Using risk factors in P&C insurance pricing: a data driven strategy with GAMs, regression trees and GLMs.

Roel Henckaerts[1], Katrien Antonio[1,2], Maxime Clijsters and Roel Verbelen[1]

[1]Faculty of Economics and Business, KU Leuven, Belgium.
[2]Faculty of Economics and Business, University of Amsterdam, The Netherlands.

November 10, 2016

Abstract

We present a fully data driven strategy to deal with continuous risk factors and geographical information in an insurance tariff. We develop a framework that aligns the statistical advantages of flexible modelling to the practical requirements of an insurance company. We start by fitting flexible generalized additive models (GAMs) to the claims data. These models contain categorical risk factors, smooth effects of continuous risk factors and a spatial effect that captures geographical information. The goal is to bin the continuous risk factors and the spatial effect in order to transform them to categorical risk factors and incorporate them in generalized linear models (GLMs). The spatial effect is binned by the Fisher-Jenks algorithm and the continuous risk factors are binned by evolutionary trees. In a last step, GLMs are fitted to the claims data with the resulting categorical risk factors.

Keywords: P&C insurance pricing, continuous risk factors, geographical information, data driven binning, generalized additive models (GAMs), evolutionary trees, Fisher-Jenks, generalized linear models (GLMs)

1 Introduction

This paper presents a fully data driven framework to deal with continuous risk factors and geographical information in a practical insurance tariff. We start from flexible generalized additive models (GAMs) because of their statistical advantages. Driven by the practical requirements of an insurance company, we transform these GAMs to generalized linear models (GLMs). The GAMs contain categorical risk factors, smooth effects of continuous risk factors and a spatial effect that captures geographical information. The continuous risk factors and spatial effect are transformed to categorical risk factors (*binning*) to be able to incorporate them in GLMs. This appraoch allows us to develop a GLM which is easy to interpret and takes into account the statistical properties of the data in an efficient way.

This paper should be framed in between two existing approaches to handle different types of risk factors in the literature on P&C ratemaking. On the one hand some authors develop GAMs for tarification with flexible effects of continuous and spatial risk factors (see Denuit and Lang (2004), Klein et al. (2014)). This approach leads to models with good statistical properties, but they are often too complex for usage in a practical insurance tariff. On the other hand many authors use predefined bins for the continuous risk factors (see Frees and Valdez (2008), Antonio et al. (2010)). Dougherty et al. (1995) gives an overview of methods that can be used to bin continuous risk factors in advance. A disadvantage of this approach is that the response variable is not taken into account in the binning process. What is lacking is a practical framework that aligns the statistical advantages of flexible modeling with GAMs to the requirements of a production environment in an insurance company. Rating models should be interpretable, intuitive, easy to program, deployable and adjustable to marketing needs and benchmark studies with competitors. This paper tries to fill the gap between the two existing approaches by starting from GAMs with good statistical properties and transforming these into GLMs that satisfy the practical needs of an insurance company.

Policyholders differ in their risk factors, implying that different policyholders have different risk profiles. An insurance portfolio comprises many policyholders and is therefore rather heterogeneous. In order to reduce this heterogeneity, insurance companies partition policyholders in several risk categories. This process is known as risk classification (see Denuit et al. (2007), Antonio and Valdez (2012)) and the goal is to group policyholders with similar risk profiles in the same risk category. A distinction can be made between a priori and a posteriori risk factors. A priori risk factors are known at the start of the insurance contract, like for example the age of the policyholder or the amount of horsepower of the car. The same a priori premium is charged to policyholders in the same risk category. A posteriori risk factors are only known after a certain period of time since the start of the insurance contract. An example of an a posteriori risk factor is the bonus-malus level, which gives an indication of the claim history of the policyholder. The bonus-malus level is typically used as an a posteriori correction in credibility models or bonus-malus schemes (see Lemaire (1995)).

Three types of risk factors are considered in this paper: categorical, continuous (both single and interaction) and spatial effects. Categorical risk factors have a discrete number of possible outcomes or levels. Examples are the type of coverage and the gender of the policyholder. Continuous risk factors can attain all real values in a specific interval. Examples of continuous single effects are the age of the policyholder and the amount of horsepower of the car. Continuous interaction effects capture how different risk factors interact with each other and can be interpreted as a correction term to the single effects. An example is the interaction between the age of the policyholder and the amount of horsepower of the car. The spatial effect takes into account the geographical area where the policyholder resides. Examples are the longitude and latitude coordinates of districts and postal codes. This information can serve as a proxy for the region where a policyholder drives his car.

Actuarial models for property and casualty (P&C) ratemaking put focus on two components: a model for the frequency of claims and a model for the severity of claims (see Denuit et al. (2007), Frees et al. (2014) and Parodi (2014)). Frequency is defined as the number of claims per unit of exposure. Exposure, as described in McClenahan (1990), can be seen as a rating unit and measures how long the policyholder is exposed to the insured risk. An example of exposure is the fraction of the year for which premium has been paid and therefore coverage is provided. Severity is the average claim cost and can be calculated as the ratio of the total loss to the corresponding number of claims (causing this total loss) over a specific period of insurance. An actuary develops separate regression models for the frequency and severity component. These two components are typically assumed to be independent and the risk premium can therefore be calculated as the product of the expected value of the frequency and the expected value of the severity (see Klugman et al. (2012)). Alternatives for this independence assumption are investigated in the literature, allowing dependence between frequencies and severities (see Gschlößl and Czado (2007), Czado et al. (2012)).

Generalized linear models (GLMs), developed by Nelder and Wedderburn (1972), have become the industry standard to produce regression models for insurance applications (see Haberman and Renshaw (1997), Denuit et al. (2007), De Jong and Heller (2008)). GLMs allow the response variable to follow any distribution of the exponential family. The Poisson distribution is particularly interesting for claim frequency models whereas the Gamma distribution is often used for claim severity modelling. The relationship between a resonse and the covariates is often multiplicative rather than additive (see Ohlsson and Johansson (2010)). GLMs therefore model a transformation of the mean of the response by a linear predictor. The logarithm is a well known example of such a link function. Since covariates enter the model through a linear predictor, GLMs are not well suited for continuous risk factors that relate to the response in a non-linear way. Generalized additive models (GAMs), developed by Hastie and Tibshirani (1990), extend the framework of GLMs and allow for non-parametric smooth functions in the predictor structure. GAMs therefore enable one to model smooth effects of continuous risk factors. This extra flexibility results in a statistically more elegant model, though less transparent. For marketing and ICT departments within insurance companies it is not easy to work with GAMs and therefore GLMs are favored. Actuaries prefer to work with GLMs in comination with categorical risk factors, because this allows them to work via dummy variables (see Denuit et al. (2007)). It is however unclear how to bin continuous risk factors in order to obtain categorical risk factors. Insurance companies often use expert knowledge or simple binning procedures, like quantile binning. This paper proposes a more advanced and fully data driven alternative.

2 Claims data set

The data set covers a motor third party liability (MTPL) insurance portfolio with a Belgian insurer during one specific year (see Denuit and Lang (2004), Klein et al. (2014) where a sample from this data set is analyzed). Every policyholder occurs only once in this data set such that we consider the risk factors to be constant and registered at the start of the period of insurance. The data set contains 163231 policyholders and the available risk factors are listed in Table 1.

Variable	Description
`nclaims`	The number of claims filed by the policyholder in 1997.
`exp`	The fraction of the year 1997 that the policyholder was exposed to the risk.
`amount`	The total amount claimed by the policyholder in 1997.
`coverage`	Type of coverage provided by the insurance policy (only third party liability (TPL), partial omnium (PO) which covers TPL and limited material damage or full omnium (FO) which covers TPL and comprehensive material damage).
`fuel`	Type of fuel of the vehicule (gasoline or diesel).
`sex`	Gender of the policyholder (male or female)[1].
`use`	Main use of the vehicule (private or work).
`fleet`	Indicator whether the vehicule is part of a fleet (yes or no).
`ageph`	Age of the policyholder (on January 1, 1997).
`power`	Horsepower of the vehicule in kilowatt (on January 1, 1997).
`agec`	Age of the vehicule (on January 1, 1997).
`bm`	Level occupied in the former compulsory Belgian bonus-malus scale (on January 1, 1997). Going from 0 to 22, a higher level indicates a worse claim history (see Lemaire (1995)).
`long`	Longitude coordinate of the center of the district where the policyholder resides.
`lat`	Latitude coordinate of the center of the district where the policyholder resides.

Table 1: `MTPL`: an enumeration and short description of all the risk factors.

Figure 1 indicates how the risk factors from Table 1 are distributed in the `MTPL` data set. Most policyholders, nearly 88.79% of them, are claim-free in the year 1997. Approximately 10.14% of the policyholders file one claim and the remaining 1.07% of policyholders file two, three, four or five claims in the year 1997. Around 77.33% of the policyholders have an exposure of 1 and are therefore covered by the insurance and exposed to the risk during the entire year. The exposure of the other 22.67% of the polichholders is approximately uniformly distributed between 0 and 1. Policyholders with an exposure lower than 1 have surrendered their policy during the year or they started their policy in the course of the year. The overall claim frequency of the portfolio in the `MTPL` data set, calculated as the total number of claims divided by the total exposure, is equal to 13.93%. The average claim severity of the portfolio in the `MTPL` data set, calculated as the total amount of claims divided by the total number of claims, is equal to 1620.06 EUR.

The `MTPL` data set contains five categorical risk factors: `coverage`, `fuel`, `sex`, `use` and `fleet`. Approximately 58.28% of the policyholders have only TPL coverage, which means that only their liability with respect to a third party (that is: another person) is covered. As in many developed countries, this coverage is compulsory in Belgium. The other policyholders have chosen for a policy that covers material damage on top of the TPL. More specifically 28.17% of them chose for a partial omnium (PO) which covers limited material damage, while 13.54% chose for a full omnium (FO) which gives comprehensive coverage. Two types of fuel are used in the cars of the policyholders: gasoline by 69.12% of them and diesel by the other 30.88%. Approximately 73.55% of the policyholders are males, 95.17% of the policyholders use their car mainly for private reasons rather than professional reasons and around 96.83% of the cars are not part of a fleet.

The `MTPL` data set contains four continuous risk factors:, `ageph`, `power`, `agec` and `bm`. Almost all

[1] Article 5(1) of Directive 2004/113/EC, also known as the European Gender Directive, prohibits the use of gender to discriminate between males and females regarding tarification. This covariate is therefore only investigated for an internal, technical tariff, but can not be used in a commercial tariff.

policyholders, 93.53% of them, are between the ages of 25 and 75, which means that there are few young and old people in the insurance protfolio. Most of the cars in the insurance portfolio have less than 100 kilowatt of horsepower (97.35%) and are younger than 20 years old (99.53%). The rather low range of horsepower is nowadays outdated, but fits the less powerful cars from back in 1997. More than half of the policyholders reside in the two lowest bonus-malus scales, 37.77% in level 0 and 16.52% in level 1. Most of the other polichholders, approximately 42.90%, have a bonus-malus scale between 2 and 11 and almost no policyholders (2.81%) occupy a bonus-malus level higher than 11. It should be noted that the bonus-malus level is usually not incorporated as a risk factor to develop an a priori tariff. However, we keep the variable in the analysis to investigate the information contained in this risk factor, much in line with the work of Denuit and Lang (2004). Investigating the interaction between `ageph` and `power` confirms that most of the policyholders have an age between 25 and 70 and a car with horsepower between 40 and 70.

The `MTPL` data set contains geographical information in the form of longitude and latitude coordinates: `long` and `lat`. The map of Belgium in Figure 1 gives an idea of the total exposure in each district. White districs are those where the insurer has no policyholders and therefore also no exposure to the risk of filing a claim. Light blue districts represent the lowest 20% of total exposure and dark blue districts represent the highest 20% of total exposure. The other districts therefore contain an 'average' amount of exposure. The south-east part of Belgium, the Ardennes, does not contain a lot of exposure because this area is only sparsely populated. A lot of exposure is present in some of the big cities of Belgium; Brussels, Liège, Namur and Charleroi.

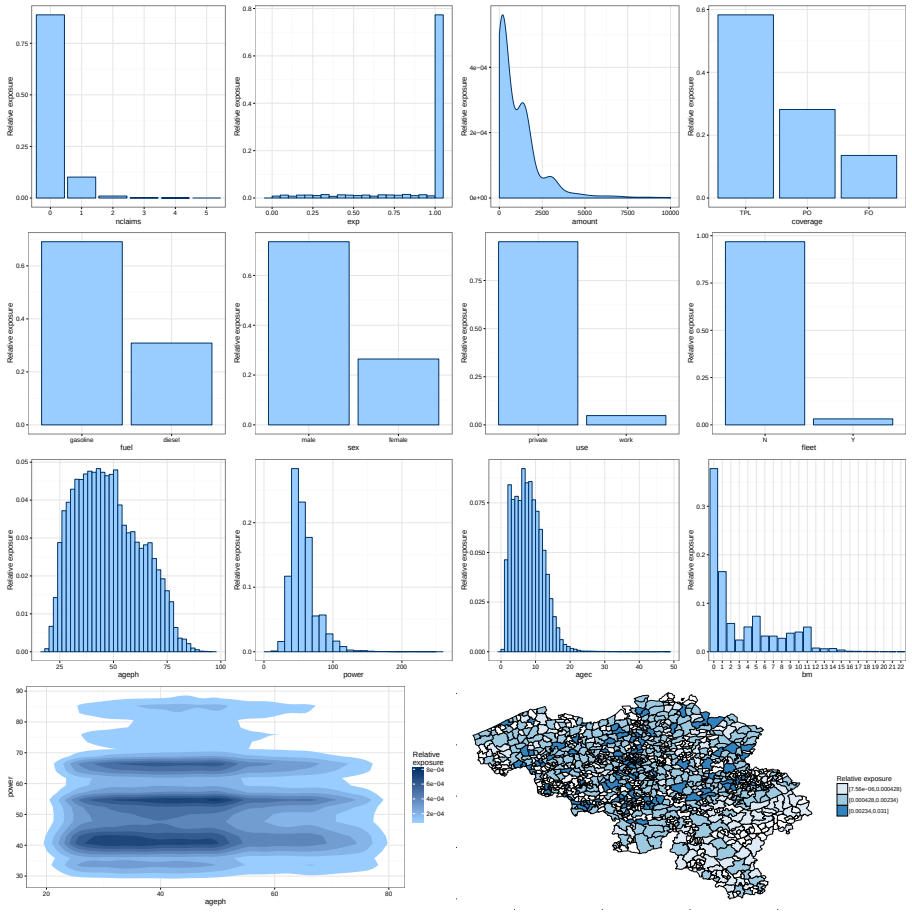

Figure 1: `MTPL`: relative exposure of the risk factors in Table 1. Row 1: `nclaims`, `exp`, `amount` and `coverage`. Row 2: `fuel`, `sex`, `use` and `fleet`. Row 3: `ageph`, `power`, `agec` and `bm`. Row 4: interaction between `ageph` and `power` and a map of Belgium with exposure per district.

3 Flexible models for P&C ratemaking

The pure premium for policyholder i ($\mathbb{E}[P_i]$) is calculated as the product of the expected values of the frequency ($\mathbb{E}[F_i]$) and the severity ($\mathbb{E}[S_i]$). The frequency for policyholder i (F_i) is defined as the ratio of the number of claims (N_i) and the exposure (e_i). The severity for policyholder i (S_i) is defined as the ratio of the total claim amount (L_i) and the number of claims (N_i). The pure premium can therefore be decomposed as follows:

$$\mathbb{E}[P_i] = \mathbb{E}[F_i] \times \mathbb{E}[S_i] = \mathbb{E}[\frac{N_i}{e_i}] \times \mathbb{E}[\frac{L_i}{N_i}]. \tag{1}$$

GAMs are the preferred tool for actuarial regression modelling from a statistical point of view. These models allow for the incorporation of flexible smooth effects of the continuous single, interaction and spatial elements. The general formulation of a GAM is as follows:

$$\eta(\mathbb{E}(X_i)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}^d + \sum_{j=1}^{q} f_j(x_{ij}^c) + \sum_{j=1}^{r} f_j(x_{ij}^c, y_{ij}^c) \tag{2}$$

where the response X follows a particular distribution from the exponential family. A link function η is used to relate the explanatory variables to the response. The 0/1-valued dummy variables x^d represent the typical way to code categorical risk factors in the GLM framework (an n-level categorical risk factor has 1 reference level and $n-1$ dummy levels). The accompanying coefficients β_j capture the additional effect on the response for the dummy levels (with respect to the reference level). GAMs extend GLMs by including smooth functions of continuous risk factors. Single effects are captured by the univariate smooth functions $f(x^c)$, while interaction and spatial effects are captured by bivariate smooth functions $f(x^c, y^c)$. The intercept β_0 captures the risk for policyholders with their categorical risk factors in the reference levels and the smooth functions evaluating to zero.

We fit the GAMs to the claims data following a two-step approach (for computational reasons). The first step performs an exhaustive search for the optimal GAM without taking into account interactions between the risk factors. The second step adds interactions between continuous risk factors to the resulting model from step one and performs an exhaustive search for the optimal GAM including those interactions. A measure has to be chosen to be able to compare different GAMs in the search for the optimal model. The Akaike information criterium (AIC) and Bayesian information criterion (BIC) are often used to compare non-nested models. Both are defined as follows:

$$\begin{aligned} \text{AIC} &= -2 \cdot \log \mathcal{L} + 2 \cdot k \\ \text{BIC} &= -2 \cdot \log \mathcal{L} + \log(n) \cdot k \end{aligned} \tag{3}$$

where $\log \mathcal{L}$ is the log-likelihood of the model, k is the number of parameters of the model and n is the number of observations in the data set. Both AIC and BIC measure the goodness of fit by minus two times the log-likelihood supplemented with a complexity penalty. The BIC penalty is more severe compared to the AIC penalty and BIC will therefore favor less complex models compared to AIC. We continue our study with the BIC as performance criterion in order to obtain compact GAMs on which the binning methods can be demonstrated in the following sections.

We use `R` and the `mgcv` package developed by Wood (2006) to fit our flexible GAMs. This package makes use of penalized maximum likelihood estimation in order to avoid overfitting the functions f from Equation (2). We use penalized thin plate regression splines to represent the smooth functions f, as described in Wood (2003). Thin plate regression splines are low rank approximations of the natural thin plate splines of Duchon (1977). Thin plate splines in n dimensions are defined as follows (with $\mathbf{x}$ an n-dimensional vector):

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \sum_{k=1}^{K} \alpha_k \|\mathbf{x} - \mathbf{x}_k\|^2 \log(\|\mathbf{x} - \mathbf{x}_k\|). \tag{4}$$

The level of smoothness of the splines is controlled by a smoothing parameter. This parameter will make a trade-off between penalizing a bad fit to the data and penalizing the 'wiggliness' of a function. Generalized Cross Validation (GCV) is used to automatically estimate the smoothing parameters, as explained in Craven and Wahba (1978).

3.1 Frequency

In this section we focus on developing a flexible regression model for claim frequencies. The goal is to explain the number of claims *nclaims*, given exposure *exp*, based on different types of risk factors. Starting point is a Poisson GAM which includes the categorical risk factors `coverage`, `fuel`, `sex`, `use` and `fleet` together with single effects of the continuous risk factors `ageph`, `power`, `agec` and `bm` and a spatial effect based on `long` and `lat`. This full GAM without interactions is displayed in Equation (5):

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \beta_3 \texttt{fuel}_{diesel} + \\ &\beta_4 \texttt{sex}_{female} + \beta_5 \texttt{use}_{work} + \beta_6 \texttt{fleet}_{Y} + f_1(\texttt{ageph}) + f_2(\texttt{power}) + \\ &f_3(\texttt{agec}) + f_4(\texttt{bm}) + f_5(\texttt{long}, \texttt{lat}).\end{aligned} \tag{5}$$

A Poisson distribution is assumed for the number of claims and a logarithmic link function is used. The logarithm of exposure is included in the model as an offset, such that the number of claims reported is expressed per unit of time exposure. The five categorical risk factors are coded with dummy variables by taking the level with the most exposure as reference level ($\texttt{coverage}_{TPL}$, $\texttt{fuel}_{gasoline}$, $\texttt{sex}_{male}$, $\texttt{use}_{private}$ and $\texttt{fleet}_{N}$). The continuous risk factors and spatial effect are included through non-parametric smooth functions. The functions f_1, f_2, f_3 and f_4 are univariate smooth effects of continuous risk factors. The spatial effect, f_5, is a bivariate smooth function of the latitude and longitude coordinates .

An exhaustive search over all the possible combinations of explanatory variables is performed in order to find the best GAM fit. The full model in Equation (5) contains 10 risk factors (5 categorical, 4 continuous and 1 spatial). All 1024 different models that can be formed by including or excluding these 10 risk factors are evaluated. This operation takes approximately 20 hours on a MacBook Pro with a 2.7 GHz Intel Core i5 processor and 8 GB 1867 MHz DDR3 RAM. The model with the lowest BIC value of all 1024 investigated models is the one displayed in Equation (6). Two categorical variables (`coverage` and `fuel`), three continuous risk factors (`ageph`, `power` and `bm`) and the spatial effect are included in the optimal model.

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \beta_3 \texttt{fuel}_{diesel} + \\ &f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{bm}) + f_4(\texttt{long}, \texttt{lat}).\end{aligned} \tag{6}$$

We now investigate whether the model in Equation (6) can further be improved by adding interaction effects between the continuous risk factors. Interaction effects are not considered in the studies of Denuit and Lang (2004) and Klein et al. (2014). We want to define the interaction effects as corrections on the single effects and therefore the possible interactions that can be supplemented to Equation (6) are `ageph-power`, `ageph-bm` and `power-bm`. Incorporating the interaction effect `ageph-power` shows a slight decrease in the BIC compared to the model without the interaction effect. The final model is displayed in Equation (7) and incorporates the interaction between `ageph` and `power` next to the effects already included in Equation (6).

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \beta_3 \texttt{fuel}_{diesel} + \\ &f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{bm}) + f_4(\texttt{ageph}, \texttt{power}) + f_5(\texttt{long}, \texttt{lat}).\end{aligned} \tag{7}$$

Figure 2 displays the five fitted smooth functions ($\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$ and $\hat{f}_5(\texttt{long}, \texttt{lat})$) from Equation (7). The top row shows the smooth effects for the risk factors `ageph`, `power` and `bm` in solid lines. The dashed lines represent the 97.5% pointwise confidence intervals, which are wider in regions with scarce data. Young policyholders are clearly very risky drivers. This riskiness decreases over increasing ages and stabilizes around the age of 35. It even increases slightly going from age 45 to age 50. This observation can be explained by the fact that children of policyholders in their late 40s - early 50s start to drive with their parents' car. After age 50 the riskiness decreases again until the age of 70, after which it starts increasing again. This implies that seniors start to have more car accidents once they grow older. The smooth effect for `power` shows a steep increase over the interval from 0 to 50 kilowatt and a more gradual increase from 50 kilowatt onwards. This implies that policyholders driving a more powerful vehicle are more likely to report a claim. The smooth effect for `bm` shows a steady increase

over increasing bonus-malus levels. This effect is in line with our intuition, since policyholders with high bonus-malus levels have worse claim histories compared to policyholders with low bonus-malus levels.

The interaction effect between `ageph` and `power` is displayed in the bottom left graph of Figure 2. This interaction effect can be seen as a correction on top of the single effects of both `ageph` and `power`. A negative (positive) correction, coloured in light blue (dark blue), indicates that the combined single effects of `ageph` and `power` overestimate (underestimate) the annual expected claim frequency. The combinations low `ageph` - low `power` and high `ageph` - high `power` receive a negative correction and are therefore less risky than the two single effects predict. The combinations high `ageph` - low `power` and low `ageph` - high `power` receive a positive correction and are therefore more risky than the two single effects predict. The results of our preferred GAM show for example that young policyholders driving a more powerful car imply a high risk for the insurer, at least in terms of the claim frequency.

The spatial effect is displayed in the bottom right graph of Figure 2. Note that this map does not indicate the districts where the car accidents took place, but the districts where the policyholder who had a car accident resides. Although a policyholder can have an accident in any other district than where he resides, we can assume that he will drive quite often in his own district. Further on, this district can be interpreted as a proxy for socio-economical characteristics that define the neighborhood where the policyholder resides. The districts are colour coded where light blue (dark blue) indicates a district where policyholders reside which have, on average, few (many) car accidents. The region around Brussels, in the center of Belgium, is clearly the most dangerous area to live and drive a car. Traffic is very dense in this area, which is reflected in a higher expected annual claim frequency for policyholders who live here. Another risky area in Belgium is the region around the city of Liège, in the eastern part of Belgium. The south-eastern, north-eastern and western parts of Belgium are safer regions to live and drive a car. These areas are less densely populated, which is reflected in a lower expected annual claim frequency for policyholders who live here.

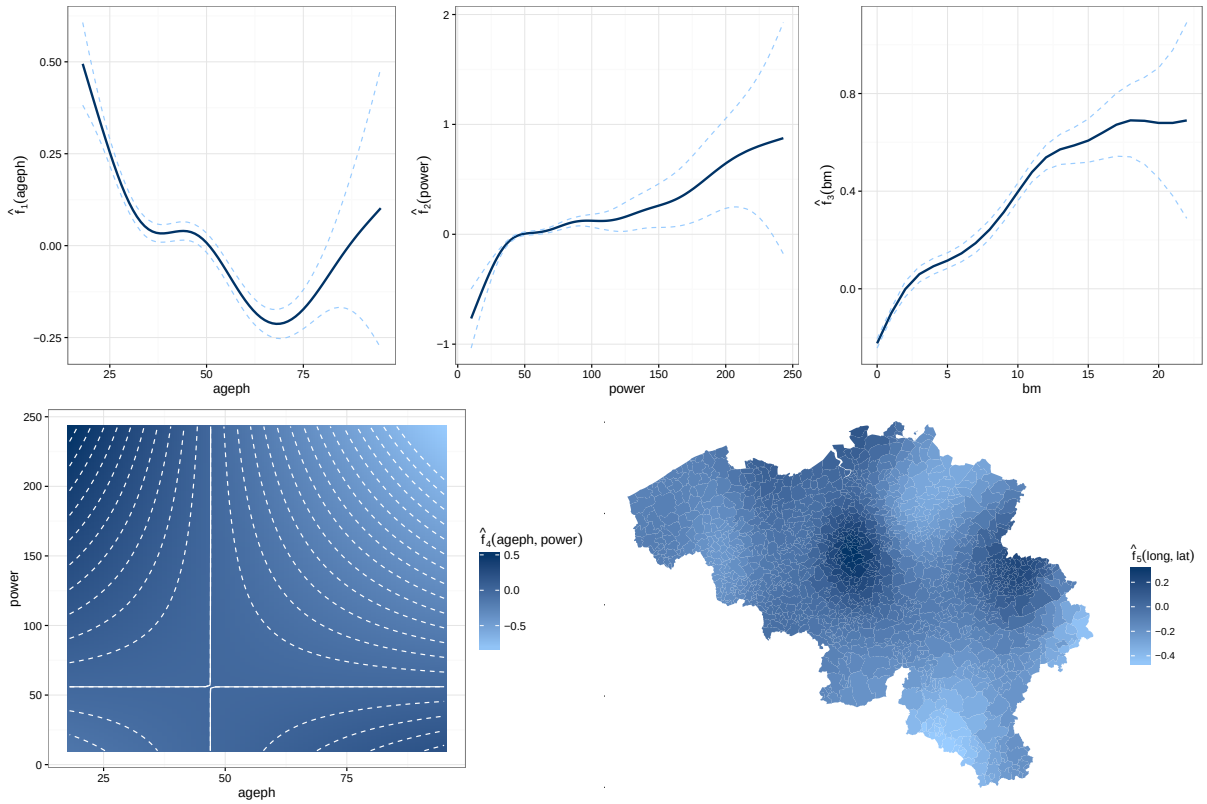

Figure 2: `MTPL`: fitted smooth GAM effects from Equation (7). Top row: single effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$ and $\hat{f}_3(\texttt{bm})$. Bottom row: interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$ and spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$.

3.2 Severity

In this section we focus on developing a flexible regression model for claim severities. The goal is to explain the average cost of a claim (`avg`), which is defined as the ratio of the total amount (`amount`) and the number of claims (`nclaims`), based on different types of risk factors. We follow the same fitting procedure as with the claim frequency by fitting an optimal model without interactions and investigating afterwards whether interaction effects can improve that fit. We can only use data of policyholders that actually filled a claim in this fitting procedure, which accounts for 18276 policyholders in our `MTPL` data set. The final model is the lognormal GAM in Equation (8):

$$\mathbb{E}(\log(\texttt{avg})) = \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + g_1(\texttt{ageph}) + g_2(\texttt{bm}). \tag{8}$$

A Gaussian distribution is assumed for the response $\log(\texttt{avg})$, such that the average amount of a claim follows a lognormal distribution. Only one categorical risk factor, `coverage`, and two continuous risk factors, `ageph` and `bm`, are selected in the optimal model. No interaction effects or spatial effect are present in this optimal model. It is a well known fact in actuarial modelling that usually less covariates are able to explain claim severities compared to frequencies.

Figure 3 displays the two fitted smooth functions ($\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$) from Equation (8). Going from ages 18 to 35, we can observe a decrease of the average claim cost. This might indicate that young drivers over this range are involved in more severe car accidents. The average claim cost starts to increase again for policyholders older than 35, stabilizes in the age interval 45 to 60 after which it starts to increase again. This might be explained by the fact that older policyholders drive more expensive cars and repairing costs increase. The average claim cost increases over increasing bonus malus levels. There is however a stabilizing region around level 5 and the average claim cost decreases from bonus malus level 13 onwards. Because of the scarceness of data for the high bonus malus levels one can not conclude much about this region (note the widening confidence bounds).

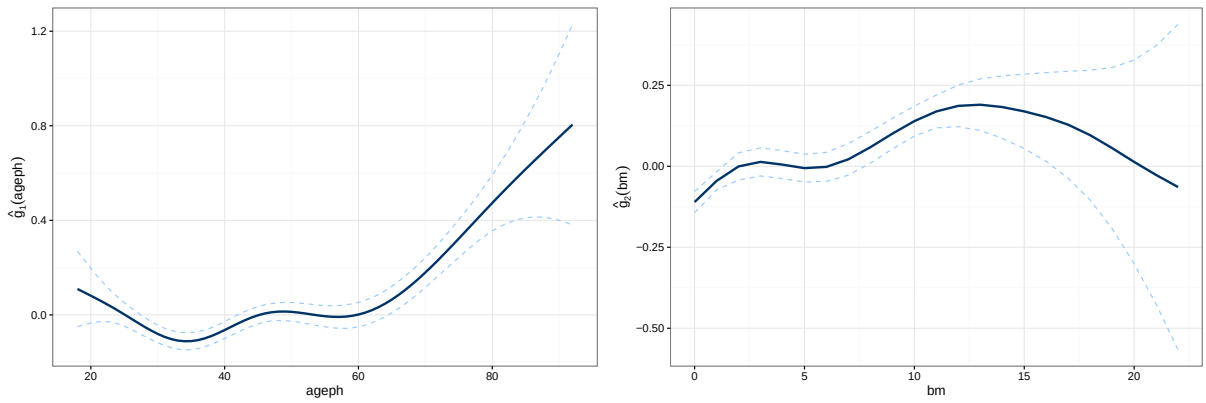

Figure 3: `MTPL`: fitted smooth GAM effects ($\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$) from Equation (8).

4 Binning methods for the smooth GAM effects

The GAMs in Equations (7) and (8) are optimal models, according to BIC, for claim frequency and severity in the `MTPL` data set. These models offer a high degree of flexibility for the continuous risk factors and spatial effect, which is very good from a statistical point of view. In practice, for marketing and ICT purposes, insurers typically prefer to have a model with a discrete number of classes (or: bins) for every risk factor. Such a pricing model is also more easy to explain to the regulator[2]. In this section we will describe a fully data driven approach to bin the continuous risk factors and spatial effect from Equations (7) and (8) in a discrete number of bins. Once all the risk factors are categorical, a GLM can be estimated where the risk factors are coded by dummy variables.

4.1 Spatial effect

We first put focus on binning the fitted continuous spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), in a discrete number of bins. We use the `classInt` package in `R`, developed by Bivand (2015), to compare four different methods to bin $\hat{f}_5(\texttt{long}, \texttt{lat})$ in non-overlapping intervals:

- **Equal intervals** The data range is divided in k parts of equal length $\frac{\max - \min}{k}$, where max and min respectively indicate the maximum and minumum value of the data being binned. This approach can give good results for uniformly distributed data, but tends to perform poorly for skewed data.
- **Quantile binning** Intervals containing approximately the same number of observations are created such that each bin will contain approximately $\frac{n}{k}$ elements where n indicates the total number of data points and k the number of classes. This method is often the default in statistical software packages, though it can give very misleading results. Similar observations can be assigned to different bins in order to make sure that each bin contains the same number of observations.
- **Complete linkage** This method, as described in Kaufman and Rousseeuw (2009), performs agglomerative hierarchical clustering where initially each data point forms its own bin. In every iteration the two bins closest to each other are merged where the distance between bins is defined as the distance of the two observations that are farthest away from each other: distance between bins A and B is calculated as $d(A, B) = \max_{a \in A, b \in B}(a, b)$. Bins with a remote pair of observations will only be merged in a late stage of the iteration process. This approach therefore results in bins with small within-bin variance.
- **Fisher's natural breaks** This iterative algorithm, developed by Fisher (1958) and discussed in Slocum et al. (2005), maximizes the homogeneity within bins. Bins are created such that every observation is as close as possible to the mean of its bin. This is done by minimizing the total distance between data points and the means of their respective bins. Formally, this can be expressed as minimizing the sum of squared distances between observations x_{ij} and their respective bin means $\bar{x}_j$: $\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$ where j runs over the different bins and i runs over the observations within each bin (k bins and n_j observations within bin j). This approach minimizes the within-bin variance and can therefore be expected to give good results.

We compare the results of the different binning methods using two measures: goodness of variance fit (GVF) and tabular accuracy index (TAI). The GVF and TAI are defined as follows (see Armstrong et al. (2003)):

$$\text{GVF} = 1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9}$$

$$\text{TAI} = 1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} |x_{ij} - \bar{x}_j|}{\sum_{i=1}^{n} |x_i - \bar{x}|} \tag{10}$$

where k, n_j and n indicate the number of bins, the number of districts in bin j and the total number of districts. In our setting the variable x indicates the spatial effect of the districts, namely $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7). The denominator of the fraction measures the deviation of each district from the global average. The numerator of the fraction measures the deviation of each district from the bin

[2] insurers are obliged to disclose their tarification structure (link naar wet)

average. The fraction will have a small value when the variance within the bins is small compared to the global variance. For both measures a value closer to 1 therefore indicates a more homogeneous binning of the data.

Table 2 compares the performance of the four binning methods when $k = 5$ bins are created (why five bins are chosen is explained further on). This table shows that Fisher's natural breaks method outperforms the other three methods. Figure 4 shows the empirical cumulative distribution function of the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), in combination with the bins produced by the different methods. Figure 5 shows the groupings of districts produced by the different methods on a map of Belgium. The different methods clearly result in very different groupings of districts. The method with intervals of equal length is performing rather well regarding its simplicity. Quantile binning assigns too many districts to both the most positive and most negative bins. Surprisingly, the complete linkage method performs worst of all four.

	GVF	TAI
Equal	0.913	0.675
Quantile	0.894	0.694
Complete	0.892	0.657
Fisher	**0.927**	**0.724**

Table 2: `MTPL`: The GVF and TAI for the four different methods to bin the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), into five bins.

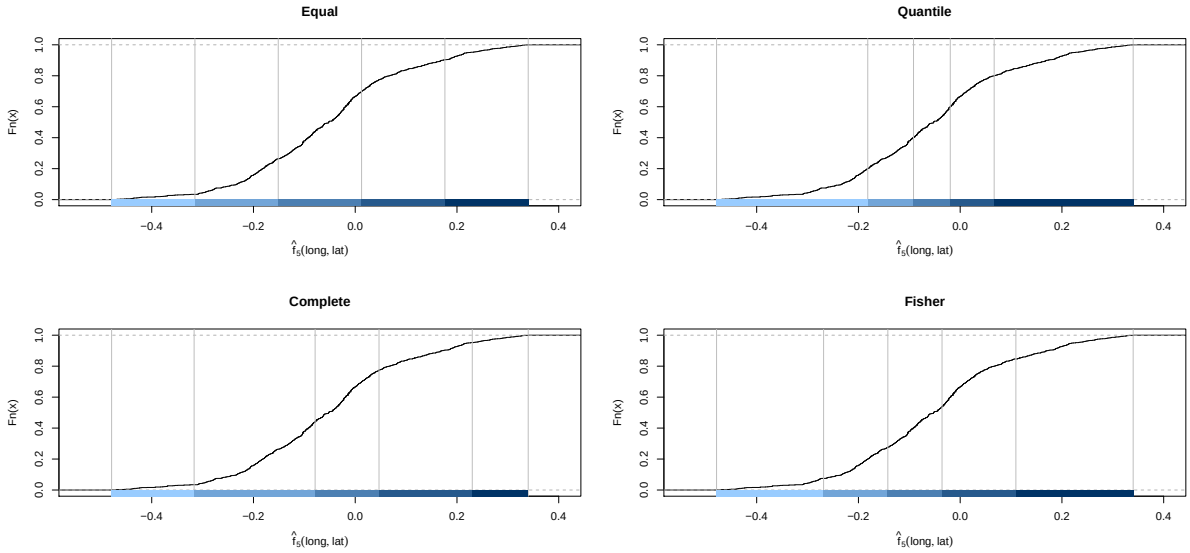

Figure 4: `MTPL`: the empirical cumulative density function of the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), in combination with the bins produced by the four different binning methods.

As motivated in Section 1, pricing actuaries ultimately prefer a GLM where all types of risk factors are coded with binary dummy effects. Therefore, we propose to choose the optimal number of bins for the spatial effect by focussing on the resulting GLM. We tune the number of bins by considering a set {2,3,4,5,6,7} for the possible number of spatial bins. For each value in this grid the procedure listed in Table 3 is applied. The results of this procedure are displayed in Table 4. Choosing five bins results in the lowest BIC and AIC for the refitted GAM. For the frequency model we continue with a GAM which specifies the spatial effect as a categorical risk factor, as specified in Equation (11).

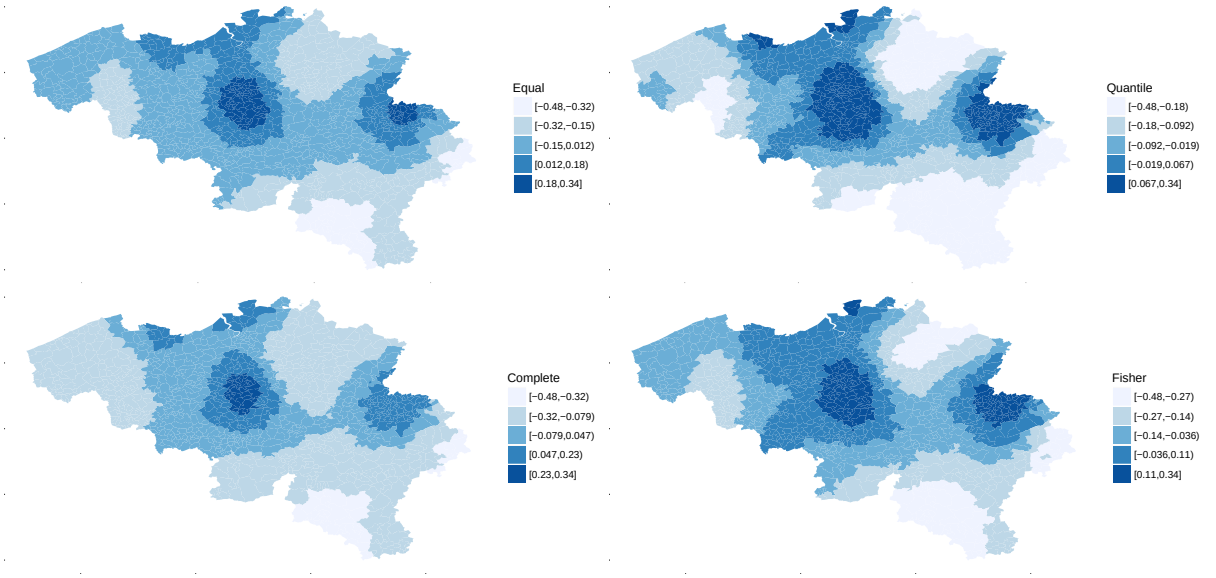

Figure 5: `MTPL`: maps of Belgium with the districts grouped into five distinct bins based on the intervals produced by the four different binning methods for the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7).

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \beta_3 \texttt{fuel}_{diesel} + \\ &\beta_4 \texttt{geo}_{[-0.48,-0.27)} + \beta_5 \texttt{geo}_{[-0.27,-0.14)} + \beta_6 \texttt{geo}_{[-0.036,0.11)} + \beta_7 \texttt{geo}_{[0.11,0.34]} + \\ &f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{bm}) + f_4(\texttt{ageph}, \texttt{power}).\end{aligned} \tag{11}$$

Procedure:	Find the optimal number of bins for the spatial effect
Step 1	Apply Fisher's algorithm to calculate the bin intervals for the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), where the number of bins is chosen equal to the current value of the predefined set of values. These bin intervals are used to transform the continuous spatial effect into a categorical spatial effect.
Step 2	Estimate a new GAM where we use a predictor structure that is similar to Equation (11).
Step 3	Calculate the BIC and AIC of the newly fitted GAM.

Table 3: Procedure to find the optimal number of bins for the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7), via the newly estimated GAM.

# bins	BIC	AIC
2	125047.6	124778.9
3	125023.9	124753.1
4	124928.4	124652.3
5	**124907.2**	**124621.3**
6	124921.6	124627.7
7	124942.9	124639.1

Table 4: `MTPL`: BIC and AIC of the newly fitted GAM after binning the spatial effect via Fisher's algorithm for different number of bins. The lowest BIC and AIC are attained for 5 bins.

4.2 Continuous risk factors

We now put focus on binning the continuous risk factors: the single effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$ and the interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$ in Equation (11) and the single effects $\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$ in Equation (8). We want to create bins where consecutive values of a risk factor are grouped together. The approach followed for binning the spatial effect is therefore no longer appropriate, since it might create a bin where policyholders from age 30 and age 80 are grouped together without that bin also containing all the intermediary ages. We propose decision trees as a technique to perform the binning since these models produce intuitive splits in the data in line with our requirement. We use evolutionary trees from the `R` package `evtree`, developed by Grubinger et al. (2014), which combine the framework of regression trees with genetic algorithms. Classical regression tree methods, as discussed in Breiman et al. (1984), are recursive partitioning methods that fit a tree in a forward stepwise search. Splits are chosen to maximize homogeneity at every step and these splits are kept fixed in all the following steps. This is an efficient heuristic, but the results are only locally optimal. The big advantage of evolutionary trees over recursive trees is the fact that earlier splits can still be adapted in a later stage of the fitting procedure. Thanks to this extra flexibility, evolutionary trees are capable of finding a global optimum.

The evolutionary tree algorithm follows a process of natural selection. An initial population (of 100 trees by default) is set up by choosing a random split in every single tree. In each iteration, all the trees from the current population, the parents, are altered by one of the following operations to produce children: split, prune, mutate or crossover. A split, prune or mutate operation respectively adds, removes or alters a split. The crossover operation exchanges subtrees between two individuals of the population. Every parent has to compete with its respective child in order to survive in the population and the surviving tree becomes a parent in the following iteration. This implies that the population size remains constant during the algorithm. The evaluation function to measure the performance of a tree has the following form:

$$N \cdot \log(\text{MSE}) \; + \; 4 \cdot \alpha \cdot (M+1) \cdot \log(N) \tag{12}$$

where N is the number of observations, M is the number of leaf nodes in a tree and α is a tuning parameter. The first part measures the accuracy of the tree by means of the mean squared error (MSE) while the second term represents a complexity penalty in terms of the size of the tree. A tuning parameter, α, is intoduced to make the trade-off between accuracy and complexity. Choosing $\alpha = 0.25$ results in an evaluation function that is equivalent to the BIC used by Fan and Gray (2005). The fitting process terminates when the quality of the top 5% performing trees in the population stabilizes for 100 iterations, but only after a minimum of 1000 iterations in total. The best performing tree, according to the evaluation function in Equation (12), is chosen as final model.

The data that serves as input to the evolutionary trees are the single and interaction effects from the GAMs in Equations (11) and (8): $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$, $\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$. Six evolutionary trees are developed, one for each GAM effect. It is important to take the composition of the insurance portfolio into account when deciding where to split the risk factors. Consider for instance policyholders older than 75; the smooth effect $\hat{f}_1(\texttt{ageph})$ in Figure 2 is strongly increasing for these ages, but Figure 1 indicates that the portfolio does not contain many policyholders aged over 75. It is therefore not desirable to obtain a lot of splits in this region, because it will only affect a very small portion of the entire portfolio. The number of policyholders in the portfolio who possess a specific risk factor is taken into account as a weight for that factor. There are for example 393 policyholders with an age of 20 in the portfolio. The smooth effect for $\texttt{ageph} = 20$, $\hat{f}_1(20)$, will therefore obtain a weight of 393. This implies that the mean squared error in Equation (12) will be calculated as a weighted mean squared error. As such, the evaluation criterion takes the composition of the portfolio into account when deciding which trees perform best. One extra constraint is imposed to make sure that bins are not too sparsely populated: each bin should at least contain 5% of the policyholders in the entire portfolio. This parameter gives insurers some flexibility to decide how coarse the bins should be.

The number of observations N in Equation (12) is equal to the sum of all the weights. This implies that $N = 163231$ for the four trees that bin the frequency effects ($\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$), since this is the total number of observations in the `MTPL` data set. The evaluation function in Equation (12) is therefore comparable over the four frequency trees and the same tuning parameter α can be used in each tree. Only 18276 policyholder file a claim in the `MTPL` data set such

that $N = 18276$ for the two trees that bin the severity effects ($\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$). This makes the evaluation function comparable over the two severity trees and the same α can be used in each tree. This implies that we have two tuning parameters which can be optimized independently of each other: one for the four frequency trees (α_{freq}) and one for the two severity trees (α_{sev}). Tuning these α's follows the same approach as tuning the number of bins for the spatial effect in Section 4.1. We consider a set $\{1{,}1.5{,}\ldots{,}9.5\} \times \{1{,}10{,}100\}$ as the possible values for α_{freq} and α_{sev}. For each value the procedure listed in Table 5 is applied. Note that the AIC measure is chosen over BIC since it uses a softer penalty and therefore allows for more splits in the trees. This is in line with commercial needs since insurance companies, driven by competition in the market, try to differentiate premiums. Figure 6 shows the splits produced by the evolutionary trees that deliver the best GLM fits ($\alpha_{freq} = 550$ and $\alpha_{sev} = 70$).

Procedure:	Find the optimal tuning parameters α_{freq} and α_{sev} for the evolutionary trees
Step 1	Fit an evolutionary tree to every single and interaction effect, $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$, $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$, where α is chosen equal to the current value of the predefined set of values. The splits produced by these trees are used to transform the continuous single and interaction effects into categorical effects.
Step 2	Estimate a frequency and severity GLM with the resulting categorical risk factors from the frequency and severity trees respectively.
Step 3	Calculate the AIC of the frequency GLM and the severity GLM.

Table 5: Procedure to find the optimal tuning parameters α_{freq} and α_{sev} to bin the single and interaction effects, $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$, $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$, via the resulting GLM.

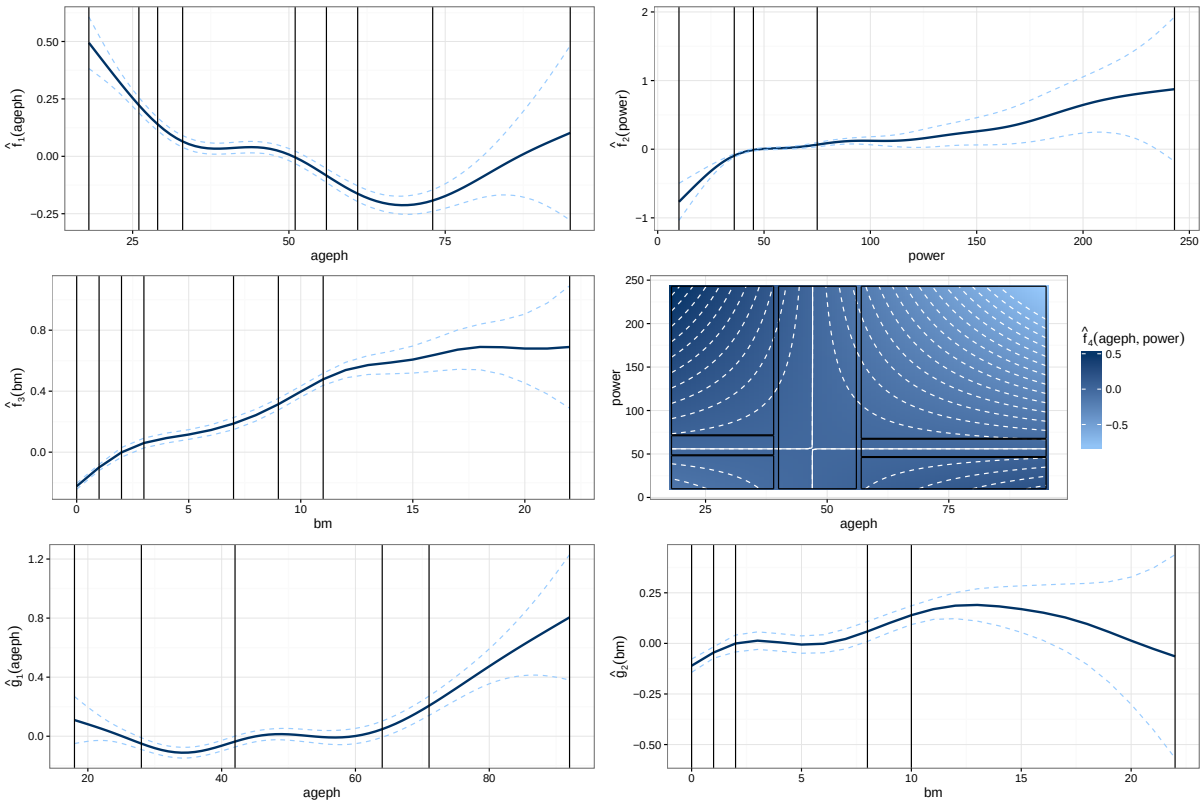

Figure 6: `MTPL`: Top and middle row: binning intervals for the continuous single and interaction effects, $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$ and $\hat{f}_4(\texttt{ageph}, \texttt{power})$, obtained by the evolutionary trees with $\alpha_{freq} = 550$. Bottom row: binning intervals for the continuous single effects, $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$, obtained by the evolutionary trees with $\alpha_{sev} = 70$.

The single effect $\hat{f}_1(\texttt{ageph})$ is split into eight bins. All policyholders with an age in the interval $[33, 51)$ are grouped together since $\hat{f}_1(\texttt{ageph})$ is rather flat over this interval. Younger policyholders have a higher risk profile and three bins are formed for policyholders younger than 33: $[18, 26)$, $[26, 29)$ and $[29, 33)$. The first bin is chosen wider because there are very few young policyholders present in the portfolio. The smooth effect $\hat{f}_1(\texttt{ageph})$ decreases for policyholders aged over 51. These policyholders typically have more driving experience and therefore a lower risk profile. Two bins are created in this region: $[51, 56)$ and $[56, 61)$. The smooth effect $\hat{f}_1(\texttt{ageph})$ stabilizes after age 61 before it starts increasing again for senior policyholders. This results in two bins; $[61, 73)$ for the stabilizing region and $[73, 95]$ for the senior policyholders with a higher risk profile.

The single effect $\hat{f}_2(\texttt{power})$ is split into four bins. The first bin, $[10, 36)$, groups policyholders driving less powerful cars. Even though $\hat{f}_2(\texttt{power})$ is increasing over this region, there are not enough policyholders in the portfolio to justify more splits. The next two bins, $[36, 45)$ and $[45, 75)$, group the bulk of policyholders in the portfolio. The seond of these two is chosen wider because $\hat{f}_2(\texttt{power})$ is more flat over this region. The last bin, $[75, 243]$, groups all policyholders driving powerful cars. This bin is so large because there is only a small amount of policyholders driving very powerful cars in the portfolio.

The single effect $\hat{f}_3(\texttt{bm})$ is split into seven bins. The first three bonus-malus levels end up in separate bins: $[0, 1)$, $[1, 2)$ and $[2, 3)$. The next bin, $[3, 7)$, is wider because $\hat{f}_3(\texttt{bm})$ increases less steeply over this range. The next two bins, $[7, 9)$ and $[9, 11)$, are less wide because $\hat{f}_3(\texttt{bm})$ starts to increase more steeply again over this region. The higher bonus-malus levels are grouped into one bin: $[11, 22]$. This bin is so wide because of two reasons: the slope of $\hat{f}_3(\texttt{bm})$ decreases for higher bonus-malus levels and only a few policyholders have such high bonus-malus levels.

The interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$ is split into seven bins. The solid white contour lines indicate where $\hat{f}_4(\texttt{ageph}, \texttt{power}) = 0$. These combinations of $\texttt{ageph}$ and $\texttt{power}$ therefore indicate a neutral region where the single effects of $\texttt{ageph}$ and $\texttt{power}$ give a full indication of the risk. One can see that our method detects these neutral regions. The vertical bin for $40 \leqslant \texttt{ageph} < 57$ indicates the neutral zone around the vertical contour line. Two horizontal bins, one for $49 \leqslant \texttt{power} < 72$ on the left and one for $47 \leqslant \texttt{power} < 68$ on the right, take the neutral zone around the horizontal contour line into account. Two bins represent policyholders with a lower risk profile than the neurtral zones: $\texttt{ageph} < 40, \texttt{power} < 49$ and $\texttt{ageph} \geqslant 57, \texttt{power} \geqslant 68$. The weighted averages of the interaction effect in these regions are respectively equal to -0.029 and -0.052. Two bins represent policyholders with a higher risk profile than the neurtral zones: $\texttt{ageph} < 40, \texttt{power} \geqslant 72$ and $\texttt{ageph} \geqslant 57, \texttt{power} < 47$. The weighted averages of the interaction effect in these regions are respectively equal to 0.047 and 0.039. This figure shows that our evolutionary tree approach is also good at discriminating between areas of increased and decreased risk in a two-dimensional setting. One might argue that the neutral zones are not necessary and only four bins are sufficient. These neutral zones are added because of the fact that there are a lot of policyholders in the portfolio in these regions. It is therefore important to assess the risk of these policyholders properly. The bottom left graph of Figure 1 shows the two-dimensional distribution of the number of policyholders over $\texttt{ageph}$ and $\texttt{power}$. This figure justifies why the neutral zone in the vertical direction is chosen as largest; most policyholders can be captured that way.

The single effect $\hat{g}_1(\texttt{ageph})$ is split into five bins. All policyholders with an age in the interval $[42, 64)$ are grouped together since $\hat{g}_1(\texttt{ageph})$ is rather flat over this interval. Younger policyholders have a lower risk profile and two bins are formed for policyholders younger than 42: $[18, 28)$, $[28, 42)$. The second bin is chosen wider because $\hat{g}_1(\texttt{ageph})$ stabilizes in this region. The smooth effect $\hat{g}_1(\texttt{ageph})$ increases for policyholders aged over 64. Two bins are created in this region: $[64, 71)$ and $[71, 95]$. The second bin is chosen wider because there are fewer policyholders in this range.

The single effect $\hat{g}_2(\texttt{power})$ is split into five bins. The first two bonus-malus levels end up in separate bins: $[0, 1)$ and $[1, 2)$. The bin $[2, 8)$ is wider because $\hat{g}_2(\texttt{bm})$ is rather flat over this range. The next bin, $[8, 10)$, is less wide because $\hat{g}_2(\texttt{bm})$ starts increasing again in this region. The higher bonus-malus levels are grouped into one bin: $[10, 22]$. This bin is so wide because of two reasons: $\hat{g}_2(\texttt{bm})$ stabilizes for these bonus-malus levels and only a few policyholders have such high bonus-malus levels.

5 Fitting GLMs with the binned risk factors

The previous section described how the spatial and continuous effects from the GAMs in Equation (11) and (8) can be transformed into categorical risk factors. We now fit a Poisson GLM (for the frequency) and a lognormal GLM (for the severity) to the claims data with the resulting categorical risk factors.

Figure 7 compares the original GAM effects with the resulting GLM coefficients. The first two rows show the comparison for the continuous single effects in the frequency and severity model respectively. The third and fourth row respectively show the comparison for the continuous interaction effect and spatial effect in the frequency model. Note that the GLM coefficients in the first two rows of Figure 7, indicated by purple dots, are a rescaled version of the actual coefficients. This rescaling is performed to bring the GAM effects and GLM coefficients on a comparable range. A GAM effect is estimated in such a way that the weighted mean, with the number of policyholders as weights, of the smooth effect is equal to zero. For the smooth effect of `ageph` for example this means that:

$$\frac{\sum_{i=1}^{K_{\texttt{ageph}}} n_{\texttt{ageph}_i} \hat{f}_1(\texttt{ageph}_i)}{\sum_{i=1}^{N_{\texttt{ageph}}} n_{\texttt{ageph}_i}} = 0 \tag{13}$$

with $K_{\texttt{ageph}}$ the number of different ages in the sample, $n_{\texttt{ageph}_i}$ the number of policyholders with $\texttt{ageph} = i$ and $\hat{f}_1(\texttt{ageph}_i)$ the GAM effect for a policyholder with $\texttt{ageph} = i$. This implies that the range for the GAM effects is always centered around zero. Within a specific categorical risk factor, the GLM coefficient for the reference class is equal to zero and the GLM coefficients of the other classes are expressed relative to this reference class. The range of the GLM coefficients therefore depends on the choice of the reference class within each categorical risk factor. We rescale the GLM coefficients such that the weighted mean of the rescaled GLM coefficients is equal to zero. We do this by computing the weighted mean of the original GLM coefficients per categorical risk factor, with the number of policyholders as weights, and then subtracting this value from the GLM coefficients for that specific categorical risk factor. This causes the rescaled GLM coefficients to have a range centered around zero, just like the GAM effects. This rescaling is of course only performed to compare the GAM effects and GLM coefficients with each other; the actual GLM model is not adjusted in any way.

The piecewise constant functions formed by the GLM coefficients in the first two rows of Figure 7 approximate the smooth GAM effects very closely. This indicates that our resulting GLMs are a good approximation of the original GAMs. We trade flexibility for simplicity and some mismatches are therefore impossible to avoid. For example: both $\hat{f}_1(\texttt{ageph})$ and $\hat{g}_1(\texttt{ageph})$ are underestimated by the GLM coefficients for the youngest and oldest policyholders. Other underestimations happen with $\hat{f}_2(\texttt{power})$ for high powered cars and $\hat{f}_3(\texttt{bm})$ for high bonus-malus levels. Two overestimations are visible: $\hat{f}_2(\texttt{power})$ for low powered cars and $\hat{g}_2(\texttt{bm})$ for high bonus-malus levels. Note however that these mismatches only happen in the extreme ends of the ranges of the risk factors, where the exposure in the portfolio is very low. It is therefore normal that the GLM coefficients for these bins are tuned more towards the policyholders that contribute most to the total exposure of that bin. The approximation of the GLM coefficients to the GAM effects is really good for bins with high exposure in the portfolio. The GLM coefficients for bins $[0, 1)$ and $[1, 2)$ approximate both $\hat{f}_3(\texttt{bm})$ and $\hat{g}_2(\texttt{bm})$ very well for example.

The third row of Figure 7 shows the GAM interaction effect between `ageph` and `power` in the frequency model together with the approximation by GLM coefficients. The vertical neutral zone is chosen as the reference class and has a GLM coefficient of zero. To the left/right of this neutral zone the GLM coefficients respectively increase/decrease for increasing `power`. We can therefore interpret the + region as a neutral zone with the top left and bottom right (bottom left and top right) indicating regions with increased (decreased) risk. The fourth row of Figure 7 shows the GAM estimate for the spatial effect in the frequency model together with the approximation by GLM coefficients. The fourth bin is chosen as reference class because it contains most exposure and it contains the zero GAM effect (see Figure 5). We obtain three bins with negative GLM coefficients and one bin with a positive GLM coefficient. As expected, we can observe a monotonic increase in the GLM coefficients for the spatial effect.

Table 6 shows a comparison of the AIC and BIC from the original GAM and resulting GLM. The GAMs attain a lower AIC for both the frequency and severity models, whereas the GLMs attain a lower BIC for both the frequency and severity models. This clearly shows the tade-off between flexibility and simplicity in the modelling process.

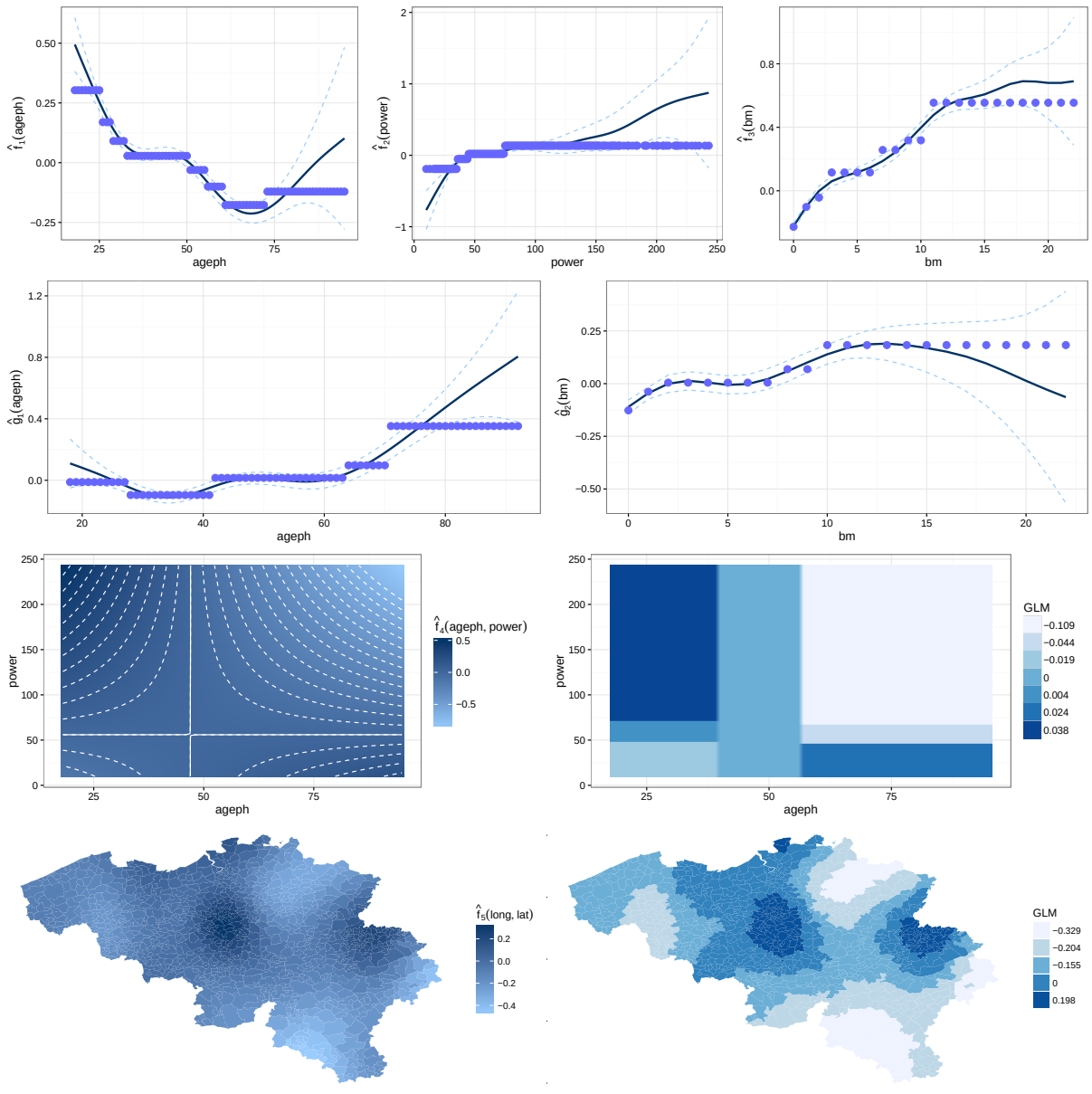

Figure 7: `MTPL`: First row: a comparison between the GAM effects and GLM coefficients for the risk factors `ageph`, `bm` and `power` in the frequency model. Second row: a comparison between the GAM effects and GLM coefficients for the risk factors `ageph` and `bm` in the severity model. Third row: a comparison between the GAM effect and GLM coefficients for the interaction between `ageph` and `power` in the frequency model. Fourth row: a comparison between the GAM effect and GLM coefficients for the spatial effect in the frequency model.

AIC	Frequency	Severity
GAM	**124630**	**65593**
GLM	124647	65603

BIC	Frequency	Severity
GAM	125121	65706
GLM	**124947**	**65696**

Table 6: `MTPL`: a performance comparison between the original GAMs and resulting GLMs via both the AIC and BIC measures. The frequency and severity GAMs attain lower AIC values whereas the GLMs attain lower BIC values.

6 Discussion and conclusions

This paper presents a fully data driven strategy to deal with geographical information and continuous risk factors in an insurance tariff. Both a model for claim frequencies and severities is developed in this paper. Combining both these models will allow the actuary to calculate a pure risk premium for an insurance contract. The proposed strategy is summarized in Table 7.

Procedure:	Strategy to deal with geographical information and continuous risk factors
Step 1	Fit the GAMs in Equation (7) and (8) to the `MTPL` data set for respectively the frequency and severity component.
Step 2	Apply Fisher's natural breaks algorithm to bin the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$, and refit the GAM with a categorical spatial effect. The number of bins is chosen according to the lowest AIC for the refitted GAM.
Step 3	Fit evolutionary trees to the single and interaction effects, $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$, $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$, and fit GLMs with the resulting categorical risk factors. The tuning parameters of the evolutionary trees, α_{freq} and α_{sev}, are chosen according to the lowest AIC of the fitted GLMs.
Step 4	Fit GLMs with the resulting categorical risk factors to the `MTPL` data set for respectively the frequency and severity component.

Table 7: Summary of the strategy proposed in this paper to deal with geographical information and continuous risk factors in an insurance tariff.

The first step in our approach is to fit flexible GAMs to the `MTPL` data set for both the frequency and severity component. GAMs are chosen because these models allow for smooth effects of continuous risk factors. These smooth effects indicate how the risk is distributed within every continuous risk factor, while smoothing out some irregularities that are present in the data. An exhaustive search over all possible models is performed and the GAMs in Equation (7) and (8) are chosen because these models result in the lowest BIC. Another measure could have been used to evaluate the fit of the GAMs, but BIC is chosen to obtain a good and compact model. Using AIC would lead to a more complex GAM, but this brings no added value to the demonstration of our strategy. If the goal is to develop a model for insurance ratemaking in practice, then it is advisable to use AIC in order to get a more detailed model.

The seond step in our approach is to bin the spatial effect present in the frequency model: $\hat{f}_5(\texttt{long}, \texttt{lat})$ from Equation (7). Four different binning methods are compared (equal intervals, quantile binning, complete linkage and Fisher's algorithm) and we observe that the different methods result in very different groupings of districts. The binning results of Fisher's natural breaks algorithm give the highest accuracy measures (GVF and TAI). This method minimizes the within-cluster variance while maximizing the between-cluster variance and can therefore be expected to perform better than more simple methods like quantile binning. This analysis shows that it is worthwhile to use advanced binning methods in order to assess the risk properly instead of using the simple default methods often supplied in statistical software packages. Once the spatial effect is binned, a new frequency GAM is estimated with a predictor structure of Equation (11). The number of bins for the spatial effect is chosen according to the lowest AIC of the newly estimated GAM. This choice is made because, at the end, it is our goal to develop the best possible GLM. Other approaches coud be followed, like for example trying to detect the number of clusters from the input data $\hat{f}_5(\texttt{long}, \texttt{lat})$ directly. By following such an approach, the number of bins for the spatial effect is no longer tuned towards developing the best possible GLM. Although our approach is more time consuming, since a GAM has to be estimated for every possible value of the number of bins considered, we believe it is worth the effort to obtain the best GLM in the end.

The third step in our approach is to bin the continuous single and interaction effects: $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$ from Equation (11) and $\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$ from Equation (8). Decision trees are proposed because these models give rise to consecutive bins. More specifically, evolutionary trees are used because these models offer much higher flexibility than the classical recursive partitioning trees. The trees are very good at binning the single effects. They focus on splitting the smooth effect in regions where it has a large (positive or negative) slope, but only if sufficient policyholders are present

in these regions. The composition of the current portfolio is taken into account and our approach will automatically adapt to changing portfolio compositions. The fact that the composition of the portfolio plays a role in choosing the splits is a major advantage of this approach. The trees also perform well at binning the interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$. The tree recognizes that it is important to define a neutral zone where a lot of policyhoders with very similar risk profiles are grouped. The regions containing policyholders with increased or decreased risk profiles are also very well delineated by the tree. A downside of using decision trees in two dimensions is the fact that these models can only produce splits parallel with the x- and y-axis. The resulting bins will therefore always have a rectangular shape. It is for example not possible to produce a split along a non-linear contour line in Figure 2. More complex models, like support vector machines (SVMs), are needed to obtain non-linear splits. The problem with such models however is interpretability; a non-linear split is much harder to explain than the straight splits produced by a decision tree. We choose to put interpretability first and keep on using the decision trees for binning the interaction effect.

The fourth and last step in our approach is to fit GLMs for the frequency and severity component with the categorical risk factors resulting from the previous steps. These GLMs originate from flexible GAMs of which the spatial effect and continuous risk factors are binned in a fully data driven way. No ad hoc decisions are made to develop these GLMs and all the parameters in the process, the number of bins for the spatial effect and the tuning parameters α_{freq} and α_{sev}, are tuned towards producing the best possible GLMs.

References

Antonio, K., Frees, E. W., and Valdez, E. A. (2010). A multilevel analysis of intercompany claim counts. *Astin Bulletin*, 40(01):151–177.

Antonio, K. and Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, 96(2):187–224.

Armstrong, M. P., Xiao, N., and Bennett, D. A. (2003). Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers*, 93(3):595–623.

Bivand, R. (2015). *classInt: Choose Univariate Class Intervals*. R package version 0.1-23.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305.

De Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press, Cambridge.

Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie*. Economica.

Denuit, M. and Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627–647.

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

Dougherty, J., Kohavi, R., Sahami, M., et al. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the twelfth international conference*, volume 12, pages 194–202.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Constructive Theory of Functions of Several Variables*, pages 85–100.

Fan, G. and Gray, J. B. (2005). Regression tree analysis using target. *Journal of Computational and Graphical Statistics*, 14(1):206–218.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798.

Frees, E. W., Derrig, R. A., and Meyers, G. (2014). Predictive modeling in actuarial science. *Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques*, page 1.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, 103(484):1457–1469.

Grubinger, T., Zeileis, A., and Pfeiffer, K.-P. (2014). evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software*, 61(1):1–29.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225.

Haberman, S. and Renshaw, A. (1997). Generalized linear models and actuarial science. *Insurance Mathematics and Economics*, 2(20):142.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. CRC Press.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

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

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. John Wiley & Sons.

Lemaire, J. (1995). Bonus-malus systems in automobile insurance. *Insurance Mathematics and Economics*, 3(16):277.

McClenahan, C. L. (1990). Ratemaking. *Encyclopedia of Actuarial Science*.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384.

Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer, Berlin.

Parodi, P. (2014). *Pricing in General Insurance*. CRC Press.

Slocum, T., McMaster, R., Kessler, F., and Howard, H. (2005). *Thematic cartography and geographic visualization*. Upper Saddle River, New Jersey: Pearson Prentice Hall, Second.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

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