Data science in insurance

Pricing analytics with GLMs and GAMs

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

September, 2019

P&C insurance pricing models

- In a pricing model, identify for each insured i:

  N_i: number of claims during (period of) exposure d_i

  L_i: aggregate loss over N_i claims.

- The pure premium or risk premium π_i:

$$\pi_i = E\left(\frac{N_i}{d_i}\right) \cdot E\left(\frac{L_i}{N_i} | N_i > 0\right) = \underbrace{E(F_i)}_{\text{frequency}} \cdot \underbrace{E(\text{Sev}_i)}_{\text{severity}}.$$

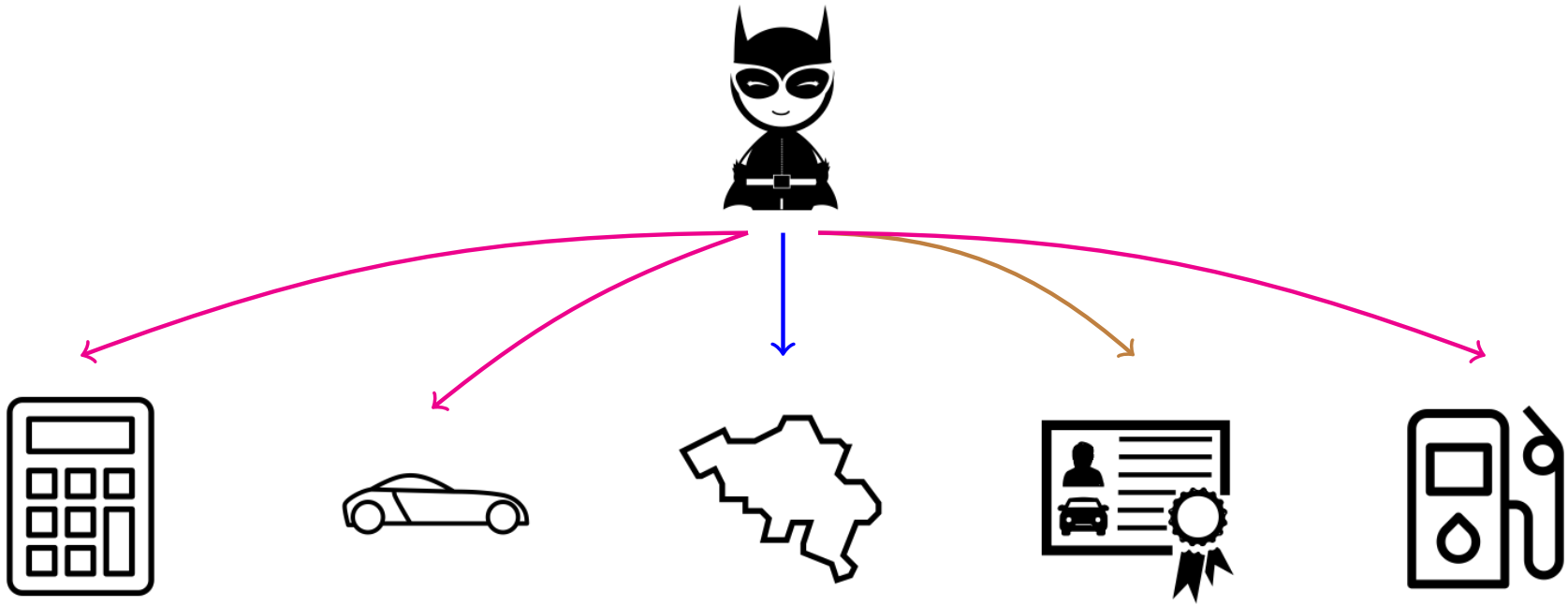
- Classify risks using a priori measurable characteristics:

  risk classification or segmentation.

A data driven strategy for the construction of insurance tariff classes

Henckaerts, Antonio, et al., 2018 (SAJ)

Motivation



Claim frequency and claim severity

as function of

nominal / numeric $\sim$ ordinal / spatial

features

Research questions

- Generalized Linear Models (GLMs) for frequency ($\sim$ Poisson) and severity ($\sim$ gamma).
- How to:

  (1) select risk factors or features?

  (2) cluster (or bin or fuse) levels within a risk factor?

  age groups / postal code clusters / clusters of car models
- Procedure should be data driven, scalable to large (big) data.
- End product is interpretable, within actuarial comfort zone.


a little less
conversation
a little more
ACTION
PLEASE
Elvis Presley

R workshop: import and EDA

- You will now:
  - import the data
  - explore the data.

MTPL data set

Variable	Description
`nclaims`	The number of claims filed by the policyholder.
`exp`	The fraction of the year that the policyholder was exposed to the risk.
`amount`	The total amount claimed by the policyholder.
`coverage`	Type of coverage provided by the insurance policy (TPL, PO, FO). (TPL = only third party liability, PO = TPL and limited material damage, FO = TPL and comprehensive material damage).
`fuel`	Type of fuel of the vehicle (gasoline or diesel).
`sex`	Gender of the policyholder (male or female).
`use`	Main use of the vehicle (private or work).
`fleet`	The vehicle is part of a fleet (yes or no).
`ageph`	Age of the policyholder.
`power`	Horsepower of the vehicle in kilowatt.
`agec`	Age of the vehicle.
`bm`	Level occupied in the former compulsory Belgian bonus-malus scale. Going from 0 to 22, a higher level indicates a worse claim history.
`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.

Response variables: frequency and severity

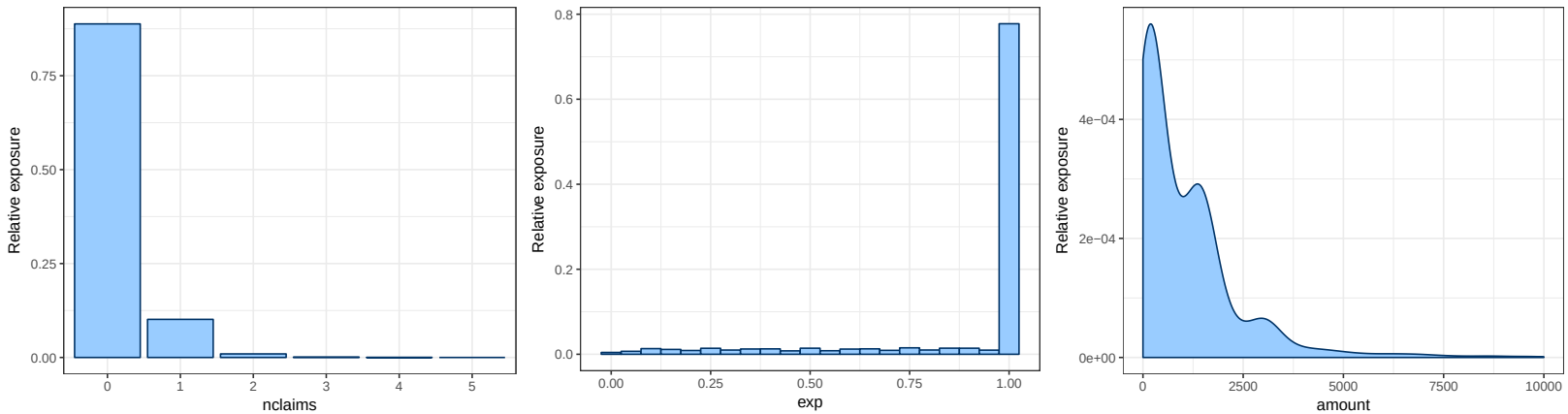
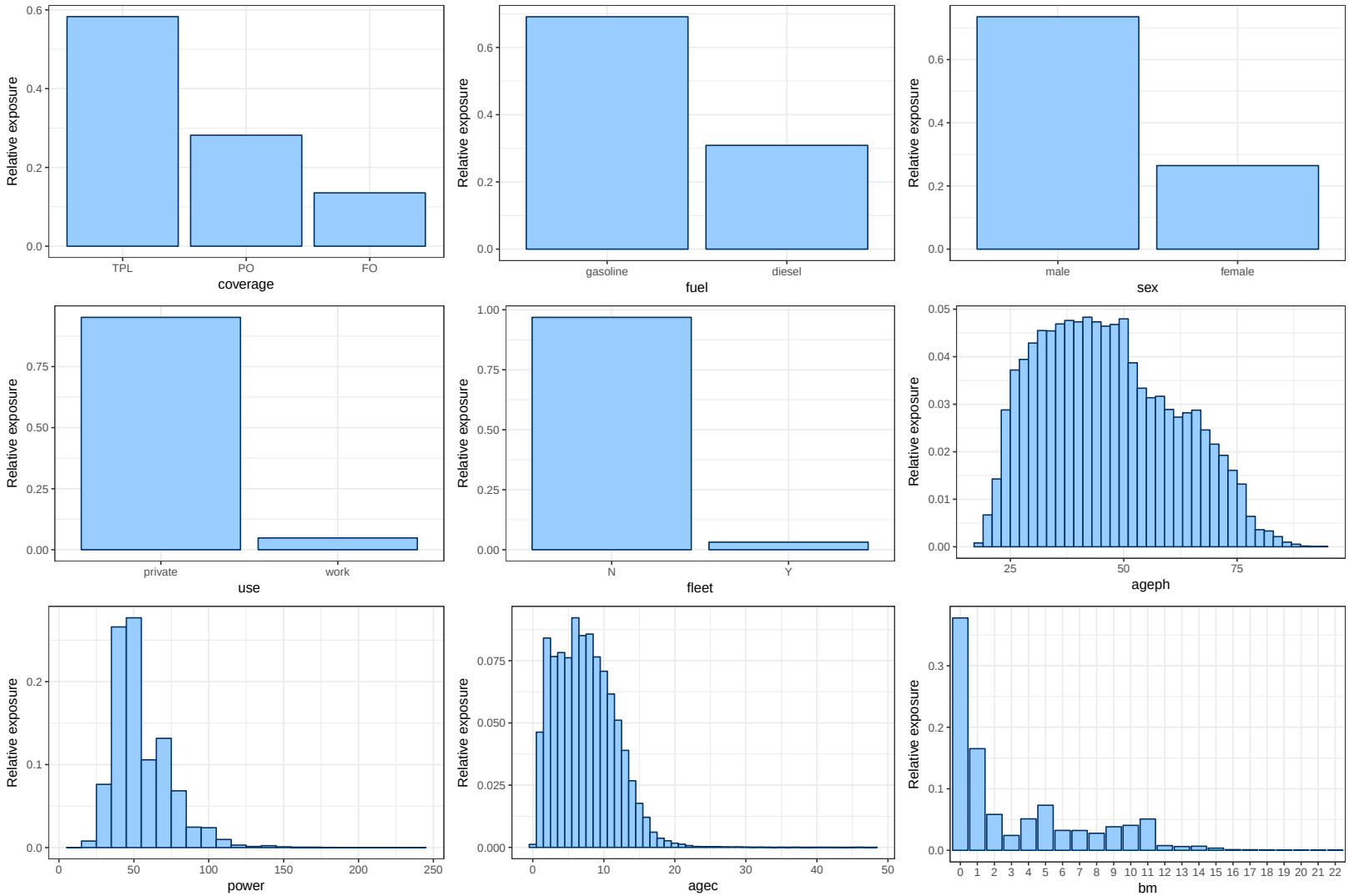
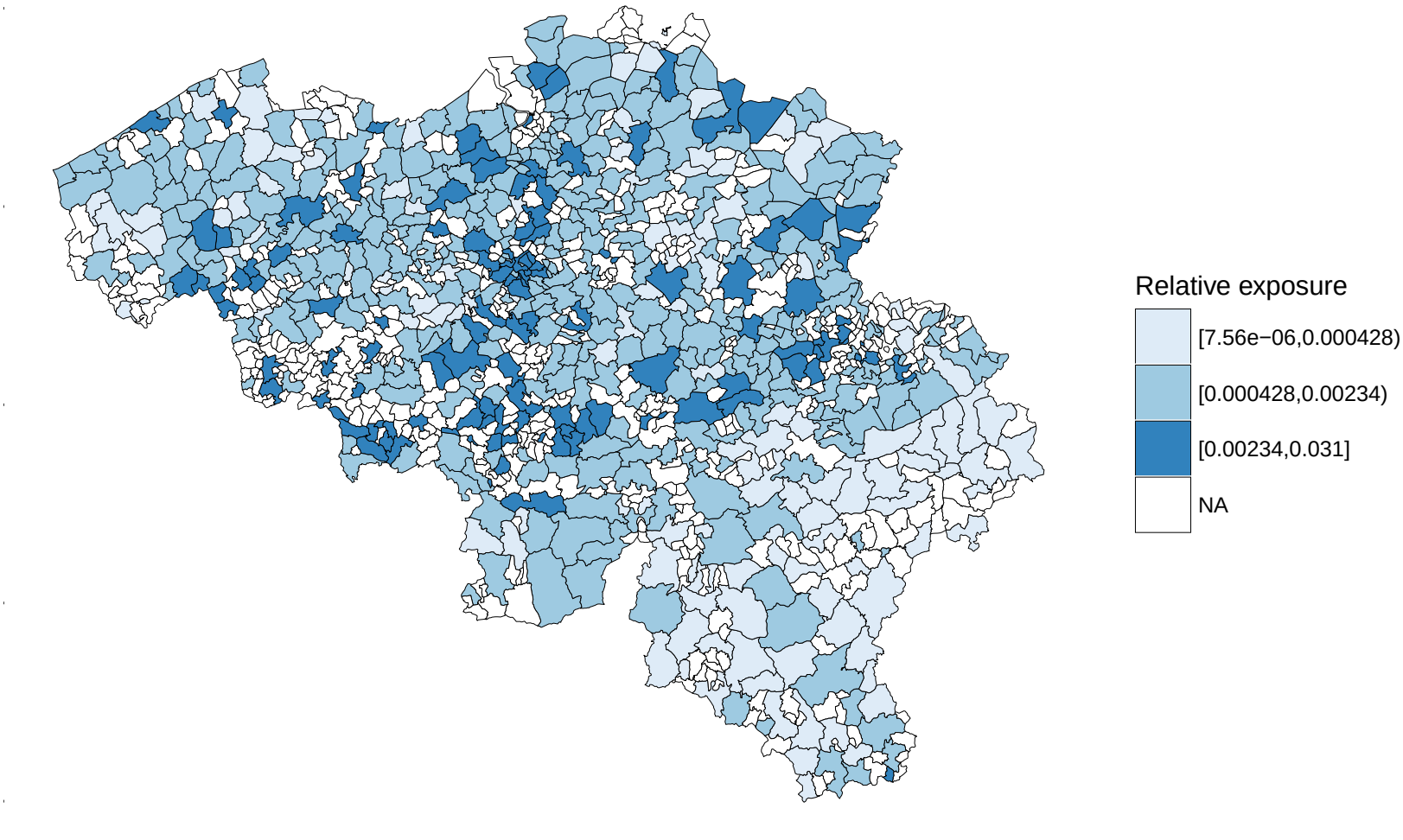


Figure: Frequency (left), exposure (middle) and severity (right).

Risk factors: categorical and continuous



Risk factors: spatial



Theoretical fundamentals: GLMs and GAMs

GLMs: the Poisson regression model

- To explain GLMs, start with a one way analysis to assess impact of a single (factor) covariate. (A toy example)
- We consider:
  - N_i: the number of claims reported by policyholder i (with $i = 1, \ldots, n$)
  - d_i: the exposure to risk of policyholder i
  - Age(i): the age category to which policyholder i belongs, i.e. factor variable with (e.g.) 4 levels.
- Frequency model for claim counts:

$$N_i \quad \sim \quad \text{POI}(d_i \cdot \lambda_{\text{Age}(i)}).$$

GLMs: the Poisson regression model

- We assume: $N_i \sim \text{Poi}(d_i \cdot \lambda_{\text{Age}(i)})$.
- Thus:

$$P(N_i = y) \quad = \quad \exp\left(-d_i \cdot \lambda_{\text{Age}(i)}\right) \frac{\left(d_i \cdot \lambda_{\text{Age}(i)}\right)^y}{y!},$$

with $y = 0, 1, 2, \ldots$.

GLMs: the Poisson regression model

- Use Maximum Likelihood Estimation (MLE):

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) &= \prod_{i=1}^{n} P(N_i = y_i) = \prod_{i=1}^{n} \exp\left(-d_i \cdot \lambda_{\text{Age(i)}}\right) \frac{\left(d_i \cdot \lambda_{\text{Age}(i)}\right)^{y_i}}{y_i!} \\ &\propto \prod_{j=1}^{4} \exp\left(-\lambda_j \sum_{i|\text{Age}(i)=j} d_i\right) \lambda_j^{\sum_{i|\text{Age}(i)=j} y_i}.\end{aligned}$$

- Differentiate $L(\boldsymbol{\lambda}) = \ln \mathcal{L}(\boldsymbol{\lambda})$ with respect to λ_j and set derivative equal to 0.

- Then, $\hat{\lambda}_j = \frac{\sum_{i|\text{Age}(i)=j} y_i}{\sum_{i|\text{Age}(i)=j} d_i}$, for $j = 1, 2, 3, 4$, with $\hat{\lambda}_j$ the expected annual claim frequency.

GLMs: the Poisson regression model

- With multiple risk factors, instead of just one ...
- Denote $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ the vector with observable characteristics of policyholder i.
- The Poisson regression model:

$$
\begin{aligned}
E[N_i|\boldsymbol{x}_i] &= d_i \cdot \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) \\
N_i &\sim \text{Poi}\left(d_i \cdot \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right),
\end{aligned}
$$

with $i = 1, \ldots, n$.

GLMs: the Poisson regression model

- Use MLE to obtain $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{\boldsymbol{p}})$.
- Solve the maximum likelihood equations numerically

  (e.g. with Newton-Raphson or Fisher scoring).

GLMs: the Poisson regression model

- Variable and model selection tools:
  - drop in deviance analysis (with nested models)
  - Wald test
  - AIC or BIC:

$$\text{AIC} = -2 \cdot \log \mathcal{L}(\boldsymbol{\beta}) + 2 \cdot \dim(\boldsymbol{\beta})$$
$$\text{BIC} = -2 \cdot \log \mathcal{L}(\boldsymbol{\beta}) + \log(n) \cdot \dim(\boldsymbol{\beta}),$$

  where smaller is better.

GLMs: the Poisson regression model

- The Poisson deviance then becomes

$$\begin{aligned} D(\boldsymbol{y}, \hat{f}(\boldsymbol{x})) &= 2 \cdot \ln \prod_{i=1}^{n} \exp(-y_i) \frac{y_i^{y_i!}}{y_i!} \\ &\quad -2 \cdot \ln \prod_{i=1}^{n} \exp(-\hat{f}(\boldsymbol{x}_i)) \frac{\hat{f}(\boldsymbol{x}_i)^{y_i}}{y_i!} \\ &= 2 \sum_{i=1}^{n} \left(y_i \cdot \ln \frac{y_i}{\hat{f}(\boldsymbol{x}_i)} - (y_i - \hat{f}(\boldsymbol{x}_i)) \right), \end{aligned}$$

where $\hat{f}(\boldsymbol{x}_i) = d_i \cdot \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})$.

GLMs: general framework

- Response Y_i and independent variables $x_i^{'} = (x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$.

(1) Random component: Y_i independent with density from the exponential family, i.e.

$$f(y; \theta, \phi) \quad = \quad \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

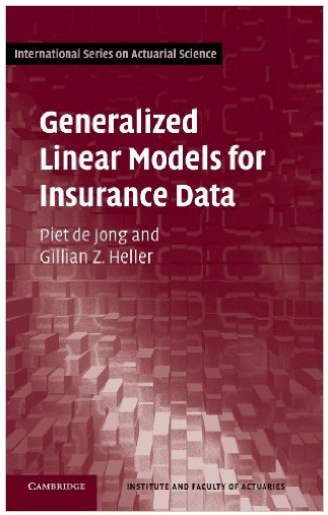
Here ϕ is a dispersion parameter and functions $b(.)$, $a(.)$ and $c(.,.)$ are known.

(2) Systematic component: $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ the linear predictor, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ regression parameters.

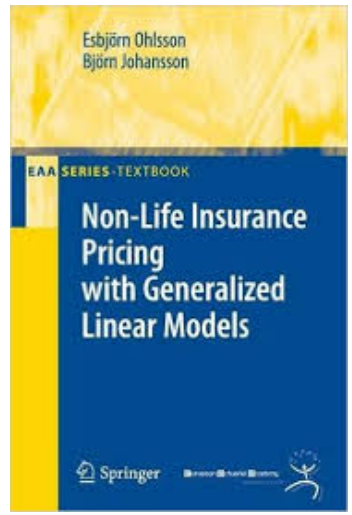
(3) Link function: $g(\mu_i) = \eta_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}$ links the linear predictor to the mean $\mu_i = E[Y_i]$.

GLMs: references

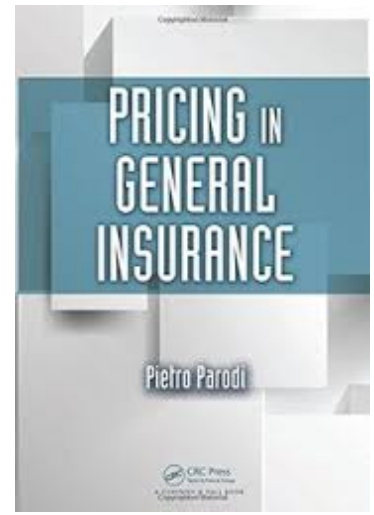
- Non–life insurance pricing with GLMs:



de Jong & Heller



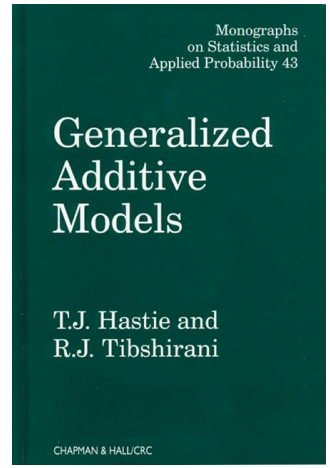
Ohlsson & Johansson

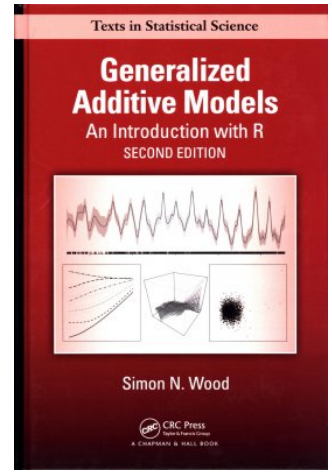


Parodi

- See also the lecture sheets by prof. Claudia Czado on GLMs.

From GLMs to GAMs





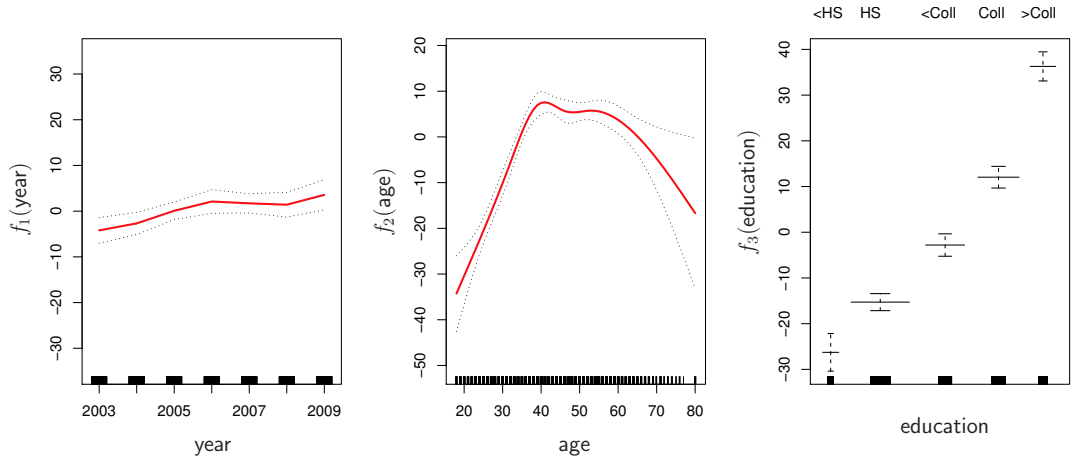
- Generalized Linear Models (GLMs):
  - transformation of the mean modelled by a linear predictor (say $\boldsymbol{x}_i'\boldsymbol{\beta}$)
  - not well suited for continuous risk factors that relate to the response in a non-linear way.
- Generalized Additive Models (GAMs):
  - allow for smooth effects of continuous risk factors and spatial effect in the predictor.

Moving beyond linearity

- The goal is to relax the linearity assumption while still attempting to maintain as much interpretability as possible.
- We sketch possible approaches:
  - polynomial regression and step functions
  - regression splines and smoothing splines
  - Generalized Additive Models.

Moving beyond linearity

Generalized Additive Models



- The goal:
  - flexibly predict Y as a function of $X_1, \ldots, X_p$
  - instead of a single predictor.
- GAMs allow non-linear functions of the predictors, while maintaining additivity.
- Use methods discussed earlier as building blocks for fitting an additive model.

Specifying GAMs

- Generalized Additive Model with predictor:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}^d + \sum_{j=1}^{q} f_j(x_{ij}^c) + \sum_{j=1}^{r} f_j(x_{ij}^s, y_{ij}^s),$$

  where μ_i is the mean of the response (e.g. claim frequency or severity).

- The predictor includes:
  - dummy-coded factor variables
  - smooth functions $f_j(.)$ of one-dimensional continuous x_{ij}^c
  - smooth functions $f_j(.,.)$ of two-dimensional continuous (x_{ij}^s, y_{ij}^s).

Fitting GAMs

- Estimate with cubic regression splines.
- Represent $f_j(x)$ as $\sum_{m=1}^{M} \beta_{jm} \cdot b_{jm}(x)$ with basis functions $b_{jm}(x)$.
- To avoid overfitting, a wiggliness penalty is added to the log-likelihood. That is

$$\int f_j''(x)^2 dx \quad = \quad \beta_j^t \boldsymbol{S}_j \beta_j.$$

Fitting GAMs

- Estimate with thin plate splines.
- Represent $f_j(x,y)$ as $\sum_{n=1}^{N} \gamma_{jn} \cdot \tilde{b}_{jn}(x,y)$ with basis functions $\tilde{b}_{jn}(x,y)$.
- To avoid overfitting, a wiggliness penalty is added to the log-likelihood. That is

$$\int\int \left(\frac{\partial^2 f_j}{(\partial x)^2}\right)^2 + 2\left(\frac{\partial^2 f_j}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f_j}{(\partial y)^2}\right)^2 dxdy \quad = \quad \gamma_j^t \boldsymbol{T}_j \gamma_j.$$

Fitting GAMs

- The penalized log-likelihood

$$\log \mathcal{L}(\boldsymbol{\beta}) - \sum_j \lambda_j \beta_j^t \boldsymbol{S}_j \beta_j - \lambda_s \gamma^t \boldsymbol{T} \gamma.$$

- Smoothing parameters λ_j control trade-off goodness of fit and degree of smoothness.

- Select smoothing parameters using (a.o.) generalized cross-validation or Akaike's Information Criterion (AIC).

- Information criteria for variable and model selection:

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

  balancing goodness of fit and complexity.

Construction of tariff classes: GAM as starting point

GAM as a starting point

- MTPL data set from Denuit & Lang (2004), 163 231 records.
- Lowest BIC among exhaustive search with 1 024 fitted models:

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

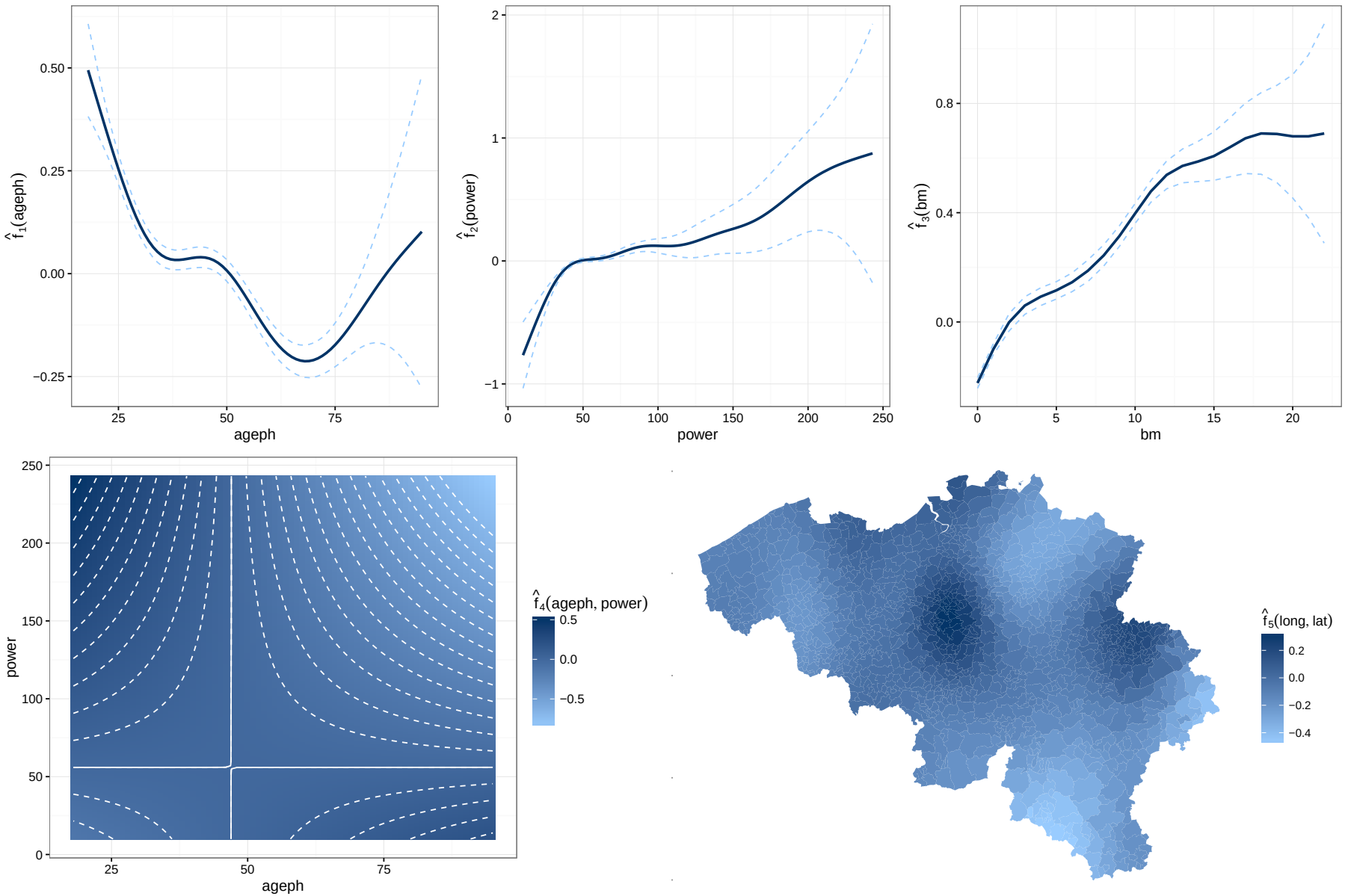
which combines offset and

categorical $\sim$ nominal continuous $\sim$ ordinal

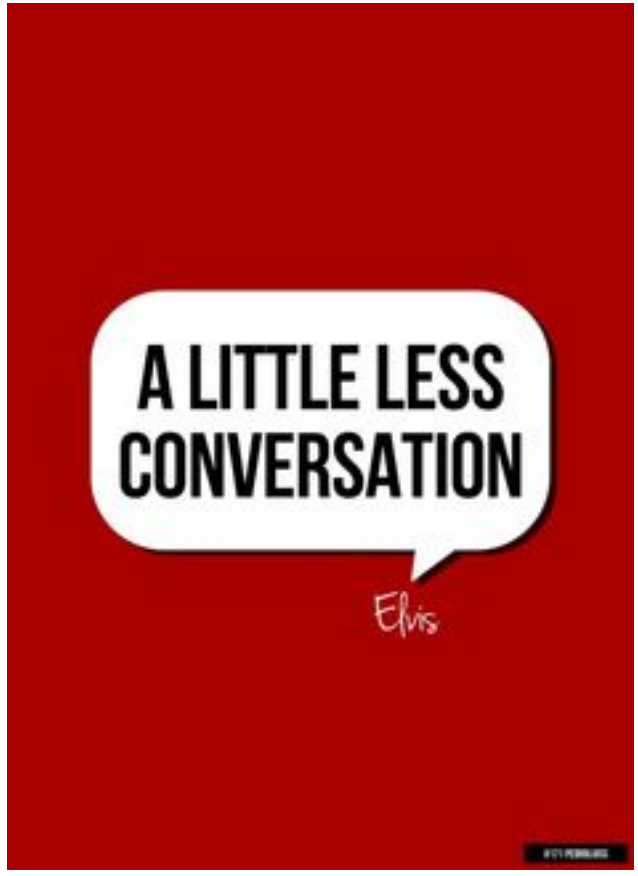
interactions spatial

risk factors.

GAM as a starting point



R demos: GLMs and GAMs



R workshop: fitting GAMs on frequency data

- You will now focus on:
  - fitting GAMs for frequency data, including:

    factor variables, smooth effect of continuous risk factors, interaction effects and spatial effect
  - extracting and visualizing the smooth terms.


Construction of tariff classes: bin the spatial effect

Simplify the GAM: bin the spatial effect

- Use the fitted spatial effect $\mathbf{f_5}(\texttt{long}, \texttt{lat})$.
- Split or bin using one of the following: (also see `classint` package in `R`)
  - equal intervals
  - quantile binning
  - complete linkage (see Kaufman & Rousseeuw, 1990)
  - Fisher's natural breaks (see Fisher, 1958 and Slocum et al., 2005).
- We compare the homogeneity of the class intervals ('the bins') using the goodness of variance fit (GVF) and the tabular accuracy index (TAI) (see Armstong et al., 2003).

Simplify the GAM: bin the spatial effect

- The methods in `classint` applied to $s_i = \mathbf{f_5}(\texttt{long}_i, \texttt{lat}_i)$ for $i = 1, \ldots, 1\ 146$.
  - equal intervals: k bins of equal length $\frac{\max(s_i) - \min(s_i)}{k}$
  - quantile binning: each bin contains $\approx \frac{1\ 146}{k}$ municipalities
  - complete linkage ($\sim$ hierarchical clustering):

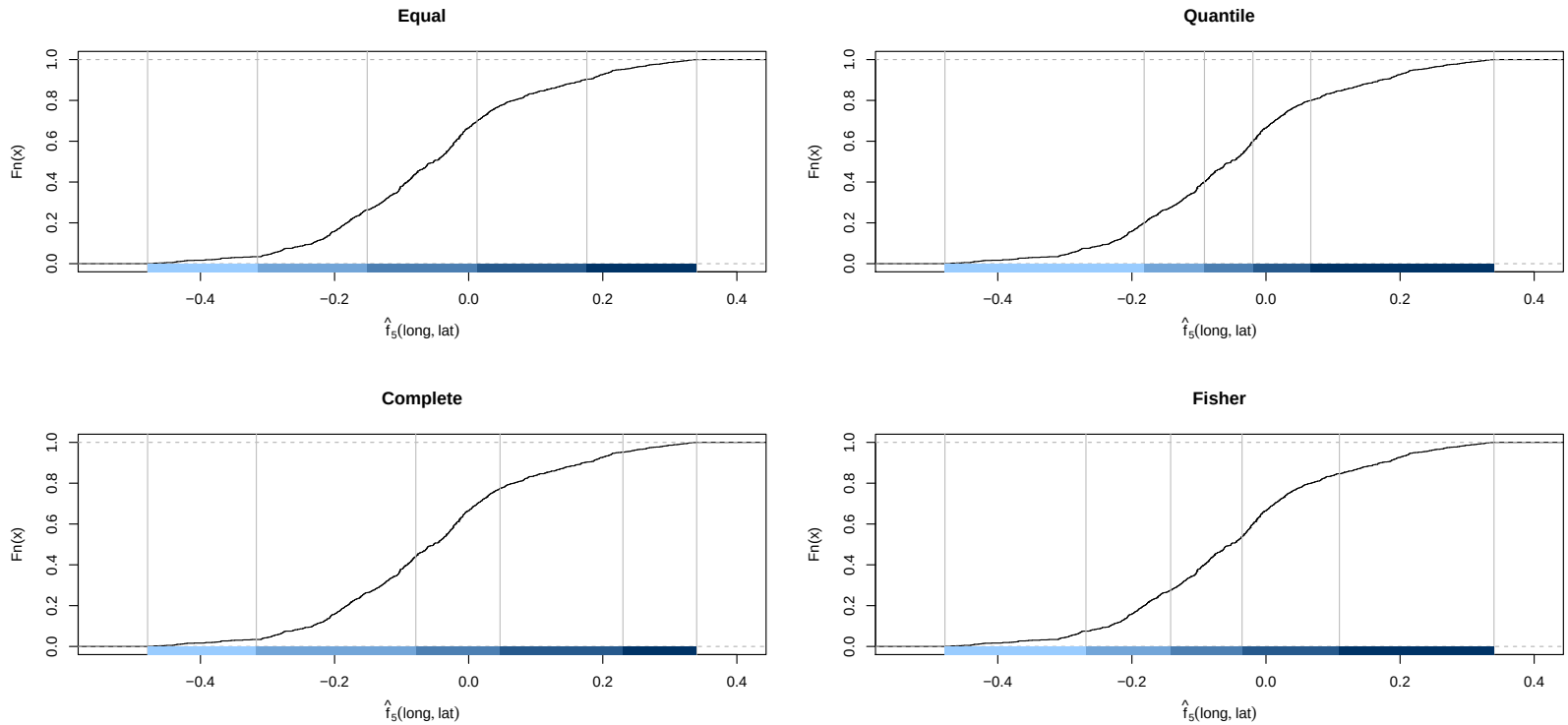
    each municipality initially forms its own bin

    then merge bins closest to each other
  - Fisher's natural breaks ($\sim$ K-means clustering):

    minimize the sum of squared distances between observations $s_u^{(i)}$ and the bin means

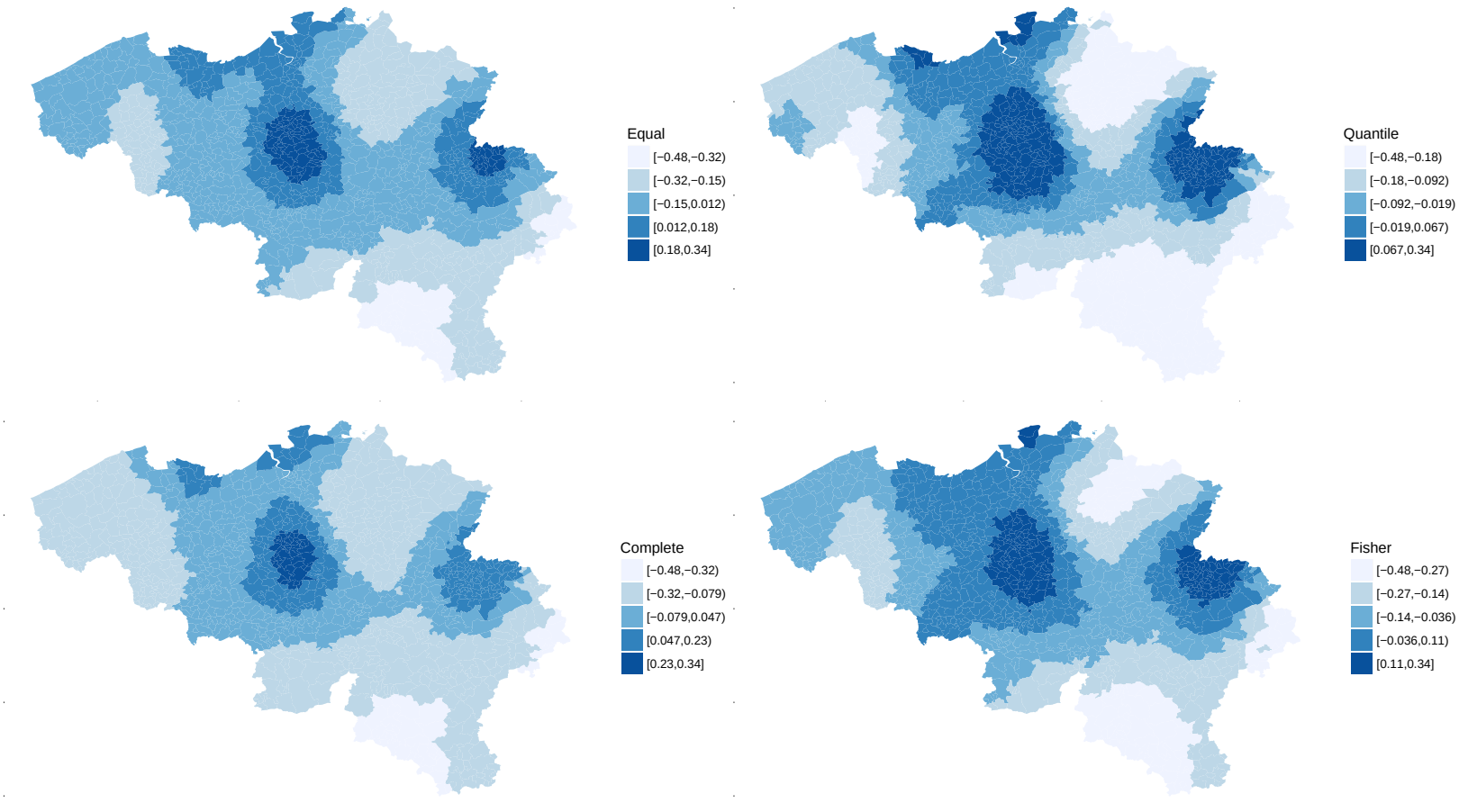
$$\sum_{i=1}^{k} \sum_{u=1}^{n_i} (s_u^{(i)} - \bar{s}^{(i)})^2.$$

Simplify the GAM: bin the spatial effect



Use Fisher's natural breaks, but how many bins?

Simplify the GAM: bin the spatial effect



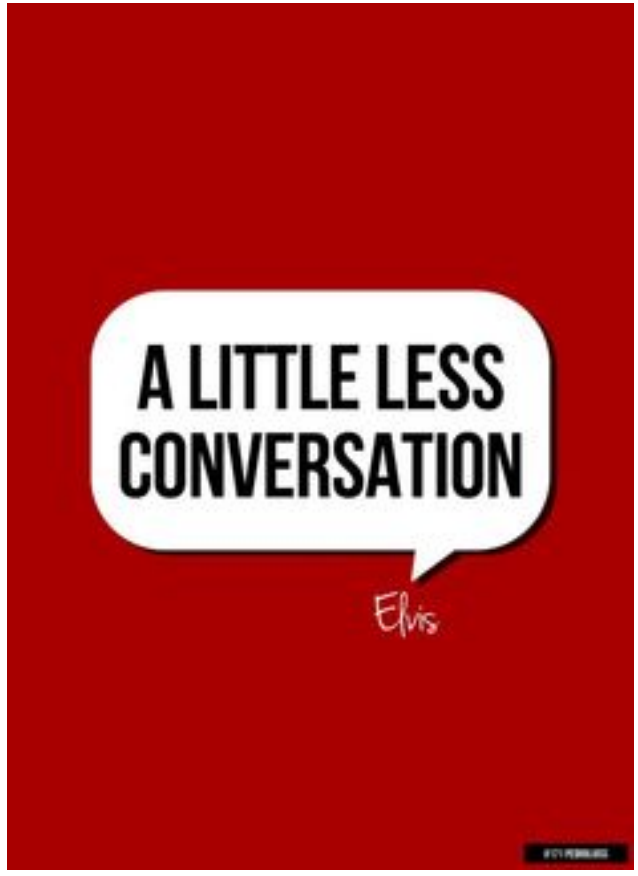
Use Fisher's natural breaks, but how many bins?

Simplify the GAM: bin the spatial effect

Procedure:	Find the optimal number of bins for the spatial effect
Step 1	Apply Fisher's algorithm to calculate the class interval breaks for the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$, for a given number of bins. These class interval breaks are used to transform the continuous spatial effect into a categorical spatial effect.
Step 2	Estimate a new GAM with bins of the spatial effect.
Step 3	Calculate the BIC and AIC of the newly fitted GAM.

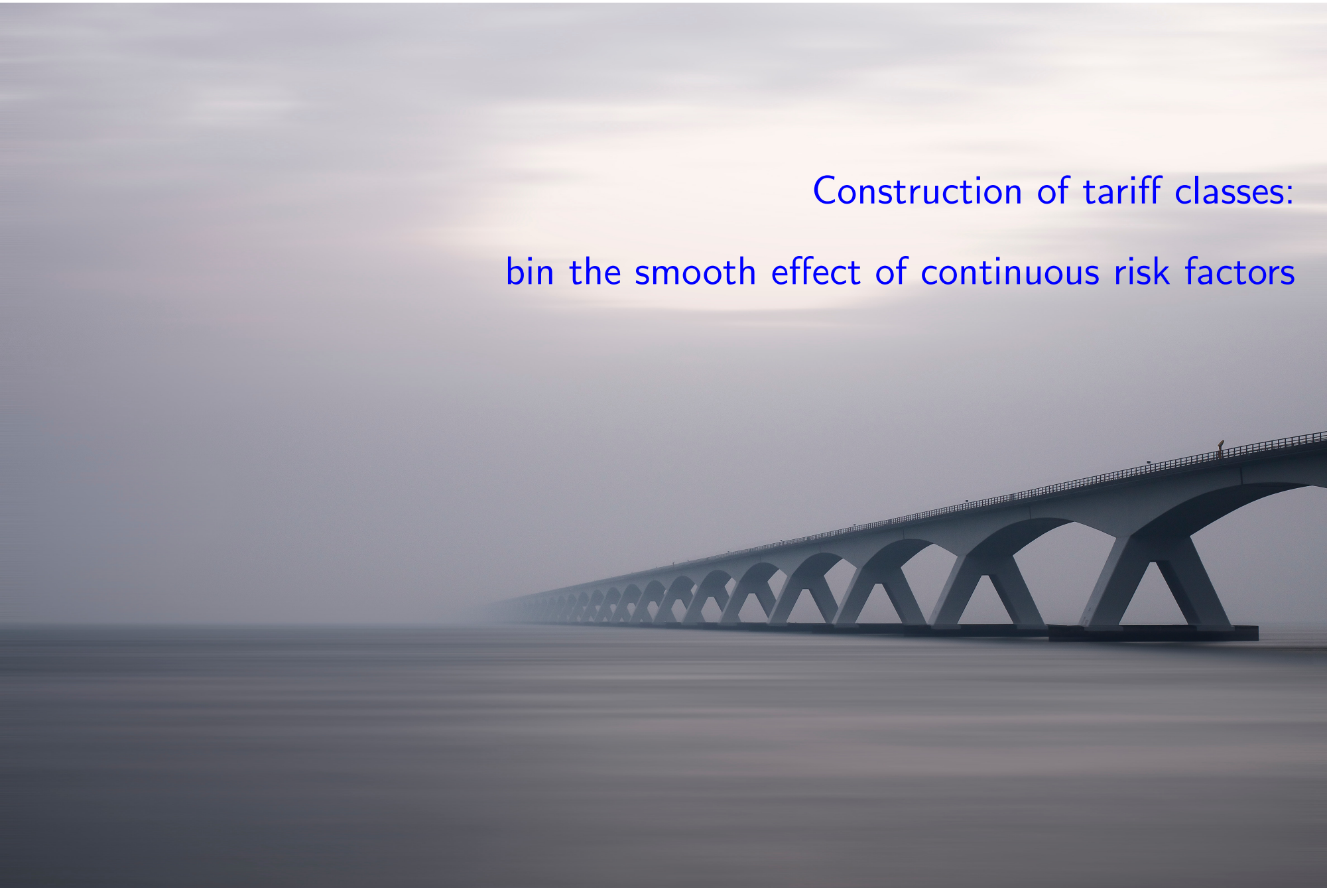
# bins	BIC	AIC
2	125047.6	124778.9
3	125023.9	124753.1
4	124928.4	124652.3
5	**124907.2**	**124621.3**
6	124921.6	124627.7
7	124942.9	124639.1

R demos: bin the spatial effect



R workshop: bin the spatial effect

- You will now focus on:
  - using the `classInt` package to bin the fitted spatial effect;
  - store the resulting geographical zones as factor information in the data frame.



Construction of tariff classes:

bin the smooth effect of continuous risk factors

Simplify the GAM: bin the continuous risk factors

- We want consecutive intervals for the continuous risk factors
  - method to bin or split the spatial effect is not applicable.
- We use evolutionary trees, combining decision trees with genetic algorithms:
  - in contrast to the greedy approach of recursive partitioning trees, splits can be changed;
  - global optimum obtained.
- Important to take the composition of the insurance portfolio into account:
  - use the number of policyholders as weights.

Simplify the GAM: bin the continuous risk factors

- We fit evolutionary trees to the single and interaction effects:

$$\hat{f}_1(\texttt{ageph}) \qquad \hat{f}_2(\texttt{power}) \qquad \hat{f}_3(\texttt{bm}) \qquad \hat{f}_4(\texttt{ageph}, \texttt{power}),$$

- Evaluation criterion:

$$n \cdot \log(\text{MSE}) \; + \; \alpha \cdot \text{complexity penalty},$$

where

  - n is the number of observations (or: the total sum of weights);
  - MSE is the Mean Squared Error (or: a weighted MSE);
  - α is a tuning parameter;
  - the complexity of the tree is its number of leaf nodes.

Simplify the GAM: bin the continuous risk factors

- In our setting:

Covariate: ageph	Response: $\hat{f}_1(\text{ageph})$	Weight: w
18	0.495	16
19	0.459	116
20	0.424	393

and

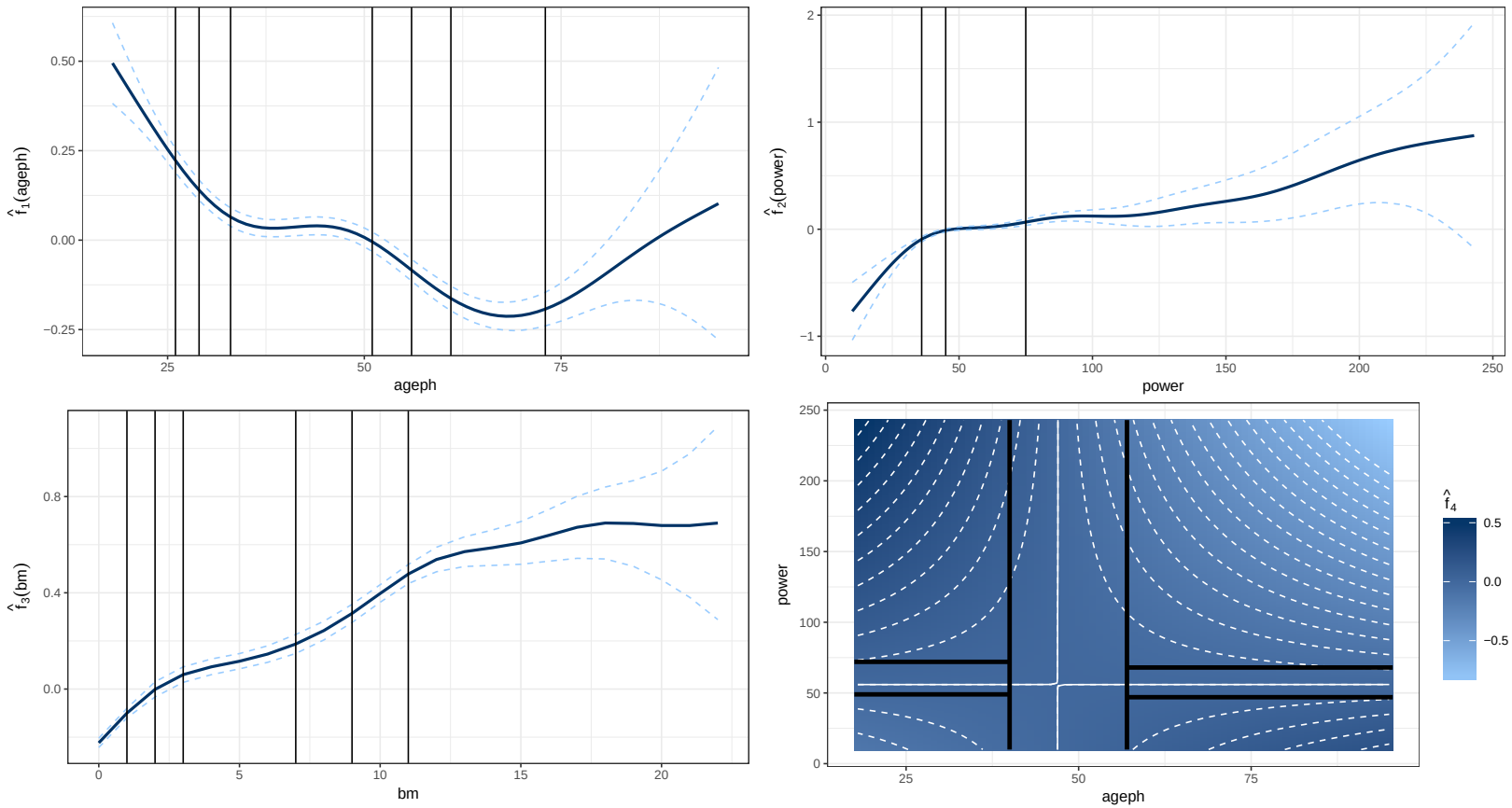
$$\text{MSE} = \frac{\sum_{i=\min(\text{ageph})}^{\max(\text{ageph})} w_{\text{ageph}_i} (\hat{f}_1(\text{ageph}_i) - \hat{f}_1^b(\text{ageph}_i))^2}{\sum_{i=\min(\text{ageph})}^{\max(\text{ageph})} w_{\text{ageph}_i}}.$$

Simplify the GAM: bin the continuous risk factors

- We propose a strategy to tune the parameter α.
- This tuning process then determines the optimal number of splits or bins per fitted effect.
- Hence, we obtain a fully data-driven procedure to split the continuous risk factors.

Procedure:	Find the optimal tuning parameter α for the evolutionary trees
Step 1	Fit an evolutionary tree to every single and interaction effect, $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$ and $\hat{f}_4(\texttt{ageph}, \texttt{power})$, for a given value of α. The splits produced by these trees are used to transform the continuous single and interaction effects into categorical effects.
Step 2	Estimate a new GLM with all risk factors in categorical format.
Step 3	Calculate the AIC of the GLM.

Simplify the GAM: bin the continuous risk factors

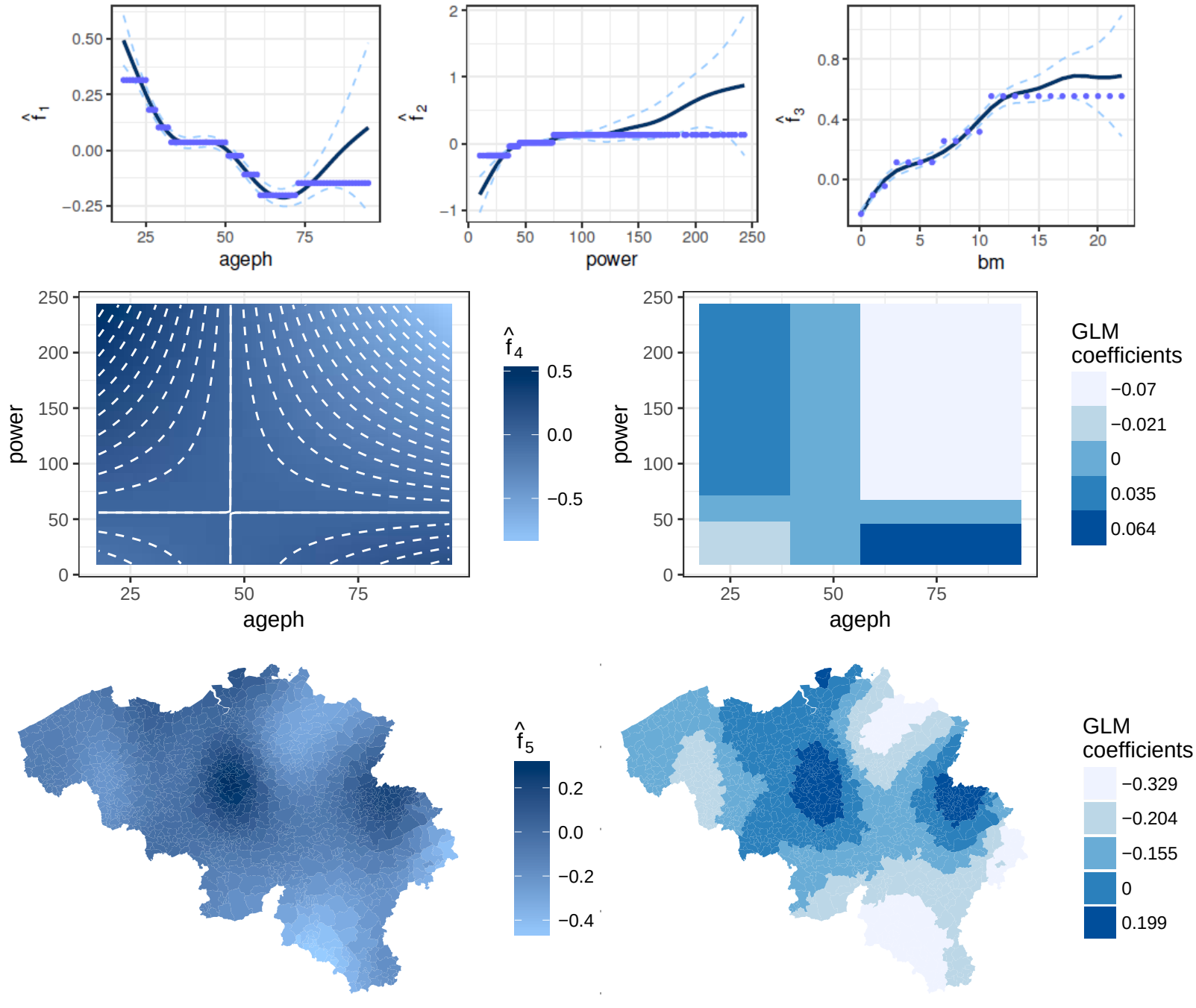


R demos: bin the continuous risk factors

R workshop: bin the smooth effects of continuous risk factors

- You will now focus on:
  - the use of `evtree` to bin a fitted smooth effect
  - interpreting the resulting trees and using them to transform a continuous variable into a factor one.

Construction of tariff classes: putting it all together

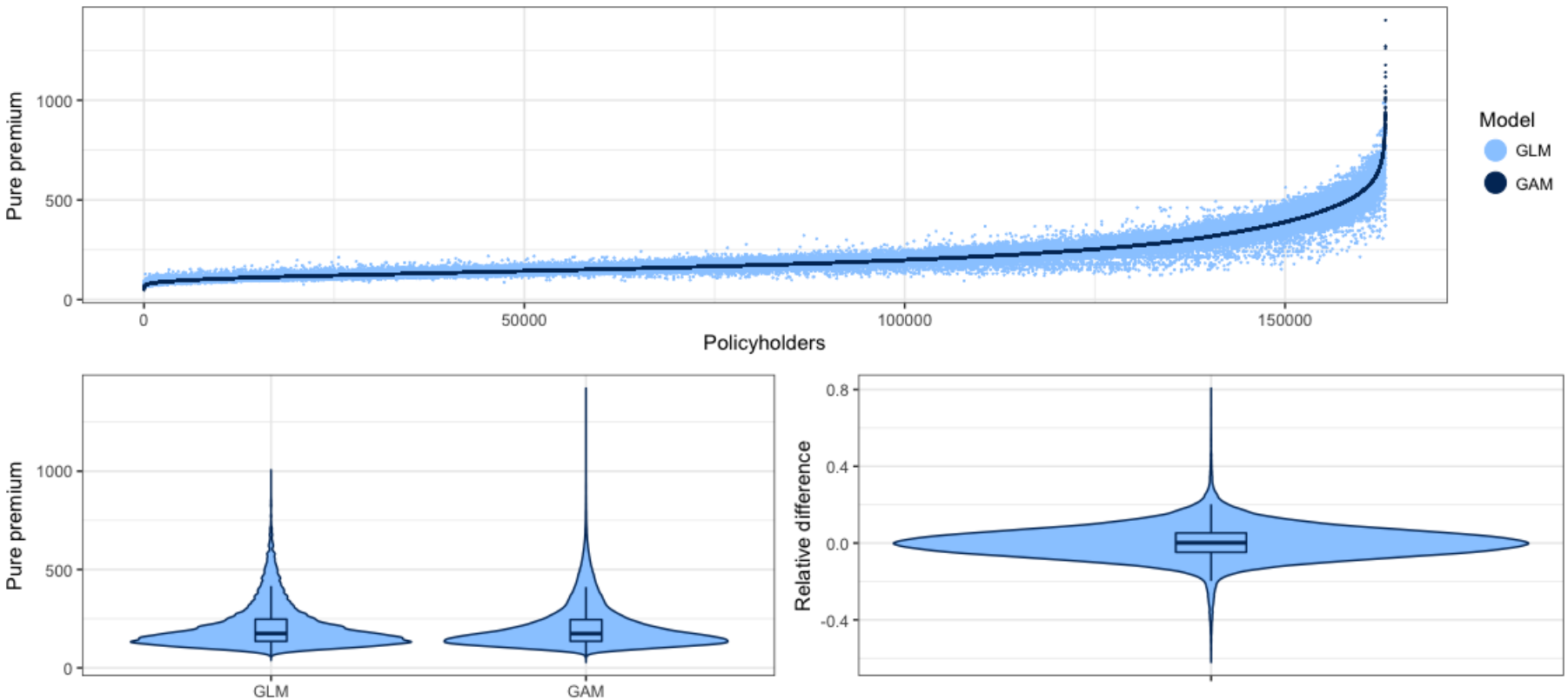

$\hat{f}_1$
ageph
$\hat{f}_2$
power
$\hat{f}_3$
bm
power
ageph
$\hat{f}_4$
power
ageph
GLM coefficients
−0.07
−0.021
0
0.035
0.064
$\hat{f}_5$
GLM coefficients
−0.329
−0.204
−0.155
0
0.199

Comparison GAM vs GLM

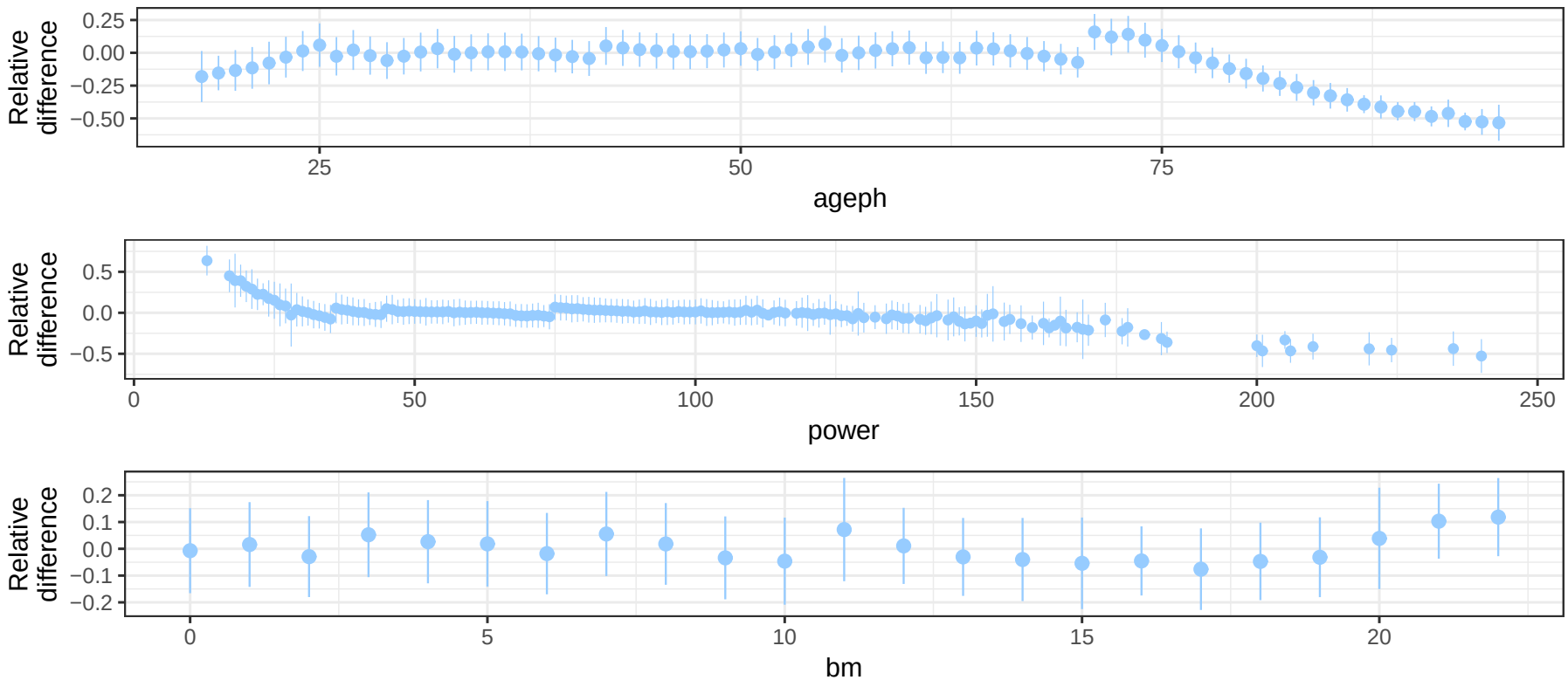
- At portfolio level, total pure premium is 29 867 565 (GLM) and 29 865 859 (GAM) euro.
- Actual losses in the portfolio is 26 464 970 euro (with very large losses excluded).
- Consider relative premium differences:

$$\frac{\pi_i^{GLM} - \pi_i^{GAM}}{\pi_i^{GAM}}.$$

Comparison GAM vs GLM

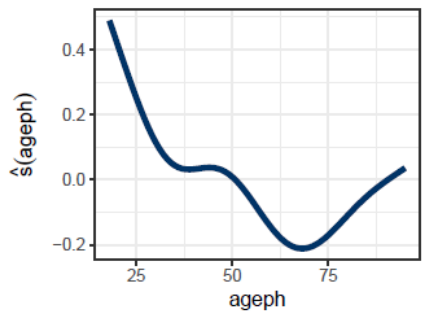


Comparison GAM vs GLM

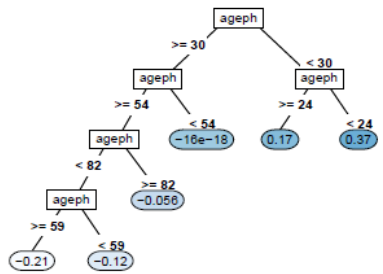


Research contributions Katrien's lab

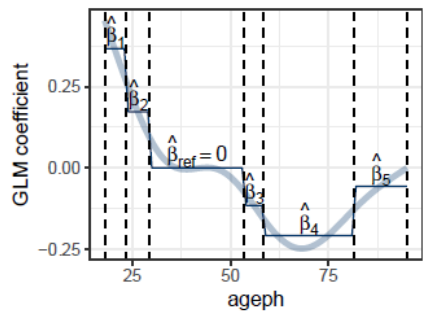
A data driven binning strategy for the construction of insurance tariff classes by Henckaerts, Antonio et al. (2018)



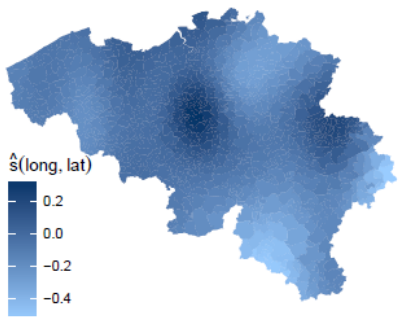
(1a) Smooth continuous effect



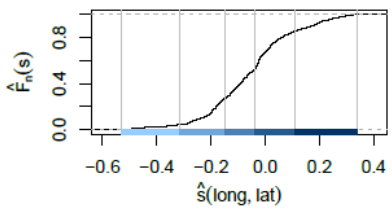
(1b) Supervised decision tree



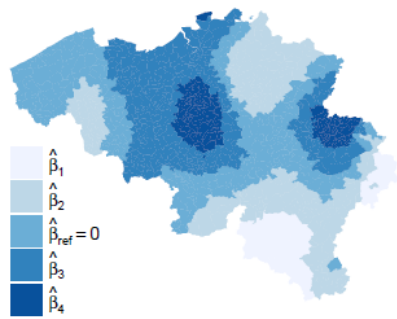
(1c) Binned continuous effect



(2a) Smooth spatial effect



(2b) Unsupervised clustering



(2c) Binned spatial effect

Research contributions Katrien's lab

Sparse regression with multi-type feature modelling by Devriendt, Antonio et al. (2018)

- automatic feature selection and binning of risk factors via regularization
- R package `smurf` on CRAN
- end product is a GLM!

Research contributions Katrien's lab

Boosting insights in insurance tariff plans with tree-based machine learning
by Henckaerts, Côté, Antonio et al. (2019)

- GLMs, GAMs, decision trees, random forests and gradient boosting machines
- R packages extended to Poisson and gamma deviance
- tuning strategy, interpretability tools and managerial insights
- supplementary material on GitHub.

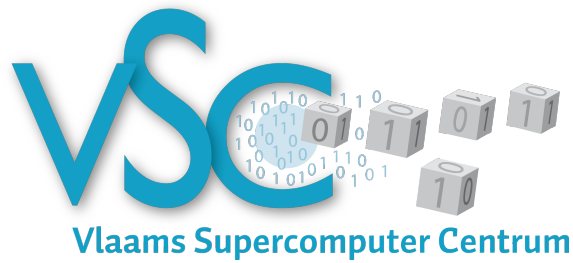
More information

For more information, please visit:

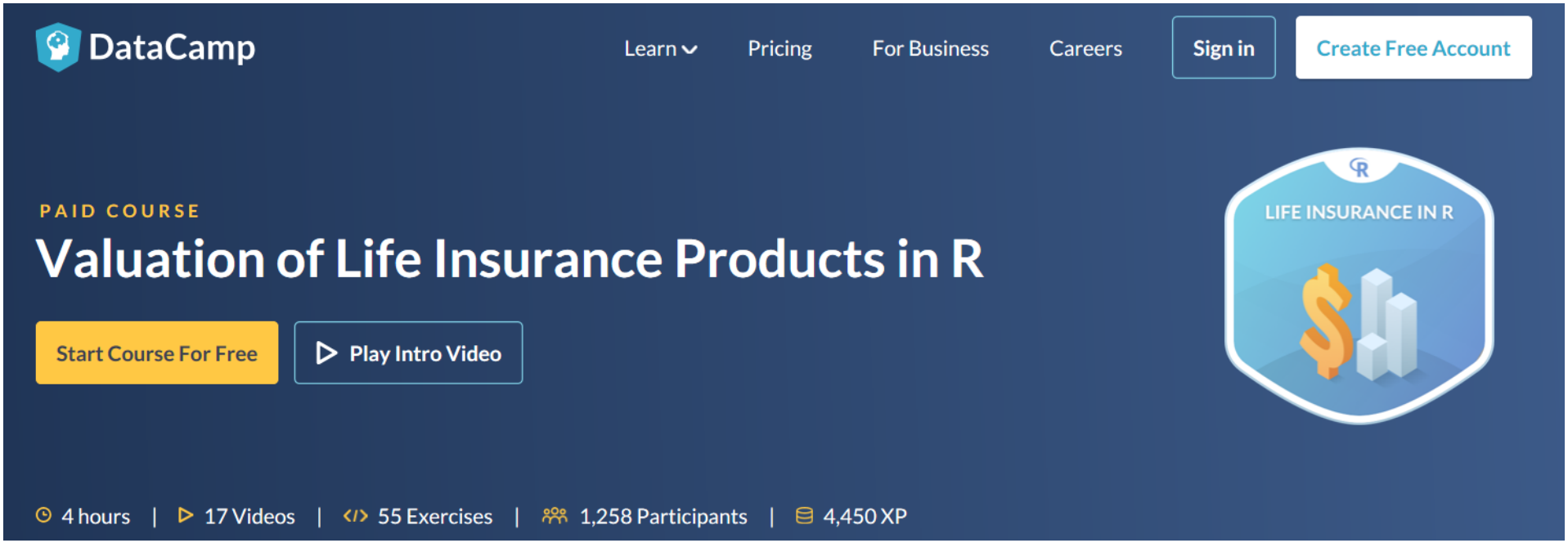
LRisk website, www.lrisk.be

https://katrienantonio.github.io

Thanks to



Research Foundation
Flanders
Opening new horizons



Online course with DataCamp on Valuation of Life Insurance Products in R

designed by Katrien Antonio & Roel Verbelen

http://www.datacamp.com/courses/2333

References


Henckaerts, R., Antonio, K., Clijsters, M. and Verbelen, R. (2018)
A data driven strategy for the construction of insurance tariff classes.
Scandinavian Actuarial Journal


Wood, S. (2006)
Generalized additive models: an introduction with R.
Chapman and Hall/CRC Press.


Gertheiss, J. and Tutz, G. (2010).
Sparse modeling of categorial explanatory variables.
The Annals of Applied Statistics, 4(4), 2150-2180.


Oelker, M. and Gertheiss, J. (2017).
A uniform framework for the combination of penalties in generalized structured models.
Advances in Data Analysis and Classification, 11(1),97-120.

References


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.


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


Parikh, N. and Boyd, S. (2013).
Proximal algorithms.
Foundations and Trends in Optimization, 1(3):123-231.


Hastie, T., Tibshirani, R. and Wainwright, M. (2015)
Statistical learning with sparsity: the Lasso and generalizations.
Chapman and Hall/CRC Press.

References


Breiman, L. (2001).
Random forests.
Machine learning, 45(1):5–32.


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


Friedman, J. H. (2001).
Greedy function approximation: a gradient boosting machine.
Annals of statistics, pages 1189–1232.