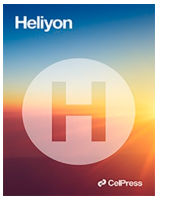
Heliyon 10 (2024) e36501

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Pricing weekly motor insurance drivers' with behavioral and contextual telematics data

Montserrat Guillen [a,b,*], Ana M. Pérez-Marín [a,b], Jens P. Nielsen [c]

[a] Departament d'Econometria, Estadística i Economia Aplicada, Universitat de Barcelona (UB), Av. Diagonal, 690, 08034, Barcelona, Spain
[b] RISKcenter-Institut de Recerca en Economia Aplicada (IREA), Universitat de Barcelona (UB), Av. Diagonal, 690, 08034, Barcelona, Spain
[c] Bayes Business School. City, University of London, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom

ARTICLE INFO

Keywords:
Motor insurance
Near-miss
Traffic accident
Highway
Speed

ABSTRACT

Telematics boxes integrated into vehicles are instrumental in capturing driving data encompassing behavioral and contextual information, including speed, distance travelled by road type, and time of day. These data can be amalgamated with drivers' individual attributes and reported accident occurrences to their respective insurance providers. Our study analyzes a substantial sample size of 19,214 individual drivers over a span of 55 weeks, covering a cumulative distance of 181.4 million kilometers driven. Utilizing this dataset, we develop predictive models for weekly accident frequency. As anticipated based on prior research with yearly data, our findings affirm that behavioral traits, such as instances of excessive speed, and contextual data pertaining to road type and time of day significantly aid in ratemaking design. The predictive models enable the creation of driving scores and personalized warnings, presenting a potential to enhance traffic safety by alerting drivers to perilous conditions. Our discussion delves into the construction of multiplicative scores derived from Poisson regression, contrasting them with additive scores resulting from a linear probability model approach, which offer greater communicability. Furthermore, we demonstrate that the inclusion of lagged behavioral and contextual factors not only enhances prediction accuracy but also lays the foundation for a diverse range of usage-based insurance schemes for weekly payments.

1. Introduction

Data providers that collect telematics from vehicles in motion usually do not have access to evidence from accidents, which would easily be retrieved from insurance records. At the same time, insurers make little use of the massive amounts of material gathered by telematics boxes and they do not look at detailed telematics information as they mostly resort to driving mileage only. The dissociation between information suppliers comes together with the reluctance of insurance companies to reveal the nature of their rating factors to external parties. All in all, this has considerably slowed down research on measurable driving behavior and operating circumstances that explain a driver's proneness to cause a traffic accident, in spite of a massive amount of information that is known to have been recorded somewhere. Therefore, data inaccessibility and the lack of synergies are the reasons why we do not expect to see major transformations in usage-based insurance in the market in the short future. We do, however, find pay-as-you-drive schemes being

* Corresponding author. Departament d'Econometria, Estadística i Economia Aplicada, Universitat de Barcelona (UB), Av. Diagonal, 690, 08034, Barcelona, Spain.
E-mail addresses: mguillen@ub.edu (M. Guillen), amperez@ub.edu (A.M. Pérez-Marín), jens.nielsen.1@city.ac.uk (J.P. Nielsen).

https://doi.org/10.1016/j.heliyon.2024.e36501
Received 7 December 2023; Received in revised form 9 July 2024; Accepted 16 August 2024
Available online 18 August 2024

commercialized all over the world. Under those systems, drivers pay a constant fee plus some cost per mile. Typically, these schemes do not consider where or when the driving distance has been driven or how drivers are effectively managing their vehicles.

Our contribution aims to bridge the gap of the literature on the analytic methods to assess the role of behavioral and contextual driving data in predicting accident expected frequency on a weekly basis, something that has a direct implication on the expansion of usage-based insurance. Our method proposes an algorithm to estimate risky driving scores that can be used to provide feedback to drivers about their performance on the wheel or to construct insurance tariffs based not only on how many miles are driven, but also on when and where distance is driven and how a driver is operating their car. Contrary to the existing literature, we offer a comprehensive analysis of advanced driving risk assessment in weekly periods. Our approach has the potential to be integrated in new vehicles by innovative manufacturers to provide feedback to the driver, or to serve as the basis for insurance ratemaking that includes behavioral and contextual driving data. Certainly, one may assert formally that the consideration of weekly ratemaking in motor insurance becomes particularly intriguing in the context of carsharing vehicles or rental cars, wherein drivers undergo frequent and dynamic changes. The fluid nature of driver rotations in such scenarios necessitates a nuanced approach to ratemaking, accommodating the variability in driver profiles and usage patterns. This underscores the importance of a weekly rate structure that can effectively adapt to the evolving nature of the driver pool, ensuring a fair and economically viable insurance framework for both the service providers and the transient drivers involved.

Unlike previous existing contributions we are able to disclose predictors and examine the role of timely data collection. We therefore take full advantage of the fact that telematics data provide a continuous source of information. We argue that a driving risk score can be formulated weekly and that lagged information from the previous week is informative about future accident occurrence. We provide the analysis of a unique dataset of 19,214 drivers observed over more than one year, with a total kilometer distance covered of 181.392.006 km, making this analysis the most complete existing study that can be found in the literature up to now. Our telematics data contain behavioral information on speeding events counts, that is, the number of times in a week and per type of road that a driver exceeded the maximum posted speed. Data also contain distance driven per type of road and distance driven in the night.

Contrary to most analysis which focus on yearly data, our proposed methodology can be generalized to cope to time interval frequencies under one year, such as daily or monthly data, and it can also be implemented by trip, so that a driving risk score is provided after a voyage is finished, but unlike the driving score that we see in some modern cars, we are able to adapt the score to a variety of context information. For example, driving in an urban area at low speed is not necessarily a sign of precautious behavior, but it may be the consequence of heavy traffic congestion and, consequently, more risk of having an accident.

We show that incorporating behavioral and contextual data about the driver's experience improves the prediction performance of classical frequency models used in accident analysis, compared to not including this information, when considering telematics information from the same period. This is a well-known fact, but we also show that lagged information, i.e. telematics data from the previous period, helps anticipating accident frequency in the subsequent period. This opens the door to establishing warning scores that can help drivers identify how their probability of suffering an accident changes along time.

We analyze a unique combination of weekly information on individual drivers and their insurance provider records. Telematics boxes collect data on distance travelled, type of roads, time of day, and speeding events, which are then integrated with accident occurrence records. By amalgamating information on driving style and contextual data with traditional ratemaking factors such as gender, age, and vehicle power, we confirm a substantial enhancement in the ability to predict accident frequency. We further demonstrate how a multiplicative scheme derived from a Poisson model specification or an additive scheme resulting from a linear probability model specification may alter the existing usage-based insurance pricing in the current marketplace, primarily predicated on distance driven, irrespective of the manner and location of driving.

Through our contribution, we emphasize that accident frequency can be anticipated by considering when, where, and how a driver behaves behind the wheel. Previous research has offered only partial solutions, either due to the lack of merging accident data with telematics data or the non-disclosure of factors influencing accident frequency by insurance companies. Our contribution provides a comprehensive perspective, introducing new insights that can significantly enhance usage-based insurance schemes and payment models based on weekly data, incorporating contextual information beyond behavioral patterns.

2. Background

The literature on usage-based insurance is extensive and it has been intensively developed in last twenty years. Eling and Kraft [1] conducted a thorough review of numerous academic studies and industry papers spanning nearly two decades, from 2000 to 2019. These works predominantly focused on investigating the pivotal telematics variable for estimating claim frequency: distance driven. Lemaire et al. [2] highlighted the significance of annual mileage as a potent predictor of at-fault claims. More recently, Gao et al. [3] provided a survey of telematics driving data research in actuarial science. The authors provided a thorough description the nature of telematics driving data, received second by second, and the difficulties one faces dealing with such information.

The concept of usage-based insurance (UBI) initially revolved around assessing insurance rates based on the distance covered, as elucidated in Litman's discussion [4] on various distance-based insurance price structures. The role of mileage and its correlation with claim frequency was examined in conjunction with other factors. Boucher et al. [5] concluded that although total distance driven is a pertinent variable, the relationship between distance driven and accident occurrence might not be strictly linear due to the "learning effect." Essentially, this means that individuals who drive twice as much as others with identical characteristics have fewer than twice the accident claims. Moreover, it can be argued that covering more distance might indicate superior driving skills or a propensity to use safer roads like highways, which are typically associated with long-distance trips. These roads tend to have a decreasing marginal effect on the probability of accidents occurring.

Boucher et al. [6,7] employed a generalized additive model approach to scrutinize the impact of both distance driven and the duration of insurance contracts on claim frequency. Surprisingly, they discovered that neither distance nor contract duration exhibited a linear relationship with claim frequency. Adding to this, Guillen et al. [8] incorporated yearly distance travelled as an offset within a zero-inflated Poisson model to account for excess zeros in claim frequency counts. They also noted a non-linear effect in their dataset regarding this variable. More recently, Boucher and Turcotte [9] utilized GAMLSS (Generalized Additive Models for Location, Scale, and Shape) and GAMs (Generalized Additive Models) with fixed effects to analyze telematics count data in a panel setting. Their findings contradicted earlier perceptions of non-linearity, suggesting that the relationship does appear to be linear. They attributed the apparent non-linearity to residual heterogeneity, effectively captured by the GAMs.

Beyond just considering mileage, a plethora of evidence suggests that various telematics variables hold a strong causal link with accidents. Consequently, these variables can significantly enhance the predictive accuracy of frequency models utilized in automobile insurance. For instance, Verbelen et al. (2018) [10]contend that telematics data empowers the tailoring of automobile insurance pricing based on policyholders' driving behavior. They devised a statistical modeling approach for claim frequency using telematics variables and demonstrated that such variables bolster the model's predictive capability. Consequently, gender as a discriminating rating variable becomes obsolete. A similar finding was echoed by Ayuso et al. [11]. In a more recent study, Ayuso et al. [12] constructed a frequency model adaptable to updates with telemetric data. Their research affirmed that not only the distance covered but also driver habits significantly impact the anticipated number of at-fault accident claims [13]. This revelation underscores that the cost of insurance coverage can be personalized. Telemetry enables insurers to consider factors identified by traffic authorities as associated with risky driving, including traffic violations. So et al. [14] delved into the integration of telematics data into a classification model to ascertain driver heterogeneity, utilizing data gleaned from a Canadian telematics program. Their investigation revealed that evaluating driving behavior is markedly enhanced when employing telematics in comparison to traditional risk factors.

In this context, identifying telematics variables with significant predictive power for accident frequency is pivotal. Modern telematics technologies in car insurance generate vast amounts of data, obtained from high-frequency GPS location data (measured per second) from individual car drivers and trips, leading to the proliferation of big data in the insurance industry. Paefgen et al. [15] noted the complexity and data volume associated with usage-based insurance pricing, emphasizing its challenge in actuarial decision-making. They analyzed real raw location data, considering 15 predictor variables, and compared logistic regression, neural network, and decision tree classifiers. Their study demonstrated that while neural networks exhibited superior classification performance, logistic regression was more favorable from an actuarial perspective due to its ease of interpretation and direct effect quantification. Their results highlighted the potential of high-resolution exposure data in simplifying usage-based insurance pricing. Baecke and Bocca [16] explored risk assessment models integrating driving behavior data using three distinct data mining techniques. They concluded that including standard telematics variables significantly enhanced customer risk assessment, enabling insurers to tailor their products to individual risk profiles. The study also emphasized the importance of incorporating easily interpretable data mining techniques mandated by regulators before advancing to more complex predictive models. Moreover, they demonstrated that telematics-based insurance products could be swiftly implemented, requiring only three months of data for reliable risk estimations.

Huang and Meng [17] utilized logistic regression and four machine learning techniques as risk probability models and Poisson regression as a claim frequency model. They established tariff classes with substantial predictive effects, proposing a pricing framework that improved both interpretability and predictive accuracy. Their empirical results reaffirmed the considerable potential of driving behavior variables in automobile insurance. Pesantez et al. [18] also highlighted logistic regression as a suitable model for predicting claim frequency using telematics information, given its interpretability and good predictive capacity. Despite implementing modern machine learning modeling approaches, they observed that XGBoost necessitated extensive model-tuning procedures to match logistic regression's predictive performance and required more effort for interpretation.

In the realm of machine learning, numerous contributions focused on driving pattern recognition, which can be leveraged to determine accident safety scores and enhance insurance pricing. Weidner et al. [19] identified maneuver patterns, trips, trip segments, and the total insurance period as significant indicators of individual driving behavior. Wüthrich [20] utilized high-frequency GPS location data and innovative algorithms to classify distinct driving styles, demonstrating their applicability in regression analysis for car insurance pricing. Gao and Wüthrich [21] introduced speed and acceleration heatmaps, categorized using the K-means algorithm to differentiate varying driving styles. Gao et al. [22] further explored telematics covariates extracted from car driving data, affirming their superior predictive power for claim frequencies compared to traditional pricing factors like driver's age. Gao and Wüthrich [23] extracted feature information from high-frequency GPS location data, utilizing it to allocate individual car driving trips to specific drivers. Geyer et al. [24] defined a driving factor based on overall distance driven, the number of car rides, and speeding, identifying a significant impact of speed driving factor on risk. Meng et al. [25] calculated risk scores using a supervised driving risk scoring neural network model, demonstrating improved prediction performance for claim frequency when incorporating these risk scores.

Arumugan and Bhargavi [26] conducted a survey on driving behavior in usage based insurance using big data. They proposed a solution that finds the risk posed by aggressive driving and road rage incidents by considering the behavioral and emotional factors of a driver. Ziakopoulos et al. [27]. claimed that telematics pricing entails crash reductions of 20%–43 % and harsh event reductions of 10%–52 % are reported. However, they also noted that telematics-based research might have biases stemming from data availability. The usefulness of telematics-supported driver behaviour analysis is addressed by Ziakopoulos et al. [28] and Siami et al. [29].

Pérez-Marín and Guillen [30] investigated telematics information for risk quantification and safety in vehicles with speed control capabilities, emphasizing the potential to reduce accident claims by addressing excess speed. Guillen et al. [31] identified relevant risk factors to streamline telematics information necessary for risk classification, introducing the concept of near miss events in usage-based insurance, i.e recorded risky events such as braking/accelerating/cornering or smart phone use that are positively correlated with accident occurrence. Their analysis revealed that near-miss events, even if no accident is recorded, offer valuable

insights for dynamic risk monitoring through telematics. Recently, Alrassy et al. [32] investigated driver behavior obtained from large-scale telematics data and its relationship with crash data. They found that hard braking is more indicative of higher collision rates on highways, and hard acceleration is a stronger risk indicator on non-highways urban roads. Guillen et al. [33] integrated telematics data in UBI pricing schemes that penalize near miss occurrence. In their analysis, the authors compensate the lack of claims during the period when telematics information was collected with past claim history of insureds. This is a common limitation in actuarial research dealing with telematics data, specifically that the accident history does not match with telematics data collection period. Similarly, Moosavi and Ramnath [34] investigated driver's styles and also used past-at fault traffic accidents and citations as risk indicators of clusters of drivers with similar driving behavior. Masello et al. [35] found that the driving context has significant power in predicting driving risk.

Tesla presented its Predicted Collision Frequency (PCF) formulas, shedding light on risk score components like forward collision warnings, hard braking, aggressive turning, unsafe following, and forced autopilot disengagement (see Ref. [36]). This transparency contributes to the ongoing discussion on model opacity and showcases the relevance of driving behavior variables in assessing risk. Several car manufacturers [37] have introduced similar safety score systems, emphasizing acceleration, braking, cornering behavior, and distance driven as key metrics to calculate driving performance scores.

Regarding the effectiveness of telematics-based feedback in improving driving behavior, Li et al. [38] remark that post-trip interventions have a limited effect if they are not part of a risk mitigation strategy able to improve long-term behavior. In that sense, the authors proposed to provide a personalized feedback and realistic and actionable suggestions for policyholders. Malekpour et al. [39] found that only providing feedback has a minuscule impact in reducing speeding behavior, and financial incentives are necessary. Similarly, Meuleners et al. [40] concluded that personalized feedback does not seem to produce a significant change in overall driving scores of young drivers (they only found some improvements for specific drivers).

The Appendix A1 provides a summary of telematics variables utilized in the literature for driving risk assessment, encompassing factors beyond distance travelled, such as speed, road type, time of vehicle usage, and the inclusion of near miss events. These insights collectively advance the understanding of telematics variables and their role in shaping insurance products and pricing strategies [41–47].

3. Methods

Our strategy consists in predicting the expected frequency of accidents for driver i in period t, in a sample on n drivers each observed a total of T_i time periods. In our application we observe weekly data, so that T_i is the total number of observed weeks for driver i. We define the maximum observation frame, $T = \max_i T_i$.

Our objective is to model the conditional mathematical expectation of accident frequency for driver i, in period t, denoted as $E(y_{it})$, as a function of J dynamic predictors z_{jit}, where $j = 1,\ldots,J$, which change over time and K static predictors x_{ki}, where $k = 1,\ldots,K$, which do not change over time, including a constant intercept.

Generalized linear models specify a link of the linear predictor, $h(x_{ki}, z_{jit})$, and the output $E(y_{it})$. A statistical distribution in the exponential family for the response random variable y_{it} is also specified. Parameter estimates of the linear predictor can easily be found by likelihood maximization. Other machine learning methods are more flexible in the specification of the combinations of static and dynamic factors, but they require establishing a loss-minimization principle. Usually, Random Forest or XGBoost methods provide interesting and accurate predictive algorithms at the expense of interpretability and analytical expression for the expected accident frequency depending on the predictors [18,48].

In the pre-processing phase we transform some of the telematics information as risky events recorded as part of the dynamic predictors z_{jit}. This is done similar to Guillen et al. [33] where near-miss events (based on hard braking/acceleration and smartphone usage) are considered. We usually denote by D_{it}, distance driven by driver i in period t, i.e. the exposure, or the model offset. Dynamic predictors can be divided in two groups. On one side, we consider continuous predictors like total distance driven in a certain condition (type of road and nighttime/daytime). On the other side, we consider event counts. For example, the sum of excess speed occurrence by type of road.

A simple approach to calculating the impact on the expected frequency of accidents of behavioral and contextual predictor is provided. In order to convert the occurrence of telematics risky events or dangerous distance driven into a simple scoring, we may consider a linear probability model specification or a more general input function $h(x_{ki}, z_{jit})$, which may later be linearized. This linear approximation may not be necessary if we only aim at producing a risk score to inform the driver. However, a linear formulation provides a straightforward way to design usage-based insurance schemes that are easy to communicate.

When risky events have a direct linear impact, an insurance rate per week can be expressed as a flat rate, plus some additional

Table 1
Accident risk scoring formulae for static and dynamic predictive factors with distance driven as exposure.

Risk scoring formula specification	Communication
$h(x_{ki})$	Expected accident frequency depends on a function of driver characteristics only
$D_{it}h(x_{ki})$	Expected accident frequency is proportional to current period distance driven times a combination of driver characteristics
$D_{it}h(x_{ki}, z_{jit})$	Expected accident frequency is proportional to current period distance driven times a combination of driver characteristics and dynamic factors

charges for distance driven and risky event occurrence. Charges can be homogeneous or depending on contextual data, so the cost can vary depending on context and behavioral information. For example, if distance is driven exceeding speed limits, in the weekend, in the night or in congestion areas (urban driving), the impact on accident risk and the final cost, differs from driving without speeding events, during weekdays, during the day and in non-urban areas.

Several possibilities for static and dynamic scoring are presented in Table 1. Note that even if Table 1 only aims at modelling accident frequency, usage-based insurance schemes can follow directly from frequency models, once average cost and general insurance charges are imputed proportionally to expected frequencies. Table 2 presents possible linearized specifications of driving scores.

Our results explore basic classical generalized linear models. Similar conclusions can be found for other possibilities. Poisson models with a log-link were estimated using SAS software and R software. The link in the Poisson model equals $h(x_{ki}, z_{jit}) = \exp\left(\sum_{k=1}^{K}\alpha_k x_{ki} + \sum_{j=1}^{J}\beta_j z_{jit}\right)$. Logistic regression and linear probability models are estimated too. Their corresponding links are $h(x_{ki}, z_{jit}) = 1/\left[1+exp\left\{-\left(\sum_{k=1}^{K}\alpha_k x_{ki} + \sum_{j=1}^{J}\beta_j z_{jit}\right)\right\}\right]$ and $h(x_{ki}, z_{jit}) = \sum_{k=1}^{K}\alpha_k x_{ki} + \sum_{j=1}^{J}\beta_j z_{jit}$, respectively. Details on linearization approximations for the Poisson model can be found in Guillen et al. [33]. Model comparison was done using AIC so that a model with a lower AIC is preferable to another model with a higher AIC.

The predicted average frequency of claims is interpreted as a driving risk score. To design a premium rating, we multiply weekly expected frequency of claims by the claim cost. To compare the differences between premium rates calculated under different models the Gini coefficient is employed [49]. In interpreting the results of the Gini index, it is imperative to consider that small Gini values signify a portfolio with premiums that closely resemble one another. Conversely, a Gini index approaching one signifies substantial inequality, where one policyholder bears the entire cost burden while others contribute comparatively trivial premiums. An intermediate Gini index value indicates a moderate level of personalization in premiums, tailored to policyholder characteristics.

4. Data

Anonymous data were provided by a Spanish insurer that commercializes pay-as-you drive-insurance since 2009. Specifically, our data contain 19,214 drivers observed in Spain from the 9th week of 2018 to the 17th week of 2019. Nevertheless, 8 of these weeks were finally not considered in the analysis because there was a failure in the IT telematics recording system and, as a result, there were too few observations. A total of 922 accidents at fault were observed for the sampled drivers. Table 3 shows variable definitions and Table 4 presents some basic descriptive analysis.

Our data are observed by weeks. This is a unique feature of this particular data set, as usually only yearly analysis is generally available. Reig-Torra et al. [50] analyzed a subset of these data and included weather information from external sources. In this data set we estimated an average claim cost equal to 2331.4 Euros.

Table 4 shows some descriptive statistics. There are 44.30 % men and 55.70 % women in the sample. The average age of drivers is 28.73 years old (standard deviation 4.67), and age ranges between 17 and 74 years of age. The average vehicle power is 102.63 Hp. (standard deviation 29.88). The total number of drivers-week observations is 790,698. Regarding telematics variables, we observe that the total distance travelled per week by one driver ranges between 0.001 and 5974 km, with an average of 229 km/week. When a driver does not drive for one week, that week is excluded from the sample. We also observe that, on average, 20.29 km/week are travelled in the night. The weekly number of speed events (in any type or road) is 3.19, with a maximum of 61. In urban roads the weekly number of speed events is 1.847, with a maximum of 23. Note that the mean weekly frequency of 0.001 corresponds to an expected level of annual claim rate for at-fault accidents (0.001 multiplied by 52 weeks). This rate level is not surprisingly high, given that the portfolio is slightly biased, comprising predominantly novice and young drivers. We observe 0.117 % of the weekly observations there is one at fault claim (in the remaining 99.883 % there is no claim), this corresponds to a yearly frequency of 4.80 %, which lies within the range of similar studies when only accidents at fault are being considered.

Fig. 1 shows the evolution of mean distance driven for the drivers in this data set and Fig. 2 presents the frequency of at-fault claims over time. Fig. 3 shows histograms and bar charts of the variables in the data set. The sharp drop observed after the age of 35 in the sample can be attributed to the fact that Pay-as-You-Drive schemes were primarily marketed to young drivers, resulting in fewer older individuals participating in such pricing schemes. This demographic limitation should be acknowledged in our study. The small

Table 2
Accident risk scoring linear formulae for static and dynamic predictive factors considering distance or log-distance driven.

Linear risk scoring formula specification	Communication
$\sum_{k=1}^{K}\alpha_k x_{ki} + \gamma D_{it} + (1-\gamma)\sum_{j=1}^{J}\beta_{2j} z_{jit}$	Expected accident frequency is approximated (or bounded for pricing purposed) by a static part that depends on a combination of driver characteristics plus a linear combination of distance driven and same-period dynamic factors
$\sum_{k=1}^{K}\alpha_k x_{ki} + \gamma lnD_{it} + (1-\gamma)\left[\sum_{j=1}^{J}\beta_{2j} z_{jit} + \sum_{l=1}^{L}\beta_{3l} z_{lit-1}\right]$	Expected accident frequency is approximated (or bounded for pricing purposed) by a static part that depends on a combination of driver characteristics plus a linear combination of log-distance driven and same- and previous-period dynamic factors

Table 3
Variable definition in telematics weekly data set, Spain 2019.

Variable	Description
VEHICLE_POWER	Vehicle power (in Hp)
AGE	Age of the driver
GENDER	1 = Male, 0 = Female
TOTAL_DISTANCE_DRIVENMK	Thousands of kilometers travelled during the week
KM_NIGHTMK	Thousands of kilometers travelled in the night during the week
SPEED_EVENT	Number of trips when the driver exceeded the posted speed limit on the road during the week
SPEED_EVENT_URBAN	Number of trips when the driver exceeded the posted speed limit on an urban area during the week
PERC_URBAN	Percentage of kilometers driven in urban roads
CLAIM_AT_FAULT	Number of claims at fault during the week

Table 4
Descriptive statistics in telematics weekly data set, Spain 2019.

Variable	Mean	Standard Deviation	Minimum	Maximum
Characteristics of the driver				
AGE	28.727	4.667	17.000	74.000
Characteristics of the vehicle				
VEHICLE_POWER	102.625	29.876	34.000	450.000
Telematics variables (referred to weeks)				
TOTAL_DISTANCE_DRIVENMK	0.229	0.204	0.000	5.974
KM_NIGHTMK	0.020	0.058	0.000	4.064
SPEED_EVENT	3.191	3.709	0.000	61.000
SPEED_EVENT_URBAN	1.847	2.104	0.000	23.000
PERC_URBAN	33.265	23.766	0.000	100.000
CLAIMS_AT_FAULT	0.001	0.034	0.000	1.000

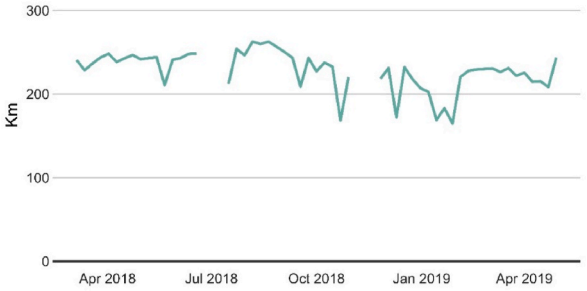


Fig. 1. Mean distance driven per week in the telematics weekly data set, Spain, 2019.

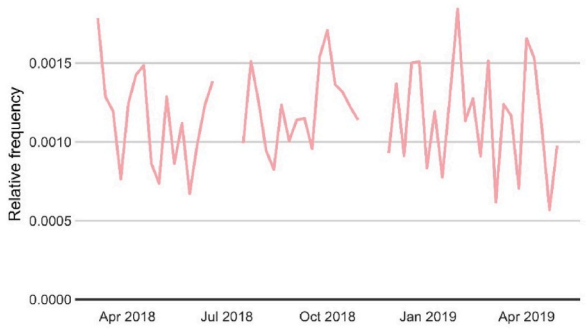
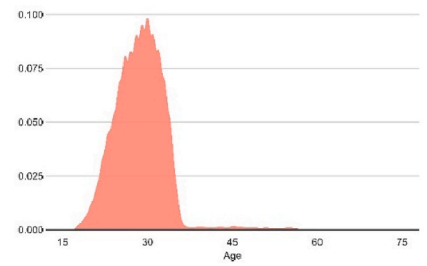
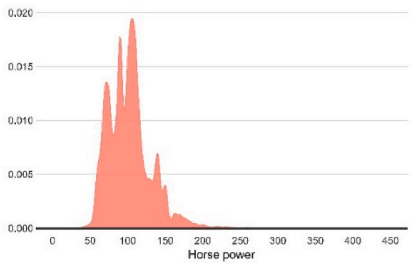


Fig. 2. Average frequency of at-fault claims in the telematics weekly data set, Spain, 2019.

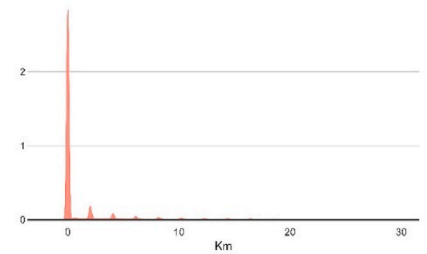
increase observed after the 95th percentile for driving in urban areas reflects the presence of drivers who primarily use their vehicles within their own city, without venturing onto highways or national interurban roads. While this does not constitute a limitation, it is an important aspect that merits discussion in our analysis.



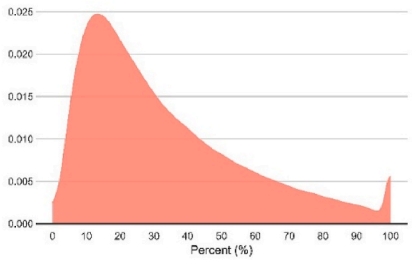
(a) Policy holder's age



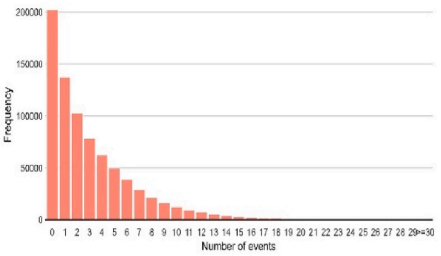
(b) Vehicle capacity



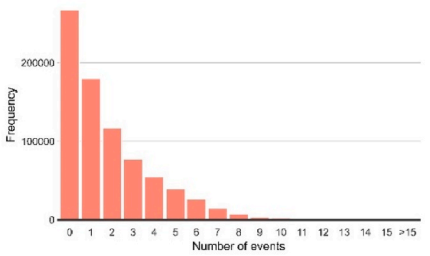
(c) Mean weekday Km at night



(d) Percent driving in urban area



(e) Number of speed events in all areas



(f) Number of speed events in urban areas

Fig. 3. Descriptive histograms of covariates in the telematics weekly data set, Spain, 2019.

5. Results

Following Guillen et al. [33], in Table 5 we present an initial model (Model 0) where we only use driver's characteristics to predict the expected frequency of claims at fault with a Poisson model. Note that the dependent variable is the number of claims in the same

Table 5
Poisson Regression Models: Model 0 (non-telematics), Model 1 (following [33], one near-miss event) and Model 2 results (with speed events as a contextual and risky event), in telematics weekly data set, Spain 2019.

Variable	MODEL 0		MODEL 1		MODEL 2	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−6.502	<0.001	−6.599	<0.001	−6.680	<0.001
Vehicle_power	0.002	0.069	0.001	0.238	0.002	0.149
Gender	0.180	0.009	0.163	0.018	0.163	0.018
Age	−0.019	0.010	−0.016	0.028	−0.016	0.034
Speed_event (wherever) 10^{-1} lag			0.284	<0.001		
Speed_event_urban_lag					0.065	<0.001
AIC	14287		14277		14269	
Pseudo R^2 (%)	0,139		0,220		0,279	
Residual deviance (Null deviance 12455)	12435		12423		12415	

weekly period when telematics information will be introduced. The results of Model 0 set the baseline performance level. The vehicle's engine power, gender and age of the driver conform to the static explanatory variables in all the models considered here. The parameter estimates indicate that a significantly higher accident frequency is expected for powerful cars, younger drivers and males. The reason why we retained the covariate related to car power is to be consistent with previous research by Guillen et al. [31], where hard-braking and acceleration events as well as smartphone use while driving increase the cost of insurance, while conditioning on vehicle horse power.

To demonstrate that the main conclusions remain solid even when all types of reported accidents are considered, we have rerun the analysis for all claims. The results are available from the authors upon results. We acknowledge that our analysis does not include contextual information such as traffic congestion, road conditions, and external factors, nor does it account for assisted driving technologies that may be available in some insured vehicles.

In Model 1, a behavioral risk event count is introduced in the Poisson model: the lagged number of speed events no matter what type of road or circumstance. Our conclusions coincide with those of Guillen et al. [33] who provided much less sophisticated data. We conclude that more speed events positively correlate (p-value$<$0.001) with a higher expected frequency of at fault claims. In Model 2 only the lagged number of speed events are considered if they occur in urban roads. The parameter for this risky event factor is positive and significant (p-value$<$0.001). The AIC of Model 2 is lower than the previous two models.

In Table 5 we consider total distance travelled, which is the most basic information provided by telematics. Specifically, we consider the total number of kilometers, that we introduce in the model with a logarithm transformation. The log of the total number of kilometers is introduced in three different ways: as an offset (Model 3), as an explanatory variable (Model 4a), and, finally, we also introduced the lagged log of the total number of kilometers as an explanatory variable (Model 4b). When the log of the distance is introduced as an explanatory variable, the corresponding parameter is positive and significant, therefore, travelling more kilometers increases the frequency of claims. The model with the lowest AIC is the one where the lagged log of the total number of kilometers is introduced as an explanatory variable.

In Table 6, we introduced more sophisticated behavioral and contextual information, apart from the lagged log of the total distance. Specifically, in Model 5a the lagged number of speed events in urban roads is considered, and it has a positive and significant parameter (p-value = 0.003), therefore, more speed events in urban roads are associated to a higher frequency of claims. In Model 5b we introduce the lagged percentage of urban driving, which it has a positive and significant parameter (p-value $<$0.001). Therefore, we conclude that driving in urban roads increases the frequency of claims. It is important to remark that these two variables should not be included in the model at the same time, as they are highly correlated (the more kilometers are travelled in urban roads, the more speed event in urban roads occur). As the AIC is lower for Model 5b compared to Model 5a, we decide to keep the lagged percentage of urban driving in the model. In Model 6, the lagged log of the total number of kilometers travelled at night is also included in the model, and it does not have a significant effect. Nevertheless, in Model 7 we observe that when we introduce the log of the current number of kilometers travelled at night, then the coefficient is positive and significant (p-value = 0.027) and the AIC is lower than the one obtained for Model 6. Therefore, we conclude that the distance travelled at night should be included in the model by using the information of the current week, and it has a positive effect on the frequency of claims: driving at night increases the accident risk. Additionally, in Model 7 we also observe that men and younger drivers have a higher claim frequency, while vehicle power does not have a significant effect (p-value = 0.076). Moreover, the lagged log of the total distance, lagged percentage of urban driving and log of the total distance travelled at night increase the claim frequency. Finally, in Model 8 we introduce the same variables as in Model 7, and, additionally, the log of the current total distance travelled. We observe that the AIC slightly increases with respect to Model 7, which means that the effect of the log of the current total distance travelled is not significant even at the 10 % significance level (p-value = 0.166).

Logistic regression and linear probability models are presented in Appendix A2. Interpretations and additional discussion are available from the authors for these two other models.

All results show that distance driven and risky event information help to improve the predictive performance of model for the expected frequency of at-fault accidents on a weekly basis using the AIC. We show that contextual information and lagged data is even more informative, showing that location and time of the day where driving improves the model.

The significance of our conclusions is important for implementing usage-based risk assessment digital tech devices. Our risk

Table 6
Poisson Regression Models: Model 3 (all non-telematics and log of distance travelled as offset), Model 4a (all non-telematics and log of distance travelled) and Model 4b (all non-telematics and lagged log of distance travelled) in telematics weekly data set, Spain 2019.

Variable	MODEL 3		MODEL 4a		MODEL 4b	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−4.863	<0.001	−6.258	<0.001	−6.212	<0.001
Vehicle_power	−0.001	0.491	0.002	0.138	0.002	0.155
Gender	0.099	0.149	0.171	0.013	0.169	0.014
Age	−0.013	0.066	−0.018	0.014	−0.018	0.015
Ln(Total distance drivenKM)	offset	–	0.126	<0.001		
Ln(Total distance drivenKM)_lag					0.150	<0.001
AIC	14789		14275		14270	
Pseudo R^2 (%)	0,042		0,233		0,270	
Residual deviance (Null deviance 12455)	12937		12421		12416	

assessment scores, especially those based on linear approximations, are convenient to create simple pricing mechanisms for insurance ratemaking. However, traditional ratemaking in motor insurance has been based on the Poisson regression model and they are multiplicative. Our results have the limitation that data on the same driver might be correlated, thus calling for using panel data analysis. If the number of days in a week that the driver uses their car is introduced in the model, then the model performance may improve, but the associated parameter is not significant due to correlations with other telematics variables, in particular the total distance.

6. Pricing schemes

In this section, we present a comparative analysis of ratemaking approaches within telematics insurance schemes. We provide distinct examples that illuminate various models and frameworks employed to determine insurance premiums based on telematics data. Through the examination of these examples, our objective is to elucidate the versatility and effectiveness of different ratemaking strategies, contributing to a deeper understanding of how telematics technology influences insurance pricing and policy design. Additionally, we illustrate how scores can be translated into prices, facilitating their integration into insurance ratemaking to either reward good drivers or penalize poor ones.

Fig. 4 shows the evolution of the mean premium paid by insureds that suffered a claim at fault (red) and those that did suffer a claim at fault (green). We see that independently of when the claim was reported, the average premium of those who reported at least one claim at fault is higher than for the rest. In Table 8 we calculate in the same dataset what the insured policy holders in our sample would pay per week (and then find the annual equivalent) under the schemes in model 0 (in Table 5) and models 3, 4a, 4b (in Table 6), 5a, 5b, 6, 7 and 8 (in Table 7). In our dataset there are 922 claims, and the average number of claims per week is 0.00117. The total sum of costs equals 2,152,040.2. Therefore, on average each claim costs 2334.1 Euros. This results in an average weekly premium of 2.72 Euros.

In Model 0 (Table 8) we calculate the pure premium based on average cost times expected claims that only depend on age, gender and vehicle power. The weekly premium ranges between 1.2 and 6.5 Euros and the yearly premium between 63.8 and 338.0 Euros. Model 3 corresponds to the Pay As You Drive scheme, where the premium is proportional to distance driven. In that case, the weekly premium ranges between 0 (for parked cars) and 73.62 Euros. Note that insurers would usually establish a minimum premium even if the vehicle is parked and is not driven. Models 4a and 4b correspond to the Pay As You Drive scheme, but distance is in logs (not lagged and lagged, respectively) and so, it is not a linear model. Now the weekly premium ranges between 0.5 and 7.2 Euros for Model 4a and 0.4 and 7.3 Euros for Model 4b. Model 5a and 5b correspond to the Pay-How-You-Drive scheme, as lagged speed events in urban areas (Model 5a) or lagged percentage of urban driving (Model 5b) are considered. In that case, the weekly premium ranges between 0.6 and 12.2 Euros for Model 5a and between 0.02 and 20.4 Euros for Model 5b. Model 6 corresponds to the Usage-Based Insurance scheme. Now, apart from the lagged percentage of urban driving, also information about time of driving is considered, specifically, the lagged distance travelled at night. In that case, the weekly premium ranges between 0.02 and 20.4 Euros. Model 7 corresponds to the same Usage-based Insurance scheme considered in Model 6, but now we consider the current log of the number of kilometers travelled at night. In that case, the weekly premium ranges between 0.02 and 21.2 Euros. Finally, in Model 8 we consider the same Usage-based Insurance scheme considered in Model 7 but the model also includes the current log of the total distance travelled. In that case, weekly premiums range between 0.02 and 19.0 Euros. In all scenarios, the total premiums paid are equal to the total sum of costs.

The Gini index in Table 8 escalates as the insurance model gains sophistication, except for Model 3, where the proportionality to distance travelled results in a stark contrast between the maximum premium and the mean premium. In this case, the Gini coefficient reaches 0.443. This discrepancy cautions against employing Poisson multiplicative specifications with distance as an offset. Instances arise where policyholders with extensive distances travelled may face exorbitant premiums, underscoring the observed inequality in premium distribution, as evidenced by the Gini coefficient.

In Appendix A3, we present calculations akin to those in Table 8, encompassing both the Linear Probability Model, characterized by

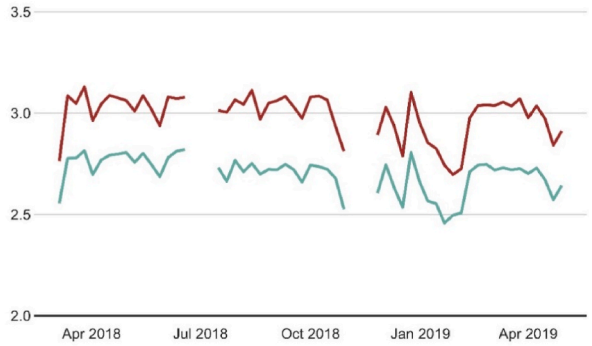


Fig. 4. Average pure premium per week under Model 8 by group of drivers in the telematics weekly data set, Spain, 2019. Those who suffered an accident during the observation period (red) and those who did not suffer an accident during the observation period (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 7
Poisson Regression Models: Model 5a (all non-telematics, distance, urban speed events), Model 5b (all non-telematics, distance, percentage of urban driving), Model 6 (all non-telematics and lagged telematics variables), Model 7 (all non-telematics, lagged total distance, lagged percentage of urban driving and current total distance travelled at night) and Model 8 all non-telematics, current and lagged total distance, lagged percentage of urban driving and current total distance travelled at night) in telematics weekly data set, Spain 2019.

Variable	MODEL 5a		MODEL 5b		MODEL 6		MODEL 7		MODEL 8	
	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−6.429	<0.001	−6.336	<0.001	−6.340	<0.001	−6.289	<0.001	−6.246	<0.001
Vehicle_power	0.001	0.203	0.002	0.097	0.002	0.097	0.002	0.103	0.002	0.116
Gender	0.160	0.020	0.156	0.023	0.158	0.022	0.144	0.037	0.144	0.038
Age	−0.016	0.031	−0.016	0.026	−0.017	0.025	−0.014	0.054	−0.014	0.052
Ln(Total distance drivenKM)									0.060	0.166
Ln(Total distance drivenKM)_lag	0.105	0.007	0.376	<0.001	0.379	<0.001	0.361	<0.001	0.330	<0.001
Speed_event_urban_lag	0.047	0.003								
Perc_urban_lag			0.013	<0.001	0.013	<0.001	0.013	<0.001	0.013	<0.001
Ln(km_nightMK)							0.010	0.027	0.009	0.059
Ln(km_nightMK)_lag					−0.001	0.802				
AIC	14263		14219		14220		14216		14216	
Pseudo R^2 (%)	0,331		0,644		0,644		0,678		0,692	
Residual deviance (Null deviance 12455)	12407		12363		12362		12358		12356	

Table 8
Range of weekly and yearly premium for different ratemaking schemes (based on the Poisson model) and Gini index for the telematics weekly data set, Spain, 2019 (in Euros).

	Weekly premium		Yearly premium		Gini
	Min	Max	Min	Max	
Model 0: Traditional risk factors	1.227	6.500	63.791	337.989	0.082
Model 3: PAYD (proporcional distance)	0.000	73.616	0.001	3828.046	0.443
Model 4a: PAYD (distance in logs)	0.541	7.155	28.149	372.058	0.107
Model 4b: PAYD (lagged distance in logs)	0.403	7.279	20.97	378.518	0.115
Model 5a: PHYD (lagged speed events urban)	0.640	12.227	33.285	635.812	0.126
Model 5b: PHYD (lagged percentage urban)	0.017	20.444	0.902	1063.113	0.175
Model 6: UBI (lagged telematics info)	0.017	20.356	0.876	1058.505	0.175
Model 7: UBI (lagged telematics info, but current log of km in the night)	0.020	21.164	1.035	1100.538	0.180
Model 8: UBI (lagged telematics info, but current log of km in the night and also current log of total km)	0.015	19.013	0.763	988.670	0.182

The sum of premiums always equals 2,152,040.2.
PAYD: Pay As You Drive. PHYD: Pay How you Drive. UBI: Usage based Insurance.

an additive structure, and the Logistic Probability Model, which adopts a non-linear framework. On one hand, we observe that the minimum weekly premiums across various scenarios are set at 0, while the maximum weekly premium in the linear models reaches 11.3. Hence, the maximum is notably lower compared to the Poisson approach, arising from the multiplicative scheme.

On the other hand, the weekly premiums derived from logistic models exhibit a range of values between 0.02 and 37.5 Euros. Despite the logistic regression model featuring a non-linear specification, it appears to strike a reasonable balance between the multiplicative and additive schemes.

7. Conclusions

In this study, we present compelling evidence that dynamic telematics factors offer easily interpretable and transparent algorithms that represent the future of dynamic driving safety assessment. We commence by considering a foundational scenario where traditional risk factors are the sole focus and subsequently compare these results with a scenario where speed events' occurrence is integrated into the model, akin to the approach taken by Guillen et al. [33]. Our distinctive contribution lies in the analysis of claims that occurred during the same time as the collection of telematics information, differentiating our approach from Guillen et al. [33] where claims were derived from historical claim data. We acknowledge the potential for correlated data within our dataset, which can present challenges for the reliability of our statistical models.

End users are typically less concerned with the intricacies of scoring calculations. Nevertheless, we advocate for a linear specification, which is often more transparent and easier to communicate to consumers. We have incorporated this suggestion into our

manuscript. Additionally, in the limitations section, we acknowledge that while more sophisticated tree-based algorithms, capable of capturing non-linear effects and interactions between factors, may be technically more accurate, they are perceived as less transparent and can be more complex. These adjustments help clarify the focus on transparency and the trade-offs between linear and tree-based models.

Our findings substantiate the significance of risky events in predicting accident occurrence, aligning with the conclusions drawn by Guillen et al. [33]. Near miss or risky events emerge as potent indicators of risky driving, encapsulating critical insights into potential accidents. Nevertheless, our study underscores the continued relevance of incorporating classical telematics variables, notably distance, time of driving, and type of road. Specifically, we ascertain that the distance travelled on urban roads and during nighttime significantly correlates with a heightened risk of accidents and should therefore be factored into the safety assessment. We acknowledge that the focus on a specific geographic region (Spain) and time period (2018–2019) may limit the generalizability of our findings to other countries and time periods. We applied three distinct modeling approaches (Poisson, Logistic and Linear Probability models), and the three of them converge on these crucial conclusions. In discussing the implications of our findings, it is essential to acknowledge the growing adoption of telematics-based insurance schemes, particularly Pay-as-You-Drive policies. These policies, which charge premiums based on actual driving behavior and distance travelled, offer potential benefits such as cost savings for low-mileage drivers. Our study underscores the importance of understanding and leveraging telematics data to refine risk assessment models and enhance insurance pricing accuracy. However, it is also critical to address ethical considerations and privacy concerns associated with the collection and use of such data, despite our use of anonymized information in this analysis. As telematics technology continues to evolve, future research should explore its broader implications for insurance practices and policyholder preferences.

Additionally, we explored models incorporating lagged predictors and demonstrated their comparable efficacy to models utilizing concurrent information. This finding is pivotal in encouraging safe driving practices, as it implies that recent historical data can be as influential as real-time data in predicting and promoting safe driving behaviors. These insights provide a comprehensive framework for leveraging telematics data in dynamic driving safety assessment, emphasizing the importance of both traditional and dynamic factors for a comprehensive understanding of risk in the domain of automobile insurance.

Data availability

The authors do not have permission to share data.

Funding

This work was supported by the Spanish Ministry of Science [grant number PID2023-146845NB C21]; [TED2021-130187B–I00]; Institució Catalana de Recerca Avançada [grant number ICREA Academia].

CRediT authorship contribution statement

Montserrat Guillen: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ana M. Pérez-Marín:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Jens P. Nielsen:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used OpenAI in order to improve language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

M.G. has received funds from insurance companies, but the funding organisations had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors declare no other potential conflicts of interest.

Acknowledgements

We acknowledge the support by Iain Robinson for English Language editing.

Appendix A1

Table A1
Summary table with the revised list of references where telematics information was used.

Authors	Year	Variables
Alrassy, P., Smyth, A. W. & Jang, J.	[32]	GPS trajectòries Timestamps Speed Hard braking/acceleration
Gao, G., Meng, S. & Wüthrich, M. V.	[48]	Latitude Longitude Heading (approaching direction of the vehicle in decimal degree) Speed Positional_Quality (indicator of the GPS signal quality) Engine_RPM Lateral acceleration Longitudinal acceleration Vertical acceleration
Li, H. J., Luo, X. G., Zhang, Z. L., Jiang, W. & Huang, S. W.	[38]	Mileage Timestamp Speed Acceleration Gear State Engine speed Tire pressure Fuel consumption Abnormal vehicle status information
Malekpour, M. R., Ghamari, S. H., Ghasemi, E., Hejaziyeganeh, S., Abbasi-Kangevari, M., Bhalla, K., Rezaei, N., Shahraz, S., Dilmaghani-Marand, A., Heydari, S. T., Rezaei, N., Lankarani, K. B. & Farzadfar, F.	[39]	GPS location Speed 3-axis acceleration data
Meuleners, L., Fraser, M., Stevenson, M. and Roberts, P.	[40]	Distance travelled Speed Harsh deceleration/acceleration
Moosavi, S. & Ramnath, R.	[34]	GPS trajectòries Speed Acceleration Angular speed Road type Day light
Gao, G., Wang, H., & Wüthrich, M. V.	[48]	Average driving time per week Speed Acceleration
Henckaerts, R., & Antonio, K.	[51]	Mileage Road type Time of day Harsh events
Meng, S., Wang, H., Shi, Y., & Gao, G.	[25]	Latitude Longitude Heading (approaching direction of the vehicle in decimal degree) Speed Positional_Quality (indicator of the GPS signal quality) Engine_RPM Lateral acceleration Longitudinal acceleration Vertical acceleration.
Duval, F., Boucher, J. P., & Pigeon, M.	[52]	Average daily distance Average daily number of trips Median of the average speeds of the trips Median of the distances of the trips Median of the maximum speeds of the trips Maximum of the maximum speed of the trips Proportion of long trips (>100 km) Time of day Week day
Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M.	[33]	Acceleration event (three intensities: 1, 2, and 3) Braking event (three intensities: 1, 2, and 3) Smartphone usage event (usage in seconds)

(*continued on next page*)

Table A1 (*continued*)

Authors	Year	Variables
Guillen, M., Pérez-Marín, A. M., & Alcañiz, M.	[53]	Number of kilometres driven at speeds above the posted limit during 2010 Total number of kilometres driven during 2010 % of kilometres driven on urban roads during 2010 % of kilometres driven at night (between midnight and 6 a.m.) during 2010
Mao, H., Guo, F., Deng, X., & Doerzaph, Z. R.	[54]	Driving hours in the study Annual Mileage (mile) Run red lights past 12 months (*) Drive sleepy past 12 months (*) Impatiently pass on right (*) Brake aggressively (*) Involved in racing (*) Nod off while driving (*) (*) Never; Rarely; Sometimes; Often; NA
So, B., Boucher, J. P., & Valdez, E. A.	[55]	Annualized percentage of time on the road Total distance driven in miles Percent of driving day x of the week: mon/tue/ …/sun Percent vehicle driven within x hrs: 2hrs/3hrs/4hrs Percent vehicle driven during x: wkday/wkend Percent of driving during x rush hours: am/pm Mean number of days used per week Number of sudden acceleration 6/8/9/ …/14 mph/s per 1000 miles Number of sudden brakes 6/8/9/ …/14 mph/s per 1000 miles Number of left turn per 1000 miles with intensity 08/09/10/11/12 Number of right turn per 1000 miles with intensity 08/09/10/11/12
Sun, S., Bi, J., Guillen, M., & Pérez-Marín, A. M.	[56]	Frequency of driving speed greater than 100 km/h Frequency of braking when the driving speed is greater than 90 km/h Frequency of cases when acceleration is greater than 6 m/s2 Frequency of cases when acceleration is less than 6 m/s2 Total driving distance (km) Total fuel consumption (L) Total number of brakes Range of driving (geographical units) Mean of speed (km/h) Mean of revolutions per minute (r/min) Mean of acceleration pedal position (%) Mean of engine fuel rate (%)
Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., & Elpidorou, V.	[31]	Percentage of kilometers travelled during night hours Percentage of kilometers travelled in urban areas Percentage of kilometers travelled at speeds above the limits Total number of kilometers travelled over one week Number of observed accelerating events over one week. Number of observed braking events over one week. Number of observed maneuvering events over one week. Percentage of kilometers travelled during night hours
Seacrist, T., Douglas, E. C., Hannan, C., Rogers, R., Belwadi, A., & Loeb, H.	[57]	Secondary tasks preceding crashes and near crashes Seven types of incident: (1) rear-ends, (2) road departures, (3) intersections, (4) side-swipe, (5) head-on, (6) animal, and (7) pedestrian/cyclist
Sun, S., Bi, J., Guillen, M., & Pérez-Marín, A. M.	[58]	Brake counts with speed >40 km/h Mean of acceleration pedal position (%) Cumulative driving distance (km) Mean of Speed (km/h) Mean of RPM Range of driving (geographical units)
Geyer, A., Kremslehner, D. and Mürmann, A.	[24]	Speeding index (average speeding above legal speed limits) Distance driven Number of car rides per day Interaction between the previous two
Pérez-Marín, A. M., Guillén, M., Alcañiz, M., & Bermúdez, L. (2019).	[59]	Number of kilometers driven at speeds above the posted limit during 2010 Total number of kilometers driven during 2010 % of kilometers driven on urban roads during 2010 % of kilometers driven at night (between midnight and 6 a.m.) during 2010
Guillén, M., Nielsen, J. P., Ayuso, M. & Pérez-Marín, A. M.	[8]	Total kilometres travelled per year Percentage of kilometres travelled at night during the year Percentage of kilometres travelled during the year above the limit Percentage of kilometres travelled in urban areas during the year

(*continued on next page*)

Table A1 (*continued*)

Authors	Year	Variables
Gao, G., & Wüthrich, M. V.	[23]	Average speed Average acceleration Braking Average change in direction (angle), by second, trip and driver.
Hu, X., Zhu, X., Ma, Y.-L., Chiu, Y.-C., & Tang, Q.	[60]	Vehicle miles travelled Average relative speed measures weighted on the duration of each trip among all the trips for one user Average speeding measures weighted on the duration of each trip among all the trips for one user Average smooth measures weighted on the duration of each trip among all the trips for one user Average hard brake measures weighted on the duration of each trip among all the trips for one user Average hard start measures weighted on the duration of each trip among all the trips for one user Worst relative speed measures among all the trips for one user Worst speeding measures among all the trips for one user Worst smooth measures among all the trips for one user Worst hard brake measures among all the trips for one user Worst hard start measures among all the trips for one user Average congestion level among all the trips for one user Average percentile duration on freeway among all the trips for one user Average percentile duration on arterial among all the trips for one user Average percentile duration on the local street among all the trips for one user Average percentile duration at peak hour among all the trips for one user Average percentile duration at off peak hour among all the trips for one user Average percentile duration at midnight among all the trips for one user
Huang, Y., & Meng, S.	[17]	Annual mileage Range of usual driving regions Irregularity of travel routes Median of trip distances Fractions of longtime driving (over 2 h) Exposure fraction between 0 and 30 km/h Exposure fraction between 30 and 60 km/h Exposure fraction between 60 and 90 km/h Exposure fraction between 90 and 120 km/h Exposure fraction on workdays Exposure fraction on peak workday mornings (7–9 a.m.) Exposure fraction on peak workday evenings (5–8 p.m.) Exposure fraction at noon (11 a.m.–2 p.m.) Exposure fraction in the evenings (8–12 p.m.) Exposure fraction at night (0–6 a.m.) Median of the average speed of trips (unit: km/h) Maximum of the recorded speed (unit: km/h) Median of the driving instability of trips Median of the comfort score of trips Bad driving events per km Lane changes per km Fractions of lane changes at high speeds (over 30 km/h) Sudden accelerations per km Fractions of sudden accelerations Sudden brakes per km Fractions of sudden brakes Sudden left turns per km Fractions of sudden left turns Sudden right turns per km Fractions of sudden right turns
Pérez-Marín, A. M., & Guillen, M.	[30]	Distance travelled during the year measured in kilometers Speed (% of kilometers travelled at speeds above the limit) Urban (% of kilometers travelled on urban roads)
Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M.	[18]	Total kilometers travelled % of total kilometers travelled in urban areas % of total kilometers above the mandatory speed limit % of total kilometers travelled at night
Gao, G., Meng, S. and Wüthrich, M.V.	[22]	GPS speed and vehicle sensor speed

(*continued on next page*)

Table A1 (*continued*)

Authors	Year	Variables
		They use velocity-acceleration heat maps. Covariates extracted from these maps are introduced in classical models.
Bian, Y., Yang, C., Zhao, J. L., & Liang, L.	[61]	Total mileage per month Nighttime driving hours per month Workday driving hours per month Monthly average speed Times of over speed (the vehicle speed is higher than road speed limits) Times of acceleration, maximum deceleration and sharp turn
Gao, G., & Wüthrich, M. V. and Gao, G., Wuthrich, M. V., & Yang, H.	[21, 62]	Speed Acceleration
Verbelen, R., Antonio, K., & Claeskens, G.	[10]	Distance driven during the policy period Number of trips (key-on, key-off) during the policy period Distance in meters driven on average during one trip Division of the distance 4 road types (urban, other, motorways and abroad) Division of the distance into 5 time slots (6h–9h30, 9h30-16h, 16h–19h, 19h–22h and 22h–6h) Division of distance into week (Monday–Friday) and weekend (Saturday and Sunday)
Ma, Y.L., Zhu, X., Hu, X. and Chiu, Y.C.	[63]	Hard brake Hard start Speeding when congestion Speeding when no congestion Speed limit $\geq$60 Speed limit $<$60 Faster relative speed Link speed $<$40 mph Link speed $\geq$40 mph Slower relative speed Link speed $<$40 mph Link speed $\geq$40 mph
Wüthrich, M. V.	[20]	Average acceleration/braking Total distance Average distance per trip (in km) Total time (in h) Average time per trip (in min) Average speed (in km/h) Median speed over trips (in km/h)
Baecke, P. and Bocca, L.	[16]	Total distance Total trip time Location distance (city, highway, abroad, other) Day time distance (low night, high AM, low day, high PM, low PM) Telematics crash Telematics crash G-force Night trip (Friday, Saturday) Rush hours trip (morning, evening) Rush hours trip start (morning, evening)
Ayuso, M., Guillén, M. & Pérez-Marin, A.M.	[11]	Distance travelled % of urban driving % of nighttime driving % of the total kilometers travelled above the mandatory speed limits
Weidner, W., Transchel, F. W., & Weidner, R.	[19]	Velocity Longitudinal and lateral acceleration
Makov, U. and Weiss, J.	[64]	Time (UTC) Latitude Longitude Average speed (MPH, since prior obs.) Accelerometer axis x readings (g-force) Accelerometer axis y readings (g-force) Accelerometer axis z readings (g-force)
Ellison, A.B., Bliemer, M.C.J. & Greaves, S.P.	[65]	Speed limit (40, 50, 60, 70, 80, 90, 100, 110 (km/h)) School Zone (binary) Rain Temporal (Binary) Time of day (1: Morning, 2: Day 3: Afternoon, 4: Night) Weekend (Binary) Outcomes: Driver Behaviour Profiles (DBPs) constructed from sec-by-sec data for speeding, accelerating and braking
Ellison, A. B., Greaves, S. P., & Bliemer, M. C.	[66]	Speed limit of road School zone Rain

(*continued on next page*)

Table A1 (*continued*)

Authors	Year	Variables
		Signalized intersection (within 25m)
		Non-signalized intersection (within 25m)
		Roundabout (within 25m)
		Time of Day: Morning, Day, Afternoon or Night
		Weekend: Saturday or Sunday
Wahlström, J., Skog, I. and Händel, P.	[67]	Speed
		Acceleration
		Strength G
Ayuso, M., Guillén, M. and Pérez-Marín, A.M.	[68, 69]	Distance travelled
		% of urban driving
		% of nighttime driving
		% of the total kilometers travelled above the mandatory speed limits
Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J. and Ohlsson, M.	[70]	Number of rapid acceleration events and their harshness
		Number of harsh braking events and their harshness
		Amount of absolute speeding
		Amount of speeding relative a location dependent limit
		Long-term speed variations around a nominal speed
		Number of abrupt steering maneuvers and their harshness
		Number of events when turning at too high speed and their harshness
		Instantaneous or trip-based energy consumption or carbon footprint
		Time duration of the trip
		Distance of the trip
		Actual time of day when making the trip
		Geographical location of the trip
Boucher, J.P., Pérez-Marín, A.M. & Santolino, M.	[5]	Number of kilometers driven by the insured in the year 2011
Paefgen, J., Kehr, F., Zhai, Y. and Michahelles, F.	[71]	Speed
		Longitudinal and lateral acceleration
Gerpott, T.J. and Berg, S.	[72]	Kilometrage: proportion with at least 20,000 km
		Type of road: proportion of highway >25 %
		Time of travel: Proportion of rush-hourc >10 %
Bolderdijk, J.W., Knockaert, J., Steg, E.M. and Verhoef, E.T.	[73]	Speeding (percentage of total distance travelled at 6 % or more above the local speed limit across all five road types)
		Type of road: (30, 50, 80, 100 and 120 km/h)
		Total distance travelled
		Distance driven on weekend nights
Farmer, C., Kirley, B. and McCartt, A.	[74]	Sudden braking/acceleration events per 100 miles driven.
		Miles not using seat belts
		Speeding by more than 10 mph per 100 miles driven
Toledo, T., Musicant, O. and Lotan, T.	[75]	Trip start and end times
		Acceleration of the vehicle (both in the lateral and longitudinal directions)
		Speed
		Vehicle location measured
		Vehicle on-board diagnostics system in order to obtain additional engine parameters
Musicant, O., Lotan, T. and Toledo, T.	[76]	Vehicle movement (longitudinal and lateral accelerations and the speed of the vehicle)
		Driver control (engine throttle and brake application and wheel-angle)
		Engine parameters (such as RPM)
		State of the vehicle safety systems (air bags, seat belts, ABS and traction control)
		Vehicle location
		Time
		Visual documentation (both inside and outside the vehicle)

Appendix A2

Table A2.1
Linear Regression Models: Model 0 (non-telematics), Model 1 (following [31], one near-miss event) and Model 2 results (with speed events as a contextual and risky event), in telematics weekly data set, Spain 2019 (coefficients have been multiplied by 10,000).

Variable	MODEL 0		MODEL 1		MODEL 2	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	14.358	<0.001	13.254	<0.001	12.356	<0.001
Vehicle_power	0.025	0.065	0.016	0.236	0.020	0.147

(*continued on next page*)

Table A2.1 (*continued*)

Variable	MODEL 0		MODEL 1		MODEL 2	
	Coef	p-val	Coef	p-val	Coef	p-val
Gender	2.120	0.009	1.914	0.018	1.915	0.018
Age	−0.216	0.009	−0.186	0.025	−0.180	0.030
Speed_event (wherever) 10^{-1} lag			3.873	<0.001		
Speed_event_urban_lag					0.866	<0.001
AIC	−3097503		−3097514		−3097523	

Table A2.2
Linear Regression Models: Model 3 (all non-telematics and log of distance travelled as offset), Model 4a (all non-telematics and log of distance travelled) and Model 4b (all non-telematics and lagged log of distance travelled) in telematics weekly data set, Spain 2019 (coefficients have been multiplied by 10,000).

Variable	MODEL 3		MODEL 4a		MODEL 4b	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	13.268	<0.001	17.19	<0.001	17.726	<0.001
Vehicle_power	0.021	0.125	0.021	0.128	0.02	0.143
Gender	2.013	0.013	2.036	0.012	2.021	0.012
Age	−0.209	0.011	−0.208	0.012	−0.207	0.012
Total distance drivenKM	5.949	0.002				
Ln(Total distance drivenKM)			1.384	<0.001		
Ln(Total distance drivenKM)_lag					1.637	<0.001
AIC	−3097511		−3097514		−3097519	

Table A2.3
Linear Regression Models: Model 5a (all non-telematics, distance, urban speed events), Model 5b (all non-telematics, distance, percentage of urban driving), Model 6 (all non-telematics and lagged telematics variables), Model 7 (all non-telematics, lagged total distance, lagged percentage of urban driving and current total distance travelled at night) and Model 8 all non-telematics, current and lagged total distance, lagged percentage of urban driving and current total distance travelled at night) in telematics weekly data set, Spain 2019 (coefficients have been multiplied by 10,000).

Variable	MODEL 5a		MODEL 5b		MODEL 6		MODEL 7		MODEL 8	
	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	15.058	<0.001	16.339	<0.001	16.304	<0.001	17.081	<0.001	17.603	<0.001
Vehicle_power	0.018	0.195	0.024	0.085	0.024	0.085	0.023	0.09	0.022	0.104
Gender	1.899	0.019	1.916	0.018	1.925	0.018	1.767	0.029	1.766	0.029
Age	−0.182	0.028	−0.195	0.018	−0.197	0.018	−0.171	0.04	−0.172	0.039
Ln(Total distance drivenKM)									0.676	0.149
Ln(Total distance drivenKM)_lag	1.078	0.011	3.870	<0.001	3.886	<0.001	3.706	<0.001	3.379	<0.001
Speed_event_urban_lag	0.656	0.001								
Perc_urban_lag			0.147	<0.001	0.148	<0.001	0.145	<0.001	0.147	<0.001
Ln(km_nightMK)							0.128	0.020	0.112	0.046
Ln(km_nightMK)_lag					−0.008	0.893				
AIC	−3097527		−3097566		−3097564		−3097570		−3097570	

Table A2.4
Logistic Regression Models: Model 0 (non-telematics), Model 1 (following [31], one near-miss event) and Model 2 results (with speed events as a contextual and risky event), in telematics weekly data set, Spain 2019.

Variable	MODEL 0		MODEL 1		MODEL 2	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−6.501	<0.001	−6.598	<0.001	−6.678	<0.001
Vehicle_power	0.002	0.069	0.001	0.238	0.002	0.149
Gender	0.180	0.009	0.163	0.018	0.163	0.018
Age	−0.019	0.010	−0.016	0.028	−0.016	0.033
Speed_event (wherever) 10^{-1} lag			0.284	<0.001		
Speed_event_urban_lag					0.065	<0.001
AIC	14286		14276		14268	

Table A2.5
Logistic Regression Models: Model 3 (all non-telematics and log of distance travelled as offset), Model 4a (all non-telematics and log of distance travelled) and Model 4b (all non-telematics and lagged log of distance travelled) in telematics weekly data set, Spain 2019.

Variable	MODEL 3		MODEL 4a		MODEL 4b	
	Coef	p-val	Coef	p-val	Coef	p-val

(*continued on next page*)

Table A2.5 (*continued*)

Variable	MODEL 3		MODEL 4a		MODEL 4b	
	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−6.589	<0.001	−6.257	<0.001	−6.21	<0.001
Vehicle_power	0.002	0.126	0.002	0.137	0.002	0.155
Gender	0.171	0.013	0.171	0.013	0.170	0.014
Age	−0.018	0.013	−0.018	0.014	−0.018	0.015
Totaldistance drivenKM	0.442	0.002				
Ln(Total distance drivenKM)			0.126	<0.001		
Ln(Total distance drivenKM)_lag					0.150	<0.001
AIC	14279		14274		14269	

Table A2.6

Logistic Regression Models: Model 5a (all non-telematics, distance, urban speed events), Model 5b (all non-telematics, distance, percentage of urban driving), Model 6 (all non-telematics and lagged telematics variables), Model 7 (all non-telematics, lagged total distance, lagged percentage of urban driving and current total distance travelled at night) and Model 8 all non-telematics, current and lagged total distance, lagged percentage of urban driving and current total distance travelled at night) in telematics weekly data set, Spain 2019

Variable	MODEL 5a		MODEL 5b		MODEL 6		MODEL 7		MODEL 8	
	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val	Coef	p-val
Intercept	−6.427	<0.001	−6.334	<0.001	−6.338	<0.001	−6.287	<0.001	−6.244	<0.001
Vehicle_power	0.001	0.203	0.002	0.097	0.002	0.097	0.002	0.103	0.002	0.116
Gender	0.160	0.020	0.157	0.023	0.158	0.022	0.144	0.037	0.144	0.038
Age	−0.016	0.031	−0.017	0.025	−0.017	0.025	−0.014	0.054	−0.014	0.052
Ln(Total distance drivenKM)									0.060	0.165
Ln(Total distance drivenKM)_lag	0.105	0.007	0.376	<0.001	0.379	<0.001	0.361	<0.001	0.330	<0.001
Speed_event_urban_lag	0.047	0.003								
Perc_urban_lag			0.013	<0.001	0.014	<0.001	0.013	<0.001	0.013	<0.001
Ln(km_nightMK)							0.010	0.027	0.009	0.058
Ln(km_nightMK)_lag					−0.001	0.802				
AIC	14262		14217		14219		14214		14215	

Appendix A3

Table A3.1

Range of weekly and yearly premium for different ratemaking schemes (based on the linear probability model) and Gini index for the telematics weekly data set, Spain, 2019 (in Euros).

	Weekly premium		Yearly premium		Gini
	Min	Max	Min	Max	
Model 0: Traditional risk factors	0.635	5.247	33.033	272.824	0.083
Model 3: PAYD (proporcional distance)	0.397	11.013	20.646	572.658	0.099
Model 4a: PAYD (distance in logs)	0	5.473	0	284.58	0.106
Model 4b: PAYD (lagged distance in logs)	0	5.516	0	286.845	0.114
Model 5a: PHYD (lagged speed events urban)	0	7.302	0	379.693	0.128
Model 5b: PHYD (lagged percentage urban)	0	7.876	0	409.534	0.167
Model 6: UBI (lagged telematics info)	0	7.868	0	409.123	0.167
Model 7: UBI (lagged telematics info, but current log of km in the night)	0	8.008	0	416.396	0.173
Model 8: UBI (lagged telematics info, but current log of km in the night and also current log of total km)	0	7.747	0	402.847	0.175

The sum of premiums always equals 2,152,040.2.

PAYD: Pay As You Drive. PHYD: Pay How you Drive. UBI: Usage based Insurance.

Table A3.2

Range of weekly and yearly premium for different ratemaking schemes (based on the logistic regression model) and Gini index for the telematics weekly data set, Spain, 2019 (in Euros).

	Weekly premium		Yearly premium		Gini
	Min	Max	Min	Max	
Model 0: Traditional risk factors	1.226	6.497	63.773	337.821	0.082
Model 3: PAYD (proporcional distance)	1.136	37.481	59.061	1949.025	0.096
Model 4a: PAYD (distance in logs)	0.541	7.150	28.124	371.81	0.107
Model 4b: PAYD (lagged distance in logs)	0.403	7.274	20.946	378.254	0.115
Model 5a: PHYD (lagged speed events urban)	0.640	12.202	33.262	634.527	0.126
Model 5b: PHYD (lagged percentage urban)	0.017	20.342	0.897	1057.803	0.175
Model 6: UBI (lagged telematics info)	0.017	20.255	0.872	1053.254	0.175

(*continued on next page*)

Table A3.2 (*continued*)

	Weekly premium		Yearly premium		Gini
	Min	Max	Min	Max	
Model 7: UBI (lagged telematics info, but current log of km in the night)	0.020	21.053	1.030	1094.779	0.180
Model 8: UBI (lagged telematics info, but current log of km in the night and also current log of total km)	0.015	18.928	0.759	984.278	0.182

The sum of premiums always equals 2,152,040.2.
PAYD: Pay As You Drive. PHYD: Pay How you Drive. UBI: Usage based Insurance.

References

[1] M. Eling, M. Kraft, The impact of telematics on the insurability of risks, J. Risk Finance 21 (2) (2020) 77–109, https://doi.org/10.1108/JRF-07-2019-0129.
[2] J. Lemaire, S.C. Park, K.C. Wang, The use of annual mileage as a rating variable, ASTIN Bulletin 46 (1) (2016) 39–69, https://doi.org/10.1017/asb.2015.25.
[3] G. Gao, S. Meng, M.V. Wüthrich, What can we learn from telematics car driving data: a survey, Insur. Math. Econ. 104 (2022) 185–199, https://doi.org/10.1016/j.insmatheco.2022.02.004.
[4] T. Litman, Distance-based vehicle insurance feasibility, costs and benefits, Victoria Transport Policy Institute 11 (2007). https://vtpi.org/dbvi_com.pdf. (Accessed 9 September 2023).
[5] J.P. Boucher, A.M. Pérez-Marín, M. Santolino, Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident, Anales Del Instituto de Actuarios Españoles 19 (3) (2013) 135–154.
[6] J.-P. Boucher, S. Côté, M. Guillen, Exposure as duration and distance in telematics motor insurance using generalized additive models, Risks 5 (4) (2017) 54, https://doi.org/10.3390/risks5040054.
[7] J.-P. Boucher, M. Denuit, M. Guillén, Risk classification for claim counts: a comparative analysis of various zeroinflated mixed Poisson and hurdle models, North Am. Actuar. J. 11 (4) (2007) 110–131, https://doi.org/10.1080/10920277.2007.10597487.
[8] M. Guillen, J.P. Nielsen, M. Ayuso, A.M. Perez-Marin, The use of telematics devices to improve automobile insurance rates, Risk Anal. 39 (3) (2019) 662–672, https://doi.org/10.1111/risa.13172.
[9] J.-P. Boucher, R. Turcotte, A longitudinal analysis of the impact of distance driven on the probability of car accidents, Risks 8 (3) (2020) 91, https://doi.org/10.3390/risks8030091.
[10] R. Verbelen, K. Antonio, G. Claeskens, Unravelling the predictive power of telematics data in car insurance pricing, J. Roy. Stat. Soc. C Appl. Stat. 67 (5) (2018) 1275–1304, https://doi.org/10.1111/rssc.12283.
[11] M. Ayuso, M. Guillen, A.M. Pérez-Marín, Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's, Risks 4 (2) (2016) 10, https://doi.org/10.3390/risks4020010.
[12] M. Ayuso, M. Guillen, J.P. Nielsen, Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data, Transportation 46 (3) (2019) 735–752, https://doi.org/10.1007/s11116-018-9890-7.
[13] M. Ayuso, M. Guillén, A.M. Pérez-Marin, Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance, Transport. Res. C Emerg. Technol. 68 (2016) 160–167, https://doi.org/10.1016/j.trc.2016.04.004.
[14] B. So, J.P. Boucher, E.A. Valdez, Synthetic dataset generation of driver telematics, Risks 9 (4) (2021) 58, https://doi.org/10.3390/risks9040058.
[15] J. Paefgen, T. Staake, F. Thiesse, Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach, Decis. Support Syst. 56 (2013) 192–201, https://doi.org/10.1016/j.dss.2013.06.001.
[16] P. Baecke, L. Bocca, The value of vehicle telematics data in insurance risk selection processes, Decis. Support Syst. 98 (2017) 69–79, https://doi.org/10.1016/j.dss.2017.04.009.
[17] Y. Huang, S. Meng, Automobile insurance classification ratemaking based on telematics driving data, Decis. Support Syst. 127 (2019), https://doi.org/10.1016/j.dss.2019.113156.
[18] J. Pesantez-Narvaez, M. Guillen, M. Alcañiz, Predicting motor insurance claims using telematics data—xgboost versus logistic regression, Risks 7 (2) (2019) 70, https://doi.org/10.3390/risks7020070.
[19] W. Weidner, F.W.G. Transchel, R. Weidner, Classification of scale-sensitive telematic observables for risk individual pricing, European Actuarial Journal 6 (1) (2016) 3–24, https://doi.org/10.1007/s13385-016-0127-x.
[20] M.V. Wüthrich, Covariate selection from telematics car driving data, European Actuarial Journal 7 (1) (2017) 89–108, https://doi.org/10.2139/ssrn.2887357.
[21] G. Gao, M.V. Wüthrich, Feature extraction from telematics car driving heatmaps, European Actuarial Journal 8 (2) (2018) 383–406, https://doi.org/10.2139/ssrn.3070069.
[22] G. Gao, S. Meng, M.V. Wüthrich, Claims frequency modeling using telematics car driving data, Scand. Actuar. J. 2 (2019) 143–162, https://doi.org/10.1080/03461238.2018.1523068, 2019.
[23] G. Gao, M.V. Wüthrich, Convolutional neural network classification of telematics car driving data, Risks 7 (1) (2019) 6, https://doi.org/10.3390/risks7010006.
[24] A. Geyer, D. Kremslehner, A. Muermann, Asymmetric information in automobile insurance: evidence from driving behavior, J. Risk Insur. 87 (4) (2020) 969–995, https://doi.org/10.1111/jori.12279.
[25] S. Meng, H. Wang, Y. Shi, G. Gao, Improving automobile insurance claims frequency prediction with telematics car driving data, ASTIN Bulletin 52 (2) (2022) 363–391, https://doi.org/10.1017/asb.2021.35.
[26] S. Arumugam, R. Bhargavi, A survey on driving behavior analysis in usage based insurance using big data, J. Big Data 6 (2019) 1–21, https://doi.org/10.1186/s40537-019-0249-5.
[27] A. Ziakopoulos, V. Petraki, A. Kontaxi, G. Yannis, The transformation of the insurance industry and road safety by driver safety behaviour telematics, Case Stud. Transport Pol 10 (4) (2022) 2271–2279, https://doi.org/10.1016/j.cstp.2022.10.011.
[28] A. Ziakopoulos, A. Kontaxi, G. Yannis, Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning, Accid. Anal. Prev. 181 (2023) 106936, https://doi.org/10.1016/j.aap.2022.106936.
[29] M. Siami, M. Naderpour, J. Lu, A mobile telematics pattern recognition framework for driving behavior extraction, IEEE Trans. Intell. Transport. Syst. 22 (3) (2020) 1459–1472, https://doi.org/10.1109/TITS.2020.2971214.
[30] A.M. Pérez-Marín, M. Guillen, Semi-autonomous vehicles: usage-based data evidences of what could be expected from eliminating speed limit violations, Accid. Anal. Prev. 123 (2019) 99–106, https://doi.org/10.1016/j.aap.2018.11.005.
[31] M. Guillen, J.P. Nielsen, A.M. Pérez-Marín, V. Elpidorou, Can automobile insurance telematics predict the risk of near-miss events? North Am. Actuar. J. 24 (1) (2020) 141–152, https://doi.org/10.1080/10920277.2019.1627221.
[32] P. Alrassy, A.W. Smyth, J. Jang, Driver behavior indices from large-scale fleet telematics data as surrogate safety measures, Accid. Anal. Prev. 179 (2023) 106879, https://doi.org/10.1016/j.aap.2022.106879, 2023.
[33] M. Guillen, J.P. Nielsen, A.M. Pérez-Marín, Near-miss telematics in motor insurance, J. Risk Insur. 88 (3) (2021) 569–589, https://doi.org/10.1111/jori.12340.
[34] S. Moosavi, R. Ramnath, Context-aware driver risk prediction with telematics data, Accid. Anal. Prev. 192 (2023) 107269, https://doi.org/10.1016/j.aap.2023.107269, 2023.

[35] L. Masello, G. Castignani, B. Sheehan, M. Guillen, F. Murphy, Using contextual data to predict risky driving events: a novel methodology from explainable artificial intelligence, Accid. Anal. Prev. 184 (2023) 106997, https://doi.org/10.1016/j.aap.2023.106997.
[36] TESLA, Safety score beta. https://www.tesla.com/support/safety-score#version-2.0, 2023. (Accessed 9 September 2023).
[37] Toyota, What is drive pulse?. https://support.toyota.com/s/article/What-is-Driver-Score-10536?language=en_US, 2023. (Accessed 14 February 2023).
[38] H.J. Li, X.G. Luo, Z.L. Zhang, W. Jiang, S.W. Huang, Driving risk prevention in usage-based insurance services based on interpretable machine learning and telematics data, Decis. Support Syst. 172 (2023) 113985, https://doi.org/10.1016/j.dss.2023.113985, 2023.
[39] M.R. Malekpour, S.H. Ghamari, E. Ghasemi, S. Hejaziyeganeh, M. Abbasi-Kangevari, K. Bhalla, N. Rezaei, S. Shahraz, A. Dilmaghani-Marand, S. Taghi Heydari, N. Rezaei, K.B. Lankarani, F. Farzadfar, The effect of Real-Time feedback and incentives on speeding behaviors using Telematics: a randomized controlled trial, Accid. Anal. Prev. 191 (2023) 107216, https://doi.org/10.1016/j.aap.2023.107216, 2023.
[40] L. Meuleners, M. Fraser, M. Stevenson, P. Roberts, Personalized driving safety: using telematics to reduce risky driving behavior among young drivers, J. Saf. Res. 86 (2023) 164–173, https://doi.org/10.1016/j.jsr.2023.05.007.
[41] X. Che, A. Liebenberg, J. Xu, Usage-based insurance—impact on insurers and potential implications for InsurTech, North Am. Actuar. J. 26 (3) (2022) 428–455, https://doi.org/10.1080/10920277.2021.1953536.
[42] J. Cheng, F.Y. Feng, X. Zeng, Pay-as-you-drive insurance: modeling and implications, North Am. Actuar. J. 27 (2) (2023) 303–321, https://doi.org/10.1080/10920277.2022.2077220.
[43] M. Eling, M. Lehmann, The impact of digitalization on the insurance value chain and the insurability of risks, Geneva Pap. Risk Insur. - Issues Pract. 43 (2018) 359–396, https://doi.org/10.1057/s41288-017-0073-0.
[44] E.W. Frees, F. Huang, The discriminating (pricing) actuary, North Am. Actuar. J. 27 (1) (2023) 2–24, https://doi.org/10.1080/10920277.2021.1951296.
[45] R. Henckaerts, M.P. Côté, K. Antonio, R. Verbelen, Boosting insights in insurance tariff plans with tree-based machine learning methods, North Am. Actuar. J. 25 (2) (2021) 255–285, https://doi.org/10.1080/10920277.2020.1745656.
[46] M. Lindholm, R. Richman, A. Tsanakas, M.V. Wüthrich, Discrimination-free insurance pricing, ASTIN Bulletin: J. IAA 52 (1) (2022) 55–89, https://doi.org/10.1017/asb.2021.23.
[47] R. Turcotte, J.P. Boucher, GAMLSS for Longitudinal Multivariate Claim Count Models, North American Actuarial Journal, 2023, pp. 1–24, https://doi.org/10.1080/10920277.2023.2202707.
[48] G. Gao, H. Wang, M.V. Wüthrich, Boosting Poisson regression models with telematics car driving data, Mach. Learn. 111 (1) (2022) 243–272, https://doi.org/10.1007/s10994-021-05957-0.
[49] E.W. Frees, G. Meyers, A.D. Cummings, Summarizing insurance scores using a Gini index, J. Am. Stat. Assoc. 106 (495) (2011) 1085–1098, https://doi.org/10.1198/jasa.2011.tm10506.
[50] J. Reig Torra, M. Guillen, A.M. Pérez-Marín, L. Rey Gámez, G. Aguer, Weather conditions and telematics panel data in monthly motor insurance claim frequency models, Risks 11 (3) (2023) 57, https://doi.org/10.3390/risks11030057.
[51] R. Henckaerts, K. Antonio, The added value of dynamically updating motor insurance prices with telematics collected driving behavior data, Insur. Math. Econ. (2022), https://doi.org/10.1016/j.insmatheco.2022.03.011.
[52] F. Duval, J.P. Boucher, M. Pigeon, How much telematics information do insurers need for claim classification? North Am. Actuar. J. 26 (4) (2022) 570–590, https://doi.org/10.1080/10920277.2021.2022499.
[53] M. Guillen, A.M. Pérez-Marín, M. Alcañiz, Percentile charts for speeding based on telematics information, Accid. Anal. Prev. 150 (2021) 105865, https://doi.org/10.1016/j.aap.2020.105865.
[54] H. Mao, F. Guo, X. Deng, Z.R. Doerzaph, Decision-adjusted driver risk predictive models using kinematics information, Accid. Anal. Prev. 156 (2021) 106088, https://doi.org/10.1016/j.aap.2021.106088.
[55] B. So, J.P. Boucher, E.A. Valdez, Cost-sensitive multi-class adaboost for un-derstanding driving behavior based on telematics, ASTIN Bulletin: J. IAA 51 (3) (2021) 719–751, https://doi.org/10.1017/asb.2021.22.
[56] S. Sun, J. Bi, M. Guillen, A.M. Pérez-Marín, Driving risk assessment using near-miss events based on panel Poisson regression and panel negative binomial regression, Entropy 23 (7) (2021) 829, https://doi.org/10.3390/e23070829.
[57] T. Seacrist, E.C. Douglas, C. Hannan, R. Rogers, A. Belwadi, H. Loeb, Near crash characteristics among risky drivers using the SHRP2 naturalistic driving study, J. Saf. Res. 73 (2020) 263–269, https://doi.org/10.1016/j.jsr.2020.03.012.
[58] S. Sun, J. Bi, M. Guillen, A.M. Pérez-Marín, Assessing driving risk using internet of vehicles data: an analysis based on generalized linear models, Sensors 20 (9) (2020) 2712, https://doi.org/10.3390/s20092712.
[59] A.M. Pérez-Marín, M. Guillén, M. Alcañiz, L. Bermúdez, Quantile regression with telematics information to assess the risk of driving above the posted speed limit, Risks 7 (3) (2019) 80, https://doi.org/10.3390/risks7030080.
[60] X. Hu, X. Zhu, Y.-L. Ma, Y.-C. Chiu, Q. Tang, Advancing usage-based insurance – a contextual driving risk modelling and analysis approach, IET Intell. Transp. Syst. 13 (3) (2019) 453–460, https://doi.org/10.1049/iet-its.2018.5194.
[61] Y. Bian, C. Yang, J.L. Zhao, L. Liang, Good drivers pay less: a study of usage-based vehicle insurance models, Transport. Res. Pol. Pract. 107 (2018) 20–34, https://doi.org/10.1016/j.tra.2017.10.018.
[62] G. Gao, M.V. Wuthrich, H. Yang, Driving risk evaluation based on telematics data, Available at: SSRN. (November 21, 2018), https://ssrn.com/abstract=3288347, 2018. (Accessed 21 November 2023).
[63] Y.L. Ma, X. Zhu, X. Hu, Y.C. Chiu, The use of context-sensitive insurance telematics data in auto insurance rate making, Transport. Res. Pol. Pract. 113 (2018) 243–258, https://doi.org/10.1016/j.tra.2018.04.013.
[64] U. Makov, J. Weiss, Predictive modeling for usage-based auto insurance, Predictive Modeling Applications in Actuarial Science (2016) 290–308, https://doi.org/10.1017/CBO9781139342681.012.
[65] A.B. Ellison, M.C.J. Bliemer, S.P. Greaves, Evaluating changes in driver behaviour: a risk profiling approach, Accid. Anal. Prev. 75 (2015) 298–309, https://doi.org/10.1016/j.aap.2014.12.018.
[66] A.B. Ellison, S.P. Greaves, M.C. Bliemer, Driver behaviour profiles for road safety analysis, Accid. Anal. Prev. 76 (2015) 118–132, https://doi.org/10.1016/j.aap.2015.01.009.
[67] J. Wahlström, I. Skog, P. Händel, Detection of dangerous cornering in GNSS-data-driven insurance telematics, IEEE Trans. Intell. Transport. Syst. 16 (6) (2015) 3073–3083, https://doi.org/10.1109/TITS.2015.2431293.
[68] M. Ayuso, M. Guillén, A.M. Pérez-Marín, The driving habits based on gender in pay-as-you-drive or usage-based insurance, Anales Del Instituto de Actuarios Españoles 20 (3) (2014) 17–32.
[69] M. Ayuso, M. Guillén, A.M. Pérez-Marin, Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance, Accid. Anal. Prev. 73 (2014) 125–131, https://doi.org/10.1016/j.aap.2014.08.017.
[70] P. Handel, I. Skog, J. Wahlstrom, F. Bonawiede, R. Welch, J. Ohlsson, M. Ohlsson, Insurance telematics: opportunities and challenges with the smartphone solution, IEEE Intelligent Transportation Systems Magazine 6 (4) (2014) 57–70, https://doi.org/10.1109/MITS.2014.2343262.
[71] J. Paefgen, F. Kehr, Y. Zhai, F. Michahelles, Driving behavior analysis with smartphones: insights from a controlled field study, in: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, 2012, https://doi.org/10.1145/2406367.2406412.
[72] T.J. Gerpott, S. Berg, Preferences for pay-as-you-drive insurance offers among residential customers in Germany – a conjoint-analytical investigation, Int. J. Serv. Technol. Manag. 17 (1) (2012) 22–53, https://doi.org/10.1504/IJSTM.2012.048037.
[73] J.W. Bolderdijk, J. Knockaert, E.M. Steg, E.T. Verhoef, Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment, Accid. Anal. Prev. 43 (3) (2011) 1181–1186, https://doi.org/10.1016/j.aap.2010.12.032.

[74] C. Farmer, B. Kirley, A. McCartt, Effects of in-vehicle monitoring on the driving behavior of teenagers, J. Saf. Res. 41 (1) (2010) 39–45, https://doi.org/10.1016/j.jsr.2009.12.002.
[75] T. Toledo, O. Musicant, T. Lotan, In-vehicle data recorders for monitoring and feedback on drivers' behaviour, Transport. Res. C Emerg. Technol. 16 (3) (2008) 320–331, https://doi.org/10.1016/j.trc.2008.01.001.
[76] O. Musicant, T. Lotan, T. Toledo, Safety correlation and implications of an in-vehicle data recorder on driver behaviour, in: Preprints of the 86th Transportation Research Board Annual Meeting, 2007. https://shorturl.at/apxC5. (Accessed 9 September 2023).