Data driven strategies for the construction of insurance tariff classes

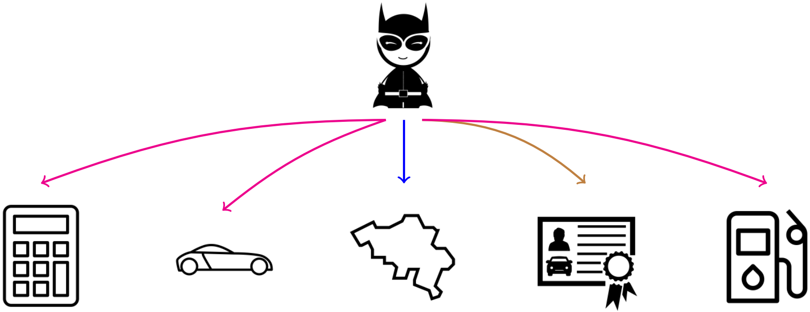
SAA Annual Meeting 2018 in Zurich

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

August 31, 2018


UNIVERSITEIT VAN AMSTERDAM
AMSTERDAM SCHOOL OF ECONOMICS
Economics
KU LEUVEN
LRISK

Motivation



Claim frequency and claim severity

as function of

nominal / numeric ∼ ordinal / spatial

features

Research questions

- Comfort zone:

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

Research questions

- Comfort zone:

  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

Research questions

- Comfort zone:

  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.

Research questions - rephrased

- Comfort zone:

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

Research questions - rephrased

- Comfort zone:

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

- How to:

  (1) avoid overfitting with too many risk factors or levels?

  (2) avoid underfitting with a priori binning/selection?

Research questions - rephrased

- Comfort zone:

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

- How to:

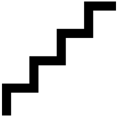
  (1) avoid overfitting with too many risk factors or levels?

  (2) avoid underfitting with a priori binning/selection?

- Procedure should be data driven, scalable to large (big) data, and automatic!

Research contributions

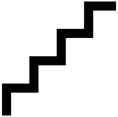
Research contributions



step-by-step

best subset
selection

Research contributions



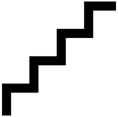
step-by-step

best subset
selection

SMuRF

sparsity
regularization

Research contributions



step-by-step

best subset selection

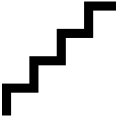
SMuRF

sparsity regularization

tree-based

CART, random forest, gradient boosting

Research contributions



step-by-step

best subset selection

SMuRF

sparsity regularization

tree-based

CART, random forest, gradient boosting

Statistical Learning

GLMs and GAMs

Machine Learning

A data driven strategy for the construction of insurance tariff classes

Henckaerts, Antonio, Clijsters & Verbelen, 2018, Scandinavian Actuarial Journal

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

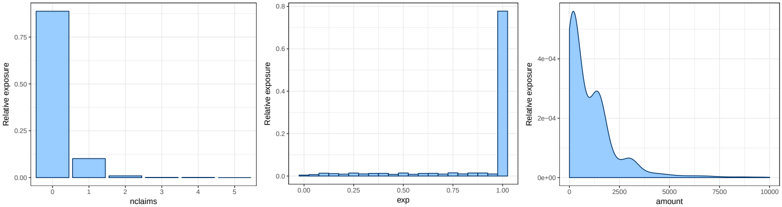
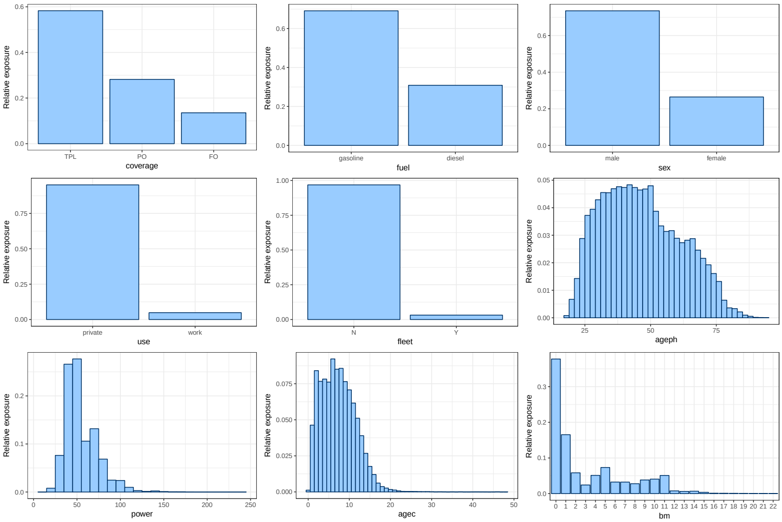
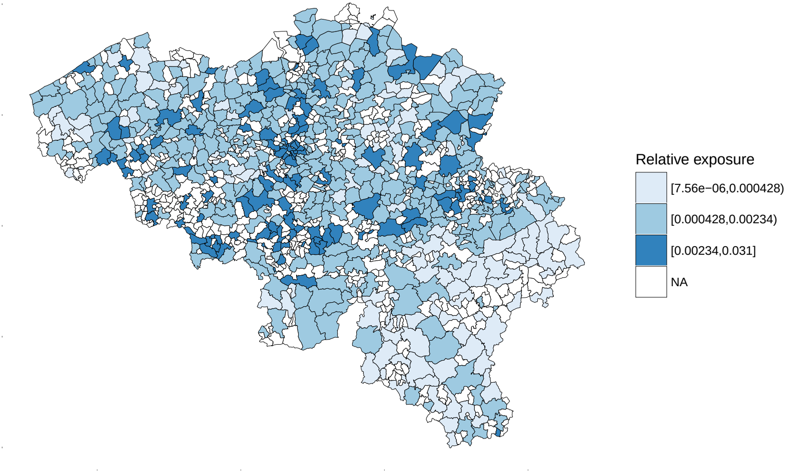


Figure: Frequency (left), exposure (middle) and severity (right).

Risk factors: factor and continuous



Risk factors: spatial



On GLMs and GAMs

- Generalized Linear Models (GLMs):
  - transformation of the mean ($g(\mu_i)$) modelled by a linear predictor ($\boldsymbol{x}_i^{'}\boldsymbol{\beta}$);
  - not well suited for continuous risk factors that relate to the response in a non-linear way.

On GLMs and GAMs

- Generalized Linear Models (GLMs):
  - transformation of the mean ($g(\mu_i)$) modelled by a linear predictor ($\boldsymbol{x}_i^{'}\boldsymbol{\beta}$);
  - 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 and spatial risk factors in the predictor.

GAM as a starting point

- Generalized Additive Model with predictor: (R package mgcv)

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

- Information criteria:

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

balancing goodness-of-fit and complexity.

- Best subset selection strategy!

MTPL data set: step-by-step solution

- Lowest BIC among exhaustive search with 1 024 fitted models:

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

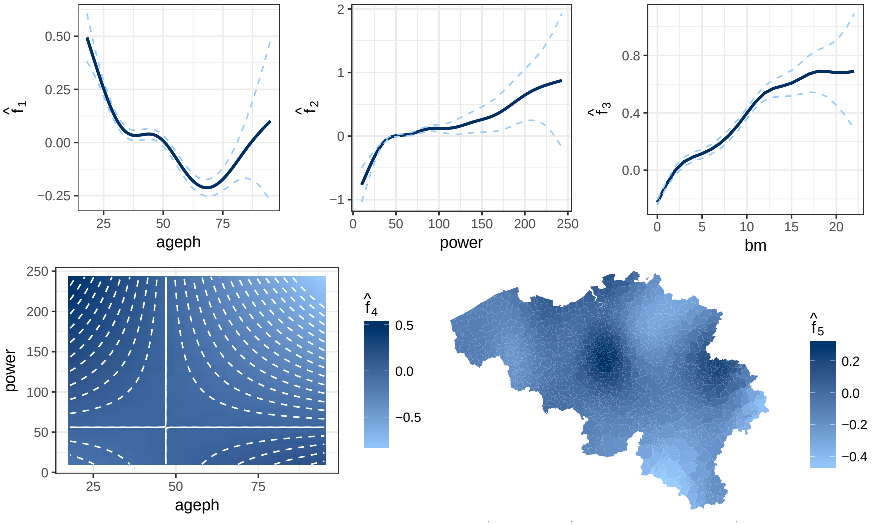
which combines offset and

categorical ∼ nominal   continuous ∼ ordinal

interactions   spatial

risk factors.

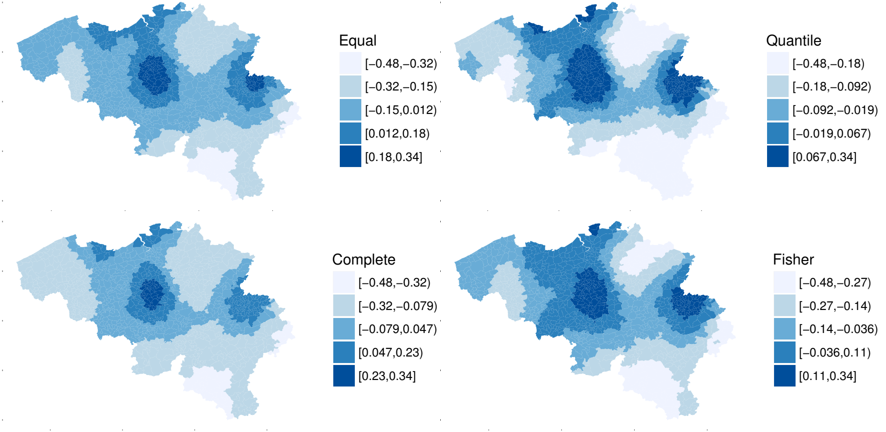
MTPL data: step-by-step solution



Bin smooth GAM effects - spatial

- Bin or cluster $\hat{f}_5(\texttt{long}_i, \texttt{lat}_i)$ for $i \in \{1, ..., 1\ 146\}$.
- We use: (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 two measures: the goodness of variance fit (GVF) and the tabular accuracy index (TAI).

Bin smooth GAM effects - spatial



Bin smooth GAM effects - spatial

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

Bin smooth GAM effects - continuous

- We want consecutive intervals for the continuous risk factors
    - method to bin or split the spatial effect is not applicable.

Bin smooth GAM effects - continuous

- 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 regression trees with genetic algorithms:
  - in contrast to the greedy approach of recursive partitioning (`rpart`) trees, splits can be changed;
  - global optimum obtained.

Bin smooth GAM effects - continuous

- 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 regression trees with genetic algorithms:
  - in contrast to the greedy approach of recursive partitioning (`rpart`) trees, splits can be changed;
  - global optimum obtained.
- We take the composition of the insurance portfolio into account:
  - use the number of policyholders as weights.

Bin smooth GAM effects - continuous

- 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:

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

where

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

Bin smooth GAM effects - continuous

- In our setting:

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

and

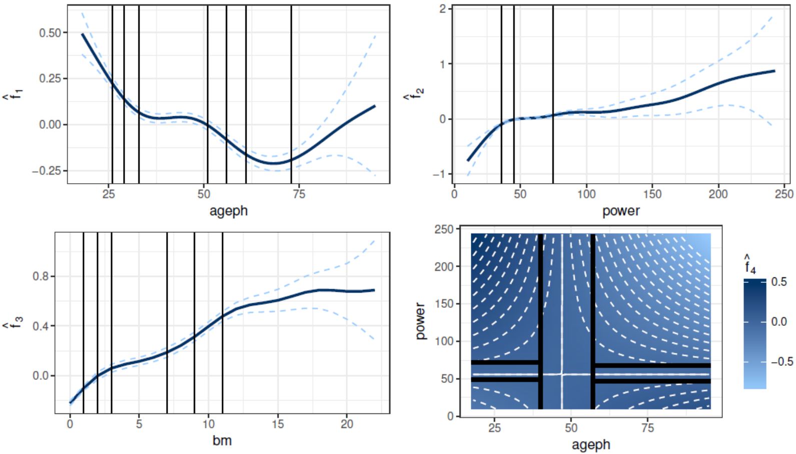
$$\text{wMSE} = \frac{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i} (\hat{f}_1(\texttt{ageph}_i) - \hat{f}_1^{b}(\texttt{ageph}_i))^2}{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i}}.$$

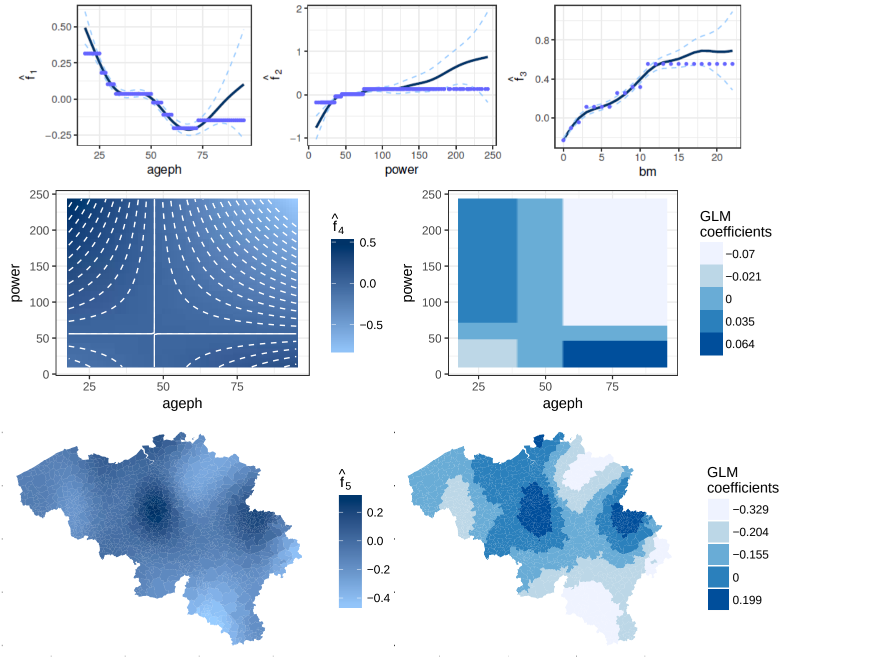
Bin smooth GAM effects - continuous

- Tuning process for α 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.

MTPL data: step-by-step solution





$\hat{f}_1$
0.50
0.25
0.00
−0.25
25
50
75
ageph
$\hat{f}_2$
2
1
0
−1
0
50
100
150
200
250
power
$\hat{f}_3$
0.8
0.4
0.0
0
5
10
15
20
bm
power
250
200
150
100
50
0
25
50
75
ageph
$\hat{f}_4$
0.5
0.0
−0.5
power
250
200
150
100
50
0
25
50
75
ageph
GLM coefficients
−0.07
−0.021
0
0.035
0.064
$\hat{f}_5$
0.2
0.0
−0.2
−0.4
GLM coefficients
−0.329
−0.204
−0.155
0
0.199

Sparsity with multi-type Lasso penalties

Devriendt, Antonio, Frees, Reynkens & Verbelen, 2018 (in progress)

Lasso

- Less is more: (Hastie, Tibshirani & Wainwright, 2015)

  a sparse model is easier to estimate and interpret than a dense model.

- Regularize (with budget constraint t, or regularization parameter λ):

$$\min_{\beta_0,\boldsymbol{\beta}} \{-\log \mathcal{L}\} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t,$$

  or equivalenty (L_1 or lasso penalty)

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{ -\log \mathcal{L} + \lambda \cdot \sum_{j=1}^{p} |\beta_j| \right\}.$$

  Shrinks coefficients and even sets some to zero.

Lasso and friends

- Adjust lasso regularization to the type of risk factor:
  - Determine type (nominal / numeric $\sim$ ordinal / spatial)
  - Allocate logical penalty.
- Thus, for J risk factors, each with convex regularization term $g_j(.)$, we want to optimize:

$$-\log \mathcal{L}\left(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J\right) + \lambda \cdot \sum_{j=1}^{J} g_j\left(\boldsymbol{\beta}_j\right).$$

A multi-type regularized predictive model!

Regularization with multi-type penalty

- Continuous or binary risk factors: lasso

$$g_{\text{Lasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i}|\beta_{j,i}|.$$

- Ordinal risk factors: fused lasso

$$g_{\text{fLasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i}|\beta_{j,i+1} - \beta_{j,i}| = ||\boldsymbol{D}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1$$

- Nominal risk factors: generalized fused lasso

$$g_{\text{gflasso}} = \sum_{(i,l)\in\mathcal{G}} w_{j,il}|\beta_{j,i} - \beta_{j,l}| = ||\boldsymbol{G}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1$$

SMuRF

Sparse Multi-type Regularized Feature modeling

- SMuRF unifies penalty-specific (machine learning) literature with statistical (or: actuarial) literature!
- Efficient algorithm (with proximal operators).
- Scalable to large (big) data (splits into smaller sub-problems).
- Flexible regularization
  - penalty takes type of risk factor into account
  - works for all popular penalties.

MTPL data: Poisson with multi-type penalty

- Model claim frequencies with regularized Poisson GLM

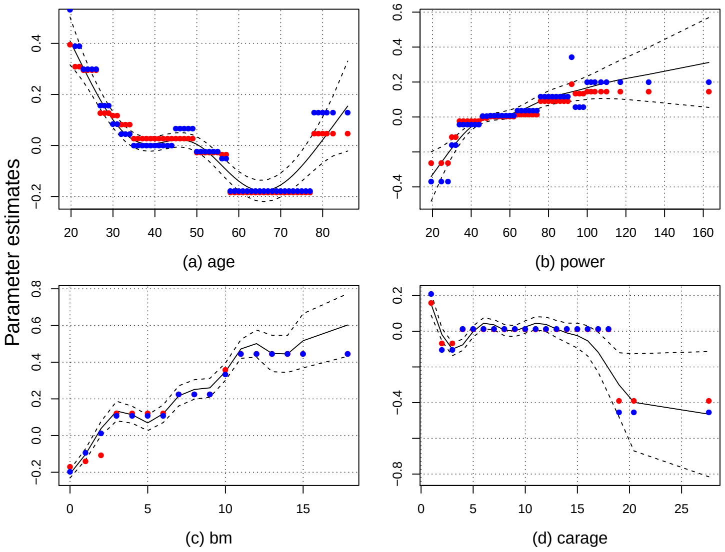
$$-\frac{1}{n}\log \mathcal{L}(\beta; \boldsymbol{X}, \boldsymbol{y}) + \lambda \left(\sum_{j \in \text{bin}} |w_j \beta_j| + \sum_{j \in \text{ord}} ||\boldsymbol{D}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1 + ||\boldsymbol{G}(\boldsymbol{w}_{\text{muni}})\boldsymbol{\beta}_{\text{muni}}||_1 \right).$$

- Incorporate multi-type penalty, with:
  - standard Lasso for binary `use`, `fleet`, `mono`, `four`, `sports`, `sex` and `fuel`
  - fused Lasso for ordinal `payfreq`, `coverage`, `ageph`, `bm`, `power`, `agec`
  - generalized fused Lasso for spatial `muni`.

MTPL data: Poisson with multi-type penalty

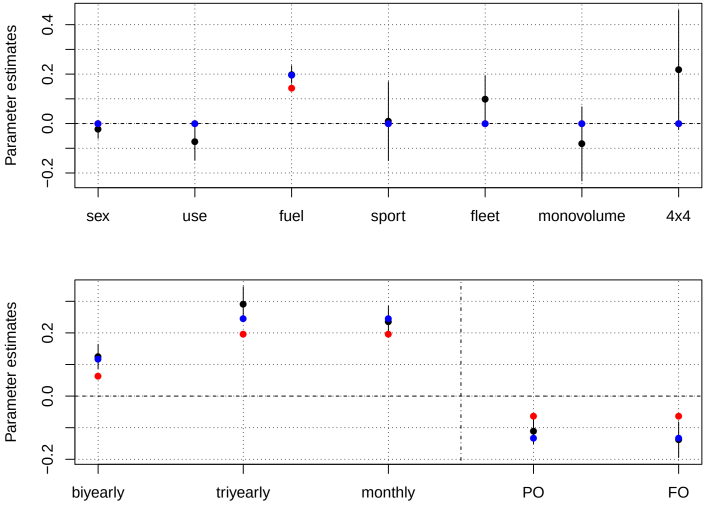
- Settings:
  - incorporate adaptive (GLM) and standardization weights for better consistency and predictive performance
  - tune λ with 10-fold stratified cross-validation where the deviance is used as error measure and the one-standard-error rule is applied
- Re-estimate the final sparse GLM with standard GLM routines (from 422 to 71 params.).

MTPL data: Poisson with multi-type penalty



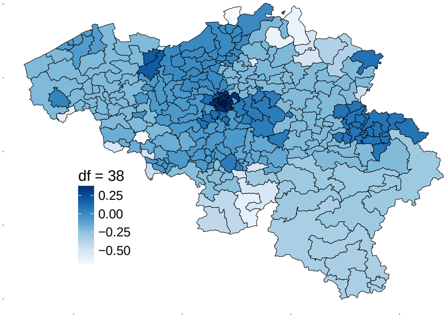
GAM fit, penalized GLM fit, GLM refit with new bins

MTPL data: Poisson with multi-type penalty

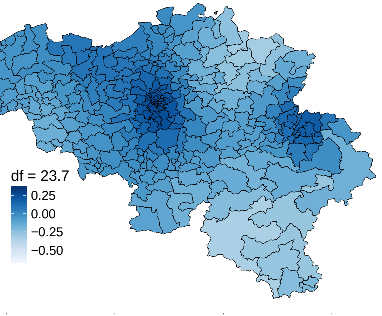


GAM fit, penalized GLM fit, GLM refit with new bins

MTPL data: Poisson with multi-type penalty



(a) SMuRF estimates



(b) GAM estimates

Wrap-up

- From multi-step (published in SAJ, `R` code upon request) to less is more.
- Flexible regularization can help predictive modeling tasks.
- `SMuRF` package, vignette and working paper forthcoming.

More information

For more information, please visit:

LRisk website, www.lrisk.be

www.feb.kuleuven.be/katrien.antonio

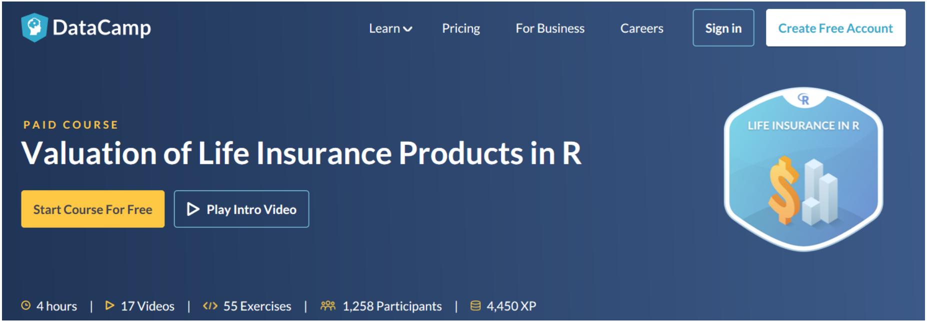
Thanks to

ageas

ARGENTA

VSC Vlaams Supercomputer Centrum

fwo 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):532.

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 11891232.