Actuaries and predictive modeling: past, present and future

Katrien Antonio

Faculty of Economics and Business
LRisk Research Center
KU Leuven & UvA
katrien.antonio@kuleuven.be

ACIS symposium 'Big Data, digitale innovatie en de gevolgen voor de verzekeringssector'

October 18, 2016






Goals of this talk

Focus on recent research in insurance rating a blend of analytic techniques.

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

by Antonio, Clijsters, Henckaerts & Verbelen.

(2) Unraveling the predictive power of telematics data in car insurance pricing.

by Verbelen, Antonio & Claeskens.

A blend of techniques/learning outcomes/buzz words from

(recent) past, present and future?

Goals of this talk

Focus on recent research in insurance rating a blend of analytic techniques.

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

by Antonio, Clijsters, Henckaerts & Verbelen.

(2) Unraveling the predictive power of telematics data in car insurance pricing.

by Verbelen, Antonio & Claeskens.

A blend of techniques/learning outcomes/buzz words from

(recent) past, present and future?

Goals of this talk

Focus on recent research in insurance rating a blend of analytic techniques.

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

by Antonio, Clijsters, Henckaerts & Verbelen.

(2) Unraveling the predictive power of telematics data in car insurance pricing.

by Verbelen, Antonio & Claeskens.

A blend of techniques/learning outcomes/buzz words from

(recent) past, present and future?

Goals of this talk

Focus on recent research in insurance rating a blend of analytic techniques.

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

by Antonio, Clijsters, Henckaerts & Verbelen.

(2) Unraveling the predictive power of telematics data in car insurance pricing.

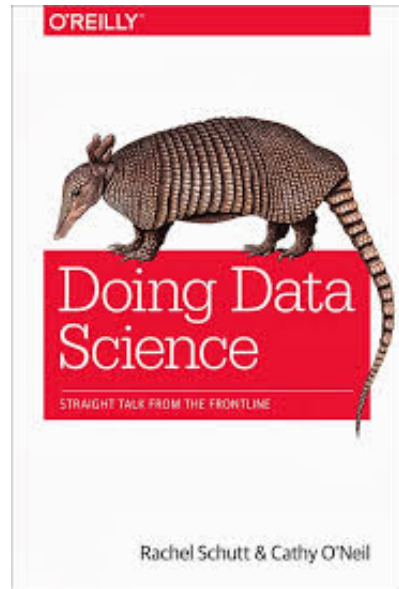
by Verbelen, Antonio & Claeskens.

A blend of techniques/learning outcomes/buzz words from

(recent) past, present and future?

Data science and predictive modeling

(1) Schutt & O'Neil (2013), *Doing data science - Straight talk from the frontline.*



What is the eyebrow-raising about big data and data science?

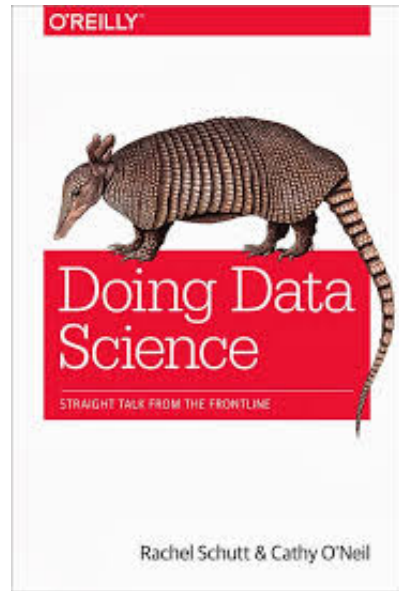
> 'The hype is crazy.'

Getting past the hype?

> 'There might be some meat in the data science sandwich';
>
> 'Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.'

Data science and predictive modeling

(1) Schutt & O'Neil (2013), *Doing data science - Straight talk from the frontline.*



What is the eyebrow-raising about big data and data science?

> 'The hype is crazy.'

Getting past the hype?

> 'There might be some meat in the data science sandwich';
>
> 'Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.'



(2) Prof. David Donoho (2015), *50 years of data science*.

Actuarial pricing models in P&C insurance

- In an actuarial pricing model, identify for each insured i:
  - N_i: number of claims/frequency during (period of) exposure e_i;
  - Y_{ij} loss/severity corresponding to each claim made ($j = 1, \ldots, N_i$);
  - Aggregate loss is $L_i := Y_{i1} + \ldots + Y_{iN_i} = \sum_{j=1}^{N_i} Y_{ij}$.
- Actuaries price risks using a priori measurable characteristics:
  - risk classification or segmentation;
  - a predictive model for **frequency** and **severity**;
  - incorporate exposure-to-risk measure.

Actuarial pricing models in P&C insurance

Actuarial pricing models in P&C insurance

- **(Past)**

  One-way and two-way analysis, minimum bias (Bailey & Simon, 1960).

Actuarial pricing models in P&C insurance

- **(Past)**

  One-way and two-way analysis, minimum bias (Bailey & Simon, 1960).

- **(Present)**

  Risk classification in competitive markets using Generalized Linear Models for frequency and severity.

Actuarial pricing models in P&C insurance

- **(Past)**

  One-way and two-way analysis, minimum bias (Bailey & Simon, 1960).

- **(Present)**

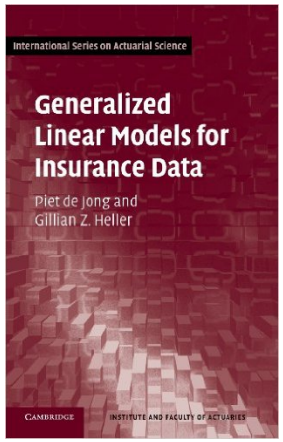
  Risk classification in competitive markets using Generalized Linear Models for frequency and severity.

- **(Future)** Challenges?

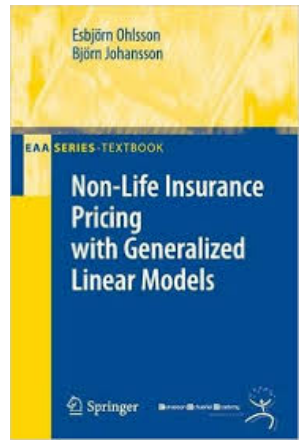
  - high dimensional variables (e.g. territory, vehicle groups)
  - (structured and unstructured) telematics data;
  - keep model explainable to clients, regulators, ICT, ...;
  - be aware of actuarial features!!

Actuarial pricing models in P&C insurance: a blend of?

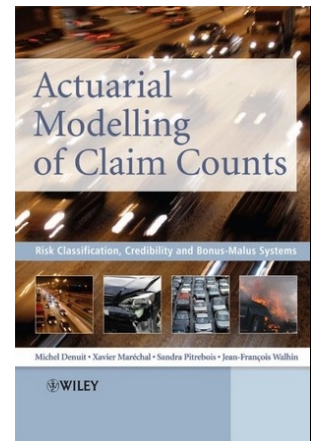
Actuarial pricing models in P&C insurance: a blend of?



de Jong & Heller

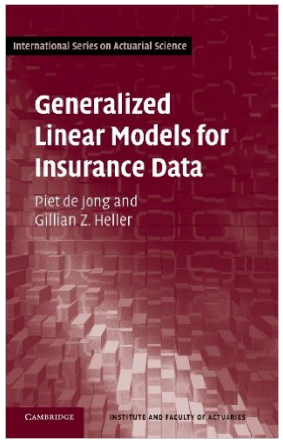


Ohlsson & Johansson

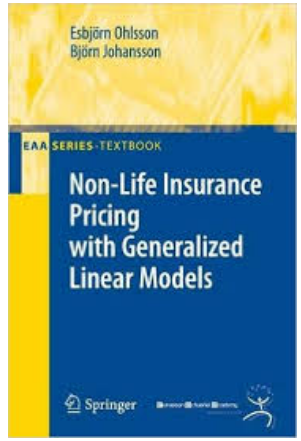


Denuit et al.

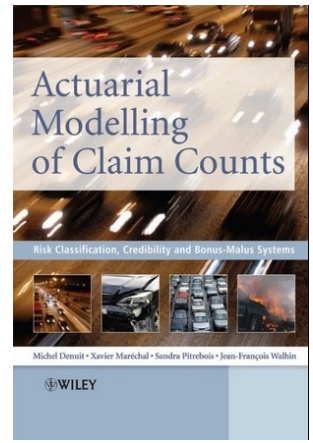
Actuarial pricing models in P&C insurance: a blend of?



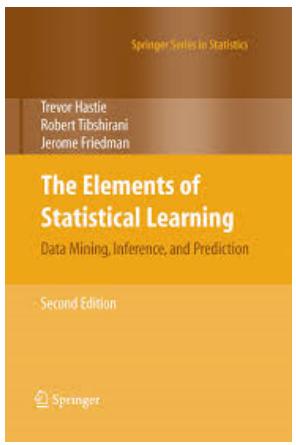
de Jong & Heller



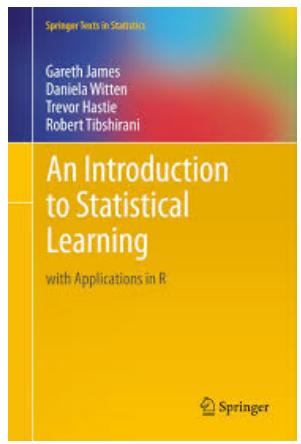
Ohlsson & Johansson



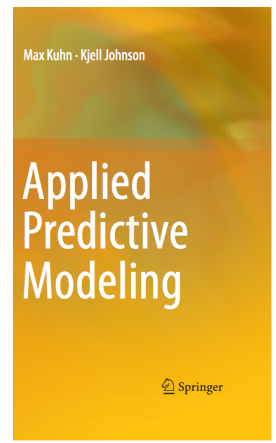
Denuit et al.



Hastie, Tibshirani & Friedman



James et al.



Kuhn & Johnson

Unraveling the predictive power of telematics data in car insurance pricing.



Roel Verbelen
KU Leuven



Katrien Antonio
KU Leuven & UvA



Gerda Claeskens
KU Leuven

Telematics insurance: the future?

- The Economist, February 23 2013, *How's my driving?*
- *"Underwriters have traditionally used crude demographic data such as age, location and sex to separate the testosterone-fuelled boy racers from their often tamer female counterparts. [...] By monitoring their customers' motoring habits, underwriters can increasingly distinguish between drivers who are safe on the road from those who merely seem safe on paper. Many think that telematics insurance will become the industry norm."*


The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Insurance and telematics

How's my driving?

Gizmos that track driving habits are changing the face of car insurance

Feb 23rd 2013 | From the print edition | Timekeeper

New rating variables due to telematics technology

Telematics data collected in each trip: driving habits

- the distance driven;
- the time of day;
- how long you have been driving;
- the location;



New rating variables due to telematics technology

Telematics data collected in each trip: driving habits and driving style

- the distance driven;
- the time of day;
- how long you have been driving;
- the location;



- the speed/speeding;
- harsh or smooth breaking;
- aggressive acceleration or deceleration;
- your cornering and parking skills.

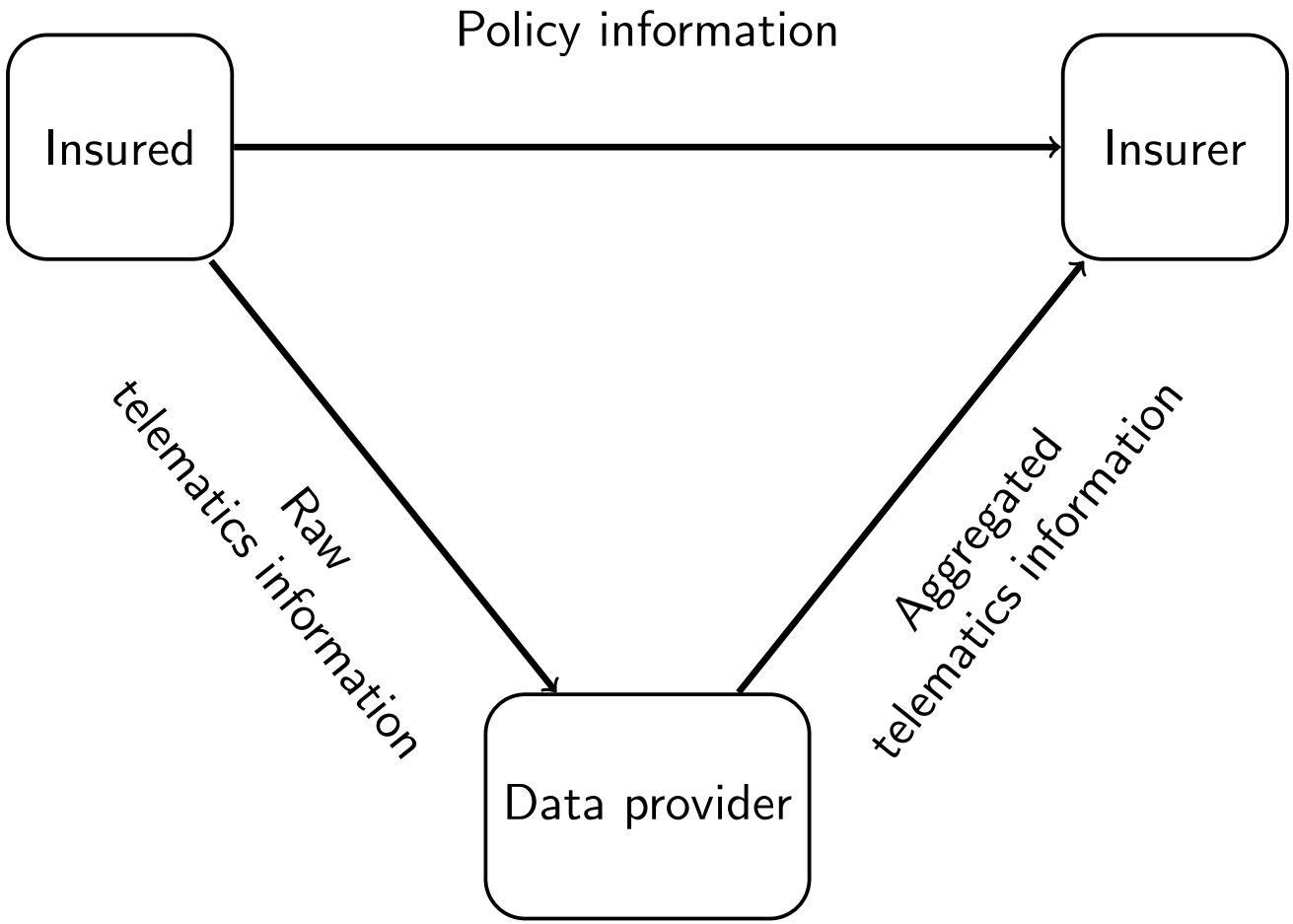
Possibly combined with:

- road maps;
- weather information;
- traffic information.

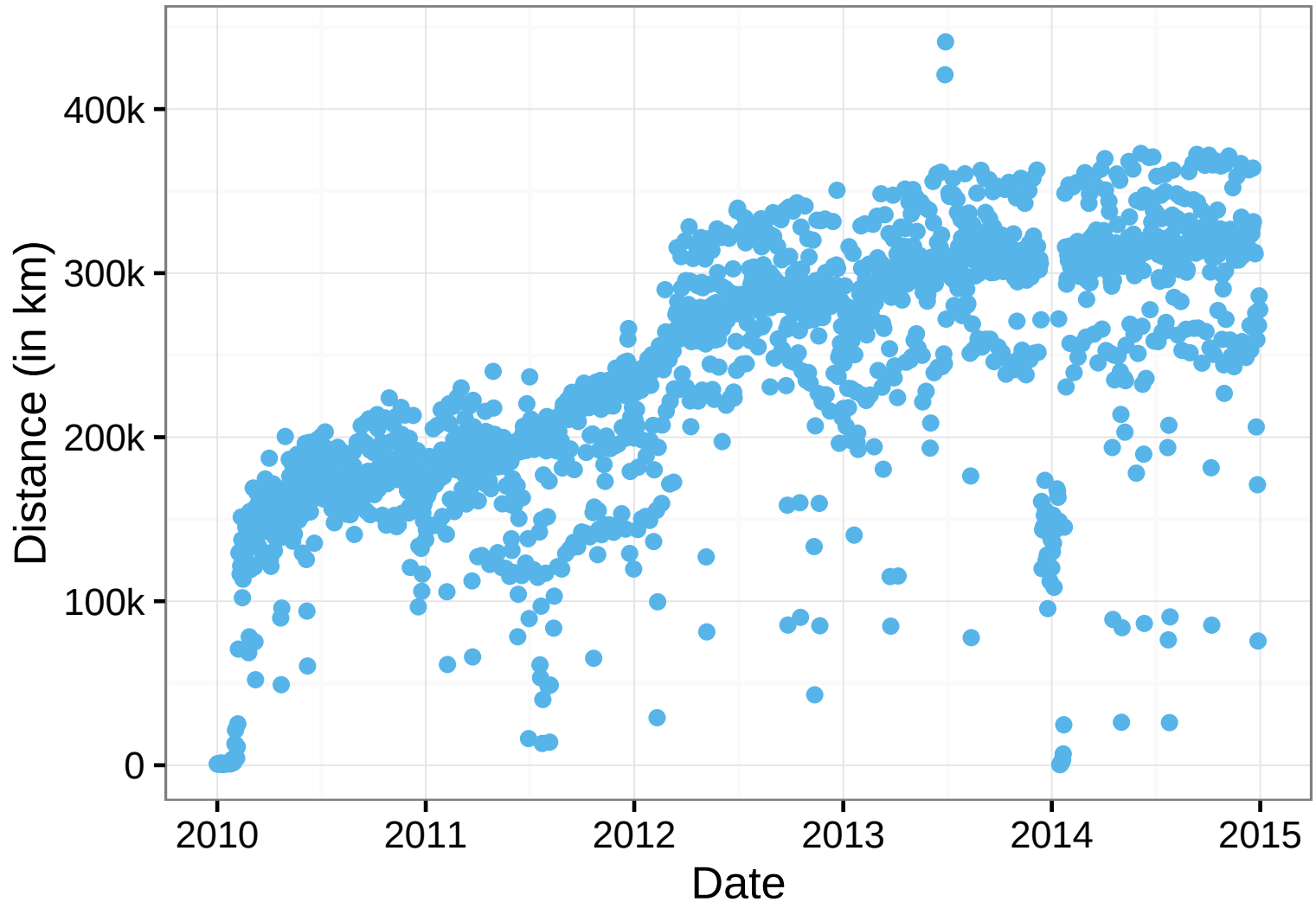
Unique telematics data set from a Belgian insurer

- Telematics data collected in between 2010 and 2014.
- Belgian MTPL product with telematics black box targeted to young drivers.
- Daily CSV-files with trip info, aggregated on daily basis:
  - number of trips;
  - meters traveled (in total) and
    - divided by time slot: 6u-9u30, 9u30-16u, 16u-19u, 19u-22u, 22u-6u;
    - divided by road type: motorways, urban area, abroad, any other type.

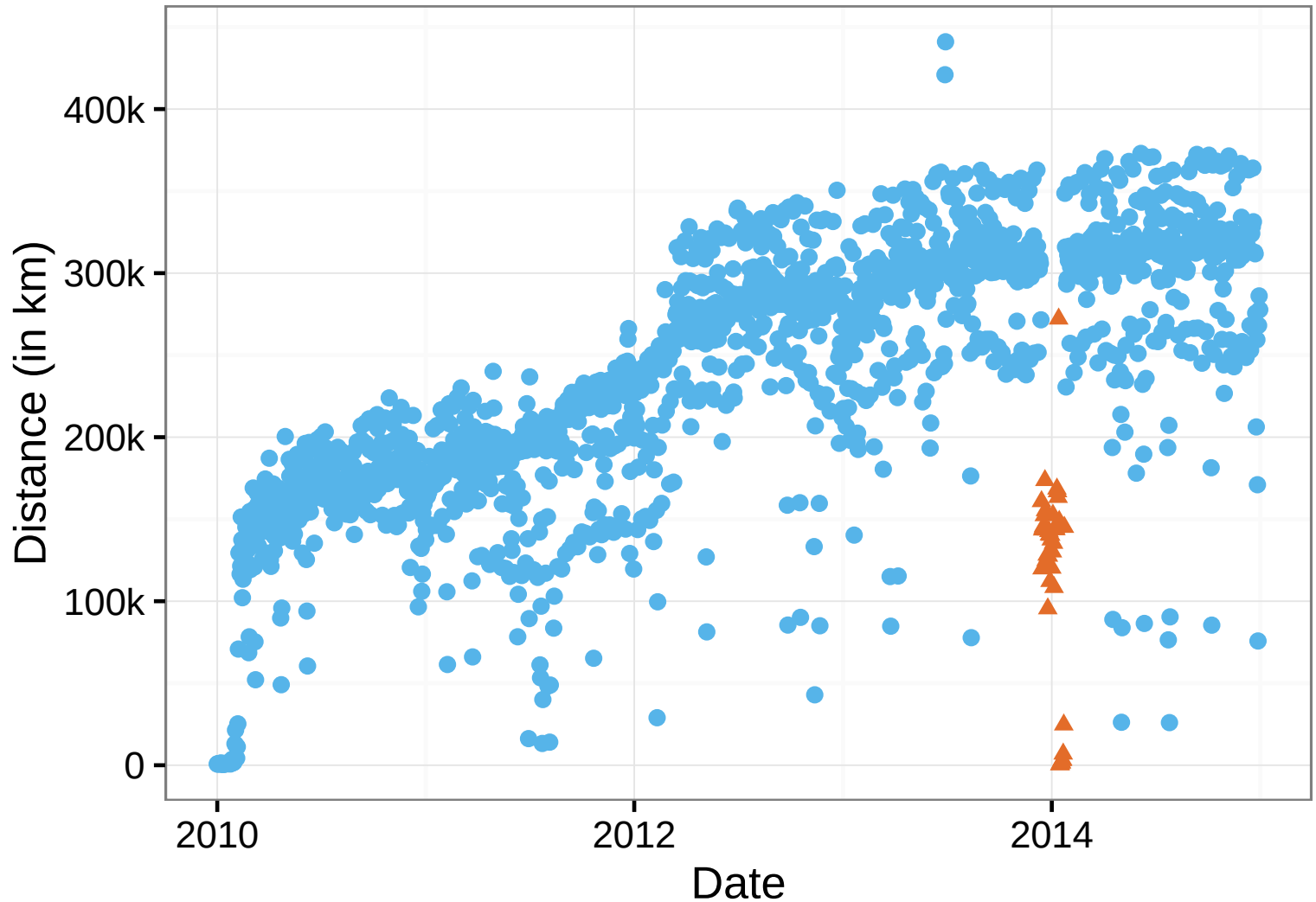
Unique telematics data set from a Belgian insurer



Unique telematics data set from a Belgian insurer



Unique telematics data set from a Belgian insurer

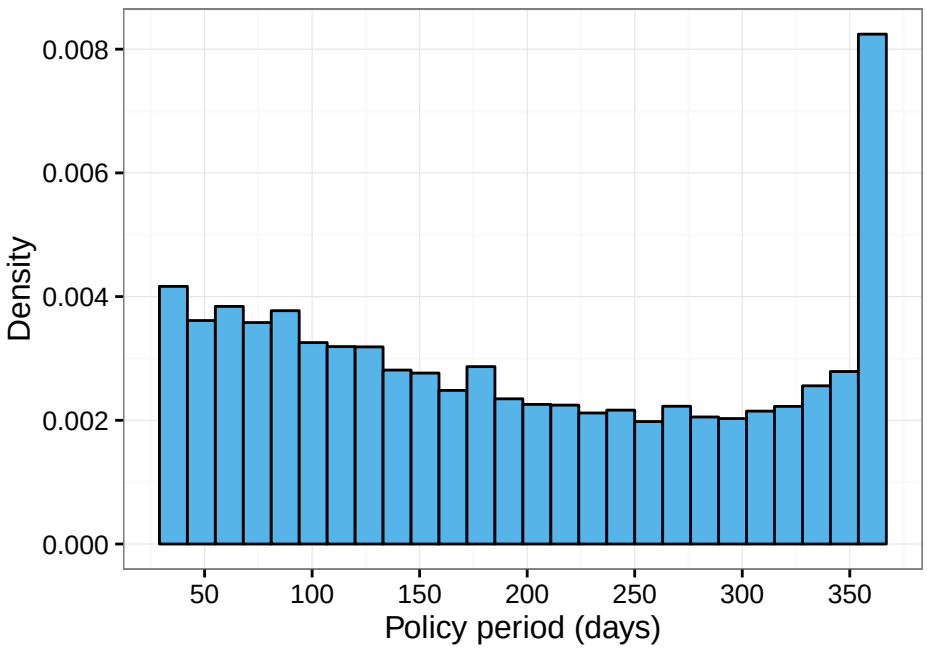


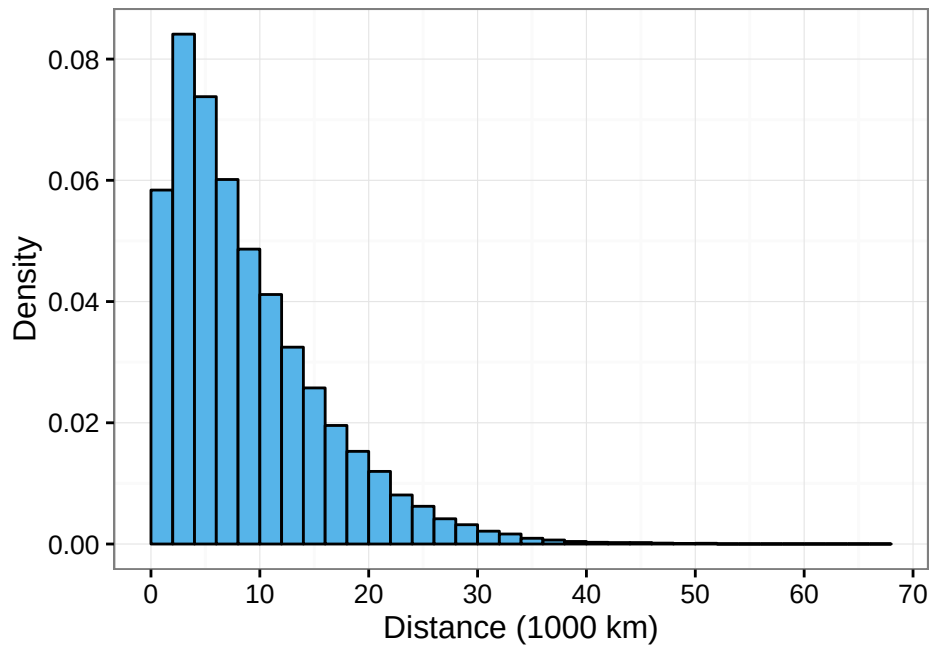
Description of the data

The resulting data set has 33 259 observations:

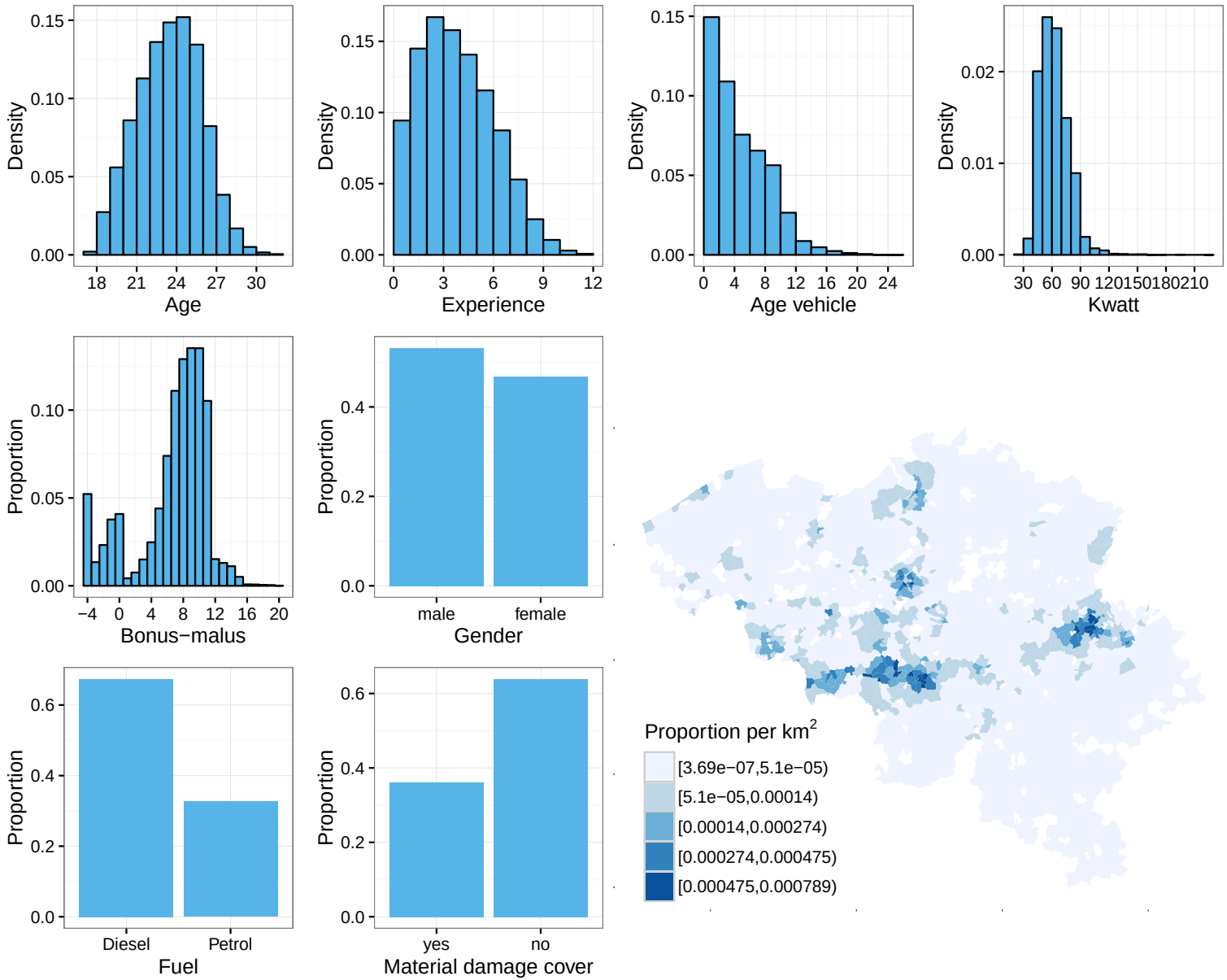
- 10 406 unique policyholders;
- 17 681 years of insured periods;
- 0.0838 claims per insured year;
- 1481 MTPL claims at fault;
- 297 million kilometers driven;
- 0.0499 claims per 10 000 km.

What is the best measure of exposure to risk?

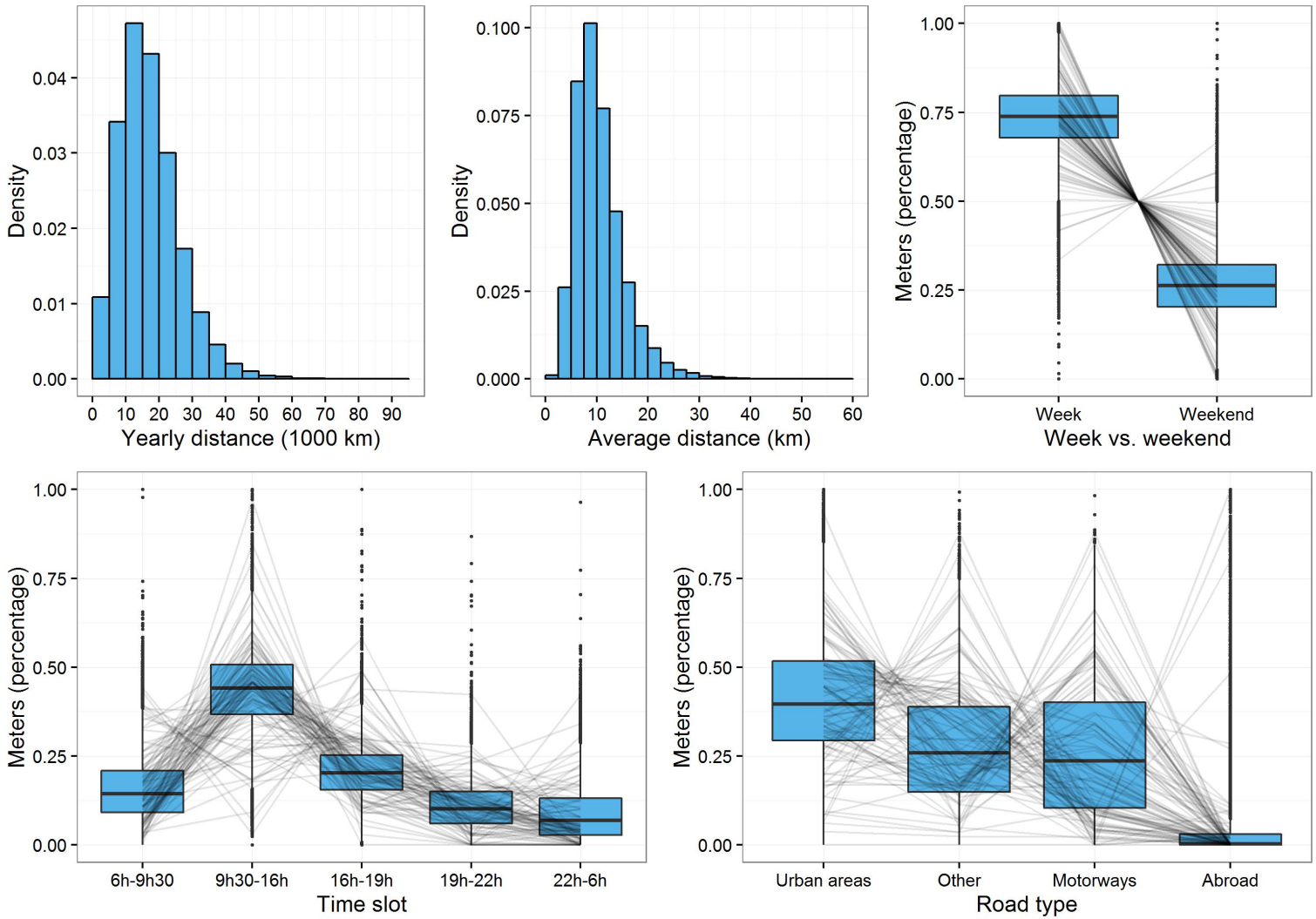




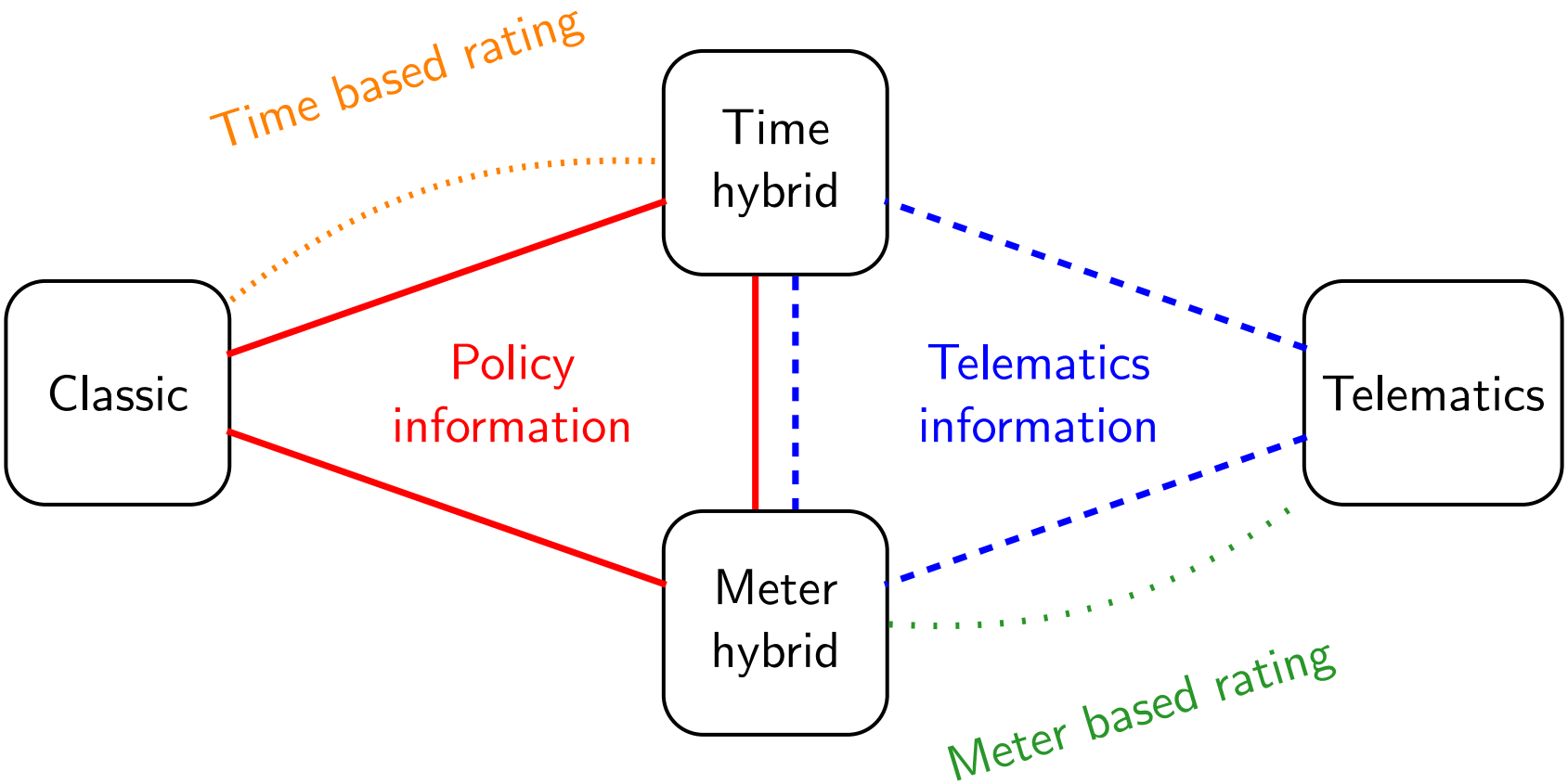
Policy information



Telematics information



Predictor sets



Generalized additive models

We use GAMs (Wood, 2006):

$$N_{it} \sim \text{POI}(\mu_{it} = \exp(\eta_{it}))$$

$$\eta_{it} = \text{offset} + \eta_{it}^{\textbf{cat}} + \eta_{it}^{\textbf{cont}} + \eta_{it}^{\textbf{spatial}} + \eta_{it}^{\text{re}} + \eta_{it}^{\textbf{comp}}$$

$$\eta_{it}^{\textbf{cat}} + \eta_{it}^{\textbf{cont}} + \eta_{it}^{\textbf{spatial}} = \boldsymbol{z}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_{jit}) + f_{spatial}(lat_{it}, long_{it}),$$

We combine:

categorical + **continuous** + **spatial** + **compositional (new!!)**

risk factors.

Model selection and assessment

- Exhaustive search with AIC as a global goodness-of-fit measure.

$$\text{AIC} = -2 \cdot \log \mathcal{L} + 2 \cdot EDF$$

where EDF is the effective degrees of freedom.
retained.

- Predictive performance is assessed using proper scoring rules for count data (Czado et al., 2009) with 10-fold cross validation

$$S = \frac{1}{\sum_{i=1}^{I} T_i} \sum_{i=1}^{I} \sum_{t=1}^{T_i} s(\widehat{P}_{it}^{-\kappa_{it}}, n_{it}) \,,$$

where $\widehat{P}_{it}^{-\kappa_{it}}$ the predictive count distribution for observation n_{it} estimated with the κ_{it}th part of the data removed.

Results: model selection

	Predictor	Classic	Time hybrid	Meter hybrid	Telematics
Policy	Time	×	×		
	Age				
	Experience	×	×	×	
	Sex	×			
	Material	×	×	×	
	Postal code	×	×	×	
	Bonus-malus	×	×	×	
	Age vehicle	×	×	×	
	Kwatt		×	×	
	Fuel	×	×	×	
Telematics	Distance			×	×
	Yearly distance		×		
	Average distance		×	×	
	Road type 1111		×	×	×
	Road type 0111		×	×	×
	Time slot		×	×	×
	Week/weekend		×	×	×

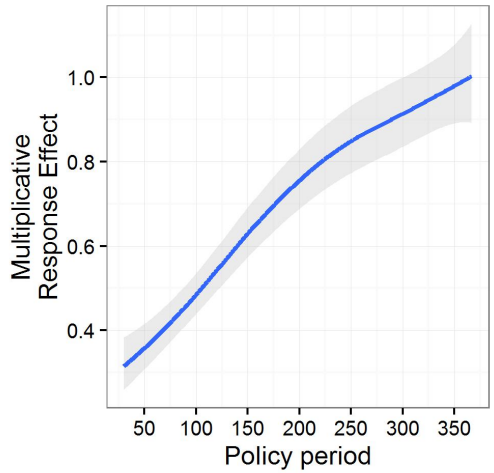
Results: model assessment

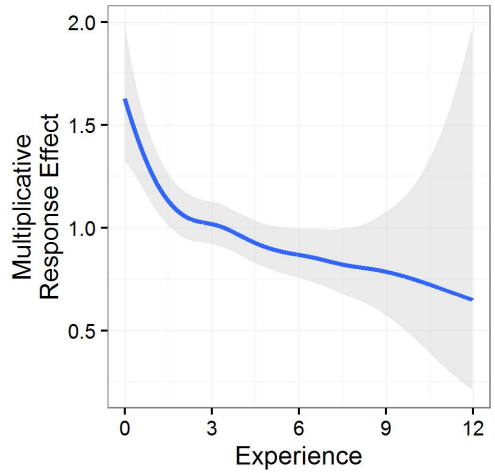
Predictor set	EDF	AIC		logS		QS		SphS	
		value	rank	value	rank	value	rank	value	rank
Classic	32.15	11 896	4	0.1790	4	−0.918 58	4	−0.958 22	4
Time hybrid	39.66	11 727	1	0.1764	1	−0.919 10	1	−0.958 37	1
Meter hybrid	41.47	11 736	2	0.1766	2	−0.919 08	2	−0.958 36	2
Telematics	18.05	11 890	3	0.1787	3	−0.918 60	3	−0.958 22	3

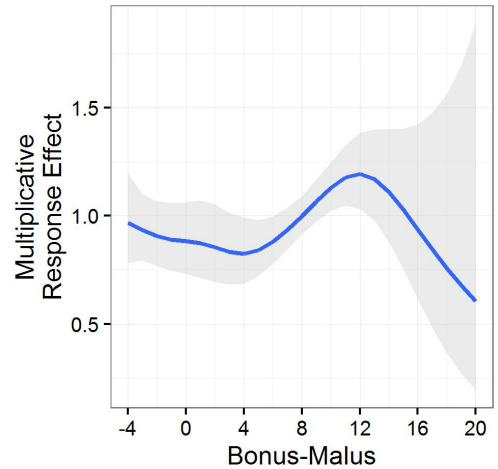
- Significant impact of the use of telematics data;
- Time hybrid is the best model according to AIC and all proper scoring rules;
- Using only telematics predictors is even better than the use of traditional rating variables.

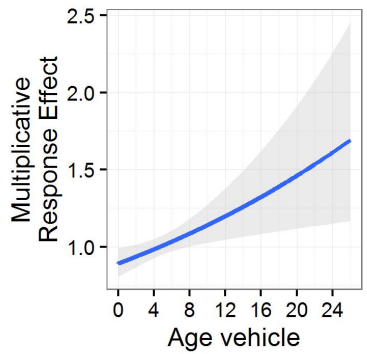
Time hybrid - Policy information

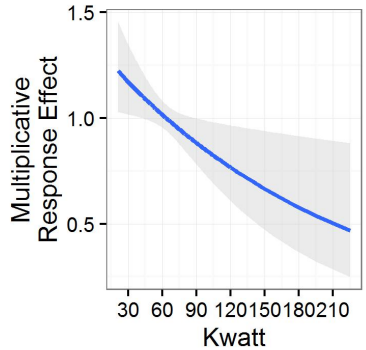
	Predictor
Policy	Time
	Age
	Experience
	Sex
	Material
	Postal code
	Bonus-malus
	Age vehicle
	Kwatt
	Fuel

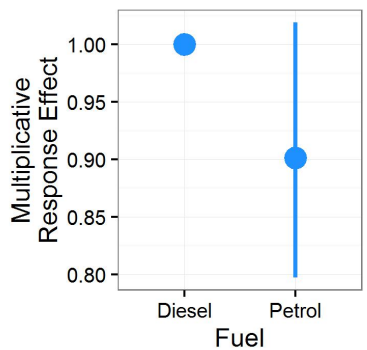


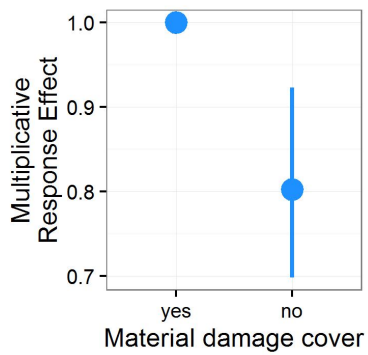


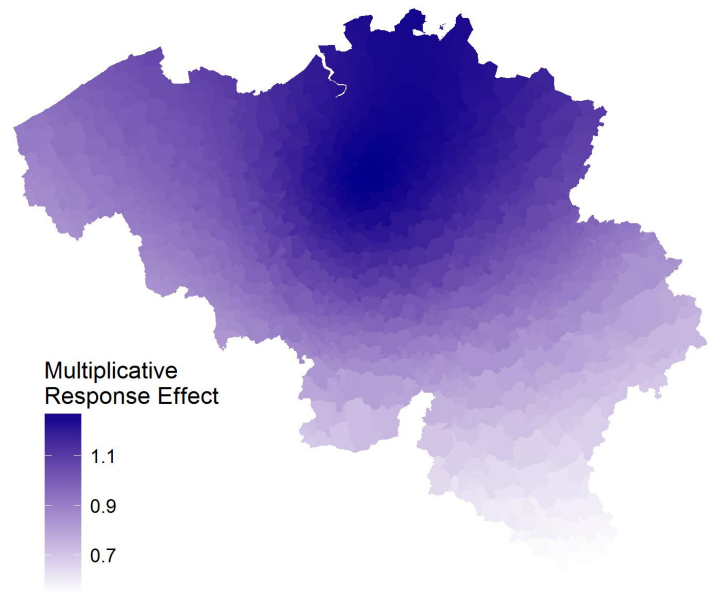






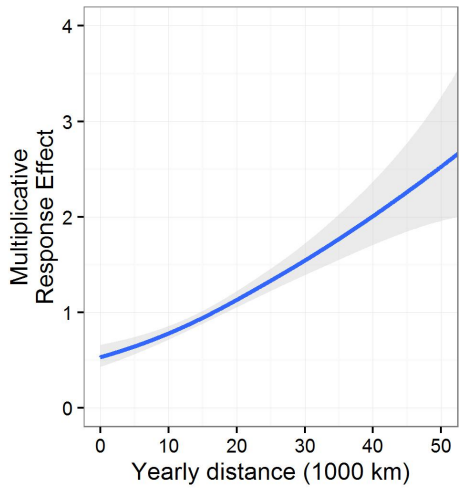


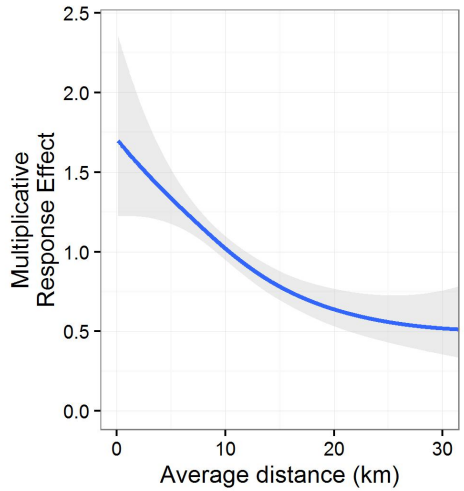


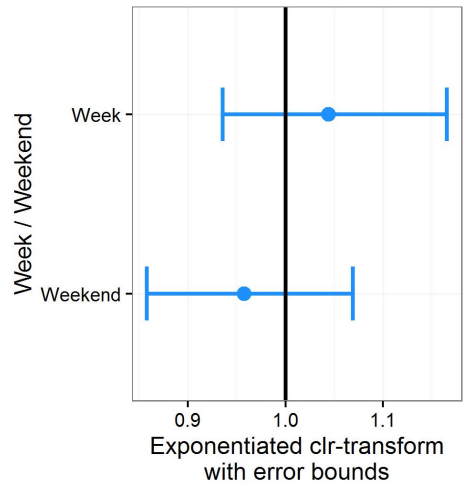


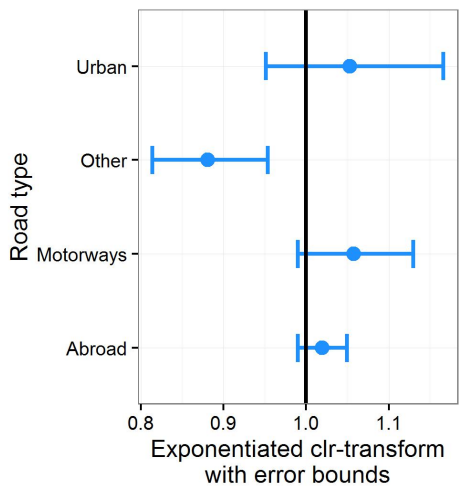
Time hybrid - Telematics information

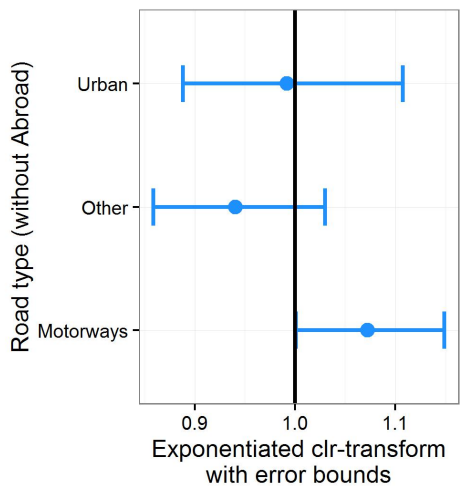
	Predictor
Telematics	Distance
	Yearly distance
	Average distance
	Road type 1111
	Road type 0111
	Time slot
	Week/weekend

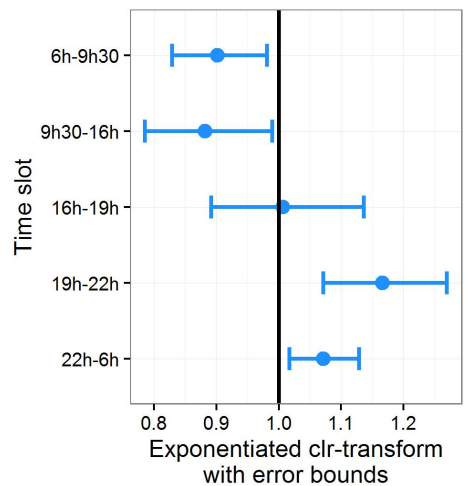












Results: discussion

- Telematics information improves predictive power.
  - Time hybrid model incorporating telematics through additional risk factors is optimal.
  - Classic approach performs worse.
  - Gender plays no role anymore in models incorporating telematics information (cfr. Gender Directive).
  - Spatial heterogeneity decreases.
  - Experience is preferred above age of the driver.
  - Compositional driving habits have significant impact on riskiness.
- Similar results using negative binomial regression and using exposure as offset.

Outlook

I encourage the blending idea ...

- of techniques (from machine learning, statistical modeling, actuarial science);
- of disciplines (from computer science, statistics, actuarial science, but also law);
- of people from practice and academia;

... to tackle the challenges imposed by structured and unstructured data in order to create insurance analytics, products and risk management of the future.