B-KUL-D0N55A, Advanced Non-Life Insurance Mathematics, Home Assignment 2

Dieter Verbeke

December 2016

1 Introduction

This text covers the second assignment for the course Advanced Non-Life Insurance Mathematics. It addresses the construction of a tariff structure for a car insurance product. The approach taken here is completely analogous to the one in [1]. Since the problem at hand, as well as the available data are very similar, this way of dealing with the task is certainly reasonable.

2 Data

The data concerns a motor third party liability (MTPL) insurance portfolio with a Belgian insurer during one specific year. The data set comprises observations of 163657 policyholders. Table 1 lists the available risk factors.

2.1 Augmented data set

The original dataset contained only one continuous risk factor, namely `ageph`. However, the geographical factor was transformed to a continuous effect by relating the postal code to spatial coordinates, introducing two additional factors `long` and `lat`. This could easily be done by comparing the codes in the current data set with those in the data set of [1].

2.2 Exploratory data analysis

Figure 1 reveals how the risk factors in the `MTPL` data set are distributed.

The majority of the policyholders, more than 88% of them, are claim-free during the year of observation. Approximately 10% of the policyholders filed one claim and the remaining few percentage of policyholders filed two, three, four or five claims. Almost 80% of the policyholders have an exposure of 1, meaning they were covered by the insurance during the entire year. The exposure of the rest is scattered between 0 and 1.

Table 1: `MTPL` risk factors. The superscripts 1 and 2 denote respectively continuous and categorical factors. In bold are the two geographical coordinates that are derived from the postal code and added to the original dataset.

Value	**Description**
exp	Exposure: fraction of the year the insured was covered.
nclaims	Number of claims filed by the policyholder during the year.
amount	Total amount claimed by the policyholder.
ageph^1	Age of the policyholder
agec^2	Age of the vehicle.
sex^2	Gender of the policyholder.
fuel^2	Type of fuel of the vehicle: gasoline or diesel.
split^2	Split of the premium: montly, once, twice, three times per year.
use^2	Use of the car: private or professional.
fleet^2	Car belonging to a fleet: yes or no.
sport^2	Sports car: yes or no.
cover^2	Coverage: MTPL, MTPL+, MTPL+++.
power^2	Power class of the vehicle: < 66, 66-110, >110.
pc^2	Belgian postal code.
long^1	Longitude coordinate corresponding to the postal code.
lat^1	Latitude coordinate corresponding to the postal code.

The overall claim frequency of the portfolio, calculated as the total number of claims divided by the total exposure, is equal to 13.93 %. The average claim severity of the portfolio, calculated as the total amount of claims divided by the total number of claims, is equal to 1622 EUR.

The MTPL data set contains eight categorical risk factors: coverage, fuel, sex, use, fleet, age of policyholder, power and age of vehicle.

Approximately 60% of the policyholders have MTPL coverage. Only their liability with respect to a third party is covered. This type of coverage is mandatory in Belgium. The other policyholders have opted for additional protection through MTPL+ and MTPL+++ plans. The policyholder's vehicles are fueled by either gasoline (app. 70%) or diesel (app. 30%). There is an apparent gender imbalance of roughly 75% male versus 25% female policyholders. Nearly all of the policyholders (>95%) use their vehicle for private purposes rather than professional, and almost no vehicles are part of a fleet. Most vehicles, around 70%, fall in the lowest power class. Few of them have more power than 110 hp. Approximately 70% of the vehicles was between 2 and 10 years old.

The `MTPL` data set comprises one continuous risk factor, namely the age of the policyholder. The majority of policyholders are between 25 and 75 years old. The portfolio comprises few young and old people.

The data set contains geographical information in the form of postal codes. These were transformed to longitude and latitude coordinates prior to further processing.
The map of Belgium gives an idea of the total exposure in each district. White districts have no policyholders living in them and do not have any exposure to the risk of filing a claim. The color scale is such that darker blue signifies higher relative exposure. From the map it is evident that the south-east of Belgium has considerably less exposure than the more densely populated areas around the larger cities in the center and north of the country.

Figure 2 investigates the interaction between five categorical factors and the continuous age factor. These histograms allow some precarious observations. It seems that younger policyholders are more inclined to split premium payments. The average age of female policyholders appears to be lower than the average age of male policyholders. The average age of policyholders driving a sports car is ostensibly lower than the average age of policyholders driving a regular car. Though the latter is not entirely clear from the graph due to scaling.

3 Strategy

Three types of risk factors are considered here: categorical, continuous (only single, no interaction) and spatial effects. I followed the data driven strategy of [1] to handle continuous risk factors and geographical information. It consists of two stages.
Firstly, generalized additive models (GAMs) are fitted to the claims data. These models contain categorical risk factors, smooth effects of continuous risk factors and the spatial effect that captures geographical information. The goal is to bin the continuous risk factors and the spatial effect thereby transforming them into categorical risk factors. The spatial effect is binned by the Fisher-Jenks algorithm and the continuous risk factors are binned by evolutionary trees.
Secondly, the categorical factors are incorporated in generalized linear models (GLMs).

GAMs possess some favourable statistical properties, while GLMs are more suited to an insurance company's production environment. The data driven strategy bridges the gap between, on the one hand, GAM based tarification, and on the other hand, GLM based tarification with predefined bins. For details the reader is referred to aforementioned reference.

The ratemaking model has two components: a claim frequency model and a claim severity model. Frequency is defined as the number of claims per unit of exposure. The exposure measures how long the policyholder is exposed to the insured risk. In this case it is the fraction of the year for which coverage was provided. Severity is the average claim cost. It is calculated as the ratio of

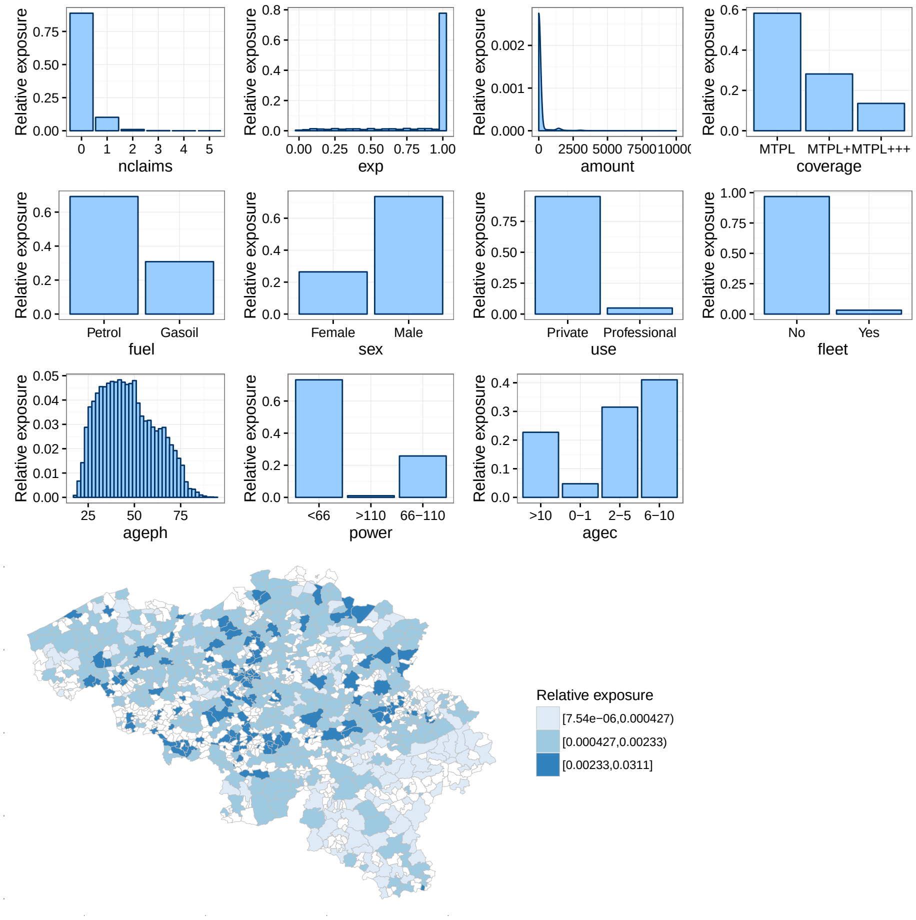


Figure 1: Relative exposure in the `MTPL` data set of the risk factors in Table 1.

the total loss and the number of claims (causing this loss) in the corresponding period.

The 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.

A Poisson distribution was considered for the claim frequency, and a log-normal distribution for the claim severity.

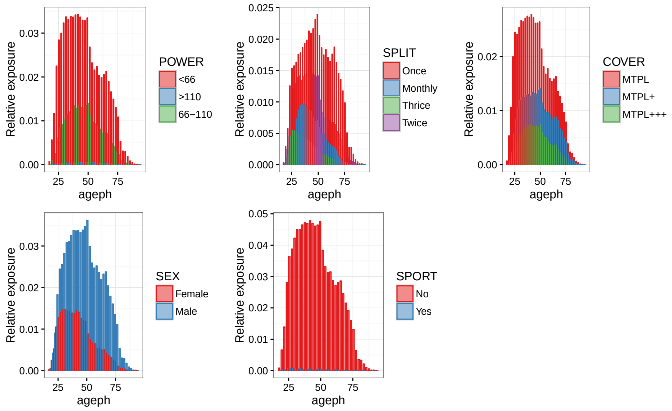


Figure 2: Interactions between five categorical factors and the continuous age factor.

4 Ratemaking model

4.1 Generalized additive model

From a statistical point of view, GAMs are the preferred tool for actuarial regression. The general formulation of a GAM is

$$\eta(E[X_i]) = \beta_0 + \sum_{j=1}^{p} B_j x_{ij}^d + \sum_{j=1}^{r} f_j(x_{ij}^c) + \sum_{j=1}^{r} f_j(x_{ij}^c, y_{ij}^c). \tag{1}$$

The response X follows a particular distribution from the exponential family. The link function η relates the explanatory variables to the response. The binary variables x^d encode categorical risk factors in the GLM framework. Single continuous risk factors 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 represents the risk for policyholders with their categorical risk factors in the reference levels and the smooth functions evaluating to zero.

4.2 Quality measure

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were chosen as quality measures to compare different GAMs/GLMs. Their respective definitions are

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

Table 2: Strategy from [1] to deal with geographical information and continuous risk factors in a motor third party liability (MTPL) insurance tariff

Procedure	Strategy to deal with geographical information and continuous risk factors
Step 1	Fit independent GAMs to the MTPL claims data for respectively the frequency and severity component.
Step 2	Apply Fisher's natural breaks algorithm to bin the spatial effect, and refit a GAM with a categorical spatial effect. The number of bins is chosen according to the lowest BIC for the refitted GAM.
Step 3	Fit evolutionary trees to the continuous single and interaction effects. Then fit GLMs with the resulting categorical risk factors. The tuning parameters of the independent evolutionary trees for frequency and severity, 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.

where $\log(\mathcal{L})$ is the log-likelihood of the model, k is the number of parameters of the model and n the number of observations. Compared to the AIC, the BIC favors less complex models.

4.3 Approach

In [1] a GAM is fitted to the claims data in two steps. In the first step, an exhaustive search for the optimal GAM is performed without taking into account interaction effects. In a second step, interactions between continuous risk factors are added to the outcome of the first step. Then again, an exhaustive search for the optimal GAM is performed.

4.4 Frequency model

Following the procedure explained above the starting point would be a GAM including all categorical risk factors, continuous risk factors and the spatial effect but excluding interactions. The full model would contain nine categorical risk factors, one continuous factor, and one spatial effect. That would require the evaluation of 2^{11} models formed by including or excluding these factors, and comparing their quality measures. Because of limited resources and the great similarity between data sets I took a different approach. I selected a reference model that comprises all factors that are in the optimal model of [1] and that are present in both data sets. These are `cover`, `fuel`, `ageph`, `power`, `long` and `lat`. Then I investigated whether the model could be improved by adding the remaining factors `sport` and `split`. The model with the lowest BIC value was eventually

$$\begin{aligned}\log(\mathbf{E}[\texttt{nclaims}]) =&\log(\texttt{exp}) + \beta_0 + \beta_1\texttt{cover}_{\text{MTPL+}} + \beta_2\texttt{cover}_{\text{MTPL+++}} +\\ &\beta_3\texttt{fuel}_{\text{gasoil}} + \beta_4\texttt{power}_{\text{66-110}} + \beta_5\texttt{power}_{>110} +\\ &\beta_6\texttt{split}_{\text{twice}} + \beta_7\texttt{split}_{\text{thrice}} + \beta_8\texttt{split}_{\text{monthly}} +\\ &f_1(\texttt{ageph}) + f_2(\texttt{long}, \texttt{lat}). \qquad (3)\end{aligned}$$

The number of claims was assumed to have a Poisson distribution and a logarithmic link function was used. The offset by the logarithm of the exposure ensures that that the number of claims reported is expressed per unit of time exposure. The categorical risk factors are coded with dummy variables by taking the level with the most exposure as reference level ($\texttt{cover}_{\text{MTPL}}$, $\texttt{fuel}_{petrol}$, $\texttt{power}_{<66}$, $\texttt{split}_{\text{once}}$). The continuous risk factor and spatial effect are captured by non-parametric smooth functions. f_1 determines the univariate smooth age effect. f_2 is a bivariate smooth function of the coordinates.

Carrying on the outlined approach, the next step would be to add interaction effects between the continuous factors. Since there is only one continuous risk factor, i.e. `ageph`, this step was omitted. Nonetheless, it is possible to incorporate interactions between categorical and continuos factors in a GAM. Three such interactions were considered (`ageph-power`,`ageph-split`,`ageph-sport`) but none of them achieved a decrease in BIC value.

Figure 3 (upper left) displays the fitted smooth function $\hat{f}_1(\texttt{ageph})$ from equation (1). The dashed lines represent the 97.5% pointwise confidence interval. In regions with few observations the interval is wider.
From this graph it is evident that young policyholders exhibit risky behaviour. The risk declines with increasing age and flattens out between the age of 38 and 46. Then the risk decreases at a moderate pace until the age of 59. From that age onwards, the risk soars rather sharply but remains considerably below the levels of the youngest drivers.
The lower half of Figure 3 shows the fitted spatial effect $\hat{f}_2(\texttt{long}, \texttt{lat})$. Note that the map indicates the regions where policyholders resided and not the sites where accidents took place. It is not unreasonable to assume that a large portion of distance travelled will be in the neighbourhood of one's home. Moreover, a policyholder's place of residence is a proxy for his socio-economical characteristics. The map clearly pinpoints the capital region as a district with elevated risk compared to more peripheral regions in the south-east, north-east and west. The disparity is due to a difference in population density. A higher population density is reflected in a higher claim frequency of the residing policyholders.

4.5 Severity model

The same procedure as before can be applied to build a claim severity model. Note that only policyholders that actually filed a claim, can be used for that

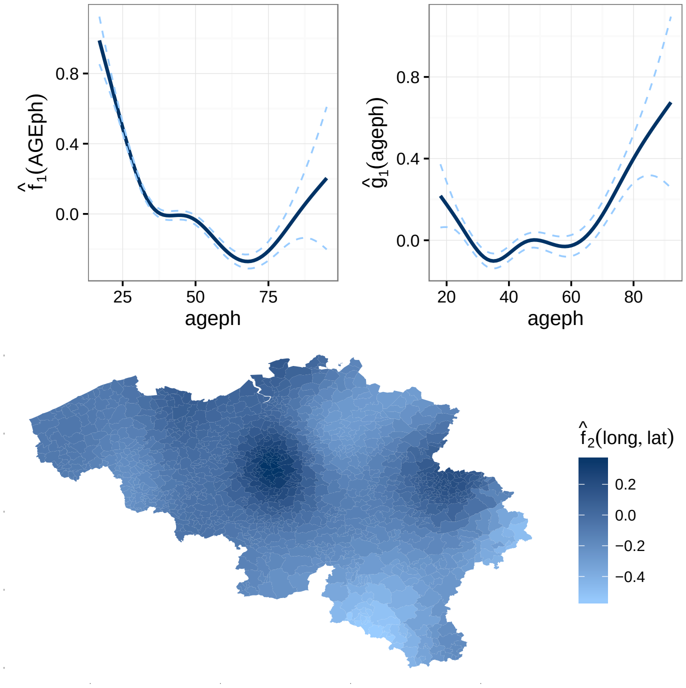


Figure 3: Fitted smooth GAM effects. Top left: single effect $\hat{f}_1(\texttt{AGEph})$. single effect $\hat{g}_1(\texttt{AGEph})$. Top right: Bottom: spatial effect $\hat{f}_2(\texttt{LONG}, \texttt{LAT})$.

purpose. This reduces the size of the data set to 18345 observations. The optimal model is the following GAM

$$\begin{aligned}\log(\mathbf{E}[\texttt{avg}]) =& \log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{cover}_{\text{MTPL}} + \beta_2 \texttt{cover}_{\text{MTPL+}} + \\ & \beta_3 \texttt{cover}_{\text{MTPL+++}} + \beta_8 \texttt{split}_{\text{once}} + \\ & \beta_9 \texttt{split}_{\text{twice}} + \beta_{10} \texttt{split}_{\text{thrice}} + \\ & f_1(\texttt{ageph}). \end{aligned} \tag{4}$$

The response $\log(\texttt{avg})$ was assumed to have a Gaussian distribution, such that the average claim amount obeys a lognormal distribution. Comparison of the frequency and severity models reveals that the latter one includes fewer risk factors. In actuarial modeling it is commonly observed that fewer covariates suffice to explain claim severities.
Figure 3 (upper right) displays the fitted smooth function $\hat{g}_1(\texttt{ageph})$ from equation (4). Between ages 18 and 35 the average claim cost decreases. This might indicate that less experienced drivers suffer more severe accidents. The average claim cost then starts increasing again, stabilizes between 45 and 60, and eventually rises once more. A potential explanation is that older policyholders drive more expensive cars implying higher repair costs.

5 Binning the smooth GAM effects

The second major stage in the data driven tarification strategy consists of binning the continuous effects in the previously obtained GAMs.

5.1 Spatial effect

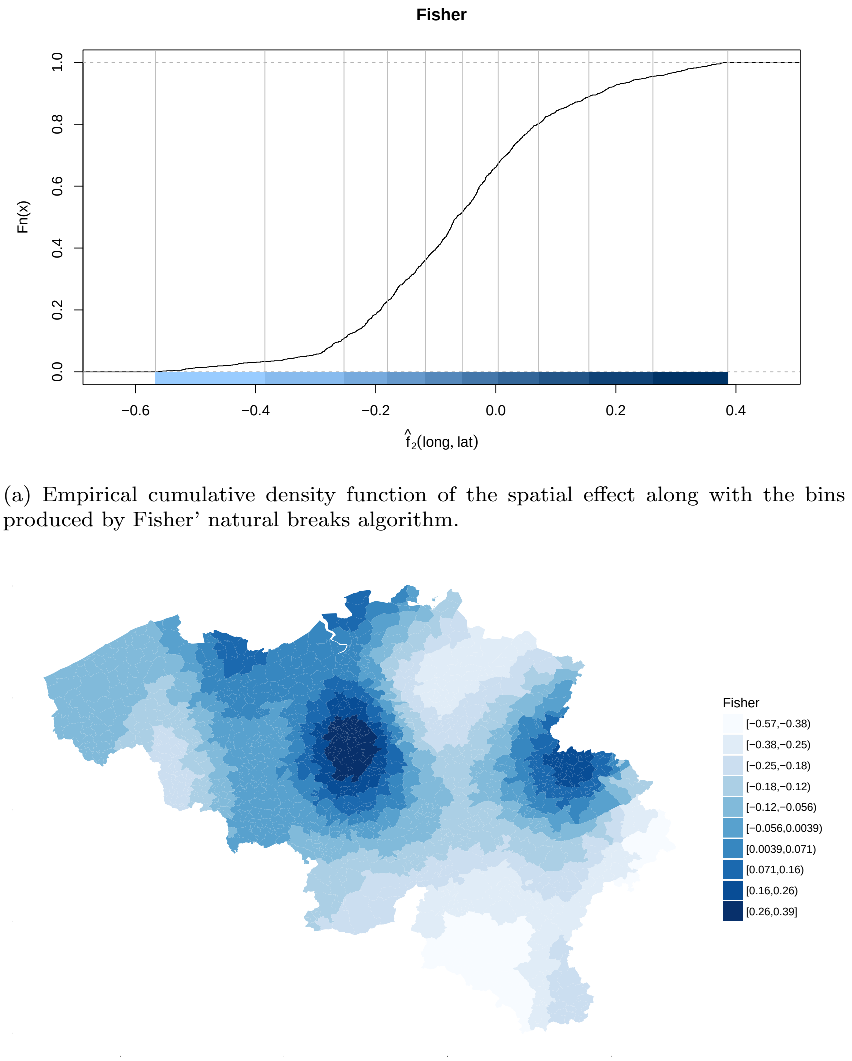
I applied the procedure of [1] to find the optimal number of non-overlapping bins for the spatial effect. Table 3 summarizes the procedure and table 4 lists the BIC and AIC values for different numbers of bins. The lowest BIC value is attained for 10 bins. Figure 4 visualizes the resulting districts on the map.

Table 3: Procedure from [1] to find the optimal number of bins for the spatial effect, $\hat{f}_2$ starting from the optimal GAM in equation (1).

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, $f_2(\texttt{long}, \texttt{lat})$ from equation 1, 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 incorporates the categorical spatial risk effect.
Step 3	Calculate the BIC and AIC value of the newly fitted GAM.

Table 4: BIC and AIC of the newly fitted GAM after binning the spatial effect via Fisher's algorithm for different numbers of bins. The lowest BIC value is attained for 10 bins.

# bins	**BIC**	**AIC**
6	125810.7	125609.7
7	125790.1	125579.2
8	125784.9	125563.8
9	125782.0	125551.0
10	**125765.9**	125524.7
11	125776.2	125525.1
12	125783.4	**125522.2**



(a) Empirical cumulative density function of the spatial effect along with the bins produced by Fisher' natural breaks algorithm.

(b) Districts based on the bins produced by Fisher' natural breaks algorithm.

Figure 4: Output of Fisher's natural breaks algorithm with 10 bins.

5.2 Continuous risk factors

Binning the continuous risk factors $f_1(\texttt{agph})$ and $g_1(\texttt{agph})$ is quite different from binning the spatial effect since consecutive values of the risk factor need to be grouped together. The binning was performed with evolutionary trees. These combine the regression trees framework with genetic algorithms. The evaluation function to measure the performance of a tree has the form

$$N \cdot \log(\mathrm{MSE}) + 4 \cdot \alpha \cdot (M+1) \cdot \log(N) \tag{5}$$

where N is the number of observations, M the number of leaf nodes in a tree and α is a tuning parameter.

Again, the method of [1] was applied to find the optimal tuning parameters α_{freq} and α_{sev} of the trees binning the age effect. Table 5 summarizes the procedure. α_{freq} and α_{sev} were assumed to lie in a predefined set $\{1, 1.5 \ldots, 9.5\} \times \{1, 10, 1000\}$ which was explored through a bisection search. The AIC measure was in general preferred over BIC since it uses a softer penalty and therefore allows for more splits in the trees.

Table 6 shows some relevant results. α_{freq} between 100 and 170, and α_{sev} between 30 and 90 resulted in a minimal AIC value for respectively the frequency and severity GLMs. Ultimately, α_{freq} was set to 170 and α_{sev} to 90.

Table 5: Procedure from [1] to find the optimal tuning parameters α_{freq} and α_{sev} to bin the continuous effect of age.

Procedure	Find the optimal tuning parameters α_{freq} and α_{sev} for the evolutionary trees.
Step 1	Fit an evolutionary tree to the single effects,f_1(`agph`) and g_1(`agph`), 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 6: AIC and BIC values for relevant values of the tuning parameters α_{freq} and α_{sev}.

α_{freq}	**AIC**
75	125557.9
$[100, \mathbf{170}]$	**125554.9**
$[172.5, 190]$	125561.9

α_{sev}	**AIC**
$[10, 20]$	66124.33
$[30, \mathbf{90}]$	**66121.45**
100	66138.03

6 Generalized linear models with binned risk factors

6.1 Discussion

With all effects expressed as categorical risk factors a Poisson GLM was fitted to the claims data for the frequency component, and a lognormal GLM for the severity component.

Figure 5 compares the original GAM effects with the final GLM coefficients of the factor intervals. The GLM coefficients of the age effect are rescaled versions of the actual coefficients as explained in section 5 of [1]. The rescaling operation brings the GAM effects and GLM coefficients to a comparable range. Remark that the actual GLM is not altered in any way.

The piecewise constant functions formed by the GLM coefficients approximate the GAM effects rather well. However, trading flexibility for simplicity inevitably leads to some mismatch.

The mismatches mainly occur in ranges where the exposure is very low. The GLM coefficients of these bins will consequently be tuned towards the policyholders that contribute most to the exposure of that bin. The approximation is much better for bins with high exposure. Underestimates of the smooth effects occur for both $\hat{f}_1$ and $\hat{g}_1$ in the youngest and oldest age category.

In the range between ages 54 and 58 the GLM coefficient is overestimating the smooth effect $\hat{f}_1$. It even introduces a qualitative behaviour that is not seen in the GAM: an increase and subsequent decrease of the effect between the ages of 50 and 62.

Even more interesting is the fitted GLM of the spatial effect in the bottom row of the figure. It displays a qualitative feature that is not at all present in the original smooth effect $\hat{f}_2$: a drop and subsequent increase of the effect in the peripheries around Brussels and Liége. This phenomenon was encountered for #bins ≥ 9, and for all αs in the predefined set. To avoid the occurrence of these artefacts the number of bins for the spatial effect was fixed to eight. With this new number of bins I performed another search for the optimal α_{freq}.

Table 7 provides the result of this search. The tuning parameter with the lowest associated AIC value, again seems to originate an aberrant qualitative feature that is not observed in the GAM. Figure 6 shows a rise and subsequent decline of the GLM coefficient between 53 and 56. $\alpha_{sev} = 100$ is the tuning parameter with the lowest AIC and BIC value, not demonstrating this issue. Table 8 shows a comparison of the AIC and BIC from the original GAM and two fitted GLMs. In case of the severity model both the BIC and AIC values are lower for the GAM. In case of the frequency model it is the other way around: the GLM has better quality measures. That is not what is anticipated. As expected GLM$_1$ outperforms GLM$_2$. However, because of the unusual fact that the original GAM has unfavourable AIC/BIC values, and the surprising qualitative features of GLM$_1$, it is not entirely clear which model is preferred. If one sticks to the philosophy adopted throughout the text of using the BIC/AIC as the quality measure, one would end up with GLM_1.

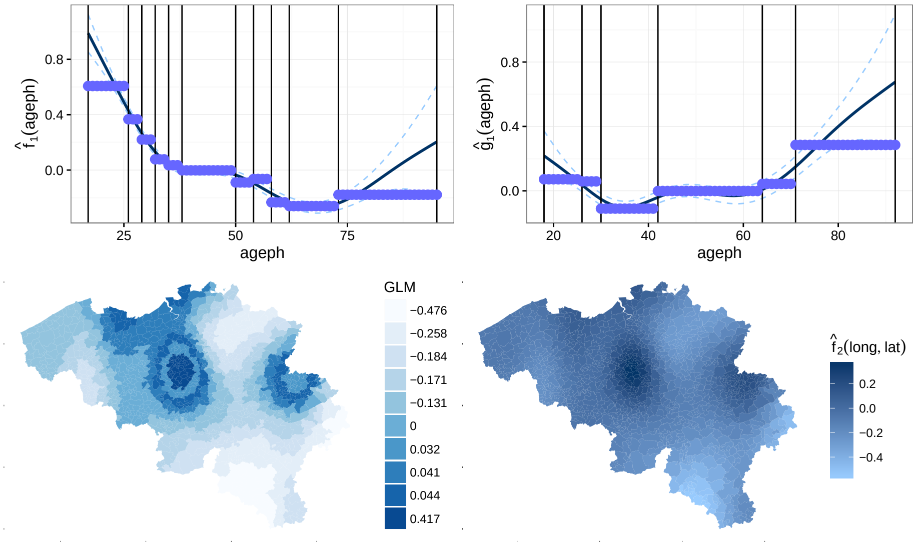


Figure 5: Claim frequency and severity GLMs obtained with the procedure described above.

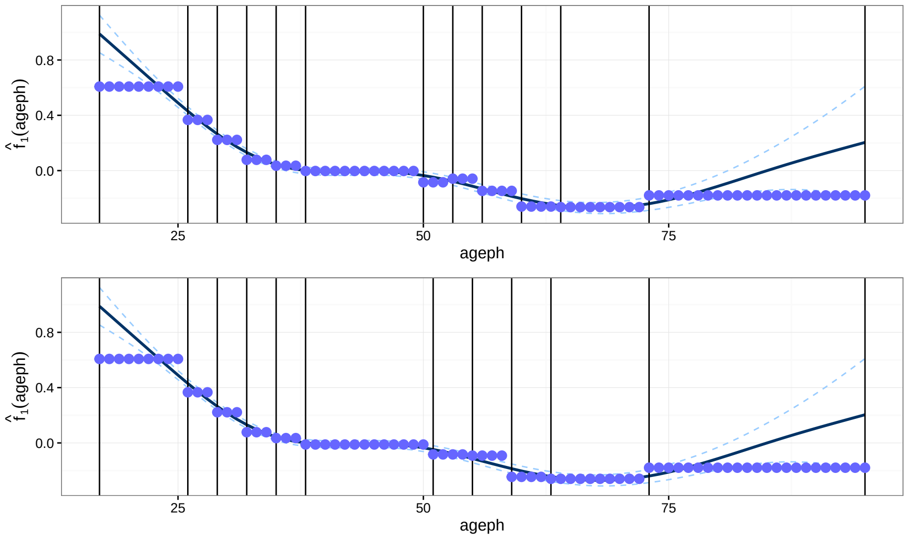


Figure 6: Claim frequency GLMs obtained with “optimal” $\alpha_{sev} = 87.5$ (top), and $\alpha_{sev} = 100$ (bottom). Note the different behaviour between ages 50 and 60.

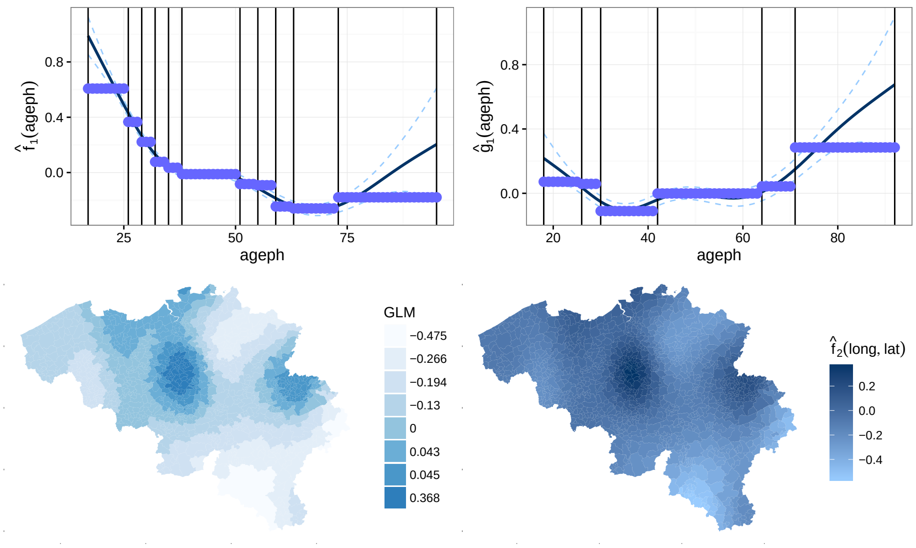


Figure 7: Alternative claim frequency and severity GLMs (eight spatial bins, $\alpha_{freq} = 100$).

Table 7: In search of the optimal tuning parameter α_{freq} for a GLM with eight bins for the spatial effect.

α_{freq}	**AIC**
$[1, 10]$	125598.5
$[25, \mathbf{87.5}]$	**125596.3**
$[100, 125]$	125863.4

Table 8: Comparison of the AIC and BIC values for the original GAMs, frequency GLM$_1$ with 10 bins and $\alpha_{freq} = 170$, frequency GLM$_2$ with 8 bins and $\alpha_{freq} = 100$ (left), and the severity GLM with $\alpha_{sev} = 90$ (right).

Model	**BIC**	**AIC**
GAM	126023.4	125610.4
GLM$_1$	125835.0	125554.9
GLM$_2$	125863.4	125603.3

Model	**BIC**	**AIC**
GAM	66214.93	66116.56
GLM$_3$	66215.25	66121.45

6.2 Some output

The reference levels for the frequency (GLM$_1$) and severity (GLM$_3$) are $\texttt{cover}_{\text{MTPL}}$, $\texttt{fuel}_{petrol}$, $\texttt{power}_{<66}$, $\texttt{split}_{\text{once}}$, $\texttt{geo}_{[-0.06,0.018)}$, and $\texttt{ageph}_{[38,51)}$ (GLM$_1$) and `ageph`$_{[42,64)}$ (GLM$_3$).

Tables 9 and 10 give the GLM summaries obtained in `R`. From these results it may be concluded that not all coefficients associated with the dummy variables are significantly different from zero. This is the case for some of the binned age and spatial categories.

From the ANOVA output in Table 11 it may be concluded that all the covariates in the GLMs are significant.

7 Implementation

The regression analysis was done in the `R` language for statistical computing. The implementation strongly relies on the scripts accompanying [1]. A list of essential packages is given here.

`mgcv`: Generalized additive models with integrated smoothness estimation.

`classInt`: Fisher's natural breaks algorithm.

`evtree`: Evolutionary trees.

All data and scripts used to produce the results presented in this report can be found in the supplementary material.

8 Composition of the pure premium

The pure premium for policyholder i, $E[P_i]$, is calculated as the product of the expected value of the claim frequency ($E[F_i]$) and the expected value of the claim severity $E[S_i]$. The frequency of policyholder i (F_i) is equal to the ratio of the number of claims (N_i) and the exposure (e_i). The severity for policyholder i (S_i) is equal to the ratio of the total claim amount (L_i) and the number of claims (N_i). The pure premium is then calculated as

$$E[P_i] = E[F_i] \cdot E[S_i] = E\Big[\frac{N_i}{e_i}\Big] \cdot E\Big[\frac{L_i}{N_i}\Big]. \tag{6}$$

Figure 8 gives the boxplot comparison of the pure premiums predicted for the policyholders in the `MTPL` data set with the obtained GAMs and GLMs. Table 12 exhibits the Tukey five-number summaries for the GAM and GLM predictions.

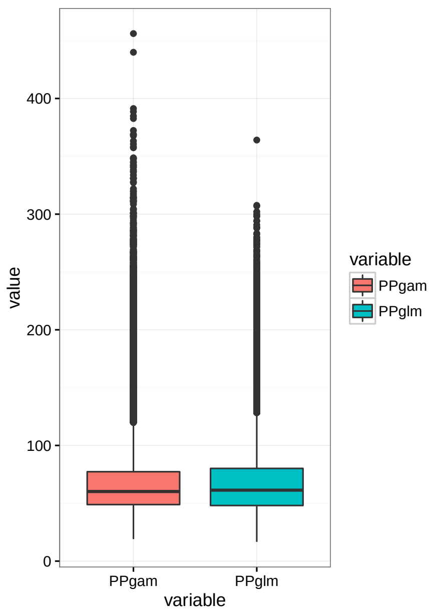


Figure 8: Boxplot of the pure premiums predicted for the policyholders in the `MTPL` data set with the obtained GAMs and GLMs respectively.

9 Conclusion

This report covered the construction of an insurance tariff based on the strategy of [1] to deal with continuous risk factors and geographical information. Two independent models were obtained, one for the claim frequency, and one for the claim severity. Combining both allows to calculate the pure risk premium for an insurance contract. The approach lead to satisfactory results, although some anomalous effects were observed.

References

[1] Roel Verbelen, Katrien Antonio, Maxime Clijsters, and Roel Verbelen. Using risk factors in P&C insurance pricing: a data driven strategy with GAMs, regression trees and GLMs., 2016. Working paper.

Table 9: Frequency GLM summary.

```
> summary(glm.freq)

Family: poisson
Link function: log

Formula:
NCLAIMS ~ COVER + FUEL + POWER + SPLIT + AGEph + GEO

Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.15433    0.02195 -98.135  < 2e-16 ***
COVERMTPL+       -0.16159    0.01667  -9.696  < 2e-16 ***
COVERMTPL+++     -0.17441    0.02204  -7.914 2.48e-15 ***
FUELGasoil        0.19351    0.01519  12.737  < 2e-16 ***
POWER>110         0.28005    0.06694   4.184 2.87e-05 ***
POWER66-110       0.08632    0.01637   5.274 1.33e-07 ***
SPLITMonthly      0.30814    0.02181  14.131  < 2e-16 ***
SPLITThrice       0.37842    0.02438  15.520  < 2e-16 ***
SPLITTwice        0.17488    0.01693  10.331  < 2e-16 ***
AGEph[17,26)      0.61887    0.02724  22.719  < 2e-16 ***
AGEph[26,29)      0.37770    0.02807  13.457  < 2e-16 ***
AGEph[29,32)      0.23321    0.02844   8.201 2.38e-16 ***
AGEph[32,35)      0.08890    0.02895   3.071 0.002136 **
AGEph[35,38)      0.04637    0.02922   1.587 0.112438
AGEph[51,55)     -0.07168    0.02897  -2.475 0.013331 *
AGEph[55,59)     -0.08021    0.03239  -2.476 0.013282 *
AGEph[59,63)     -0.23411    0.03577  -6.546 5.92e-11 ***
AGEph[63,73)     -0.24810    0.02665  -9.311  < 2e-16 ***
AGEph[73,95]     -0.16774    0.03686  -4.550 5.36e-06 ***
GEO[-0.57,-0.38) -0.47465    0.13810  -3.437 0.000588 ***
GEO[-0.38,-0.23) -0.26614    0.04596  -5.790 7.04e-09 ***
GEO[-0.23,-0.14) -0.19403    0.02640  -7.350 1.99e-13 ***
GEO[-0.14,-0.06) -0.12960    0.02204  -5.881 4.07e-09 ***
GEO[0.018,0.12)   0.04303    0.02020   2.131 0.033130 *
GEO[0.12,0.24)    0.04543    0.02938   1.546 0.122028
GEO[0.24,0.39]    0.36750    0.02437  15.082  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1


R-sq.(adj) =  0.0154   Deviance explained = 2.73%
UBRE = -0.4637  Scale est. = 1         n = 163657
```

Table 10: Severity GLM summary.

```
> summary(glm.sev)

Family: gaussian
Link function: identity

Formula:
log(AVG) ˜ COVER + SPLIT + AGEph

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.17068    0.02239 275.552  < 2e-16 ***
COVERMTPL+   -0.18133    0.02505  -7.238 4.73e-13 ***
COVERMTPL+++  0.13921    0.03276   4.250 2.15e-05 ***
SPLITMonthly -0.24064    0.03238  -7.431 1.12e-13 ***
SPLITThrice   0.15981    0.03692   4.329 1.51e-05 ***
SPLITTwice    0.10557    0.02556   4.130 3.65e-05 ***
AGEph[18,26)  0.07299    0.04029   1.811   0.0701 .
AGEph[26,30)  0.05983    0.03714   1.611   0.1072
AGEph[30,42) -0.10929    0.02634  -4.149 3.36e-05 ***
AGEph[64,71)  0.04423    0.04442   0.996   0.3194
AGEph[71,92]  0.28620    0.04798   5.964 2.50e-09 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1


R-sq.(adj) =  0.0153   Deviance explained = 1.58%
GCV = 2.3083  Scale est. = 2.3069    n = 18345
```

Table 11: ANOVA output.

```
> anova(glm.sev)

Family: gaussian
Link function: identity

Formula:
log(AVG) ~ COVER + SPLIT + AGEph

Parametric Terms:
      df     F  p-value
COVER  2 45.16  < 2e-16
SPLIT  3 42.09  < 2e-16
AGEph  5 15.64 2.18e-15


> anova(glm.freq)

Family: poisson
Link function: log

Formula:
NCLAIMS ~ COVER + FUEL + POWER + SPLIT + AGEph + GEO

Parametric Terms:
      df  Chi.sq  p-value
COVER  2  125.86  < 2e-16
FUEL   1  162.24  < 2e-16
POWER  2   42.07 7.32e-10
SPLIT  3  365.30  < 2e-16
AGEph 10 1081.82  < 2e-16
GEO    7  524.79  < 2e-16
```

Table 12: Tukey five-number summaries for the GAM and GLM predictions of the pure premium.

```
> fivenum(DT.Premium$PPgam)
[1]  19.03713  48.86116  60.17536  77.28541 456.11727
> fivenum(DT.Premium$PPglm)
[1]  16.60847  48.04979  61.33106  80.16688 364.07957
```