the real portfolio for confidentiality reasons. Moreover, the axes and legends of figures are blanked out.
4. Personal data are left out of the dataset to ensure the data cannot be traced back to private individuals.

Chapter 2

Context description

In this chapter, we describe the context in which this research is conducted. In order to improve the risk classification of the frequency model, the structure of the legislation concerning car insurance in the Netherlands and the different kind of products should be clear. Therefore, in Section 2.1, we outline the Dutch legislation relating to car insurance and the car insurance products. In Section we elaborate on the current pricing scheme.

2.1 Car insurance

Insurances in the Netherlands can be divided into life and non-life. According to the Dutch Motor Insurance Liability Act (Wet aansprakelijkheidsverzekering motorrijtuigen (WAM) in Dutch), car insurances fall within the non-life sector of the insurer [16]. The non-life sector represents the losses that are incurred from a specific financial event and are compensated to the insured. Along with car insurance, non-life insurance covers home, theft, fire and travel insurances. These non-life insurance plans are characterized by their short term (often one year) and in the the case of a claim, the reimbursement is normally a one-time payment. In the Netherlands, it is regulated by law that all cars must have at least Third Party Liability insurance [16]. Extra vehicle insurance for fire, theft, vandalism is optional. In general, there are three forms of auto insurance [17]:

Third Party Liability
Third Party Liability (TPL) insurance is the minimum coverage and is required by law. The third-party insurance covers the damage your car causes to others and their property. To protect others, this insurance has been made mandatory by the Dutch government.

Third Party Liability+
Here, damage to your the policyholder's car is partly insured. This coverage consists of two parts, TLP and Limited Casco, and covers damage to others as well as damage to your own car as a result of theft, fire, broken windows, storm, hail and collision with animals. Damage you cause to your own vehicle is excluded.

All Risk
This is the most comprehensive coverage. The insurance contains the components TLP and Casco, also called All Risk. All Risk covers not only the damage to others caused by you, but also any damage to your own car, even if it is your own fault. The premium for the full comprehensive insurance depends on, among others, the catalog value of the car. All Risk insurance is mostly used for newer cars (up to about 6 years old).

2.2 The pure premium

Setting the optimal premium for a car insurance product involves an understanding of costs, price elasticity, consumer preferences, and the strategic actions of competitors. The price of the premium of an insurance is based on three underlying components, namely operational expenses, profits, and the variable costs [18]. The operational expenses are estimates and covers, among others, the cost of salaries, agents' compensation, rent, legal fees, and postage. The profits are underwriting the earning margin of the insurances. Within this research however, we focus on the variable cost, which is the amount needed to cover the expected losses.

The pure premium is defined by the Dutch insurance company as the expected cost of all claims that a policyholder will file during a coverage period. The question here is how high the premium should be to cover the cost of possible claims. The premium of all drivers combined should be at least equal to the expected cost of all filed claims that policyholders will make during the policy term [19].

In order to explain the equation of pure premium, we first introduce the claim frequency F_i and claim severity S_i of each policyholder i. The claim frequency is the ratio of the total the number of claims N_i per time unit t_i (also known as the exposure) for which premium has been paid. Therefore, the claim frequency is calculated with the following formula,

$$F_i = \frac{N_i}{t_i}, \tag{2.2.1}$$

where the exposure refers to the unit of time for which the policyholder is covered (e.g., exposure is the fraction of a year, so 1 represents a complete year and 1/12 represents one month). This time unit is used since not all policies are active during the entire year [20]. This is necessary, since some of the policies are contracted for only couple of months or even days. Besides, it can happen that a policy is closed and a new policy is opened when the characteristics of the policy in question change. This concept is explained in Section 3.1.2.

Claim severity S_i per policyholder indicates the average claim cost, and is expressed as total loss L_i over a time period divided by the number of filed claims N_i from the same time period. Therefore, the formula for the claim severity is as follows:

$$S_i = \frac{L_i}{N_i} \tag{2.2.2}$$

Based on Equations 2.2.1 and 2.2.2, we can define the pure premium p_i for a policyholder i. Assuming that the claim frequency and severity are independent, we can define the pure premium by modeling the claim frequency and severity separately. The pure premium of a policyholder is simply given by the product of both models, i.e. for a given year, the risk of a specific policyholder is given by the estimated frequency times the estimated severity for that year. The pure premium can thus be calculated through the following formula:

$$E(p_i) = E(F_i)E(S_i) \tag{2.2.3}$$

This method where the claim frequency and severity are modeled separately is called the frequency-severity method [3]. An alternative method to determine the pure premium is by modeling the premium directly. However, we opt to choose the frequency-severity method and assume independence between both models since this method has become the industry standard to calculate the pure premium [18]. Additionally, this approach has been shown to model a proper representation of the reality [21]. Furthermore, this method has a number of advantages [22] [5]. Firstly, the frequency-severity approach provides a deeper understanding of the underlying prior variables that contribute to the severity and frequency of claims. The interesting effects could disappear when we model the premium directly. An example of such effect is a contrary effect on its component (e.g., a variable that is negatively influencing the claim frequency but is equally positively influencing the claim severity) would be completely overlooked. Secondly, this method allows for the selection of a distribution for the claim frequency and severity separately. Usually, the the distributions of the claim frequency and severity are different. For predicting the claim frequency, the Poisson distribution is often used, while for predicting the claim severity, the gamma distribution is especially interesting. In Section 4.2 we further elaborate on on why these specific distributions are used. Lastly, claim frequency is usually far more stable than claim severity, and therefore claim frequency can be determined with better precision.

Regression models are commonly used to predict both the expected frequency and severity based on historical data. As explained in Section 1.2, it is not feasible for car insurers to determine a premium for all drivers individually. Therefore, by using the estimates of the response variables, the pure premium can be determined per risk level. It is critical for insurers to provide an accurate prediction of claim frequency and severity. Using the expected claim frequency and severity, insurers set a premium that should exceed the expected level of loss of the total claims filed. When the realized loss of claims is higher than the expected loss it receives in premiums, then the insurer makes a loss and may eventually go bankrupt.

Chapter 3

Data preparations

This chapter describes processes of preparing the data for our methodology. In Section 3.1 we describe the available dataset by analyzing the size and distribution of the different variables in the dataset[1]. Subsequently, in Section 3.2, we review literature on handling missing values in the data in general. Concluding, we describe how we deal with missing values in the dataset in Section 3.3.

3.1 Available data

The dataset considered have been provided by the Dutch insurance company. The data, spanning from 2015 to 2020, includes records of over 1.7 million policyholders from one specific car insurance policy. The years included in the dataset resulted from a trade-off between more observations and therefore more reliable results on the one hand, and the relevance of older observations as well as computational performance on the other hand. By considering a single insurance policy, a degree of homogeneity is assured to be present between the individuals. This homogeneity derives from the fact that selection bias, i.e. intrinsic variations across groups based on their insurance choices, can be avoided when just one type of policy is considered.

Each record within the dataset corresponds to a unique policyholder during a period of exposure. The characteristics of the policyholder are registered from the beginning of the policy period and do not change during this period of exposure. When a characteristic changes, a new record is created. The dataset can be split on a policyholder- as well as a claim-level basis. Note that we assume that the data provided by the Dutch insurance company is reliable and checked by the company itself. Therefore, we have not examined how reliable the dataset is. However, it is important to keep in mind that the presence of outliers in the dataset impacts the model to great extent since outliers increase the variability in your data, which decreases statistical power. Several diagnostic aspects and methods to detect and handle outliers in logistic regression are discussed in literature [23].

[1] Note that the data shown in the figures in this chapter, representing the distribution of the different variables, contains noise and therefore does not represent the real portfolio for confidentiality reasons. Moreover, both axis labels and legends are blanked out.

3.1.1 Policyholder data

The policyholder dataset is essential since we try to predict the occurrence of a claim based on differentiated policyholder information with higher accuracy. The dataset includes data on the policyholder and is described in Table 7.2. The dataset has been extended by means of two external datasets, one based on geographical factors, the other based on the characteristics of the car. The geographical variables have been provided by the municipalities as well as the waterboards of the Netherlands. As mentioned earlier, these geographical variables have not been provided on an individual basis but instead indicate the averages in a particular postal code. Since we have to protect the data of the policyholder according to the GDPR we removed the data such as telephone number, license plate, name of the policyholder, are removed. These data are left out of the dataset to ensure they cannot be traced back to private individuals.

Feature type	Examples
Policyholder data	Gender, age, postalcode, damage-free years, ect.
Geograhpical data	Province, urbanisation, ect.
Vehicle data	Brand, weight, fuel type, catalogue value, vehicle age, ect.
Insurance information	Starting end dates, payment clause, minimal insured amount, ect.

Table 3.1: Examples of the characteristics per data type field.

Figure 3.1 illustrates the visualization of three example independent risk factors (the age of the policyholders, the age of the car and the bonus-malus (BM)) and how these are distributed in the dataset. The distribution of the driver age appears similar to a normal distribution. The majority of the policyholders (85.23%) are aged between 25 and 70, indicating that the insurance portfolio contains few young and old policyholders. The vehicle age, however, has a long right tail. Most of the cars in the portfolio are from a younger age and the number of policyholders decreases substantially when the age of the car increases. The average age of the cars is left out due to the sensitivity of this company-specific information. Bonus-malus (BM) is a system that insurers use to reward damage-free driving [24]. If you drive without damage, you get a premium discount (bonus), but if you claim many damages, you get a premium surcharge (malus). Every year that you do not claim any damage, you get more discount on the premium of your car insurance. The BM level follows a distribution with a long left tail with most of the policy holders being within the highest BM level.

The map of the Netherlands in Figure 3.2 represents the exposure (as mentioned earlier, exposure refers to the unit of time for which the policyholder is covered) in each municipality. The map of the Netherlands in Figure 3.2 represents the exposure in each municipality, with the light orange to dark red gradient representing the lowest and highest relative exposures, respectively. The explored insurance product is mainly active near some big cities; Rotterdam, Eindhoven, Enschede, etc. It is apparent that few policyholders live in the northern part of the Netherlands. In addition, the insurance product is active in all the municipalities of the Netherlands).

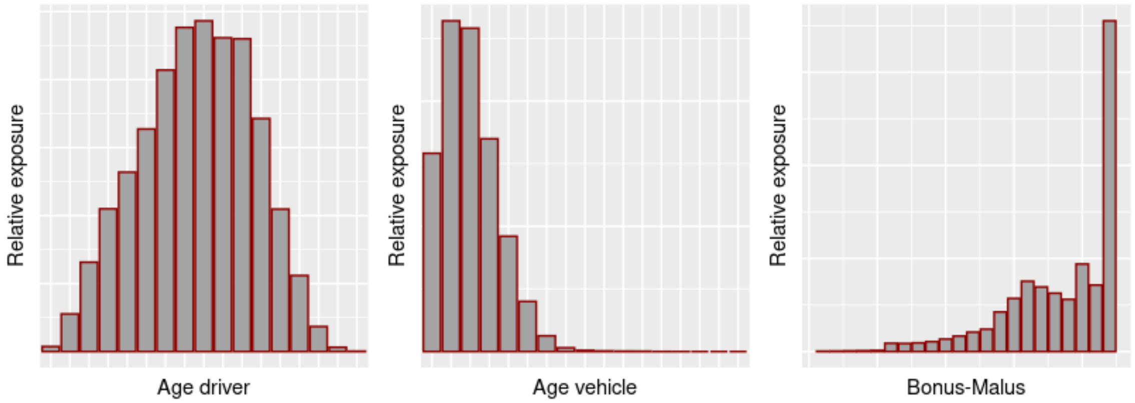

Figure 3.1: Histograms of the relative exposures of three example variables, the driver age, vehicle age as well as the bonus-malus level.

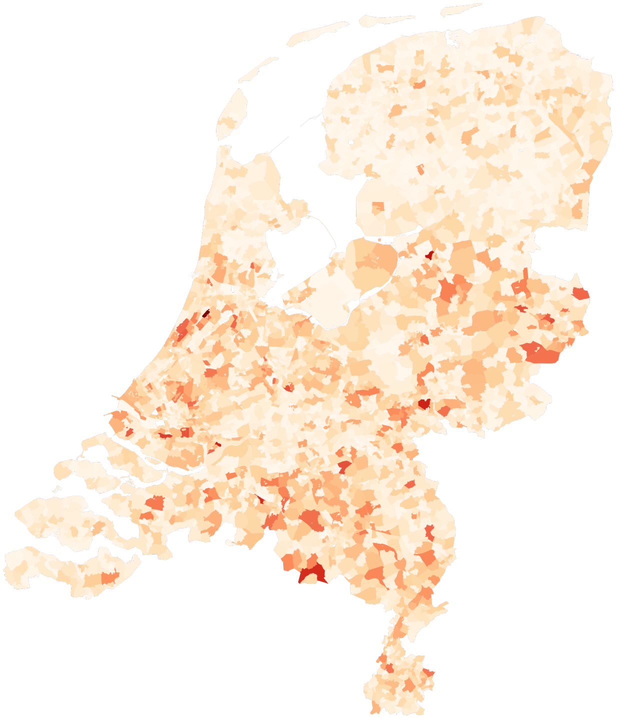

Figure 3.2: Map of the Netherlands with exposure per municipality.

3.1.2 Claim level data

Claim data, stored in a different data warehouse, contain the amount of money to cover the claim as well as the date of the payments. To obtain a dataset applicable for training a regression model, policyholder data from different years had to be merged with the claim data, since we want to have, for instance, the number of claims and the characteristics of a policyholder during a period of exposure. The primary key connecting the claim data to a specific record is the contract number. Most records do not contain claim data. In the insurance industry, claim data are often unbalanced, i.e. a large majority of the claim frequency observations are zeros. Specifically, in a given year, around 95% of the observations are zeros. However, this does not mean that only 5% of the policyholders files a claim on average per year, since there can be more than one observation for a specific policyholder in a year. That is, there is one observation per exposure with those specific characteristics, meaning that each time a characteristic changes within a year, a new observation is added. This is not only the case when a policyholder obtains an additional claim-free year (or it resets to zero), but also when a policyholder buys a new car or moves to a new house. Table 3.2 shows an illustration of the data. Notice that in each year, every observation corresponds to a set of characteristics with weights given by the exposure. The exposure weighted mean is equal to 9.0%, meaning that, on average, a policyholder files a collision claim approximately once every 11 years.

index	policy number	year	age	no claim	brand	district	sub-claims	exposure
1	123456	2020	35	10	Mercedes	urban	0	0.138
2	123456	2020	36	11	Mercedes	urban	0	0.862
3	123458	2020	45	30	Tesla	urban	1	0.379
4	123458	2020	45	0	Tesla	rural	0	0.099
5	123458	2020	46	0	Audi	rural	0	0.632
6	...	...	...	...	...	...	...	...

Table 3.2: Example of observations in the data.

Figure 3.3.A illustrates how the number of claims are distributed and shows that the claim data is imbalanced. Most of the policyholders do not fill a claim during their insured period and the number of policyholders decreases substantially when the total number of claims increases. Therefore, the number of claims are right-tail distributed. Figure 3.3.B shows a slight decrease in the number of claims from 2015 onward. This can possibly be explained by that, over the years, cars have become more safe. However, in 2020 we can see an ever steeper drop in the number of claims, caused by the COVID-19 pandemic, which caused less driving on average resulting in fewer claims[25]. In Figure 3.3.C we visualize the differences between the average claim frequency of the different policyholders' age groups . The overall claim frequency was the highest for the older age groups. Furthermore, this figure indicates a difference among male and female policyholders in terms of the average claim frequency. In younger and the older age groups, the claim frequency for males is higher. For the younger age groups this could be explained by the fact that a younger male is more reckless and consequently more likely be involved in an accident [26].

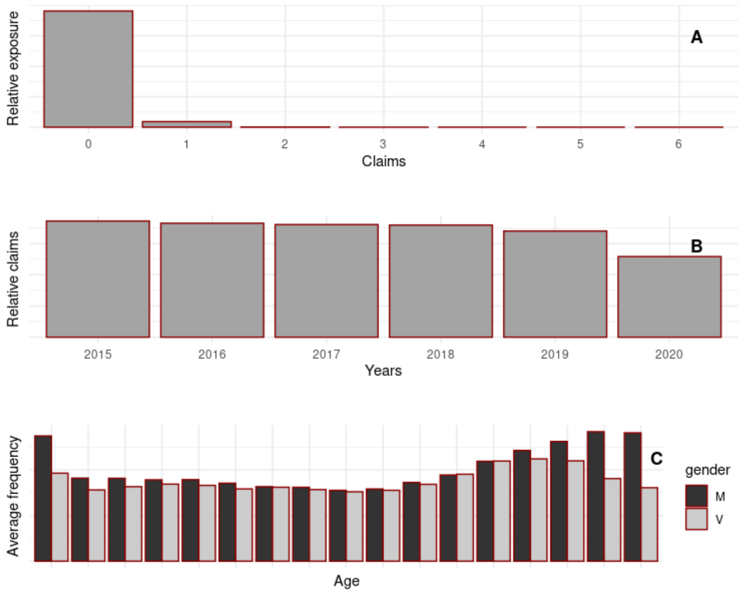

Figure 3.3: A: Distribution of the relative frequency of the total number of claims per policy period. B: Distribution of the total number of claims over the years 2015 - 2020. C: Distribution of the average claim frequency of gender and age.

3.2 Handling missing values in a dataset

After obtaining the dataset, an important step of the data pre-processing stage is dealing with missing values [10]. For regression or correlation analyses, a complete dataset is needed and therefore it requires data preparation and cleaning for useful knowledge extraction [27]. When missing values are not handled correctly, it could lead to undesirable effects, such as reduced statistical power of the analysis, bias of different statistical parameters, and a decrease in the degree to which the data describe reality [28] [29]. This may result in drawing incorrect inferences about the data. Furthermore, the approach used to deal with missing values can have a significant influence on data interpretation. Therefore, it has to be chosen carefully.

In statistics, list-wise deletion is a standard method for handling missing data [30]. This method removes all observations from your data that have a missing value in one or more variables. To produce appropriate results for this method, data must be missing completely at random (MCAR) (e.g., the missing values within the dataset must be unrelated to the observed values) [31]. When the data are only missing values at random (MAR), e.g, missing values occur randomly in a specific variable but the probability of being missed depends on values of one or more other observed variables, the estimates of this method will be biased [31]. Furthermore, although it is a common method used, It may result in the loss of critical information, which can be influencing the outcomes, particularly in imbalanced or small datasets.

Techniques for replacing missing data with calculated 'estimates' are known as missing value imputing [32]. These methods are used to retain as much data as possible. Simplistic methods are mean or median imputation, where missing values are replaced by the mean or median of the column in which the value was found. Another widely used method to impute missing values is the k-Nearest-Neighbor (kNN) method [32]. This method identifies the neighboring points through a distance function. The missing values can be estimated using completed values of neighboring observations. It is a popular technique as it is intuitive. Note that, this method is not able to handle categorical data directly, meaning that data transformation is needed. To use this method, a reference dataset is needed. This is why the kNN algorithm falls in the supervised approach area of machine learning [33]. This method does not need any complex parameter tuning. Although, the number of neighbors for k needs to be determined [32]. A method introduced in recent years uses the Random Forest algorithm for imputing missing values, named MissForest [34]. MissForest is a another non-parametric supervised machine learning-based data imputation algorithm. This algorithm determines the mean, then for each variable with missing values, the model fits a random forest on the observed part and then predicts the missing part. This method can be used for data that are continuous, discrete, ordinal and categorical, which makes it useful for dealing with all kind of missing data.

Armitage et al. [35], compared imputing methods as median, minimum, and kNN. These methods were compared based on how the imputed data influenced the original distribution of a dataset from which random values were removed. In this research, the kNN algorithm was determined as the optimal method for impute missing data [35]. In a different investigation performed by Penone et al. [36] the MissForest technique was compared to the kNN algorithm. It was found that MissForest can handle large multivariate datasets without tuning parameters nor assumptions about distributional aspects of the data. This research concluded that the MissForest outperforms the kNN since the MissForest had significantly lower error rates compared to kNN.

3.3 Missing data handling

After analyzing the dataset, it became clear that there are missing values in the dataset. Over the years, the Dutch Insurance company has added variables to the frequency analysis. The product has become more complex, and as a result the dataset has grown. This has resulted in missing data points, mainly in older policy records. With regression analysis, every program's default setting is to get rid of any record when one or more of the variables have missing data. This result in the loss of critical information, which can be influencing the outcomes [37]. Section 3.2 describes that the method to handle these missing data can have a large impact on the interpretation of the data.

In order to handle the missing values, the team at the Dutch Insurance company specialized in the car insurance product, was consulted about the variables in the dataset. Based on expert opinion, we evaluated whether a value was MAR, MCAR or left out with a specific reason. However, it was not always possible to identify for each variable correctly what the origin was of the missing values. When dealing with variables which were left out intentionally we assigned an own class label, e.g. Unknown, to the missing values.

We used an imputation algorithm for the remaining variables. In Section 3.2, several imputation methods are discussed, along with their benefits and drawbacks. Because to the dataset's imbalanced character, the distribution of the data is essential in predicting samples belonging to a minority. Furthermore, only a few samples are available from the minority class. Based on these two reasons as well as the outcomes of the different studies outlined in Section 3.2, we do not consider the List-Wise deletion and mean imputation to be able to accurately handle the missing values in the dataset. Furthermore, compared to other imputations methods, the MissForest (MF) algorithm performs well and can easily be used for data that are continuous or categorical (the kNN imputation method is not able to handle categorical data directly). Besides, this method can be applied for classification and regression purposes. Therefore, we selected the MF as method to impute the missing values.

3.3.1 MissForest

We followed the MF algorithm proposed by Stekhoven & Buhlmann [34] in order to impute the missing values. This algorithm predicts the missing values directly by using a Random Forest trained on the observed parts of the dataset. By using mean imputation for numerical missing values and the mode for categorical missing values, the algorithm begins with an initial approximation for the missing values. Then, the algorithm order the variables based on the amount of missing values by ascending number of missing values. After this, the algorithm train a random forest classifier on the prediction classifier using response variables and predictors. The dataset is divided into two different parts based on whether the variable is observed or missing the original dataset. The observed part is used as the training set, and the missing part is used as the prediction set. Based on the trained RF the algorithm predict the missing values and impute these for each variable. The algorithm is running iteratively, continuously updating the dataset variable-wise. When all the variables with missing values have been imputed then one imputation iteration is completed. After each iteration the difference between the previous and the new imputed dataset is measured for the numerical and categorical variables. The algorithm keep iterating until the stopping criterion v is reached. The stopping criterion is met when the difference between the latest and prior imputed data matrix is at least as great as the the previous iteration measured for both variables types, if present. The stopping criterion v is defined by:

$$v = \frac{\sum_{s\in H}(y_{new}^{imp} - y_{old}^{imp})^2}{\sum_{s\in H}(y_{new}^{imp})^2} + \frac{\sum_{s\in C}\sum_{i=1}^{b} I_{y_{new}^{imp} \neq y_{old}^{imp}}}{\#NA} \tag{3.3.1}$$

where s is the collection of variables and $i = 1, ..., b$ observations. For a set numerical variables H we look at the first part of the equation, the difference between the newly imputed values y_{new}^{imp} and values of the previous iteration y_{old}^{imp} is divided by the sum of y_{new}^{imp}. For the set of categorical variables C (the second part of the equation), divides the number of changes I made in one iteration by the number of missing values $\#NA$. In both situations, great performance yields a value close to 0, whereas poor performance yields a value of about 1. As explained, the imputation process is stopped as soon as both differences have become at least as great as the previous iteration. In this case the last iteration is generally less accurate than the previous one. The Normalized Root Mean Squared Error (NRMSE)is used to evaluate performance after missing values are imputed. For the numerical values the PFC is used [34]. For a more in-depth explanation of the MissForest method, we refer to the description of the pseudo algorithm in [34].

3.4 Validation of MissForest

To implement the algorithm, the MissForest package in R is used. However, due to the size of the dataset, it is impossible to directly implement the method. To overcome this problem, MissForest is applied on the the different types of variables separately (car, person and geographical location), given that the provided information for a particular person is not directly related to the characteristics of their car.

Since we do not have the reference dataset, it is not possible to know what the true values are of the missing values. Therefore, is not possible to determine the predictive error of the MissForest algorithm'. In order to circumvent this problem we randomly delete 1300 values from a specific column (the catalog value of the car). By running the MF algorithm, the deleted values can be predicted from the existing values. Afterwards, we can measure the performance of the algorithm since we know the true values of the deleted values.

In Figure 3.4 and Table 3.3 represents the results of the MissForest algorithm applied to the synthetic dataset. Figure 3.4 shows the the NRMSE and the stopping criteria of the simulated test. We can see that after the first iteration, the stopping criterion decrease to almost zero and does not change so significantly afterwards, indicating that there is not much change in the imputed values after this iteration. The stopping criterion v keeps decreasing until iteration three. The algorithm is stopped after the fourth iteration since v is lower for the third iteration compared to the fourth iteration. Furthermore, we plotted the NRMSE of the imputed (predicted) catalog values of the cars and the true values. The NRMSE decrease after the third iteration. This indicates that the fourth iteration is probably a less accurate imputation than the third. Therefore, the imputed values from the third iteration were used.

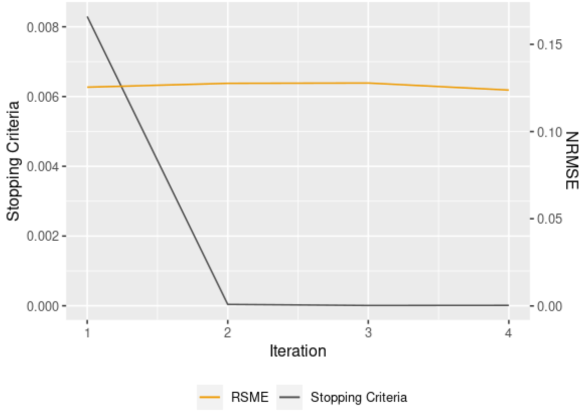

Figure 3.4: NRMSE and stopping criterion v of the MissForest algorithm.

Table 3.3 shows the results of the predicted values compared with true values of the deleted values. It shows that most of the the predictions are very accurate. However, some of the records, for instance record 1296, the predicted value may not be accurate. Furthermore, we observed that this method keeps the original distribution intact. Based on the outcomes of this simulated test, we conclude that we can use this method for handling the missing values.

Index	True value	Predicted value
1	91390	91399
2	44780	44533
3	54395	54755
4	28990	28573
...	...	...
1296	73331	52882
1297	15040	16035
1298	34781	34522
1299	12110	13248

Table 3.3: Predicted catalogue value of car by the MissForest algorithm.

Chapter 4

Regression models

In this chapter we describe the theory that lays the groundwork for the choice of regression models used to determine premiums based on risk within non-life insurance. Regression models are models commonly used to predict the pure premium by modeling the expected claim frequency and claim severity separately. We introduce the standard linear model (LM), generalized linear model (GLM), and generalized additive model (GAM).

A linear model (LM) is the simplest form of linear regression, while a general linear model (GLM) is a more flexible generalization of the LM. When using these linear models, the underlying goal is to investigate the relationship between a random variable (also known as the response variable) and one or more predictor variables (known as covariates). The covariates are fundamentally modeled in the format of a linear predictor. The main reasons for using linear models to solve regression problems is the easy interpretability and reasonable computational speed [34]. Although these models are very efficient in handling categorical covariates, they also have limitations. In fact, the models are unable to detect nonlinear effects for numerical variables directly. However, many nonlinear specifications can be converted to linear form by performing transformations on the variables in the model. Furthermore, it is inefficient to take many values into account separately. The presence of many risk levels results in long computation time. Additionally, since each of these risk levels likely contains very few policyholders, it leads to statistically insignificant results. Therefore, it is important to consider alternatives suitable for including numerical covariates in the modeling. Here, we specifically focus on the geographical area in which a policyholder resides, interpreted as a discreet variables with 454.267 different postal codes. As described in literature, a general additive model (GAM) is one of the suitable alternatives that can be used to model possible nonlinear effects by using smooth functions in order to represent the effects of numerical variables, and thus, more flexible relations can be made [38].

To explore the concepts and differences of each regression technique in determining the premium of a car insurance product, in this chapter the LM, GLM and GAM are analyzed. To this end, in Section 4.1, we review the theoretical background of the LM as it is important to know the basics of this model in order to understand the models used in this sequel. In Section 4.2, the concept of the GLM is covered, including an explanation on how the coefficients are estimated. Furthermore, some limitations of this model are discussed. Section 4.3 discusses the concepts of the GAM with the application of GAM in car insurance. The same section briefly explains the smooth splines and also notes some shortcomings of the GAM.

4.1 Linear Model

Regression analysis has an important role in statistical modeling, and the overall goal is to examine the effect of one or more explanatory variables on a response variable. The model describes the linear relationship between the response variable y and the covariates X. Given a dataset $(y_i, x_{i1}, ..., x_{ip})$ for $i = 1, ..., b$ statistical units. A LM takes on the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + .. + \beta_p x_{ip} + \epsilon_i \tag{4.1.1}$$

where, x_{ij}, are the covariates for statistical unit i and β_j constitute the (partial) regression coefficients for $j = 0, ..., p$. The coefficient β_1 indicates how much y_i increases when x_{i1} increases by one unit, taking into account (adjusted for) the influence of the other x covariates by keeping them constant. Thus, in this way, we can analyze the impact of x_{i1} on y_i independently of the influence of the other x covariates. The ϵ_i stands for the error term. The error term is used to explain the variability in y_i that cannot be explained by the linear relationship between the covariates x_ij and the response variable y_i. If ϵ_i were not present, it could mean that knowing x would provide enough information to determine the value of y. These b statistical units are stacked together and written in matrix form as,

$$y = X\beta + \epsilon. \tag{4.1.2}$$

where X is an $b \cdot p$ matrix of regressors and β is a $p \cdot 1$ vector of coefficients. The models estimators $\hat{\beta}$ are computed from a sample to estimate the population parameters β in a linear regression model. The linear predictor $X\hat{\beta}$ is the linear combination of estimated β. The optimal coefficients $\hat{\beta}$ weights minimizing the squared error values, obtained non-iteratively by $\hat{\beta} = (X'X)^{-1}X'y$ where X' denotes the transpose of X. One of the assumptions of linear regression is that the errors ϵ are independent and normally distributed with mean zero, conditional on the covariates. Based on $E(\epsilon) = 0$, it is given that $E(y) = X\beta$. Furthermore, the linear regression model assumes that the vector of response variable y is normally distributed conditioned on the predictors. To overcome this problem would be to identify a method that normalizes the data so that the LM can be used [39].

4.2 The current industry-standard GLM

Despite transformation of the data, the LM provides limited flexibility. In cases where more flexibility is needed, it could be desirable to use a method that has been further developed to enable modeling of non-normally distributed data and more complex relationships between covariates. Nelder & Wedderburn [40] extended the linear regression model to the generalized linear model by replacing the normal distribution of the response variable with a distribution within the exponential family (e.g., the Gamma, Binomial, and Poisson distributions). The GLM is a widely used regression model to examine the relationship between the response variable and covariates within the insurance industry and has been studied thoroughly [40].

We aim to model the claim frequency of a set policyholders i for $i = 1, ..., b$, which is defined as the number of claims per unit of exposure. The claim frequency is commonly modeled by a GLM with a log-link function. A log-link function converts the expected value of the response value to a linear predictor, i.e.

$$E(Y) = \lambda = g^{-1}(n), \tag{4.2.1}$$

where Y is the stochastic dependent variables, with mean λ or stated differently, the expected value $E(Y)$. Note that the expected value of Y, λ, is linked to the linear predictor n by the link function $g(\cdot)$. For policyholder i, the n_i is denoted as the linear predictor $x_i'\beta$, a linear combination of unknown parameters β, where x_i' is a p-dimensional transposed vector (i.e., $n_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$).

We examine a Poisson distribution for a vector of count data. There is only one parameter to be considered, λ_i, since the mean and variance are equal. Since, $n_i = g(\lambda_i) = x_i'\beta$ and for the Poisson, the link function $g(\cdot)$, is the natural log. Therefore relates the log of λ_i to the linear predictor. For the log function $g(\lambda_i) = ln(\lambda_i)$, we have

$$\begin{aligned} \lambda_i &= exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \qquad (4.2.2) \\ &= exp(\beta_0) \cdot exp(\beta_1 x_{i1}) \cdot ... \cdot exp(\beta_p x_{ip}) \qquad (4.2.3) \end{aligned}$$

where x_{ij} are the covariates of policyholder i and β_j constitutes the (partial) regression coefficient for $j = 0, ..., p$. Hence, the log-link function assumes that the effects of the covariates are multiplicative and the GLM then estimates natural logarithms of multiplicative effects rather than additive effects as in a regular linear model.

When studying the probability distributions for claim frequency, several possibilities are available. One possibility might be to calculate the claim frequencies from the data. Since the claim frequency is defined as the number of claims per unit of time, we can simply divide the number of claims for a given year by the corresponding exposure. However, the problem here is that the domain of the Poisson distribution includes only discreet numbers and the claim frequency has a continuous distribution.

As a solution to this, the claim count is more frequently used and usually modeled using non-negative discrete probability distributions, since the number of claims is discrete and non-zero [41]. The Poisson distribution is widely used for the claim count, and also used in the Dutch insurance company as the standard distribution for claim counts. Therefore, we make use of the Poisson distribution when modeling the claim count.

However, when modeling claim count, the problem is that the expected value of the number of claims increases with exposure, but that the exposure is not accounted for within the claim count. For instance, two different records with different lengths of coverage (for instance, 2 months and 1 year) can have the same claim count. Because the predicted number of claims is related to the period of coverage, insurance companies' should not model these two policies' loss experiences identically. As a result, it is important to predict claim counts while accounting for total exposure. Therefore, we have to weigh the observations by their corresponding exposure. One way to do this is to use an offset ξ equal to the logarithm of the exposure t. Adding this offset to the linear predictor n_i. This gives the following formula,

$$n_i = x_i'\beta + \xi_i = \beta_0 + x_i'\beta + ln(t_i), \tag{4.2.4}$$

For the expectation of y_i, this results in,

$$E(y_i) = \lambda_i = g^{-1}(x_i'\beta + ln(t_i)) = exp(x_i'\beta)t_i. \tag{4.2.5}$$

That is, by adding an offset term to the linear predictor n_i, we weigh the expected claim counts by the corresponding exposure, which results in estimating each claim count for the same unit of time.

4.2.1 Estimation of the coefficient

In this section we want to estimate the coefficients β. We can do this using maximum likelihood estimation. This method seeks to find the combination of coefficients that produce the observed data with the highest probability, given the assumed model form. The likelihood is defined as the product of probabilities of observing each claim count. The Poisson distribution has a density function as given:

$$P(Y = y_i|x_i) = \frac{\lambda_i^{y_i} exp(-\lambda_i)}{y_i!}, \quad \text{for} \quad y_i = 1, 2, 3, ... \tag{4.2.6}$$

where parameter λ_i denotes the event rate and y_i the number of events over a given time for $i = 1, ..., b$ statistical units. For a certain single unit of time, both the mean and the variance are equal to $E[y_i] = Var[y_i] = \lambda_i$. The likelihood function is then similar to the joint density function of each observation, i.e. the product of individual densities. However, the joint density is a function of the data given the parameters, while for the likelihood it is the other way around. For the Poisson likelihood, we get

$$L(\beta) = \prod_{i=1}^{b} \frac{\lambda_i^{y_i} exp(-\lambda_i)}{y_i!} \tag{4.2.7}$$

It is convenient to take the natural logarithm of the likelihood and since a log-transformation does not change the optimum, we can do it without influencing the maximum likelihood we are interested in. Taking logarithms, we get the following log-likelihood $L = ln(L)$:

$$L(\beta) = \sum_{i=1}^{b} (y_i ln(\lambda_i) - \lambda_i - ln(y_i!)). \tag{4.2.8}$$

With the log-link function $\lambda_i = exp(x_i'\beta)t_i$. we can estimate the β by maximizing the following log-likelihood function,

$$L(\beta) = \sum_{i=1}^{b}(y_i(x_i'\beta + ln(t_i)) - exp(x_i'\beta)t_i - log(y_i!)). \tag{4.2.9}$$

Finding β such that $L(\beta)$ is maximized is equivalent to finding the solution $\hat{\beta}$ by taking the derivatives with respect to β and equate them to zero. We get the following formula,

$$0 = \frac{\delta L(\beta)}{\delta \beta} = \sum_{i=1}^{b} x_i(y_i - exp(x_i'\beta)t_i) \tag{4.2.10}$$

which can be solved numerically to get the optimum using the Newton-Raphson method to find the estimates for each coefficient β. The Newton-Raphson method is a root-finding algorithm able to approximate the roots of a real-valued function such as a Equation 4.2.9 [42]. For details about the Newton-Raphson method, see Appendix A.

4.2.2 Different type of risk factors

The independent variables can either be numerical (e.g. age or the catalog value of the car) or categorical (e.g. car brand). Before modeling, it is necessary to consider how the independent explanatory variables should enter the model. In the case of categorical variables, several factor levels can be clustered together in order to achieve a more parsimonious model. In some cases, especially in levels with low data volume, unstable parameter estimates or similar characteristics, levels can be combined to achieve a more parsimonious model. For example, the variable car brand has many levels, the majority of which are unknown brands with very few observations, and therefore unreliable results. These brands can be combined into the factor 'other', containing all infrequently used brands which appear to be similar in terms of claim frequency. Additionally, similar levels can be grouped.

Despite the fact that the GLM model is a frequently-used regression model in the industry of non-life insurance, this model has a major drawback. Many researchers, including Denuit & Lang [43] and Klein et al. [20], argue that the main problem arising from the use of GLM is the fact that the explanatory variables are modeled as linear predictors. For categorical explanatory variables, this is not a problem because they are modeled by binary variables, but GLMs are too restrictive when there are nonlinear effects in the explanatory variables. An example of a nonlinear effect would be the effect of the variable age on claim frequency, as some age groups (such as very young or very old drivers) are expected to make more damage than others. Adding such variables into a GLM will most likely lead to incorrect assumptions that are taken based on such a model. More practically speaking, in the case of car insurances, this would result in unfair premiums given the policyholder's true risk factors. These numerical variables can be incorporated into the GLM if they are suitably transformed where their effect on the response variables is represented as truthfully as possible. According to Denuit, the problem here is that it is not always clear how the numerical variables are to be transformed and, in addition, the question arises as to what extent the transformation gives a true estimate [43].

In practice, researchers often solve this problem by applying polynomials to approximate the nonlinear effects as closely as possible. Polynomial regression is a regression algorithm that fits a nonlinear relationship to the data. Nonlinear in the way that the explanatory variables are nonlinear correlated with conditional mean of the response variable. However, the equation $f(x, \beta)$ itself is still linear since for linear regression the linearity constraint only applies to the coefficients β. For details about the polynomial regression method, see Castro [44]. For small nonlinear effects this is a good solution, but for large nonlinear effects this approach is not sufficient. According to Klein et al. [20], finding the right degree of the polynomial can be challenging, as low-order polynomials are often not flexible enough to approximate the variability in the data, and polynomials of a large order have the problem of providing unstable estimates, especially for extreme values of the explanatory variables. Additionally, since the method of using polynomials is easily affected by outliers, the results can easily be skewed by the presence of extremely high or low values. In view of all this, there may be cases where polynomial regressions are appropriate, but the issue at hand is that the nonlinear relationships in question here might not be as easily specified.

Alternatively, rather than treating the variable as a numerical variable in the GLM, it can be treated as a categorical variable that has a high number of levels, where for example, each level represents a single age. The downside of this method is that it would lead to many different levels where, for example, not all ages would be made up of a similar amount of policyholders, resulting in statistically insignificant results and possible over-fitting. To overcome this problem, one could cluster numerical variables like age into a categorical variables. For example, the variable age contains many levels; one for each discrete number of years present in the data. As expected, two adjacent years are often not significantly different from each other. Therefore, age classes containing consecutive years can be formed, which then get significantly different effects. An algorithm to automatically implement this is given in Appendix B. However, as mentioned in the problem description, a limitation for this method would be that the effect on the claim frequency for two policyholders with a different (but close) variable values may have substantially different premiums to pay, when these values fall within different clusters. Additionally, this method results in a loss of information due to the fact that the numerical variable is represented by a step-wise constant function. Insurers, however, often prefer the use of a model containing categorical variables, as those are easier to interpret. Here, the choice of the number of clusters is essential and should be balance between choosing a number large enough to accurately reflect the effect of the numerical variable, and choosing a number small enough to allow for enough observations within a cluster in order to obtain statistically significant results.

4.3 Generalized Additive Models

Generalized Additive Models (GAM), introduced by Hastie & Tibshirani [45], are GLMs that accommodate for nonlinear effects of numerical covariates in the linear predictor. These models are well-suited for regression modeling in actuarial and financial applications due to their flexibility in handling different types of risk factors, as the effects of numerical risk factors can be smoothly incorporated. Because of this, GAMs are more statistically flexible than GLMs. Despite this advantage, GLMs tend to be the preferred model by insurance companies due to their simplicity, since premium models in the insurance industry should be easily interpreted and programmed, tunable to marketing needs, as well as intuitive and easily explainable to stakeholders. Additionally, due to their wide-spread use, they can be easily compared in benchmark studies other insurance companies.

4.3.1 The model

For the GAM, consider a sample of the response variable y belonging to the exponential distribution family $y \sim ExpoFam(\lambda, \cdot)$, where y_i, $i = 1, ..., b$, are assumed to be independent, and the predictor $E(y) = \lambda$, which relates the expected value of y, to the covariates through a link function $g(\cdot)$. This is given by,

$$g(\lambda_i) = \underbrace{\beta_0 + \beta_1 x_{i1}}_{parametric} + \underbrace{f_1(x_{i2}) + f_2(x_{i3}x_{i4})}_{non-parametric}. \tag{4.3.1}$$

The first, parametric part is known from the GLM. The second, non-parametric part defines two functions. The univariate function f_1 takes x_2 as numerical predictor, and bivariate function f_2 expresses interactions and spatial effects between two numerical predictors, x_3 and x_4, as argument. Rather than polynomial functions, the GAM use smoothing splines for functions such as f_1 and f_2, this will allow for nonlinear relationships between the covariates and the target variable y.

4.3.2 Smoothing splines

Smoothing splines provide an excellent estimation and inference for the smooth functions. A spline is a piece-wise polynomial constrained to match at certain points (knots) and are considered as a non-parametric fit type. Similar to polynomials, splines have an order. The order of a spline is equal to the maximum order of the piece-wise polynomials it consists of.

For univariate numerical variables, smoothing splines are applied and are constructed from cubic polynomials. For our purpose, smoothing splines are often sufficient to model most nonlinear effects. A smoothing spline is more flexible and, in contrast to linear splines, is able to fit curved functions, instead of having straight lines, by forcing the first and second order derivatives of the function to agree at the knots (see Figure 4.1).

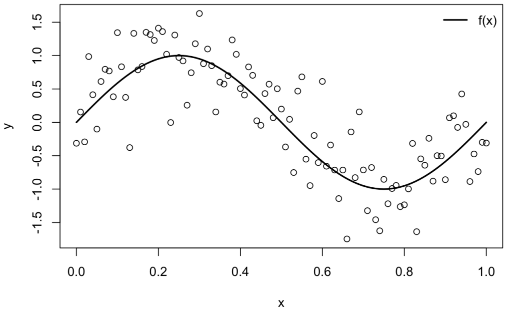

Figure 4.1: Example of modeling cubic splines.

For the modeling of multi-dimensional relations, such as the latitude and longitude within this research (i.e., the combined effects of numerical variables), thin-plate splines are used. Thin-plate smoothing splines are based on cubic smoothing splines and similarly to a cubic smoothing spline, a thin-plate smoothing spline aims to model the data as closely as possible, but taking into account two (or more) predictor variables. So, instead of a single curve, thin plate splines are defined as a smoothed surface in a two- or multi-dimensional space (Figure 4.2).

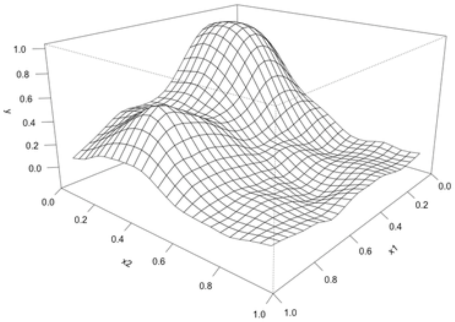

Figure 4.2: Example of modeling thin plate Splines [46].

To define the smoothness of the splines, smoothing parameters (the number and location of the knots) need to determined. Smoothing splines has the property of circumventing the knot selecting by considering all the inputs as knots. To estimate the coefficients $\hat{\beta}$ within smoothing splines, the aim is to minimize the penalized least squares estimation function, which is modified by adding a roughness penalty which penalizes it for roughness (wiggliness) and high variance. The penalized least squares estimation function is defined as,

$$\hat{\beta} = \sum_{i=1}^{b} (y_i - f(x_i))^2 + \delta f"(x_i)^2 dx, \tag{4.3.2}$$

where x_i are defined as smoothing covariates, and y_i is the dependent variable. Assuming that the relation between x_i and y_i is unknown. where $f"$ is the penalty on the roughness of f and is defined, in most cases, as the integral of the square of the second derivative of f. Furthermore, δ is defined as the smoothing parameter, which controls the trade-off between the fit and smoothness. The first therm measures the goodness of fit. The second term of the equation is the wiggliness penalty which is added to avoid over-fitting. For the thin-plate smoothing spline, the penalized least squares estimation function is adapted in order to handle the bivariate coefficients [46]. A larger value of the smoothing parameter results in more smoothing or regularization, which constrains the model more and reduces the effective degrees of freedom

The generalized cross validation (GCV) criterion can be used in order to determine the δ. The optimal δ is estimated in order to get the right fit. When the δ is not chosen right this could result in overfitting or underfitting. This GCV method determine the δ by minimizing the following formula,

$$\delta = b^{-1} \frac{(\sum_{i=1}^{b} (y_i - \hat{y}_i)}{(1 - b^{-1} EDF)^2} \tag{4.3.3}$$

where b is the number of data points, y_i is the i-th response value, and $\hat{y}_i$ is the i-th fitted value. The numerator represents the residual deviance and $i = 1, ..., b$ stands for the number of observations. Futhermore, the EDF is the effective degrees of freedom. The number of data points in the model can also affect the effective degrees of freedom. Other factors that can influence the effective degrees of freedom include the number and complexity of the predictors in the model, the amount of noise in the data, and the nature of the relationship between the predictors and the response. For a more in-depth overview of splines, see for example of the smoothing spline Harrell [47] and for the thin-plate spline Wood [46].

4.3.3 Differences between GAM and the current industry-standard GLM

As we have described here, GAMs, using these smooth functions, are able to represent the possible nonlinear effect of a numerical variable. This is contrary to GLMs, which make use of a parametric method such as polynomial regression. Compared to the use of numerical variables in these linear models, GAMs are much more flexible as well as completely automatic, since it's not necessary to pre-define a certain function. The major downside of using a GAM is that the estimates are more difficult to interpret compared to a GLM.

Furthermore, the GAM model is a flexible alternative to the standard regression model, since it can not only use distributions from the exponential family and it additionally estimates each parameter with its own predictor and corresponding link function. However, the problem is that a GAM model requires the estimation of several variables. As a result, the GAM model a high degree of flexibility, but at the same time it means that there must be efficient strategies, such as the GCV as described in the previous section, to avoid overestimating the data and to produce models that contain only the most relevant covariates for each distribution parameter.

Chapter 5

Clustering methods

In chapter 4, we discussed the foundation of different regression models for predicting the claim frequency and severity in order to obtain the pure premium. Additionally, we highlighted the possible difficulties when using numerical variables and investigated the ways of implementing these kind of variables in to the models. Two strategies were discovered as feasible ways to include these variables in the models. The first method is the current industry standard and is based on the GLM, to cluster the continuous and spatial variables in order to get categorical variables which can be implemented into the GLM. The second method is using a GAM and provides a flexible way in order to model the numerical variables with nonlinear effects. As discussed in section 1.2, the first method is often preferred by the insurance companies. However, the downside of this method is that there is a loss of information, both within clusters and between clusters. The manner in which the clusters are determined impacts whether an effect can be observed. Therefore, it essential to find the most optimal clustering algorithm to cluster the numerical variables in such way that there is a balance between a cluster that is small enough so that the cluster accurately represents all the values contained in it, but at the same time large enough so that there are enough data points so that the clustered group is statistically significant.

In this chapter, we elaborate on the strategy we use to cluster our spatial variable into categorical levels, in order to implement the spatial variables in a parametric model. In Section 5.1 we discuss the approach for transforming numerical variables on which our strategy is based. Furthermore, in Section 5.2 we discuss different clustering algorithms. In Section 5.3 we discuss the methodology to compare the different algorithms. In Section 5.4 we discuss how the suitable number of clusters can be determined.

5.1 An alternative approach: numerical variables and the GLM

Clijsters [38] introduced a methodology to deal with numerical variables in a GLM. She suggests starting by implementing the numerical variable in a GAM, which allows us to model the numerical variable in semi-parametrical way. Based on the outcomes of the GAM, it becomes possible to investigate the smoothing effect of the variable as well as the potential non-linear effect. Following that, based on a clustering algorithm, the variable can be clustered. As a result of this approach, the categorical variable can easily be implemented in a parametric model such as the GLM.

Based on the proposed method of Clijsters [38], here, we model the spatial variable using a GAM in order to obtain the effect of each postal code on the claim frequency. We implement the spatial variable in the model by using the smooth bivariate effect of the center of the postal codes (the center can be obtained by using longitude and latitude). Additionally, we include all the other variables. The outcomes of the GAM reflect the effect of the spatial variable and the other variables. The mgcv R-package is used to model the GAM with the smooth bivariate effect [46] [48]. The objective of this approach is to group postal codes with similar risk levels together. In chapter 6, the results of this approach will be shown.

5.2 Clustering methods

This section elaborates on the different clustering algorithms we use in order to convert the spatial variable into distinct categorical clusters. For this, it is important that the categorical clusters appropriately reflect the effect of the spatial variable by defining risk as homogeneous groupings of the variable. A well-suited R package for 1-dimensional clustering methods is the ClassInt package, as introduced by Chi [49], and comprises commonly used methods for clustering different postcodes. Other clustering methods, such as k-means clustering, are popular clustering methods but are aimed at clustering multidimensional datasets, which is why the clustering methods in the ClassInt package are better suited for clustering the spatial variable.

1. **Quantile clustering**
   The goal of quantile clustering is to allocate the same number of observations (in this case, postal codes) to each cluster. As a result, each cluster should have, $postalcodes/k$, postal codes, where k is the predefined number of clusters. This algorithm is frequently used as the standard in statistics when the values of a variable follows a normal distribution. However, it can produce extremely misleading clusters, since it is possible that when many observations have the same values, they are allocated to different clusters to ensure that each cluster has the same number of observations.

2. **Equal-width intervals**
   Equal-width interval clustering is the most basic approach for discretizing data and is frequently used to transform numerical values into categorical ones [50]. It involves sorting the observed dataset of a numerical variable and dividing the variable's range of observed values into k equally sized clusters. This algorithm produces decent results with uniformly distributed data.

3. **Jenks natural breaks**
   The Jenks natural breaks optimization algorithm, also known as the Jenks natural breaks classification technique. This method is commonly used in thematic maps, especially choropleth maps. Choropleth maps provide an easy way to visualize how a variable varies across a geographic area. It is a data clustering approach meant to categorize near values into groups with the maximum distance to other groups based on natural breaks between the groups of close values. This is accomplished by iteratively attempting to locate the best feasible classes of near values in order to decrease each class's average deviation from the class mean while

maximizing each class's deviation from the mean of the other classes. In other words, the technique aims to minimize variance within classes while increasing variance across classes [51].

5.3 Selection of clustering method

There are several ways to compare allocations produced by a clustering algorithm. The main distinction between these ways are the so-called internal and external performance measures. An external performance measure uses more inputs than just the dataset and an allocation of the dataset into clusters. In empirical applications, external performance measures are not always useful because the additional information required by the performance measure is not always available, but in simulations, they can be used to compare methods [52]. Internal performance measures are statistics where only the allocation to be evaluated and the dataset itself are assessed. A good allocation should generally satisfy two things. First, it should minimize the variance within classes, meaning that the observations placed in the same cluster must be similar to each other. Secondly, it should increase the variance between classes, meaning that observations not placed in the same cluster must not resemble each other. In this way, there is a clear distinction between the clusters formed. These two requirements are called compactness and separation, respectively [53]. Since in this research, no knowledge is assumed outside the dataset and the allocation, only internal performance measures are considered.

Because of the combination complexity of the compactness and separation requirements, Jenks & Caspall [54] assert that a complete enumeration of the solution space is infeasible and suggest ways to search for a good solution using two error measures, the goodness of variance fit (GVF) and the tabular accuracy index (TAI). The TAI and GVF approaches are widely documented in several cartography textbooks in order to cluster geographical data [55]. These error measures do not describe the map itself, but a distributive set of the data to be represented on the map. The ClassInt R-package used in this research returns values of these two indices for assessing class intervals. The formulas of the GVF and TAI by Jenks & Caspall [54] follows by:

$$GVF = 1 - \frac{\sum_{j=1}^{p}\sum_{i=1}^{b_j}(z_{ij}-\hat{z}_j)^2}{\sum_{i=1}^{b}(z_i-\hat{z})^2}, \qquad (5.3.1)$$

$$TAI = 1 - \frac{\sum_{j=1}^{p}\sum_{i=1}^{b_j}|z_{ij}-\hat{z}_j|}{\sum_{i=1}^{b}|z_i-\hat{z}|} \qquad (5.3.2)$$

Where, the z_i, $i = 1, ..., b$ are the observed values, $j = 1, ..., p$ are the different classes with p as the number of classes, $\hat{z}_j$, the mean of of the observed values in class j, and N_j the number of observed values in class j. The numerator of the TAI is the sum of the absolute deviations of the values classified into the classes, and the denominator is the sum of the absolute deviations of the entire classified set. The numerator of the GVF is the sum of the square deviations of the values classified into the classes, and the denominator is the sum of the square deviations of the entire classified set. The better the class classification reflects the nature of the data, the higher the value of the indicators ranges from 0 to 1 [56].

We use these two metrics to compare the results of the three clustering methods discussed in Section 5.2. Figure 5.1 compares the performance of the three methods for $n = 10 : 100$ clusters using the GVF and the TAI. The perfect model is achieved when the quantities of both metrics are maximized. The results of both metrics hardly differ between the allocations of the different methods. Notable is that, in both cases, the Jenks natural breaks algorithm scored the best of the different clustering methods, regardless of the performance metric of the the number of clusters.

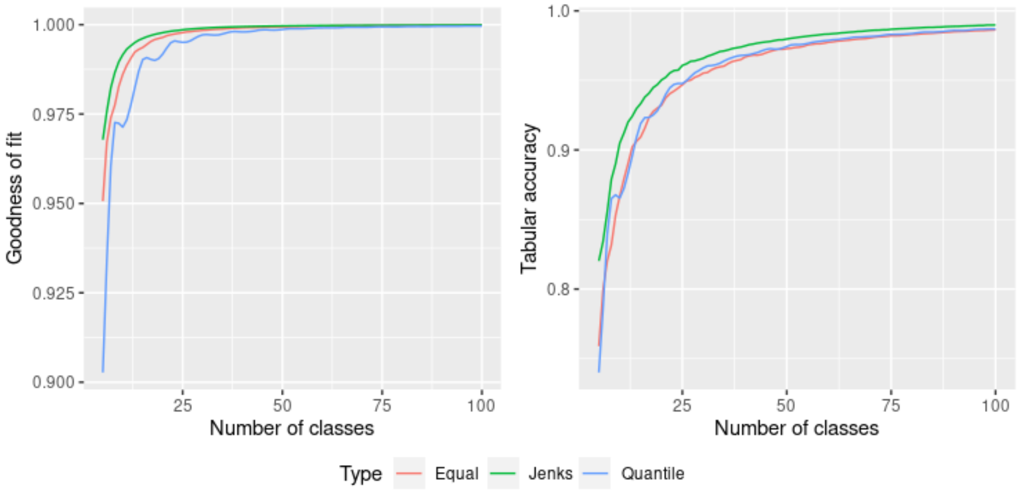

Figure 5.1: The left (right) plot represents the result of the GVF (TAI) per number of clusters.

We compare the results of the three different clustering methods in two additional manners. For this comparison, we set the number of clusters to 10. Figure 5.2 illustrates the cumulative distribution function (CDF) of the effect of this spatial variable on the claim frequency and the intervals of the clusters obtained by the different algorithms. As seen in this figure, the equal interval algorithm clearly divides the range of the spatial effect into ten clusters with equally-sized intervals. This means that, when the cumulative distribution function becomes steeper, more postal codes will be grouped together in a cluster. Within the figure this means that the right clusters contain many more postal codes than the left clusters. Quantile clustering produces very wide clusters in the left ends of the support where data are scarce and divides the postal codes in more equal clusters with respect to cluster size than the equal algorithm. Jenks' algorithm produces the most homogeneous clusters and appears to be a mixture of the equal and quantile algorithms.

Figure 5.3 visualizes the map of the Netherlands with each postal code assigned to one of ten clusters, produced by each of the three clustering methods. In the picture, it can be seen that the groups of colors differ both in color and shape and reflects the effect of the variable postal code on the claim frequency. A coefficient below one is represented in blue and indicates that the claim frequency of the spatial coefficient is lower than the base-level, and a coefficient above one is represented in orange, which indicates that the claim frequency of the spatial coefficient is higher than the base level. Furthermore, it is clear that especially the north of the Netherlands has a negative effect (blue) on the claim frequency compared to the base level and in the south-west of the Netherlands has a positive effect (orange).

Therefore, the three different algorithms result in distinct clusters for the spatial variable, resulting in different postal code groups. In addition, the map with the Jenks algorithm seems to have the most gradual changes between clusters, based on the transition between colors. This seems to be the most natural way in which the risk would spread. Based on the results of the GVF and TAI performance metrics, as well as the visualizations in figures 5.2 and 5.3, the Jenks algorithm is selected as the most appropriate method for clustering the spatial variable for the frequency model.

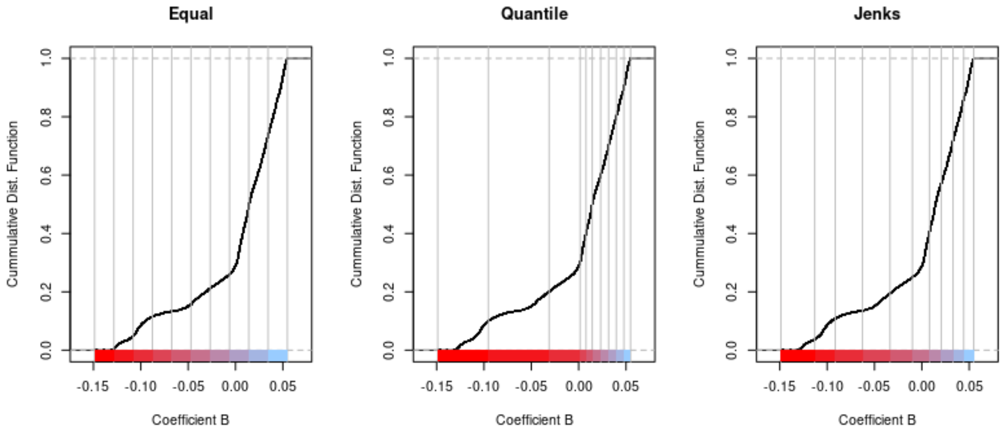

Figure 5.2: CDF of the spatial variable and the ten intervals produced by the three clustering algorithms.

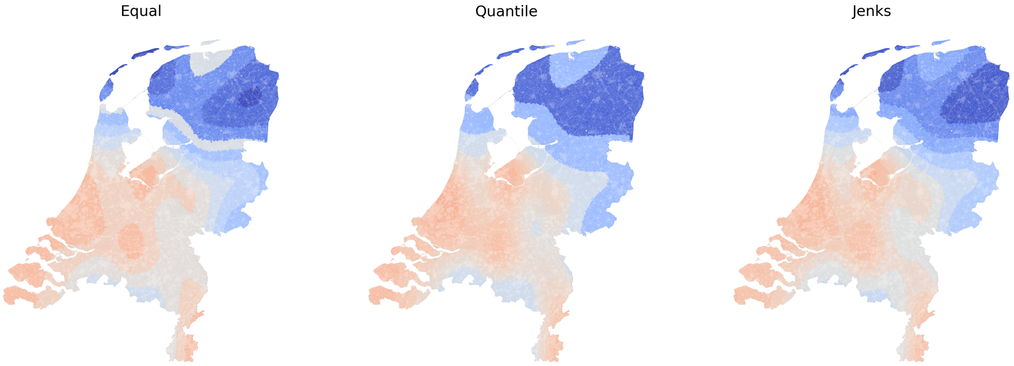

Figure 5.3: Maps of Netherlands with the spatial variable (postal codes) clustered into ten different groups based on the intervals produced by the three clustering algorithms.

5.4 Tuning the number of clusters

In order to select the suitable number of clusters for a discreet variable that would properly reflect the distributional differences across the variable dimensions without having an excessive amount of parameters, multiple methods are described in literature. Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are widely used model selection criteria[57]. These methods can be referred to as a measure of the goodness of fit. The formulas of both methods are represented as,

$$AIC = 2 * K - 2 * ln(L) \tag{5.4.1}$$

$$BIC = ln(K) * -2 * ln(L) \tag{5.4.2}$$

where $ln(L)$ is the log-likelihood of the model and K is the number of model parameters that constitutes the "penalty" on the number of parameters. The AIC is designed to find the model that explains the most variation in the data, while penalizing for models that use an excessive number of parameters[58]. The lower the AIC or BIC, the better the model fit. The difference between AIC and BIC is that the penalty for an additional parameters is higher for the BIC than the AIC, due to the log of K being taken. In other words, the BIC will tend too choose a more simplistic model, compared to the AIC. In order to choose the most optimal model, the results of the AIC and BIC are compared.

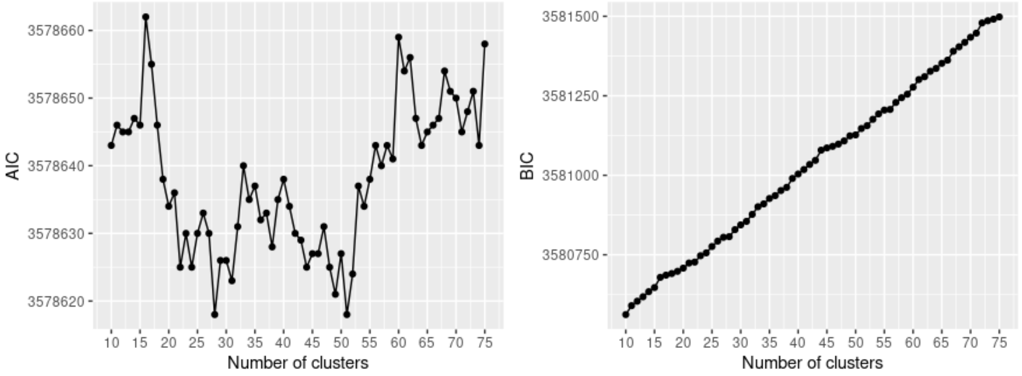

Figure 5.4: AIC and BIC for the GLM including the spatial variable over k number of clusters.

Based on the method described in Section 5.1, we transform the numerical variable into a categorical variable that can be implemented in the GLM in order to reflect the effect of the spatial variable on the claim frequency. We determine the suitable number of clusters by taking into account a large range of potential values for the number of clusters of the spatial variable, e.g. $k \in 10$ until 75, where k defines the number of clusters. We start by applying the Jenks' natural break algorithm to calculate the intervals of the clusters for the spatial effect from the GAM. After this, we use these breaks to transform the numerical variable into a categorical spatial factor variable, and use this categorical variable in combination with the covariates of the current industry-standard model to estimate a new GLM. Based on the outcomes of the GLM model we calculate the AIC and BIC. Finally, we iterate this process for the numbers within the set k. Based on the AIC and BIC, we determine the most optimal number of clusters.

Following these steps, we select the number of clusters for the spatial variable that yields the lowest AIC. Figure 5.4 shows that based on the AIC, the range between 20 and 55 clusters seems to contain the most optimal number of clusters, with 28 and 52 clusters resulting in the lowest AIC. The BIC values of the GLMs with a clustered spatial impact are also presented. However, there is not a very clear-cut conclusion that can be drawn from the figure based on AIC, since the maximum difference of the AIC is ± 40 given the magnitudes. In addition, we would have expected a slightly more smooth line. Furthermore, the figure shows that the BIC might not be a good measure for these data either, due to the fact that it only continues to increase. However, since 28 clusters have a lower BIC than the one with 42 clusters, we choose 28 clusters for transferring the spatial variable into a categorical variable. Appendix C1 shows the estimated parametric components of the GLM including the clustered spatial variable with 28 coefficients (these are left out for confidentiality reasons). Of these spatial coefficients, almost all are statistical significant with a p-value lower than 0.001. However, for few coefficients of the spatial variable the p-value is higher than 0.05.

In this chapter, we proposed a strategy to cluster our spatial variable into categorical levels, in order to implement the spatial variable in a parametric model. Based on the metrics used, we conclude that it is suitable to cluster the spatial effect using the Jenks natural breaks algorithm in 28 clusters. This results in a factor variable that reflects the differences across the variable dimensions without adding an excessive amount of parameters to implement in the GLM. We are continuing our research using a GLM that implement the spatial effect as a factor variable with 28 levels for the claim frequency model. The cluster with the interval $[6.88e-05, 0.00545)$ is chosen as the reference cluster within the GLM because it has the highest exposure.

Chapter 6

Frequency analysis

In this chapter, we describe the results of the claim frequency models, with the response variable being the number of claims. Section 6.1 describes the estimation of claim frequency using the original GLM model. We work with the prepared dataset that contains 23 independent variables about policyholder information available in the portfolio. Subsequently, Section 6.2 shows the results of estimating claim frequency using the GAM model by implementing the 23 independent variables and the spatial variable through the smooth bivariate effect of the center-points of the postal codes. The outcomes of the smooth bivariate effect of the spatial effect are clustered based on the methodology described in the previous chapter. In Section 6.3 we estimate the claim frequency by using a GLM that implements the 23 independent variables and the spatial variable as a categorical variable with 28 levels.

6.1 The current industry-standard GLM

In order to be able to compare the models developed to the GLM which is the current industry-standard model. The GLM is estimated using 23 independent variables (which can be split in policyholder information, car information, and geographical information) selected by the Dutch insurance company. Some variables are categorical, with some levels clustered together, while others are numerical, for which some nonlinear effects are estimated using polynomials. Most variables are individual-specific, e.g. the number of claim-free years and the age of the driver. Additionally, there are car-specific variables, e.g. the car brand and the vehicle age. Lastly, there are a few postal code-specific variables, e.g. the province and the urbanization level. For confidentiality reasons, we cannot report more details about the characteristics of this model. Within the Dutch insurance company, the selection of the variables included in the model is mainly based on the AIC and expert judgment.

For most variables the interpretation is rather straightforward. For example, the categorical variable gender has two levels: ’male’ and ’female’. The male group has the most observations and is therefore chosen as the base category. The GLM has then estimated one coefficient; the one corresponding to female, β_2, compared to the base category. The coefficient corresponding to the male (and others) category is set to zero to prevent perfect multicollinearity. We have a log-link function:

$$\lambda_i = exp(x_i'\beta) \tag{6.1.1}$$

We first have to transform the coefficients before we can interpret them. That is, to observe the effect on the frequency, we take the exponent of the coefficients (Equation 6.1.1). By taking the exponent of 0, this results in a base level 1. For example, the females' claim frequency is $exp(\beta_2) = 1.078$ times as high compared to males (and others), ceteris paribus. Note, this is not the real factor due to competitive sensitivity and figures only as an example

Most of the numerical variables are interpreted similarly. However, for some numerical variables, polynomial regression is included in the model to capture nonlinear effects. For these variables, the interpretation is a little more challenging as multiple coefficients have to be taken into account. One example of such a variable is the policyholder's age. For convenience, let x denote the policyholder age and let β_4, β_5 and β_6 be the coefficients corresponding to the polynomials of degree one, two and three, respectively. Then on average, a given policyholder with age x_i files

$$exp(\beta_4 * x_i + \beta_5 * x_i^2 + \beta_6 * x_i^3) \tag{6.1.2}$$

times as many claims, compared to the base rate (which is equal to 1), while holding all remaining factors equal. Equation 6.1.3 is shown visually in Figure 6.1. Here we see that young policyholders of around 18 years old have a relatively high risk of collision. In general, due to their inexperience on the road, these drivers are more likely to be involved in an accident than older age groups. In comparison to older age groups, insurance firms may often charge a slightly higher price for coverage. The higher premium, in effect, covers this higher risk. As experience increases over time, the risk decreases from 18 years onward until it slightly goes up again with a small local peak at 54 years old. This effect is probably explained by children of the policyholders becoming 18 around the parents' age of 50, and after the children obtain their driver's license, they start driving in their parents' car. An additional cause might be the midlife crisis of a policyholder, although this effect is probably smaller than the previous hypothesis. When we increase age even further and the just explained effect has declined again, the risk keeps increasing with age as expected, because impairment of both sensory and cognitive skills are also known to increase with age for elderly people. Lastly, we observe the risk of claim starts to decline after the age of 88. This can be explained by the decline in mileages of older people and therefore has lest risk on the road.

Appendix C2 shows the estimated parametric components of the GLM[1]. These findings indicate the presence of spatial heterogeneity within the relationship under study. More precisely, we found that if a policyholder lives in all the other provinces of the Netherlands, then the expected claim frequency is lower compared to the expected claim frequency for the base category (province Zuid-Holland and Zeeland, who are grouped together as a coefficient). This frequency is the lowest if a policyholder lives in the North. The variable representing urbanization grade allows us to capture further elements of spatial heterogeneity. The degree of urbanization reflects the size of a place: it

[1] The estimated parametric components of the different models are left out for confidentiality reasons

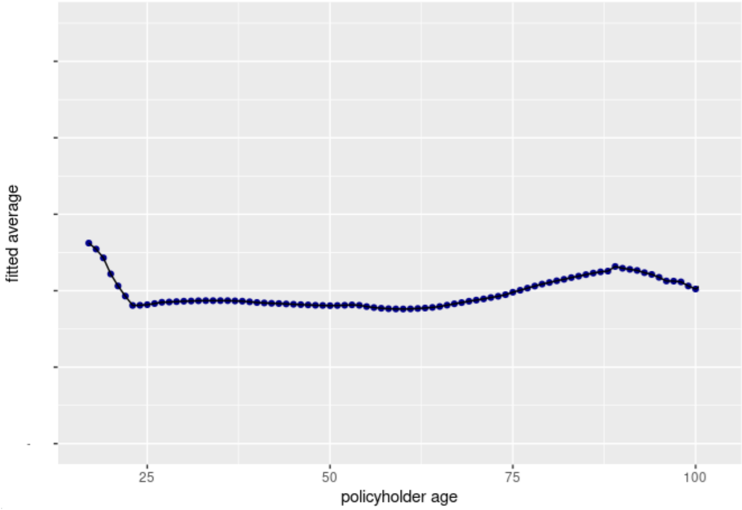

Figure 6.1: Estimated marginal effect of policyholder age on the claim frequency. Note that again some noise is added and the legends are left out for confidentiality reasons.

is a classification based on the number of people living in an area (e.g., urbanization level 1 are big cities like Rotterdam and Amsterdam with more than 250.0000 persons living, and urbanization level 6 are areas with less than 5.000 persons living in it). Within the model, the reference category is urbanization level 4 (20.000 - 50.000 persons). The estimates suggest that the higher the urbanization level, the higher the expected claim frequency of the policyholder. These results are consistent with, which indicate within literature [59].

The total spatial effect of postal codes, based on the GLM, is visualized in Figure 6.2. The geographical effect is based on multiple demographic and social economic variables, such as province and urbanization level. These variables are not provided on an individual basis but instead indicate the averages in a particular postal code. Let z_i denote the policyholder postal code and let β_7, β_8 and β_9 be the coefficients corresponding to the demographic and social economic variables of this postal code. Then on average, a given policyholder with postal code files

$$\beta_7 * \beta_8 * \beta_9 \tag{6.1.3}$$

times as many claims, compared to the base rate.The interval of this total spatial effect ranges from $[0.851698; 1.568864]$. Note that within the current industry-standard model we do not implement the spatial variable we calculated in the previous chapter. Figure 6.2 visualizes the presence of spatial heterogeneity within the relationship under research. As already mentioned, A coefficient below one is represented in blue and indicates that the claim frequency of the spatial coefficient is lower than the base-level, and a coefficient above one is represented in orange, which indicates that the claim frequency of the spatial coefficient is higher than the base level[2]. It is noticeable that the demographic and social-economic variables like urbanization grade and the province influence the

[2] This applies to all figures wherein the map of the Netherlands with the effect on claim frequency is shown

overall geographical effect on the claim frequency. In particular, the urbanization level 1 and 2 (i.e., Rotterdam, Amsterdam, Utrecht and Den Haag) have a strong positive effect (the red areas) on the claim frequency than the base level (the gray areas). Furthermore, it is clear that especially the north of the Netherlands has a strong negative effect (blue) compared with the base level. Noticeably, there are clear jumps in effect on claim frequency at the borders of the provinces Groningen and Friesland. This is consistent with our expectations in the problem statement in Section 1.2. Additionally, the rest of the Netherlands is fairly similar to the base level. In conclusion, the categories urbanization level 1 and the north of the Netherlands play an important role in determining the expected claim frequency.

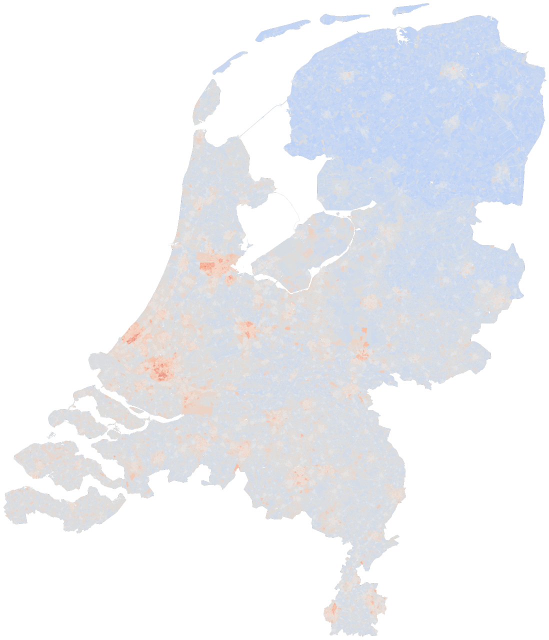

Figure 6.2: Estimated effect of the postal codes on the claim frequency based on the demographic and social economic variables in the current situation (GLM). A coefficient below one is represented in blue (red) and indicates that the claim frequency of the spatial coefficient is lower (higher) than the base-level.

The model explained in this section, a GLM with 23 independent variables (only categorical and polynomial variables), can be considered as the start of this claim frequency analysis and will be named the current industry-standard model. In the next section, we discuss the results obtained by estimating the model using a semi-parametric specification within a GAM framework. We emphasize once more that a GAM allows for the flexible modeling of numerical bivariates without imposing any variable assumptions (e.g., linear or nonlinear). We achieve this by employing thin-plate regression splines, which allow us to more realistically model the impact that the spatial effect have on the expected claim frequency.

6.2 The GAM including the spatial variable

The GAM extends the current industry-standard model by including the effect of the numerical spatial variable captured by the bivariate smooth effects the new spatial variable (the longitude and latitude). For each of the 454.267 postal codes in the Netherlands we have estimated the effect on the expected claim frequency. The GAM has estimated the postal codes compared to one base-level. Since we have a log-link function, by transforming the coefficients as before, we can interpret them as Equation 6.1.2. Therefore, the base level becomes $\beta = 1$. In relative terms, a spatial claim frequency is $exp(\beta_i) = 1.14$ times as high compared to base-level, ceteris paribus.

Figure 6.3 represents the marginal spatial effect of the bivariate spatial variable on postal code level[3]. The interval of this spatial coefficient ranges between $[0.861707; 1.056312]$. This figure indicates the presence of a significant spatial effect on the expected claim frequency on postal code level. The spatial variable exhibits a weak negative impact on expected claim frequency in the North of the Netherlands and a weak positive impact in the South-West.

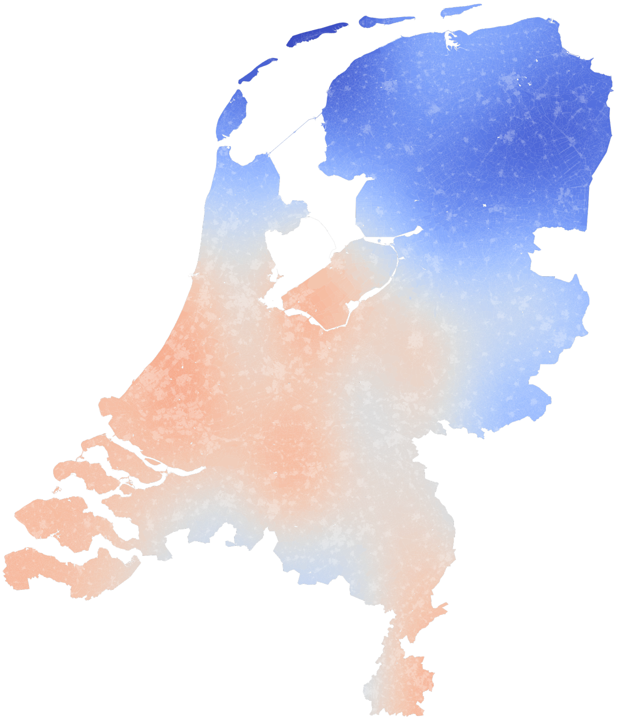

Figure 6.3: Estimated marginal smooth effect of the new bivariate spatial variable on the claim frequency.

[3] Note that, even though the colors in the other figures represent the same effect, the color scale in figure 6.3 represents a different interval and thus the colors are of a different intensity compared to other figures.

Figure 6.4 represents the total spatial effect of each postal code on the expected claim frequency where we combine the new spatial variable with the other demographic variables modeled by the GAM. Left are the results of the total spatial effect modeled by the current industry-standard GLM as discussed and visualized in Section 7.1 (Figure 6.2). On the right side we visualize, together with the variables of demographic variables, the smooth bivariate effect of the new spatial variable based on the outcomes of the GAM model. The interval of the total spatial effect of the GAM model is ranges between $[0.845669; 1.635882]$. We see that, using the bivariate spatial variable in the GAM, the spatial effect changes more gradually across the Netherlands, rather than the hard boundaries that can be observed in the current-industry standard model. This is in line with what was expected, as the model more accurately models the effect of the spatial effect on claim frequency. In more detail, we can see that using the bivariate variable in the GAM model, the overall spatial effect is weaker negative in the north of the Netherlands and a stronger negative effect in some parts of the province (the North of Noord-Holland en the East of Gelderland) compared to the current industry-standard model. Also, there is a stronger positive effect in the south of the Netherlands because more areas are orange. In other words, the current industry-standard GLM is overestimating relative to the new model in the north and underestimating in the south.

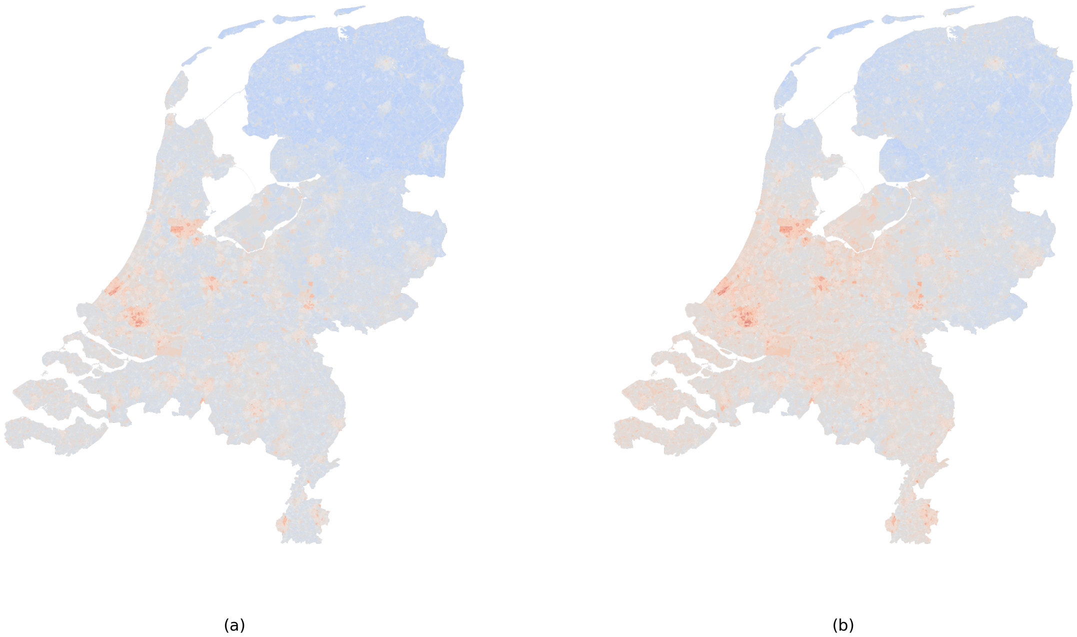

Figure 6.4: (a) The estimated marginal effect of the postal code based on demographic and social economic variables of the current industry-standard model (GLM). (b) The effect based on combined effect of the demographic variables with the smooth bivariate effect of the new spatial variable (GAM).

Furthermore, Appendix C3 show the estimated parametric and non-parametric components, respectively, of the model. Most of the parametric coefficients for GAM are similar to the GLM. However, province (i.e. the residing province of policyholders) is no longer a statistically significant factor (based on the regular statistical tests). This can be explained by the fact that the province variable is a very general effect of thousands of postal codes combined. However, we now have a spatial variable at the postal code level. This variable on postal code level allows the effect to be estimated more accurately and better represents reality, whereby the effect of the province variable is reduced.

6.3 The improved GLM with the clustered spatial variable

In this section, we continue our research using the methodology discussed in Chapter 5 that transforms the numerical bivariates of the postal codes (obtained by the GAM model and visualized in Figure 6.4) as a factor variable with 28 clusters.

By using the Jenks's natural breaks algorithm we cluster the postal codes with a comparable spatial effect on the claim frequency together. We select 28 clusters as the desirable number of clusters to transform the numerical spatial variable, based on the AIC of the GLM by modeling a set of different numbers of clusters. Figure 6.5 illustrates the cumulative distribution function (CDF) of the effect of this spatial variable on the claim frequency, and the vertical lines are the breaks of clusters obtained through the Jenks binning method.

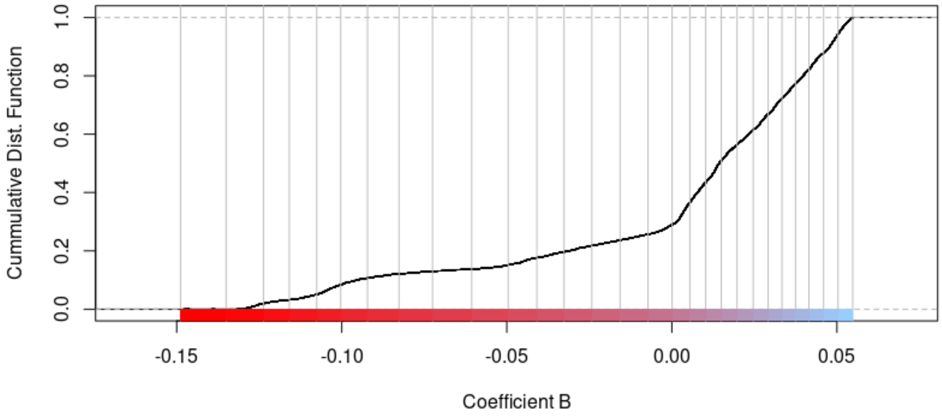

Figure 6.5: The CDF of the spatial variable and the 28 clusters produced by the Jenks' natural breaks algorithm.

In Figure 6.6 we visualize the 28 clusters that reflect the spatial effect on the map of Netherlands[4]. Once again, a coefficient below one is represented in blue and indicates that the claim frequency of the spatial coefficient is lower than the base-level, and a coefficient above one is represented in orange, which indicates that the claim frequency of the spatial coefficient is higher than the base level. The interval of this clustered spatial coefficient ranges between [0.867621; 1.054641]. Based on visual inspection, the clustered spatial variable closely resembles the smooth bivariate effect of the spatial variable computed from the original GAM (in Figure 6.4b).

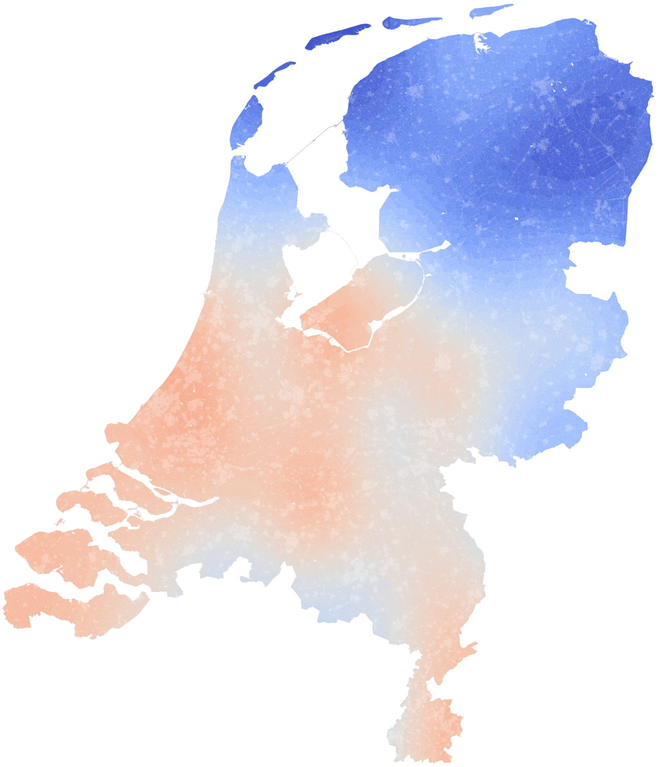

Figure 6.6: Map of Netherlands with the marginal effect of the clustered bivariate spatial variable with 28 different levels based on the breaks produced by the Jenks' natural breaks algorithm.

After clustering the smooth bivariate spatial variable, We model a GLM that implements the spatial effect as a categorical risk factor in combination with the other independent variables of the current industry-standard model. Figure 6.7 represents the spatial effect of the postal codes. Left is the current industry-standard model that we aim to improve (the same as Figure 6.2). On the right side we include, together with the demographic variables, the clustered smooth bivariate effect of the new spatial variable in the GLM. The interval of the total spatial effect of the latter model ranges between [0.800164; 1.602179], compared to a range between [0.851698; 1.568864] for the current

[4] Note that, the color scale in Figure 6.6 represents the same scale as Figure 6.3 in order to compare the marginal smooth effect of the bivariate spatial variable with the clustered version of this bivariate variable.

industry-standard model described in Section 6.1. Comparing these intervals, it can be observed that the range of the intervals of the improved GLM is larger compared to the base GLM, which, in practice, translates to larger differences in premiums for car insurances. Furthermore, if we compare the total spatial effect of the GLM with the GAM visualized in the figure 6.4 and the intervals of the improved GLM with the GAM (which had a interval range between [0.845669; 1.635882]), we find that the total spatial effect closely approaches those computed from the original GAM. This means that the outcomes of new GLM model closely resemble the effect described by the GAM. Furthermore, Appendix C1 show the estimated parametric and non-parametric components, respectively, of the model. Again, almost all variables are statistically significant and the effect of the province variable is reduced.

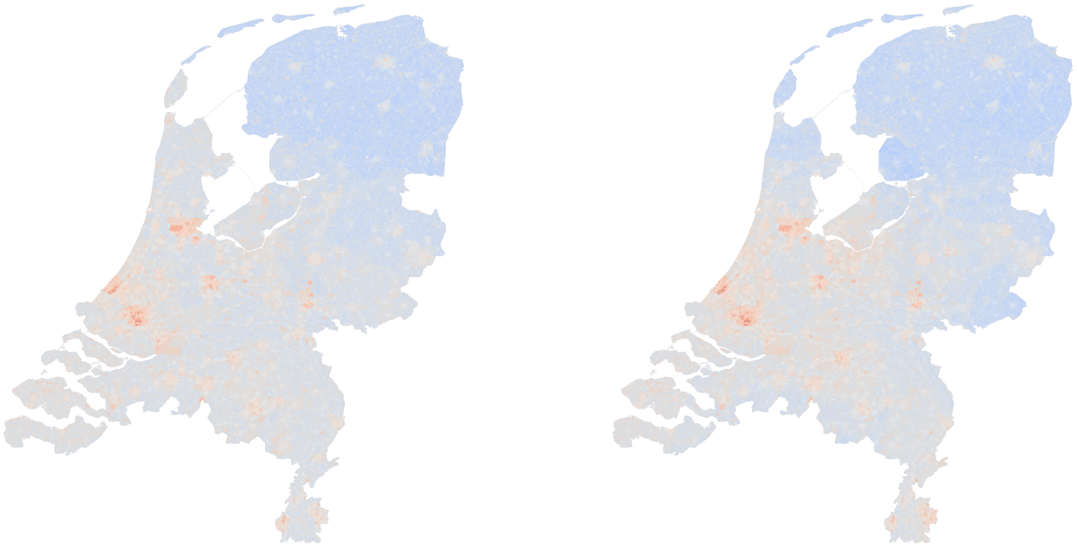

Figure 6.7: (a) The estimated marginal effect of the postal code based on demographic variables of the current industry-standard model (GLM). (b) The combined effect of the demographic variables with the clustered smooth bivariate effect of new spatial variable (GLM).

Chapter 7

Model comparison

In the previous chapter, we produce a model which has introduced a strategy to produce adequate class separation and provide a method to handle numerical variables within a GLM to determine the expected claim frequency. One of the research questions was to see whether it is possible to improve the expected claim frequency estimates by using more flexible methods in order to implement numerical spatial variables. Here, in order to answer this question, statistical metrics are used to analyze and compare the different models. The models are assessed and compared on its ability to predict outcomes and its goodness of fit to the available data. In the first scenario, we want to know if our model is capable of predicting new data accurately, and in the second scenario, we want to know if our model is able to accurately characterize the relationships between our present data. We therefore compare the different models, in Section 7.1 based on the fit of on the current dataset. Furthermore, we assess the predictive accuracy in Section 7.2. The table and figure in this chapter visualize parts of the dashboard created to assess different models and the spatial effect on the claim frequency. Therefore, in Section 7.3 we describe how the implementation of the entire investigative procedure is carried out.

7.1 The fit of the model

A core objective in model comparisons from a statistical perspective is to choose the model that finds the correct balance between goodness of fit (GOF) and parsimony [60]. The GOF indicates how well the model fits a set of observations, and measures of GOF summarize the discrepancy between observed values and the values predicted by the model under consideration. A parsimonious model is one that uses the fewest feasible predictor variables possible to provide the required degree of prediction. Models with low parsimony (i.e., those with many parameters) usually have a better fit. However, the risk is that the models have too many parameters and are therefore overfitting on the data. This means that so many parameters are used that the model fits the training data too well and therefore cannot predict closely on other datasets. On the other hand, once model possess too little parameters (high parsimony), there is a risk that it does not represent the correct relationships, has a high bias and does not fit the data enough. Bias often declines and variance normally rises as model complexity grows [61]. The concepts of underfitting and overfitting are visualized in Figure 7.1. The choice of model complexity is informed the right balance between those to concepts with the goal of minimizing the total error.

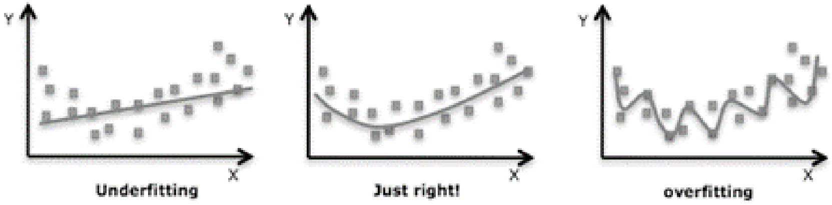

Figure 7.1: Visualization of the different possible fits of a model [61].

In order to find the right balance between GOF and parsimony there are multiple statistical measures available.The generalized Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) are used as standard for this purpose [60]. Therefore, we make use of the AIC and BIC as a measurement instrument. In Chapter 5.4 these metrics are described in more detail. These information criteria compares the quality of a set of statistical models to each other. Based on the AIC and BIC of each model, we can rank the models from best to worst, where higher AIC and BIC values correspond to better models. The best model will be the one that has the best fit compared to the others.

Table 7.1 shows the results of the information criteria for each model to assess on the basis of flexibility and the GOF. The best model would have lowest possible AIC and BICs. It can be seen that the AICs and BICs of the three models do not differ much, but it does show that the models which include the new spatial variable result in a lower AIC and BIC than the current industry-standard GLM. This indicates that including a (clustered) bi-variate spatial variable in the model is advantageous, resulting in better estimates of the claim frequency. In addition, the results show that the GLM with clustered spatial variable has the lowest BIC. In contrast, the GAM has the lowest AIC. This can be explained by the fact that a GAM uses nonlinear functions to estimate the numerical variables, and therefore the number of degrees of freedom used is higher than in the GLM model. This indicates that there is a trade-off between the AIC choosing rather a more simple model while the BIC prefers a more complex model. However, the relative difference between the BIC of the models using the (clustered) spatial variable compared to the industry-standard model is much smaller than the relative difference between the AIC of those two models. One could find arguments for choosing either of the two models, but we conclude that the GLM with the clustered spatial variable has the optimal balance between a good fit and complexity compared to the other models.

index	Frequency model	AIC	BIC
1	Current industry-standard GLM	3,585,490	3,583,689
2	GAM with the spatial variable	3,578,312	3,582,645
3	GLM with clustered spatial variable	3,578,618	3,580,811

Table 7.1: The AIC and BIC of the frequency models.

7.2 Prediction accuracy

Given the difficulties of assessing the quality of regression models, we used cross-validation techniques in order to asses the predictive accuracy of the models on new data. Cross validation (CV) is a common technique used to test the effectiveness of a regression models [62]. There are several CV techniques, but they basically consist of separating the data into training and testing subsets. The most simple form of CV is the hold-out method. This method separates the dataset into two groups without overlapping, a trainset and a testset. The trainset is used to train the model and the remaining part is used to test the predictive power of the model based on some metrics, which will be discussed later. However, the hold-out method score could be influenced by how the data is divided into the train- and test set. k-Fold CV is generally preferred since it allows your model to train on numerous train-test splits [63]. This provides you a better idea of how your model will perform on untested data and avoids the bias of the single split could produce. Therefore the k-fold method produces more reliable outcomes. The hold-out method is good to use, simpler, and needs less computational power and time. However, we chose to use the k-fold CV since it produces more reliable results.

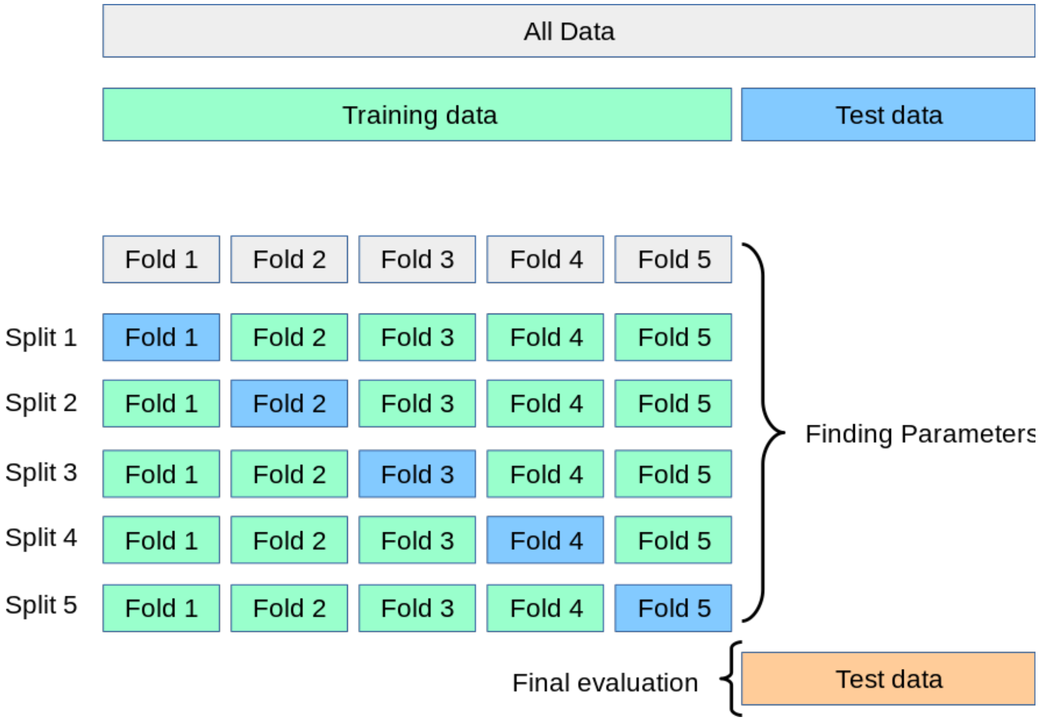

Figure 7.2: Splitting Data 5-fold Cross Validation [62].

Even though CV is a common method of evaluating models, it is not without drawbacks. Running multiple iterations of a model on the same dataset for CV can provide good results just by coincidence [39]. Therefore we have implemented an additional step. In order to avoid, or at least identify, overfitting of the model on the training data, we are first dividing our dataset into two separate test- and training datasets. The training dataset will be used to perform k-fold CV, and using the test data, we can assess the model's capability to generalize to unseen data (Figure 7.2). Splitting data by time into chronological sets (i.e., Going one year back in time and predict the data for that year). In our case a training set with data of 2015 until 2019 and a test set with data of 2020. This makes it possible to check for parameter stability over time. However, as described in Section 3.1.2, in 2020 there were substantially fewer claims caused by the COVID-19 pandemic and therefore distributions differs of the different datasets. So, if the data were to be split based on this chronological method, this effect would not be included in the model. Therefore, we split our dataset by random assignment; a train set containing (80%), and a test set containing (20%) of the data. The train set is used for the K-fold CV to initially estimate the models, which we explain later. The test set is used to provide an unbiased assessment of the model. It is important that the test set has never been used for training, as otherwise, the results might be biased [64]. Note that after splitting the data set, it is important to check for possible coincidental differences between the subsets. Therefore, we compare the means and quantiles and look for deviations.

We use the R package Caret to perform the k-fold CV. The process of k-fold CV starts with splitting the original data into k subsamples (e.g., k = 5). After that, models are estimated using k-1 subsamples for each fold, with the k^{th} sub sample serving as the validation sample. This process repeats until every subsample has served as the validation data and model results can be averaged across folds.The choice of k is usually 5 or 10, but there is no standard rule [65]. We chose $k = 5$.

There are many standard metrics in order to assess model prediction performance and to compare various models based on predicted accuracy. We use the Mean Squared Error (MSE) and Mean Squared Error represents (MAE) metrics since they are mainly used to evaluate the prediction error rates and model performance in regression analysis [66]. The MSE measures the average of the squared difference between the original and predicted values in the dataset, whereas the MAE measures the average of the squared difference between the original and predicted values in the dataset. These metrics can be expressed as follows:

$$MAE = \frac{1}{b}\sum_{i=1}^{b} |y_i - \hat{y}_i| \tag{7.2.1}$$

$$MSE = \frac{1}{b}\sum_{i=1}^{b} (y_i - \hat{y}_i)^2 \tag{7.2.2}$$

where $\hat{y}$ represents the predicted value of y for $i = 1, ..., b$ predictions. All the models were evaluated using these metrics.

After training the model, the model needs to be evaluated. This is where the test data come in. The test data are taken as input for the trained model and compared using metrics discussed above to asses the prediction accuracy. When those outcomes do not match with the cross-validated ones, it means that the models do not adapt well to new data.

Using the mentioned the two measures, we assess the improved prediction of the claim frequency models. Table 7.2 compares the outcomes of the performance metrics, in two distinct ways, produced by the the different frequency models, namely the current industry-standard GLM (with 23 variables), the GAM model (with 23 variables and the smooth bivariate spatial variable), and the improved GLM (with 23 variables and the clustered smooth bivariate spatial variable with 37 levels). The first two columns are based on the mean of the performance of each iteration of the CV to evaluate the model. The last two columns are obtained after the model was trained, by using the model to execute by the out-of-sample test set. When comparing the results there does not appear to be notable difference between the metrics of the trained CV set and the test-set. Furthermore, when we compare the metrics there does not appear to be a notable difference between the results of of the models that include the new spatial variable. However, the GAM has a slightly larger MSE and a smaller MAE than the new GLM including the spatial variable. This means that the difference between the true and predicted values using the GAM was, for some predictions, very large, whereas the predictions by the new GLM were on average closer to the true value without large outliers. Furthermore, the results indicates that the models that include the new spatial variable result in a more accurate predictions on the claim frequency compared to the current industry-standard GLM. Note that, it may seem like these are relatively small improvements in the predictive performance of the models. However, based on a portfolio of 1.7 million policies, this can have a huge impact on the premiums. In addition, we have to keep in mind that we add 1 covariate on a model of 23 variables. All together, this indicates that implementing a numerical spatial variable in the model is advantageous, resulting in better estimates of the claim frequency.

Type of model	MSE (CV)	MAE(CV)	MSE(test-set)	MAE(test-set)
Current GLM	0.1864273	0.06505922	0.1853024	0.06368340
GAM with spatial variable	0.1824066	0.06271937	0.1822836	0.06267688
GLM with spatial variable	0.1824083	0.06271836	0.1822849	0.06267605

Table 7.2: Mean values of the MSE and MAE in the CV step and the metrics of the test-set step.

7.3 Implementation

In this section, we outline the implementation of the data preparation and modeling process. The aim is to automate and simplify the entire process and present it in a clear and logical sequence, while also allowing for interpretability and flexibility.

To automate the data preparation steps outlined in Section 3, we utilize the SASPy package in Rstudio to establish a connection between Rstudio and the data warehouse of the Dutch insurance company. Furthermore, such connection helps prevent data leakage by eliminating the need to download CSV files and store them on local computers. To establish the connection, the user is prompted to enter their SAS (SQL database) username and password within Rstudio. After establishing the connection the required datasets can be imported based on a SQL query. By creating an identical key the required datasets are merged. Upon completion of the merging process, the created dataset is returned to the data warehouse via the established connection, which removes the need for local copies of the data.

The merged dataset contains columns which require manual processing because some columns contain values which are not missing at random or need to be completely removed. Additionally, the variables must be modified or clustered in order to be incorporated into the model. If following the protocol as described here, the data preparation tasks have already been automated and do not need to be customized. Furthermore, the MissForest algorithm (described in Section 3.3.1) is modeled in such way that it can be applied to any dataset. The missing values are automatically removed during the process.

The model we designed is highly flexible and can be used with any combination of response variables and covariates. It is also easy to adjust the model's parameters as needed. Additionally, we chose to program the model from scratch, which enables manual interpretation of the underlying formulas. Furthermore, the computation time of the model can be significantly reduced by paralleling the training and validation steps. This is accomplished by dividing the training data into k-fold sets and training the model in parallel on each set.

Furthermore, we developed a dashboard to facilitate the visualization and understanding of the insights discussed in this chapter. This dashboard include the visualization of expected claim frequency per postal code visualized on the map of the Netherlands, the expected versus true plots and the visualizations of the metrics to asses the models. The geographic plots of the Netherlands allow us to explain the spatial risk per postal code in a interpretable way to multiple stakeholders of the car insurance product. Previously, this was not possible for the Dutch insurance company.

Chapter 8

Conclusions and recommendations

8.1 Conclusions

In order to accurately represent the heterogeneity of risks within portfolios, insurance companies differentiate premiums based on the policyholders' risk profiles. Within a risk level, policyholders pay an equal premium that is meant to reflect their inherent risk. The current-industry standard for modeling risk levels is using a GLM. However, the main disadvantage of a GLM is that numerical risk factors cannot be included in the model directly. Here, we present a strategy that allows for the implementation of a numerical spatial variable (the geographical locations of policyholders' residences) in a GLM, which allows for more accurate prediction of the claim frequency of the car insurance model used within the Dutch insurance company. Combining this improved claim frequency model with the current claim severity model allows the Dutch insurance company to more accurately calculate the pure premium, ultimately benefiting both policyholders as well as the company itself.

GAMs are models that can easily accommodate nonlinear effects of numerical variables. Additionally, GAMs are much more flexible than GLMs and completely automatic, as there is no need to specify a certain function in advance. The major downside of using GAMs is that the estimates are more difficult to interpret compared to GLMs. Since one of the objectives of this study was to retain the advantages of the GLM and complement the current pricing model used within the Dutch insurance company, we used Jenks' natural breaks algorithm to cluster the smooth spatial effect obtained by the GAM into categorical levels, to implement it in a parametric GLM. Subsequently, the optimal number of clusters for this specific dataset was determined. Overall, the proposed approach resulted in an easy-to-understand predictive model formulated within the well-known framework of generalized linear models. We therefore end up with a simpler model that can be deployed in practice as a close substitute for a complex, flexible model.

When comparing the results of the industry-standard GLM to the proposed approach, it is clear that our model leads to more accurate prediction, with a decrease of the MSE of 1.65% and the MAE with 1.60% based on the out-of-sample test. Additionally, our approach results in an improved representation of the spatial variable, since the effect is more gradual. This leads to fairer premiums for policyholders, which is not only important in the context of car insurances, but can be extended to many similar problems caused by step-wise functions. For example, during the energy crisis in 2022, the Dutch government financially supported entrepreneurs if their costs increased beyond a certain cut-off due to the rising energy prices, leading to unfair compensation for entrepreneurs that fell just outside of this bracket [67]. It may seem like these are small improvements in the predictive performance of the models. However, based on a portfolio of 1.7 million policies, this can have a huge impact on the premiums. Naturally, the proposed approach remains a clustering method, meaning that the variables will still need to be divided into groups. However, using our data-driven approach, these clusters are less random and the differences between adjacent clusters are more gradual compared to the industry-standard model. Additionally, this research has led to important insights into the effect of the geographical location on claim frequency, mainly due to the visualization using the map of the Netherlands. This is a first step into the right direction and can give actuaries valuable insights on the data on which the premiums are based.

Concluding, the proposed strategy is a GLM that more accurately predicts the claim frequency compared to the current industry-standard GLM and provides valuable insights into the geographical variables. This fulfills the objectives of this study, to create an approach that retains the advantages of the GLM whilst improving upon the current claim frequency model for the car insurance product used within the Dutch insurance company.

8.2 Discussion and recommendations

Based on our findings, we recommend implementing our approach in car insurance policies in general as it shows promising results. However, we also recognize that there is potential for further improvement and recommend extending this research to refine the approach, in order to fully understand the potential and limitations. In this section, we identify and highlight some of the shortcomings of this study that can serve as starting points for further research.

Firstly, it is important to note that the addition of a new variable, such as the spatial variable, also affects other variables in the model. For example, with the addition of the spatial variable, the province factor in the model was no longer statistically significant. This is most likely due to the fact that the spatial variable captures the same effect as the province variable but allows the effect to be estimated more accurately. Further research should be done on feature selection, in order to reveal whether it matters that this method weakens other factors and whether those should be left out of the model. It should be noted that, if the approach uses an different combination of covariates or assumes different underlying assumptions, it may lead to different conclusions.

In order to calculate the premium for a car insurance, both the claim frequency and claim severity need to be taken into account. Here, we only investigated the effect of the spatial variable on the frequency model, since we expect that the effect of the spatial variable on the severity model is less pronounced. However, future research could investigate whether implementing the spatial variable is relevant to the severity model as well. Furthermore, the proposed approach can be extended to any other numerical variable, such as age or weight of the car, that are currently grouped based on expert opinion. We expect our data-driven approach to be applicable to these variables as well. Additionally, the bi-variate smooth function of the GAM can be used to model interaction effects as well. For example, Su and Bai [68], have found a significant interaction between the driver and vehicle age, and argue that the vehicle age has a significant effect on the claim frequency for young drivers, and that this effect decreases when the driver's age increases.Additionally, this approach is not only applicable to the specific case of car insurances, or insurances in general. It can be applied in many business environments, where there is a need for the approximation of the effects of numerical variables with easily explainable and simple to program predictive models, such as credit scoring, because the need for transparency and ease of explanation of the credit model. The research presented here allows for further investigation on these and similar cases, and allows for further optimization of insurance products.

We focused on developing a model that is intuitive and explainable, which means that we have to face the trade-off between flexibility and interpretability. Therefore, we chose to base our work on GAMs, which offer a middle ground between easily interpretable but inflexible models such as linear models, and flexible but complex deep learning models like neural networks [69]. These models can fit complex, nonlinear relationships and make accurate predictions, but at the same time, the underlying structure of the models can be easily understood and explained, and one can still perform inferential statistics using these models. However, we would recommend looking beyond regression models and perform research on more flexible machine learning methods as these methods might be able to predict with higher accuracy. Knowing the limitations of the machine learning methods, we propose to focus future research on a hybrid system in which actuaries obtain some information from machine learning model as input for the statistical models. This would be an efficient way to combine the quality benefits of an ML model with the standard-industry regression models [70].

There are a few points of discussion when that affect the accuracy of the predictions of our proposed approach. Firstly, we assume that our data is distributed according to the Poisson distribution, in which the variance is equal to the mean. This is the standard distributions for claim frequency models of the car insurance products of the Dutch insurance company. However, in the real world, the mean is never known precisely. Therefore, it is not known if the data accurately fits this distribution, or if a different distribution (such as the negative binomial distribution where the variance is greater than the mean) might more accurately represent the data. The choice of distribution will depend on the nature of the data and the research question being addressed. Therefore, we recommend to do further research to find the optimal distribution for the claim frequency. Secondly, in order to avoid reducing the information content in an incomplete dataset, we used the MissForest imputation method to handle missing values. This was selected based on a literature review. In order to completely investigate the extent to which the different methods perform well, we recommend conducting an experimental study and analyzing different imputation methods. Additionally, even though the MissForest method does not lead to the loss of data, it should be kept in mind that imputing data with missing values does not increase the information contained in this dataset either. Lastly, the Jenks' natural breaks algorithm was chosen based on three clustering methods evaluated. The AIC and BIC were used to select the most appropriate number of clusters for this dataset. However, these metrics did not lead to a clear optimal number of clusters. In addition to using the AIC and BIC, the Goodness of Variance Fit (GVF) measure can be used to guide the choice for the optimal number of clusters, as the GVF is the metric that the Jenks natural breaks algorithm aims to optimize. Therefore, we recommend doing more research on improving the tuning strategy for the optimal number of clusters.

Bibliography

[1] Yang, Y., Qian, W., & Zou, H. (2018). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, *36*(3), 456–470.

[2] Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory: Using r* (Vol. 128). Springer Science & Business Media.

[3] Frees, E. W. (2014). Frequency and severity models. *Predictive modeling applications in actuarial science*, *1*, 138–164.

[4] Xie, Y., Lv, H., Sun, X., Mao, Y., & Yang, J. (2019). Study on the transform method of estimating discrete frequency from continuous variable: Ratemaking for car repair insurance based on sas system coding. *Cluster Computing*, *22*(6), 15493–15503.

[5] Tufvesson, O., Lindström, J., & Lindström, E. (2017). Spatial statistical modeling of insurance risk an epidemiologist approach to car insurance.

[6] Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681–705.

[7] Guven, S. (2004). Multivariate spatial analysis of the territory rating variable. *Colorado Springs, Colorado: Casualty Actuarial Society*.

[8] Ohlsson, E., & Johansson, B. (2010a). *Non-life insurance pricing with generalized linear models* (Vol. 2). Springer.

[9] Denuit, M., & Trufin, J. (2019). Effective statistical learning methods for actuaries.

[10] Maimon, O., & Rokach, L. (2009). Introduction to knowledge discovery and data mining. *Data mining and knowledge discovery handbook* (pp. 1–15). Springer.

[11] Fayyad, U. (1997). Data mining and knowledge discovery in databases: Implications for scientific databases. *Proceedings. Ninth International Conference on Scientific and Statistical Database Management* (*Cat. No. 97TB100150*), 2–11.

[12] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*, 161–168.

[13] Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan kaufmann.

[14] Loi, M., & Christen, M. (2021). Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, *34*(4), 967–992.

[15] *Algemene verordening gegevensbescherming* (*avg*). (n.d.). Retrieved May 11, 2022, from https://autoriteitpersoonsgegevens.nl/wetten/algemene-verordening-gegevensbescherming-avg

[16] Netherlands Enterprise Agency, RVO. (2022). Third-party liability insurance for motor vehicles. Retrieved May 10, 2022, from https://business.gov.nl/regulation/vehicle-insurance/

[17] Philips, R. (2021). The third party litigation funding law review: Netherlands. *Redbreast Associates*.

[18] Ohlsson, E., & Johansson, B. (2010b). The basics of pricing with glms. *Non-life insurance pricing with generalized linear models* (pp. 15–38). Springer.

[19] Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

[20] Klein, N., Denuit, M., Lang, S., & Kneib, T. (2014). Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, *55*, 225–249.

[21] Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2016). *Generalized linear models for insurance rating* (Vol. 5).

[22] Brockman, M. J., & Wright, T. (1992). Statistical motor rating: Making effective use of your data. *Journal of the Institute of Actuaries*, *119*(3), 457–543.

[23] Nurunnabi, A., & West, G. (2012). Outlier detection in logistic regression: A quest for reliable knowledge from predictive modeling and classification. *2012 IEEE 12th international conference on data mining workshops*, 643–652.

[24] Kafková, S. (2015). Bonus-malus systems in vehicle insurance. *Procedia economics and finance*, *23*, 216–222.

[25] Yasin, Y. J., Grivna, M., & Abu-Zidan, F. M. (2021). Global impact of covid-19 pandemic on road traffic collisions. *World journal of emergency surgery*, *16*(1), 1–14.

[26] Swedler, D. I., Bowman, S. M., & Baker, S. P. (2012). Gender and age differences among teen drivers in fatal crashes. *Annals of Advances in Automotive Medicine/Annual Scientific Conference*, *56*, 97–116.

[27] Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147–177.

[28] Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L., & Goodacre, R. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, *4*(2), 433–452.

[29] Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

[30] Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406.

[31] Kellermann, P., Travathan, D., & Kromrey, J. (2016). Missing data and complex sample surveys using sas®: The impact of listwise deletion vs. multiple imputation on point and interval estimates when data are mcar and mar.

[32] García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl. Based Syst.*, *98*, 1–29.

[33] Patwardhan. (2022, June 23). *Simple understanding and implementation of knn algorithm!* Retrieved June 5, 2022, from https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/

[34] Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

[35] Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á., & Barbas, C. (2015). Missing value imputation strategies for metabolomics data. *Electrophoresis*, *36*(24), 3050–3060.

[36] Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970.

[37] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

[38] Clijsters, M. (2015). Dealing with continuous variables and geographical information in non-life insurance ratemaking: Practical solutions applied to a car insurance data set. *KU Leuven master thesis*.

[39] De Jong, P., Heller, G. Z. et al. (2008). Generalized linear models for insurance data. *Cambridge Books*.

[40] McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

[41] Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the frequency and severity of auto insurance claims using statistical distributions. *Journal of Mathematical Finance*, *08*, 137–160.

[42] Bakari, H., Adegoke, T., & Yahya, A. (2016). Application of newton-raphson method to non-linear models. *International Journal of Mathematics and Statistics Studies*, *4*(4), 21–31.

[43] Denuit, M., & Lang, S. (2004). Non-life rate-making with bayesian gams. *Insurance: Mathematics and Economics*, *35*(3), 627–647.

[44] De Castro, Y., Gamboa, F., Henrion, D., Hess, R., & Lasserre, J.-B. (2019). Approximate optimal designs for multivariate polynomial regression. *The Annals of Statistics*, *47*(1), 127–155.

[45] Hastie, T., & Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical methods in medical research*, *4*(3), 187–196.

[46] Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.

[47] Harrell, F. E. et al. (2001a). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. 608). Springer.

[48] Wood, S. N. (2006). *Generalized additive models: An introduction with r*. chapman; hall/CRC.

[49] Chi, G. (2013). Applied spatial data analysis with r. *Spatial Demography*, *1*, 227–228.

[50] Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Machine learning proceedings 1995* (pp. 194–202). Elsevier.

[51] Slocum, T. A., McMaster, R. M., Kessler, F. C., Howard, H. H., & Mc Master, R. B. (2008). Thematic cartography and geographic visualization.

[52] van de Velden, M., D'Enza, A. I., & Palumbo, F. (2017). Cluster correspondence analysis. *Psychometrika*, *82*(1), 158–185.

[53] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. *2010 IEEE international conference on data mining*, 911–916.

[54] Jenks, G. F., & Caspall, F. C. (1971). Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers*, *61*(2), 217–244.

[55] Smith, R. M. (1986). Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Geographer*, *38*(1), 62–67.

[56] Pasławski, J., Korycka-Skorupa, J., Nowacki, T., & Opach, T. (2012). Choropleth maps and diagram maps in atlas of cartographic presentation methods. *Miscellanea Geographica. Regional Studies on Development*, *16*(1), 49–56.

[57] Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

[58] Harrell, F. E. et al. (2001b). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. 608). Springer.

[59] Ministerie van Infrastructuur en Waterstaat. (2022). Difference in car-dependency between urban and non-urban areas is growing in the Netherlands. https://english.kimnet.nl/latest-news/feature/2022/02/22/difference-in-car-dependency-between-urban-and-non-urban-areas-is-growing-in-the-netherlands

[60] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

[61] Khan, M., & Srivastava, K. (2020). Regression model for better generalization and regression analysis. *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 30–33.

[62] Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

[63] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

[64] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

[65] Kuhn, M., Johnson, K. et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

[66] Silveira Kupssinskü, L., Thomassim Guimarães, T., Menezes de Souza, E., C. Zanotta, D., Roberto Veronez, M., Gonzaga Jr, L., & Mauad, F. F. (2020). A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. *Sensors*, *20*(7), 2125.

[67] Rijksoverheid. (2022). Kabinet vergoedt vanaf 1 november kosten energie-intensief mkb. Retrieved May 12, 2022, from https://www.rijksoverheid.nl/actueel/nieuws/2022/10/14/kabinet-vergoedt-vanaf-1-november-kosten-energie-intensief-mkb

[68] Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PloS one*, *15*(8), e0238000.

[69] Gopalani. (2021). Generalized Addictive Models (GAMs) - The Startup. Retrieved July 12, 2022, from https://medium.com/swlh/generalized-addictive-models-gams-72fc77bb0b29

[70] Nia, S. M. (2021). The capability of machine learning for predicting disability probabilities. *University of Twente master thesis*.

[71] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2009). Practitioner's guide to generalized linear models.

Appendix A

Newton-Raphson method

Analytical solutions for complex GLMs are usually not easily available, and therefore numerical methods are necessary. The Newton-Raphson method can be used to approximate solutions when no closed-form exists, and it can converge quickly.

When we want to find the root of the function $L(\beta)$, and we have initial guess β, this method makes use of a linear approximation, i.e. tangent line, to update the initial guess. This process repeats itself until the algorithm converges. If $\hat{\beta}_j$ is the current estimate, then the next estimate $\hat{\beta}_{j+1}$ is given by

$$\hat{\beta}_{j+1} = \hat{\beta}_j - \frac{\delta^2 L(\hat{\beta})}{\delta \beta^2}^{-1} \frac{\delta L(\hat{\beta})}{\delta \beta} \tag{A.0.1}$$

Our goal is to find the system of partial log-likelihood derivatives with respect to the parameters β in equation 4.2.10. So to compute equation A.0.1, we need the derivative of 4.2.10, i.e. the second order partial derivatives of the log-likelihood. These are obtained by taking the derivative of 4.2.11 once more, we get

$$\frac{\delta^2 L(\beta)}{\delta \beta \delta \beta'} = -\sum_{i=1}^{n} x_i x_i' exp(x_i' \beta) t_i. \tag{A.0.2}$$

Thus by substituting equations 4.2.10 and A.0.2 in to equation A.0.1, we can predict the parameters β. We keep iterating until convergence, that is, until the difference ϵ between two consecutive iterations is very small, for example $|x_n - x_{n-1}| < \epsilon = 0.00001$. For a more in depth overview, see for example Bakari et al. [42].

Appendix B

Factor categorization

One automated approach within the GLM framework to categories a single factor with many levels that have a natural ordering is given in Anderson et al. [71]. Here, the idea is to replace the single factor with many levels with a series of binary factors and test these binary factors for significance. For example, instead of modeling driver age, the following factors can be created:

- (binary factor 1) age less than 18;
- (binary factor 2) age less than 19;
- ...
- (binary factor 22) age less than 39;
- (age 40 is the base level in this example)
- (binary factor 23) age less than 41;
- ...
- (binary factor 82) age less than 100.

Then the binary factors can be tested for significance and the least significant factors are iteratively removed from the model until all factors remaining in the model are significant. The implied parameter estimates for each age could then be obtained by summing the appropriate binary factors. For example, the implied parameter estimate for age 41 would be the sum of the parameters corresponding to binary factors 24 to 82.