Advanced Non-Life Insurance Mathematics Assignment 2: Pricing

Pieter Pijls[*] *Pieter-Jan Dubois*[†] *Eva Thienpondt*[‡]

Academic year 2017-2018

Introduction

In this assignment we analyze a real-life pricing data set (third party liability insurance portfolio) on claim frequencies and sevirities. We include different risk factors (factor,continuous,spatial,interaction) in our models. First, we perform a data exploration. Second, we develop a model for the frequency using GAMs and GLMs. Finally, we construct a severity model.

Block 1) Load and inspect the data

Question 1

In this section we reproduce the severity data from Henckaerts et al (2017) by calculating the average cost of claim equal to the ratio of total claim amount and the number of claims. Next, we exclude observations with no claims. In addition, we remove observations with an average cost of claim higher than EUR 81.000 using Extreme Value Theory (EVT). The R code we used is as follow.

```r
# Henckaerts et al (2017), page 11: specification of the severity data
DT.sev <- as.data.frame(cbind(DT$AMOUNT/DT$NCLAIMS,DT$AMOUNT))
# Exclude NA
DT.sev <- as.data.frame(DT.sev[complete.cases(DT.sev[,1]),])
# Exclude large claims
DT.sev <- as.data.frame(DT.sev[!(DT.sev$V1>81000),])
colnames(DT.sev) <- c("Severity","AMOUNT")
```

Question 2

In this section we explore a third party liability (MTPL) insurance portfolio from a Belgian insurer in 1997. In total, the data set contains 163 231 policyholders. Figure 1 illustrates the distribution of the number of claims (`nclaims`), fraction of year the policyholders was exposed(exp) and the total amount claimed by the policyholder in euro (`amount`).

Table 1: Basic statistics

Overall claim frequency	Average claim amount	Fraction of claims higher as 10000
0.1393352	1620.055	0.0197322

[*] r0387948 (Faculty of Economics and Business, KU Leuven, Leuven, Belgium)

[†] r0382187 (Faculty of Economics and Business, KU Leuven, Leuven, Belgium)

[‡] r0639885 (Faculty of Economics and Business, KU Leuven, Leuven, Belgium)

In figure 1 the vertical red dotted line is equal to the mean of the respective variable. Around 90% of the policyholders are claim-free, while around 10% files one claim. On average a policyholders files in 0.12 claims per year. Around 1% of the policyholders files more than one claim. Around 77% of the policyholders have an exposure equal to one. The average exposure is 0.89 year. We calculated the claim frequency as the ratio of the total number claims and the total exposure in years. Most claims involve only small amounts, where around 2% of the claims exceed 10.000 EUR. In addition, we calculated the overall claim frequency, equal to 13.93%, as the ratio of the total number of claims and the total exposure in years.

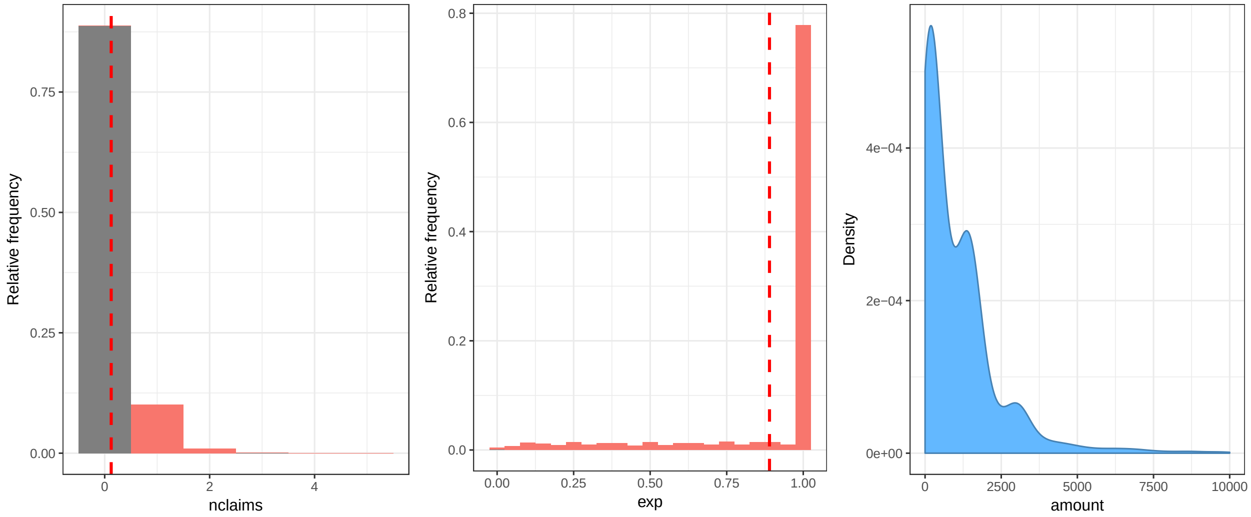

Figure 1: Relative frequency of the risk factors of nclaims and exp and density estiamte amount

Next, we create a population pyramid of the policyholders in our data set. Figure 2 illustrates that most policyholders are men (blue). For every female (pink) policyholder there exist around two male policyholders. The population pyramid shows that most policyholders are between 30 and 70 years old. We only observe a few young and old drivers in the population pyramid.

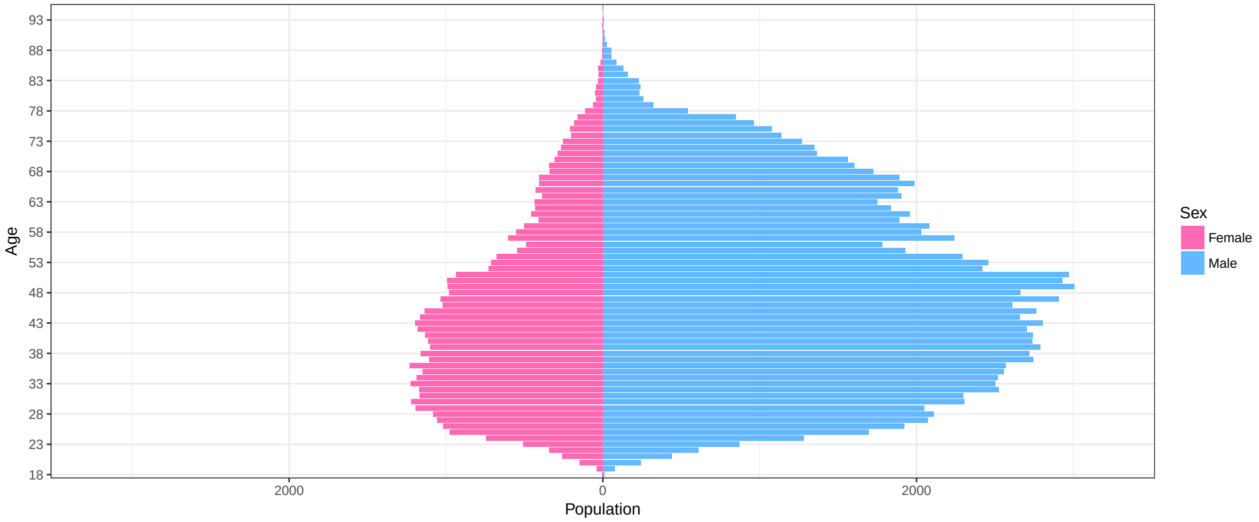

Figure 2: Population Pyramid

Figure 3 illustrates how the four categorical risk factors: `coverage`, `fuel`, `use` and `fleet` are distributed. Around 60% of the policyholders only have a TPL coverage, while around 28% has a partial omnium and around 13% has a full omnium. Most policyholders' type of fuel is gasoline (69.12%), while diesel accounts for 30.88%. Most policyholders use their car mainly for private reasons (95%) and most cars are not part of a fleet (97%).

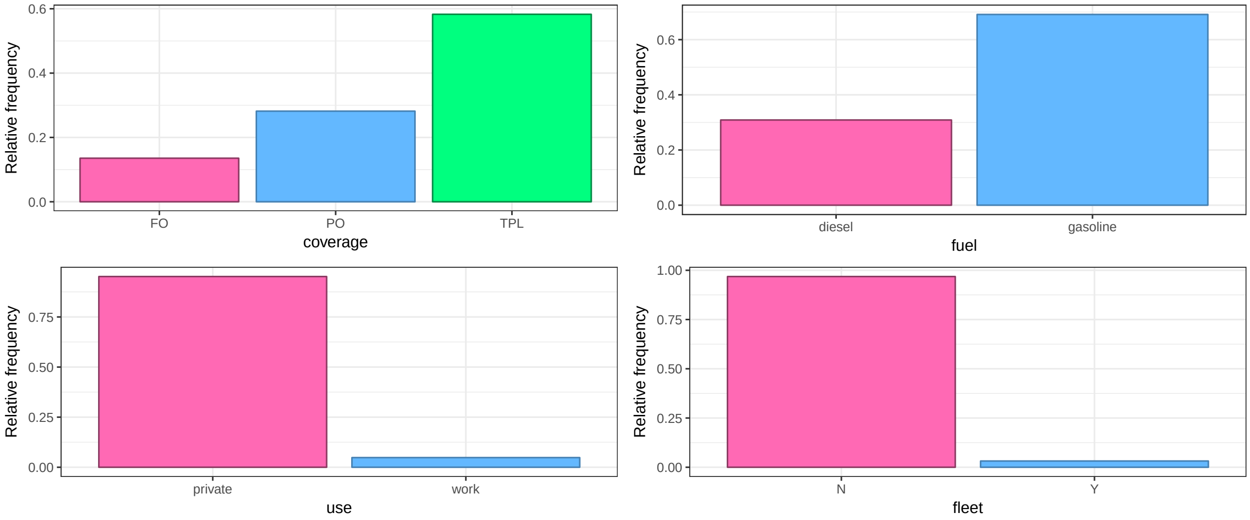

Figure 3: Relative frequency of coverage, fuel, sex, use and fleet

Figure 4 illustrates the distribution of three continuous risk factors: `power`, `agec` and `bm`. The histograms of the variables power and agec are split into five groups using the quantiles 0%, 20%, 40%, 60%, 80% and 100%. The vertical red dashed line is equal to the variable mean. The average horsepower of a car is around 55 kilowatt. Almost all insured cars (97.35%) have less than 100 kilowatt of horsepower. The average age of the vehicles is 7.37 years and the average bonus-malus level is 3.3. Most policyholders have a bonus-malus level of 0 (37.77%) or 1 (16.52). Only a small fraction of the policyholders (2.81%) has a higher bonus-malus level than 11.

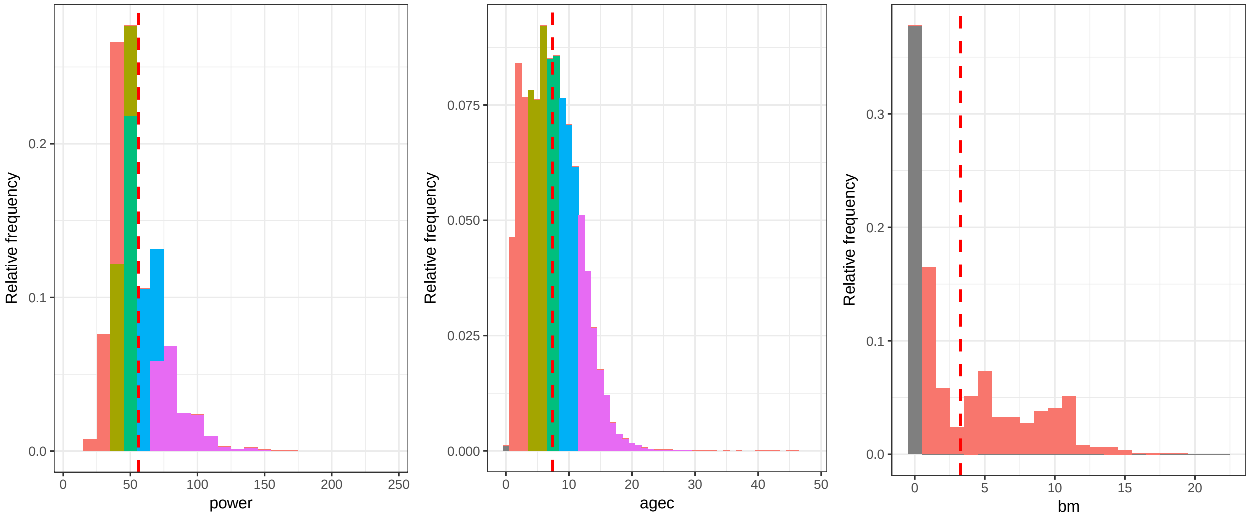

Figure 4: Relative frequency of power, agec and bm

The left panel of Figure 5 illustrates the two-dimensional estimate for ageph and power. The right panel shows the map of Belgium. The exposure of each municipality relative to the area of the municipality is grouped into three sets: low, average and high **(for more details see Q4)**.

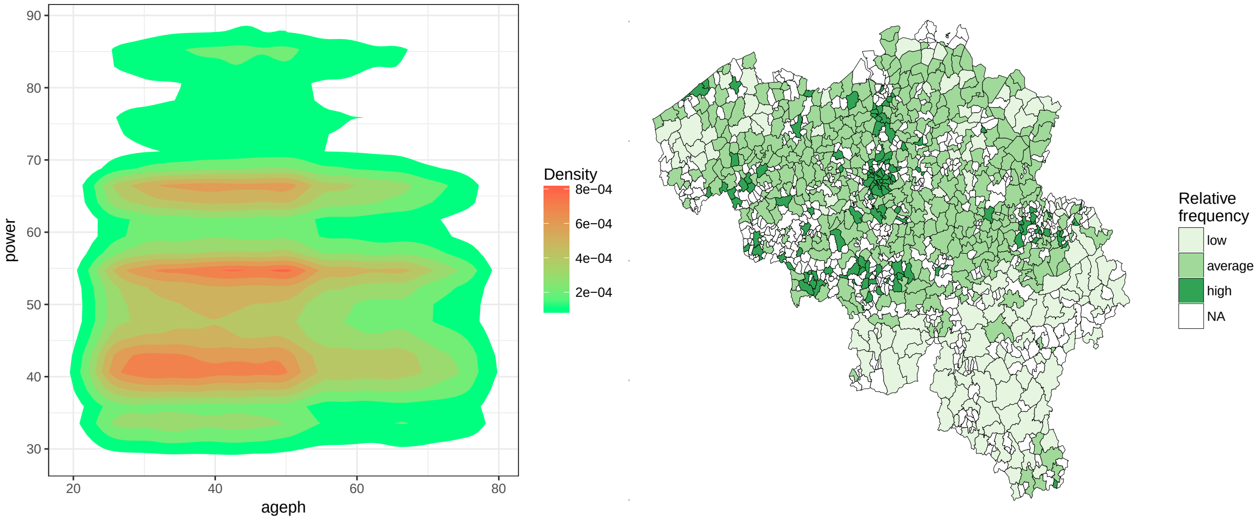

Figure 5: Density estimate of ageph-power (left) and map of Belgium wth exposures (right)

Question 3

In this section we reproduce the figures from Henckaerts et al (2017). See R code for further details.

Question 4

In this section we explain how the three clusters `Low`, `Average`, `High` are created. The fiction `aggregate()` aggregates the exposure data `EXP` by the postal code `PC`. The second command binds the column `N` to the data frame `DT_PC` and contains again the exposure data `EXP`.

In the following graph three clusters Low, Average, High are created using the quantiles of the relative exposure. The relative exposure for all the municipalities is calculated. The relative exposure is defined as the total exposure in the municipality over the total exposure of all the municipalities and is corrected for the total shape area of the municipality. The cluster Low contains 20% of the municipalities with the lowest relative exposure, the cluster Average contains the municipalities with a relative exposure between the quantile 20% and 80% and the cluster High represents the 20% of municipalities containing the highest relative exposure.

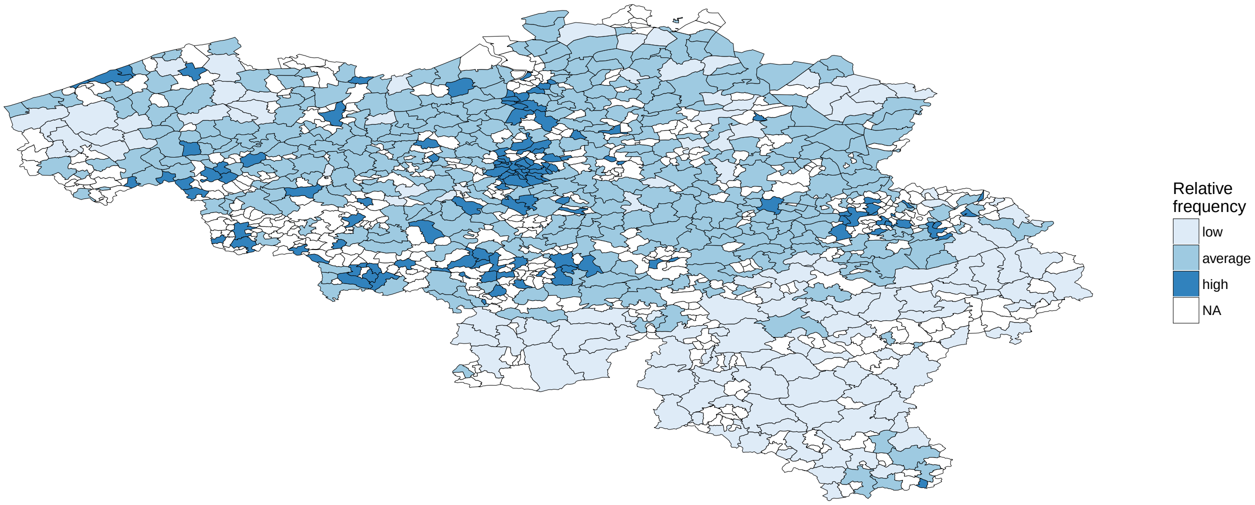

Figure 6: Map of Belgium wth exposures

Block 2) Fit a flexible GAM for frequency to the data

Question 5

In this section we fit a GLM with only a factor variable. We selected the factor variable `fuel` to calibrate a Poisson GLM with an offset and a linear predictor that only uses the factor variable. We isolate the role of `fuel` in the model. The regression model is formulated as follows: the number of claims is dependent on the type of `fuel`. The latter is a dummy variable that is zero if the `fuel` type is diesel and one if the `fuel` type is gasoline. The model with one factor variable `fuel` looks as follows

$$log(E(nclaims)) = log(exp) + \beta_0 + \beta_1 coverage_{PO} + \beta_2 fuel_{gasoline}$$

.

The parameters of the model are given in Table 2. The coefficient of gasoline is -0.18451. This means that driving a gasoline-fueled car results in less claims than driving a diesel-fueled car as the coefficient is negative. Given that the p-value is smaller than 0.001, this effect can be considered statistically significant.

Table 2: GLM with only a factor variable

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8466884	0.0119259	-154.84678	0
FUELgasoline	-0.1845086	0.0147633	-12.49776	0

Next, we create a GAM model where `AHEPH` is a continuous numeric risk factor. For more details on this model consult the R code.

Question 6

In this section we explore the differences and connection between ‘pred’ with ‘type’ is response’, ‘link’ and ‘terms’. When the type is set equal to ‘terms’ we get an intercept β_0 which is equal to the estimate of the intercept calculated from the GAM. We also get a fitted value for every age which indicates the deviation from the intercept. For example, an eighteen-year old has a deviation from the intercept β_0 of 0.91108. This positive coefficient is making an eighteen-year-old more risky. The `pred` function also calculates a standard error for every fitted value, for every age. If we want to calculate the average number of accidents for a particular age. We calculate it as follows:

$$\text{Average number of accidents} = \exp(intercept_{terms} + fit_{age.terms}).$$

For an eighteen-year-old this comes down to: $\exp(-1.995266 + 0.911082126) = 0.33$. When type is set equal to ‘link’ we only get a fitted value and a standard error for every age. The relation between ‘terms’ and ‘links is the following. The fitted value calculated for a particular age when type equal to ’link’ = the fitted value for the same age when + the intercept β_0 when type is equal to ‘terms’. Thus $fit_{age.link} = fit_{age.terms} + intercept_{terms}$. We then calculate the average number of claims as the $\exp(fit_{age.link})$.

The last type is ‘response’ which directly calculates the predicted number of claims that a person of age X will report. Thus the average number of claims is given by $fit_{age.resp}$. Therefore the connection between all three methods can be described as follows:

$$fit_{age.resp} = \exp(intercept_{terms} + fit_{age.terms}) = \exp(fit_{age.link})$$

Question 7

In this section we explore four different ways to incorporate a continuous variable. Also we will try to incorporate the continuous variable `age` in four different ways, as a linear effect, age as a factor, five year bins of age and a smooth effect of `age`. In the following graph the solution of these four approaches is shown. In the graph of the factor effect of `age` we did not include the ages 93 until 95 because these were not significant. In each graph we see that the oldest people have the largest standard deviation and thus the largest confidence interval. The large standard error for older ages is because we have less data of people of those ages. We also observe that the confidence interval is the largest in the factor effect of `age` for every age. While it is the smallest for every age when taking into account the linear effect of `age`. Again because the factor effect of `age` uses per age less data points. While the linear effect of `age` uses all the data to plot a global relation. Making the confidence intervals smaller.

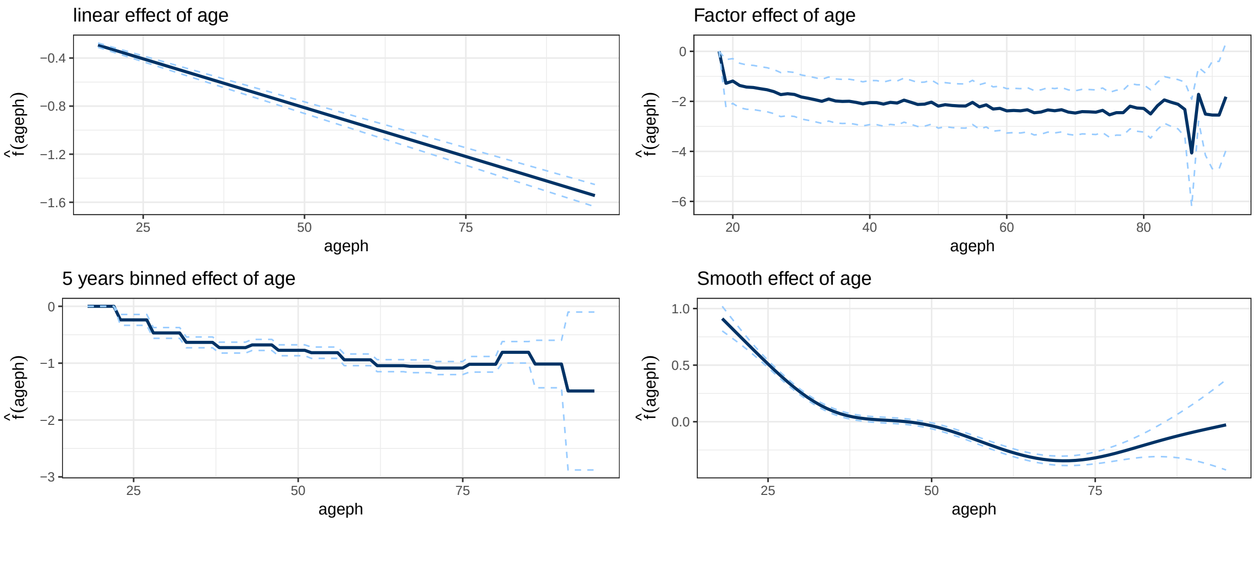

Figure 7: linear effect of ‘age’, ‘age’ as factor, ad hoc ‘age’ bins and smooth effect of ‘age’

In the Table 3 we compare the expected claim frequency between the four models for a 23 year old over the period of one year. We can see that the expected claim frequency for all four models are close to each other. The linear model predict the lowest claim frequency while the factor model predicts the highest. The standard error is the highest for the factor model and the lowest for the linear model. The binning and smoothing approach report results in between the factor and the linear model.

Table 3: Expected claim frequency for a 23 year old

	linear	factor	bins	smooth
Expected claim frecquency	0.2013828	0.2548571	0.2224151	0.2534732
Standard error	0.0025266	0.0147883	0.0045998	0.0054235

Question 8

In this section we calibrate the optimal frequency model from Henckaerts et al. (2017) (check R code for more details). We visualize the fitted effects.

The top row of the figure shows the smooth effects of the respective risk factors. The bottom left panel shows the interaction effect between ageph and power. The bottom right panel shows the fitted spatial effect.

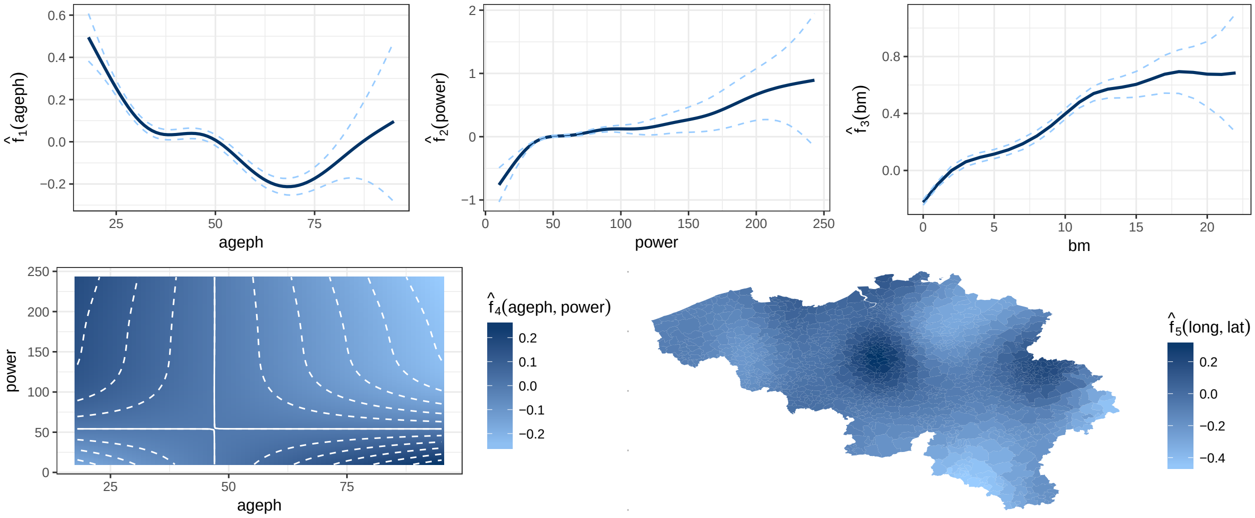

Figure 8: Top row

Block 3) Bin the spatial effect

Question 9

In this section we use different clustering strategies for binning the spatial effect. We highlight the differences both graphical and statistical. The first clustering approach, "equal intervals", divides the range of the spatial effect in $k = 5$ bins of equal length. The five bins each have a range of $(0.3405166 - (-0.4808574))/5 = 0.1642748$ and they have a number of 39, 263, 498, 235 and 111 municipalities respectively. Thus the majority of the municipalities is put into the middle bin, whereas the first and last bin almost have no municipalities.

The second clustering approach, “quantile binning”, results in $k = 5$ bins, where each bin contains approximately $1146/k$ municipalities. The resulting five bins thus have 229, 229, 230, 228 and 230 municipalities respectively. Comparing this approach to the first one, we see some major differences. Municipalities with similar spatial riskiness are not grouped together, because the bins are too wide in the extreme ends of the support (where data is scarce). These ends represent the respective least and most riskiness thus they should be binned more narrowly together.

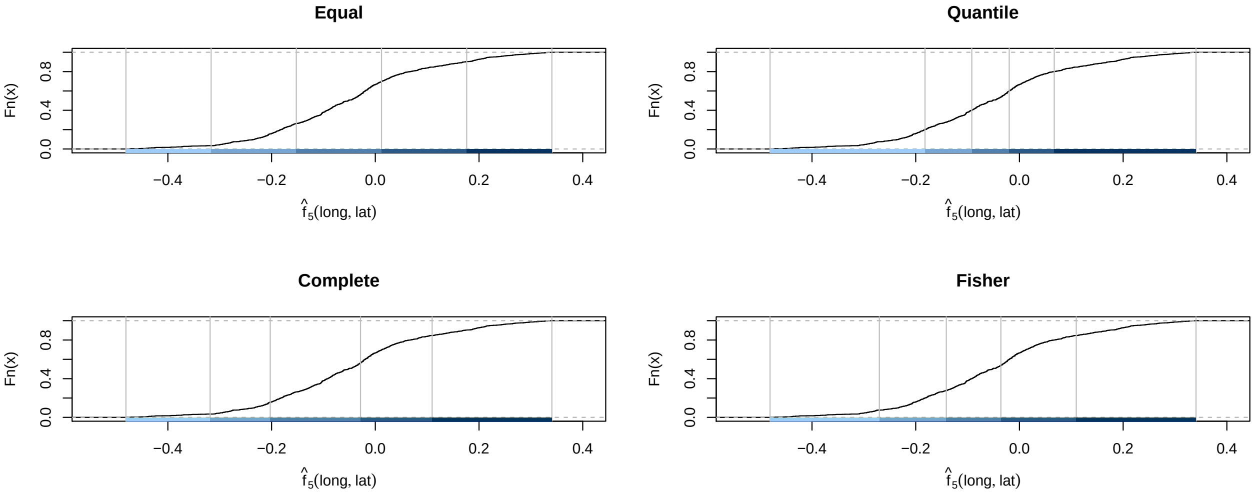

Figure 9: the empirical cumulative density function of the fitted spatial effect in combination with the fice bins produced by the four different binning methods.

The third approach, “complete linkage”, performs hierarchical clustering. This means that initially each municipality forms its own bin and in every iteration the two bins closest to each other are merged. The resulting five bins k have 39, 140, 466, 325 and 176 municipalities respectively. The majority of the municipalities is grouped into the third bin, whereas few municipalities are grouped into the first bin. This approach resembles the first one in output.

The fourth approach, “Fisher’s natural breaks” maximizes the homogeneity within bins. The resulting five bins have 86, 231, 299, 354, 176 municipalities respectively. These bins k seem to be the most homogeneous and a good compromise between the first and second approach.

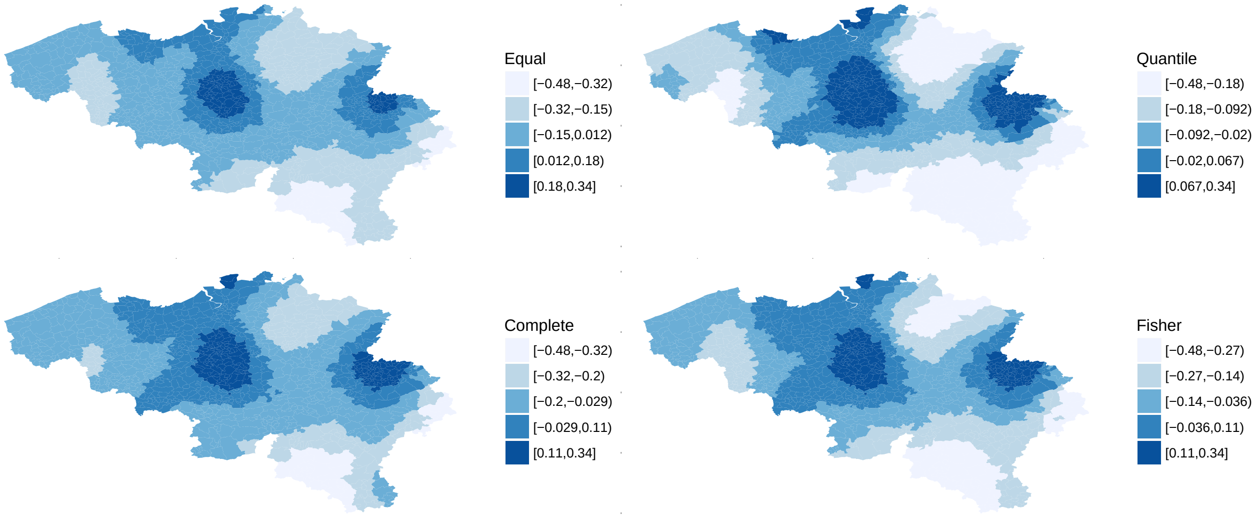

Figure 10: maps of Belgium with the municipalities grouped into five distint bins based on the intervals produced by the four different binning methods for the spatial effect.

In Table 4 we calculate the goodness-of-fit (GVF) and tabular accuracy index (TAI) using $k = 5$ bins. The Fisher's natural breaks algorithm outperforms the other models for both the GVF and TAI. Therefore, Fisher's algorithm is our preferred method to bin the spatial effect from the frequency model.

Table 4: The GVF and TAI for the different binning methods with five bins

	GVF	TAI
Equal	0.9131629	0.6751719
Quantile	0.8936280	0.6940038
Complete	0.9090685	0.6790104
Fisher	0.9273669	0.7238903

Question 10

In this section we bin the fitted spatial effect using 5 bins and Fisher-Jenks (check R code for more details). Next, we re-calibrate the `gam` when the smooth spatial effect is replaced with the clusters. This is model (9) in Henckaerts et al. (2017).

Block 4) Bin the continuous and interaction effects

In this section we construct the bins based on the fitted smooth functions on the continuous risk factors.

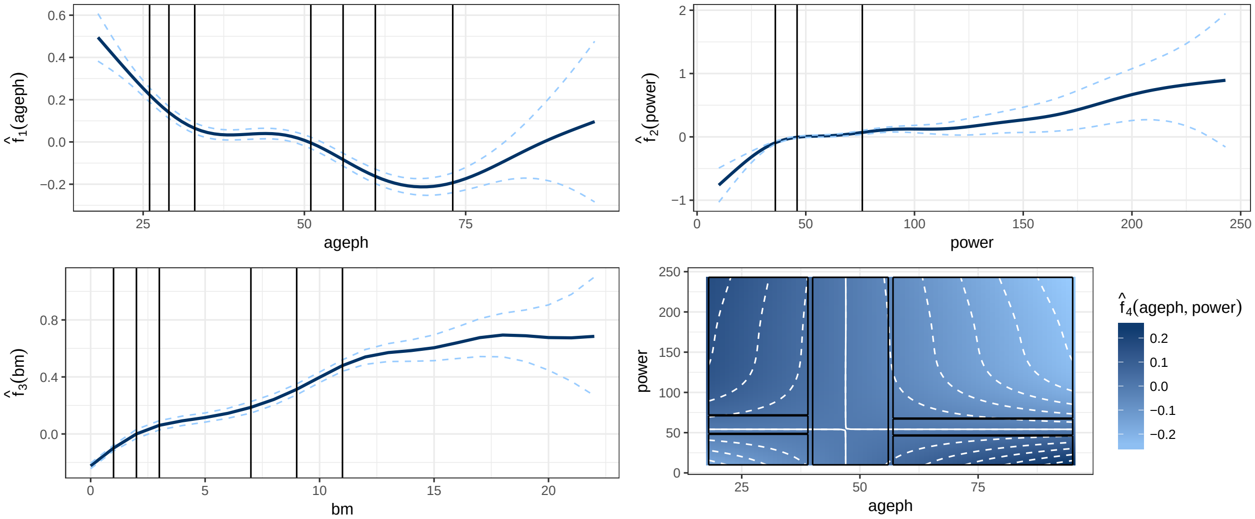

Figure 11: Binning intervals for the continuous effects

Question 11

In this section we calibrate the final `glm` proposed in Henckaerts et al. (2017) (see Appendix). Within each risk factor the level with the highest exposure is used as the reference level. We use the `relevel` instruction to accomplish this.

Block 5) Severity modeling

Question 12

Finally, we create our own severity model. First we will try to replicate the lognormal approach as is done in the research paper. Second we will propose our own severity model using the gamma distribution.

Before we calibrate the severity models we need to construct a new data set. We use the `avg` claim which is defined as the ratio of `amount` and `nclaims`. We can only take into account policyholders who actually filed a claim. Therefore, we discard all data for which the number of claims was zero. Since we do not want the very large claims to affect the tariff structure we exclude all losses above 81000. Next, we fit the following GAM's.

$$\mathbf{E}(log(\text{avg})) = \gamma_0 + \gamma_1 \text{coverage}_{PO} + \gamma_2 \text{coverage}_{FO} + g_1(\text{ageph}) + g_2(\text{bm})$$

The first model above is the same model as in the research paper. In this model we assume a Gaussian distribution for the log(avg). The following model is our new proposed model (see Appendix) which assumes a Gamma distribution for `avg`.

$$\mathbf{E}(\text{avg}) = \gamma_0 + \gamma_1 \text{coverage}_{PO} + \gamma_2 \text{coverage}_{FO} + g_1(\text{ageph}) + g_2(\text{bm})$$

Notice that we take the same reference levels as the research paper did for the lognormal severity model. After fitting the GAM's we transformed them to GLM's by binning the variables `age` and `bm` using evolutionary trees as explained in the frequency case. The only difference is that we set the tuning parameter α equal to 150. Here we see large difference. While the lognormal uses five bins to classify `age`, the gamma model only uses three bins. The same is true for `BM` where the lognormal uses five bins, while the gamma only uses three.

This huge difference might by because we held the tuning parameter α equal at 150 for the callibration of the bins for the gamma model. We did not, as in the paper, optimized the tuning parameter. Next, we create the GLM's from the binned data. The reference levels we introduced for the gamma model are TPL, $[31, 72)$ and $[0 - 1]$ for `coverage`, `age` and `BM` respectively.

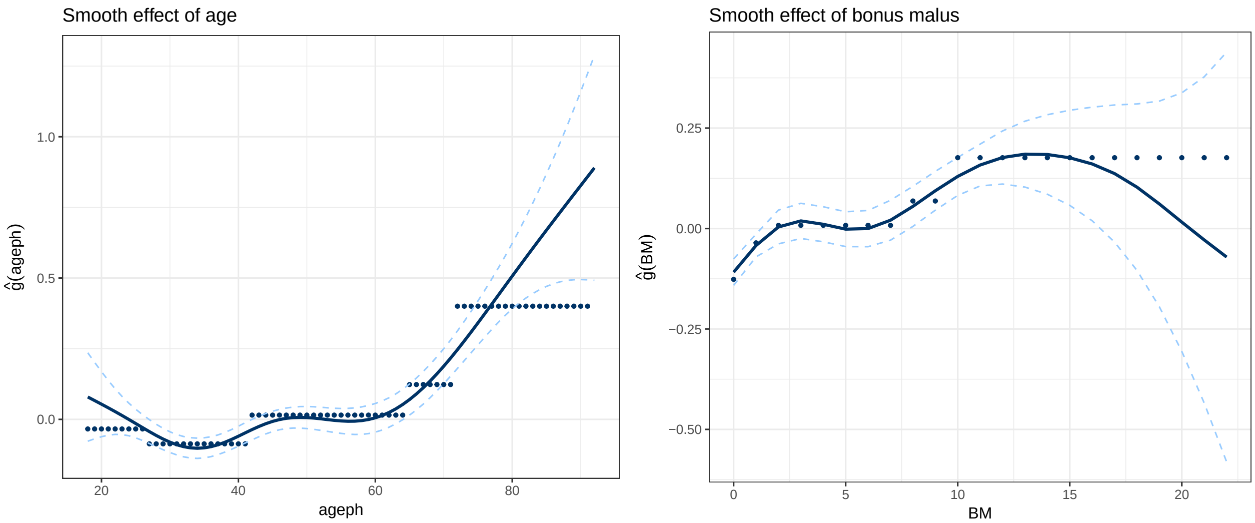

Figure 12: Severity model with a lognormal model

We also replicate the graphs of the smooth effect for `age` and `bm`. Figure 13 illustrates the smooth effect from the Gamma model, we see a completely different picture compared to the lognormal model.

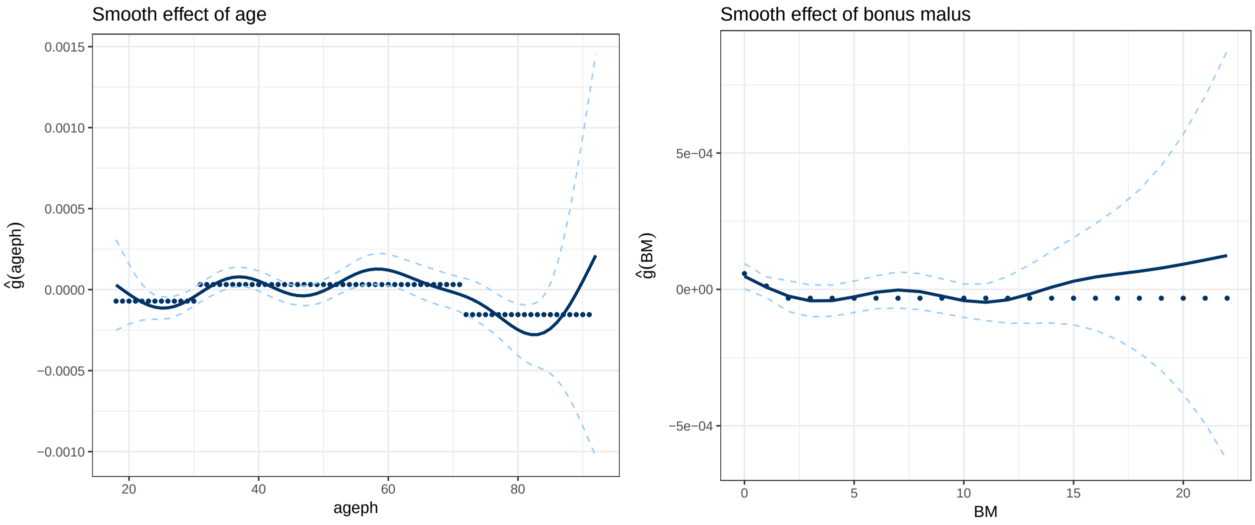

Figure 13: Severity model with a Gamma model

Finally, we calculate the AIC and BIC of both lognormal and gamma model. We observe that the lognormal provides a better fit as a GLM as well as a GAM. In conclusion, modeling the severity by a lognormal distribution gives the best fit.

Table 5: AIC and BIC for the severity models

	Lognormal~GAM~	Lognormal~GLM~	Gamma~GAM~	Gamma~GLM~
AIC	65535.51	65523.13	294492.3	294428.0
BIC	65535.14	65522.76	294492.0	294427.6

References

Henckaerts, Roel and Antonio, Katrien and Clijsters, Maxime and Roel, Verbelen, A Data Driven Binning Strategy for the Construction of Insurance Tariff Classes (May 12, 2017). Available at SSRN: https://ssrn.com/abstract=3052174

Appendix

Table 6: Frequency model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1739692	0.0213010	-102.0593375	0.0000000
COVERAGEFO	-0.1139780	0.0218718	-5.2111894	0.0000002
COVERAGEPO	-0.1205597	0.0167915	-7.1797878	0.0000000
FUELdiesel	0.1766412	0.0157002	11.2508558	0.0000000
AGEPH18-26	0.2770104	0.0296748	9.3348596	0.0000000
AGEPH26-29	0.1439304	0.0292465	4.9212850	0.0000009
AGEPH29-33	0.0652045	0.0261161	2.4967176	0.0125349
AGEPH51-56	-0.0607267	0.0264841	-2.2929502	0.0218509
AGEPH56-61	-0.1633508	0.0333041	-4.9048247	0.0000009
AGEPH61-73	-0.2483999	0.0296106	-8.3888885	0.0000000
AGEPH73-95	-0.1895626	0.0408371	-4.6419219	0.0000035
POWER10-36	-0.2080235	0.0330213	-6.2996716	0.0000000
POWER36-45	-0.0650021	0.0226528	-2.8694900	0.0041113
POWER75-243	0.1201028	0.0277360	4.3302153	0.0000149
BM1-2	0.1245501	0.0229708	5.4221083	0.0000001
BM2-3	0.1846636	0.0331444	5.5714869	0.0000000
BM3-7	0.3429337	0.0212489	16.1388590	0.0000000
BM7-9	0.4837291	0.0297941	16.2357191	0.0000000
BM9-11	0.5446250	0.0270306	20.1484633	0.0000000
BM11-22	0.7806171	0.0259584	30.0718548	0.0000000
GEO(-0.48,-0.27]	-0.3314731	0.0537611	-6.1656643	0.0000000
GEO(-0.27,-0.14]	-0.2035579	0.0232654	-8.7493812	0.0000000
GEO(-0.14,-0.036]	-0.1555569	0.0194400	-8.0019010	0.0000000
GEO(0.11,0.34]	0.1987907	0.0182225	10.9091003	0.0000000
AGEPHPOWER-0.052	-0.0690783	0.0458919	-1.5052402	0.1322623
AGEPHPOWER-0.029	-0.0250592	0.0260903	-0.9604784	0.3368145
AGEPHPOWER0.04	0.0582144	0.0396838	1.4669553	0.1423882
AGEPHPOWER0.047	0.0376408	0.0364468	1.0327593	0.3017166

Table 7: Severity model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0008467	3.40e-05	24.9324659	0.0000000
COVERAGEFO	-0.0001358	3.73e-05	-3.6374618	0.0002761
COVERAGEPO	0.0001908	4.00e-05	4.7719363	0.0000018
AGEPH[18,31)	-0.0001025	3.56e-05	-2.8757110	0.0040358
AGEPH[72,92)	-0.0001857	5.71e-05	-3.2510634	0.0011518
BM[1,2)	-0.0000456	4.99e-05	-0.9134074	0.3610404
BM[2,22)	-0.0000902	3.79e-05	-2.3803108	0.0173082