KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Unraveling the predictive power of telematics data in car insurance pricing

Verbelen R, Antonio K, Claeskens G.

KBI_1624

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.

January 18, 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 discriminating 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. The aim in this paper 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. Pricing through risk classification or segmentation is the mechanism insurance companies use to compete and to reduce the price of insurance contracts. Insurance Europe, the European insurance and reinsurance federation, reports[1] a total motor premium income amounting to €124 billion in 2014. Car insurance is the most widely purchased non-life insurance product in Europe, accounting for 27.3% of non-life premiums. To avoid lapses in this competitive market many rating factors are used to classify risks and differentiate prices. Besides the fierce competition, high acquisition and retention costs, low

[1] http://www.insuranceeurope.eu/european-motor-insurance-markets-addendum

customer engagement, no brand loyalty and a high cost of retention have put a huge pressure on the car insurance industry. 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 present patterns of driving habits and the driving style, and consequently tariff cells are still quite heterogeneous.

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.[2] 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.[3] 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). In addition, some positive externalities are to be expected (Parry, 2005; Litman, 2015; Tselentis et al., 2016). Telematics insurance gives a high incentive to change the current driving pattern and stimulates more responsible driving. 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 includes *Pay-as-you-drive* (PAYD) and *pay-how-you-drive* (PHYD) schemes (Tselentis et al., 2016). PAYD focuses on the driving habits, e.g. the driven distance, the time of day, how long the insured has been driving, and the location. PHYD goes even further by also considering the driving style, e.g. the speed, harsh or smooth braking, aggressive acceleration or deceleration, cornering and parking skills. Furthermore, the telematics data collected can be enriched using other sources of data, for example road maps with corresponding speed limitations to infer road types and speeding violations.

Telematics insurance started as a niche market when the technology first surfaced more than 10 years ago. The high implementation costs and its complexity limited its success. Advances in technology and telecommunication have however reduced the cost substantially. Early adopters of UBI were seen primarily in the United States (US), Italy and the United Kingdom (UK) due to the higher premiums, particularly for young drivers, the highly competitive markets, and a higher incidence of fraudulent claims and vehicle theft. Monti's decree of 2012[4], encouraging Italian insurers to provide a telematics option, has made Italy the most active country in Europe

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

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

[4] Law Decree of 24 January 2012, n.1 "Urgent provisions for competition, infrastructure development and competitiveness" (the so-called "Cresci Italia"), converted by law 24 March 2012, n.27.

in telematics insurance, with the overall penetration level around 15% in June 2016.[5] Ptolemus further reports that at that moment insurance companies have launched 292 telematics programs or active trials worldwide (see Husnjak et al., 2015, for some examples of UBI solutions implemented worldwide). The number of UBI policies is over 7.9 million in the US, over 5 million in Italy and over 860 000 in the UK.[6] Moreover, 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.[7] However, legislation also gives rise to legal concerns and challenges in the telematics insurance market. In particular, insurers have to comply with the aspects of data protection and privacy in the evolving legal environment.

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 statistical methodology, including in particular the challenges when incorporating the divisions of the driven distance by road type and time slots as predictors in the model, is presented in Section 4. In Section 5, we present the results and, finally, in Section 6, we conclude.

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). The pure premium represents the expected cost of the claims a policyholder will declare during the insured period. 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.

A priori pricing refers to the statistical problem of pricing without incorporating the claim history of the policyholder, thus neither frequency nor severity of past claims is taken into account. The construction of an a priori tariff traditionally relies on a frequency-severity modeling framework in which the claim frequency and severity components are typically modeled separately using regression techniques (Frees, 2014). A policyholder's pure premium is obtained by multiplying the

[5] http://www.ptolemus.com/ubi-study/telematics-insurance-infographic/

[6] Ptolemus Consulting Group (2016). Usage-based insurance (global study), free abstract.

[7] 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.

expected claim frequency and expected claim severity, given the observable risk factors. 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. Modeling the claim severities is difficult, since only those observations corresponding to policyholders who filed a claim can be used to estimate the claim severity model and due to the complexity of the phenomenon (Denuit and Charpentier, 2005). On the one hand, there is a long delay to assess the cost of bodily injury and other severe claims and on the other hand the cost of an accident is, for most part, beyond the control of the driver. In practice, covariates are much less informative to predict claim amounts than to predict frequency (Boucher and Charpentier, 2014).

A posteriori pricing refers to experience rating systems which penalize or reward policyholders based on (usually) the number of claims reported in the past. The idea is that, over time, insurers try to refine their a priori risk classification and restore fairness using no-claim discounts and claim penalties. A bonus-malus system is a typical example (Lemaire, 1995). From a statistical point of view a posteriori rating requires the analysis of multilevel data (Gelman and Hill, 2007).

In car insurance, the duration of the policy period during which coverage is provided, is referred to as the exposure-to-risk, the basic rating unit underlying the insurance premium. The expected number of claims is in practice modeled directly proportional to the exposure. The logic behind this is to make the premiums proportional to the length of coverage. As such, a premium related to an insured period of 6 months will be half of the one-year premium, for a given risk profile. From a theoretical point of view, this can also 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. They suggest that possibly high-kilometer drivers are more experienced, have newer and safer vehicles, or drive more on low-risk motorways rather than high-risk urban areas.

Data collected using telematics technology offers more insight in the driving habits. Instead of relying only on the self-reported annual number of driven kilometers, pay-as-you-drive insurance can also account for the type of road and the time of the day when an insured has been driving. A next step is to also take data on driving style into account, leading to a pay-how-you-drive insurance (Weiss and Smollik, 2012). Statistical analysis of these types of data has been the subject of limited academic scrutiny.

Ayuso et al. (2014, 2016) study the traveled time and distance to the first accident using Weibull regression models involving both policy and telematics predictors. 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, combining the new telematics variables with traditional policy(holder) information through a careful model and variable selection process as well as recognizing the compositional structure in the analysis are main focus points in our research, see Section 3.2.

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 have so far no effect on the (future) premium levels of the insureds and do 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 areas, 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.

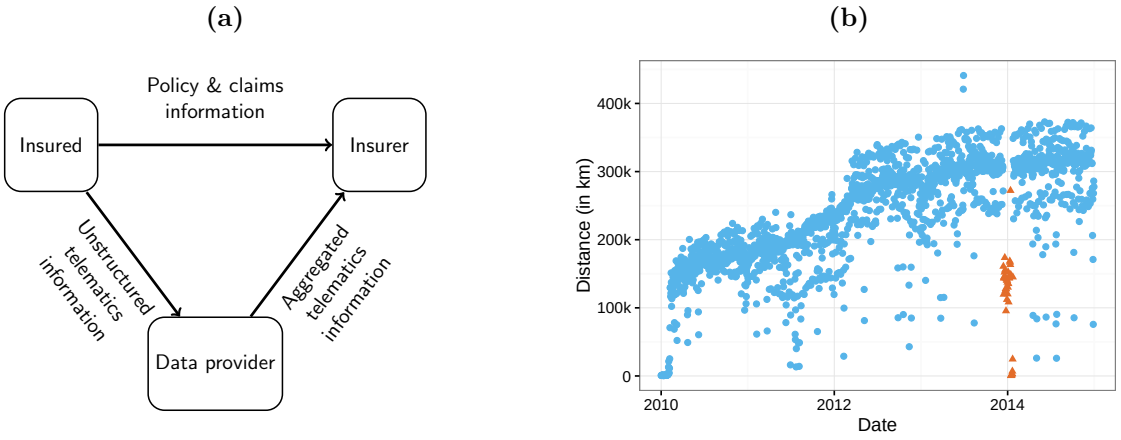

Figure 1: (a) A schematic overview of the flow of information. (b) The number of registered kilometers on each day on an aggregate, portfolio level for the telematics data observed between January 1, 2010 and December 31, 2014. The outliers by the turn of the year 2014, corresponding to a technical malfunction,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

The telematics data are linked with 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 that date 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.

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 link 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 amount of 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.

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 (*motorways, urban areas, abroad* and *other*)
`time slot`	division of the `distance` into 5 time slots (*22h-6h, 6h-9h30, 9h30-16h, 16h-19h* and *19h-22h*)
`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.

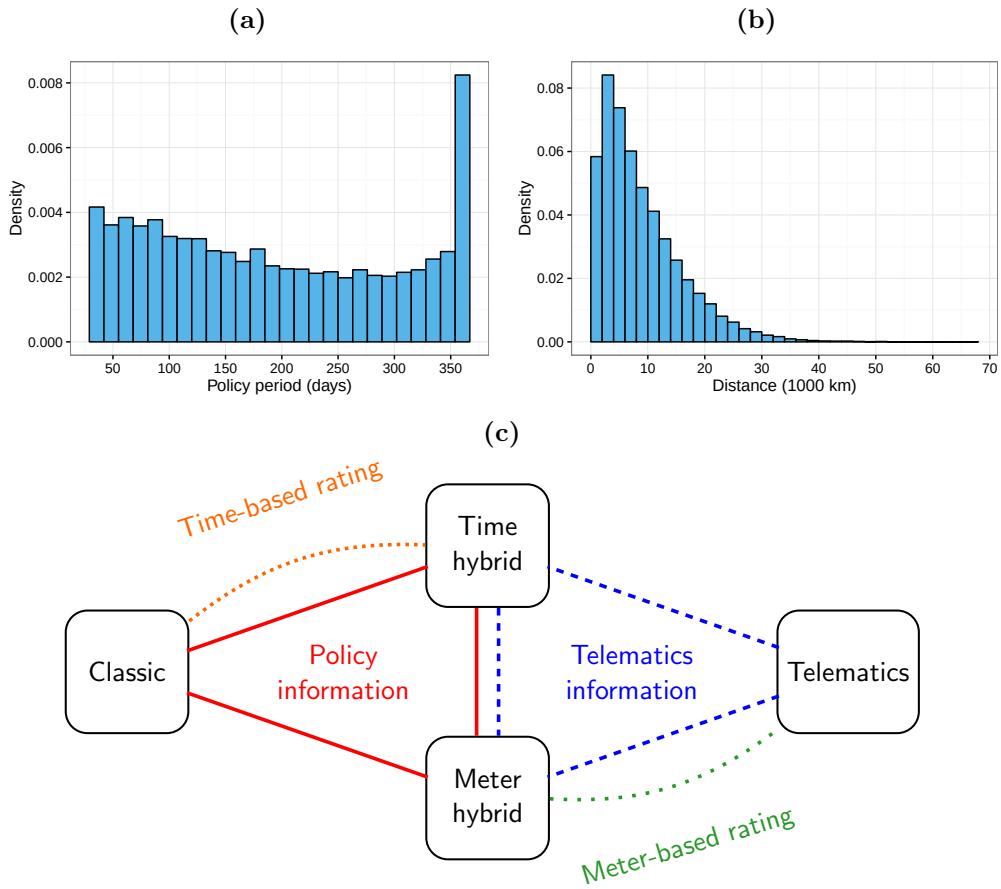

Figure 2: Histogram of (a) the duration (in days) of the policy period (at most one year) and (b) the driven distance (in 1000 km) during the policy period. (c) A graphical representation of the similarities and differences between the four predictor sets.

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 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 as exposure and the use of distance as exposure 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. Even though, the bonus-malus scale level is not a covariate of the same type as the other a priori variables, we

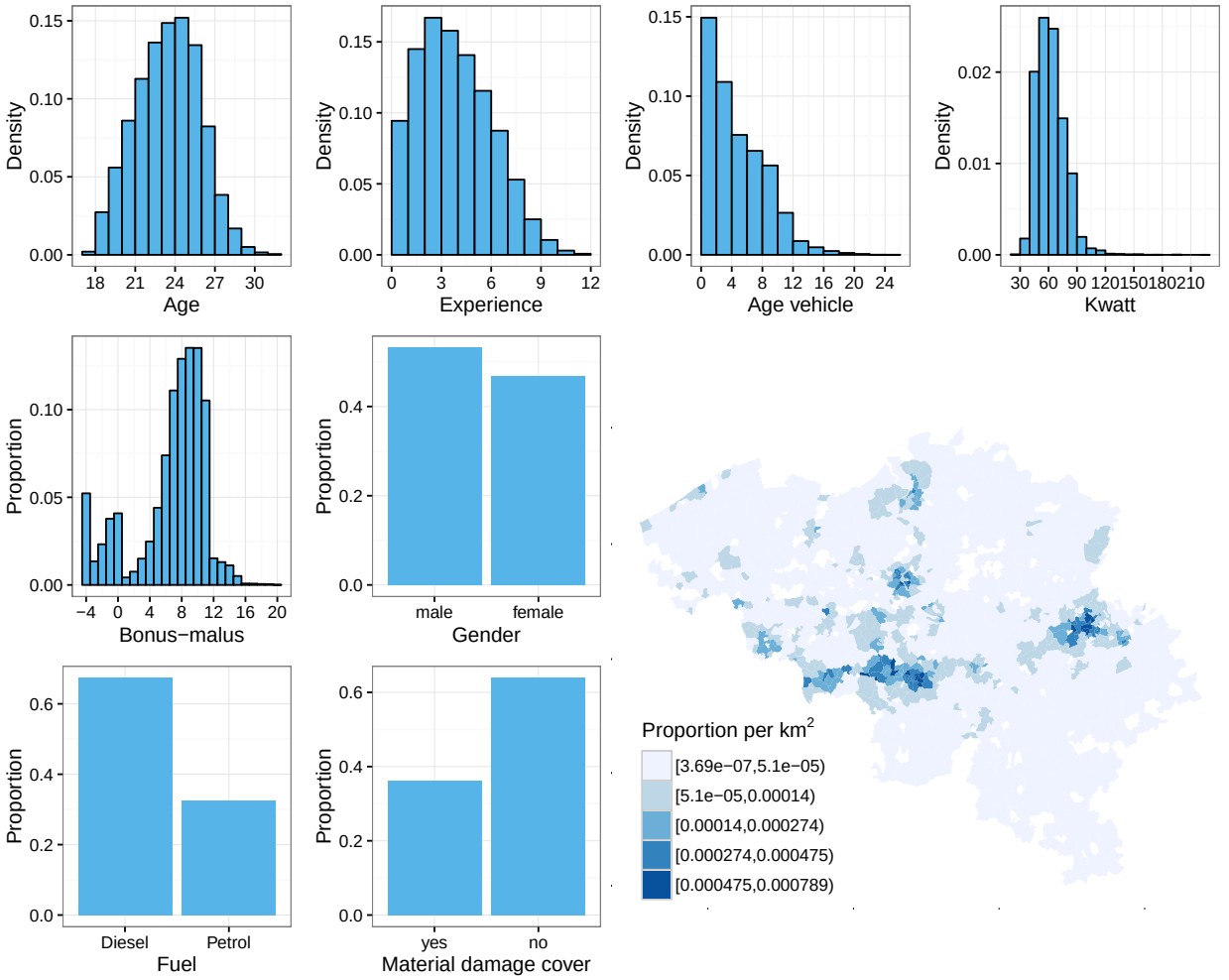

Figure 3: Histograms and bar plots of the continuous and categorical policy variables contained in the data set. The map in the lower right depicts the geographical information by showing the proportion of insureds per squared kilometer living in each of the different postal codes in Belgium. The five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

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 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 driven distance 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 in our data set. 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

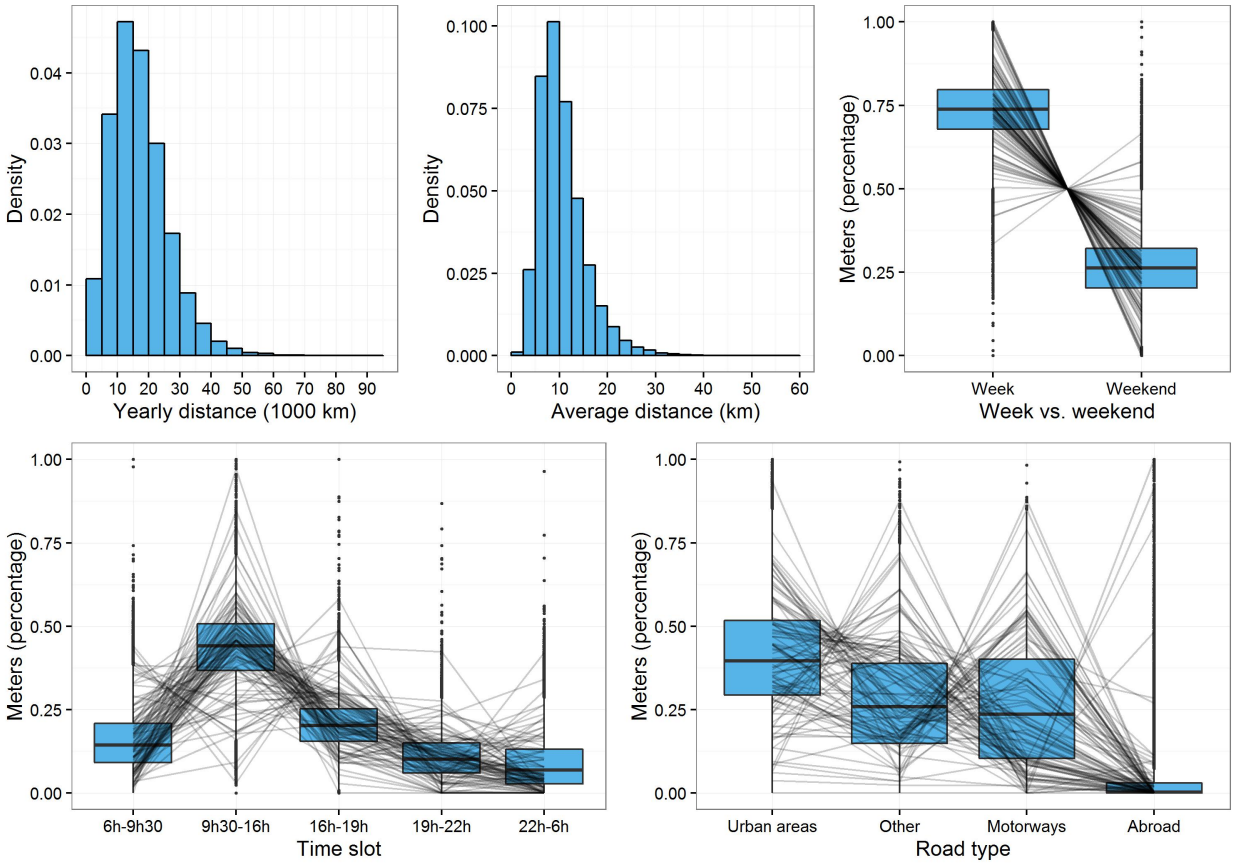

Figure 4: Graphical illustration of the telematics variables contained in the data set. For the yearly and average distance, we construct histograms. For the division of the driven distance by road types, time slots and week/weekend, we construct box plots of the relative proportions. To highlight the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

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 another key issue in this research.

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

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

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_{it}^{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 it is not realistic to assume that the sample is coming from a mixture of two sorts of drivers: one group of drivers whose number of claims are generated by the standard regression model, and another group of drivers who have a zero probability of a claim count greater than 0. Moreover, such models are also 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 predictors 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 parametric terms for the categorical predictors 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 f_j can be represented as $f_j(x) = \sum_{k=1}^{q} \gamma_{jk} B_{jk}(x)$. 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).

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_{it}^{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 in 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. Incorporating these divisions in a predictor also containing the total distance leads to a perfect multicollinearity problem. Furthermore, the corresponding model parameter estimators are not invariant to the ordering of the components: the statistical inference changes when permuting the components 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 of adding up to the total distance.

The total distance in meters is used as a continuous predictor in the telematics models and its effect is modeled using a smooth function. Since the divisions of the distance only contribute additional relative information, we divide all components of each split by the total driven distance, see Figure 4. We obtain what is known as *compositional data* (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). Such data are represented by real vectors with constant sum equal to one and positive components. The space of representations of compositions

is called the simplex of D parts, denoted $\mathcal{S}^D$, defined by

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

Only relative information is important, and multiplication of the vector of positive components by a positive constant does not change the ratios between the components. When data are considered compositional, classical statistics, that do not take the special geometry of the simplex into account, are not appropriate. Extending the current literature, we propose a new way of quantifying and interpreting the effect of the compositional explanatory variables on the outcome and propose an approach to deal with structural zeros.

4.2.1 The Aitchison geometry of the simplex

The vector space structure of the mathematical simplex was discovered by Aitchison (1986) who defined operations on compositional data leading to the Aitchison geometry of the simplex. Perturbation plays the role of addition on the simplex and is defined as a closed component-wise product $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)^t$, where the closing operation $\mathcal{C}$ ensures a total sum of one, 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 $\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)^t$, for $\alpha \in \mathbb{R}$. The Aitchison inner product for compositions is

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

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. $\ominus \boldsymbol{y} = \oplus((-1) \odot \boldsymbol{y})$. 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.

Egozcue et al. (2003) constructed orthonormal bases for this Euclidean space and deduced corresponding isometries between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, called isometric logratio transformations (*ilr*). One possible ilr transformation 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

$$z_i = \mathrm{ilr}_i(\boldsymbol{x}) = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1 . \tag{3}$$

As the ilr transformation is 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. Let V be the $D \times (D-1)$ matrix with elements

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

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 rewrite this ilr transform and its inverse in matrix notation as

$$\boldsymbol{z} = \mathrm{ilr}(\boldsymbol{x}) = V^t \ln \boldsymbol{x} , \quad \text{and} \quad \boldsymbol{x} = \mathrm{ilr}^{-1}(\boldsymbol{z}) = \mathcal{C}(\exp(V \boldsymbol{z})) , \tag{4}$$

where the logarithmic and exponential function apply componentwise.

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 this aspect in a statistical context can lead to incompatible or incoherent results. The compositional nature of the data must not be ignored. The principle of working on coordinates in statistics (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 setting, it is key to incorporate the compositional data arising from the divisions of the distances into different categories as predictors in the claim count regression models. Hron et al. (2012) propose to first apply the isometric log ratio transform (3) to map the compositions in the D-part Aitchison simplex to a $(D-1)$ Euclidean space. Then, these terms are used as explanatory variables in a linear regression model. 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.

$$\eta^{\text{comp}} = \beta_1 z_1 + \ldots + \beta_{D-1} z_{D-1}\,. \tag{5}$$

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. Only the first regression parameter, β_1, however has a comprehensible interpretation since z_1 explains relevant information about x_1. The remaining coefficients are not straightforward to interpret and hence Hron et al. (2012) suggest to permute the indices in formula (3) and construct D regression models, each time with a different component first for which we can interpret the corresponding coefficient. Having to refit the model multiple times is undesirable, especially 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 strategy to include compositional predictors and interpret their effect.

By using the inverse ilr transform on the model coefficients, i.e. set $\boldsymbol{b} = \text{ilr}^{-1}(\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{D-1})^t$, we can rewrite the compositional predictor as

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

since the ilr transform preserves the inner product (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). 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 increase $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a}$, 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.

In order to overcome this shortcoming in interpretation and to develop a graphical representation for compositional explanatory variables, 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 $(\alpha, 1, \ldots, 1)^t$. This leads to a change of the predictor given by

$$\langle \boldsymbol{b}, (\alpha, 1, \ldots, 1)^t \rangle_a = \ln(b_1)\ln(\alpha) - \frac{1}{D}\left(\sum_{i=1}^{D} \ln(b_i)\right)\ln(\alpha) = \mathrm{clr}_1(\boldsymbol{b})\ln(\alpha)\,, \tag{6}$$

which is independent of the original composition $\boldsymbol{x}$ and where

$$\mathrm{clr}_i(\boldsymbol{b}) = \ln\left(\frac{b_i}{g_m(\boldsymbol{b})}\right), \qquad g_m(\boldsymbol{b}) = \left(\prod_{i=1}^{D} b_i\right)^{1/D}, \qquad i = 1, \ldots, D$$

denotes the centered log-ratio (*clr*) transform of $\boldsymbol{b}$ (Egozcue et al., 2011). 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. 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 interpret the effect on the level of the expected outcome in the Poisson and NB models, we can transform these confidence intervals using the exponential function. The exponentiated clr transform of $\boldsymbol{b}$ has to be compared to one and the effect of a relative ratio change of α in component $i = 1, \ldots, D$ is 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.

The *structural zeros* patterns (Pawlowsky-Glahn et al., 2015) are listed in Appendix A. The presence of zeros is most prominent for splitting distance by road types as 40% of the drivers did not go abroad. Unlike *rounded zeros*, when certain components may be unobserved because their true values are below the detection limit (cfr. geochemical studies), or *count zeros*, when the zero values are due to the limited size of the sample in compositional data arising from count data, structural zeros are truly zero. 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.

We propose to treat the structural zero patterns of the compositional predictors as different subgroups within the data and model the effect conditional on the zero pattern. 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). We introduce indicator variables for each zero pattern and use these in the compositional predictor term η^{comp} of the regression model to specify the effect on the outcome separately for each zero pattern. More specifically, we define the variables

$$d_{(i_1,\ldots,i_k)} = \begin{cases} 1 & \text{if components } i_1,\ldots,i_k \text{ of } \boldsymbol{x} \text{ are nonzero and all other are zero}\,, \\ 0 & \text{otherwise} \end{cases}$$

for all $k = 1,\ldots,D$ and $1 \leqslant i_1 < \ldots < i_k \leqslant D$. Conditional on the zero pattern $(i_1,\ldots,i_k)$ of the compositional data vector $\boldsymbol{x}$, the contribution to the predictor is given by the Aitchison inner product $\langle \boldsymbol{b}_{(i_1,\ldots,i_k)}, \boldsymbol{x}_{(i_1,\ldots,i_k)} \rangle_a$ of the subcomposition $\boldsymbol{x}_{(i_1,\ldots,i_k)}$ existing of the nonzero components of $\boldsymbol{x}$ and a subcompositional simplicial gradient $\boldsymbol{b}_{(i_1,\ldots,i_k)}$, which is different for each zero pattern. In case of only one nonzero component, the contribution is given by a simple categorical effect $b_{(i)}$. Note that the subscript $(i_1,\ldots,i_k)$ has a different interpretation for the dummy variable, simplicial gradient and compositional data vector. The proposed compositional predictor reads

$$\eta^{\text{comp}} = \sum_{i=1}^{D} d_{(i)} b_{(i)} + \sum_{k=2}^{D} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant D} d_{(i_1,\ldots,i_k)} \langle \boldsymbol{b}_{(i_1,\ldots,i_k)}, \boldsymbol{x}_{(i_1,\ldots,i_k)} \rangle_a \,.$$

Zero pattern specific intercepts can be added in the second term if deemed necessary.

4.3 Model selection and assessment

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

$$\text{AIC} = -2 \cdot \widehat{\ell} + 2 \cdot \text{EDF} \tag{7}$$

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. The EDF is defined as the trace of the hat or smoothing matrix in the corresponding working linear model at the last P-IRLS iteration (Hastie and Tibshirani, 1990). As such, (7) 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. We limit ourselves 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 division of the number of meters in different categories, 10 structural zero patterns occur for the road types, 20 for the time slots, and 3 for week/weekend. However, based on their relative frequencies, we only allow an additional compositional predictor 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. The most comprehensive compositional predictor term we allow to be selected in the hybrid and telematics models is

$$\begin{aligned} \eta_{it}^{\text{comp}} =& d_{(1111)}^{\text{road}} \langle \boldsymbol{b}_{(1111)}^{\text{road}}, \boldsymbol{x}_{(1111)} \rangle_a + d_{(1110)}^{\text{road}} \langle \boldsymbol{b}_{(1110)}^{\text{road}}, \boldsymbol{x}_{(1110)} \rangle_a + (1 - d_{(1111)}^{\text{road}} - d_{(1110)}^{\text{road}}) b_0^{\text{road}} \\ &+ d_{(11111)}^{\text{time}} \langle \boldsymbol{b}_{(11111)}^{\text{time}}, \boldsymbol{x}_{(11111)} \rangle_a + (1 - d_{(11111)}^{\text{time}}) b_0^{\text{time}} \\ &+ d_{(11)}^{\text{week}} \langle \boldsymbol{b}_{(11)}^{\text{week}}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d_{(11)}^{\text{week}}) b_0^{\text{week}} \,. \end{aligned}$$

In total, 165 888 model specifications are estimated under both the Poisson and the negative binomial framework.

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 negative binomial 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 $\kappa_{it} \in 1, 2, \ldots K$ 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

5.1 Model selection

All computations are performed with `R` 3.2.5 (R Core Team, 2016) and, in particular, the `R` package `mgcv` version 1.8-11 (Wood, 2011) is used for the parameter estimation in the GAMs. 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. 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. Both hybrid models drop the `fuel` term in the best offset variants.

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. The newly introduced telematics predictors are selected in both the hybrid and the telematics models and hence contribute to the quality of these models.

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 `fuel` and `kwatt` can easily be left out of the hybrid models without deteriorating the fit.

	Predictor	Classic		Time-hybrid		Meter-hybrid		Telematics	
Policy	Time	×	offset	×	offset				
	Age								
	Experience	×	×	×	×	×	×		
	Sex	×	×						
	Material	×	×	×	×	×	×		
	Postal code	×	×	×	×	×	×		
	Bonus-malus	×	×	×	×	×	×		
	Age vehicle	×	×	×	×	×	×		
	Kwatt			×	×	×	×		
	Fuel	×	×	×		×			
Telematics	Distance					×	offset	×	offset
	Yearly distance			×	×				
	Average distance			×	×	×	×		
	Road type 1111			×	×	×	×	×	×
	Road type 1110			×	×	×	×	×	×
	Time slot			×	×	×	×	×	×
	Week/weekend			×	×	×	×	×	×

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

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. 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 restriction conclude the top four according to all criteria except 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.

Predictor set	Offset	EDF	AIC		logs		qs		sphs		rps		dss		ses	
			value	rank	value	rank	value	rank	value	rank	value	rank	value	rank	value	rank
Classic	no	32.15	11 896	6	0.1790	6	−0.918 58	6	−0.958 22	6	0.042 24	6	−2.206	5	0.045 35	6
	yes	27.27	11 995	7	0.1804	7	−0.918 39	7	−0.958 16	7	0.042 34	7	−2.130	7	0.045 46	7
Time-hybrid	no	39.66	11 727	1	0.1764	1	−0.919 10	1	−0.958 37	1	0.041 95	1	−2.275	1	0.045 01	1
	yes	36.22	11 811	3	0.1777	3	−0.918 90	3	−0.958 31	3	0.042 06	3	−2.212	4	0.045 14	3
Meter-hybrid	no	41.47	11 736	2	0.1766	2	−0.919 08	2	−0.958 36	2	0.041 96	2	−2.266	2	0.045 02	2
	yes	36.23	11 856	4	0.1784	4	−0.918 80	4	−0.958 27	4	0.042 12	4	−2.158	6	0.045 22	4
Telematics	no	18.05	11 890	5	0.1787	5	−0.918 60	5	−0.958 22	5	0.042 22	5	−2.224	3	0.045 32	5
	yes	14	12 061	8	0.1813	8	−0.918 16	8	−0.958 07	8	0.042 50	8	−2.066	8	0.045 80	8

Table 4: Model assessment of the best models according to AIC under each of the four predictor sets. The second row of each predictor set refers to the model with the offset restriction 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.

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. The rankings according to AIC are the same as in Table 4. 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 in Table 4 of the best models.

5.3 Visualization and discussion

The effects of each predictor variable in the best time-hybrid model without offset restriction are graphically displayed in Figure 5 for the policy model terms and Figure 6 for the telematics model terms. 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 exponentiated clr transform of the corresponding model parameters with 95% confidence intervals along with a reference line at one (see Section 4.2.2). 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.

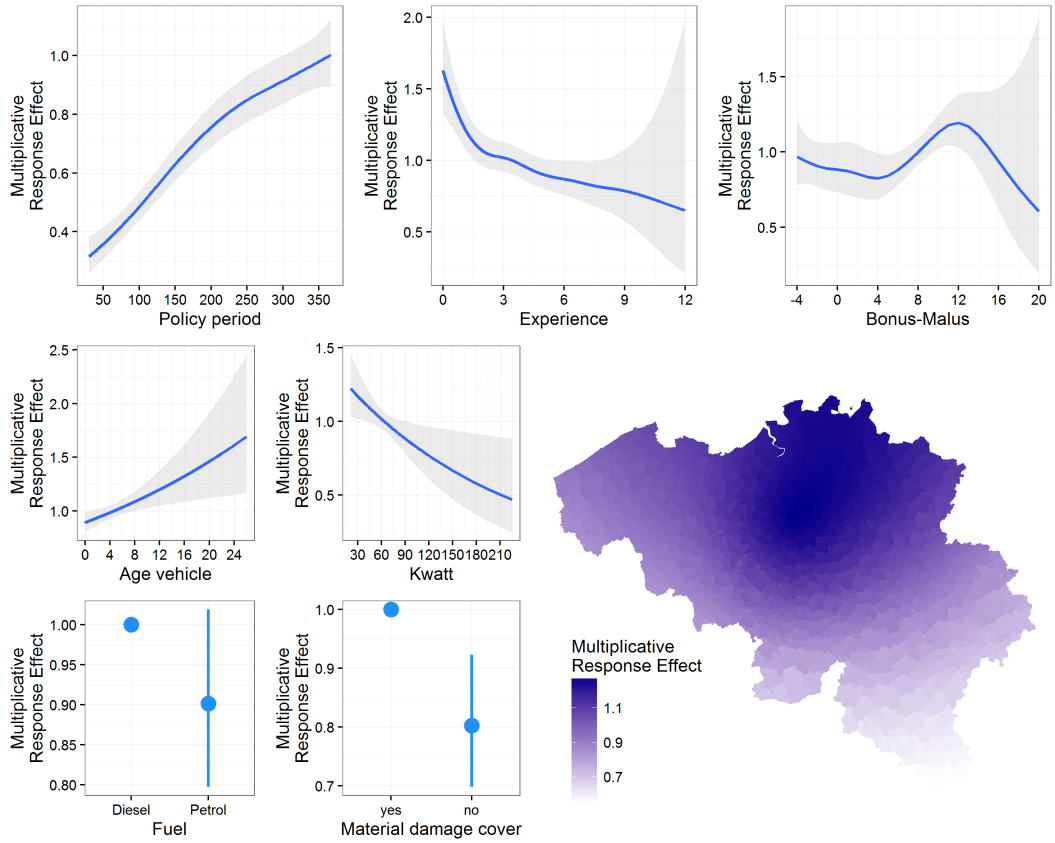

Figure 5: Multiplicative response effects of the policy model terms of the time-hybrid model.

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 components are 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 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. (2016). 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 level 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, although this effect is of lesser importance for the model fit as indicated earlier. The categorical model term `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.

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 6.4 in both hybrid models. 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 (such as the proportion driven on urban roads) 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

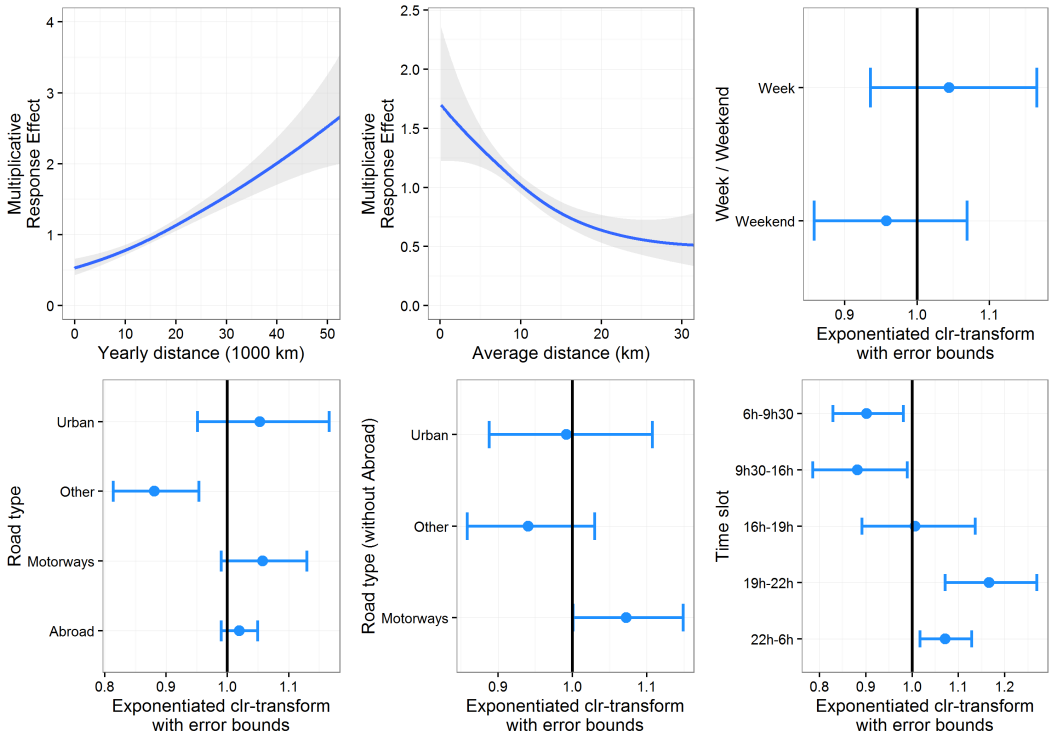

Figure 6: Multiplicative response effects of the telematics model terms of the time-hybrid model.

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, the estimated `average distance` effect shows lower claim frequencies for insureds who on average drive long distances.

The exponentiated 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 relative more on the road type 'other' have lower claim frequencies. The same interpretation holds for insureds who do not drive abroad during the policy period. In the hybrid models, these effects are less outspoken but heading 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 (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 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. In Paefgen et al. (2014), the accident risk is considered to be lower during the daytime (between 5 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.

Driving more in the week than in the weekend increases the probability of having a claim. An increased accident risk in case of more driving in the week is also found in Paefgen et al. (2014), though they define weekend from Friday to Sunday. The compositional effect of `week/weekend` is retained in both hybrid models as well as the telematics model according to AIC even though it is not statistically significant. This is due to a highly significant and positive estimated categorical effect b_0^{week} for the 73 observations with structural zeros belonging to the rest group, see Table A.4 of Appendix A. These drivers have jointly driven 58 000 kilometers during a combined insured policy period of 16.5 years and reported the remarkably high number of 5 claims.

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, in the evening or at night and during the week contributes to a riskier driving pattern. The best hybrid models highlight that certain popular pricing factors (gender, fuel, postcode) 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 acknowledge support 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). We would also like to thank our 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., Guillen, M., and Pérez-Marín, A. M. (2016). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

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. and Charpentier, A. (2014). General insurance pricing. In *Computational Actuarial Science with R*, pages 475–510. Chapman and Hall/CRC.

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 Charpentier, A. (2005). *Mathématiques de l'assurance non-vie. Tome II: Tarification et provisionnement*. Collection "Economie et statistiques avancées". Economica.

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.

Gelman, A. and Hill, J. (2007). *Applied Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

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. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

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.

Husnjak, S., Peraković, D., Forenbacher, I., and Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, 100:816–825.

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.

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

Lemaire, J. (1995). *Bonus–malus systems in automobile insurance*. Springer–Verlag, New York.

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 (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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. 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.

Weiss, J. and Smollik, J. (2012). Beginner's roadmap to working with driving behavior data. *Casualty Actuarial Society E-Forum*, 2:1–35.

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 semi-parametric 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.

A Structural zero patterns of the compositional telematics 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 M observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$ and $i = 1, \ldots, M$ is defined as

$$\overline{\boldsymbol{x}} = \frac{1}{M} \odot \bigoplus_{i=1}^{M} \boldsymbol{x}_i = \mathcal{C}\left(\left(\prod_{i=1}^{M} x_{i1} \right)^{1/M}, \ldots, \left(\prod_{i=1}^{M} x_{iD} \right)^{1/M} \right)^t \tag{8}$$

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

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

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 $\mathcal{S}^D$. In the paper, infrequently observed patterns are bundled into a residual group when incorporating the compositional variables as predictors in the claim count models leading to the distinguished structural zero patterns of Table A.4.

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 in road types.

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 in time slots.

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 in week and weekend.

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 in road types, time slots and week/weekend as recognized in the models in the paper.

B Functional forms of the selected best models

The functional form of the predictor in the preferred classic model can be written as

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

The predictor in the best time-hybrid model is

$$\begin{aligned}\eta_{it}^{\text{time-hybrid}} =&\beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{time})_{it} + f_2(\texttt{experience}_{it}) \\ &+ f_3(\texttt{bonus-malus}_{it}) + f_4(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) \\ &+ f_5(\texttt{yearly distance}_{it}) + f_6(\texttt{average distance}_{it}) + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a \\ &+ d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a + (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a \\ &+ (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} + d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}}\,,\end{aligned}$$

and for the preferred meter-hybrid model we have

$$\begin{aligned}\eta_{it}^{\text{meter-hybrid}} =&\beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{experience}_{it}) + f_2(\texttt{bonus-malus}_{it}) \\ &+ f_3(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) + f_4(\texttt{distance}_{it}) \\ &+ f_5(\texttt{average distance}_{it}) + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a + d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a \\ &+ (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a + (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} \\ &+ d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}}\,.\end{aligned}$$

Finally, the predictor in the best telematics model is

$$\begin{aligned}\eta_{it}^{\text{telematics}} =&\beta_0 + f_1(\texttt{distance})_{it} + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a + d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a \\ &+ (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a + (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} \\ &+ d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}}\,.\end{aligned}$$

C Graphical displays of the multiplicative response effects

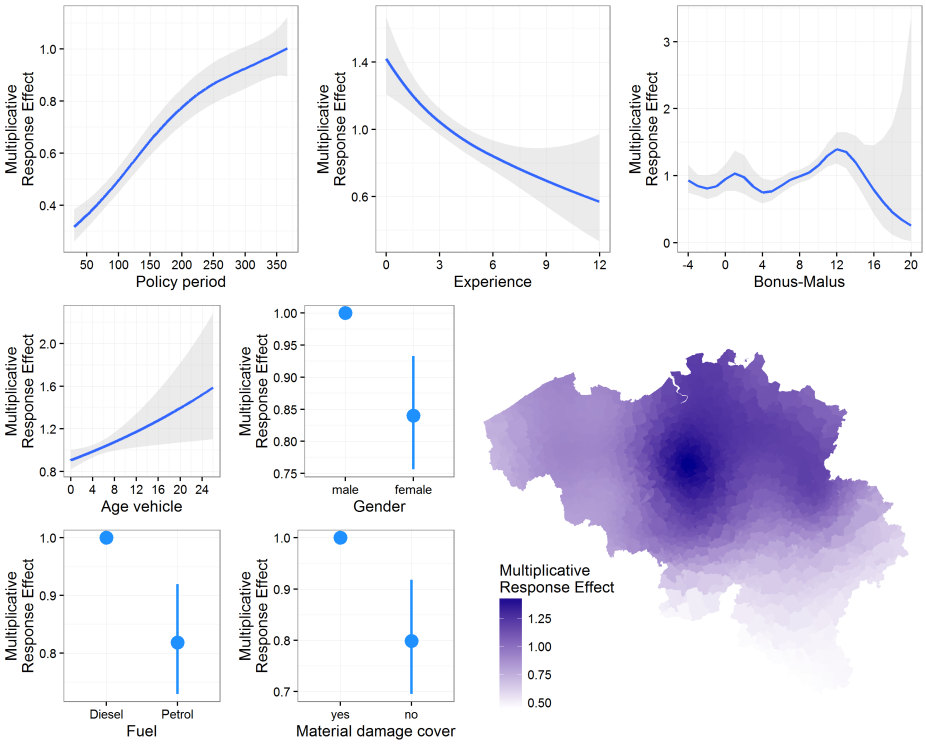

Figure C.1: Multiplicative response effects of the model terms of the classic model.

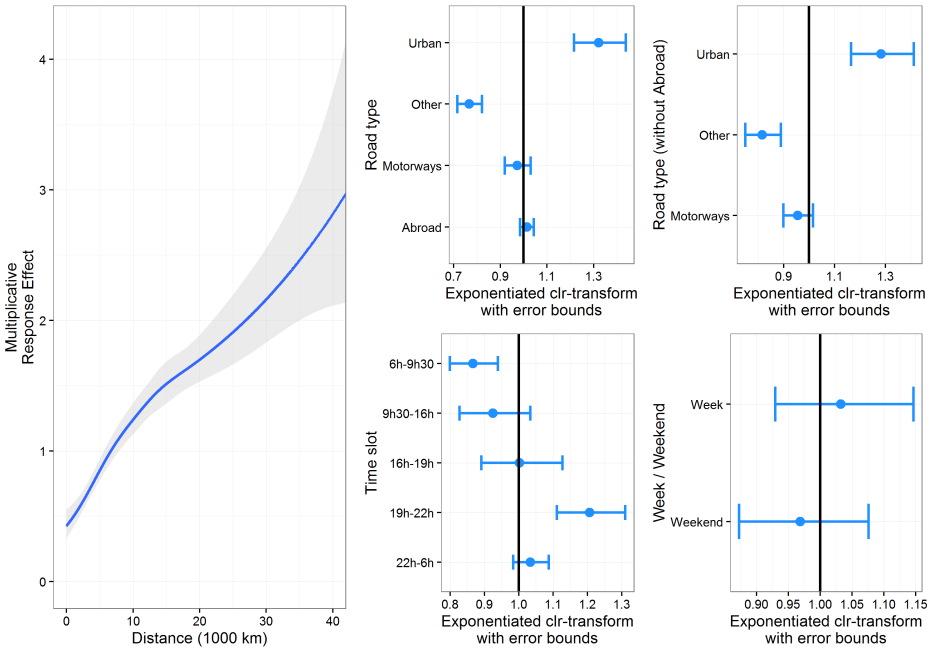

Figure C.2: Multiplicative response effects of the model terms of the telematics model.

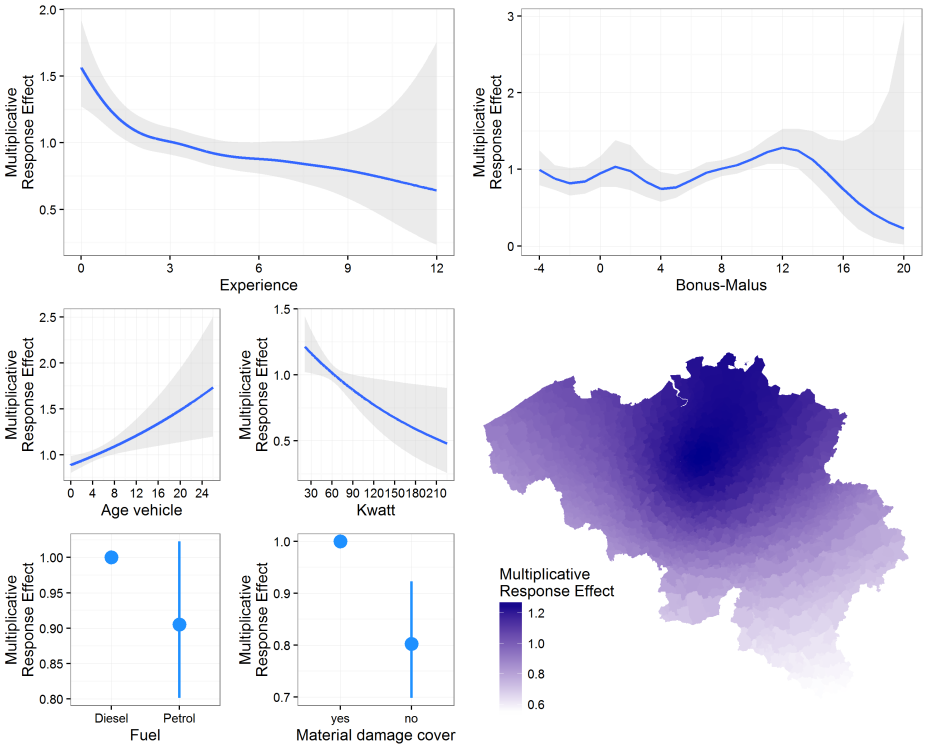

Figure C.3: Multiplicative response effects of the policy model terms of the meter-hybrid model.

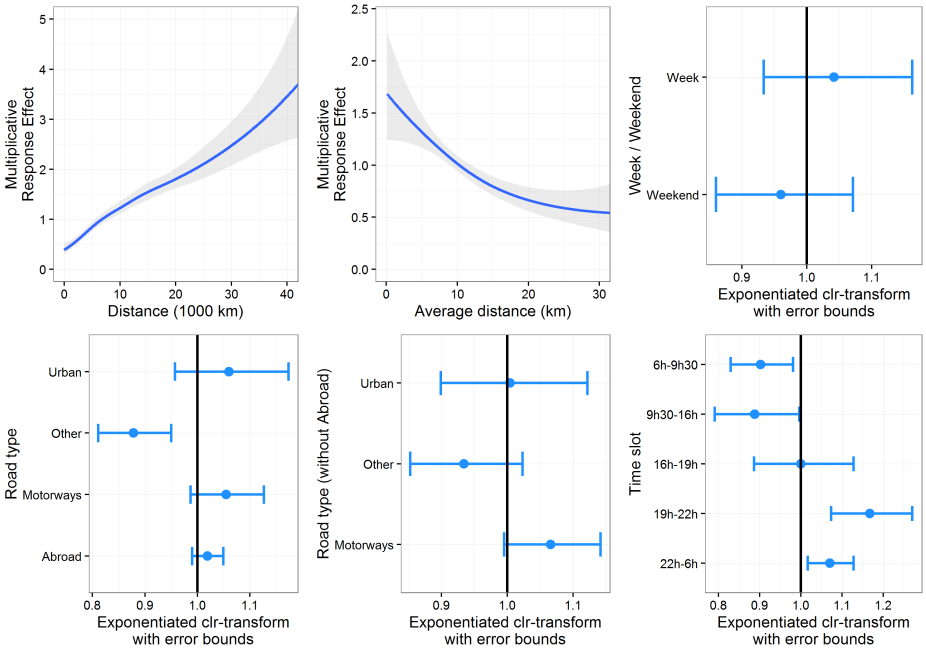

Figure C.4: Multiplicative response effects of the telematics model terms of the meter-hybrid model.

D Relative importance of the predictors

To assess the relative importance of these variables in the model, we construct histograms of the multiplicative effects by predictor for each observation in the data set. This is done for the classic model in Figure C.1, for the telematics model in Figure C.2, for the time-hybrid model in Figures D.7 and D.8 and for the meter-hybrid model in Figures C.3 and C.4. For the hybrid models, we constructed separate graphs for the model terms derived from the policy and telematics information. For categorical predictors this 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 term $\langle \hat{\boldsymbol{b}}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a$ for all observations with pattern 11111. With the division in road types, we consider simultaneously the terms related to patterns 1111 and 1110. 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.

	Predictor	Classic		Time-hybrid		Meter-hybrid		Telematics	
Policy	Time	0.36	0.69	0.37	0.69				
	Age								
	Experience	0.18	0.14	0.16	0.11	0.15	0.12		
	Gender	0.09	0.09						
	Material	0.11	0.11	0.11	0.10	0.11	0.10		
	Postal code	0.21	0.20	0.14	0.14	0.14	0.16		
	Bonus-malus	0.16	0.18	0.11	0.15	0.14	0.15		
	Age vehicle	0.08	0.10	0.09	0.10	0.10	0.11		
	Kwatt			0.07	0.06	0.07	0.08		
	Fuel	0.09	0.09	0.05		0.05			
Telematics	Distance					0.44	0.95	0.41	0.95
	Yearly distance			0.30	0.36				
	Average distance			0.23	0.25	0.21	0.32		
	Road type			0.13	0.14	0.12	0.15	0.22	0.33
	Time slot			0.20	0.20	0.20	0.18	0.21	0.19
	Week/weekend			0.03	0.03	0.03	0.04	0.02	0.02

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

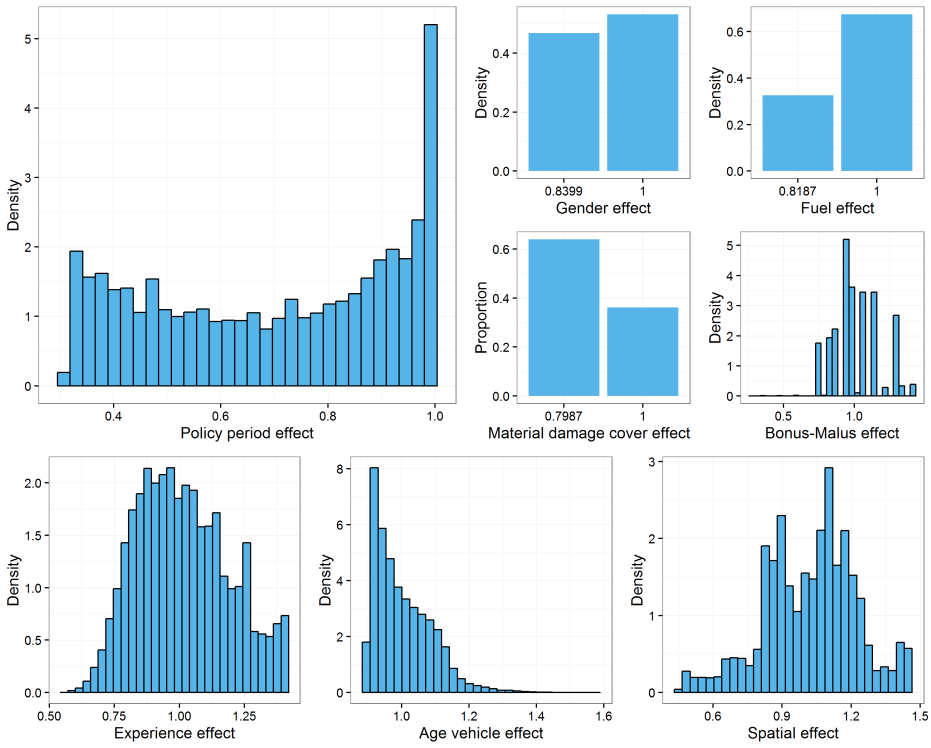

Figure D.5: Relative frequencies of the multiplicative response effects of the model terms of the classic model.

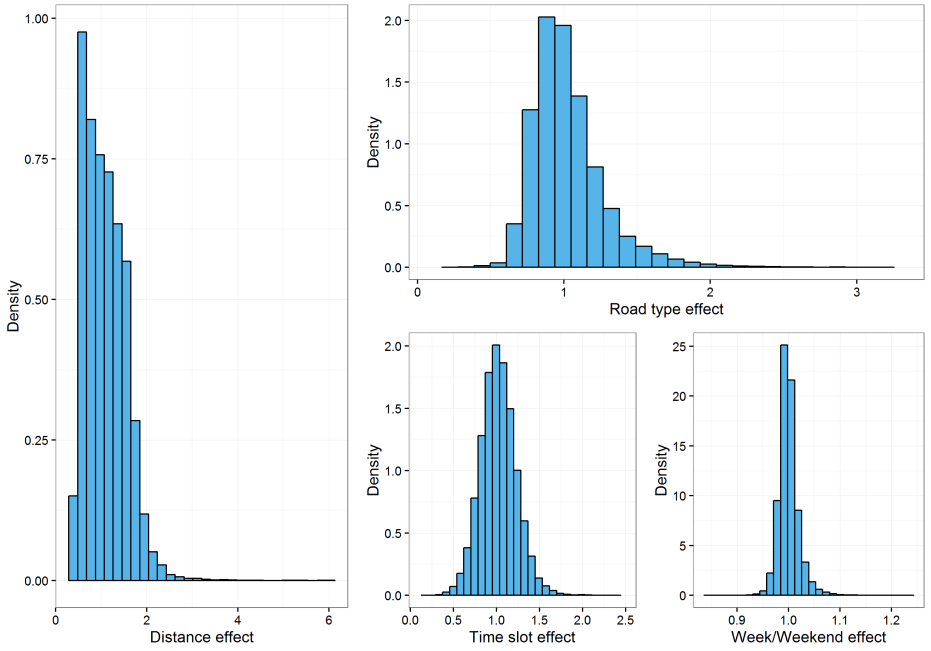

Figure D.6: Relative frequencies of the multiplicative response effects of the model terms of the telematics model.

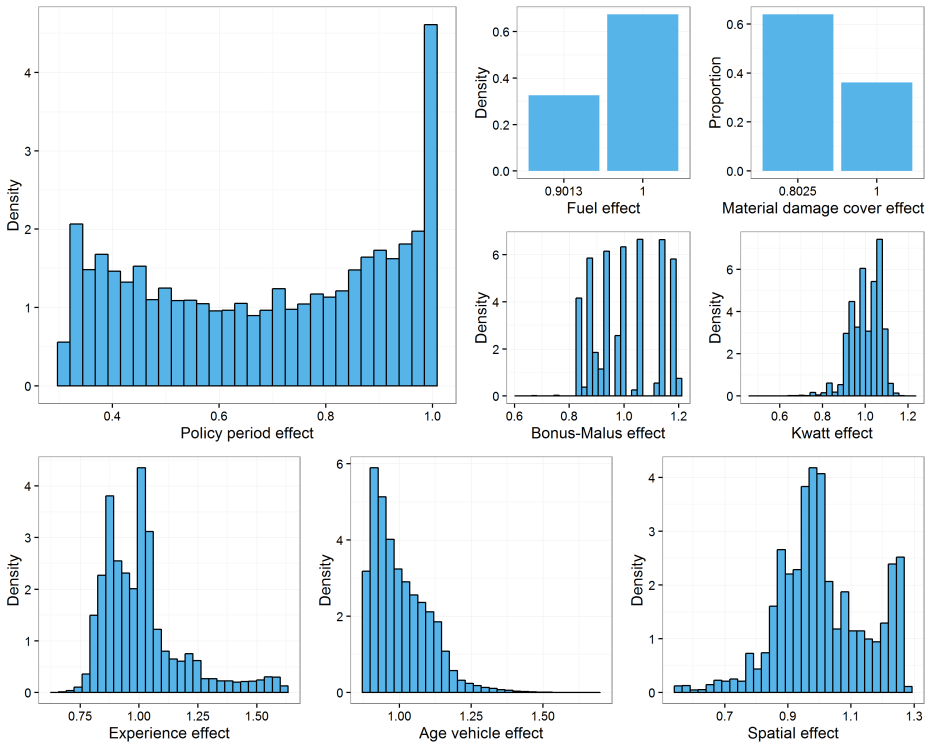

Figure D.7: Relative frequencies of the multiplicative response effects of the policy model terms of the time-hybrid model.

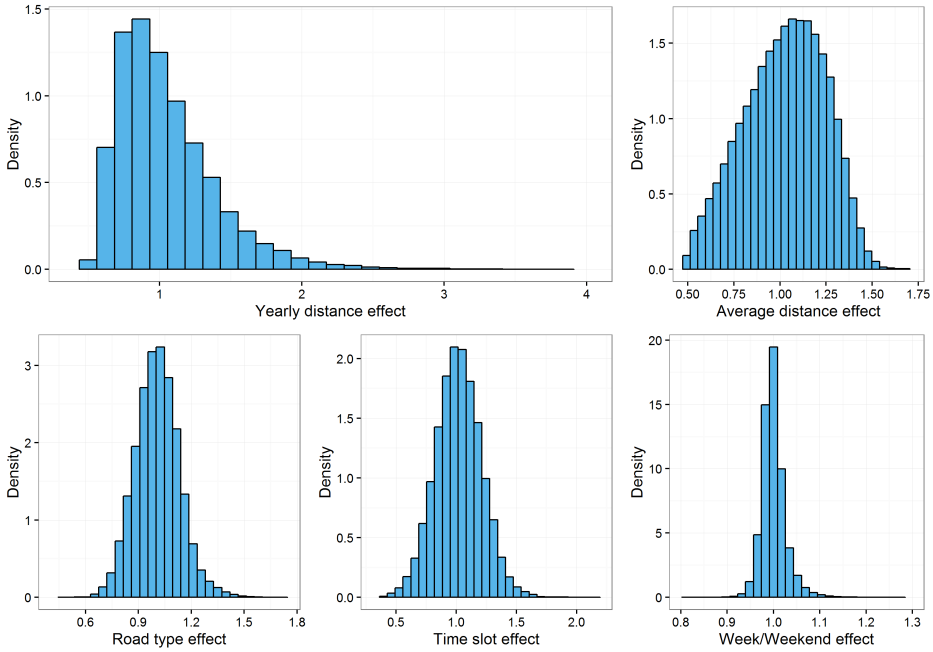

Figure D.8: Relative frequencies of the multiplicative response effects of the telematics model terms of the time-hybrid model.

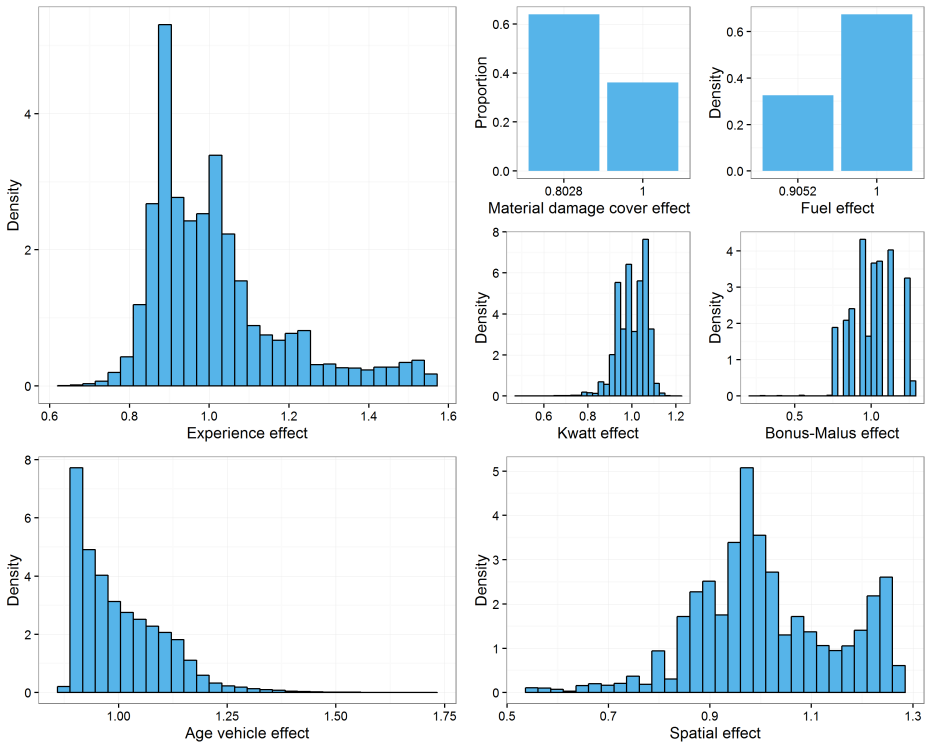

Figure D.9: Relative frequencies of the multiplicative response effects of the policy model terms of the meter-hybrid model.

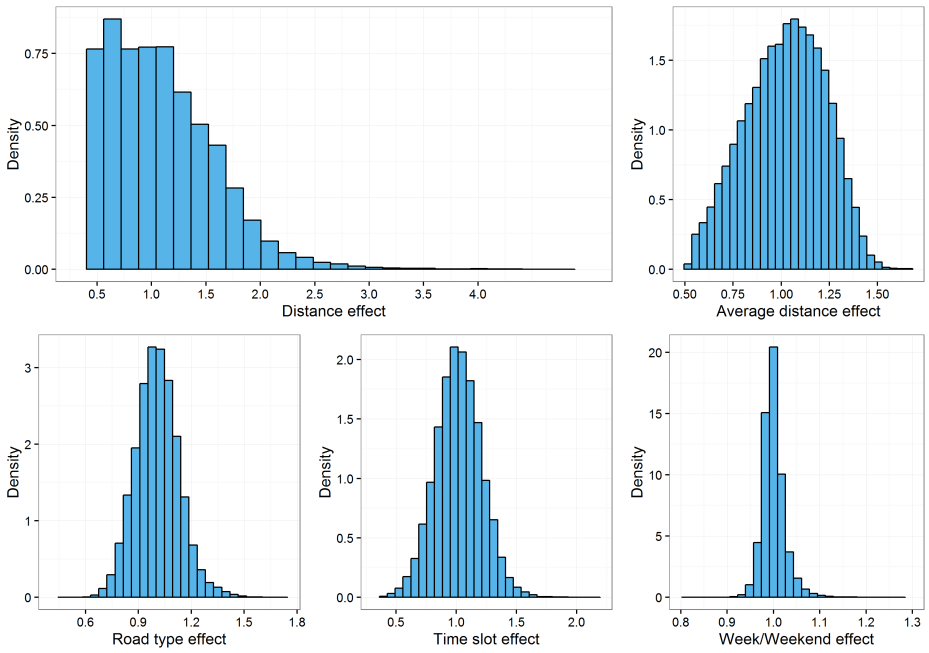

Figure D.10: Relative frequencies of the multiplicative response effects of the telematics model terms of the meter-hybrid model.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be