A statistical modeling approach for car insurance pricing with telematics data

Roel Verbelen

joint work with Katrien Antonio and Gerda Claeskens

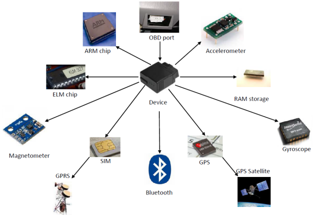
Faculty of Economics and Business
KU Leuven, Belgium
roel.verbelen@kuleuven.be
feb.kuleuven.be/roel.verbelen

R in Insurance 2016
Cass Business School, London

July 11, 2016




What is telematics insurance?



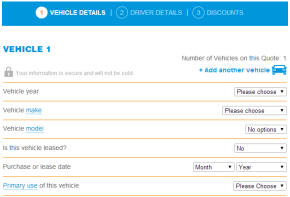
Synonyms: usage-based insurance (UBI)
pay-as-you-drive (PAYD)
pay-how-you-drive (PHYD)

- telematics is the integrated use of telecommunications and informatics;
- black-box device is installed in the vehicle;
- real driving behavior is monitored;
- allows for better risk assessment and personalized premiums based on individual driving data;
- drives down the cost for low-mileage clients and good drivers;
- may fundamentally change the car insurance industry.

Traditional rating variables

Self-reported information, including:

- age;
- age driver's license;
- vehicle year, make and model;
- catalog value;
- engine power;
- use of the vehicle;
- type of coverage;
- postal code;
- claims history.



⇒ only proxy variables for the accident risk;

⇒ does not reflect the present pattern of driving behavior;

⇒ a lot of heterogeneity between drivers remains.

Additional rating variables due to telematics technology

Telematics data collected in each trip:

- the distance driven;
- the time of day;
- how long you have been driving;
- the location;
- the speed;
- harsh or smooth braking;
- aggressive acceleration or deceleration;
- your cornering and parking skills.

Possibly combined with:

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

Research goals

Goals of our contribution (see Verbelen, Antonio & Claeskens):

(1) set-up data merge, cleaning, quality checks to combine traditional and telematics rating variables; (all coded in open source `R: data.table`)

(2) develop the statistical methodology for pricing car insurance policies based on the high dimensional telematics data collected while driving;

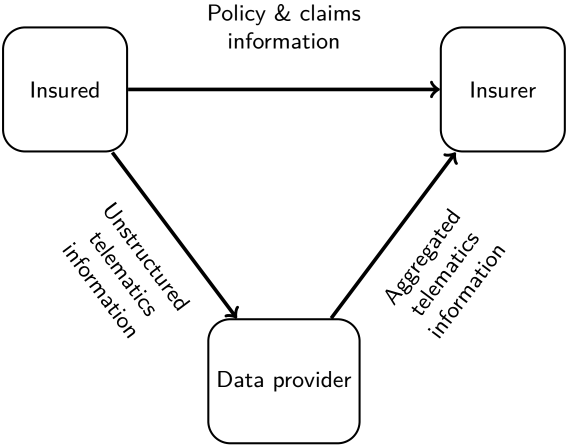
(3) combine traditional rating variables and telematics information in the claim frequency model;

- → compare the performance of different sets of predictor variables (e.g. traditional vs purely telematics);
- → discover the relevance and impact of adding telematics insights;
- → contrast the use of time and distance as exposure to risk.

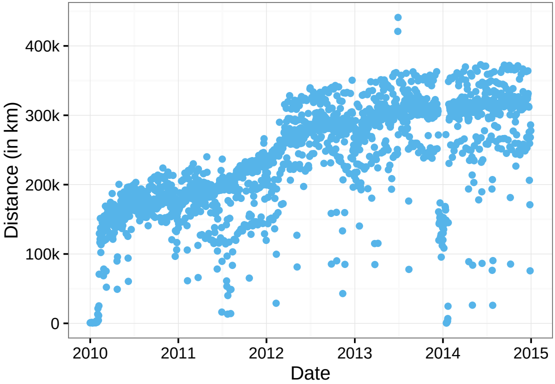
Telematics data set from a Belgian insurer

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

Flow of information



Data quality



Combined with policy information and claim counts

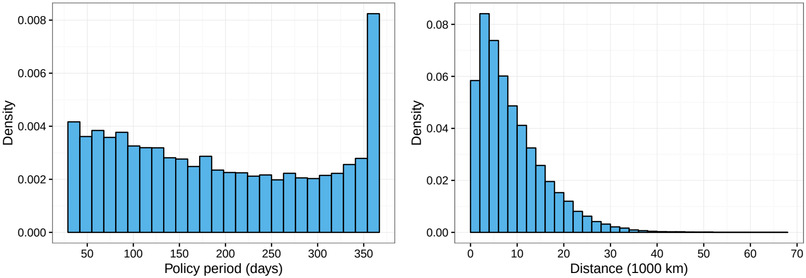
- Merged with traditional policy(holder) information by policy number and policy period:
  - policy: policy period, material damage cover;
  - policyholder: age, experience, sex, bonus-malus, postal code;
  - car: age vehicle, kwatt, fuel.
- Policy period is restricted to the time period in which telematics data is being captured.
- Technical failure at the turn of the year 2014 taken into account in these restrictions.
- Minimum policy duration of 30 days to be kept in the analysis;
- Linked with claim counts of MTPL claims at fault falling in between the restricted policy durations.

Description of the data

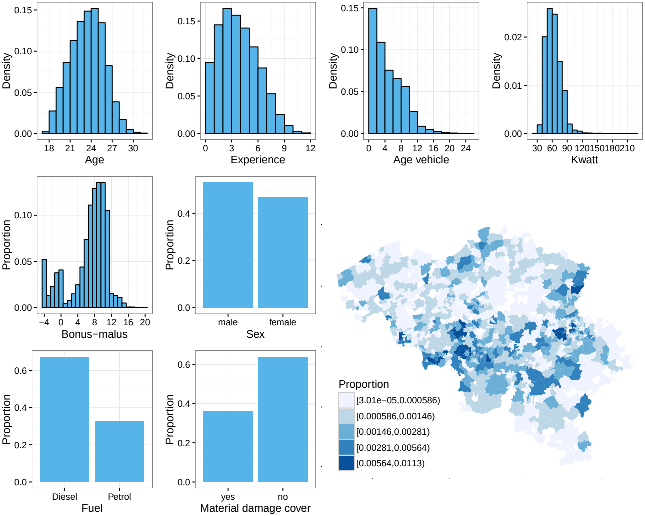
The resulting data set has 33 259 observations:

- 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?

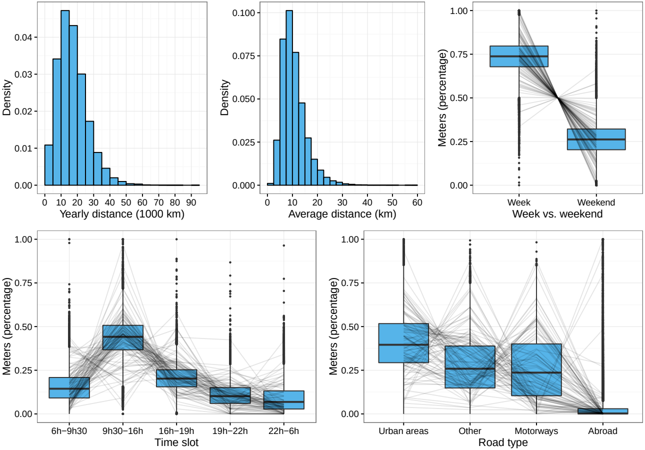


Policy information

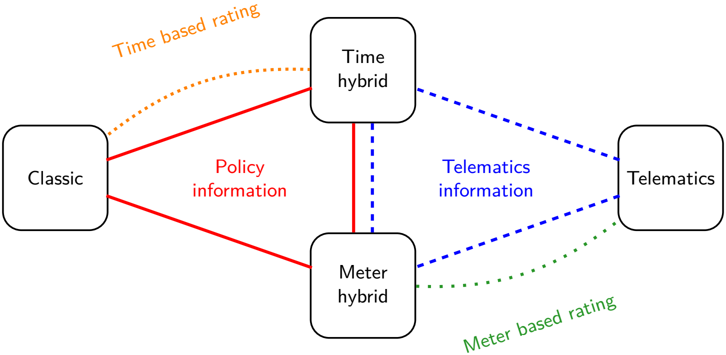


R: ggplot2, rgdal

Telematics information



Predictor sets



Claim count modeling

We model the frequencies of claims by constructing Poisson regression models (Denuit et al., 2007).

- N_{it}: number of claims for policyholder $i = 1, \ldots, I$ in policy period $t = 1, \ldots, T_i$.
- $N_{it} \sim \text{Poisson}(\mu_{it})$ with

$$P(N_{it} = n_{it}) = \frac{\exp(-\mu_{it})\mu_{it}^{n_{it}}}{n_{it}!}.$$

- log linear relationship between the mean and the predictor variables

$$E(N_{it}) = \mu_{it} = \exp(\eta_{it}).$$

with η_{it} is a predictor function of the available explanatory variables.

Generalized additive models

We use GAMs (Wood, 2006, `R: mgcv`) to define nonparametric relationships between the response and predictors

$$
\begin{aligned}
\eta_{it} &= \beta_0 + \text{offset} + \eta_{it}^{\text{cat}} + \eta_{it}^{\text{cont}} + \eta_{it}^{\text{spatial}} + \eta_{it}^{\text{re}} + \eta_{it}^{\text{comp}} \\
&= \beta_0 + \text{offset} + \boldsymbol{Z}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_{jit}) + f_s(\text{lat}_{it}, \text{long}_{it}) + \eta_{it}^{\text{re}} + \eta_{it}^{\text{comp}} ,
\end{aligned}
$$

- parametric model terms for all categorical predictors;
- penalized cubic regression spline components f_j for all continuous variables;
- spatial term f_s as a smooth bivariate function of the coordinates of the postal code;
- random effect term and compositional predictors;
- estimation using penalized iteratively reweighted least squares (P-IRLS);
- smoothing parameters selected using AIC.

Compositional data

- Divisions of the total distance driven in the different categories: road type (4), time slot (5), week/weekend (2)
  - $\rightarrow$ highly correlated with and sums up to total distance driven;
  - $\rightarrow$ perfect multicollinearity problem;
  - $\rightarrow$ standard regression interpretation does not hold.
- We divide the divisions by the total distance since they only contribute relative information;
  - $\rightarrow$ positive components that sum to one;
  - $\rightarrow$ compositional data (`R: compositions`);
  - $\rightarrow$ classical statistical techniques incoherent on compositions;
  - $\rightarrow$ special vector space structure has to be taken into account.

Compositional predictors

From a methodological point of view this is the novelty of our work.

- We show how to include the compositional data as predictors in the regression,
- . . . and how to interpret their effect on the average claim frequency;
- We present a solution for structural zeros as predictors;
- As such, we extend both the actuarial pricing literature as well as the statistical literature on regression with compositional data.

Model selection and assessment

- AIC is used as a global goodness-of-fit measure.

$$\mathrm{AIC} = -2 \cdot \log \mathcal{L} + 2 \cdot \mathrm{tr}(\boldsymbol{H})$$

  where $\boldsymbol{H}$ denotes the hat or smoothing matrix.

- For each predictor set, variables are selected using an exhaustive search over all the possible combinations. The best model according to AIC is retained.

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

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

  where s is a scoring rule and $\widehat{P}_{it}^{-\kappa_{it}}$ is the predictive distribution of the observed claim count 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		×	×	×

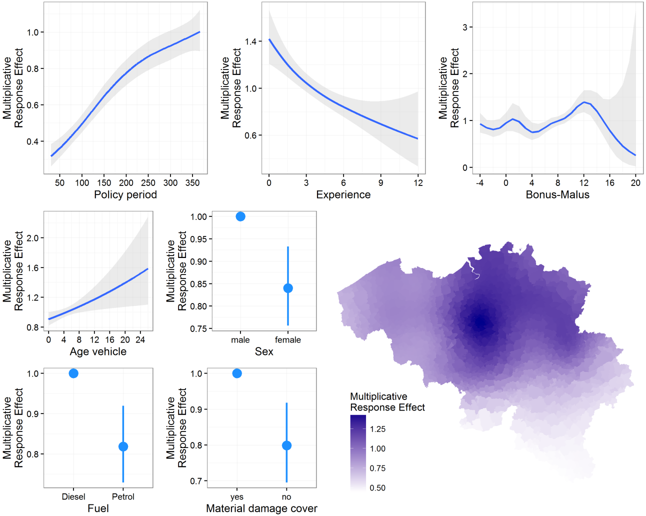
Results: model assessment

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.

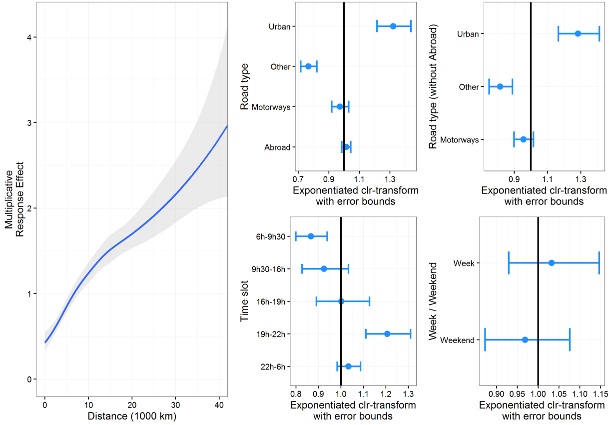
Classic

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



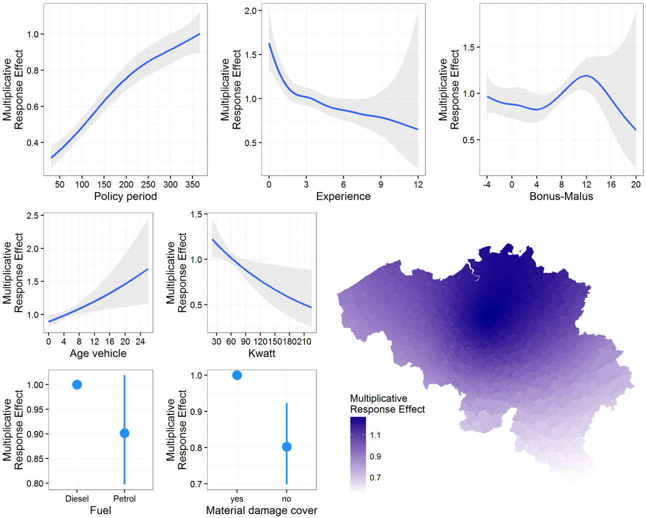
Telematics

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



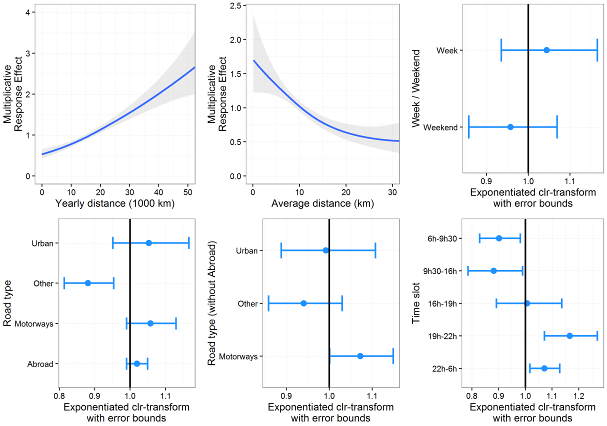
Time hybrid - Policy information

	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



Conclusions

- Statistical methodology developed to incorporate new data structures provided through telematics in models for claim frequencies.
- Telematics information improves predictive power.
  - Gender plays no role anymore in models incorporating telematics information (cfr. Gender Directive).
  - Spatial heterogeneity decreases.
  - Time hybrid model incorporating telematics through additional risk factors is optimal.
  - Classic approach performed worse.
- Similar results using negative binomial regression and using exposure as offset.

References

Verbelen, R., Antonio, K., and Claeskens, G. (2016)
A statistical modeling approach for car insurance pricing with telematics data.
Working paper.

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

Hron, K., Filzmoser, P., and Thompson, K. (2012)
Linear regression with compositional explanatory variables.
Journal of Applied Statistics, 39(5):1115-1128.

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013)
Analyzing compositional data with R.
Springer.