DETRA NOTE
2018-2

TELEMATICS

Usage-based motor insurance pricing with behavioral data

By Montserrat Guillen, Michel Denuit et Julien Trufin

MULTIVARIATE CREDIBILITY MODELING FOR USAGE-BASED MOTOR INSURANCE PRICING WITH BEHAVIOURAL DATA

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
UCLouvain
Louvain-la-Neuve, Belgium

Montserrat Guillen
Riskcenter, Department of Econometrics
Universitat de Barcelona
Barcelona, Spain

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium

Abstract

Pay-How-You-Drive (PHYD) or Usage-Based (UB) systems for automobile insurance provide actuaries with behavioural risk factors, such as the time of the day, average speeds and driving habits. These data are collected while the contract is in force thanks to telematic devices installed in the vehicle. They thus fall in the category of a posteriori information that becomes available after contract initiation. For this reason, they must be included in the actuarial pricing by means of credibility updating mechanisms instead of being included in the score as ordinary a priori observable features. We propose the use of multivariate mixed models to describe the joint dynamics of telematics data and claim frequencies. Future premiums, incorporating past experience can then be determined using the predictive distribution of claim characteristics given past history. This approach allows the actuary to deal with the variety of situations encountered in insurance practice, ranging from new drivers without telematics record to contracts with different seniority and drivers using their vehicle to different extent, generating varied volumes of telematics data.

Keywords: Risk classification, premium calculation, driving behavior, internet of things, count data models.

1 Introduction

The classical approach to motor insurance pricing can be summarized as follows (see Denuit et al. 2007, for an extensive presentation). The claim frequency is often the main target in actuarial pricing, both from an "a priori" perspective (supervised learning model including policyholder's characteristics as well as information about his or her vehicle and about the type of coverage selected, among others) and from an "a posteriori" perspective based on credibility models (mixed models linking past to future claims, inducing serial dependence with the help of random effects accounting for unexplained heterogeneity), sometimes simplified into a bonus-malus scale for commercial purposes.

Technological advances have now supplemented these classical risk factors with new ones, reflecting the policyholder's actual behaviour behind the wheel. Telematics is a branch of information technology that transmits data over long distances. Examples of telematics data include the global position system (GPS) data and the in-vehicle sensor data. The main data source for the aforementioned parameters are the automotive diagnostic systems (or OBD, for On-Board Diagnostics), installed in the vehicle and/or the smart phone held by drivers. We refer the reader to Boucher et al. (2013) and Tselentis et al. (2017) for reviews of current practices and emerging challenges in UB motor insurance pricing.

Telematics insurance data offer the opportunity to base actuarial pricing on policyholder's behaviour. With Pay-How-You-Drive (PHYD) or Usage-Based (UB) motor insurance, premium amounts are based on the total distance traveled, the type of road, the time of the day, average speeds and other driving habits. Thus, premiums are based directly on driver's behaviour. Several insurance companies have launched pilot projects to market new products with such innovative premiums, especially towards young, inexperienced drivers.

UB actuarial pricing ties the amount of insurance premium to the risk level associated with the actual driving behaviour of the policyholder. For instance, if increased mileage and speeding are associated with larger expected claim frequencies then they result in a higher insurance premium. This system of variable premiums offers an alternative to the current system of fixed insurance premiums exclusively based on proxies for risk such as age and gender, rather than on the actual driving behaviour of policyholders. UB pricing can integrate a multitude of risk factors, including distance travelled (annual mileage) and driving style (speeding or non-fluent driving i.e. frequent acceleration and deceleration, for instance), as well as other factors (e.g. time of driving).

Contrarily to standard risk factors, such as age, gender or place of residence, telematics data evolve over time in parallel to claim experience, progressively revealing the actual behaviour of the policyholder behind the wheel. The information contained in past telematics data differs between individuals. For newly licensed drivers, no record is available. For those observed over the past, telematics data are available for the time they were subject to the UB system which may vary among policyholders. Moreover, the reliability of the information is also heterogeneous. Indeed, telematics data are recorded while the policyholders are driving, and some of them regularly use their car (providing a rich information about their driving habits) whereas other ones use their car to a much lesser extent (resulting in limited volume of telematics data). In order to get the multivariate dynamics across insurance periods, past telematics data should not be included in the score like ordinary risk factors but must preferably be modelled jointly with claim experience. This is exactly the purpose of credibility models (also called mixed models, in statistics), except that here they apply to a random vector joining telematics data and claim experience. The approach proposed in this paper provides the actuary with a powerful alternative to the inclusion of behavioural traits as additional features in supervised learning (e.g. Baecke and Bocca, 2017, Ayuso et al., 2018, Verbelen et al., 2018, Jin et al., 2018) or the unsupervised classification of driving styles into a few categories that can then supplement traditional risk factors in supervised learning (e.g. Weidner et al., 2016, 2017, Wüthrich, 2017, Gao et al., 2018).

The approach proposed in this paper is illustrated by means of a real driving data recorded by GPS over three calendar years. These data relate to the portfolio of a Spanish insurance company offering UB motor insurance to young drivers. The information available is a panel that describes yearly claim numbers and the driving patterns for each driver. The driver's habits are summarized into three signals recorded thanks to telemetry: in addition to the number of kilometres driven in each year, the insurer collects information on the number of kilometres driven at night, the number of kilometres driven in an urban area, and the number of kilometres driven at excess speed. Annual mileage is considered as an exposure to risk and as such enters the multivariate models as an offset. The signals are treated as entire numbers, by rounding excess speed, nighttime driving and urban driving in natural units and a multivariate mixed Poisson model is used to describe their joint dynamics, together with yearly claim counts.

The remainder of this paper is organized as follows. Section 2 describes multivariate credibility models for random vectors joining signals and claim counts. This approach is applied to a real data set in Section 4, and the results are compared with those obtained according to the classical actuarial approach. Section 5 discusses the results and briefly concludes the paper.

2 Multivariate credibility model

2.1 Mixed Poisson model for annual claim frequencies

Let N_{it} be the number of claims reported by policyholder i during the period t, $t = 1, 2, \ldots, T_i$. Compared to classical actuarial studies dealing with annual periods, insurers using telematics data generally work with shorter time periods, like a quarter or a month.

At the beginning of each insurance period, the actuary has at his disposal some information about each policyholder. Resorting to standard regression (or supervised learning) machinery, this information is integrated into the prediction of the annual expected number of claims, or claim frequency. Specifically, define

$$\begin{aligned}
\boldsymbol{x}_{it} &= \text{features for policyholder } i,\ i = 1, \ldots, n, \\
& \text{during period } t,\ t = 1, 2, \ldots, T_i \\
d_{it} &= \text{exposure-to-risk, distance driven in kilometres} \\
s_{it} &= s(\boldsymbol{x}_{it}) \\
&= \text{score for policyholder } i \text{ in period } t.
\end{aligned}$$

Then,

$$\mathrm{E}[N_{it}] = \lambda_{it} = d_{it} \exp(s_{it}) = \exp\big(\ln d_{it} + s_{it}\big).$$

Adding $\ln d_{it}$ to the score s_{it} (i.e. treating this quantity as an offset) means that the insurer's price list is expressed per kilometre, and varies according to traditional risk features included in the vector $\boldsymbol{x}_{it}$. The score s_{it} can be calibrated by means of any Poisson regression technique, ranging from basic generalized linear models (GLM) to sophisticated machine learning algorithms.

A random effect is superposed to the prediction λ_{it} to recognize the residual heterogeneity of the portfolio. We refer to Denuit et al. (2007) for more details about this classical construction. In this paper, we assume that the residual effect of all unknown characteristics relating to policyholder i is represented by a random variable Θ_i. The annual numbers of claims $N_{i1}, N_{i2}, N_{i3}, \ldots$ are then assumed to be independent given Θ_i. The latent unobservable Θ_i characterizes the correlation structure of the claim counts N_{it} for each policyholder i. Specifically, the model is based on the following assumptions:

A1 given $\Theta_i = \theta$, the random variables N_{it}, $t = 1, 2, \ldots$, are independent and conform to the Poisson distribution with mean $\lambda_{it}\theta$, which is henceforth denoted as $N_{it} \sim \mathcal{P}oi(\lambda_{it}\theta)$, i.e.

$$\begin{aligned} \mathrm{P}[N_{i1} = k_1, \ldots N_{iT_i} = k_{T_i} | \Theta_i = \theta] &= \prod_{t=1}^{T_i} \mathrm{P}[N_{it} = k_t | \Theta_i = \theta] \\ &= \prod_{t=1}^{T_i} \left(\exp(-\lambda_{it}\theta) \frac{(\lambda_{it}\theta)^{k_t}}{k_t!} \right). \end{aligned}$$

A2 at the portfolio level, the sequences $(\Theta_i, N_{i1}, N_{i2}, \ldots)$ are assumed to be independent. Moreover, the Θ_i's are non-negative random variables with unit mean: $\mathrm{E}[\Theta_i] = 1$ for all i, which means that the a priori ratemaking is correct on average as

$$\mathrm{E}[N_{it}] = \mathrm{E}\big[\mathrm{E}[N_{it}|\Theta_i]\big] = \mathrm{E}[\lambda_{it}\Theta_i] = \lambda_{it}.$$

Mixed models generally assume that the random effects Θ_i obey the LogNormal distribution, which amounts to using a Poisson-LogNormal model for claim counts. This means that Normally distributed terms are added on the score scale (when the canonical log link function is used in the Poisson regression model, as assumed here). Formally, $\Theta_i = \exp(W_i)$ where W_i are independent and Normally distributed.

If longer panels are available then the static random effects Θ_i can be replaced with dynamic ones $\Theta_{i1}, \Theta_{i2}, \ldots$ which discount past observations according to their seniority. This is easily done by replacing W_i with a random sequence $W_{i1}, W_{i2}, \ldots$ obeying a Gaussian process whose covariance structure accounts for the memory effect (AR1, for instance).

2.2 Behavioural variable, or signal

In order to predict the number of claims N_{it} filed by policyholder i during period t, the insurer has a signal S_{it} at its disposal about the policyholder's behaviour behind the wheel during the same period. This unique signal summarizes all the information collected by means of telematic devices installed in the vehicle. For commercial purposes, it may be preferable to use a unique signal as premium updating formulas are more compact and easier to understand (in the next section, several signals will be used simultaneously).

To refine risk evaluation, we now combine past claims experience with the available signal. Hence, each contract is represented by the sequence

$$(\Theta_i, \Gamma_i, N_{i1}, S_{i1}, N_{i2}, S_{i2}, N_{i3}, S_{i3}, \ldots)$$

where

Θ_i accounts for hidden information influencing claim frequencies N_{it}

Γ_i reflects the quality of driving revealed by the observed signal S_{it}.

It is important to realize here that signals are also influenced by traditional risk factors included in $\boldsymbol{x}_{it}$ so that we need to account for this effect in model design. Here is a possible model specification in case of a Gaussian signal S_{it} (notice that even if the initial signal does not obey the Gaussian distribution, it can easily be transformed to meet approximately this condition): we supplement assumptions A1-A3 stated in Section 2.1 with

A4 Given Θ_i, the counts $N_{i1}, N_{i2}, \ldots$ are independent and independent of $\Gamma_i, S_{i1}, S_{i2}, \ldots$.

A5 Given Γ_i, the signals $S_{i1}, S_{i2}, \ldots$ are independent and independent of $\Theta_i, N_{i1}, N_{i2}, \ldots$, and

$$S_{it} = \nu_{it} + \Gamma_i + \mathcal{E}_{it}$$

where ν_{it} is the signal score based on classical features $\boldsymbol{x}_{it}$, Γ_i is Normally distributed and represents the additional information contained in the signal about claim frequencies, corrected for the effect of the features $\boldsymbol{x}_{it}$ whereas the Normally distributed error terms $\mathcal{E}_{it}$ represent the noise comprised in the observed signal S_{it} which do not reveal anything about claim counts. We also make the following assumptions about the dependence structure of these random variables

- The random variables $\Gamma_i, \mathcal{E}_{i1}, \mathcal{E}_{i2}, \ldots$ are mutually independent.
- The random variables $\mathcal{E}_{i1}, \mathcal{E}_{i2}, \ldots$ are independent from $(\Theta_i, N_{i1}, N_{i2}, N_{i3}, \ldots)$.

A6 Given Θ_i and Γ_i, all the observable random variables $N_{i1}, S_{i1}, N_{i2}, S_{i2}, \ldots$ are independent.

From assumptions A4-A6, we see that only the Γ_i component involved in the signal S_{it} is relevant to predict claim frequencies: we assume that the pair (Θ_i, Γ_i) is Normally distributed, and its covariance drives the corrections brought by signals in the evaluation of future expected number of claims.

Continuous signals are certainly appealing as many embarked devices produce real measures. Another approach consists in recording a number of events, or to round a continuous signal in multiples of a natural unit. This makes the mechanism more transparent, at the cost of a negligible loss of accuracy.

If the signal counts a number of events then A4-A6 above are replaced with

A4 Given Θ_i, the claim counts $N_{i1}, N_{i2}, \ldots$ are independent and independent of $\Gamma_i, S_{i1}, S_{i2}, \ldots$.

A5 Given Γ_i, the signal counts $S_{i1}, S_{i2}, \ldots$ are independent and independent of $\Theta_i, N_{i1}, N_{i2}, \ldots$, and

$$S_{it} \sim \mathcal{P}oi\big(d_{it}\exp(\nu_{it} + \Gamma_i)\big).$$

where ν_{it} is the signal score based on classical features $\boldsymbol{x}_{it}$ and Γ_i is Normally distributed and represents the additional information contained in the signal about claim frequencies. The noise present in the observed signal S_{it} is now represented by the Poisson error structure.

A6 Given Θ_i and Γ_i, all the observable random variables $N_{i1}, S_{i1}, N_{i2}, S_{i2}, \ldots$ are independent.

2.3 Multiple signals

In case several signals $S_{it}^{(j)}$, $j = 1, 2, \ldots$, are available, the insurer may either combine them into a single one and proceed as explained above. A natural approach would consist in using a linear combination of the signals for instance, and to work with the unique, composite signal $\sum_j \alpha_j S_{it}^{(j)}$ for appropriate weights α_j (determined so to maximize the correlation with the observed claim counts). Another possibility is to extend the model from the preceding section to the multivariate case by assuming a specific dynamics for each signal as explained next.

In case of multivariate Normally-distributed signals, we supplement assumptions A1-A3 with

A4 Given Θ_i, claim counts $N_{i1}, N_{i2}, \ldots$ are independent and independent of $\Gamma_i^{(j)}, S_{i1}^{(j)}, S_{i2}^{(j)}, \ldots$ for $j = 1, 2, \ldots$.

A5 Given $\Gamma_i^{(j)}$, the signals $S_{i1}^{(j)}, S_{i2}^{(j)}, \ldots$ are independent and independent of $\Theta_i, N_{i1}, N_{i2}, \ldots$, and admit the representation

$$S_{it}^{(j)} = \nu_{it}^{(j)} + \Gamma_i^{(j)} + \mathcal{E}_{it}^{(j)}$$

where $\nu_{it}^{(j)}$ is the score for the jth signal based on classical features $\boldsymbol{x}_{it}$, $\Gamma_i^{(j)}$ is Normally distributed and represents the additional information contained in the jth signal about claim frequencies,corrected for the effect of the features $\boldsymbol{x}_{it}$ whereas the Normally distributed error terms $\mathcal{E}_{it}^{(j)}$ represent the noise comprised in the observed signal which do not reveal anything about claim counts.

We also make the following assumptions about the dependence structure of these random variables:

- The random variables $\Gamma_i^{(j)}, \mathcal{E}_{i1}^{(j)}, \mathcal{E}_{i2}^{(j)}, \ldots$ are mutually independent.
- The random variables $\mathcal{E}_{i1}^{(j)}, \mathcal{E}_{i2}^{(j)}, \ldots$, $j = 1, 2, \ldots$, are mutually independent.
- The random variables $\mathcal{E}_{i1}^{(j)}, \mathcal{E}_{i2}^{(j)}, \ldots$ are independent from $(\Theta_i, N_{i1}, N_{i2}, N_{i3}, \ldots)$.
- The random vector $(\Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots)$ is multivariate Normally distributed.

A6 Given Θ_i and $\Gamma_i^{(j)}$, all the observable random variables $N_{i1}, S_{i1}^{(1)}, S_{i1}^{(2)}, \ldots, N_{i2}, S_{i2}^{(1)}, S_{i2}^{(2)}, \ldots$ are independent.

We also assume that the random vector $(\Theta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots)$ is Normally distributed. Its covariance structure drives the corrections induced by the signals on future expected claim counts. We acknowledge here that the multivariate Normal assumption may appear to be restrictive in some applications because it constrains the dependence structure (prohibiting tail dependence, for instance). Other multivariate distributions, such as Elliptical ones can be useful to model the dependency of the signals, and a copula construction can be employed to this end.

If the signals consist in counts of different events then assumptions A1-A3 are supplemented with

A4 Given Θ_i, claim counts $N_{i1}, N_{i2}, \ldots$ are independent and independent of $\Gamma_i^{(j)}, S_{i1}^{(j)}, S_{i2}^{(j)}, \ldots$ for $j = 1, 2, \ldots$.

A5 Given $\Gamma_i^{(j)}$, the signal counts $S_{i1}^{(j)}, S_{i2}^{(j)}, \ldots$ are independent and independent of $\Theta_i, N_{i1}, N_{i2}, \ldots$, and

$$S_{it}^{(j)} \sim \mathcal{P}oi\big(d_{it} \exp(\nu_{it}^{(j)} + \Gamma_i^{(j)})\big)$$

where $\nu_{it}^{(j)}$ is the score for the jth signal based on classical features $\boldsymbol{x}_{it}$ and $\Gamma_i^{(j)}$ is Normally distributed and represents the additional information contained in the jth signal about claim frequencies, corrected for the effect of the features $\boldsymbol{x}_{it}$.

A6 Given Θ_i and $\Gamma_i^{(j)}$, all the observable random variables $N_{i1}, S_{i1}^{(1)}, S_{i1}^{(2)}, \ldots, N_{i2}, S_{i2}^{(1)}, S_{i2}^{(2)}, \ldots$ are independent.

Of course, the insurer could use a blend of continuous and integer signals so that many variants to the models proposed above can be envisaged.

Maximum likelihood estimation of generalized linear mixed models (GLMM) for panel data is implemented in `R`, where both fixed effects and random effects are specified via the model. The results presented in the case study were obtained with the `lme4` package.

3 Case study

3.1 Presentation of the data set

Our research is based on real driving data recorded by GPS, collected by a Spanish insurance company within the framework of a new form of insurance cover. Under such policies, motor insurance premiums are determined by taking into account not only the traditional risk factors but also the number of kilometres driven in a given period of time as well as information on the number of kilometers driven at night, the number of kilometres driven in an urban area, and the number of kilometres driven at excess speed. The information available is a panel that describes yearly records on the number of claims and the driving patterns for each driver measured thanks to telemetry.

Excess speed, night-time driving and urban driving are considered to be signals of the type of driving habits or skills. We treat these signals as entire numbers, by rounding excess speed, night-time driving and urban driving in natural units of 500 kilometres. Specifically, the three signals at our disposal are as follows:

$$\begin{aligned}
S_{it}^{(1)} &= \text{distance travelled in the night (in 500 Kms)} \\
S_{it}^{(2)} &= \text{distance driven above the speed limit (in 500 Kms)} \\
S_{it}^{(3)} &= \text{distance travelled in urban zones (in 500 Kms).}
\end{aligned}$$

The joint dynamics of the number of claims N_{it} filed by policyholder i during period t and the three signals $S_{it}^{(j)}$, $j = 1, 2, 3$, will be exploited to predict the future number of claims.

Notice that these are not compositional data in the sense of Verbelen et al. (2018) because as opposed to raw counts, they model percent exposure and they have to cope with the restriction that percentages need to add up to 100% at the policy holder level. Data on the total distance driven per year (in kilometres) is considered as an exposure to risk and as such enters our models as an offset. To avoid large dispersion, distance driven is expressed in hundreds of kilometres.

Let us briefly comment on the choice of these three signals. Night-time driving is usually associated to more accidents than day-time, especially at young ages (see, for instance, Williams, 1985), and the first signal captures this effect. As pointed out by Bolderdijk et al. (2011), vehicle speed is commonly considered as the major determinant of crash risk for young adults. Specifically, these authors demonstrated that reducing the amount of time spent above the speed limit, holds the potential of dramatically reducing accidents. This is exactly the information captured by the second signal, time being here measured by the actual distance driven above the speed limits (integrating the total distance travelled by means of offset). Notice that the signal excess speed records the number of kilometres travelled at a speed in excess of the posted limit. However we do not have enough information to include the amount of excess, so we cannot distinguish between a driver who drives 10% faster or 20% than the posted limit. Finally, we note that urban areas are often congested and crash risk is higher there than in sub-urban or rural zones, because of heavy traffic. The third signal records the distance travelled in the accident-prone urban areas.

3.2 Descriptive statistics

The sample is made up of 2,494 insured drivers followed over the three calendar years 2009-2011. The observation period ends on December 31, 2011. The mean age of all drivers in the sample in 2009 is 25.17 years (standard deviation 2.44). In the participating insurance company, the policies that involve collecting telematics information are only offered to young drivers (the maximum age in the sample being 30 years). Our sample comprised 51.60% of male drivers and 48.40% of female drivers.

In Table 3.1 we present descriptive statistics for telematics data observed in the sample for each year. Many contracts were discontinued because customers preferred to chose other forms of insurance payments in 2011 and this is the reason why distance driven dropped dramatically the last year. However, since we have distance driven as an offset in our model, we predict the expected number of claims per mileage, and therefore this is automatically corrected in the analysis.

	year: 2009	year: 2010	year: 2011
Total distance			
min	1.06	80.61	17.54
mean	14,062.39	13,475.16	7,170.96
median	12,777.59	12,070.94	6,404.03
(IQR)	(8,342.37, 18,590.10)	(7,934.84, 17,662.90)	(4,064.64, 9,375.69)
max	53,412.06	56,360.86	36,101.56
Km night			
min	0.00	0.00	0.00
mean	923.24	1,011.26	527.73
median	579.00	611.00	298.00
(IQR)	(235.25, 1,202.50)	(242.00, 1,290.00)	(112.00, 698.75)
max	10,989.00	11,494.00	6,526.00
Km speed			
min	0.00	0.00	0.00
mean	1,564.76	1,547.05	560.81
median	834.50	769.00	258.50
(IQR)	(343.25, 1,848.75)	(324.25, 1,879.25)	(106.00, 632.75)
max	18,160.00	23,500.00	11,836.00
Km urban			
min	1.00	45.00	0.00
mean	3,122.52	2,871.50	1,483.40
median	2,803.00	2,590.50	1,345.50
(IQR)	(1,903.00, 3,947.25)	(1,755.00, 3,637.00)	(875.00, 1,923.00)
max	15,519.00	14,732.00	6,462.00

Table 3.1: Sample statistics for raw telematic information by year ($n = 2,494$).

Our yearly responses are the number of claims, and then the number of count units of excess speed, night-time driving and urban driving (rounded in 500s kilometres). Our measure of exposure-to-risk is the distance driven measured as a continuous variable in 100s kilometres. Figure 3.1 shows a four histogram presentation of the raw telematics data in 2009.

3.3 Association between signals and claim counts

We focus specifically on the three signals $S_{it}^{(j)}$ because we expect a clear association between claims and excess speed, night driving and urban driving. We treat total driving distance as a total exposure offset. There is an extensive literature on how all these factors are associated to claiming. Ayuso et al. (2016, 2018) and showed that information on speed excess, night-time driving and urban driving improves the prediction of the number of claims, compared to classical models not using telematics information. Guillen et al. (2018) provide an extended overview on how accumulated distance driven shows evidence that drivers improve their skills, a phenomenon that is known as the "learning effect".

All this previous knowledge is the reason why we focus specifically on variables that reflect the driving habits, such as excess speed, night driving and urban driving, and for which we expect a clear association with the number claims as well as distance driven. Let us now investigate the strength of this association on our data set. Figure 3.2 shows a correlation

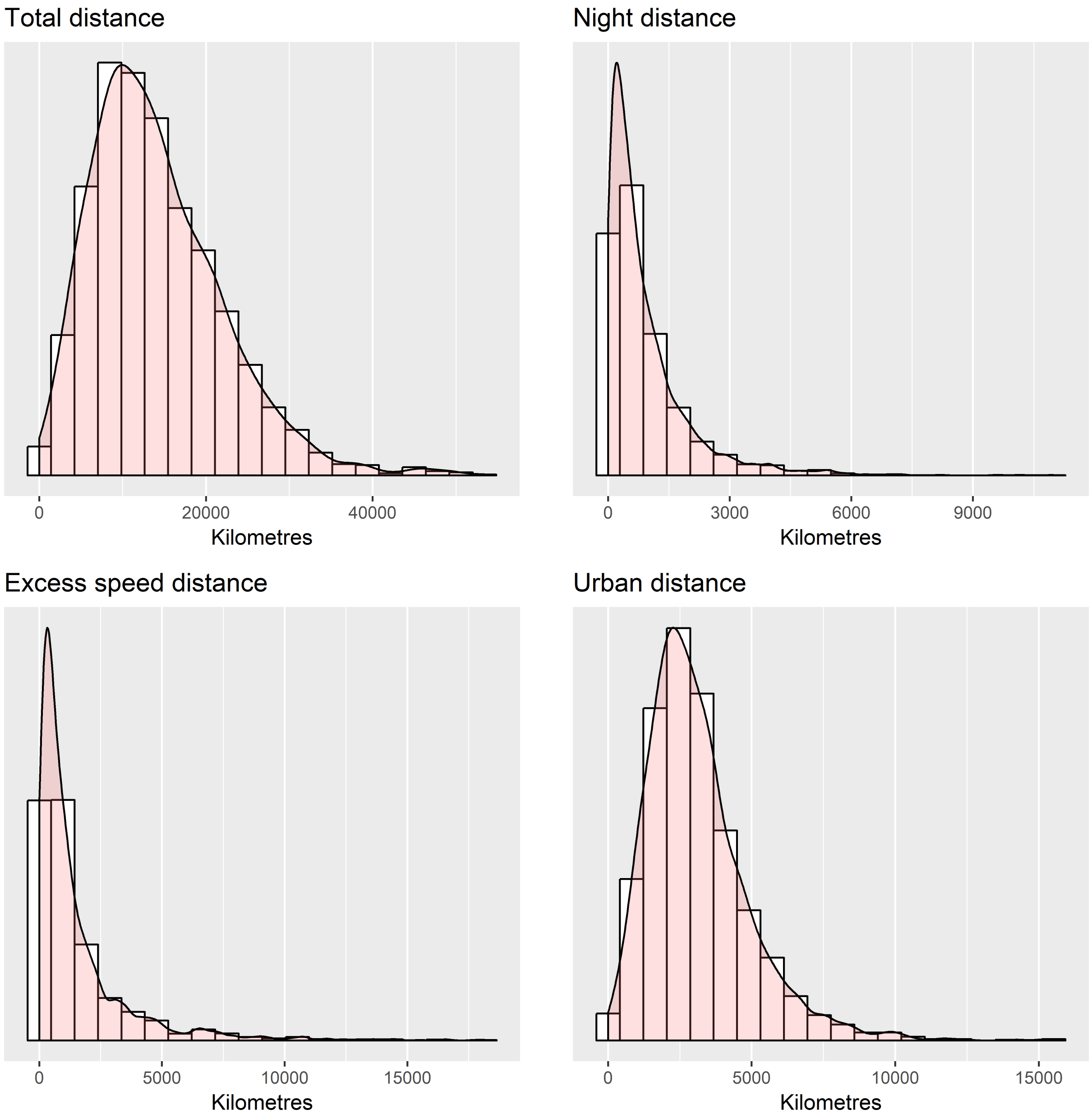


Figure 3.1: Histograms of telematic information recorded in 2009.

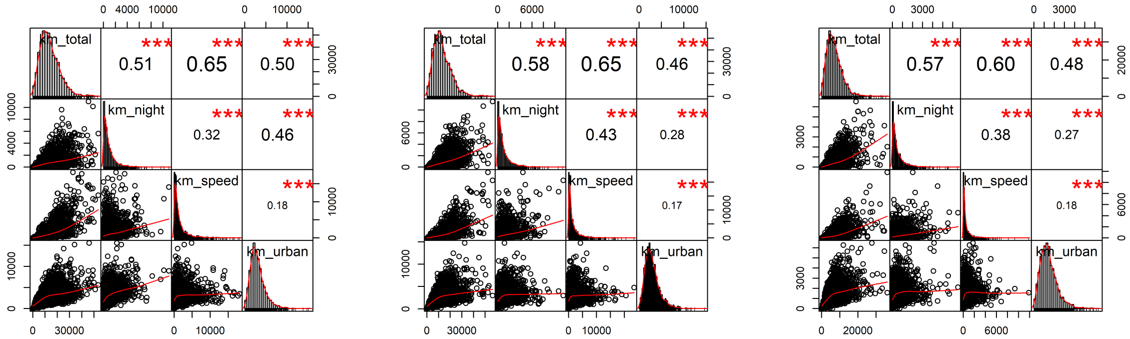


Figure 3.2: Correlation matrix of telematic information recorded in 2009, 2010 and 2011.

between the distance and the three raw indicators (speed, night-time, urban) in 2009, 2010 and 2011. Just by illustration in Figure 3.3 we also show the correlation between distance driven in the three observed years. As expected, distance driven correlates with the signals (Figure 3.2) and between consecutive years (Figure 3.3). Notice that no correction has been made for standard risk factors at this stage so that the correlation may only be apparent, being generated by the confounding effects of the standard risk factors comprised in $\boldsymbol{x}_{it}$.

The association between signals and claim counts can also be assessed by fitting a GAM model for the number of claims with the signals treated as explanatory variables. Table 3.2 shows that a GAM model for the number of claims where night, speed and urban driving are explanatory variables and measured as discrete counts in units of 500s kilometres. We can see there is a significant effect of these three signals on the expected frequency of claims.

Table 3.3 presents the counts information for the three years and the four counts once the signals of speed, night-time and urban are transformed in discrete counts in units of 500s kilometers. We can see there that the majority of claims counts as well as night-time and speed signals concentrate in low-frequency cells, whereas the counts of the signal urban is located in a higher frequency level. The information in Table 3.3 indicates that 2,004 drivers did not claim any accident in 2009 (2,038 and 2,091 in 2010 and 2011, respectively). In 2009, one policyholder claimed as much as 6 accidents, while the maximum number of claims was 4 in 2010 and 2011. A few policy holders recorded high levels of speed limit excess in 2009 and even a bit more in 2010.

3.4 Fitted models

The Poisson-LogNormal model for claim counts was fitted using the `glmer` function of the `R` package `lme4` which performs Poisson regression with random effects.

In the univariate approach (i.e. considering claim counts, or each signal, in isolation), the random effects are included by means of the component `(1|id)` where `id` denotes the policy identifier (allowing to track the same contract over time) entering model formula. In this case, only past claim experience is used to update the expected number of claims in future years. The multivariate model consider claim counts and the three signals simultaneously. We fit the multivariate model at once following the approach proposed by Faraway (2016,

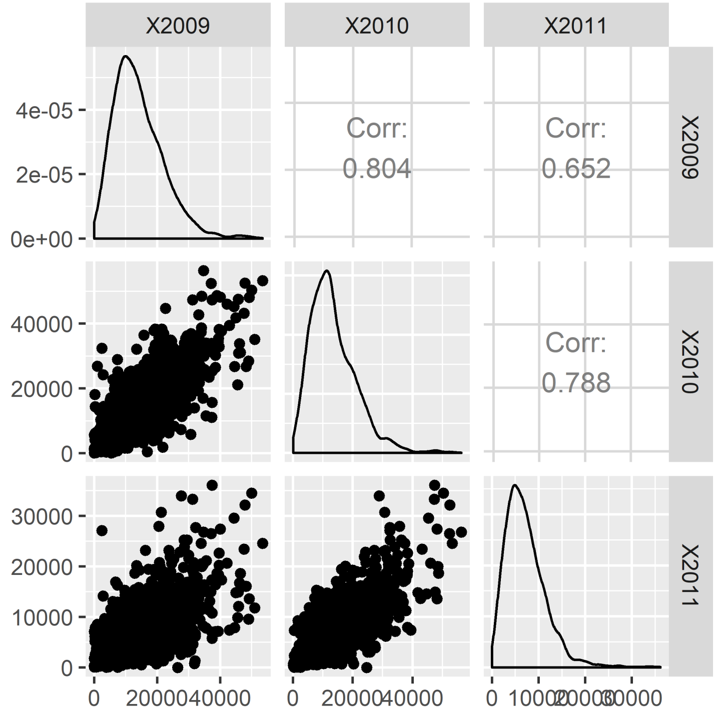


Figure 3.3: Correlation matrix of distance driven in 2009, 2010 and 2011.

	Model 1
(Intercept)	-6.156^{***}
	(0.036)
men	-0.006
	(0.050)
EDF: s(age)	0.991^{***}
	(9.000)
EDF: s(km_nightcount500)	1.892^{***}
	(9.000)
EDF: s(km_speedcount500)	4.706^{***}
	(9.000)
EDF: s(km_urbcount500)	2.827^{***}
	(9.000)
AIC	9,124.622
BIC	9,227.708
Log Likelihood	-4,547.415
Num. obs.	7,482
Num. smooth terms	4

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.2: GAM model for the number of claims 2009-2011.

Table 3.3: Counts of claims and driving signals (expressed in 500s kilometres) in 2009, 2010 and 2011.

	Year: 2009				Year: 2010				Year: 2011			
	Claims	Night	Speed	Urban	Claims	Night	Speed	Urban	Claims	Night	Speed	Urban
0	2004	652	461	18	2038	640	473	17	2091	1131	1227	59
1	370	825	705	74	350	793	755	97	318	799	732	415
2	95	428	422	181	92	409	371	199	71	298	242	642
3	18	229	234	252	11	230	231	306	11	126	113	616
4	4	133	175	343	3	128	156	381	3	69	60	368
5	2	73	102	339		91	109	367		27	41	191
6	1	49	72	309		60	83	299		19	22	94
7		28	66	272		39	67	256		11	25	56
8		27	43	183		31	38	169		6	9	27
9		14	40	147		27	38	123		3	7	10
10		10	33	102		13	32	76			2	7
11		12	15	76		6	27	50		4	5	5
12		4	12	52		6	14	53			3	2
13		2	25	45		8	21	30		1	2	2
14		3	17	20		3	12	19			1	
15		1	11	25		4	4	16				
16		1	8	15		2	11	10			1	
19		1	4	5			4	1			1	
20		1	3	9			3					
22		1	5	1			4	1				
17			6	11		1	9	5				
18			8	5		1	4	12				
21			6	3		1	3	3				
23			2	2		1	5	2				
24			2	1			1				1	
25			3	1			1	1				
26			1				5					
27			3				1					
28			1	1								
29			3				1	1				
31			1	1			1					
32			1				2					
33			2				3					
35			1				1					
36			1									
30				1								
34							1					
38							1					
41							1					
47							1					

Section 9.3). The idea is to define signal identifiers by means of a categorical feature with three levels, `S1`, `S2`, and `S3`, say, treated as fixed effects and to introduce an interaction between the signals and the other fixed effects, as well as hierarchical random effects for signals and for the insured within signal.

To ensure numerical stability of the optimization algorithms, policyholder's age has been rescaled (divided by 100). Gender is coded as 1 for male drivers and as 0 for female drivers. Also, different units have been tested for the three signals (in 100 and 1,000 kilometers, without affecting the results).

Table 3.4 presents the results of the univariate and the multivariate counts models (estimated with the three-year panel 2009-2011). The difference between the univariate approach and the multivariate approach is that the former only considers one of the signals at a time and it completely ignores the association between them. However, the reason to introduce a multivariate framework is that, for instance a claim in 2009 can influence the driver in such a way that he or she drives more carefully in 2010 in terms of excess speed and even in the total distance. This phenomenon had been noted before (see Guillen and Pérez-Marín, 2018) but it had not been studied in the way it is done here.

In the univariate modelling, the four responses N_{it}, $S_{it}^{(1)}$, $S_{it}^{(2)}$, and $S_{it}^{(3)}$ are considered to be mutually independent (but serial dependence for fixed i is taken into account in all four cases): precisely, given independent, centred, Normally-distributed random variables $\Theta_i^{\perp}, \Gamma_i^{(1),\perp}, \Gamma_i^{(2),\perp}, \Gamma_i^{(3),\perp}$, the responses are Poisson distributed with respective means

$$\begin{aligned}
\ln \mathrm{E}[N_{it}|\Theta_i^{\perp}] &= \ln(d_{it}) - 4.93 - 5.95\mathrm{age}_i - 0.09\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \ln \Theta_i^{\perp} \\
\ln \mathrm{E}[S_{it}^{(1)}|\Gamma_i^{(1),\perp}] &= \ln(d_{it}) - 4.32 - 1.33\mathrm{age}_i + 0.38\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(1),\perp} \\
\ln \mathrm{E}[S_{it}^{(2)}|\Gamma_i^{(2),\perp}] &= \ln(d_{it}) - 3.27 - 4.27\mathrm{age}_i + 0.22\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(2),\perp} \\
\ln \mathrm{E}[S_{it}^{(3)}|\Gamma_i^{(3),\perp}] &= \ln(d_{it}) - 2.30 - 3.29\mathrm{age}_i + 0.03\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(3),\perp}.
\end{aligned}$$

The prediction for the future expected number of claims is based on claim dynamics only, and integrates individual past claims frequencies. These predictions are obtained using large-sample results such as formula (3.21) on page 151 of Wood (2017) giving the a posteriori, or predictive distribution of the estimated regression coefficients and random effects (used in the `predict` function of `glmer`).

The main conclusion is that, in the univariate models and even when we control for the driving behaviour signals, age has an overall effect that is negative, meaning that the older the driver the less claims are expected. Here we chose a linear effect because the interval of ages is small for this sample of young drivers and we could not find a non-linear association. We also tried interactions between age and gender, but again we could not find significant cross-effects.

The joint dynamics of the number of claims N_{it} filed by policyholder i during period t and the three signals $S_{it}^{(1)}$, $S_{it}^{(2)}$ and $S_{it}^{(3)}$ is as follows. In the multivariate modelling, the correlation structure and the serial dependence are both taken into account for the four responses N_{it}, $S_{it}^{(1)}$, $S_{it}^{(2)}$, and $S_{it}^{(3)}$: precisely, given centred, multivariate Normally-distributed

	Multivariate	Univariate Models			
	Model	Night	Speed	Urban	Claims
(Intercept)	-4.540^{***}	-4.318^{***}	-3.266^{***}	-2.304^{***}	-4.932^{***}
	(0.273)	(0.145)	(0.159)	(0.095)	(0.319)
Night	0.179				
	(0.310)				
Speed	0.995^{**}				
	(0.305)				
Urban	2.231^{***}				
	(0.294)				
Age	-6.813^{***}	-1.330^{*}	-4.274^{***}	-3.294^{***}	-5.950^{***}
	(1.054)	(0.550)	(0.610)	(0.362)	(1.215)
Men (vs. Women)	-0.094	0.379^{***}	0.217^{***}	0.032	-0.095
	(0.055)	(0.030)	(0.036)	(0.021)	(0.063)
Night:age	5.587^{***}				
	(1.196)				
Speed:age	3.951^{***}				
	(1.174)				
Urban:age	3.507^{**}				
	(1.135)				
Night:men	0.477^{***}				
	(0.063)				
Speed:men	0.296^{***}				
	(0.062)				
Urban:men	0.124^{*}				
	(0.060)				
AIC	85254.002	21513.022	23097.651	31167.398	9085.767
BIC	85361.987	21540.703	23125.332	31195.079	9113.448
Log Likelihood	-42614.001	-10752.511	-11544.826	-15579.699	-4538.884
Num. obs.	29,928	7,482	7,482	7,482	7,482
Num. groups: signalName:idd	9976				
Var: signalName:idd (Intercept)	0.314				
Num. groups: idd		2,494	2,494	2,494	2,494
Var: idd (Intercept)		0.272	0.547	0.189	0.746

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.4: Model results for panels data on claims and driving count signals, 2009-2011.

random variables $\Theta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \Gamma_i^{(3)}$, the responses are Poisson distributed with respective means

$$\ln \mathrm{E}[N_{it}|\Theta_i] = \ln(d_{it}) - 4.54 - 6.81\mathrm{age}_i - 0.09\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \ln \Theta_i$$

$$\begin{aligned}\ln \mathrm{E}[S_{it}^{(1)}|\Gamma_i^{(1)}] &= \ln(d_{it}) + (-4.54 + 0.18) + (-6.81 + 5.59)\mathrm{age}_i \\ &\quad +(-0.09 + 0.48)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(1)} \\ &= \ln(d_{it}) - 4.36 - 1.22\mathrm{age}_i + 0.39\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(1)}\end{aligned}$$

$$\begin{aligned}\ln \mathrm{E}[S_{it}^{(2)}|\Gamma_i^{(2)}] &= \ln(d_{it}) + (-4.54 + 0.99) + (-6.81 + 3.95)\mathrm{age}_i \\ &\quad +(-0.09 + 0.30)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(2)} \\ &= \ln(d_{it}) - 3.55 - 2.86\mathrm{age}_i + 0.21\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(2)}\end{aligned}$$

$$\begin{aligned}\ln \mathrm{E}[S_{it}^{(3)}|\Gamma_i^{(3)}] &= \ln(d_{it}) + (-4.54 + 2.23) + (-6.81 + 3.51)\mathrm{age}_i \\ &\quad +(-0.09 + 0.12)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(3)} \\ &= \ln(d_{it}) - 2.31 - 3.30\mathrm{age}_i + 0.03\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(3)}.\end{aligned}$$

Compared to the univariate approach, we see that the intercept and gender effect remain almost unaffected in the multivariate model. The coefficient of age becomes even more negative. The effect of age on the score scale remains negative in the multivariate model. The prediction for the future expected number of claims is now based on both claim and signal dynamics, and integrates individual past claims frequencies and signal values.

Table 3.5 and 3.6 present the estimated covariance matrix of individual random effects for signal counts: Night, Speed, Urban and the number of claims. Correlations appear to be significant between night time driving and the number of claims (positive) and also between urban and the number of claims (negative). These results can be interpreted as follows: once information on night driving is known and included in the prediction of the number of claims, there are still other signals that impact positively on the number of claims, meaning that those drivers that drive more in the night will still be more risky than the others. However, the negative covariance in the individual random effects between urban driving and number of claims means that once urban driving is known then the rest of characteristics induces a lower number of claims, meaning that those drivers that drive more in urban areas are less risky drivers than the others.

3.5 Predictive power

Table 3.7 shows the predictive performance when analysing the number of claims for the univariate versus the multivariate method when the model is estimated for two years (2009 and 2010) and then the third year (2011) is predicted. To simplify the analysis of observed versus predicted counts, in this performance matrix, we have aggregated all frequencies that are equal to 1 or larger. The overall success rate for the univariate model is 77.5% (i.e. (1866+67)/2494) and it is 79.6% (i.e. (1945+39)/2494) for the multivariate model. Note

	Night	Speed	Urban	Claims
Night	0.064	0.002	−0.009	0.029
Speed	0.002	0.151	0.003	0.006
Urban	−0.009	0.003	0.255	−0.066
Claims	0.029	0.006	−0.066	0.169

Table 3.5: Estimated covariance matrix of random effects in the multivariate model for signals (Night, Speed, Urban) and the number of claims.

	Night	Speed	Urban	Claims
Night	1.000			
Speed	0.024	1.000		
Urban	-0.067^{**}	0.016	1.000	
Claims	0.275^{***}	0.035	-0.318^{***}	1.000

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.6: Pearson correlations of individual random effects in the multivariate model for signals (Night, Speed, Urban) and the number of claims.

that the prediction is equal to 0 if the predicted number of claims is below the fixed threshold (mean observed frequency) and it is 1, otherwise.

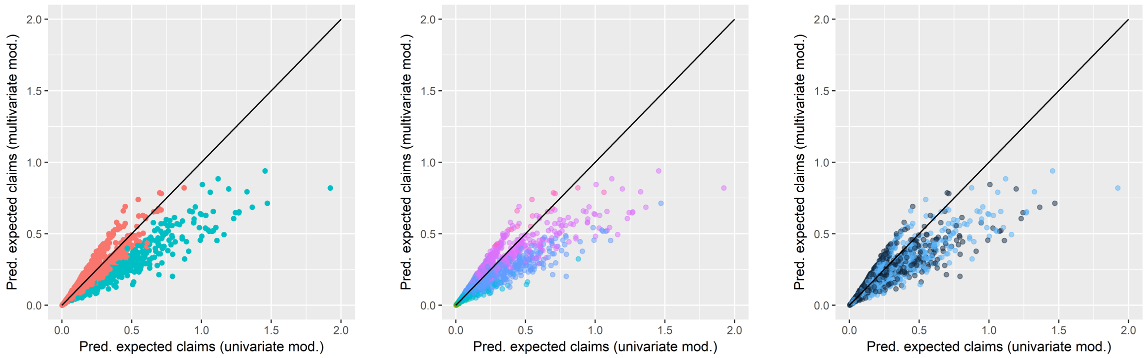
Figure 3.4 shows the difference between the prediction for 2011 of the univariate model (x-axis) and the multivariate model (y-axis). We can see that the multivariate model shows a marked difference for those insured who had an accident in 2009 and 2010 (left plot) and for those that drive more distance (centre plot). However, there is no clear pattern between the predictions of the number of claims for men and women under the univariate or the multivariate approach (right plot) possibly because their driving patterns also differ in other signals like night driving.

By looking at Figure 3.4, in general we conclude that the multivariate model predicts less expected claims than their univariate counterpart. The dots on the left-hand side of Figure 3.4 that are shown in blue, represent those policyholders who reported at least one claim either in 2009 or in 2010 or both. Those drivers who had claims in 2009 and 2010

	Predicted			
	Univariate model		Multivariate model	
Observed	0	1 or more	0	1 or more
0	1866	225	1945	146
1 or more	336	67	364	39

The threshold is set at the mean observed frequency

Table 3.7: Predicted versus observed number of claims for 2011 using model estimates 2009-2010.



Red dots indicate policy holders who did not suffer any claim in 2009 and 2010, blue dots indicate policy holders who suffered at least one claim

Purple dots indicate intensive users (longer accumulated total distance driven in 2011), blue and greenish dots indicate policy holders who drove less distance in 2011

Men (light dots) vs. women (dark dots)

Figure 3.4: Comparison between the predictions in 2011 for the multivariate model (y-axis) and the univariate model (x-axis) for the expected number of claims in 2011 using model estimates 2009-2010.

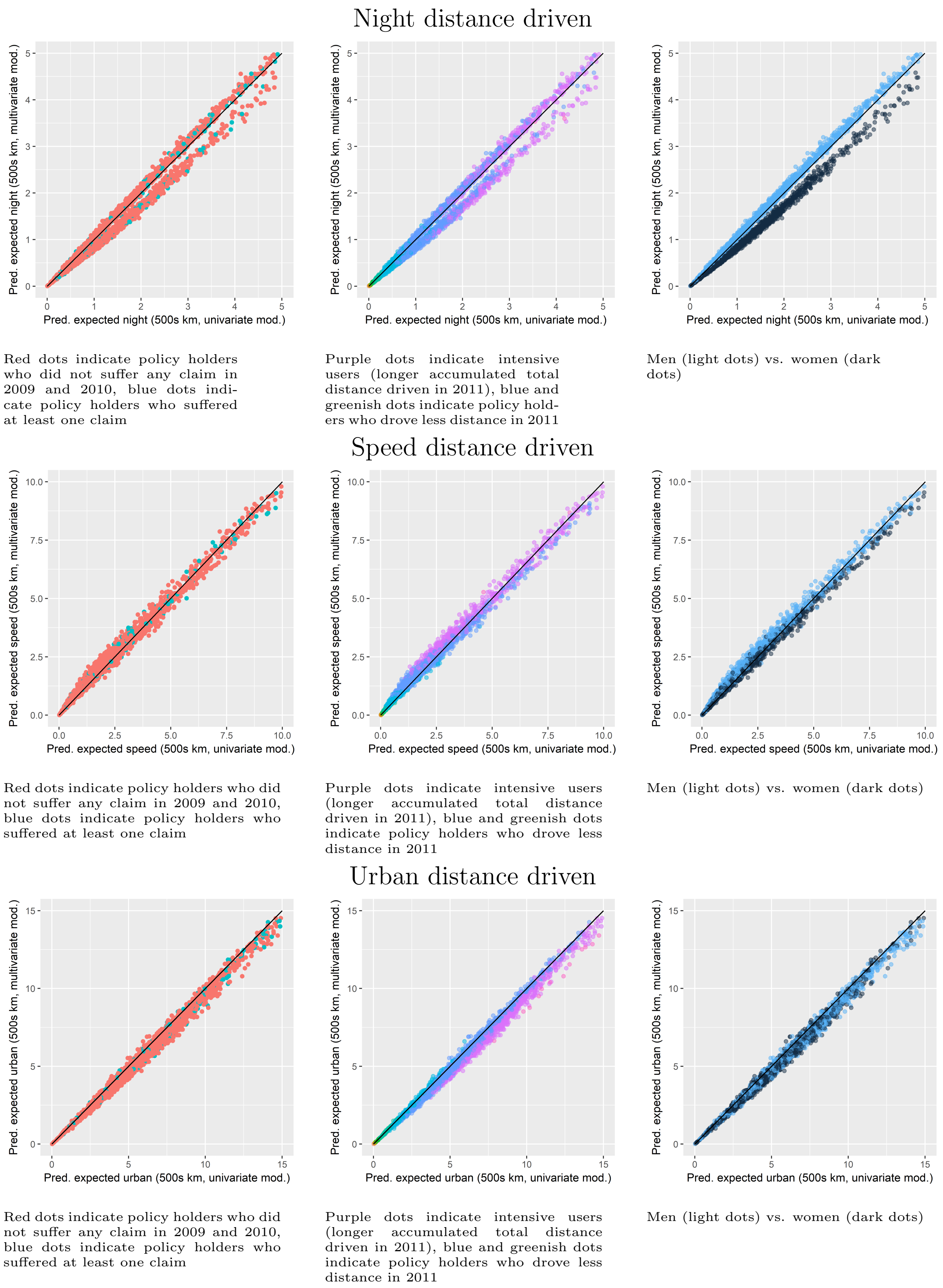
(blue dots) have a lower expected estimated number of claims with the multivariate signal panel Poisson model than with the univariate panel Poisson model. It seems that having a claim is not as important in the multivariate framework, as it is in the univariate signal study. In the multivariate model all the other signals are also considered. The central plot shows that those policyholders located in the upper part of the cloud correspond to those that drove higher distances, which means that the multivariate approach predicts for them a higher number of claims than for those who drove less. The distance driven in 2011 was used for predicting both the univariate and the multivariate estimates, we only present in Figure 3.4 the predictions for the expected frequency of claims. We note that the positive and significant association between distance driven and signals means that the longer the distance travelled, the higher is the count distance in the night, in excess speed and in urban areas. Altogether these three signals are associated with more claims, and so, the multivariate model predicts more claims for these observations. This phenomenon is ignored by the univariate model.

The comparison between predictions in the signals other than the number of claims is shown in Figure 3.5. Interestingly, for the night time driving the multivariate model and the univariate model have different predictions for men and women (see right plot), which suggests a different attitude of both gender regarding driving in the night. This phenomenon had already been found by other authors before.

4 Discussion

The approach proposed in this paper recognizes the a posteriori nature of telematics data and their heterogeneity among insured drivers. The multivariate credibility model developed in the case study captures the association between signals and claim counts, allowing the actuary to refine risk evaluations based on past history.

Bonus-malus scales, which have now become a popular experience rating scheme in motor



Night distance driven

Red dots indicate policy holders who did not suffer any claim in 2009 and 2010, blue dots indicate policy holders who suffered at least one claim

Purple dots indicate intensive users (longer accumulated total distance driven in 2011), blue and greenish dots indicate policy holders who drove less distance in 2011

Men (light dots) vs. women (dark dots)

Speed distance driven

Red dots indicate policy holders who did not suffer any claim in 2009 and 2010, blue dots indicate policy holders who suffered at least one claim

Purple dots indicate intensive users (longer accumulated total distance driven in 2011), blue and greenish dots indicate policy holders who drove less distance in 2011

Men (light dots) vs. women (dark dots)

Urban distance driven

Red dots indicate policy holders who did not suffer any claim in 2009 and 2010, blue dots indicate policy holders who suffered at least one claim

Purple dots indicate intensive users (longer accumulated total distance driven in 2011), blue and greenish dots indicate policy holders who drove less distance in 2011

Men (light dots) vs. women (dark dots)

Figure 3.5: Comparison between the predictions in 2011 for the multivariate model (y-axis) and the univariate model (x-axis) for the three signals (night, speed and urban driving) in 2011.

insurance, have been proposed to insured drivers in the 1960s. On a voluntary basis, attracting the best drivers, before becoming compulsory. We refer the reader to Lemaire (1995) for the history of this a posteriori pricing mechanism. The UB motor insurance premium systems could develop similarly.

Considering adverse selection in the vein of Rotschild and Stiglitz, individuals partly reveal their underlying risk through the contract they chose, a fact that has to be taken into account when setting an adequate tariff structure. In the presence of unobservable heterogeneity, low risk insurance applicants have interest to signal their quality, by selecting UB insurance cover for instance. As pointed out by Tselentis et al. (2017), a gradual global transition towards UB insurance can therefore be envisaged. Low-risk drivers (low-mileage, less risky drivers etc.) will first opt out of traditional insurance in favour of insurance policies with UB premium calculation. Consequently, behavioural aspects of driving are likely to be incorporated in insurance models in order to contribute towards current trends of personalized vehicle insurance.

As claims remain rare events, the standard credibility models appear to be relatively inefficient in personal insurance lines. They are even sometimes perceived as unfair by insured drivers. On the contrary, behavioural characteristics are recorded on a continuous basis, and remain for the most part under drivers' control. Premium amounts are differentiated to reflect safety, by charging higher fees for unsafe road categories and night-time driving, for instance. Moreover, insured drivers can adapt their driving style to make the amount of UB insurance premium decrease. In that respect, they appear to be superior both from an actuarial point of view (more accurate risk evaluation) and societal goal (promoting safer driving habits and decreasing traffic congestion). In this way, UB actuarial pricing also serves as a mechanism to raise drivers' awareness and improve their driving behaviour.

Acknowledgements

Michel Denuit and Julien Trufin gratefully acknowledge the financial support of the AXA Research Fund through the JRI project "Actuarial dynamic approach of customer in P&C". They warmly thank Stanislas Roth for interesting discussions on the topic dealt with in this paper. Montserrat Guillen thanks the Spanish Ministry of Economy and Competitiveness for support under FEDER grant ECO2016-76203-C2-2-P. All authors declare no conflict of interest as no sponsor has been involved in the implementation and conclusions of the research.

References

- Ayuso, M., Guillen, M., 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 and Prevention 73, 125-131.

- Ayuso, M., Guillen, M., 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, 10.

- Ayuso, M., Guillen, M., Nielsen, J. P. (2018). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. Transportation, in press.

- Baecke, P., Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. Decision Support Systems 98, 69-79.

- Bolderdijk, J. W., Knockaert, J., Steg, E. M., Verhoef, E. T. (2011). Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: Results of a Dutch field experiment. Accident Analysis and Prevention 43, 1181-1186.

- Boucher, J. P., Pérez-Marín, A. M., Santolino, M. (2013). Pay-as-you-drive insurance: The effect of the kilometers on the risk of accident. Anales del Instituto de Actuarios Espaoles 19, 135-154.

- Denuit, M., Marechal, X., Pitrebois, S., Walhin, J.-F. (2007). Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems. Wiley, New York.

- Faraway, J. J. (2016). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Second Edition. CRC, Boca Raton, FL.

- Gao, G., Meng, S., Wuthrich, M. V. (2018). Claims frequency modeling using telematics car driving data. Available at SSRN: https://ssrn.com/abstract=3102371

- Guillen, M., Nielsen, J.P., Ayuso, M. and Pérez-Marín, A.M. (2018) The use of telematics devices to improve automobile insurance rates. Risk Analysis, accepted (in press).

- Guillen, M., Pérez-Marín, A. M. (2018). The contribution of Usage-Based data analytics to benchmark semi-autonomous vehicle insurance. In "Mathematical and Statistical Methods for Actuarial Sciences and Finance" (pp. 419-423). Springer.

- Jin, W., Deng, Y., Jiang, H., Xie, Q., Shen, W., Han, W. (2018). Latent class analysis of accident risks in usage-based insurance: Evidence from Beijing. Accident Analysis and Prevention 115, 79-88.

- Lemaire, J. (1995). Bonus-Malus Systems in Automobile Insurance. Kluwer Academic Publisher, Boston.

- Tselentis, D. I., Yannis, G., Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis and Prevention, 98, 139-148.

- Verbelen, R., Antonio, K., Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society: Series C (Applied Statistics), in press.

- Weidner, W., Transchel, F. W. G., Weidner, R. (2016). Classification of scale-sensitive telematic observables for risk individual pricing. European Actuarial Journal 6, 3-24.

- Weidner, W., Transchel, F. W., Weidner, R. (2017). Telematic driving profile classification in car insurance pricing. Annals of Actuarial Science 11, 213-236.

- Williams, A. F. (1985). Nighttime driving and fatal crash involvement of teenagers. Accident Analysis & Prevention, 17(1), 1-5.

- Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. European Actuarial Journal 7, 89-108.

- Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. Second edition. Chapman and Hall/CRC.

Detralytics

info@detralytics.eu
Rue Belliard 2
1040 Brussels
www.detralytics.com