Annals of Actuarial Science, Vol. 13, part 2, pp. 378–399. © Institute and Faculty of Actuaries 2019
doi:10.1017/S1748499518000349
First published online 12 February 2019

Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science, Louvain Institute of Data Analysis and Modeling, UC Louvain, 1348 Louvain-la-Neuve, Belgium

Montserrat Guillen
Riskcenter, Department of Econometrics, Universitat de Barcelona, 08034 Barcelona, Spain

Julien Trufin*
Department of Mathematics, Université Libre de Bruxelles (ULB), 1050 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 other driving habits. These data are collected while the contract is in force with the help of 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 incorporated in the score as ordinary a priori observable features. This paper proposes 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 behaviour; Internet of things; Count data models

1. Introduction

The classical approach to motor insurance pricing can be summarised 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.

*Correspondence to: Julien Trufin. E-mail: julien.trufin@ulb.ac.be

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 source for such data is the automotive diagnostic system (or OBD, for On-Board Diagnostics) installed in the vehicle, or the driver's smartphone. We refer the reader to Boucher *et al.* (2013) and Tselentis *et al.* (2017) for reviews of current practices and emerging challenges in usage-based (UB) motor insurance pricing.

Telematics insurance data offer the opportunity to base actuarial pricing on actual policyholder's behaviour. With pay-how-you-drive (PHYD) or UB motor insurance, premium amounts are based on the total distance travelled, the type of road, the time of the day, average speeds and other driving habits. Thus, premiums are based directly on driver's behaviour behind the wheel. 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 & 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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

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 summarised into three signals recorded thanks to telemetry: in addition to the number of kilometers driven in each year, the insurer collects information on the number of kilometers driven at night, the number of kilometers driven in an urban area, and the number of kilometers 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, night-time 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 organised 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 3, and the results are compared with those obtained according to the classical actuarial approach. Section 4 discusses the results and briefly concludes the paper.

2. Multivariate Credibility Model

2.1. Mixed poisson model for annual claim frequencies

Consider an insurance portfolio comprising n policies observed during several periods. Let N_{it} be the number of claims reported by policyholder i, $i = 1,2,\ldots,n$, during 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 summarised into p features x_{itj} that may evolve over time. The a priori information $\boldsymbol{x}_{it} = (x_{it1},\ldots,x_{itp})^{\mathrm{T}}$ is recorded in the data basis under consideration. 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

$$\begin{aligned}
x_{it} &= \text{features for policyholder } i, i = 1, \ldots, n, \\
&\quad \text{during period } t, t = 1, 2, \ldots, T_i \\
d_{it} &= \text{exposure-to-risk, distance driven in kilometers} \\
\eta_{it} &= \eta(x_{it}) \\
&= \text{score for policyholder } i \text{ in period } t \\
\lambda_{it} &= d_{it} \exp(\eta_{it}) = \exp(\ln d_{it} + \eta_{it}).
\end{aligned}$$

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

A random effect Δ_i is added to the score η_{it} to recognise 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 Δ_i. The numbers of claims N_{i1}, N_{i2}, N_{i3},... are then assumed to be independent

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

given Δ_i. The latent unobservable Δ_i characterises the correlation structure of the claim counts N_{it} for each policyholder i. Specifically, the model is based on the following assumptions:

A1 given $\Delta_i = \delta$, the random variables N_{it}, $t = 1, 2, \ldots$, are independent and conform to the Poisson distribution with mean

$$\lambda_{it} \exp(\delta) = \exp(\ln d_{it} + \eta_{it} + \delta),$$

which is henceforth denoted as $N_{it} \sim \mathcal{P}oi(\lambda_{it} \exp(\delta))$. Formally,

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

A2 at the portfolio level, the sequences $(\Delta_i, N_{i1}, N_{i2}, \ldots)$ are assumed to be independent.

A3 the random effects Δ_i are independent, Normally distributed with zero mean and constant variance σ_Δ^2.

When the canonical log link function is used in the Poisson regression model, as assumed here, assumption A3 amounts to using a Poisson-LogNormal model for claim counts. Contrarily to what is generally assumed in the actuarial literature, where the random effects Δ_i are supposed to be such that $\mathrm{E}[\exp(\Delta_i)] = 1$, the statistical literature devoted to mixed models assumes that the random effects Δ_i are centred. Under assumption A3, we then have $\mathrm{E}[\exp(\Delta_i)] = \exp(\sigma_\Delta^2 / 2)$ according to the formula giving the mathematical expectation for the LogNormal distribution. Therefore, the latter factor has to be included in the calculation of the a priori expected number of claims (with a linear score η_{it}, the intercept of the regression model has thus to be modified accordingly). Formally, the a priori expected number of claims is equal to

$$\mathrm{E}[N_{it}] = \mathrm{E}[\mathrm{E}[N_{it} \mid \Delta_i]] = \lambda_{it} \mathrm{E}[\exp(\Delta_i)] = \lambda_{it} \exp(\sigma_\Delta^2 / 2).$$

Remark 2.1 *If longer panels are available then the static random effects Δ_i can be replaced with dynamic ones $\Delta_{i1}, \Delta_{i2}, \ldots$* which discount past observations according to their seniority. This is easily done by replacing Δ_i with a random sequence $\Delta_{i1}, \Delta_{i2}, \ldots$ obeying a Gaussian process whose covariance structure accounts for the memory effect (AR1, for instance).

2.2. Single behavioural variable, or signal

In order to predict the number of claims N_{it} filed by policyholder i during period t, let us assume that 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 summarises 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).

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

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

$$(\Delta_i, \Gamma_i, N_{i1}, S_{i1}, N_{i2}, S_{i2}, N_{i3}, S_{i3}, \ \dots\ )$$

where

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

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

It is important to realise 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 Δ_i, the counts N_{i1}, N_{i2},... are independent and independent of Γ_i, S_{i1}, S_{i2},....

A5 Given Γ_i, the signals S_{i1}, S_{i2},... are independent and independent of Δ_i, N_{i1}, N_{i2},..., and

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

where ν_{it} is the signal score based on a priori 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 $\mathscr{E}_{it}$ 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

(a) The random variables $\Gamma_i, \mathscr{E}_{i1}, \mathscr{E}_{i2}, \ \dots$ are mutually independent.

(b) The random variables $\mathscr{E}_{i1}, \mathscr{E}_{i2}, \ \dots$ are independent from (Δ_i, N_{i1}, N_{i2}, N_{i3},...).

A6 Given Δ_i and Γ_i, all the observable random variables N_{i1}, S_{i1}, N_{i2}, S_{i2},... 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 (Δ_i, Γ_i) is normally distributed, with zero mean vector 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 Δ_i, the claim counts N_{i1}, N_{i2},... are independent and independent of Γ_i, S_{i1}, S_{i2},....

A5 Given Γ_i, the signal counts S_{i1}, S_{i2},... are independent and independent of Θ_i, N_{i1}, N_{i2},..., and

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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

where ν_{it} is the signal score based on a priori features x_{it} and Γ_i is Normally distributed with zero mean 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 Δ_i and Γ_i, all the observable random variables $N_{i1}, S_{i1}, N_{i2}, S_{i2}, \ldots$ are independent.

Assumptions A4–A6 are in line with the traditional actuarial approach to experience rating, in that they postulate that the dependence between signal and claim counts is only apparent and results from missing information. If we had a complete knowledge of policyholder's characteristics, i.e. if we knew Δ_i, then the signal would not be needed for pricing. Because of limited knowledge about policyholder's driving style, the insurer uses the information contained in the signal that reveals the missing elements in expected claim counts. This is why the signal is separated into three components: the effect ν_{it} of the available features $\boldsymbol{x}_{it}$, the relevant information Γ_i contained in the signal, that may explain expected claim counts beyond the available $\boldsymbol{x}_{it}$, and the random noise $\mathscr{E}_{it}$. The correlation $\rho_{\Delta,\Gamma}$ between Δ_i and Γ_i can be exploited to improve the estimation of the expected number of claims by combining observed signal values with past claims history.

Notice that claim counts N_{it} and signal values S_{it} are correlated by means of the pair (Δ_i, Γ_i) of random effects. For a signal consisting in a mixed Poisson count, this is easily seen as follows:

$$\mathrm{C}[N_{it}, S_{it}] = \mathrm{C}[\mathrm{E}[N_{it} \mid \Delta_i, \Gamma_i], \mathrm{E}[S_{it} \mid \Delta_i, \Gamma_i]]$$

because the conditional covariance is zero by virtue of A6. Hence,

$$\mathrm{C}[N_{it}, S_{it}] = d_{it}^2 \exp(\eta_{it} + \nu_{it})\mathrm{C}[\exp(\Delta_i), \exp(\Gamma_i)].$$

Now, as the pair (Δ_i, Γ_i) is jointly Normal, with zero mean, variances σ_Δ^2 and σ_Γ^2, and correlation $\rho_{\Delta,\Gamma}$, we get

$$\begin{aligned}\mathrm{C}[\exp(\Delta_i), \exp(\Gamma_i)] &= \mathrm{E}[\exp(\Delta_i + \Gamma_i)] - \mathrm{E}[\exp(\Delta_i)]\mathrm{E}[\exp(\Gamma_i)] \\ &= \exp\left(\frac{\sigma_\Delta^2 + \sigma_\Gamma^2}{2}\right)(\exp(\rho_{\Delta,\Gamma}\sigma_\Delta\sigma_\Gamma) - 1)\end{aligned}$$

which is not zero unless $\rho_{\Delta,\Gamma} = 0$, that is, unless Δ_i and Γ_i are mutually independent (so that the signal brings no information about the claim counts).

2.3. Multiple signals

Assume that q signals, denoted as $S_{it}^{(j)}$, $j = 1, 2, \ldots, q$, are available in addition to the p features comprised in $\boldsymbol{x}_{it}$. In case several signals 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 $\sum_{j=1}^{q} \alpha_j S_{it}^{(j)}$ for appropriate weights α_j (determined so to maximise the correlation with the observed claim counts N_{it}). 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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

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

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

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

where $\nu_{it}^{(j)}$ is the score for the jth signal based on a priori features $\boldsymbol{x}_{it}$, $\Gamma_i^{(j)}$ is Normally distributed with zero mean 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 $\mathscr{E}_{it}^{(j)}$ 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 $\Gamma_i^{(j)}, \mathscr{E}_{i1}^{(j)}, \mathscr{E}_{i2}^{(j)}, \ldots$ are mutually independent.
- The random variables $\mathscr{E}_{i1}^{(j)}, \mathscr{E}_{i2}^{(j)}, \ldots, j = 1, 2, \ldots$ are mutually independent.
- The random variables $\mathscr{E}_{i1}^{(j)}, \mathscr{E}_{i2}^{(j)}, \ldots$ are independent from $(\Delta_i, N_{i1}, N_{i2}, N_{i3}, \ldots)$.
- The random vector $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots, \Gamma_i^{(q)})$ is multivariate Normally distributed with zero mean vector and variance-covariance matrix Σ.

A6 Given $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots, \Gamma_i^{(q)})$, 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.

The random vectors $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots, \Gamma_i^{(q)})$ are independent and all obey the same Normal distribution. The 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 Δ_i, claim counts $N_{i1}, N_{i2}, \ldots$ are independent and independent of $\Gamma_i^{(j)}, S_{i1}^{(j)}, S_{i2}^{(j)}, \ldots$ for $j = 1, 2, \ldots, q$.

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

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

where $\nu_{it}^{(j)}$ is the score for the jth signal based on a priori features $\boldsymbol{x}_{it}$ and $\Gamma_i^{(j)}$ is Normally distributed with zero mean and represents the additional information contained in the jth signal about claim frequencies, corrected for the effect of the features $\boldsymbol{x}_{it}$. Also, the random vector $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots, \Gamma_i^{(q)})$ is multivariate Normally distributed with zero mean vector and variance-covariance matrix Σ.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

A6 Given $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots \Gamma_i^{(q)})$, all the observable random variables N_{i1}, $S_{i1}^{(1)}$, $S_{i1}^{(2)}$, …, N_{i2}, $S_{i2}^{(1)}$, $S_{i2}^{(2)}$, … 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.

2.4. Credibility updating formulas

In addition to accounting for overdispersion and serial correlation, the random effects

$$(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \ldots, \Gamma_i^{(q)})$$

allow for credibility updates. In the classical actuarial approach based on claim counts, only, past numbers of claims enter the credibility formulas in addition to observable features $\boldsymbol{x}_{i,T_i+1}$ to explain N_{i,T_i+1}. Formally, the experience used to revise future premiums relates to past claims history

$$\mathscr{H}_{i,T_i}^{\text{claim}} = \{N_{it}, \quad t = 1, \ \ldots \ , T_i\}.$$

This information enters the predictive distribution, i.e. the conditional distribution of N_{i,T_i+1} given $\mathscr{H}_{i,T_i}^{\text{claim}}$. With experience rating, the a priori expectation

$$\mathrm{E}[N_{i,T_i+1}] = \lambda_{i,T_i+1}\mathrm{E}[\exp(\Delta_i)]$$

is replaced with the a posteriori expectation

$$\mathrm{E}[N_{i,T_i+1} \mid \mathscr{H}_{i,T_i}^{\text{claim}}] = \lambda_{i,T_i+1}\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,T_i}^{\text{claim}}].$$

The pricing structure is slow to adapt in personal lines because the λ_{it} are generally small.

With telematics and IoT, the past claims history $\mathscr{H}_{i,T_i}^{\text{claim}}$ can be enriched with behavioural data. This allows the pricing structure to become much more reactive but requires the development of multivariate credibility models. In this case, the policy-specific history $\mathscr{H}_{i,T_i}$ gathers all the a posteriori information

$$\mathscr{H}_{i,T_i} = \mathscr{H}_{i,T_i}^{\text{claim}} \cup \mathscr{H}_{i,T_i}^{\text{signals}} = \{N_{it}, S_{it}^{(1)}, \ \ldots \ , S_{it}^{(q)}, \quad t = 1, \ \ldots \ , T_i\}.$$

The multivariate mixed/credibility model describes the joint dynamics of $N_{it}, S_{it}^{(1)}, \ \ldots \ , S_{it}^{(q)}$, given a priori features $\boldsymbol{x}_{it}$. The predictive distribution now corresponds to the conditional distribution of N_{i,T_i+1} given $\mathscr{H}_{i,T_i}$. The a priori expectation is replaced with an a posteriori one

$$\mathrm{E}[N_{i,T_i+1} \mid \mathscr{H}_{i,T_i}] = \lambda_{i,T_i+1}\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,T_i}].$$

The factor $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,T_i}] / \mathrm{E}[\exp(\Delta_i)]$ is the credibility correction, i.e. the ratio between the a posteriori and the a priori expected numbers of claims.

3. Case Study

3.1. Presentation of the data set

In order to illustrate the approach proposed in Section 2, we perform a case study 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 kilometers driven in a given period of time as well as information on the number of kilometers driven at night, the number

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

of kilometers driven in an urban area, and the number of kilometers 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 km. Specifically, the three signals at our disposal are as follows:

$S_{it}^{(1)}$ = distance travelled in the night (in multiples of 500 kilometers)

$S_{it}^{(2)}$ = distance driven above the speed limit (in multiples of 500 kilometers)

$S_{it}^{(3)}$ = distance travelled in urban zones (inmultiples of 500 kilometers).

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. To this end, we use the modelling approach proposed in Section 2.3.

Notice that these are not compositional data in the sense of Verbelen *et al.* (2018). Such data model percent exposure and they have to cope with the restriction that percentages need to add up to 100% at the policyholder level. Here, the sum of the distances used as signals does not necessarily match the total distance travelled. Data on the total distance driven per year (in kilometers) 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 kilometers.

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 kilometers 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 $n = 2,494$ insured drivers followed over the three calendar years 2009–2011. All policyholders have been observed for three years (so that $T_i = 3$ for all i). 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.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

In Table 1, we present descriptive statistics for telematics data observed in the sample for each year. It is worth stressing that distance driven dropped 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.

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 500 s km). Our measure of exposure-to-risk is the distance driven measured as a continuous variable in 100 s km. Figure 1 shows four histograms of the raw telematics data in 2009. This helps to figure out the sample distribution of the total distance travelled as well as of the three signals entering the analysis.

Table 2 presents the counts information for the 3 years and the four counts once the signals of speed, night-time and urban are transformed in discrete counts in units of 500 s km. We can see there that the majority of claim counts as well as night-time and speed signals concentrate in low-frequency cells, whereas the counts of the signal urban are located in a higher frequency level. The information in Table 2 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 six accidents, while the maximum number of claims was four in 2010 and 2011. A few policyholders recorded high levels of speed limit excess in 2009 and even a bit more in 2010.

Table 1. Sample statistics for raw telematic information by year ($n = 2,494$).

	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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

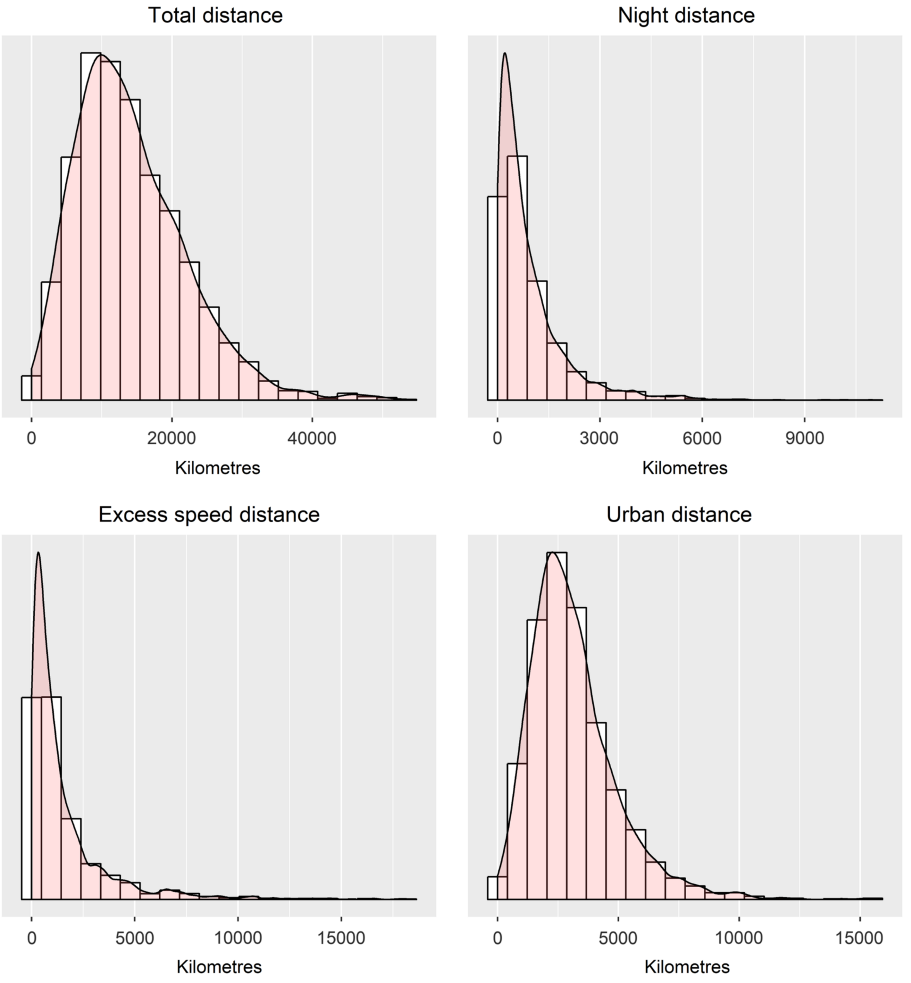

Figure 1. Histograms of telematic information recorded in 2009 ($n = 2,494$).

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-time 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) 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 of claims as well as distance driven. Let us now investigate the strength of this association on our data set. Figure 2 shows a correlation between the distance

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

Table 2. Counts of claims and driving signals (expressed in 500 s km) in 2009, 2010 and 2011.

	Year: 2009				Year: 2010				Year: 2011			
	Claims	Night	Speed	Urban	Claims	Night	Speed	Urban	Claims	Night	Speed	Urban
0	2,004	652	461	18	2,038	640	473	17	2,091	1,131	1,227	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					

and the three raw indicators (speed, night-time, urban) in 2009–2011. Just by illustration in Figure 3, we also show the correlation between distance driven in the three observed years. As expected, distance driven correlates with the signals and between consecutive years. Notice that no correction has been made for standard risk factors at this stage so that the

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

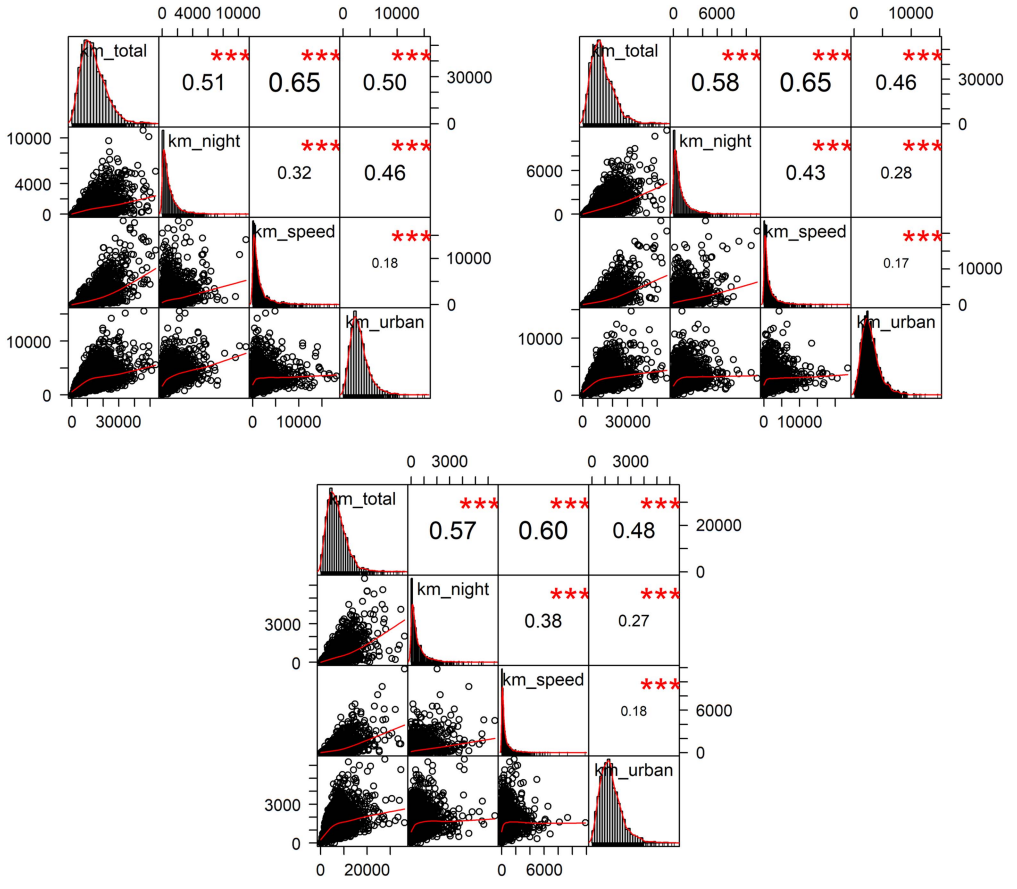

Figure 2. Correlation matrix of telematic information recorded in 2009–2011 ($n = 2,494$).

correlation may only be apparent, being generated by the confounding effects of the standard risk factors comprised in $\boldsymbol{x}_{it}$. The multivariate credibility model will precisely avoid this possible pitfall.

3.4. Fitted models

Here, we assume that the joint dynamics of $(N_{it}, S_{it}^{(1)}, S_{it}^{(2)}, S_{it}^{(3)})$, $t = 1, 2, \ldots$, is described by the multivariate mixed Poisson model described in Section 2.3. Such mixed Poisson models are particular cases of Generalised Linear Mixed-effects Models (or GLMM). The glmer function included in the R package lme4 can be used to fit a GLMM which incorporates both fixed-effects parameters and random effects in a linear predictor, via maximum likelihood. The multivariate Poisson-LogNormal model for claim and signal counts considered in the present section was fitted using the glmer function which performs Poisson regression with structured random effects.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

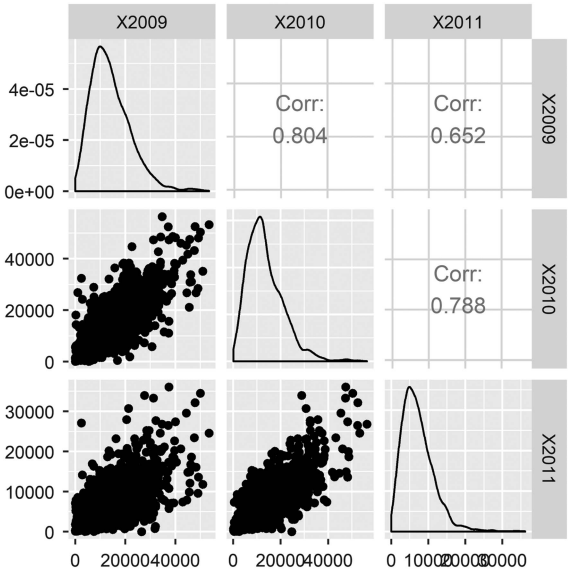

Figure 3. Correlation matrix of distance driven in 2009–2011 ($n = 2,494$).

The expression for the likelihood of a mixed-effects model involves an integral over all the random effects. In our case, the likelihood associated to the observations $(n_{it}, s_{it}^{(1)}, s_{it}^{(2)}, s_{it}^{(3)})$, $t = 1, 2, 3$, writes

$$\mathscr{L} = \prod_{i=1}^{n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \prod_{t=1}^{3} \left(\exp(-d_{it}\exp(\eta_{it}+\delta)) \frac{(d_{it}\exp(\eta_{it}+\delta))^{n_{it}}}{n_{it}!} \right.$$
$$\left. \prod_{j=1}^{3} \left(\exp\left(-d_{it}\exp(\nu_{it}^{(j)}+\gamma_j)\right) \frac{\left(d_{it}\exp(\nu_{it}^{(j)}+\gamma_j)\right)^{s_{it}^{(j)}}}{s_{it}^{(j)}} \right) \right) f_{\sum}(\delta, \gamma_1, \gamma_2, \gamma_3) d\delta d\gamma_1 d\gamma_2 d\gamma_3$$

where f_Σ is the joint probability density function of the random vector $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \Gamma_i^{(3)})$, corresponding to the assumed multivariate Normal distribution with zero mean vector and variance-covariance matrix Σ. For a GLMM, the integral must be approximated with the help of quadrature formulas. Let us mention that to achieve convergence, some care is needed and appropriate control parameters must be selected in relation with the nonlinear optimiser. To ensure numerical stability of the optimisation 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 km, without affecting the results).

Both fixed effects and random effects are specified via the model formula. The multivariate model considers claim counts and the three signals simultaneously. We fit the multivariate model at once following the approach proposed by Faraway (2016, Section 9.3). The idea is to define count and signal identifiers by means of a categorical feature signalName with four levels, N, 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 corresponding four-dimensional policyholder-specific random effects. In order to get four correlated random effects between claim counts and the three signals, we need to specify the random effect structure as (-1 + signalName|id) where id denotes the policy identifier (allowing the actuary to track the same contract over time) entering model formula.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

To illustrate the relevance of the approach proposed in this paper, we compare the multivariate credibility model described above, including past claims history as well as the three signals, with a classical credibility model, based on claim counts only. More precisely, we fit univariate mixed Poisson models for panel data, separately for each signal and the number of claims. The results for the univariate models can be considered as those obtained by replacing the covariance matrix Σ with a diagonal one, with marginal variances along the main diagonal. In the univariate modelling, the four responses N_{it}, $S_{it}^{(1)}$, $S_{it}^{(2)}$ and $S_{it}^{(3)}$ are thus considered to be mutually independent (but serial dependence for fixed i is taken into account in all four cases). 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) entering model formula. In this case, only past claim experience is used to update the expected number of claims in future years.

Table 3 presents the results of the univariate and the multivariate counts models (estimated with the 3-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.

We see that 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 random effects $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \Gamma_i^{(3)})$, the responses

Table 3. Model results for panel data on claims and driving count signals, 2009–2011.

		Univariate models			
	Multivariate Models	S1 (Night)	S2 (Speed)	S3 (Urban)	Claims
(Intercept)	−5.08(0.29)***	−4.33(0.14)***	−3.31(0.15)***	−2.34(0.09)***	−4.99(0.31)***
S1 (Night)	*0.76(0.32)				
S2 (Speed)	1.78(0.32)***				
S3 (Urban)	***2.47(0.28)				
Age	***−5.21(1.13)	−1.33(0.55)*	−4.27(0.61)***	−3.29(0.36)***	−5.95(1.22)***
Gender	−0.11(0.06)	0.38(0.03)***	0.22(0.04)***	0.03(0.02)	−0.09(0.06)
S1 (Night):Age	3.84(1.25)**				
S2 (Speed):Age	0.76(1.29)				
S3 (Urban):Age	3.02(1.11)**				
S1 (Night):Gender	0.49(0.07)***				
S2 (Speed):Gender	0.34(0.07)***				
S3 (Urban):Gender	0.14(0.06)***				

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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

are Poisson distributed with respective conditional means

$$\mathrm{E}[N_{it} \mid \Delta_i] = d_{it} \exp(-5.08 - 5.21\mathrm{age}_i - 0.11\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Delta_i)$$

$$\begin{aligned}
\mathrm{E}[S_{it}^{(1)} \mid \Gamma_i^{(1)}] &= d_{it} \exp((-5.08 + 0.76) + (-5.21 + 3.84)\mathrm{age}_i \\
&\quad + (-0.11 + 0.49)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(1)}) \\
\mathrm{E}[S_{it}^{(2)} \mid \Gamma_i^{(2)}] &= d_{it} \exp((-5.08 + 1.78) + (-5.21 + 0.76)\mathrm{age}_i \\
&\quad + (-0.11 + 0.34)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(2)}) \\
\mathrm{E}[S_{it}^{(3)} \mid \Gamma_i^{(3)}] &= d_{it} \exp((-5.08 + 2.47) + (-5.21 + 3.02)\mathrm{age}_i \\
&\quad + (-0.11 + 0.14)\mathrm{I}[\mathrm{gender}_i = \mathrm{male}] + \Gamma_i^{(3)}).
\end{aligned}$$

The estimated fixed effects are coherent between the multivariate and univariate models so that the estimated scores $\widehat{\eta}_{it}$ and $\widehat{\nu}_{it}^{(j)}$ are very similar in both cases. The main advantage of the multivariate model is to estimate the covariance matrix Σ of the random vector $(\Delta_i, \Gamma_i^{(1)}, \Gamma_i^{(2)}, \Gamma_i^{(3)})$ which connects claim counts N_{it} to corresponding signals $(S_{it}^{(1)}, S_{it}^{(2)}, S_{it}^{(3)})$.

The estimated covariance matrix $\widehat{\Sigma}$ is as follows. The marginal standard deviations are estimated to

$$\widehat{\sigma}_{\Delta} = 0.836$$

$$\widehat{\sigma}_{\Gamma,1} = 0.521$$

$$\widehat{\sigma}_{\Gamma,2} = 0.753$$

$$\widehat{\sigma}_{\Gamma,3} = 0.438.$$

The estimated correlation coefficients are given by

$$\widehat{\rho}_{\Delta,\Gamma,1} = 0.019$$

$$\widehat{\rho}_{\Delta,\Gamma,2} = -0.204$$

$$\widehat{\rho}_{\Delta,\Gamma,3} = 0.602$$

$$\widehat{\rho}_{\Gamma,1,2} = 0.026$$

$$\widehat{\rho}_{\Gamma,1,3} = 0.058$$

$$\widehat{\rho}_{\Gamma,2,3} = -0.484.$$

We see that signal 1 (night-time driving) brings little information about claim counts in our data basis. The effects of signals 2 and 3 clearly dominate with respective correlations of about 20% and 60%, exhibiting opposite signs. Signal 3 (urban driving) appears to be the most informative, and negatively correlated to signal 2 (excess speed). This can be explained by traffic congestion, reducing speed in urban areas. On our data set, the estimated correlation between Δ_i and $\Gamma_i^{(2)}$ appears to be negative. This can be attributed to the way excess speed has been recorded in the data basis, without distinctions between small and large violations of the posted speed limit.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

3.5. A posteriori corrections

The multivariate model does not outperform the classical, univariate one on aggregate. This is easily seen from Table 3, by noticing that the estimated fixed effects are very similar for the claim count component of the multivariate model and the univariate model for claim counts, only. In fact, a simple Poisson GLM with an intercept would produce predicted numbers of claims close to the observed ones (provided the portfolio experience is stationary) both at the portfolio level and within sub-portfolios. This comes from the marginal totals constraints imposed by the likelihood equations. The added value of the multivariate model proposed in this paper consists in refined individual premium corrections, as explained next.

The credibility approach consists in predicting the number of claims for next year using the conditional distribution of the response given past experience. Here, past experience gathers the observed numbers of claims filed in the past for the univariate model. In the multivariate case, it also includes the history of the three signals. Approximations for the predictions can be 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 ranef function of glmer that extracts the conditional modes of the random effects from the fitted model). Here, we prefer to implement exact formulas for a posteriori expectations in the proposed credibility model.

The expected number of claims N_{i,T_i+1} to be filed by policyholder i in year T_i+1 given past numbers of claims $N_{it}=k_{it}$, $t=1,2,\ \dots\ ,T_i$, and past values of signals $S_{it}^{(j)}=l_{it}^{(j)}$, $t=1,2,\ \dots\ ,T_i$, $j=1,2,3$, can be obtained as follows. As random effects are static, past experience is more conveniently summarised into the statistics

$$k_i=\sum_{t=1}^{T_i}k_{it}\ \text{and}\ l_i^{(j)}=\sum_{t=1}^{T_i}l_{it}^{(j)}.$$

Also, we define

$$\lambda_i=\sum_{t=1}^{T_i}\lambda_{it}=\sum_{t=1}^{T_i}\exp(\ln d_{it}+\eta_{it})\ \text{and}\ \mu_i^{(j)}=\sum_{t=1}^{T_i}\exp(\ln d_{it}+\eta_{it}^{(j)}).$$

Then,

$$\begin{aligned}
&\mathrm{E}[N_{i,T_i+1}\mid N_{i1}=k_{it},S_{it}^{(j)}=l_{it}^{(j)},t=1,2,\ \dots\ ,T_i,j=1,2,3]\\
&=\mathrm{E}[N_{i,T_i+1}\mid N_{i1}+\ \dots\ +N_{iT_i}=k_i,S_{i1}^{(j)}+\ \dots\ +S_{iT_i}^{(j)}=l_i^{(j)},j=1,2,3]\\
&=d_{i,T_i+1}\exp(\eta_{i,T_i+1})\mathrm{E}[\exp(\Delta_i)\mid N_{i1}+\ \dots\\
&\qquad +N_{iT_i}=k_i,S_{i1}^{(j)}+\ \dots\ +S_{iT_i}^{(j)}=l_i^{(j)},j=1,2,3]\\
&=d_{i,T_i+1}\exp(\eta_{i,T_i+1})\frac{A}{B}
\end{aligned}$$

where

$$\begin{aligned}
A=(k_i+1)&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(-\lambda_i\exp(\delta))(\exp(\delta))^{k_i+1}\\
&\prod_{j=1}^{3}\left(\exp\left(-\mu_i^{(j)}\exp(\gamma_j)\right)\left(\exp(\gamma_j)\right)^{l_i^{(j)}}\right)f_{\sum}(\delta,\gamma_1,\gamma_2,\gamma_3)d\delta d\gamma_1 d\gamma_2 d\gamma_3
\end{aligned}$$

$$B=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(-\lambda_i\exp(\delta))(\exp(\delta))^{k_i}$$

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

$$\prod_{j=1}^{3}\left(\exp\left(-\mu_i^{(j)}\exp(\gamma_j)\right)\left(\exp(\gamma_j)\right)^{l_i^{(j)}}\right)f_{\sum}(\delta,\gamma_1,\gamma_2,\gamma_3)d\delta d\gamma_1 d\gamma_2 d\gamma_3.$$

The updating coefficient is thus given by A/B. The integrals involved in A and B can be computed numerically with quadrature formulas as implemented in the R package MultiGHQuad.

In the univariate case, we simply get the ratio of two Mellin transforms:

$$\begin{aligned}\mathrm{E}[N_{i,T_i+1} \mid N_{i1}=k_{it}, t=1,2,\ldots,T_i] &= \mathrm{E}[N_{i,T_i+1} \mid N_{i1}+\ldots+N_{iT_i}=k_i]\\ &= d_{i,T_i+1}\exp(\eta_{i,T_i+1})\mathrm{E}[\exp(\Delta_i) \mid N_{i1}+\ldots+N_{iT_i}=k_i]\\ &= d_{i,T_i+1}\exp(\eta_{i,T_i+1})\frac{C}{D}\end{aligned}$$

where

$$C=(k_i+1)\int_{-\infty}^{\infty}\exp(-\lambda_i\exp(\delta))(\exp(\delta))^{k_i+1}f_{\sigma_\Delta^2}(\delta)d\delta$$

$$D=\int_{-\infty}^{\infty}\exp(-\lambda_i\exp(\delta))(\exp(\delta))^{k_i}f_{\sigma_\Delta^2}(\delta)d\delta$$

where $f_{\sigma_\Delta^2}$ is the probability density function of the Normal distribution with zero mean and variance σ_Δ^2.

Let us now demonstrate the added value of the multivariate model by computing individual premium corrections. The boxplots of the values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ based on the multivariate model involving the three signals and of the values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\text{claim}}]$ based on the univariate model (i.e. the classical credibility construction based on the Poisson-LogNormal model for claim counts) are displayed in Figure 4. Apart from the common increasing trend according to the number of claims $N_{i1}+N_{i2}+N_{i3}$ filed during the observation period, we see that there is more dispersion in the $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ values compared to the $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\text{claim}}]$ values, because of the variety in the signal.

Let us now compare the values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ based on the multivariate model involving the three signals to the values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\text{claim}}]$ obtained from the univariate model for claim counts, only, according to the total number of claims $N_{i1}+N_{i2}+N_{i3}$ filed during the observation period. The numerical values are displayed in Figure 5. For claim-free policyholders, we see that the univariate model always grants a discount whereas its multivariate counterpart may impose a penalty, depending on the experience with signals. When a single claim is reported, both univariate and multivariate models may still award a discount or induce a penalty. For the univariate model, it depends on the a priori features of the driver (a priori riskier drivers are less penalised when a claim is reported to the company). For the multivariate model, it depends on the a priori features as well as on the experience recorded on signals. When two claims are reported, the univariate model always imposes a penalty whereas its multivariate counterpart may still award a discount, based on favourable experience related to signals. When three claims (or more) are reported, both the univariate and multivariate models impose a penalty, but its extent also depends on the signals in the multivariate case.

Let us now consider a male policyholder with average age and driving the average annual distance. Also, we fix the signals 1 and 2 at their average value, but we let the third signal vary from 0 to its maximal value given the assumed total mileage. Based on the number of claims reported during the

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

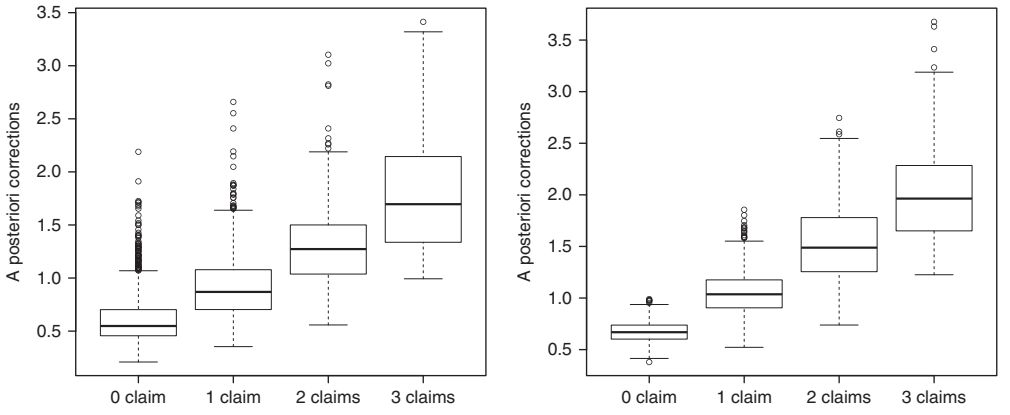

Figure 4. Boxplots of the values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ based on the multivariate model (left panel) and of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\text{claim}}]$ (right panel) obtained from the univariate model, according to the total number of claims $N_{i1}+N_{i2}+N_{i3}$ filed during the observation period.

three years, we compute the a posteriori corrections to assess the impact of the signal. The results are displayed in Figure 6. For a policyholder without claim ($N_{i1}+N_{i2}+N_{i3}=0$), we see that having a better driving style (small value of the signal) increases the discount compared to the classical credibility correction based on past claims, only (represented by the horizontal line on the graph). For a policyholder having reported a single claim, ($N_{i1}+N_{i2}+N_{i3}=1$), we see that depending on the value of the signal, the premium may increase or decrease (whereas it moderately increases using the classical credibility formula). Hence, the signal can compensate for the effect of a single claim. When two or three claims are reported, the policyholder suffers a penalty whatever the value of the signal, but the latter can attenuate the penalty compared to the classical credibility model based on past claim experience, only.

4. Discussion

The approach proposed in this paper recognises the a posteriori nature of telematics data and their variety 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 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

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

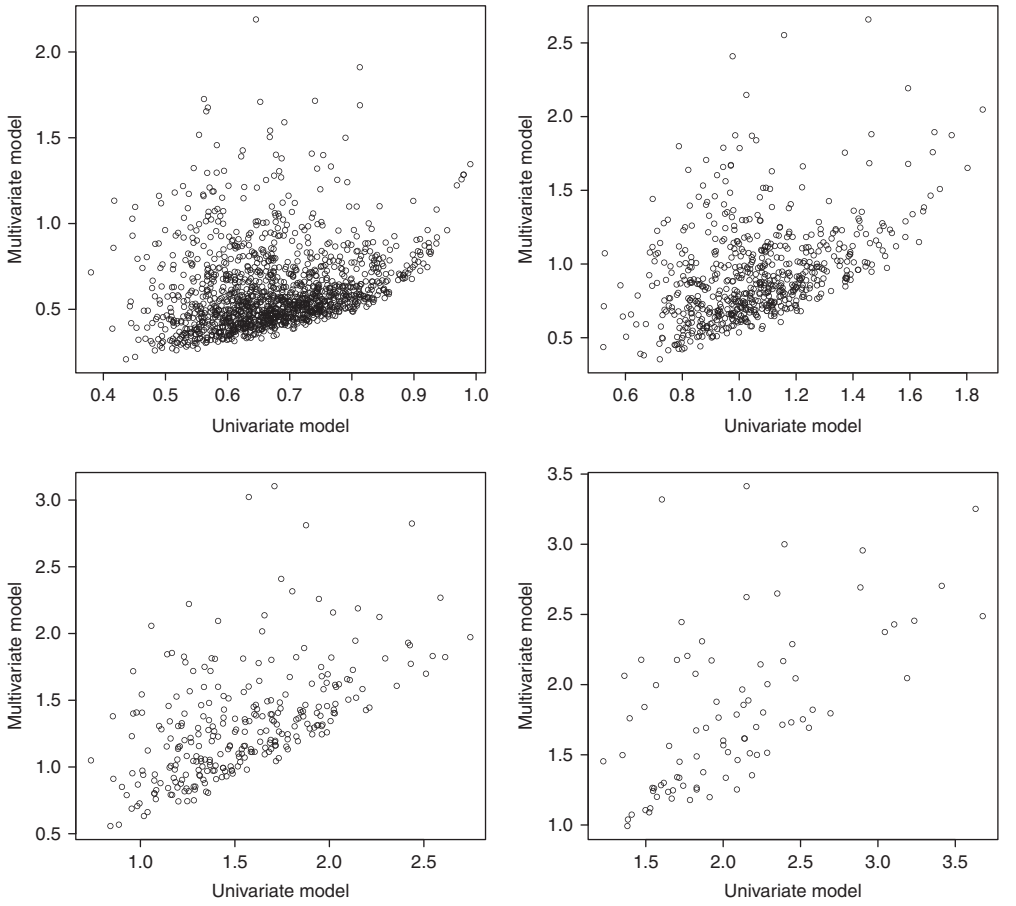

Figure 5. Values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ based on the multivariate model and of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\text{claim}}]$ obtained from the univariate model, according to the total number of claims filed during the observation period: $N_{i1}+N_{i2}+N_{i3}=0$ (top left), $N_{i1}+N_{i2}+N_{i3}=1$ (top right), $N_{i1}+N_{i2}+N_{i3}=2$ (bottom left), and $N_{i1}+N_{i2}+N_{i3}=3$ (bottom right).

aspects of driving are likely to be incorporated in insurance models in order to contribute towards current trends of personalised 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.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

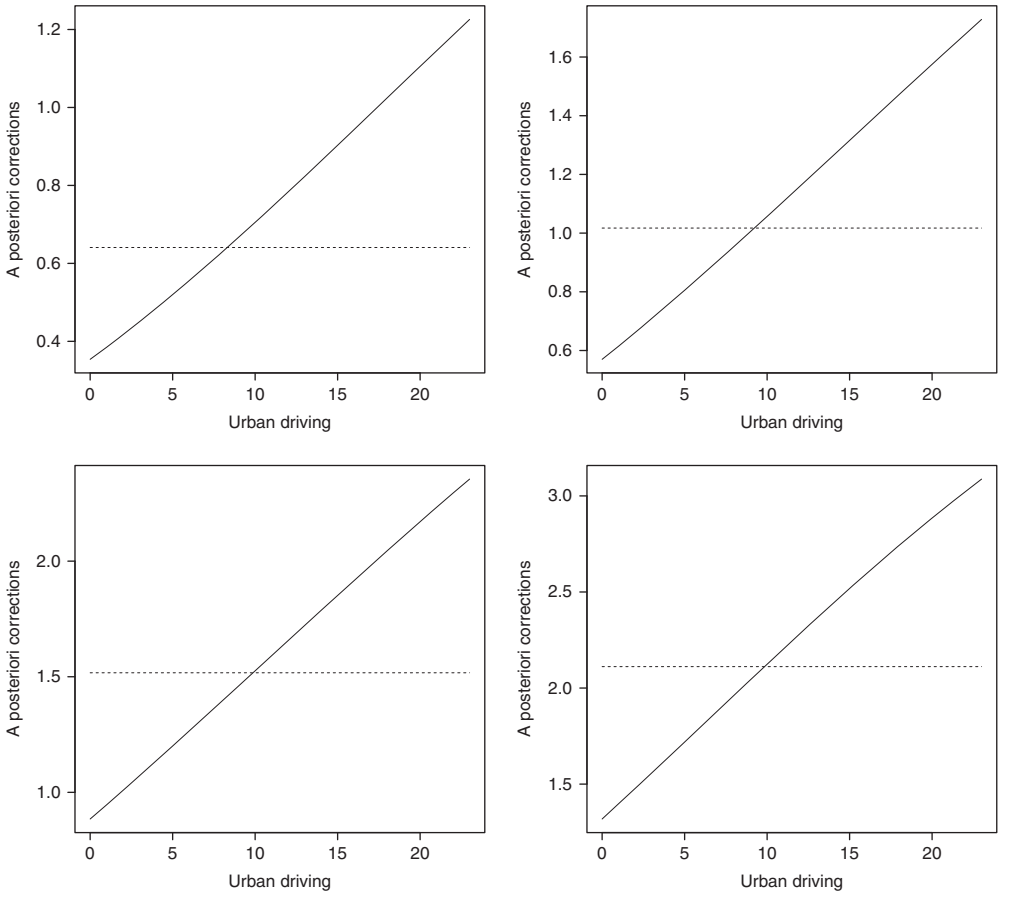

Figure 6. Values of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}]$ for an hypothetical male, mean-aged driver in function of the distance travelled in urban areas, based on the multivariate model and of $\mathrm{E}[\exp(\Delta_i) \mid \mathscr{H}_{i,3}^{\mathrm{claim}}]$ obtained from the univariate model, according to the total number of claims filed during the observation period: $N_{i1}+N_{i2}+N_{i3}=0$ (top left), $N_{i1}+N_{i2}+N_{i3}=1$ (top right), $N_{i1}+N_{i2}+N_{i3}=2$ (bottom left), and $N_{i1}+N_{i2}+N_{i3}=3$ (bottom right).

Acknowledgements

We thank the Referee for his/her careful reading of a previous version of this text. The numerous comments and suggestions contained in the review report greatly helped us to improve the present work. We also thank Florian Pechon, Researcher at UC Louvain, for useful information about numerical integration in R.

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.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press

References

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 Españoles*, **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*, 2nd 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. & 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)*, **67**, 1275–1304.

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 and Prevention*, **17**, 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*, 2nd edition. Chapman and Hall/CRC, Boca Raton, FL.

https://doi.org/10.1017/S1748499518000349 Published online by Cambridge University Press