Unraveling the predictive power of telematics data in car insurance pricing

Roel Verbelen[1, 3, 4], Katrien Antonio[1, 2, 3, 4], and Gerda Claeskens[1, 3]

[1]Faculty of Economics and Business, KU Leuven, Belgium.
[2]Faculty of Economics and Business, University of Amsterdam, The Netherlands.
[3]LStat, Leuven Statistics Research Centre, KU Leuven, Belgium.
[4]LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.

November 7, 2017

Abstract

A data set from a Belgian telematics product aimed at young drivers is used to identify how car insurance premiums can be designed based on the telematics data collected by a black box installed in the vehicle. In traditional pricing models for car insurance, the premium depends on self-reported rating variables (e.g. age, postal code) which capture characteristics of the policy(holder) and the insured vehicle and are often only indirectly related to the accident risk. Using telematics technology enables tailor-made car insurance pricing based on the driving behavior of the policyholder. We develop a statistical modeling approach using generalized additive models and compositional predictors to quantify and interpret the effect of telematics variables on the expected claim frequency. We find that such variables increase the predictive power and render the use of gender as a rating variable redundant.

Keywords: Pay-as-you-drive insurance; Usage-based insurance; Risk classification; Generalized additive models; Compositional predictors; Structural zeros.

1 Introduction

For a unique Belgian portfolio of young drivers in the period between 2010 and 2014, telematics data on how many kilometers are driven, during which time slots and on which type of roads were collected using black box devices installed in the insureds' cars. Our aim is to incorporate this information in statistical rating models, where we focus on predicting the number of claims, in order to adequately set premium levels based on individual policyholder's driving habits.

Determining a fair and correct price for an insurance product (also called *ratemaking*, *pricing* or *tarification*) is crucial for both insureds and insurance companies. Car insurance is traditionally priced based on self-reported information from the insured, most importantly: age, license age, postal code, engine power, use of the vehicle, and claims history. However, these observable risk factors are only proxy variables, not reflecting current driving habits and driving style. Telematics technology – the integrated use of telecommunication and informatics – may fundamentally change the car insurance industry. The use of this technology in insured vehicles enables to transmit and receive information that allows an insurance company to better quantify the accident risk of drivers and adjust the premiums accordingly through usage-based insurance (UBI). 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.[1] Young drivers and drivers in other high risk groups, who are typically facing hefty insurance premiums, can be judged based on how they really drive. Regulation also plays a role as the use of indirect indicators of risk is being questioned by the European Court of Justice. In 2012, a European Union (EU) ruling came into force, banning price differentiation based on gender.[2] Through telematics, women may be able to confirm that they really are safer drivers.

The use of telematics risk factors potentially enables an improved method for determining the cost of insurance. Due to a more refined customer segmentation and greater monitoring of the driving behavior, UBI addresses the problems of adverse selection and moral hazard that arise from the information asymmetry between the insurer and the policyholders (Filipova-Neumann and Welzel, 2010). Closer aligning insurance policies to the actual risks increases actuarial fairness and reduces cross-subsidization compared to grouping the drivers into too general actuarial classes (Desyllas and Sako, 2013). Telematics insurance gives a high incentive to change the current driving pattern and stimulates more responsible driving (Parry, 2005; Litman, 2015; Tselentis et al., 2016). Users' feedback on driving behavior and gamification of UBI can further enhance the customer experience by making it more interactive, gratifying and even exciting (Toledo et al., 2008). Less and safer driving is encouraged, leading to improved road safety and reduced vehicle travel with less congestion, pollution, fuel consumption, road cost, and crashes (Greenberg, 2009).

Usage-based insurance (Tselentis et al., 2016) includes *pay-as-you-drive* (PAYD) and *pay-how-you-drive* (PHYD) schemes. PAYD focuses on the driving habits, e.g. the driven distance, the time of day, how long the insured has been driving. PHYD also considers the driving style, e.g. the speed, harsh or smooth braking, aggressive acceleration or deceleration, cornering and parking skills.

Telematics insurance started as a niche market when the technology first surfaced more than 10 years ago. Early adopters of UBI were seen in the United States, Italy and the United Kingdom. On 28 April 2015 the European Parliament voted in favor of eCall regulation which forces all new cars in the EU from April 2018 onwards to be equipped with a telematics device that will automatically dial 112 in the event of an accident, providing precise location and impact data.[3]

This potentially high dimensional telematics data, collected on the fly, forces pricing actuaries to change their current practice, both from a business as well as a statistical point of view. New statistical models have to be developed to adequately set premiums based on an individual policyholder's driving habits and style and the current literature on insurance rating does not adequately address this question. In this paper, we take a first step in this direction. We use a Belgian telematics insurance data set with in total over 297 million kilometers driven. Based on how many kilometers the insured drives, on which kind of roads and during which moments in the day, we quantify the impact of individual driving habits on expected claim frequencies. Combined with a similar predictive model for claim severities, which is outside of the scope in this paper, this allows for tailor-made car insurance pricing. We first discuss how a car insurance policy is traditionally priced and relate this to the literature investigating the impact of vehicle usage on the accident risk in Section 2. The data set is described in Section 3, along with the necessary preliminary data processing steps to combine the telematics information with the policy and claims information. By constructing predictive models for the claim frequency, we compare the performance of different sets of predictor variables (e.g. traditional vs. purely telematics) and unravel the relevance and impact of adding telematics insights. In particular, we contrast the use of time and distance as exposure-to-risk measures. The novel methodological contribution of this paper, see Section

[1] How's my driving? (2013, February 23) *The Economist.* *http://econ.st/Yd5x3C*

[2] http://europa.eu/rapid/press-release_IP-11-1581_en.htm

[3] Regulation (EU) 2015/758 of the European Parliament and of the Council of 29 April 2015 concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC.

4, incorporates the divisions of the driven distance by road type and time slots as compositional predictors in the regression framework of generalized additive models and constructs a new way to interpret and visualize their effect on the average claim frequency. We develop both a conditioning and a projection approach to handle structural zeros in one or more components of a compositional predictor. We present the results in Section 5 while Section 6 concludes.

2 Statistical background and related modeling literature

Insurance pricing is the calculation of a fair premium, given the policy(holder) characteristics, as well as information on claims reported in the past (if available). Pricing relies on regression techniques and requires a data set with policy(holder) information and corresponding claim frequencies and severities, where severity is the ultimate total impact of a claim. The claim frequency and severity components are typically modeled separately using regression techniques (Frees, 2014). The current state-of-the-art (see Denuit et al., 2007; de Jong and Heller, 2008, for an overview) uses generalized linear models (GLMs; McCullagh and Nelder, 1989), with typically a Poisson GLM for the claim counts and a gamma GLM for the claim severities. In car insurance, the duration of the policy period during which coverage is provided, is referred to as the *exposure-to-risk*. The expected number of claims is in practice modeled directly proportional to the exposure, to make the premiums proportional to the length of coverage. From a theoretical point of view, this can be motivated by the probabilistic framework of Poisson processes (Denuit et al., 2007). It is however suggested (see e.g. Butler, 1993) that every kilometer traveled by a vehicle transfers risk to its insurer and hence the number of driven kilometers *(car-kilometer)* should be adopted as the exposure unit instead of the policy duration *(car-year)*. Statistical studies show how claim frequencies significantly increase with kilometers (Bordoff and Noel, 2008; Ferreira and Minikel, 2010; Litman, 2011; Boucher et al., 2013; Lemaire et al., 2016). Most of these studies show a relationship between claim frequencies and the number of driven kilometers which is less than proportional. One of the focus points in our study, see Section 3.2, is to investigate the relationship between the expected number of claims and both exposure-to-risk measures (i.e. time and distance).

Using models involving both policy and telematics predictors, Ayuso et al. (2014, 2016a) study the traveled time and distance to the first accident using Weibull regression models. Paefgen et al. (2014) investigate the relationship between the accident risk and driving habits using logistic regression models. Their case-control study design does not allow for inference on the probability of accident involvement. The difference in time exposure between the vehicles with accident involvement (6 months prior to the accident) and the control group (24 months) is however only used to obtain a per-month distance exposure, but is further neglected in the study. Traditional risk factors were not accounted for, since that information was not available, and the compositional nature of the constructed telematics predictor variables was ignored. In contrast, see Section 3.2, main focus points in our research are (i) combining the new telematics variables with traditional policy(holder) information through a careful model and variable selection process and (ii) incorporating the compositional structure of the telematics variables in the analysis.

3 Telematics insurance data

We consider data from a Belgian portfolio of drivers with motor third party liability (MTPL) insurance. MTPL insurance is the legally compulsory minimum insurance covering damage to third parties' health and property caused by an accident for which the driver of the vehicle is responsible. The special type of MTPL product we are considering, is specifically aiming for young drivers who are traditionally facing high insurance premiums. Insureds were offered a substantial discount on their premium if they agree to install a telematics black box device in their car. The telematics box collects statistics on the driving habits: how often one drives, how many kilometers, where and when. Information on the driving style (such as speeding, braking, accelerating, cornering or parking) is not registered. The telematics data did not have an effect on the (future) premium levels of the insureds and did not induce any restrictions on how much or where they can drive.

3.1 Data processing

The unstructured telematics data, collected by the telematics box installed in the vehicle, are first transmitted to the data provider who structures and aggregates these data each day and then reports them to the insurance company as a CSV file (Figure 1a). Only the structured, aggregated telematics information is available to us. Each daily file contains information on the daily driven distance (in meters) for each policyholder. This number of meters is split into 4 road types (*urban, other, motorways* and *abroad*) and 5 time slots (*6h-9h30, 9h30-16h, 16h-19h, 19h-22h* and *22h-6h*). The nature of the data does not allow for a classification of a driven meter by road type and time slot simultaneously. The number of trips, measured as key-on/key-off events, is also reported. This is a typical setup (see Paefgen et al., 2014). In this study, we analyze the telematics data collected between January 1, 2010 and December 31, 2014.

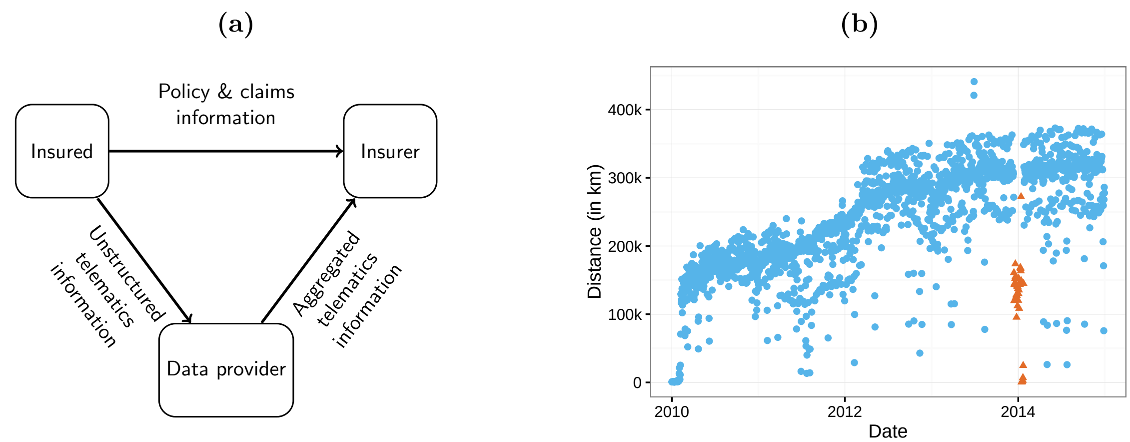

Figure 1: (a) A schematic overview of the flow of information. (b) The number of registered kilometers on each day on an aggregate, portfolio level for the telematics data observed between January 1, 2010 and December 31, 2014. The outliers by the turn of the year 2014, corresponding to a technical malfunction, are indicated as triangles.

The telematics data are linked to the policy(holder) and claims information of the insurance company corresponding to the portfolio under consideration (see Table 1 for a complete list). Policy data, such as age, gender and characteristics of the car, are directly reported by the insured to the insurer at underwriting (see Figure 1a). They are updated over time which enables us to link the claims occurring at a specific moment in time to the correct policy information. Each observation of a policyholder in the policy data set refers to a policy period over which the MTPL insurance coverage holds and contains the most recent policy information. For most insureds, this coverage period is one year, however, it can be smaller for several reasons. If for instance the policyholder decides to add a comprehensive coverage, buys a new vehicle, or changes his residence during the term of the contract, the policy period will be restricted to the date of the policy modification and an additional observation line will be added for the subsequent period. A policy period can also be split when the coverage is suspended for a certain time.

Claims information	
`claims`	number of reported MTPL claims at fault during the policy period
Policy information	
`policy period`	duration in days of the policy period (minimal 30 days and at most one year)
`age`	age of the least experienced driver listed on the policy at the start of the policy period, measured as the number of years between the birth date and the start of the policy period
`experience`	experience of the least experienced driver listed on the policy, measured as the number of years between the date when the driver's permit was obtained and the start of the policy period
`gender`	gender of the least experienced driver listed on the policy (*male* or *female*)
`material damage cover`	indicator whether the insurance policy also covers material damage (*yes* or *no*)
`postal code`	Belgian postal code where the policyholder resides
`bonus-malus`	bonus-malus level of the policy, reflecting the past individual claims experience, between -4 and 22 with lower values indicating a better history
`age vehicle`	age of the vehicle, measured as the number of years between the date when the car was registered and the start of the policy period
`kwatt`	horsepower of the vehicle, measured in kilowatt
`fuel`	fuel type of the vehicle (*petrol* or *diesel*)
Telematics information	
`distance`	distance in meters driven during the policy period
`yearly distance`	distance in meters driven during the policy period, rescaled to a full year by dividing by duration in days of the policy period and multiplying by 365
`trips`	number of trips (*key-on, key-off*) during the policy period
`average distance`	distance in meters driven on average during one trip, obtained by dividing the distance by the number of trips
`road type`	division of the `distance` into 4 road types (*urban, other, motorways* and *abroad*)
`time slot`	division of the `distance` into 5 time slots (*6h-9h30, 9h30-16h, 16h-19h, 19h-22h* and *22h-6h*)
`week/weekend`	division of `distance` into *week* (Monday to Friday) and *weekend* (Saturday, Sunday)

Table 1: Description of the variables contained in the data set arising from the different sources of information.

Using the policy number and period we first merge the telematics information on daily level with the policy data set. Next, we adjust the start and end date of the policy periods based on the first and last day at which telematics data are observed for each policy period of each insured. This ensures that the adjusted policy periods reflect time periods over which both the insurance coverage holds and telematics data are collected. Based on Figure 1b, where we plot the evolution of the driven distance on each day by all drivers of the portfolio, we suspect that technical deficiencies of the data provider can cause an underreporting of the number of meters driven on an aggregate level. The outliers indicated as triangles by the turn of the year 2014 could be linked to a serious technical failure preventing telematics information from being reported for a significant part of our portfolio. We dealt with this by removing this period of roughly one month from the policy periods of all insureds. In the remainder of the observation period between January 1, 2010 and December 31, 2014, clear causes of underreporting could not be identified and hence we did not take any other corrective action. However, this illustrates that data reliability forms a challenge for this new telematics technology. We further removed those observations with a policy duration of less than 30 days in order to avoid senseless observations of only a couple of days and retained only the complete observations with no missing policyholder information.

Next, we aggregate the telematics information by policyholder and period. This means that we sum the driven distance, their divisions into 4 road types and 5 time slots, and the number of trips made. Finally, we use the claims information to extract the number of MTPL claims at fault that occurred between the start and end date of the adjusted policy periods for each policy record.

Over the time period of this study, we end up with a data set of 33 259 observations. Table 1 gives an overview of the available variables coming from the three data sources (claims, policy, and telematics). These observations correspond to 10 406 unique policyholders, who are followed over time, have jointly driven over 297 million kilometers during a combined insured policy period of 17 681 years and reported 1481 MTPL claims at fault. Hence, on average, there were 0.0838 claims per insured year or 0.0499 claims per 10 000 driven kilometers. For over 95% of the observations no claim occurred during the corresponding policy period, whereas for 52 observations two claims occurred and for a single observation even three during the same policy period.

3.2 Risk classification using policy and telematics information

The goal of this research is to build a rating model to express the number of claims as a function of the available covariates. Two sources of information are combined which are described in detail in Table 1. First, there is the self-reported policy information which contains all rating variables traditionally used in car insurance pricing. The second source of information is derived from the telematics data. The main objective is to discover the relevance and impact of adding the new telematics insights using flexible statistical modeling techniques in combination with appropriate model and variable selection tools. One of the key questions is whether the risk transferred from the policyholder to the insurer is proportional to the duration of the policy period or the driven distance during that time. Telematics technology allows a shift to be made from time as exposure to distance as exposure. This would lead to a form of pay-as-you-drive insurance, where a driver pays for every kilometer driven. Histograms of both potential exposure variables are contrasted in Figure 2a and 2b.

In order to investigate the influence and explanatory power of the telematics variables in predicting the risk of an accident, we compare the performance of four sets of predictor variables used to model the number of claims, see Figure 2c. The *classic* set only contains policy information and uses time as exposure-to-risk. The *telematics* set only contains telematics information and uses the distance in meters as exposure-to-risk. The two other models, *time hybrid* and *meter-hybrid*, both contain policy and telematics information. Whereas the first one uses time as an exposure

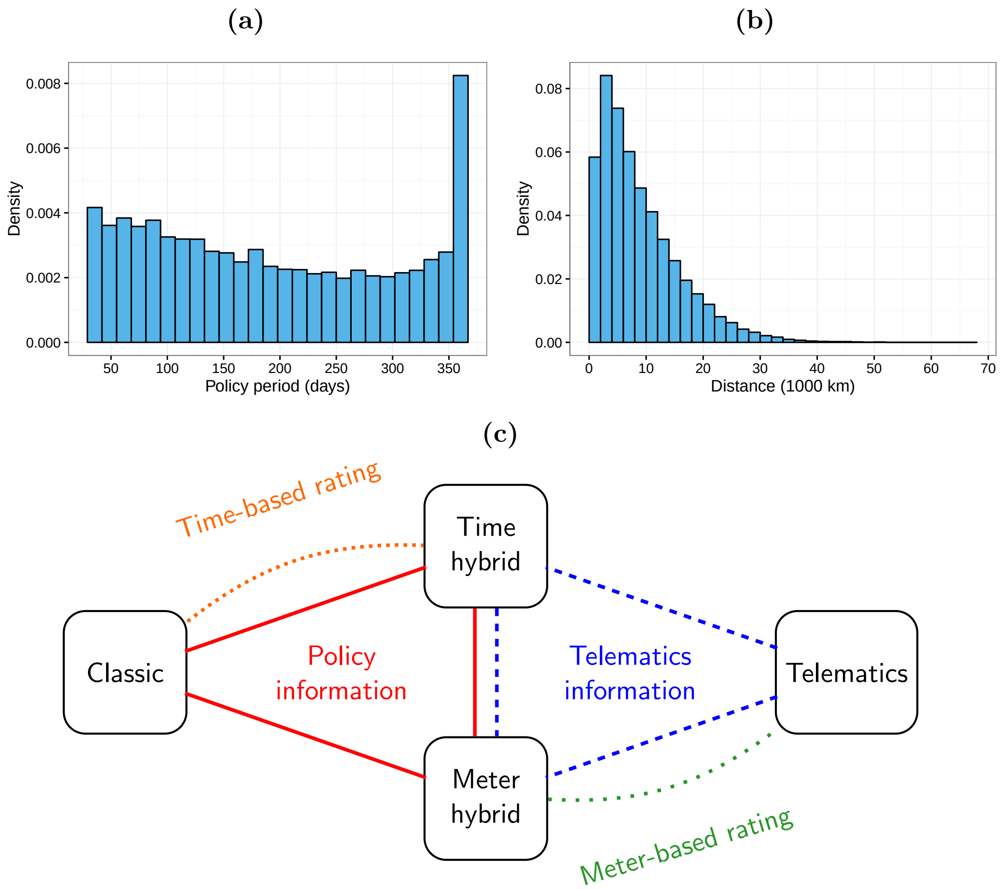

Figure 2: Histogram of (a) the duration (in days) of the policy period (at most one year) and (b) the driven distance (in 1000 km) during the policy period. (c) A graphical representation of the similarities and differences between the four predictor sets.

measure, the second one uses distance. These four predictor sets contrast on the one hand the use of traditional policy rating variables and telematics variables and on the other hand the use of policy duration versus distance as exposure measures in the assessment of the risk.

The main predictors based on the policy information besides the duration of the policy period include the age of the driver, the experience as measured using the driver's license age, the gender, characteristics of the car and the postal code where the policyholder lives. In the case of multiple insured drivers (around 18% of the observations), we select (in consultation with the insurer) the age, gender, experience and postal code belonging to the driver with the most recent permit and hence the lowest experience. This is in line with the strategy of the insurer who offers this type of insurance contract to young drivers. The bonus-malus level is a special kind of variable that reflects the past individual claims experience. It is a function of the number of claims reported in previous years with values between -4 and 22 where lower levels indicate a better history. The insurer uses a slightly modified version of the former compulsory Belgian bonus-malus system, which all companies operating in Belgium have been obliged to use from 1992 to 2002, with minor refinements for the policyholders occupying the lowest levels in the scale. Despite the deregulation, many insurers in the Belgian market still apply the former mandatory system (Denuit et al., 2007). Even though the bonus-malus scale level is not a covariate of the same type as the other a priori variables, we keep it in the analysis to have an idea of the information contained in this variable (as is also done in, for instance, Denuit and Lang, 2004). From a statistical point of view, it tries

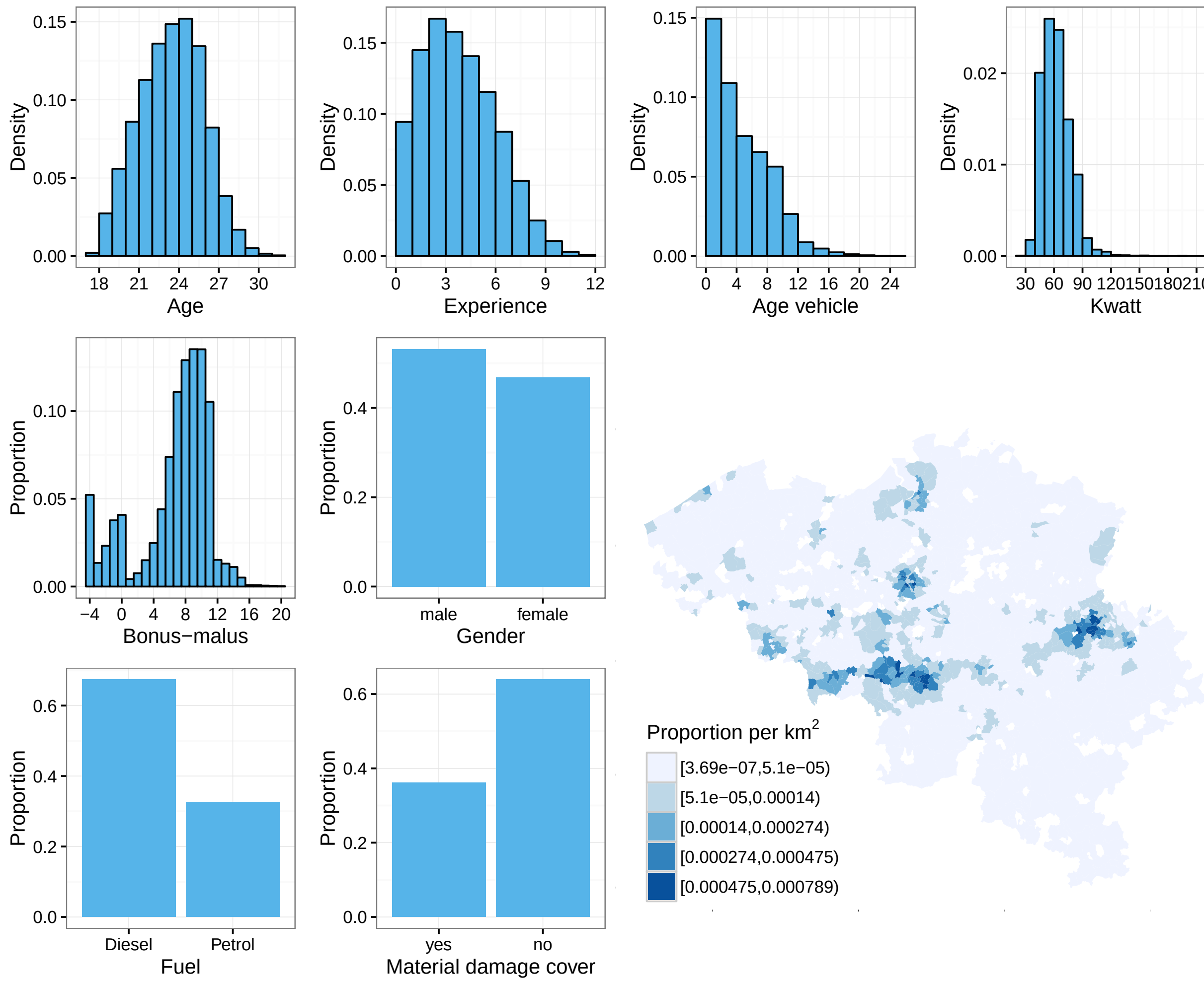

Figure 3: Histograms and bar plots of the continuous and categorical policy variables contained in the data set. The map in the lower right depicts the geographical information by showing the proportion of insureds per squared kilometer living in each of the different postal codes in Belgium. The five class intervals have been created using k-means clustering.

to structure dependencies between observations arising from the same policyholder. An overview of the policy predictor variables and their sample distributions is given in Figure 3.

In the telematics information set we use the distance driven during the policy period as a predictor but we also create two additional telematics variables, the yearly and average distance driven, see Table 1. Histograms of these variables are shown in Figure 4. The divisions of the driven distance by time slot, road type and week/weekend are highly correlated with the total driven distance as they sum up to this amount. To distinguish the absolute information measured by the driven distance in a certain policy period from the compositional information of the distance split into different categories, we consider box plots of the relative proportions in Figure 4. These relative proportions sum to one for each observation.To stress this interconnectedness present in the different splits, we show the compositional profiles of a sample of 100 drivers on top of the marginal box plots. Another important point to stress is that not all components of a certain division of the distance are present for each observation. For instance, if an insured does not drive abroad during the policy period, the relative proportion of the driven distance abroad will be zero. The use of such compositional information as predictors in statistical modeling is a key issue in this research.

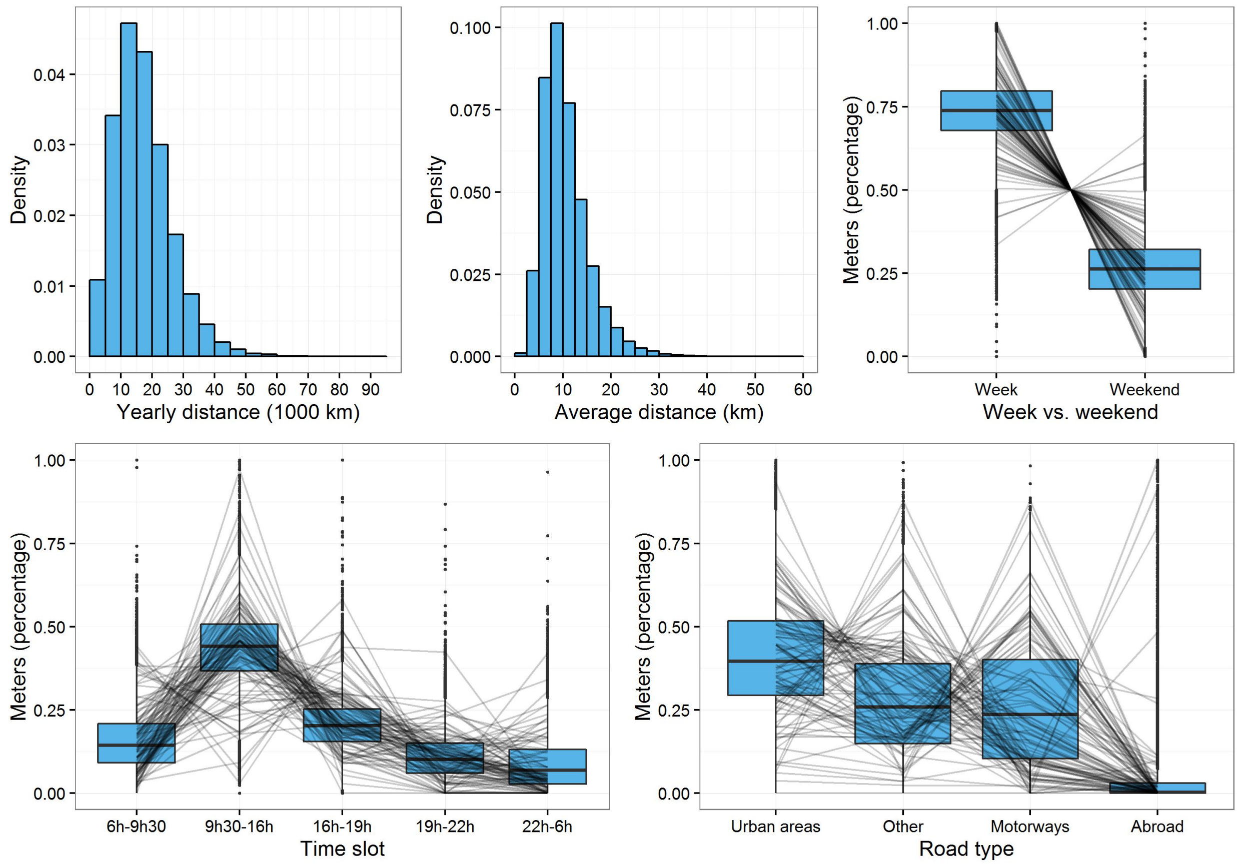

Figure 4: Graphical illustration of the telematics variables contained in the data set. For the yearly and average distance, we construct histograms. For the division of the driven distance by road types, time slots and week/weekend, we construct box plots of the relative proportions. To highlight the dependencies intrinsic to the fact that the division into different categories sums to one, we plot profile lines for 100 randomly selected observations in the data set.

4 Model building and selection

We model the frequencies of claims by constructing Poisson and negative binomial (NB) regression models. We denote by N_{it} the number of claims for policyholder i in policy period t with $i = 1, \ldots, I$ and $t = 1, \ldots, T_i$. The model is denoted by $N_{it} \sim \text{Poisson}(\mu_{it})$ or $N_{it} \sim \text{NB}(\mu_{it}, \phi)$, where $\mu_{it} = \mathbb{E}(N_{it})$ represents the expected number of claims reported by policyholder i in policy period t and ϕ is the parameter of the NB distribution such that $\text{Var}(N_{it}) = \mu_{it} + \mu_{it}^2/\phi$, allowing for overdisperion. A log linear relationship between the mean and the predictor variables is specified by the log link function. This means that we set $\mu_{it} = \exp(\eta_{it})$ where η_{it} is a predictor function of the available explanatory factors. The probability mass functions for the Poisson and the NB models are, respectively, expressed as

$$\mathbb{P}(N_{it} = n_{it}) = \frac{\exp(-\mu_{it})\mu_{it}^{n_{it}}}{n_{it}!} \quad \text{and} \quad \mathbb{P}(N_{it} = n_{it}) = \left(\frac{\phi}{\phi + \mu_{it}}\right)^{\phi} \frac{\Gamma(\phi + n_{it})}{n_{it}!\Gamma(\phi)} \left(\frac{\mu_{it}}{\phi + \mu_{it}}\right)^{n_{it}}.$$

For each of the predictor sets in Figure 2c we construct the best model using the allowed information based on AIC, see Section 4.3. Additionally, we identify the best models under the restriction that the risk is proportional to the time or meter exposure. This is accomplished by incorporating the logarithm of the exposure-to-risk, either duration of the policy period or total distance driven during the policy period, as an offset term in the predictor, i.e. a regression variable with a constant coefficient of 1 for each observation. In the most general case, the predictor has the form

$$\eta_{it} = \beta_0 + \text{offset} + \eta_{it}^{\text{cat}} + \eta_{it}^{\text{cont}} + \eta_{it}^{\text{spatial}} + \eta_{i}^{\text{re}} + \eta_{it}^{\text{comp}}\,, \tag{1}$$

where β_0 denotes the intercept, the categorical effects are bundled in η_{it}^{cat}, the term η_{it}^{cont} contains the effects of the continuous predictors, $\eta_{it}^{\text{spatial}}$ represents the geographical effect, η_{i}^{re} the policyholder-specific random effect and the term η_{it}^{comp} embodies the effects of the compositional predictors. Under the offset restriction, the continuous effect of the exposure-to-risk, either the duration of the policy period (time based rating) or the driven distance (meter based rating), gets replaced by the logarithm of the exposure-to-risk as an offset.

Zero inflated variants of these models are not considered because of interpretability reasons. Such models are not able to capture the effect of a varying exposure-to-risk in a transparent and intuitive way.

4.1 Generalized additive models

The model framework we work with in this study is the one of generalized additive models (GAMs), introduced by Hastie and Tibshirani (1986). GAMs allow to incorporate continuous covariates in a more flexible way as compared to the traditional GLMs used in actuarial practice (see e.g. Klein et al., 2014). From an accuracy standpoint, GAMs are competitive with popular black box machine learning techniques (such as neural networks, random forests or support vector machines), but they have the important advantage of interpretability. In insurance pricing it is of crucial importance to have interpretable results in order to understand the premium structure and explain this to clients and regulators. Using a semiparametric additive structure, GAMs define nonparametric relationships between the response and the continuous variables in the predictor in the following way

$$\eta_{it}^{\text{cat}} + \eta_{it}^{\text{cont}} = \boldsymbol{Z}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_{jit})\,,$$

where $\boldsymbol{Z}_{it}$ represents the row corresponding to policyholder i in policy period t of the model matrix of the categorical variables with parameter vector $\boldsymbol{\beta}$ and f_j represents a smooth function of the jth continuous predictor variable. To estimate f_j, we choose cubic spline basis functions B_{jk}, such that in our models $f_j(x) = \sum_{k=1}^{q} \gamma_{jk} B_{jk}(x)$. The knots are chosen using 10 quantiles of the unique x_j values. Cardinal basis functions parametrize the spline in terms of its values at the knots (Lancaster and Salkauskas, 1986). For identifiability, we impose constraints by centering each smooth component around zero, thus $\sum_{i=1}^{I}\sum_{t=1}^{T_i} f_j(x_{jit}) = 0$ for $j = 1, \ldots, J$. To avoid overfitting, the cubic splines are penalized by the integrated squared second derivative (Green and Silverman, 1994), which yields a measure for the overall curvature of the function. For each component, this penalty can be written as a quadratic function,

$$\int \left(f_j''(x)\right)^2 dx = \sum_{k=1}^{q}\sum_{l=1}^{q} \gamma_{jk}\gamma_{jl} \int B_{jk}''(x) B_{jl}''(x) dx = \boldsymbol{\gamma}_j^t \boldsymbol{S}_j \boldsymbol{\gamma}_j\,,$$

with $(\boldsymbol{S}_j)_{kl} = \int B_{jk}''(x) B_{jl}''(x) dx$. Given these penalty functions for each component, we define the penalized log-likelihood as

$$\ell(\boldsymbol{\psi}) - \frac{1}{2}\sum_{j=1}^{J} \lambda_j \boldsymbol{\gamma}_j^t \boldsymbol{S}_j \boldsymbol{\gamma}_j\,, \tag{2}$$

where $\ell(\boldsymbol{\psi})$ denotes the log likelihood as a function of all model parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J)^t$ and λ_j denotes the smoothness parameter that controls the tradeoff between goodness of fit and the degree of smoothness of component f_j for $j = 1, \ldots, J$. Different smoothing parameters for each component allow to penalize the smooth functions differently.

The model parameters $\boldsymbol{\psi}$ are estimated by maximizing (2) using penalized iteratively reweighted least squares (P–IRLS) (Wood, 2006). For the Poisson model, the smoothing parameters $\lambda_1, \ldots, \lambda_J$ are estimated using an unbiased risk estimator criterion (UBRE), which is a rescaled version of Akaike's information criterion (AIC; Akaike, 1974). For the negative binomial model, we estimate the smoothing parameters and the scale parameter ϕ using maximum likelihood (ML). Alternatively, the smoothing parameters can also be estimated using restricted maximum likelihood (REML; Krivobokova and Kauermann, 2007; Reiss and Ogden, 2009; Wood, 2011).

In addition to categorical and continuous covariates, the data set contains spatial information, namely the postal code where the policyholder resides. Insurance companies tend to use the geographical information of the insured's residence as a proxy for the traffic density and for other unobserved socio-demographic factors of the neighborhood. We model the spatial heterogeneity of claim frequencies by adding a spatial term $\eta_{it}^{\text{spatial}} = f_s(\texttt{lat}_{it}, \texttt{long}_{it})$ in the additive predictor η_{it}, using the latitude and longitude coordinates (in degrees) of the center of the postal code where the policyholder resides. We use second order smoothing splines on the sphere (Wahba, 1981) to model f_s. This allows us to quantify the effect of the geographic location while taking the regional closeness of the neighboring postal codes into account.

In our data set, many policyholders $i = 1, \ldots, I$ are observed over multiple policy periods $t = 1, \ldots, T_i$. This longitudinal aspect of the data can be modeled by including policyholder-specific random effects η_i^{re} in the predictor. The generalized additive model considered thus far is extended in this way by exploiting the link between penalized estimation and random effects (see e.g. Ruppert et al., 2003). We assess whether such random effects are needed to take the correlations between observations of the same policyholder into account using the approximate test for a zero random effect developed by Wood (2013).

4.2 Compositional data

The divisions of the total driven distance into the different categories – road types (4), time slots (5) and week/weekend (2), see Table 1 – are highly correlated with and sum up to the total driven distance. Hence in Figure 4 we divided all components of each split by the total driven distance. Incorporating these divisions, either in absolute or relative terms, in a predictor also containing the total distance leads to a perfect multicollinearity problem. The most straightforward way to deal with this would be to leave one component out, but this approach is not permutation invariant and the statistical inference will depend on which component is removed, making interpretations misleading. The standard regression interpretation of a change in one of the components of the distance when the other components are held constant is not possible due to the sum constraint. We introduce and further develop the necessary statistical tools to model such predictors.

In the literature, data which quantitatively describe the parts of some whole and provide only relative information between their components are called *compositional data* (van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). Typical examples include mineral compositions, molar concentrations and household budgets. In our setting, the divisions of the distance driven are compositional. Scale invariance is a key property: if a composition is scaled by a constant, the information carried is completely equivalent. Therefore compositional data can be represented by real vectors with positive components that sum to one. The space of representations of compositions is called the simplex of D parts, defined by

$$\mathcal{S}^D = \left\{ \boldsymbol{x} = (x_1, \dots, x_D)^t : x_i > 0, \sum_{i=1}^{D} x_i = 1 \right\}.$$

When data are considered compositional, classical statistics, that do not take the special geometry of the simplex into account, are not appropriate. Section 4.2.1 revises the necessary geometrical concepts to work with compositional data. Extending the current literature, we propose a new way of quantifying and interpreting the effect of the compositional explanatory variables on the outcome in Section 4.2.2. Section 4.2.3 introduces two approaches to accommodate for structural zeros in regression with compositional predictors.

4.2.1 The Aitchison geometry of the simplex

Aitchison (1986) introduced operations between compositional data vectors which define a vector space structure on the mathematical simplex known as the *Aitchison geometry of the simplex*. *Perturbation* plays the role of addition on the simplex and is defined as a closed component-wise product $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \dots, x_D y_D)^t$, where the closing operation $\mathcal{C}$ ensures a total sum of 1, i.e. the closure of $\boldsymbol{x}$ is $\mathcal{C}(\boldsymbol{x}) = \boldsymbol{x} / \sum_{i=1}^{D} x_i$. The product of a vector by a scalar is called *powering* and is defined as the closed component-wise powering of a composition by a scalar, i.e. $\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^{\alpha}, \dots, x_D^{\alpha})^t$, for $\alpha \in \mathbb{R}$. The *Aitchison inner product* for compositions,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^{D} \ln(x_i) \ln(y_i) - \frac{1}{D} \left(\sum_{i=1}^{D} \ln(x_i) \right) \left(\sum_{j=1}^{D} \ln(y_j) \right)$$

is proportional to the scalar product of the vectors formed by all possible pairwise logratios of the two compositions and induces the following norm $\|\boldsymbol{x}\|_a = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_a}$ and distance $\mathrm{d}_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a$, where $\ominus$ represents the opposite operation of $\oplus$, i.e. $\ominus \boldsymbol{y} = \oplus((-1) \odot \boldsymbol{y})$, a closed component-wise division. The simplex along with these operations then forms a $(D-1)$-dimensional Euclidean vector space $(\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$. Given this Euclidean structure, we can measure distances and angles, and define related geometrical concepts. Elementary statistical notions involving the metrics of the sample space can be adapted to the Euclidean structure of the simplex.

Compositional data are analyzed using a logratio approach and compositions $\boldsymbol{x}$ from the simplex can be represented in the real space using the *centered logratio transformation (clr)*,

$$u_i = \mathrm{clr}_i(\boldsymbol{x}) = \ln \left(\frac{x_i}{g(\boldsymbol{x})} \right), \quad i = 1, \dots, D, \quad g(\boldsymbol{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D} \tag{3}$$

where $g(\boldsymbol{x})$ denotes the geometric mean of the components. The clr transformation defines an isometry between $\mathcal{S}^D$ and the $(D-1)$-dimensional subspace of $\mathbb{R}^D$ of vectors whose components add to zero, denoted by $\mathcal{H}^D = \{ \boldsymbol{u} \in \mathbb{R}^D | \sum_{i=1}^{D} u_i = 0 \}$ and called the clr-plane. Using matrix notation we can write the clr transform and its inverse as

$$\boldsymbol{u} = \mathrm{clr}(\boldsymbol{x}) = \ln(\boldsymbol{x}/g(\boldsymbol{x})), \quad \text{and} \quad \boldsymbol{x} = \mathrm{clr}^{-1}(\boldsymbol{u}) = \mathcal{C}(\exp(\boldsymbol{u})), \tag{4}$$

where the logarithmic and exponential function apply componentwise. An orthonormal basis of $\mathcal{S}^D$ can be obtained from an orthonormal basis of $\mathcal{H}^D$ using the inverse clr transformation. A transformation between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ that provides the coordinates of any composition with respect to a given orthonormal basis is called an *isometric logratio transformation (ilr)*. The one originally defined by Egozcue et al. (2003) maps a compositional data vector $\boldsymbol{x}$ in a $(D-1)$–dimensional real vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_{D-1})^t$ with components

$$z_i = \text{ilr}_i(\boldsymbol{x}) = \sqrt{\frac{D-i}{D-i+1}} \ln \left(\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}} \right), \quad i = 1, \ldots, D-1\,. \tag{5}$$

By arranging the corresponding orthonormal basis vectors in $\mathcal{H}^D$ by columns, we obtain a $D \times (D-1)$ matrix V with elements

$$V_{ij} = \frac{D-j}{\sqrt{(D-j+1)(D-j)}} \quad \text{for } i = j, \quad \frac{-1}{\sqrt{(D-j+1)(D-j)}} \quad \text{for } i > j, \quad \text{and 0 otherwise,}$$

for which it holds that $V^t V = I_{D-1}$ and $V V^t = I_D - (1/D)\mathbf{1}_D \mathbf{1}_D^t$, where I_D is the identity matrix of dimension D and $\mathbf{1}_D$ a D-vector of ones (Egozcue et al., 2011). Then we can write the ilr transform and its inverse as

$$\boldsymbol{z} = \text{ilr}(\boldsymbol{x}) = V^t \ln \boldsymbol{x} = V^t \text{clr}(\boldsymbol{x}), \quad \text{and} \quad \boldsymbol{x} = \text{ilr}^{-1}(\boldsymbol{z}) = \mathcal{C}(\exp(V\boldsymbol{z})). \tag{6}$$

The clr and ilr transformations reflect how all relevant information of a composition is conveyed by the component logratios. In the case $D = 2$, the ilr transformation (5) is proportional to the logit function, which is used in logistic regression to transform the probability $0 < p < 1$ of a binary response into an unrestricted log-odds.

As the clr and ilr transformations are isometric, all angles and distances are preserved. This means that, whenever compositions are transformed into coordinates, the metrics and operations in the Aitchison geometry of the simplex are translated into the ordinary Euclidean metrics and operations in real space. For instance, the Aitchison inner product of two compositions is equal to the real inner product of their clr or ilr transformed vectors,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \langle \text{clr}(\boldsymbol{x}), \text{clr}(\boldsymbol{y}) \rangle = \sum_{i=1}^{D} \text{clr}_i(\boldsymbol{x}) \text{clr}_i(\boldsymbol{y}) = \langle \text{ilr}(\boldsymbol{x}), \text{ilr}(\boldsymbol{y}) \rangle = \sum_{i=1}^{D-1} \text{ilr}_i(\boldsymbol{x}) \text{ilr}_i(\boldsymbol{y}).$$

Even though the simplex $\mathcal{S}^D$ is a subset of the real space $\mathbb{R}^D$, Aitchison (1986) showed that the geometry is clearly different. Ignoring the compositional nature of the data in a statistical context can lead to incompatible or incoherent results. The principle of working on coordinates (Mateu-Figueras et al., 2011) is to first express the compositional data with respect to an orthonormal basis of the underlying vector space with Euclidean structure. Next, to apply standard statistical techniques to the vectors of coordinates and, finally, to back-transform and describe the results in terms of the simplex. Final results do not depend on the chosen basis.

4.2.2 A new interpretation for compositional predictors

In our framework, the total distance in meters is used as a continuous predictor in the telematics models and its effect is modeled using a smooth function. In addition, we propose to treat the divisions of the driven distance by road types, time slots and week/weekend as compositional data covariates in the claim count regression models. In this way, the effects of the absolute information of the total distance driven and the relative information contained in the different divisions can be structured and interpreted separately.

In the context of linear regression, Hron et al. (2012) propose to first apply the isometric logratio transform (5) to map the compositions in the D-part Aitchison simplex to a $(D-1)$ Euclidean space before including them as explanatory variables. More generally, in any regression context involving a predictor, one can add a *compositional predictor* term η^{comp} using the ilr transformed variables, i.e.

$$\eta^{\text{comp}} = \beta_1 z_1 + \ldots + \beta_{D-1} z_{D-1} \,. \tag{7}$$

The fitted model does not depend on the choice of the orthonormal ilr basis since the coordinates of $\boldsymbol{x}$ with respect to different orthonormal bases are orthogonal transformations of each other. Using the ilr transformation the model parameters can be estimated without constraints and the ceteris paribus interpretation of altering one z_i without altering any other becomes possible. A drawback is that only the first regression parameter, β_1, has a comprehensible interpretation since z_1 explains relevant information about x_1. The remaining coefficients are not straightforward to interpret. Also the suggestion by Hron et al. (2012) to permute the indices in formula (5) and construct D regression models, each time with a different component first, is undesirable. This is especially so in our case where we have more than one compositional predictor and each model fit is computationally intensive due to smooth continuous, spatial, and random effects.

Hence, we develop a new way to interpret and visualize the effect of a compositional predictor (7) without the need of refitting the model. Following van den Boogaart and Tolosana-Delgado (2013) and Pawlowsky-Glahn et al. (2015), we start by using the inverse ilr transform on the model coefficients, i.e. set $\boldsymbol{b} = \text{ilr}^{-1}(\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{D-1})^t$, such that we can rewrite the compositional predictor as

$$\eta^{\text{comp}} = \sum_{i=1}^{D-1} \beta_i z_i = \sum_{i=1}^{D-1} \text{ilr}_i(\boldsymbol{b})\text{ilr}_i(\boldsymbol{x}) = \langle \boldsymbol{b}, \boldsymbol{x} \rangle_a \,.$$

The composition $\boldsymbol{b} \in \mathcal{S}^D$ can be interpreted as the simplicial gradient of η^{comp} with respect to $\boldsymbol{x}$ (Barceló-Vidal et al., 2011) and is the compositional direction along which the predictor increases fastest. In particular, if we perturb $\boldsymbol{x}$ by a unit vector $\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a}$ in the direction of $\boldsymbol{b}$, i.e. $\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a}$, then the predictor becomes

$$\tilde{\eta}^{\text{comp}} = \langle \boldsymbol{b}, \tilde{\boldsymbol{x}} \rangle_a = \langle \boldsymbol{b}, \boldsymbol{x} \oplus \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a} \rangle_a = \langle \boldsymbol{b}, \boldsymbol{x} \rangle_a + \frac{1}{\|\boldsymbol{b}\|_a} \langle \boldsymbol{b}, \boldsymbol{b} \rangle_a = \eta^{\text{comp}} + \|\boldsymbol{b}\|_a \,.$$

When $D = 3$, the estimated regression model can be visualized as a surface on a ternary diagram (van den Boogaart and Tolosana-Delgado, 2013). For $D > 3$, a graphical representation is not straightforward.

Further, we propose to perturb the composition in the direction of each component. This offers a new interpretation for the effect of altering the composition on the predictor. For example, a relative ratio change of $\alpha > 1$ (increase) or $\alpha < 1$ (decrease) in the first component of $\boldsymbol{x}$ with constant ratios of the remaining components can be achieved by perturbing the composition $\boldsymbol{x}$ by $\mathcal{C}(\alpha, 1, \ldots, 1)^t$, i.e. $\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \mathcal{C}(\alpha, 1, \ldots, 1)^t = \mathcal{C}(\alpha x_1, x_2, \ldots, x_D)^t$. This leads to a change of the predictor given by

$$\langle \boldsymbol{b}, \mathcal{C}(\alpha, 1, \ldots, 1)^t \rangle_a = \ln(\alpha) \Big[\ln(b_1) - \frac{1}{D} \Big(\sum_{i=1}^{D} \ln(b_i) \Big) \Big] = \ln(\alpha)\text{clr}_1(\boldsymbol{b}) \,, \tag{8}$$

which is independent of the original composition $\boldsymbol{x}$. The effect of a relative increase in any of the components can hence best be understood by considering the clr transform of $\boldsymbol{b}$, of which the elements sum to zero and indicate the positive or negative effect of each component on the predictor. The difference in the predictor level between any two compositional data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be computed as

$$\langle \boldsymbol{b}, \boldsymbol{x} \ominus \boldsymbol{y} \rangle_a = \sum_{i=1}^{D} \ln\left(\frac{x_i}{y_i}\right) \mathrm{clr}_i(\boldsymbol{b}) \,,$$

a linear combination of the clr coordinates of $\boldsymbol{b}$ with weights given by the component-wise logratios. A graphical representation of the effect of a compositional predictor can be made by visualizing $\mathrm{clr}(\boldsymbol{b})$ and comparing the elements to zero. Since $\boldsymbol{\beta} = \mathrm{ilr}(\boldsymbol{b}) = V^t \ln(\boldsymbol{b}) = V^t \mathrm{clr}(\boldsymbol{b})$ and $VV^t = I_D - (1/D)\mathbf{1}_D\mathbf{1}_D^t$, the clr transform of $\boldsymbol{b}$ can be written as $\mathrm{clr}(\boldsymbol{b}) = V\boldsymbol{\beta}$. Confidence bounds can thus be constructed using the corresponding covariance matrix $V\widehat{\Sigma}V^t$ where $\widehat{\Sigma}$ is the estimated covariance matrix related to estimating $\boldsymbol{\beta}$. To quantify the influence of the compositional predictor on the level of the expected outcome in the Poisson and NB models, we exponentiate (8). The effect of a relative ratio change of α in component $i = 1, \ldots, D$ on the response scale is then given by $\alpha^{\mathrm{clr}_i(\boldsymbol{b})}$.

4.2.3 Dealing with structural zeros in compositional predictors

An additional difficulty when incorporating the compositional information as predictors in the analysis of the claim counts is the presence of proportions of a specific component that are exactly zero. In the division of the driven distance by road type, for instance, many insureds did not drive abroad during the observed policy period. Since compositional data are always analyzed by considering logratios of the components (see Section 4.2.1), a workaround is necessary.

In the compositional data literature, different types of zeros are being distinguished (Pawlowsky-Glahn et al., 2015). *Rounded zeros* occur when certain components may be unobserved because their true values are below the detection limit (cfr. geochemical studies). *Count zeros* refer to zero values due to the limited size of the sample in compositional data arising from count data. In our setting, the zero values are truly zero and are not due to imprecise or insufficient measurements. Such kind of zeros are called *structural zeros*. The structural zeros patterns in the data set are listed in Appendix A of the supplementary material. The presence of zeros is most prominent for splitting distance by road types as 40% of the drivers did not go abroad. Zeros are most often dealt with using replacement strategies (see e.g. Martín-Fernández et al., 2011, for an overview), which do not make sense for structural zeros. A general methodology is still to be developed (see e.g. Aitchison and Kay, 2003; Bacon Shone, 2003). In particular, there does not exist a method that deals with compositional data with structural zeros as predictor in regression models. Applying the ilr transform to the compositional data $\boldsymbol{x}$ and using the transformed $\boldsymbol{z}$ as explanatory variables in the predictor as discussed in Section 4.2.2 is no longer possible.

For an observation with structural zeros, we can only consider the *subcomposition* of nonzero components. We let $M \subset \{1, 2, \ldots, D\}$ denote the set of indices of the structural zeros of a composition $\boldsymbol{x}$. The subcomposition $\boldsymbol{x}_{M^c}$ of nonzero components $M^c = \{i_1, \ldots, i_m\} = \{1, 2, \ldots, D\} \setminus M$ of $\boldsymbol{x}$ is then obtained by applying the closure operation to the subvector $(x_{i_1}, \ldots, x_{i_m})$ of $\boldsymbol{x}$, i.e. $\boldsymbol{x}_{M^c} = \mathcal{C}(x_{i_1}, \ldots, x_{i_m}) = (x_{i_1}, \ldots, x_{i_m})/\sum_{i \in M^c} x_i$. The set of all possible structural zero patterns M is denoted by $\mathcal{M}$. In the most general situation, 2^D-1 possible zero patterns can occur when dealing with compositional data with D components (a structural zero for every component being excluded). To indicate the zero pattern of a composition $\boldsymbol{x}$ we introduce dummy variables

$$d_M(\boldsymbol{x}) = \begin{cases} 1 & \text{if the set of indices of the structural zeros of } \boldsymbol{x} \text{ is equal to } M\,, \\ 0 & \text{otherwise} \end{cases}$$

for all $M \in \mathcal{M}$. We propose two approaches to accommodate for structural zeros in a regression context with compositional predictors, either via conditioning on the structural zero pattern or via projection onto the orthogonal complement of the structural zero parts.

Conditioning approach We treat observations with different structural zero patterns as qualitatively different subgroups within the data and model the effect of the compositional predictor conditional on the zero pattern. The compositional predictor term η^{comp} of the regression model specifies a distinct effect $\boldsymbol{b}|_{M^c}$ for each zero pattern:

$$\eta^{\text{comp}} = \sum_{M\in\mathcal{M}} d_M(\boldsymbol{x})\langle \boldsymbol{b}|_{M^c}, \boldsymbol{x}_{M^c}\rangle_a = \sum_{M\in\mathcal{M}} d_M(\boldsymbol{x})\langle \text{ilr}(\boldsymbol{b}|_{M^c}), \text{ilr}(\boldsymbol{x}_{M^c})\rangle,$$

Conditional on the zero pattern M of the compositional data vector $\boldsymbol{x}$, the contribution to the predictor is given by the Aitchison inner product of the subcomposition $\boldsymbol{x}_{M^c}$ of nonzero components of $\boldsymbol{x}$ and a subcompositional simplicial gradient $\boldsymbol{b}|_{M^c}$ of the same dimension. The notation $\boldsymbol{b}|_{M^c}$ is used to indicate that the model parameters $\boldsymbol{\beta}|_{M^c} = \text{ilr}(\boldsymbol{b}|_{M^c})$ differ by structural zero pattern. Fitting this term requires us to compute for each compositional observation the ilr transform $\text{ilr}(\boldsymbol{x}_{M^c})$ of the subcomposition of nonzero parts and to model the compositional predictor effect separately by zero pattern. Note that in case of only one nonzero component, the Aitchison inner product is zero and there is no contribution to the linear predictor. If deemed necessary a categorical effect based on the zero pattern can be added to η^{comp}.

Projection approach The compositional regression coefficients in the conditioning approach are different for each structural zero pattern and hence only estimated using observations with that particular zero pattern. Instead of modeling the compositional predictor effect separately by zero pattern, we alternatively propose a parsimonious simplification in which the regression parameters are shared across patterns.

To this end, we regard a subcomposition $\boldsymbol{x}_{M^c}$ as an orthogonal projection of $\boldsymbol{x}$ that preserves the relative information contained in $\boldsymbol{x}_{M^c}$ and, simultaneously, filters out all the relative information involving parts in $\boldsymbol{x}_M$ (Pawlowsky-Glahn et al., 2015). The clr plane $\mathcal{H}^D$ is spanned by the non-orthogonal, non-basis vectors $\boldsymbol{w}_i = \left(\frac{-1}{D}, \dots, \frac{-1}{D}, \frac{D-1}{D}, \frac{-1}{D}, \dots, \frac{-1}{D}\right)$, for $i = 1, \dots, D$, where the component equal to $\frac{D-1}{D}$ is placed at the ith component. The subcomposition $\boldsymbol{x}_{M^c}$ of the nonzero parts can be represented by an orthogonal projection P_M of the clr transformed vector $\text{clr}(\boldsymbol{x})$ onto the null space of $\{\boldsymbol{w}_i, i \in M\}$, corresponding to the indices of the structural zeros of $\boldsymbol{x}$. van den Boogaart et al. (2006) show that the projection P_M onto the orthogonal directions to M can be computed as

$$(P_M \text{clr}(\boldsymbol{x}))_{M^c} = \text{clr}(\boldsymbol{x}_{M^c})$$

and zero otherwise. Hence, the subvector of $P_M\text{clr}(\boldsymbol{x})$ related to the nonzero parts of $\boldsymbol{x}$ equals the clr transform of the subcomposition $\boldsymbol{x}_{M^c}$ and the remaining elements of $P_M\text{clr}(\boldsymbol{x})$ related to the structural zeros of $\boldsymbol{x}$ equal zero. We can express this projected clr vector with respect to the chosen orthonormal ilr basis and define $\boldsymbol{z} = V^t P_M \text{clr}(\boldsymbol{x})$ as a *generalized isometric logratio transformation* from the simplex (allowing for zero components) to $\mathbb{R}^{D-1}$. Note that in case $\boldsymbol{x}$ has no structural zeros, the generalized ilr transform coincides with the regular ilr transform in (6). We suggest to use these generalized ilr coordinates in the compositional predictor and rewrite the term as

$$\begin{aligned}\eta^{\text{comp}} &= \sum_{i=1}^{D-1} \beta_i z_i = \text{clr}(\boldsymbol{b})^t V V^t P_M \text{clr}(\boldsymbol{x}) = \langle \text{clr}(\boldsymbol{b}), P_M\text{clr}(\boldsymbol{x})\rangle = \langle (\text{clr}(\boldsymbol{b}))_{M^c}, \text{clr}(\boldsymbol{x}_{M^c})\rangle \\ &= \langle \text{clr}(\boldsymbol{b}_{M^c}), \text{clr}(\boldsymbol{x}_{M^c})\rangle = \langle \boldsymbol{b}_{M^c}, \boldsymbol{x}_{M^c}\rangle_a, \end{aligned} \tag{9}$$

where we used the fact that the elements of a clr transform sum to zero and that $(\text{clr}(\boldsymbol{b}))_{M^c}$, the subvector of $\text{clr}(\boldsymbol{b})$ related to the nonzero parts of $\boldsymbol{x}$, and $\text{clr}(\boldsymbol{b}_{M^c})$, the clr transform of the subcomposition of $\boldsymbol{b}$ related to the nonzero parts of $\boldsymbol{x}$, only differ by a vector of equal elements.

Equation (9) shows that the compositional predictor in the projection approach is equivalent to the Aitchison inner product of the subcompositions of both $\boldsymbol{b}$ and $\boldsymbol{x}$ corresponding to the nonzero components of $\boldsymbol{x}$. In general, we can write

$$\eta^{\text{comp}} = \sum_{M \in \mathcal{M}} d_M(\boldsymbol{x}) \langle \boldsymbol{b}_{M^c}, \boldsymbol{x}_{M^c} \rangle_a \,.$$

Compared to the conditioning approach where the compositional regression coefficients $\boldsymbol{b}|_{M^c}$ are conditional on the zero pattern, the projection approach is more parsimonious. The effect of each subgroup defined by the structural zero patterns is obtained from the same model parameters $\boldsymbol{\beta}$, using subcompositions $\boldsymbol{b}_{M^c}$ of the corresponding compositional coefficient vector $\boldsymbol{b} = \text{ilr}^{-1}(\boldsymbol{\beta})$. This simplifying assumption entails that leaving out the zero components does not change the relative riskiness of the remaining components. Given an observation with structural zeros, the interpretation of the effect of a change to a nonzero component remains similar as before: if the relative ratio of the ith component of $\boldsymbol{x}_{M^c}$ changes by α, then the predictor changes by $\ln(\alpha)\text{clr}_i(\boldsymbol{b}_{M^c})$. The clr transformed subcomposition $\text{clr}(\boldsymbol{b}_{M^c})$ can be obtained by recentering the parts of $\text{clr}(\boldsymbol{b})$ corresponding to M^c around zero, i.e. $\text{clr}(\boldsymbol{b}_{M^c}) = (\text{clr}(\boldsymbol{b}))_{M^c} - \left(\frac{1}{m}\sum_{i \in M^c} \text{clr}_i(\boldsymbol{b})\right) \mathbf{1}_m$, where $m = |M^c|$ is the number of nonzero parts in $\boldsymbol{x}$. Therefore, using the projection approach, a single graphical representation of $\text{clr}(\boldsymbol{b})$ suffices to visualize and understand the effect of the compositional predictor term for each structural zero pattern.

4.3 Model selection and assessment

Using the same form as Akaike's information criterion, AIC for a GAM is defined as

$$\text{AIC} = -2 \cdot \widehat{\ell} + 2 \cdot \text{EDF} \tag{10}$$

where $\widehat{\ell}$ is the log-likelihood, evaluated at the estimated model parameters obtained using penalized likelihood maximization, and the effective degrees of freedom (EDF) is used instead of the actual number of model parameters. For details about the calculation of the EDF see Wood et al. (2016). We used the implementation as available in the `R` package `mgcv` version 1.8-18. As such, (10) measures the quality of the model as a trade-off between the goodness-of-fit and the model complexity.

For each of the four predictor sets, see Figure 2c, variables are selected by AIC using an exhaustive search over all the possible combinations of variables given in Table 1. In our analysis, model selection is done without involving policyholder-specific random effects. All model specifications are estimated under both the Poisson and the negative binomial framework. We restrict to additive regression models (i.e. no interactions) such that an exhaustive search is still feasible and the marginal impact of a single variable can be easily assessed, interpreted and visualized. Even though the 2011 EU ruling prohibits a distinction between men and women in car insurance pricing, we allow gender to be selected as a categorical predictor in the model. For the compositional predictors based on the different divisions of the driven distance, 10 structural zero patterns occur for the road types, 20 for the time slots and 3 for week/weekend, see Appendix A of the supplementary material. The model selection is performed separately for the conditioning and projection approach to the structural zeros. Following the projection approach, the three compositional predictor terms we allow to be selected in the hybrid and telematics models are

$$\begin{aligned}\eta_{it}^{\text{comp}} = & \sum_{M \in \mathcal{M}^{\text{road}}} d_M(\boldsymbol{x}_{it}^{\text{road}}) \langle \boldsymbol{b}_{M^c}^{\text{road}}, \boldsymbol{x}_{it,M^c}^{\text{road}} \rangle_a + \sum_{M \in \mathcal{M}^{\text{time}}} d_M(\boldsymbol{x}_{it}^{\text{time}}) \langle \boldsymbol{b}_{M^c}^{\text{time}}, \boldsymbol{x}_{it,M^c}^{\text{time}} \rangle_a \\ & + \sum_{M \in \mathcal{M}^{\text{week}}} d_M(\boldsymbol{x}_{it}^{\text{week}}) \langle \boldsymbol{b}_{M^c}^{\text{week}}, \boldsymbol{x}_{it,M^c}^{\text{week}} \rangle_a \,.\end{aligned}$$

Following the conditioning approach, the effect of the compositional predictor is modeled separately conditional on the structural zero pattern. However, based on the relative frequencies of the zero patterns in the data set, we only allow an additional compositional predictor term for the distinction by road type in the case that a car did not drive abroad, which occurs for 40% of the observations. All remaining zero patterns are bundled into one residual group and their effect is modeled using a categorical effect b_0, see Table A.4 of Appendix A. Using the symbolic structural zero pattern notation of Appendix A, the most comprehensive compositional predictor term in the conditioning approach can be denoted as

$$\begin{aligned}\eta_{it}^{\text{comp}} = {} & d_{1111}(\boldsymbol{x}_{it}^{\text{road}})\langle \boldsymbol{b}|_{1111}^{\text{road}}, \boldsymbol{x}_{it,1111}^{\text{road}}\rangle_a + d_{1110}(\boldsymbol{x}_{it}^{\text{road}})\langle \boldsymbol{b}|_{1110}^{\text{road}}, \boldsymbol{x}_{it,1110}^{\text{road}}\rangle_a + (1 - d_{1111}(\boldsymbol{x}_{it}^{\text{road}}) \\ & - d_{1110}(\boldsymbol{x}_{it}^{\text{road}}))b|_0^{\text{road}} + d_{11111}(\boldsymbol{x}_{it}^{\text{time}})\langle \boldsymbol{b}|_{11111}^{\text{time}}, \boldsymbol{x}_{it,11111}^{\text{time}}\rangle_a + (1 - d_{11111}(\boldsymbol{x}_{it}^{\text{time}}))b|_0^{\text{time}} \\ & + d_{11}(\boldsymbol{x}_{it}^{\text{week}})\langle \boldsymbol{b}|_{11}^{\text{week}}, \boldsymbol{x}_{it,11}^{\text{week}}\rangle_a + (1 - d_{11}(\boldsymbol{x}_{it}^{\text{week}}))b|_0^{\text{week}} .\end{aligned}$$

Predictive performance of these models is assessed using *proper scoring rules* for count data, see Table 2 (Czado et al., 2009). Scoring rules assess the quality of probabilistic forecasts through a numerical score $s(P, n)$ based on the predictive distribution P and the observed count n. Lower scores indicate a better quality of the forecast. A scoring rule is proper (Gneiting and Raftery, 2007) if $s(Q, Q) \leqslant s(P, Q)$ for all P and Q with $s(P, Q)$ the expected value of $s(P, \cdot)$ under Q. In general, we define by $p_k = \mathbb{P}(N = k)$ and $P_k = \mathbb{P}(N \leqslant k)$ the probability mass function and cumulative probability function of the predictive distribution P for count variable N. The probability mass at the observed count n is denoted as p_n. The mean and standard deviation of P are written as μ_P and σ_P, respectively, and we set $\|p\| = \sum_{k=0}^{\infty} p_k^2$.

Score	Formula
logarithmic	$\text{logs}(P, n) = -\log p_n$
quadratic	$\text{qs}(P, n) = -2p_n + \|p\|$
spherical	$\text{sphs}(P, n) = -\frac{p_n}{\|p\|}$
ranked probability	$\text{rps}(P, n) = \sum_{k=0}^{\infty}\{P_k - \mathbf{1}(n \leqslant k)\}^2$
Dawid-Sebastiani	$\text{dss}(P, n) = \left(\frac{n-\mu_P}{\sigma_P}\right)^2 + 2\log\sigma_P$
squared error	$\text{ses}(P, n) = (n - \mu_P)^2$

Table 2: Proper scoring rules for count data.

We compare the predictive performance of the best models according to AIC under the four predictor sets, with or without offset in the predictor (1), and using a Poisson or NB distribution. We apply the proper scoring rules to the predictive count distributions of the observed claim counts. We adopt a K-fold cross-validation approach (Hastie et al., 2009) with $K = 10$ and apply the same partition to assess each model specification. Let $\kappa_{it} \in \{1, 2, \dots K\}$ be the part of the data to which the observed claim count n_{it} of policyholder i in policy period t is allocated by the randomization. Denote by $\widehat{P}_{it}^{-\kappa_{it}}$ the predictive count distribution for observation n_{it} estimated without the κ_{it}th part of the data. The K-fold cross-validation score $\text{CV}(s)$ is then given by

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

where s is any of the aforementioned proper scoring rules and smaller values of $\text{CV}(s)$ indicate better forecasts.

5 Results

All computations are performed with `R` 3.4.1 (R Core Team, 2017) and, in particular, `R` package `mgcv` version 1.8-18 (Wood, 2011) is used for the parameter estimation in the GAMs and `compositions` version 1.40-1 (van den Boogaart et al., 2014) to compute the transformations of the compositional data. For the brevity and clarity of this presentation, we only show the results (tables and figures) for the Poisson models using the projection approach for the structural zeros and highlight differences with the negative binomial models and the conditioning approach, if any. As supplementary material, we provide accompanying `R` code which illustrates how to apply the methods presented in this work on simulated data with a similar structure. Following either the conditioning or the projection approach to handle structural zeros, we demonstrate how to include a compositional predictor in a GAM and how to visualize the effect.

5.1 Model selection

The variables selected for each of the predictor sets were identical for the Poisson and NB models, see Table 3. The functional forms of the selected best models are given in Appendix B of the supplementary material. The offset versions of the classic and time-hybrid model replace the term $f_1(\texttt{time}_{it})$ by $\ln(\texttt{time}_{it})$, without any regression coefficient in front. This causes the expected number of reported MTPL claims, $\mu_{it} = \mathbb{E}(N_{it}) = \exp(\eta_{it})$, to be proportional to the duration of the policy period. In the offset versions of the meter-hybrid and telematics model, the flexible term related to `distance` gets replaced by an offset $\ln(\texttt{distance}_{it})$, imposing the risk to be proportional to the distance.

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

Table 3: Variables contained in the best Poisson model for each of the predictor sets using the projection approach for structural zeros. The second column of each predictor set refers to the model with an offset for either time or meter. The best NB models were identical to the best Poisson models.

The models which are allowed to use the policyholder information prefer the use of `experience`, measured as the years since obtaining the driver's license, instead of `age` to segment the risk in young drivers. `Gender` is only selected as an important covariate in the classic models, not in any of the hybrid models, indicating that the telematics information renders the use of gender as a rating variable redundant. In the offset variants of both hybrid models the `fuel` term is dropped. The newly introduced telematics predictors `road type` and `time slot` are selected in both the hybrid and the telematics models. The `week/weekend` term is only selected in the offset variant of the telematics model.

The second best models, with only a slightly higher AIC value, show that adding `kwatt` to the classic model gives a comparable model fit and the same holds for adding `week/weekend` to the hybrid and telematics models. Furthermore, `fuel` can easily be left out of the hybrid models without deteriorating the fit.

Using the conditioning approach for structural zeros as opposed to the projection approach, the main difference in the selection of variables is that the compositional predictor `week/weekend` is always included in the hybrid and telematics models. Both the 1111 and the 1110 zero patterns of `road type` are selected. Stepwise adding more zero patterns of `road type` to the predictor did not improve AIC, and similarly for the division by `time slot` and `week/weekend`. The variables selected were again identical under the Poisson and negative binomial model specification.

For each of these best model formulations, we added a policyholder-specific random effect in the predictor (1) to account for possible dependence from observing policyholders over multiple policy periods. However, none of the added random effects were deemed necessary at the 5% significance level using the approximate test of Wood (2013).

5.2 Model assessment

Table 4 reports AIC and all 6 proper scoring rules obtained using 10-fold cross validation for each predictor set under the Poisson model specification using the projection approach for structural zeros. These performance tools unanimously indicate that the time-hybrid model without offset scores best. The meter-hybrid model is a close second. Their respective versions with an offset and the telematics model without offset conclude the top five according to all criteria except for the Dawid–Sebastiani score. This demonstrates the significant impact of telematics constructed variables on the predictive power of the model. In addition, the telematics model without offset outperforms the classic models across all assessment criteria. Hence, using only telematics predictors is considered to be better than the use of the traditional rating variables.

Across all predictor sets, the use of an offset for the exposure-to-risk, either time or meter, is too restrictive for these data. From a statistical point of view, the time or meter rating unit cannot be considered to be directly proportional to the risk. However, from a business point of view, it is convenient to consider a proportional approach due to its simplicity and explainability.

Similar results are obtained under the negative binomial model specification and using the conditioning approach for structural zeros. The rankings of the predictor sets according to AIC are the same as in Table 4 under the negative binomial model specification and/or using the conditioning approach. The AIC values for each predictor set under the NB model specification compared to their Poisson counterpart were slightly higher for the classic and hybrid models and slightly lower for the telematics models indicating that only the telematics predictor sets benefit from the additional parameter to capture overdispersion. The model assessment using proper scoring rules led to the same conclusions as before.

Beside an exhaustive search among additive terms, we have explored the use of interactions among categorical, among continuous, between categorical and continuous, and between categorical and compositional predictors. Slight marginal improvements in AIC could only be achieved in the classic model by further refining the effects of `experience`, `age vehicle` and `material` by `gender` without changing the rankings of the best models in Table 4.

Predictor set	Offset	EDF	AIC value, rank		logs ×10 value, rank		qs ×10 value, rank		sphs ×10 value, rank		rps $\times 10^2$ value, rank		dss value, rank		ses $\times 10^2$ value, rank	
Classic	no	32.15	11 896	6	1.790	6	−9.1858	6	−9.5822	6	4.224	6	−2.206	5	4.535	6
	yes	27.27	11 995	8	1.804	8	−9.1838	8	−9.5816	8	4.234	8	−2.129	6	4.547	8
Time-hybrid	no	34.11	11 734	1	1.766	1	−9.1909	1	−9.5837	1	4.196	1	−2.266	1	4.502	1
	yes	30.41	11 811	3	1.777	3	−9.1891	3	−9.5831	3	4.205	3	−2.212	4	4.512	3
Meter-hybrid	no	34.32	11 743	2	1.767	2	−9.1907	2	−9.5836	2	4.197	2	−2.259	2	4.503	2
	yes	30.37	11 866	5	1.785	5	−9.1884	4	−9.5829	4	4.209	4	−2.007	7	4.517	4
Telematics	no	15.05	11 862	4	1.784	4	−9.1871	5	−9.5826	5	4.216	5	−2.226	3	4.526	5
	yes	11.43	11 989	7	1.803	7	−9.1850	7	−9.5820	7	4.228	7	−1.965	8	4.538	7

Table 4: Model assessment of the best models according to AIC for each of the four predictor sets under the Poisson model specification and using the projection approach for structural zeros. The second row of each predictor set refers to the model with an offset for either time or meter. For each model we list the effective degrees of freedom (EDF), Akaike information criterion (AIC) and 6 cross-validated proper scoring rules: logarithmic (logs), quadratic (qs), spherical (sphs), ranked probability (rps), Dawid-Sebastiani (dss), and squared error scores (ses). For AIC and the proper scoring rules, the first column represents the value and the second column the rank.

5.3 Visualization and discussion

The effects of each predictor variable in the best time-hybrid model without an offset are graphically displayed in Figure 5 for the policy variables and in Figure 6 for the telematics variables. By exponentially transforming the additive effects, we show the multiplicative effects on the expected number of claims for each categorical parametric, continuous smooth or geographical term in the fitted model. For the categorical predictors we quantify the uncertainty of those estimates by constructing individual 95% confidence intervals based on the large sample normality of the model parameter estimators. Bayesian 95% confidence pointwise intervals are used for the smooth components of the GAM and include the uncertainty about the intercept (Marra and Wood, 2012). For the compositional data predictors, we visualize the clr transform of the corresponding model parameters with 95% confidence intervals along with a reference line at zero (see Section 4.2.2). In the supplementary material, similar graphs for the other three predictor sets, see Figure 2c, are shown in Appendix C and the relative importance of these predictors is quantified and visualized in Appendix D. In the remainder of this section, we discuss the insights and interpretations for both the policy and telematics variables in each of these models.

Policy variables The rating unit `policy period` in the classic and time-hybrid models always has a monotone increasing estimated effect. The longer a policyholder is insured, the higher the premium amount, ceteris paribus. Using the fact that the level of the nonlinear smooth component is not uniquely identifiable (see Section 4.1), we vertically translated the estimated smooth term to pass the point (365, 0) on the predictor scale (and hence (365, 1) on the response scale) for ease of interpretation.

The smooth effect of `experience` embodies the higher risk posed by younger, less experienced drivers. The increased risk is more outspoken in the first two years for the hybrid models as compared to the classic model.

In the classic model, the significant effect of `gender` indicates that women are 16% less risky drivers than men. However, when telematics predictors are taken into account in the hybrid models, the categorical variable `gender` is no longer selected as predictor. Neither did any interaction term between gender and a categorical, a continuous or a compositional predictor improve AIC. The perceived difference between women and men can hence be explained through differences in driving habits. In particular, female drivers in the portfolio drive significantly fewer kilometers on a yearly

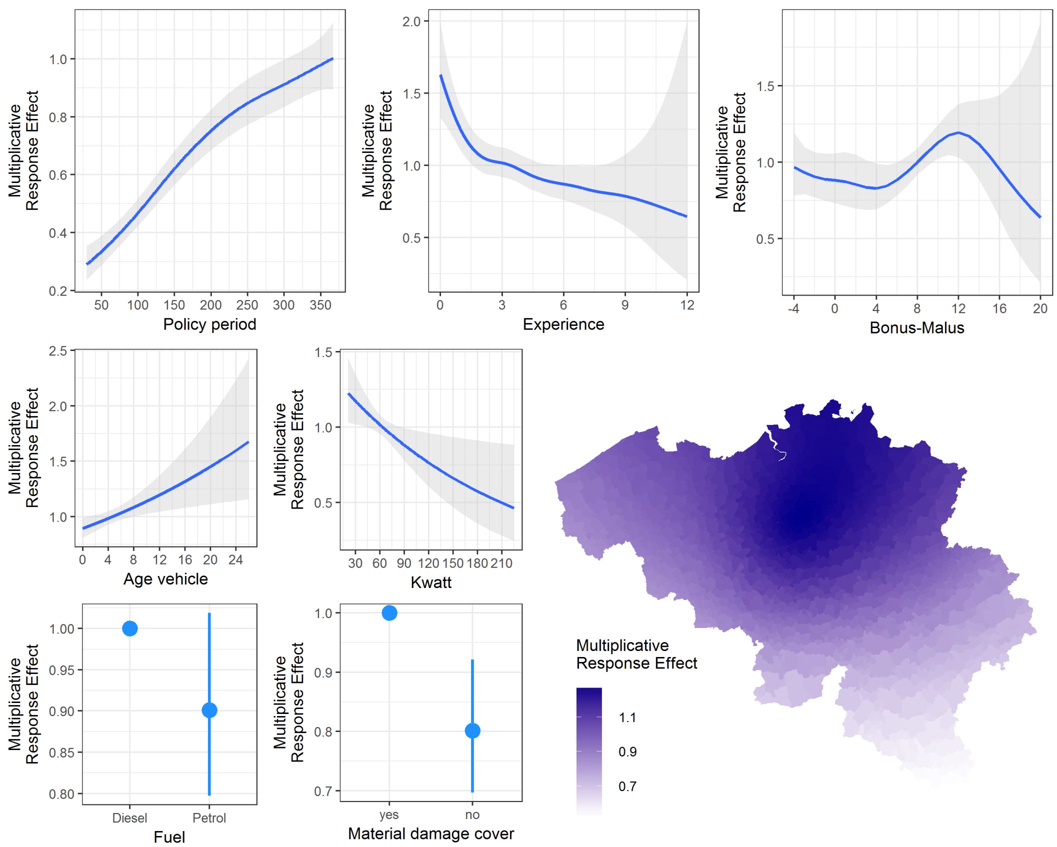

Figure 5: Multiplicative response effects of the policy predictor variables of the time-hybrid model.

basis compared to men (15 409 vs 18 570 on average, with a p-value smaller than 0.001 using a two sample t-test). Similar findings are reported in Ayuso et al. (2016a,b). In light of the EU rules on gender-neutral pricing in insurance, this shows how moving towards car insurance rating based on individual driving habits and style can resolve possible discrimination of basing the premium on proxies such as gender.

The smooth effects of `bonus-malus` in the classic and hybrid models are nonlinear and somewhat counterintuitive. Given the lack of a lengthy claim history of the young drivers of this portfolio, the BM levels of the insureds are not yet fully developed and stabilized. The majority of the drivers has a bonus-malus (BM) level between 4 and 12 for which the effect on the claim frequency is increasing. For the highest BM levels however, the effect is declining, albeit with a high uncertainty due to a lack of observations in this region. Furthermore, the effect does not decrease for the lowest BM levels. This can be explained by an improper use of the BM scale as marketing tool to attract new customers. By lowering the initial value of the BM scale, the insurer can reduce the premium a potential new policyholder has to pay.

When it comes to characteristics of the car, insureds driving older vehicles have an estimated higher risk of accidents. The smooth effect of `age vehicle` is estimated as a straight line on the predictor scale in the classic and hybrid models. The effect of `kwatt` in the hybrid models also reduced to a straight line on the predictor scale. When the insured vehicle has more horsepower, the estimated expected claims number is lower. The categorical predictor `fuel` shows that vehicles using petrol have an estimated lower risk for accidents compared to diesel. This difference is however smaller and no longer statistically significant in the hybrid models compared to the classic model, where it serves as a proxy variable for the distance driven. Indeed, vehicles using diesel as opposed to petrol drive significantly more kilometers in our portfolio (18 940 vs 13 267 on average on a yearly basis).

In both the classic and hybrid models, the policies without `material damage cover` have a 20% lower estimated expected number of claims. This may be explained by the reluctance of some insureds without additional material damage coverage to report small accidents. Due to bonus-malus mechanisms being independent of the claim amount, filing a claim leads to premium surcharges which may be more disadvantageous for policyholders than for them to defray the third party. This phenomenon is known as the hunger for bonus (Denuit et al., 2007). Insureds with an additional material damage cover are less inclined to do so since their own, first party costs are also covered making it more worthwhile to report a claim at fault. Including telematics variables in the model does not affect this discrepancy.

The geographical effect (`postal code`), plotted on top of a map of Belgium for the classic and hybrid models, captures the remaining spatial heterogeneity based on the postal code where the policyholder resides. For the classic model, the graph shows higher claim frequencies for urban areas like Brussels in the middle, Antwerp in the north and Liège in the east and lower claim frequencies in the more sparsely populated regions in the south. The geographic variation however decreases strongly in the hybrid models due to the inclusion of telematics predictors not taken into account in the classic model. The EDF corresponding to the spatial smooth reduced from 15.8 in the classic model to 4.1 and 4.4 in the time and meter hybrid model, respectively. This is satisfactory as it means, instead of overrelying on geographical proxies, the hybrid models are basing the insurance premium on actual differences in driving habits which is more closely related to the accident risk.

Telematics variables In the meter-hybrid and telematics models, `distance` is used as the rating unit. Similar to the time effect in the classic and time-hybrid model, the effect of the risk exposure is estimated as a monotone increasing function. The accident risk however does not vanish for insureds who hardly drive any kilometers during the observation period.

The `yearly distance` is used in the time-hybrid model, which uses time as exposure, to differentiate between drivers who travel many versus few kilometers on a yearly basis. In this way, the driven distance is rescaled on a yearly basis (see Section 3.2) and used as an additional risk factor having a weaker effect on the claim frequency compared to the meter-hybrid and telematics models where distance is used a rating unit. In both hybrid models and the telematics model, the estimated `average distance` effect shows lower claim frequencies for insureds who on average drive long distances.

Our modeling approach using compositional predictors separates on the one hand the effect of an overall increase in the distance driven and on the other hand the effect of a change in the division of the distance driven into different categories. This allows us to qualitatively and quantitatively interpret and visualize the impact of individual driving habits on the expected claim frequencies.

The clr transforms of the model coefficients related to the compositional `road type` predictor in the telematics model show how insureds who drive relatively more on urban roads have higher claim frequencies and insureds who drive relatively more on the 'other' road type have lower claim frequencies. In the hybrid models, these effects are headed in the same direction with the exception that motorways is perceived as riskier. The elevated accident risk for insureds driving more on urban roads is in line with Paefgen et al. (2014), where the driven distance is divided over 'highway', 'urban' and 'extra-urban' road types. The authors however neglect the compositional nature of this predictor in the analysis and do not incorporate any of the classical policy risk factors in the logistic regression model. In Ayuso et al. (2014), the percentage of urban driving is considered an important variable to predict either the time or the distance to the first accident, although percentages driven on different road types are not considered. Using either a quadratic effect or a categorical effect

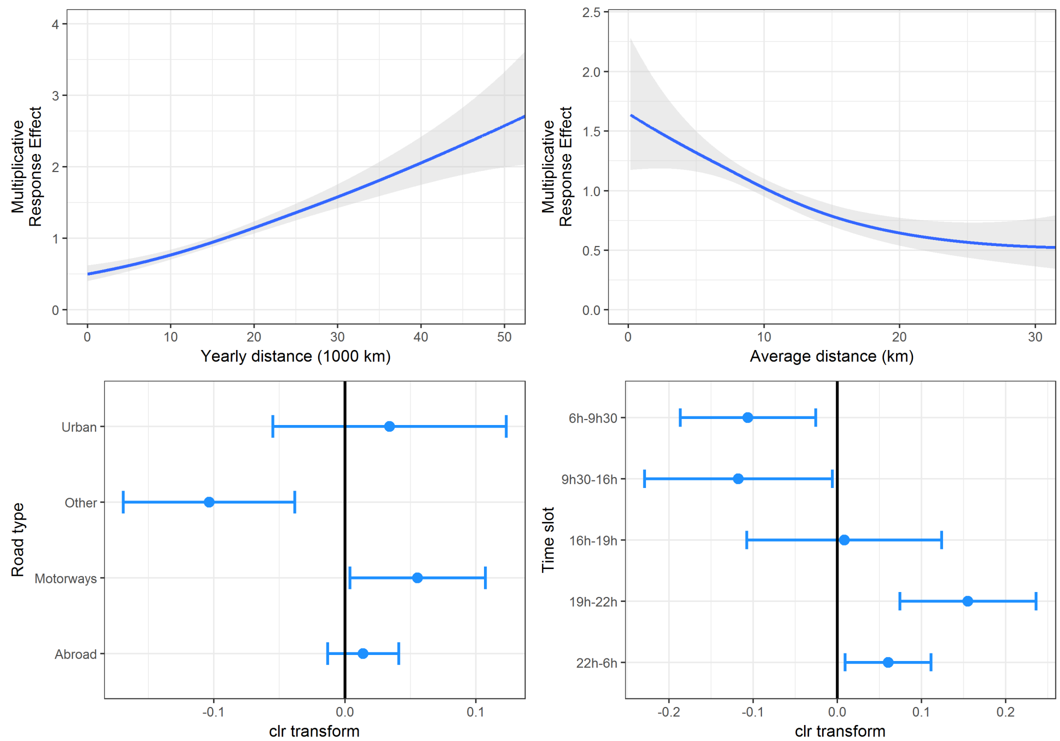

Figure 6: Multiplicative response effects of the telematics predictor variables of the time-hybrid model.

(urban driving $> 25\%$) in Weibull regression models shows how increased percentages of urban driving reduce both the expected time or distance to the first accident.

The estimated model coefficients $\boldsymbol{\beta}^{\text{road}}$ of the compositional `road type` predictor and the corresponding simplicial gradient $\boldsymbol{b}^{\text{road}}$ and clr transform $\text{clr}(\boldsymbol{b}^{\text{road}})$ are presented in Figure 7. Based on the latter we can quantitatively interpret the effect of `road type` in the time hybrid model. For instance, a relative ratio increase of 50% to the 'other' road type component, with constant ratios of the remaining components, results in a multiplicative decrease of $1.50^{-0.104} = 0.959$ for the expected number of claims. Applied to the compositional road type data vector $\boldsymbol{x} = (0.442, 0.282, 0.252, 0.024)$ of Figure 7, this relative ratio increase to the 'other' road type component would change the compositional vector to $\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \mathcal{C}(1, 1.50, 1, 1)^t = (0.387, 0.371, 0.221, 0.021)^t$. Based on our projection approach for structural zeros, the interpretation is similar for a relative ratio change to a nonzero component of observations with a certain zero pattern. In particular, for someone who did not drive abroad we base the interpretation of the effect on the clr transform of the related subcomposition, i.e. $\text{clr}(\boldsymbol{b}^{\text{road}}_{1110}) = (0.038, -0.099, 0.061)^t$, obtained by recentering $\text{clr}(\boldsymbol{b}^{\text{road}})$ without the abroad component around zero.

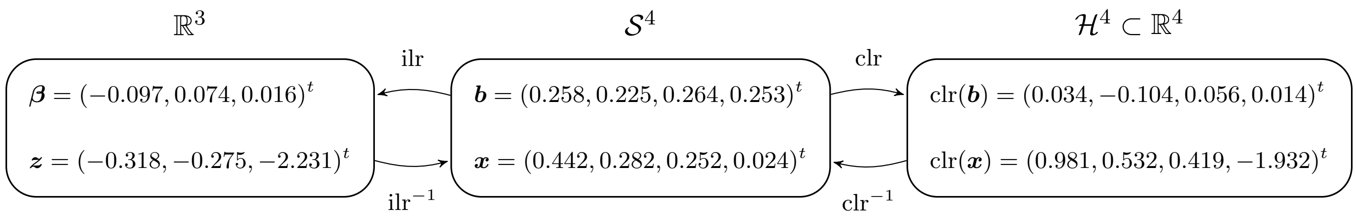

Figure 7: Representations in the ilr, simplex and clr space of the estimated regression parameters in the time hybrid model with respect to the compositional predictor term `road type` (with components *urban, other, motorways* and *abroad*) and the average compositional data vector without structural zeros.

The compositional `time slot` predictor in the hybrid and telematics models indicates that policyholders who drive relatively more in the morning have lower claim frequencies and policyholders who drive relatively more in the evening and during the night have higher claim frequencies. For instance, the multiplicative response effect of a relative ratio increase of 50% in the evening component (between 19h and 22h) is equal to $1.50^{0.155} = 1.065$. In Paefgen et al. (2014), the accident risk is considered to be lower during the daytime (between 5h and 18h) compared to the evening (between 18h and 21h), based on the estimated coefficients of linear model terms of the log transformed percentages of the driven distance in these time slots. Ayuso et al. (2014) reports how a higher percentage of driving at night reduces the expected time to a first accident, where the effect is modeled linearly, with no further distinction in time slots.

6 Conclusion

Telematics insurance offers new opportunities for insurers to differentiate drivers based on their driving habits and style. By aggregating the telematics data on the level of the policy period by policyholder and combining it with traditional policy(holder) rating variables, we construct predictive models for the frequency of MTPL claims at fault. Generalized additive models with a Poisson or negative binomial response are used to model the effects of predictors in a smooth, yet interpretive way. The divisions of the driven distance into 4 road types and 5 time slots forms a challenge from a methodological point of view that has not been addressed in the literature. We demonstrate how to include this information as compositional predictors in the regression and formulate a new way of how to interpret their effect on the average claim frequency.

Our research reveals the significant impact of the use of telematics data through an exhaustive model selection and an assessment of the predictive performance. The time-hybrid is the best model according to AIC and all proper scoring rules, closely followed by the meter-hybrid model. The model using only telematics variables is ranked higher than the best classic model using only traditional policy information.

The compositional predictors show that a further classification of the driven distance based on the location and the time is relevant. Our contribution indicates that driving more on urban roads or motorways and in the evening or at night contributes to a riskier driving pattern. The best hybrid models highlight that certain popular pricing factors (gender, fuel, postal code) are indeed proxies for the driving habits and part of their predictive power is taken over by the distance driven and the splits into different categories. Hence, we demonstrate using careful statistical modeling how the use of telematics variables is an answer to the European regulation on insurance pricing practices that bans the use of gender as a rating factor.

In the case of multiple insured drivers, it is unclear which characteristics (such as age, experience and gender) the insurer must use to determine the premium. We proceed, in consultation with the Belgian insurer providing the data, by identifying the driver with the lowest experience as the main driver and use his policyholder information as predictors in the regression for tarification purposes. In practice, when a parent adds a child as a driver in the policy, a premium surcharge is often avoided to prevent the policyholder from lapsing. By shifting towards pricing based on telematics information as we do in this research, this tarification issue becomes less of a problem because the premium will be usage-based.

Pricing using telematics data can be seen as falling in between *a priori* and *a posteriori* pricing. The driving habits and style are no traditional a priori variables since they cannot be determined before the policyholder starts to drive. Insurers now reason that available UBI products are only purchased by drivers who consider themselves to be either safe or low-kilometer drivers. This potential form of positive selection, which could not be quantified based on the studied portfolio alone, validates an upfront discount on the traditional insurance premium. Based on the telematics data collected over time, insurers can set up a discount structure to adapt the premium in an a posteriori way. The discount structure can depend on the actual driven distance, with a further personalized differentiation based on the riskiness of the profile as perceived from the driving habits of the insured. The insights provided in this paper reveal which elements can be adopted in such a structure, for instance, by making kilometers driven on urban roads or in the evening or at night more expensive.

In conclusion, telematics technology provides means to insurers to better align premiums with risk. Pay-as-you-drive insurance is a first step in which the number of driven kilometers, the type of road and the time of day are combined with the traditional self-reported information such as policyholder and car characteristics to calculate insurance premiums. A next step is pay-how-you-drive insurance, where on top of these driving habits also the driving style is considered to assess how risky someone drives by monitoring for instance speed infringements, harsh braking, excessive acceleration, and cornering style. The ideas and statistical framework presented can be extended to incorporate such additional pay-how-you-drive predictors if they are available.

Acknowledgements

The authors are grateful to the associate editor and the referees for the valuable comments and suggestions. Support is acknowledged from the agency for Innovation by Science and Technology (IWT 131173) and the Flemish Supercomputer Centre VSC. Furthermore, Katrien Antonio acknowledges financial support from the Ageas Continental Europe Research Chair at KU Leuven and from KU Leuven's research council (project COMPACT C24/15/001). Gerda Claeskens acknowledges support through the IAP Research Network P6/03 of the Belgian State (Belgian Research Policy) and the KU Leuven grant (GOA/12/14). The authors also thank their contact person at the insurance company for the smooth cooperation.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall London.

Aitchison, J. and Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In Thió-Henestrosa, S. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*. University of Girona.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, 73:125–131.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016a). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016b). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68:160 – 167.

Bacon Shone, J. (2003). Modelling structural zeros in compositional data. In Thió-Henestrosa, S. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*. University of Girona.

Barceló-Vidal, C., Martín-Fernández, J. A., and Mateu-Figueras, G. (2011). Compositional differential calculus on the simplex. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Bordoff, J. E. and Noel, P. J. (2008). Pay-as-you-drive auto insurance: A simple way to reduce driving-related harms and increase equity. The Brookings Institution. Discussion Paper.

Boucher, J.-P., Pérez-Marín, A. M., and Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles, 3ª época*, 19:135–154.

Butler, P. (1993). Cost-based pricing of individual automobile risk transfer: Car-mile exposure unit analysis. *Journal of Actuarial Practice*, 1(1):51–84.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

de Jong, P. and Heller, G. (2008). *Generalized linear models for insurance data*. Cambridge.

Denuit, M. and Lang, S. (2004). Non-life ratemaking with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627–647.

Denuit, M., Marechal, X., Pitrebois, S., and Walhin, J. (2007). *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems*. Wiley.

Desyllas, P. and Sako, M. (2013). Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. *Research Policy*, 42(1):101–116.

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L., and Mateu-Figueras, G. (2011). Elements of simplicial linear algebra and geometry. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Ferreira, J. and Minikel, E. (2010). Pay-As-You-Drive Auto Insurance In Massachusetts: A Risk Assessment And Report On Consumer. http://mit.edu/jf/www/payd/PAYD_CLF_Study_Nov2010.pdf.

Filipova-Neumann, L. and Welzel, P. (2010). Reducing asymmetric information in insurance markets: Cars with black boxes. *Telematics and Informatics*, 27(4):394–403.

Frees, E. W. (2014). Frequency and severity models. In Frees, E. W., Derrig, R. A., and Meyers, G., editors, *Predictive modeling applications in actuarial science*, volume 1. Cambridge University Press.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall.

Greenberg, A. (2009). Designing pay-per-mile auto insurance regulatory incentives. *Transportation Research Part D: Transport and Environment*, 14(6):437–445.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, second edition.

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

Klein, N., Denuit, M., Lang, S., and Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225 – 249.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480):1328–1337.

Lancaster, P. and Salkauskas, K. (1986). *Curve and surface fitting: An introduction*. London: Academic Press.

Lemaire, J., Park, S. C., and Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin*, 46:39–69.

Litman, T. (2011). Distance-based vehicle insurance feasibility, costs and benefits. Victoria Transport Policy Institute. http://www.vtpi.org/dbvi_com.pdf.

Litman, T. (2015). Pay-As-You-Drive Vehicle Insurance: Converting Vehicle Insurance Premiums Into Use-Based Charges. Victoria Transport Policy Institute. http://www.vtpi.org/tdm/tdm79.htm.

Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74.

Martín-Fernández, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing with zeros. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*, pages 43–58. John Wiley & Sons.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The principle of working on coordinates. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall, New York, second edition.

Paefgen, J., Staake, T., and Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61:27 – 40.

Parry, I. W. H. (2005). Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *American Economic Review*, 95(2):288–293.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiss, P. T. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):505–523.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Toledo, T., Musicant, O., and Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies*, 16(3):320 – 331.

Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2016). Innovative insurance schemes: Pay as/how you drive. *Transportation Research Procedia*, 14:362 – 371.

van den Boogaart, K., Tolosana-Delgado, R., and Bren, M. (2006). Concepts for handling zeroes and missing values in compositional data. In Pirard, E., editor, *Proceedings of IAMG'06, The XI annual conference of the International Association for Mathematical Geology*. University of Liège.

van den Boogaart, K. G., Tolosana, R., and Bren, M. (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1.

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

Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM Journal on Scientific and Statistical Computing*, 2(1):5–16.

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

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, 100(4):1005–1010.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

A Structural zero patterns of the compositional predictors

We give an overview of the structural zero patterns for the division of the number of meters in road types (Table A.1), time slots (Table A.2) and week/weekend (Table A.3). The pattern is represented in the first column by a code indicating which components are zero (0) or non-zero (1). For each structural zero pattern, we tabulate their absolute and relative frequency and the compositional mean of the nonzero components, which for n observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$ is defined as

$$\overline{\boldsymbol{x}} = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \boldsymbol{x}_i = \mathcal{C}\left(\left(\prod_{i=1}^{n} x_{i1}\right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD}\right)^{1/n}\right)^t \tag{11}$$

resulting in the closed componentwise geometric mean. Following the principle of working on coordinates, we can alternatively write the compositional mean as

$$\overline{\boldsymbol{x}} = \mathrm{ilr}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{ilr}(\boldsymbol{x}_i)\right),$$

where we first transform the compositional data from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ using the ilr transformation, then compute the mean in $\mathbb{R}^{D-1}$ and finally apply the inverse ilr transformation to obtain the compositional mean in $\mathcal{S}^D$.

Road type	Number	Percent	Urban	Other	Motorways	Abroad
1111	18821	0.5659	0.4421	0.2822	0.2516	0.0241
1110	13540	0.4071	0.5079	0.2782	0.2139	–
1100	481	0.0145	0.5923	0.4077	–	–
1101	258	0.0078	0.4960	0.4648	–	0.0392
0001	131	0.0039	–	–	–	1
1010	7	0.0002	0.9075	–	0.0925	–
1001	7	0.0002	0.0034	–	–	0.9966
1000	6	0.0002	1	–	–	–
0101	5	0.0001	–	0.0002	–	0.9998
0111	3	0.0001	–	0.0130	0.0833	0.9038

Table A.1: Structural zero patterns for the division of meters into road types, along with their absolute and relative frequency and the compositional mean of the subcompositions of nonzero components.

Time slot	Number	Percent	6h-9h30	9h30-16h	16h-19h	19h-22h	22h-6h
11111	31886	0.9587	0.1472	0.4699	0.2159	0.1010	0.0661
11110	991	0.0298	0.2000	0.5090	0.2323	0.0587	–
11101	130	0.0039	0.2060	0.5953	0.1296	–	0.0691
11100	110	0.0033	0.2134	0.6238	0.1628	–	–
01111	47	0.0014	–	0.5398	0.1983	0.1339	0.1280
01110	23	0.0007	–	0.5850	0.2793	0.1357	–
01100	22	0.0007	–	0.7912	0.2088	–	–
11000	16	0.0005	0.1459	0.8541	–	–	–
11001	10	0.0003	0.0697	0.8000	–	–	0.1304
01000	7	0.0002	–	1	–	–	–
01001	3	0.0001	–	0.6803	–	–	0.3197
01010	2	0.0001	–	0.3054	–	0.6946	–
10000	2	0.0001	1	–	–	–	–
01101	2	0.0001	–	0.6698	0.1744	–	0.1558
10001	2	0.0001	0.1271	–	–	–	0.8729
11011	2	0.0001	0.0653	0.5536	–	0.2762	0.1049
00100	1	0.0000	–	–	1	–	–
00110	1	0.0000	–	–	0.8200	0.1800	–
10010	1	0.0000	0.9787	–	–	0.0213	–
10110	1	0.0000	0.2451	–	0.2935	0.4614	–

Table A.2: Structural zero patterns for the division of meters into time slots, along with their absolute and relative frequency and the compositional mean of the subcompositions of nonzero components.

Week/weekend	Number	Percent	Week	Weekend
11	33186	0.9978	0.7490	0.2510
10	72	0.0022	1	–
01	1	0.0000	–	1

Table A.3: Structural zero patterns for the division of meters into week and weekend, along with their absolute and relative frequency and the compositional mean of the subcompositions of nonzero components.

Road type	Number	Percent	Urban	Other	Motorways	Abroad
1111	18821	0.5659	0.4421	0.2822	0.2516	0.0241
1110	13540	0.4071	0.5079	0.2782	0.2139	–
0	898	0.0270	–	–	–	–

Time slot	Number	Percent	6h-9h30	9h30-16h	16h-19h	19h-22h	22h-6h
11111	31886	0.9587	0.1472	0.4699	0.2159	0.1010	0.0661
0	1373	0.0413	–	–	–	–	–

Week/weekend	Number	Percent	Week	Weekend
11	33186	0.9978	0.7490	0.2510
0	73	0.0022	–	–

Table A.4: Structural zero patterns for the division of the number of meters into road types, time slots and week/weekend as recognized in the models following the conditioning approach.

B Functional forms of the selected best models

We write down the functional forms of the predictors of the selected best models under the Poisson model specification and using the projection approach for structural zeros of compositional predictors. In the preferred classic model the predictor can be written as

$$\begin{aligned}\eta_{it}^{\text{classic}} &= \beta_0 + \beta_1 \texttt{gender}_{it} + \beta_2 \texttt{material}_{it} + \beta_3 \texttt{fuel}_{it} + f_1(\texttt{time}_{it}) + f_2(\texttt{experience}_{it}) \\ &\quad + f_3(\texttt{bonus-malus}_{it}) + f_4(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it})\,.\end{aligned}$$

The predictor in the best time-hybrid model is

$$\begin{aligned}\eta_{it}^{\text{time-hybrid}} &= \beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{time})_{it} + f_2(\texttt{experience}_{it}) + f_3(\texttt{bonus-malus}_{it}) \\ &\quad + f_4(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) + f_5(\texttt{yearly distance}_{it}) + f_6(\texttt{average distance}_{it}) \\ &\quad + \sum_{M \in \mathcal{M}^{\text{road}}} d_M(\boldsymbol{x}_{it}^{\text{road}}) \langle \boldsymbol{b}_{M^c}^{\text{road}}, \boldsymbol{x}_{it,M^c}^{\text{road}} \rangle_a + \sum_{M \in \mathcal{M}^{\text{time}}} d_M(\boldsymbol{x}_{it}^{\text{time}}) \langle \boldsymbol{b}_{M^c}^{\text{time}}, \boldsymbol{x}_{it,M^c}^{\text{time}} \rangle_a\,,\end{aligned}$$

and for the preferred meter-hybrid model we have

$$\begin{aligned}\eta_{it}^{\text{meter-hybrid}} &= \beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{experience}_{it}) + f_2(\texttt{bonus-malus}_{it}) \\ &\quad + f_3(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) + f_4(\texttt{distance}_{it}) + f_5(\texttt{average distance}_{it}) \\ &\quad + \sum_{M \in \mathcal{M}^{\text{road}}} d_M(\boldsymbol{x}_{it}^{\text{road}}) \langle \boldsymbol{b}_{M^c}^{\text{road}}, \boldsymbol{x}_{it,M^c}^{\text{road}} \rangle_a + \sum_{M \in \mathcal{M}^{\text{time}}} d_M(\boldsymbol{x}_{it}^{\text{time}}) \langle \boldsymbol{b}_{M^c}^{\text{time}}, \boldsymbol{x}_{it,M^c}^{\text{time}} \rangle_a\,.\end{aligned}$$

Finally, the predictor in the best telematics model is

$$\begin{aligned}\eta_{it}^{\text{telematics}} &= \beta_0 + f_1(\texttt{distance})_{it} + f_2(\texttt{average distance}_{it}) \\ &\quad + \sum_{M \in \mathcal{M}^{\text{road}}} d_M(\boldsymbol{x}_{it}^{\text{road}}) \langle \boldsymbol{b}_{M^c}^{\text{road}}, \boldsymbol{x}_{it,M^c}^{\text{road}} \rangle_a + \sum_{M \in \mathcal{M}^{\text{time}}} d_M(\boldsymbol{x}_{it}^{\text{time}}) \langle \boldsymbol{b}_{M^c}^{\text{time}}, \boldsymbol{x}_{it,M^c}^{\text{time}} \rangle_a\,.\end{aligned}$$

C Graphical displays of the multiplicative response effects

We visualize the effects of each predictor variable of the classic model in Figure C.1, of the telematics model in Figure C.2 and of the meter-hybrid model in Figures C.3 for the policy variables and C.4 for the telematics variables. These graphs correspond to the selected best models without offset under the Poisson model specification and using the projection approach for structural zeros.

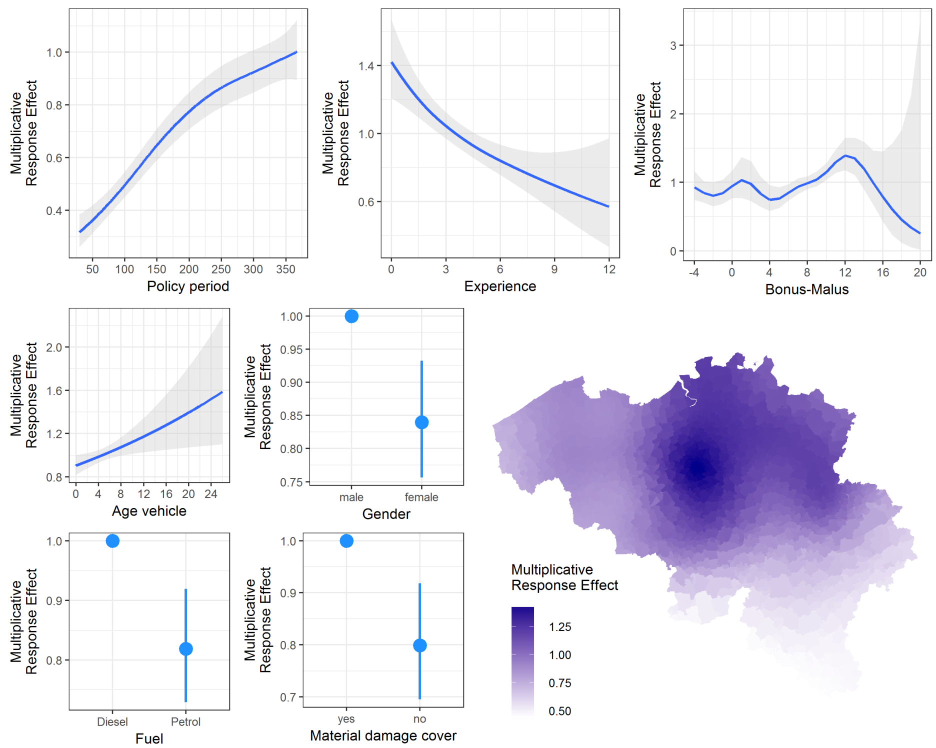

Figure C.1: Multiplicative response effects of the predictor variables of the classic model.

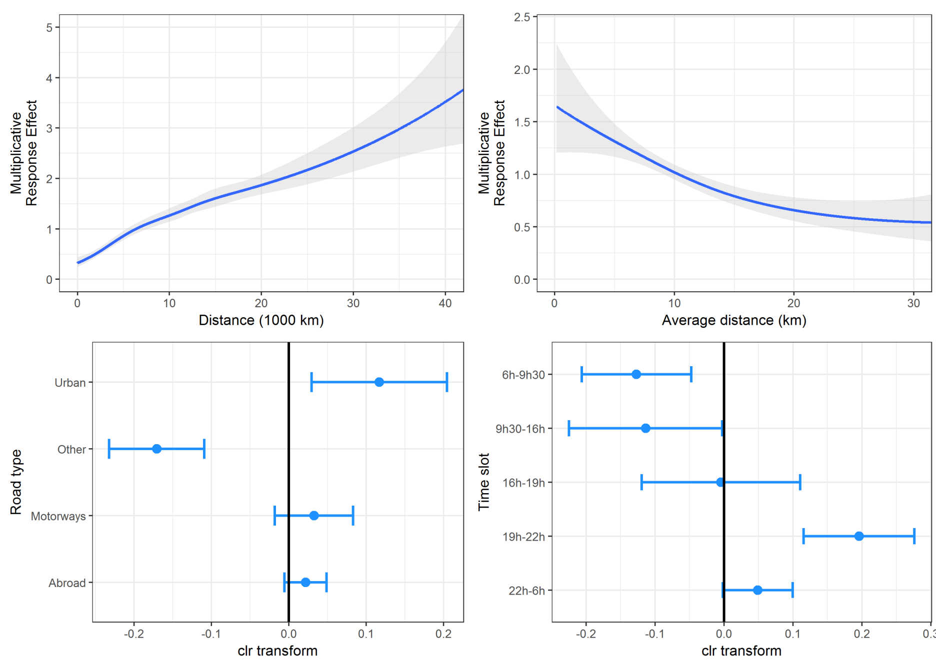

Figure C.2: Multiplicative response effects of the predictor variables of the telematics model.

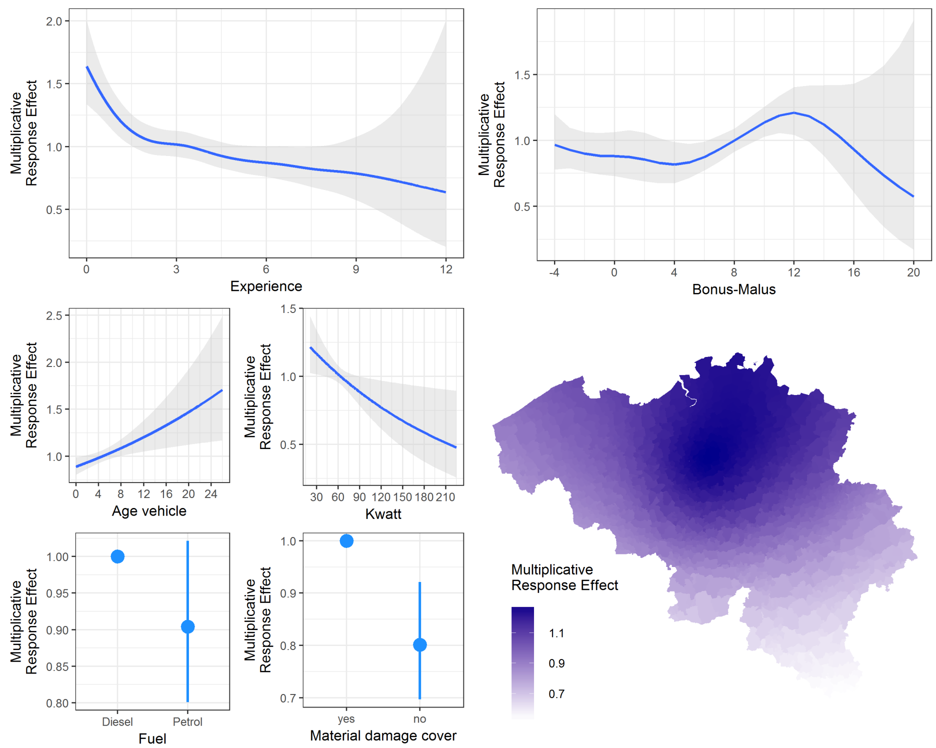

Figure C.3: Multiplicative response effects of the policy predictor variables of the meter-hybrid model.

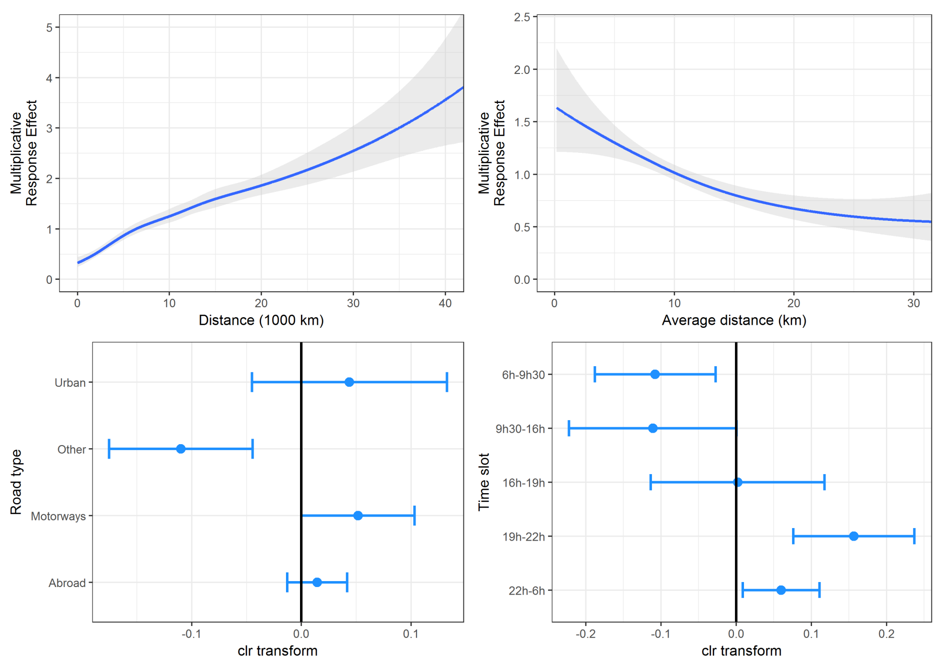

Figure C.4: Multiplicative response effects of the telematics predictor variables of the meter-hybrid model.

D Relative importance of the predictors

To assess the relative importance of the predictor variables in the model, we construct histograms of their multiplicative response effects for each observation in the data set. This is done for the selected best models under the Poisson model specification and using the projection approach for structural zeros. The graphs are shown, in the variants without an offset for time or meter, for the classic model in Figure D.5, for the telematics model in Figure D.6, for the time-hybrid model in Figures D.7 and D.8 and for the meter-hybrid model in Figures D.9 and D.10. For the hybrid models, we construct separate graphs for the predictor variables derived from the policy and telematics information. For categorical predictors the histogram reduces to a bar plot of the categorical effects and for the continuous and geographical predictors to a histogram of the exponentiated smooth effects. For a compositional predictor, such as time slot, we plot a histogram of the exponential of the related term $\sum_{M \in \mathcal{M}^{\text{time}}} d_M(\boldsymbol{x}_{it}^{\text{time}}) \langle \boldsymbol{b}_{M^c}^{\text{time}}, \boldsymbol{x}_{it,M^c}^{\text{time}} \rangle_a$. To rank the influence of the different policy and telematics variables on the claim frequency, we use the standard deviations over all observations of the effects on the predictor scale, see Table D.5. Under the offset restriction, the logarithm of time or meter is used as an explanatory variable in the predictor (without any regression coefficient in front) and we report its standard deviation.

	Predictor	Classic		Time-hybrid		Meter-hybrid		Telematics	
Policy	Time	0.36	0.69	0.39	0.69				
	Age								
	Experience	0.18	0.14	0.16	0.11	0.16	0.12		
	Gender	0.09	0.09						
	Material	0.11	0.11	0.11	0.10	0.11	0.11		
	Postal code	0.21	0.20	0.15	0.14	0.15	0.16		
	Bonus-malus	0.16	0.18	0.11	0.15	0.12	0.15		
	Age vehicle	0.08	0.10	0.09	0.10	0.09	0.12		
	Kwatt			0.07	0.06	0.07	0.08		
	Fuel	0.09	0.09	0.05		0.05			
Telematics	Distance					0.48	0.95	0.49	0.95
	Yearly distance			0.32	0.36				
	Average distance			0.22	0.24	0.20	0.31	0.22	0.34
	Road type			0.12	0.13	0.12	0.15	0.16	0.20
	Time slot			0.20	0.20	0.20	0.17	0.22	0.21
	Week/weekend								0.07

Table D.5: Standard deviations of the effects on the predictor scale in the best Poisson model for each of the predictor sets using the projection approach for structural zeros. The second column of each predictor set refers to the model with an offset for either time or meter.

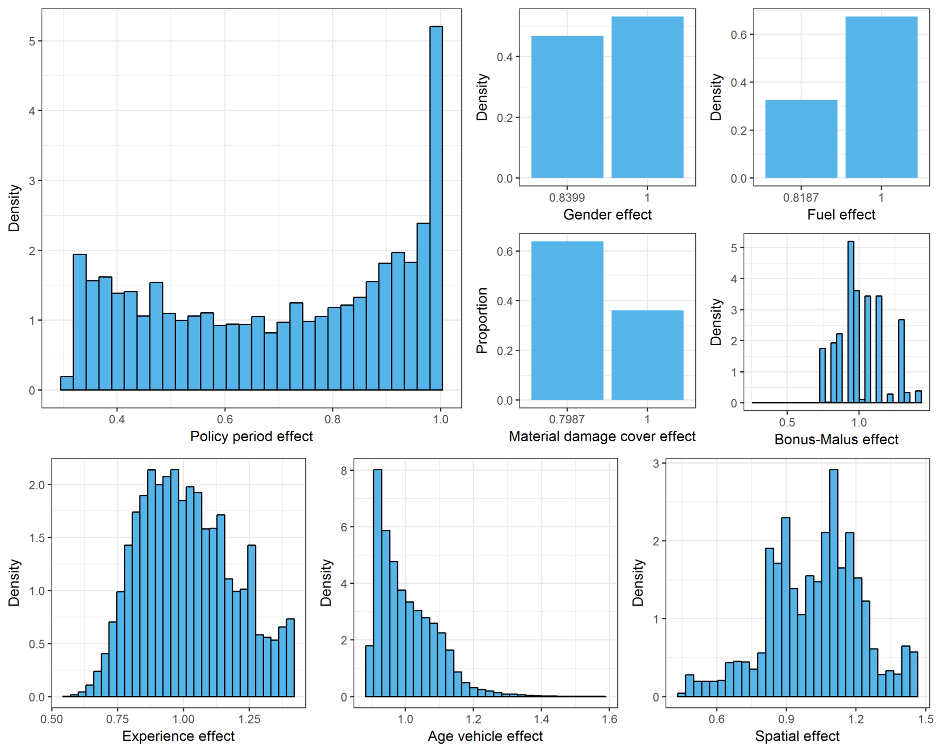

Figure D.5: Relative frequencies of the multiplicative response effects of the predictor variables of the classic model.

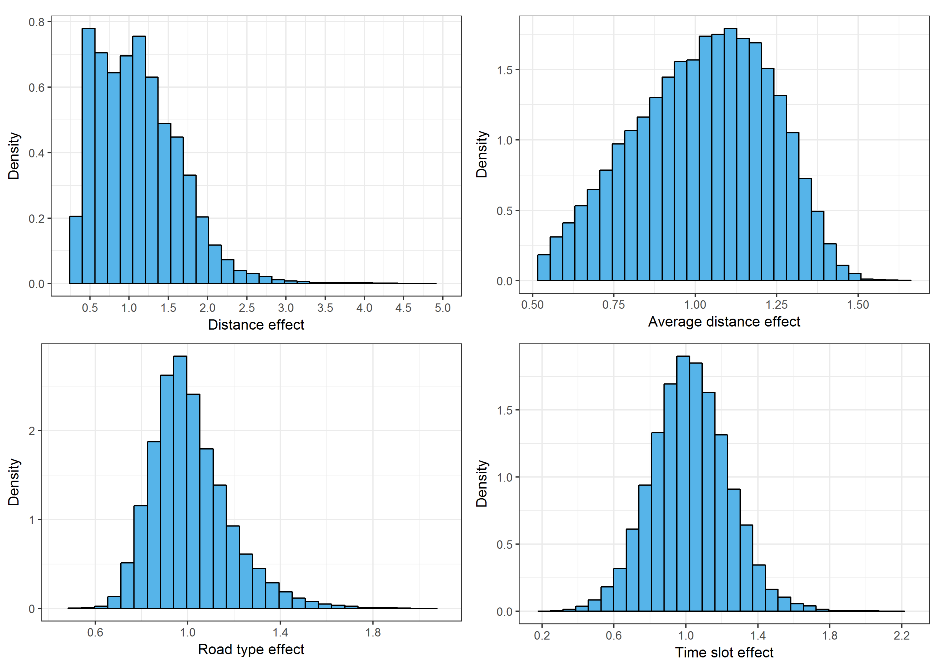

Figure D.6: Relative frequencies of the multiplicative response effects of the predictor variables of the telematics model.

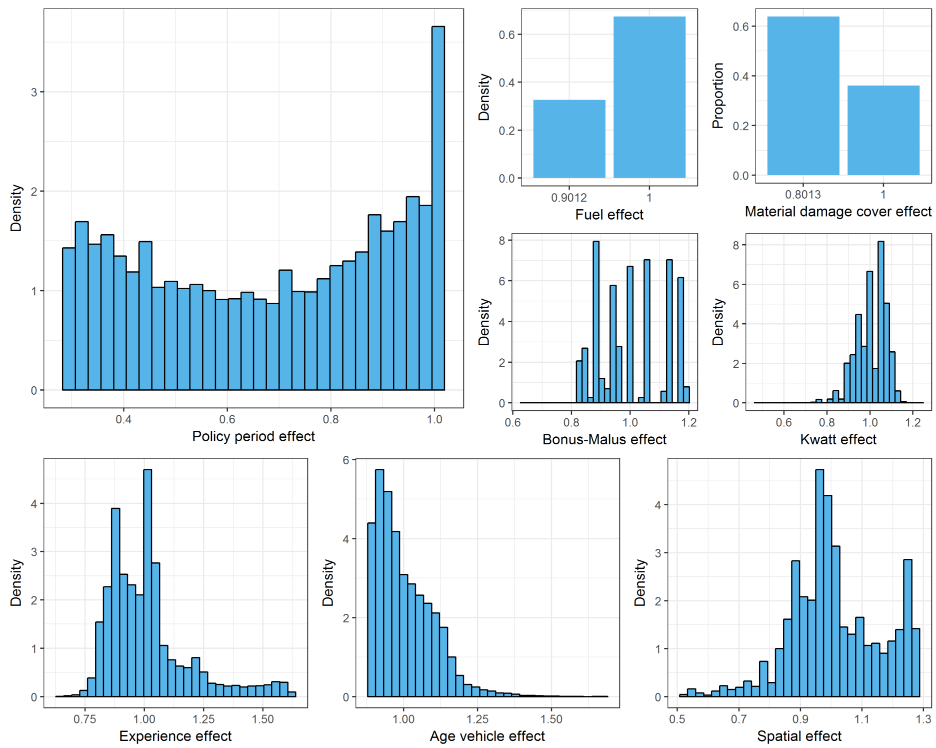

Figure D.7: Relative frequencies of the multiplicative response effects of the policy predictor variables of the time-hybrid model.

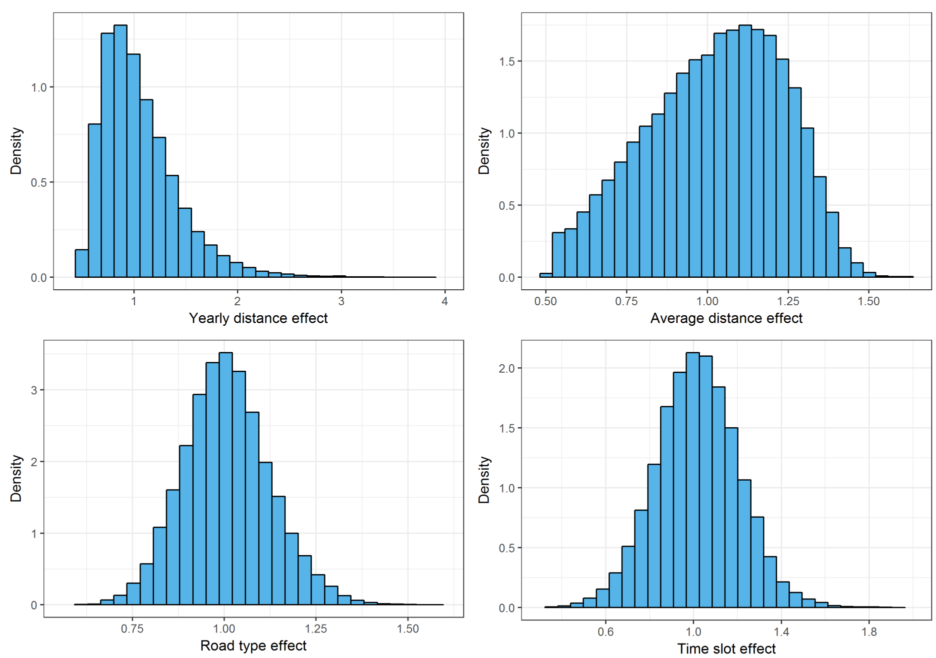

Figure D.8: Relative frequencies of the multiplicative response effects of the telematics predictor variables of the time-hybrid model.

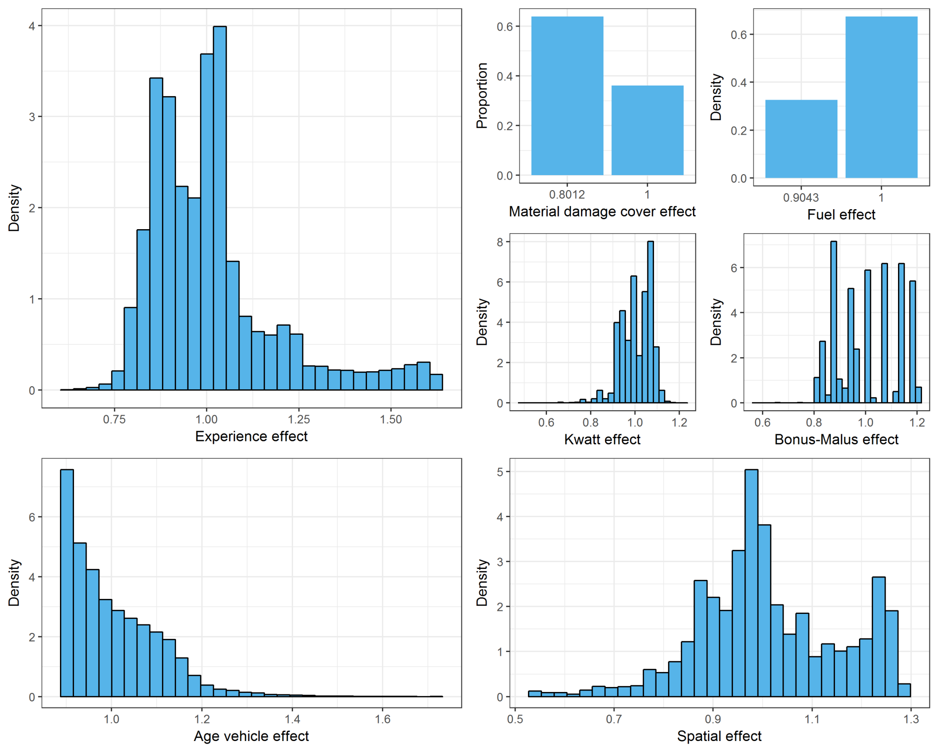

Figure D.9: Relative frequencies of the multiplicative response effects of the policy predictor variables of the meter-hybrid model.

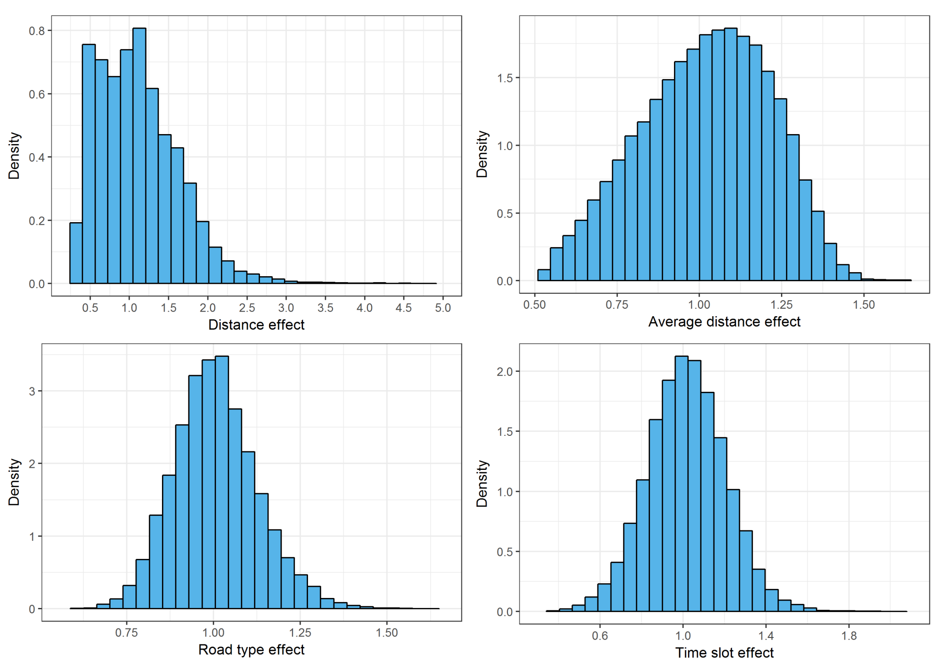

Figure D.10: Relative frequencies of the multiplicative response effects of the telematics predictor variables of the meter-hybrid model.