Can automobile insurance telematics predict the risk of near-miss events?

Montserrat Guillen
Dept. Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain; mguillen@ub.edu

Jens Perch Nielsen
Cass Business School, City, University of London, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom; jens.nielsen.1@city.ac.uk.

Ana M. Pérez-Marín
Dept. Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain; amperez@ub.edu

Valandis Elpidorou
Arch Reinsurance Europe Underwriting dac Ireland; valandis.elpidorou@archre.eu

Abstract

Telematics data from usage-based motor insurance provide valuable information – including vehicle usage, attitude towards speeding, time and proportion of urban/non-urban driving – that can be used for ratemaking. Additional information on acceleration, braking and cornering can likewise be usefully employed to identify near-miss events, a concept taken from aviation that denotes a situation that may have resulted in an accident. We analyze near-miss events from a sample of drivers in order to identify the risk factors associated with a higher risk of near-miss occurrence. Our empirical application with a pilot sample of real usage-based insurance data reveals that certain factors are associated with a higher expected number of near-miss events, but that the association differs depending on the type of near-miss. We conclude that nighttime driving is associated with a lower risk of cornering events, urban driving increases the risk of braking events and speeding is associated with acceleration events. These results are relevant for the insurance industry in order to implement dynamic risk monitoring through telematics, as well as preventive actions.

Key words: usage-based insurance, pay-how-you-drive, predictive models, acceleration, braking, speeding.

Acknowledgements: We would like to thank *InnovativeApplication* for the development of a mobile phone telematics device that provided us with our near-misses data. *InnovativeApplication* is a world leader in this type of data collection and we look forward to future co-operation in developing this approach to private auto insurance. Support from the Spanish Ministry and ERDF grant ECO2016-76203-C2-2-P is gratefully acknowledged. The first author gratefully acknowledges the financial support by ICREA under the ICREA Academia programme.

1. Introduction and motivation

Before the emergence of telematics, insurers had no verifiable information on the driving patterns and real vehicle usage of the insured. Driving circumstances and styles could only be determined, and then indirectly, in the specific case of an accident. Today, in contrast, telematics provides a novel source of data for risk classification before an accident, or even before a dangerous event, occurs, in what insurers refer to as a 'near-miss'. A near-miss – a name taken from aviation safety, where reports cover potentially dangerous practices or mistakes that could have led to a fatal accident – can be defined as a narrowly avoided accident, such as when a driver has to brake suddenly or make rapid steering operations (Arai et al., 2001). The occurrence of near-misses though seems to be related to a higher risk of being involved in future accidents.

Although, defining a near miss is straight forward enough, being able to verify its occurrence in real life is not. Insurance claims require that actual accidents be reported to the insurer, but near-misses can only be measured if they are well defined and measured on the spot. In the pilot dataset analyzed here, only near-miss events were observed, and no data on claims were available. As a matter of fact, we believe that near-misses, as defined in this study, could be used to predict real accident events, but at present we are not able to confirm this because we do not have real accidents in our sample. Future studies drawing on actual insurance data could therefore distinguish between a near-miss, as captured by telematics data, and an actual accident, which could exhibit similar characteristics to those of a near-miss.

This study focuses on near-misses identified in a sample of drivers that have a telematics sensor fitted in their vehicles. We estimate the number of three types of near-miss events: namely, accelerations, braking and cornering (see Section 4 for full definitions), as a function of two types of variable. First, we consider the traditional risk factors of age, gender, driving experience and vehicle power; and, second, we consider telematics information describing driving patterns, that is, urban and nighttime driving and speed behavior. Among the traditional risk factors, we conclude that age is relevant in predictions of near-miss events, but we do not see significant differences between men and women's expected risks of near-misses in our sample. Importantly, the impact of the risk factors on the expected number of near-miss events differs depending on the type of near-miss event being analyzed. Thus, it would be incorrect to model the sum of all near-miss events as opposed to each type of event separately, because the impacts are confounded. For instance, engine power presents a significant association with a higher frequency of cornering and acceleration events; nighttime driving is associated with a

lower risk of cornering events than daytime driving; urban driving is associated with a higher frequency of braking events; and, in line with expectations, excess speed increases the expected frequency of abnormal acceleration events.

These results are valuable for risk classification in insurance companies offering usage-based insurance (UBI) motor policies. Moreover, monitoring and predicting the risk of each near-miss can serve to construct alerts that can warn drivers when their levels are approaching a dangerous threshold or risk level. The conclusions are also of interest to traffic authorities concerned with accident prevention.

The rest of this paper is organized as follows. In the following section, we provide a brief description of the background to this study and of recent research in this field. In section 3, we outline the methods used. Section 4 describes the real dataset employed in the paper for the analysis of near-miss events. The model results are presented in section 5 and, finally, section 6 concludes.

2. Background

The use of telematics in the insurance industry provides insurers with information that can be used for risk classification. In addition to the traditional risk factors considered for insurance ratemaking (such as age, driving experience, type of vehicle, etc.), GPS-based technology provides a new wave of data with details about a driver's mileage, speeding, braking, cornering, and location as well as about road and traffic conditions. The insurance industry now faces the challenge of integrating this information correctly in its ratemaking schemes, which is far from straightforward. Apart from the high costs of the technology, insurers need to familiarize themselves with insurance telematics data (Ma et al., 2018) and the value of the information contained in data streams obtained from sensor sources. Moreover, in order to use telematics factors as rating factors, the response variable has to be at least evidently associated with accidents. In that sense, Quddus et al. (2002) found that a rapid acceleration, deceleration (braking) and sharp turns may increase driving risk and damage levels. Similarly, af Wåhlberg (2004) found evidences of a significant correlation between driver acceleration behavior and accident frequency. Jun et al. (2011) also found that drivers who had crash experiences tended to drive at higher speeds than crash-not-involved drivers, and concluded that there is a real potential to identify at-risk drivers based on in-vehicle data collection technologies. More recently, Bian et al. (2018) investigated how behavioral data of drivers affects driving risk and how driver behavior should affect UBI pricing schemes. Based on empirical data, Bian et al.

(2018) found that their driver risk classification model achieves a good accuracy in terms of risk-level classification. Additionally, the link between near-misses and accident risk has been investigated. Wang et al. (2015) carried out the assessment of driving risk associated to near-crash events. Their results indicated that the speed when braking and the potential crash type, among other factors, exerted the greatest influence on the driving-risk level of a near-crash.

Telematics-based data have been shown to be valuable for risk classification purposes in the insurance industry (Ayuso et al., 2014, 2016, 2018; Baecke, 2017; Gao and Wüthrich, 2019 and Gao et al., 2019), allowing insurers to consider the concept of risk exposure, no longer measured solely in terms of duration of policy coverage, but also of distance and time travelled. Although mileage was used as a ratemaking factor before telematics data became available (e.g. in US, France and Germany), telematics allows insurers to measure a driver's exact exposure so they do not have to rely on the insured's declaration on their initial application. In this sense, Boucher et al. (2017) show, using generalized additive models (GAM), that the simultaneous effect of distance travelled and exposure time on the risk of accident can be highly informative in the context of usage-based insurance. Likewise, Verbelen et al. (2018) recently analyzed a data set from a Belgian telematics product aimed at young drivers and report their development of generalized additive models and compositional predictors to quantify and interpret the effect of telematics variables on expected claim frequencies. They found that such variables increase the predictive power and render the use of gender as a rating variable redundant. Ayuso et al. (2016) obtained similar results in a data set for drivers in Spain. Telematics information has also been used to explain the excess of zeros observed in the frequency of claims. For example, Guillen et al. (2019) included the distance travelled per year as part of an offset in a zero-inflated Poisson model to predict the excess of zeros, which may reflect the fact that some insureds make little use of their vehicle. The authors showed the existence of a learning effect for large values of distance travelled, so that while drivers driving more should pay higher premiums, there should be a discount for drivers that accumulate longer distances over time. They also confirmed that speed limit violations and driving in urban areas increase the expected number of accident claims.

Ma et al. (2018) show that vehicle mileage, hard brakes, hard starts, peak time travel and speeding are strongly correlated with higher accident rates. They also find that contextual driving factors (such as driving at a speed significantly different from that of traffic flow) are also relevant risk factors. As a result, the authors show how second-by-second GPS data can be integrated into existing or new auto insurance pricing structures. They also analyze how usage-based insurance solution providers have chosen different measurements to evaluate driver performance. Among them, the authors describe the Progressive Insurance UBI program, where

a combination of hard braking (deceleration over 7 mph/s), number of miles driven, time and day, fast starts and trip regularity are used to calculate each driver's risk level. Ma et al. (2018) also examine the Allstate Drivewise program that rewards drivers that limit high speed driving, late-night trips and hard braking (in this case, driving at speeds above 80 mph is considered unsafe).

Recently, Stipancic et al. (2018) analyze hard braking and accelerating events and compare them with historical crash data. Both maneuvers are positively correlated with crash frequency at the link and intersection levels. Locations with more braking and accelerating are also associated with more collisions. Higher numbers of vehicle maneuvers are also related to increased collision severity, though this relationship is not always statistically significant. Previously, Wahlström et al. (2015) detected dangerous vehicle cornering events, based on statistics related to the no-sliding and no-rollover conditions.

Osafune et al. (2017) analyze aggressive driving behavior using a large dataset of accelerometer readings collected from drivers' smartphones. Their objective is to explore accident risk indexes that statistically separate safe drivers from risky drivers. They conclude that the frequency of acceleration exceeding 2.4 m/s^2, that of deceleration exceeding 1.4 m/s^2, and that of left acceleration exceeding 1.1 m/s^2 separate safe from risky drivers.

The distinction between accidents and near-misses has also been investigated in the context of car-to-cyclist crashes and near crashes (Ito et al., 2018). Here, the factors that differentiate near crashes from crashes are examined and the causes of the latter are identified. Ito et al. (2018) conclude that car-to-cyclist crashes are unavoidable when the car approaching the cyclist enters an area in which the average deceleration required to stop the car is more than 4.4 m/s^2. Finally, Sanders (2015) has analyzed the impact of near miss and collision experiences in the perceived traffic risk for cyclists.

According to Arai et al. (2001), near-miss data can be useful for diagnosing driving behavior and developing driving safety programs and driver assistance devices and, as such, near-miss events have attracted researcher attention in recent years. Indeed, it is our contention that insurance companies need to analyze occurrences of both accidents and near-misses and the circumstances in which they take place. In this regard, the expected number of near-misses should become a standard risk index for drivers, thus helping to personalize motor insurance rates.

3. Methods

We use the negative binomial (NB) regression to model the number of near-miss events observed over a period of time. The NB distribution is a Poisson-gamma mixture, i.e., the NB is a Poisson (λ) distribution, where λ is itself a random variable, distributed as a gamma distribution. Given the gamma parameter, the NB regression is a special type of generalized linear model where the mean of the dependent variable *y*, μ, depends on a set of *k* independent variables ($x_1, \ldots, x_k$) according to

$$\mu_i = km_i \cdot \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}).$$

where $i = 1,\ldots n$, where *n* is the sample size, km_i is the total distance travelled during the observation period (which is one week for all observations) and it is used as an offset variable, $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters that need to be estimated, and

$$\Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(\mathrm{y}_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(\mathrm{y}_i + 1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{\mathrm{y}_i}$$

where α is the inverse of the scale parameter of the gamma distribution. The parameter estimates of the NB regression model can be easily estimated by maximum likelihood using PROC GENMOD of SAS.

Here, we have a small sample of drivers that are observed over a maximum period of up to 15 weeks. This period varies from one driver to another in relation to their participation in the sample. We also consider a panel model in our analysis. In essence, the method is simply a generalization of the Poisson or negative binomial model, in which we consider time and individual fixed effects, in order to account for the driver correlations observed over time. For the sake of simplicity, we do not include any more details, but a complete overview of panel data for counts can be found in Frees (2004) and Boucher and Guillen (2009), and for specific applications to the insurance industry in Boucher et al. (2009).

4. The data set

A pilot study was conducted to collect telematics information on drivers in Greece during 2017. All drivers agreed to provide data from car sensors that measured all three types of near-miss event. A weekly summary allowed us to analyze the relationship between the response, defined as the observed number of near-miss counts of each type, and the explanatory factors, that is,

personal information (the traditional risk factors) and behavioral values (telematics covariates). The traditional risk factors include driver's gender, age, experience and vehicle age and engine power. Telematics covariates measure the total distance travelled per week, nighttime and urban driving, and provide information about speeding and the near-miss events of each type, that is, acceleration, braking and cornering events, as defined below. Table 1 presents and describes the variables in the dataset.

We have analyzed pre-defined what a near-miss event is for the purpose of this study. Unfortunately, the number of observations is far from enough if we want to let near-miss patterns to be naturally defined by searching for some specific structure in the data. The identification of each near-miss is based on the calculation of a severity score for each event type, which lies between [0,10]. For example, in the case of acceleration events, the calculation takes into consideration the difference between the maximum acceleration reading and the acceleration detected in the first reading above the acceleration event detection threshold (set at $6m/s^2$). This threshold was chosen in accordance with previous studies. Note that Hynes & Dickey (2008) considered 5.7m/s2 as the threshold for a low peak acceleration event during rear-end impacts. We calculate the ratio between this difference and the corresponding timestamps of the latter readings. The final severity score is a transformation of this ratio multiplied by 10, which means we obtain a final score between [0,10]. Acceleration is also used to determine the severity of braking events, given that negative acceleration can essentially be considered as deceleration. In the case of cornering events, severity depends on the ratio between the speed of a reading and the maximum speed possible during a turn for the vehicle to stay on track (note that this definition is similar to the no-sliding condition used by Wahlström et al., 2015, in their study on dangerous cornering events). Here acceleration events are considered near-misses because of the high severity of the event, but in real life, in most cases, an acceleration event results in a braking event rather than an accident. In this analysis, we also consider the total number of near-miss events, defined as the sum of acceleration, braking and cornering events.

The final dataset comprises 1,225 observations, corresponding to 157 drivers observed over an average period of 8 weeks in years 2016-2017. This means that the number of data points per driver equals on average 8 (the number of observed weeks). 75% of the drivers were observed during a period of 10 weeks as maximum. Table 2 shows the descriptive statistics and frequency tables of the non-telematics variables. There are 24.2% women in the sample. Almost all drivers do not have private overnight parking facility (only 1.91% have a car park). The average customer age is 43.76 years, with an age range from 16 to 68 years. On average, drivers have

17.93 years of driving experience and cars are 10.34 years old. The mean vehicle engine power is 159.56 hp.

Table 1. Variables in the near-misses dataset.

Variable	Description
Traditional risk factors	
CustomerGender	0 male 1 female
CustomerAge	Customer age
CustomerYearsHavingL	Number of years of driving experience.
VehicleAge	Age of the vehicle
VehicleEnginePower	Engine power
VehicleNightParking	Is the vehicle left in a car park over night? 0 no 1 yes
Telematics covariates	
Night	Percentage of kilometers travelled during night hours
Urban	Percentage of kilometers travelled in urban areas
Speed	Percentage of kilometers travelled at speeds above the limits
Km	Total number of kilometers travelled over one week
TotalNumberofAccelerationE	Number of observed accelerating events over one week.
TotalNumberofBrakingE	Number of observed braking events over one week.
TotalNumberofManeuveringE	Number of observed maneuvering events over one week.
Night	Percentage of kilometers travelled during night hours

Table 2 also shows the descriptive statistics for urban and nighttime driving and speeding. On average, 30% of the kilometers driven are done so at night, 13% on urban roads and 3% at speeds above the limits. The average distance travelled per week is 147.27 km.

Table 2. Descriptive statistics of the explanatory variables in the dataset of near-misses.

			By individuals			
Traditional ratemaking factors	Mean	Std. dev.	Min	Q1	Q3	Max
CustomerAge	43.76	10.49	16.00	37.00	51.00	68.00
CustomerYearsHavingLicense	17.93	9.70	0.00	12.00	23.00	49.00
VehicleAge	10.34	8.63	0.00	6.00	12.00	51.00
VehicleEnginePower	159.56	98.96	44.00	105.00	184.00	635.00
	%					
CustomerGender (female)	24.20					
VehicleNightParking	1.91					
			By weeks			
New telematics ratemaking factor	Mean	Std. dev.	Min	Q1	Q3	Max
Urban	0.13	0.13	0.00	0.05	0.16	1.00
Night	0.30	0.27	0.00	0.04	0.49	1.00
Speed	0.03	0.06	0.00	0.00	0.03	0.70
Km	147.27	139.91	0.18	38.15	217.62	1197.16

Figures 1, 2 and 3 represent the distribution of the total number of near miss events (by weeks) for the three types of event (acceleration, braking and cornering) considered here[1]. In approximately 73% of the weeks no acceleration events were recorded, while this percentage was 67.92% in the case of braking and 74% in the case of cornering events. Table 3 shows the descriptive statistics of the total number of near-miss events (by weeks) in the dataset. Braking was the most frequent near-miss event, followed by accelerations and cornering. Note that the standard deviation of braking and acceleration events is high, indicating that the drivers in the sample are quite heterogeneous with respect to these occurrences.

Figure 1. Distribution of the number of acceleration events per week.

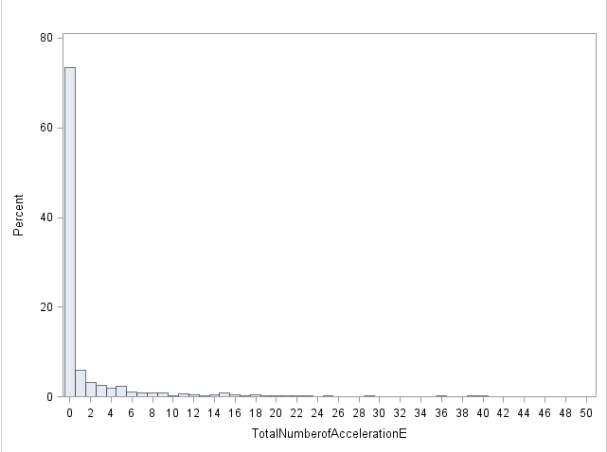
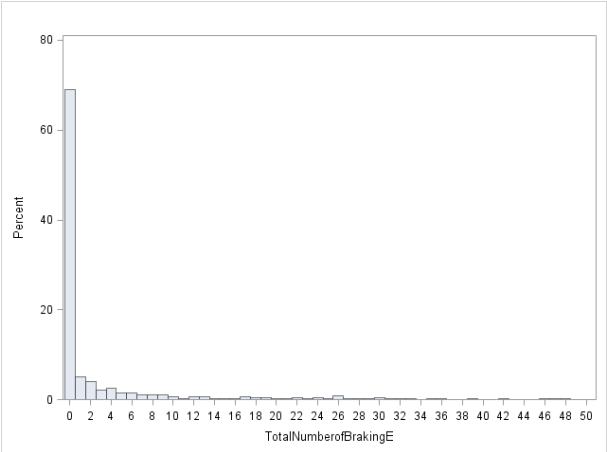


Figure 2. Distribution of the number of braking events per week.



[1] In the cases of acceleration and braking, we only considered drivers with a maximum number of near misses per week equal to 50. A few drivers presented very extreme values: thus, approximately 1% present between 50 and 290 acceleration events per week and 1.5% present between 50 and 195 braking events per week.

Figure 3. Distribution of the number of cornering events per week.

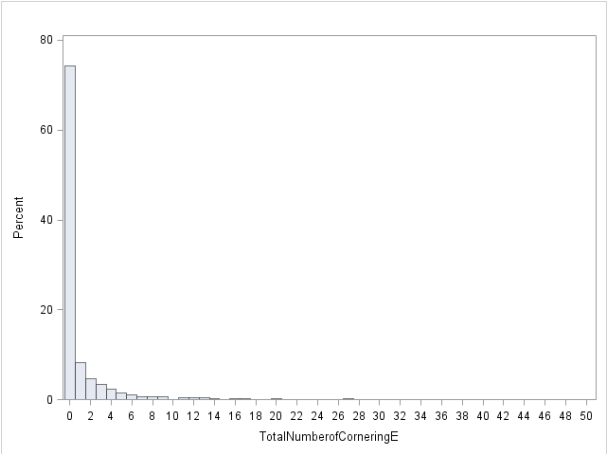


Table 3. Descriptive statistics of the total number of near-miss events in the data set.

	By weeks						
Variable	**Mean**	**Std Dev**	**Minimum**	**Lower Quartile**	**Upper Quartile**	**95th Pctl**	**Maximum**
Number of acceleration events	2.78	11.88	0	0	1	15	290
Number of braking events	4.49	12.86	0	0	2	27	195
Number of cornering events	1.10	3.14	0	0	1	6	34
Sum of the three types of event	8.37	25.03	0	0	5	44	512

5. Results

Table 4 shows the parameter estimates of the NB regression models for the acceleration, braking and cornering events, when pooling all observations in the sample. There are three model specifications: for each type of event, we consider the model with only the traditional rating factors, then with only the telematics covariates and, finally, with all the covariates. The results have been obtained by using the GENMOD procedure of SAS.

In the case of acceleration events (first three columns in Table 4), the Akaike information criterion (AIC) shows that the best model is the one that includes all the explanatory variables. Customer age and vehicle engine power are associated with a higher number of acceleration events, while vehicle age and vehicle night parking are associated with a lower number. Among the telematics variables, speed, as expected, is associated with a higher number of acceleration events. Thus, as a driver increases the percentage distance driven above the speed limits by one per cent, the expected number of acceleration events increases by about 6%. Here, the coefficient in the model with all variables is equal to 5.63, which means that $e^{5.63 \cdot 0.01} = 1.06$ which is the impact on the expected number of acceleration events. That excessive speed is associated with abrupt accelerations is unsurprising, but what is important is the magnitude of

the association when controlling for all other factors. By using the best model, the one that includes all variables, we have computed the fitted values and calculated the Chi Squared Test Statistic for a theoretical NB distribution (with an expected value for each individual equal to the corresponding fitted value), and it results in a value of 20.45.

Table 4. Negative Binomial regression model for Acceleration, Braking and Cornering Events. Parameter estimates for three models: only traditional variables, only telematics and all variables. Sample size = 1,225 observations.

	Acceleration events						Braking events						Cornering events					
	Only traditional variables		Only telematics		All variables		Only traditional variables		Only telematics		All variables		Only traditional variables		Only telematics		All variables	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	-6.1770	<.0001	-4.0349	<.0001	-6.3754	<.0001	-4.8691	<.0001	-3.5529	<.0001	-5.3354	<.0001	-6.2985	<.0001	-3.9394	<.0001	-6.0141	<.0001
CustomerGender	-0.3040	0.1916			-0.3014	0.1924	-0.0221	0.9245			-0.0186	0.9352	0.2150	0.3984			0.2256	0.3766
CustomerAge	0.0571	<.0001			0.0598	<.0001	0.0523	0.0001			0.0581	<.0001	0.0786	<.0001			0.0777	<.0001
CustomerYearsHavingL	0.0052	0.7223			-0.0038	0.8021	-0.0087	0.5331			-0.0193	0.1826	-0.0487	0.0032			-0.0526	0.0013
VehicleAge	-0.0895	<.0001			-0.0852	<.0001	-0.0798	<.0001			-0.0760	<.0001	-0.0071	0.5962			-0.0099	0.4473
VehicleEnginePower	0.0031	0.0016			0.0030	0.0018	0.0013	0.1908			0.0014	0.1449	-0.0031	0.0015			-0.0022	0.0245
VehicleNightParking	-2.5067	0.0326			-2.5970	0.0317	-1.6327	0.0297			-1.8934	0.0138	-1.5184	0.1062			-1.2649	0.1872
Night			0.0224	0.9486	-0.1044	0.7621			-0.1353	0.6764	-0.2028	0.5347			-1.3342	<.0001	-1.2201	0.0006
Urban			0.4973	0.5960	0.4976	0.5755			2.1830	0.0174	2.4071	0.0078			-0.6057	0.4662	-0.1961	0.8218
Speed			7.7295	0.0004	5.6297	0.0073			5.0161	0.0105	3.4406	0.0782			1.2529	0.5004	1.6207	0.4002
Dispersion	8.2716		9.5149		8.0622		8.0626		8.6934		7.9159		7.6532		7.7559		7.3912	
AIC	3357.7679		3426.0730		3354.6005		4205.1090		4243.9908		4199.8954		2872.5670		2883.7823		2866.5042	

In the case of braking events, the results of the parameter estimates of the negative binomial regression model are shown in the central columns of Table 4. Here, again, the model with the lowest AIC is the one that includes all the variables. It can be seen that customer age increases the number of braking events, while the older the vehicle, the lower the number of braking events. Vehicle night parking also reduces the number of braking events. The remaining traditional risk factors (CustomerGender, CustomerYearsHavingL and VehicleEnginePower) do not present a significant effect. Among the telematics variables, urban is the only factor presenting a significant effect, being associated with a higher number of braking events. This is also expected due to the density of traffic in urban areas. Again, we used the best model, the one that includes all variables, and we computed the fitted values and calculated the Chi Squared Test Statistic for a theoretical NB distribution, which results in a value equal to 20.65.

In the case of cornering events, the model presenting the lowest AIC is the one that includes both telematics and non-telematics variables. Among the traditional rating variables, customer age presents a positive and significant coefficient, indicating that cornering events are more frequent among older drivers. As expected, driving experience reduces the number of cornering events (the coefficient being significant and negative), while the greater the vehicle engine power, the lower the number of cornering events. The remaining traditional risk factors (CustomerGender, VehicleAge and VehicleNightParking) do not present a significant effect. Among the telematics variables, nighttime driving is the only factor presenting a significant effect. Driving during the night is associated with a lower expected number of cornering events, probably reflecting that drivers drive more carefully and more smoothly in the nighttime hours, compared to the daytime hours. As we did before, we used the model that includes all variables and computed the fitted values and calculated the Chi Squared Test Statistic for a theoretical NB distribution, which results in a value equal to 26.26.

When we consider the sum of near-miss events as the response variable presented in Table 5, the model results are not as clear as before and the influence of each driving pattern on the aggregate number cannot be interpreted. Table 5 shows that only the effect of excess speed is significant at the 5% level of significance for the model based on telematics covariates. Again, we used the model in Table 5 (the one that includes all variables) to compute the fitted values and calculated the Chi Squared Test Statistic for a theoretical NB distribution, which results in a value equal to 60.32. However, we recommend analyzing near-misses by type rather than in an aggregate form in order to detect the influence of urban versus non-urban driving, as well as the effects of nighttime driving.

Table 5. Negative binomial regression model for all three events considered together. Parameter estimates for three models: only traditional variables, only telematics and all variables.

	Only traditional variables		Only telematics		All variables	
	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	-4.4150	<.0001	-2.4945	<.0001	-4.6481	<.0001
CustomerGender	0.0299	0.8961			-0.0015	0.9947
CustomerAge	0.0524	<.0001			0.0598	<.0001
CustomerYearsHavingL	0.0009	0.9506			-0.0098	0.5126
VehicleAge	-0.0593	<.0001			-0.0584	<.0001
VehicleEnginePower	0.0007	0.4537			0.0009	0.3471
VehicleNightParking	-1.7454	0.0132			-1.7745	0.0157
Night			-0.4737	0.1154	-0.5404	0.0749
Urban			0.6906	0.4095	1.3074	0.1245
Speed			3.6911	0.0460	2.2134	0.2332
Dispersion	8.2819		8.8321		8.1700	
AIC	5179.1298		5216.5981		5176.9686	

As an alternative to the NB models presented here, we have also fitted a Poisson, Zero Inflated Poisson, Zero Inflated NB, Generalized Additive Model (GAM) with Poisson response and GAM regression with NB response. Tweedie was not used, as we do not have any information about costs or severities. These models have been used to estimate acceleration, braking, cornering and total number of events (four different response variables) by using traditional and telematics risk factors as explanatory variables. In the four cases, the best model was the GAM regression with NB response, as it was the one with the lowest AIC (see Table A1 in the Appendix). We also calculated the Chi Squared Test Statistic for a theoretical NB distribution with an expected value equal to the fitted values of the alternative GAM regressions. The Chi Squared Test Statistic was 34.27, 40.99 and 82.28 for acceleration, braking and total number of near-miss events, respectively. These values are higher than those corresponding to the traditional NB regression. In the case of cornering events, the Chi Squared Test Statistic was 19.71, which is lower than the corresponding to the traditional NB model. This means that, according to the Chi Squared Test Statistic, the NB model performs better than the alternative GAM regression in all cases, except for cornering events. Nevertheless, we should be careful with the interpretation of these results, as the distributions on the response variables had a heavy tail and we grouped the extreme observations in order to calculate the test statistic. As a consequence, we decided to focus on the results of the NB regression model for simplicity of interpretation of the linear component.

We also used a panel data analysis using Poisson regression, but the results are not included here (but are available from the authors on request) because, although most of the coefficient signs of the telematics variables are the same as those obtained in the regression models without a panel approach, they are quite unstable and depend heavily on the number of observed weeks considered. In most cases, no substantial changes are seen with regards to the influence of the telematics covariates, but the regressors that do not change over the weeks of observation, the case of age (in years), gender, vehicle power and age, cannot be included if individual effects have already been considered. A more sophisticated analysis with observational periods longer than 15 weeks is recommended to assess the effect of time trends on the observed responses.

6. Conclusions

The occurrence of near-miss events, and not only accidents, needs the attention of traffic authorities and insurers. Knowing the circumstances in which near-misses occur is relevant for risk quantification and also for accident prevention, given that such incidents are informative about narrowly avoided accidents and, more importantly, about the type of accident that could have occurred under a set of known circumstances.

The main conclusion to be drawn from our analysis is the different impact of a range of behavioral factors on the occurrence of different types of near-miss events. This clearly suggests that analyzing near-misses without distinguishing the type of event is likely to lead to a confounding of the factors influencing an increase in the expected number of near-misses.

In this paper we have analyzed three types of near-miss events: cornering, braking and accelerating, and we have shown that both traditional and telematics variables are relevant risk factors. Among the former, we conclude that the driver's age is associated with a higher risk of all three types of near-miss event. Specifically, older drivers present a higher risk of near-misses, perhaps owing to an excess of self-confidence at the wheel. Having said that, driving experience decreases the risk of cornering events. Among the other factors, vehicle power is associated with a higher risk of acceleration events but with a lower risk of cornering events. Finally, vehicle age is associated with a lower risk of braking and acceleration events, perhaps owing to limitations in the technical characteristics of older vehicles compared to those of newer automobiles.

Telematics risk factors have been found to be relevant for predicting the risk of each specific near-miss event. Nighttime driving is associated with a lower risk of cornering events. This is probably due to smoother driving at night, compared to daytime driving. Speeding is associated

with a higher risk of acceleration events, which is as expected. Finally, urban driving is associated with a higher risk of braking events, which is not surprising if we take into account traffic conditions in cities. We believe that these results are relevant for traffic authorities, for example, pointing to the need to promote actions encouraging drivers to maintain a safe following distance, not only on highways, but also in cities, where there is a higher risk of braking events.

Given that the average number of near-misses differs according to the event type, insurers could usefully establish benchmarks so that whenever a driver exceeds one of the factors (for instance, driving a high percentage of travelled distance in urban areas), this would trigger an alarm indicating a greater risk of near-miss events and, therefore, a higher risk of accident. However, one of the limitations of this analysis is that, while the methodology seems transferable from one portfolio to another, some of the estimated models may only be valid for the country and situation in which these data were collected.

According to the findings of this study, near-miss count data modelling shows considerable potential for the setting of personalized benchmark levels and for offering motor insurance premium rewards, based on a driver's expected number of near-misses. As such, count models can be used as predictive tools to calculate the expected level of near-miss events dynamically – that is, as the telematics measurements are processed – and drivers can be warned if the predicted levels exceed a dangerous threshold and be rewarded for good driving when near-miss counts are observed below their predicted level.

Appendix

Table A1. Parameter estimates for the GAM Regression with Negative Binomial response variable for four models: Acceleration, Braking, Cornering and Total Number of Events. All explanatory variables (traditional and telematics) are considered. Sample size = 1,225 observations.

	Acceleration Events All variables				Braking Events All variables				Cornering Events All variables				Total Number of Events All variables			
Parameter coefficients			Estimate	p-value			Estimate	p-value			Estimate	p-value			Estimate	p-value
Intercept			-4.7627	<.0001			-4.0137	<.0001			-4.9411	<.0001			-3.1317	<.0001
CustomerGender			0.2448	0.3024			0.2650	0.243			0.4173	0.0561			0.25510	0.2504
VehicleNightParking			-2.3831	0.0397			-1.9671	0.008			-2.0611	0.0237			-2.3366	0.0007
Smooth terms	EDF*	Ref. DF**	Chi Sq.	p-value	EDF*	Ref. DF**	Chi Sq.	p-value	EDF*	Ref. DF**	Chi Sq.	p-value	EDF*	Ref. DF**	Chi Sq.	p-value
CustomerAge	2.802	3.618	33.890	<.0001	3.228	4.125	40.86	<.0001	4.089	5.058	31.155	<.0001	3.636	4.547	42.76	<.0001
CustomerYearsHavingL	7.020	7.922	62.633	<.0001	7.693	8.410	76.80	<.0001	4.883	5.966	43.659	<.0001	4.982	6.038	53.78	<.0001
VehicleAge	4.215	5.003	56.889	<.0001	3.535	4.240	41.64	<.0001	1.444	1.726	2.625	0.1641	2.218	2.632	23.36	<.0001
VehicleEnginePower	7.627	8.324	82.621	<.0001	7.014	7.872	53.49	<.0001	2.728	3.433	6.698	0.1254	6.722	7.638	36.38	<.0001
Night	3.373	4.173	24.158	<.0001	3.083	3.823	22.26	0.0002	3.254	4.030	40.654	<.0001	3.333	4.117	34.15	<.0001
Urban	2.189	2.773	7.727	0.0387	3.116	3.917	27.14	<.0001	5.005	6.059	27.878	<.0001	3.582	4.480	20.14	0.0007
Speed	4.284	5.292	30.122	<.0001	4.709	5.720	19.65	0.0030	3.093	3.878	17.704	0.0014	2.940	3.666	15.80	0.0028
Dispersion	5.494				6.061				5.880				6.757			
AIC	3197.687				4072.255				2752.962				5058.856			
Chi Squared Statistic	34.267				40.096				19.710				84.280			

* Estimated degrees of freedom, ** Reference degrees of freedom.

References

af Wåhlberg, A. E. (2004). The stability of driver acceleration behavior, and a replication of its relation to bus accidents. Accident Analysis and Prevention 36, 83–92.

Arai, T., Nishimoto, T., Ezaka, Y. and Yoshimoto, K. (2001). Accidents and near-misses analysis by using video drive-recorders in a fleet test, Proceedings of the 17th International Technical Conference on the Enhanced Safety Vehicles, Amsterdam, June, 2001.

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

Ayuso, M., Guillen, M. and Perez-Marin, AM. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. Accident Analysis and Prevention, 2014; 73: 125-131.

Ayuso, M., Guillen, M. and Pérez-Marín, AM. (2016). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. Risks, 4, 2, 1-10.

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

Biana, Y., Yang, C., Zhaoc, J. L. and Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. Transportation Research Part A, 107, 20–34.

Boucher, J. P., Denuit, M. and Guillen, M. (2009). Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. Journal of Risk and Insurance, 76(4), 821-846.

Boucher, J. P. and Guillén, M. (2009). A survey on models for panel count data with applications to insurance. RACSAM-Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 103(2), 277-294.

Boucher, J.-P., Côté, S. and Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks 5 (4), 54.

Frees, E. W. (2004). Longitudinal and panel data: analysis and applications in the social sciences. Cambridge University Press.

Gao, G., Meng, S., and Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. Scandinavian Actuarial Journal, 2019(2), 143-162.

Gao, G., and Wüthrich, M. V. (2019). Convolutional neural network classification of telematics car driving data. Risks, 7(1), 6.

Guillen, M., Nielsen, J. P., Ayuso, M. and Pérez-Marín, A. M. (2019) "The use of telematics devices to improve automobile insurance rates", *Risk Analysis* 39(3), 662-672.

Hynes, L. M. and Dickey, J. P. (2008). The rate of change of acceleration: Implications to head kinematics during rear-end impacts. Accident Analysis and Prevention 40, 1063–1068.

Ito, D., Hayakawa, K.; Kondo, Y., Mizuno, K., Thomson, R., Piccinini, G. B. and Hosokawa, N. (2018). Difference between car-to-cyclist crash and near crash in a perpendicular crash configuration based on driving recorder analysis, *Accident Analysis & Prevention*, 117, 1-9.

Jun, J., Guensler, R. and Ogle, J. (2011). Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology. Transportation Research Part C, 19, 569–578.

Ma, Y. L., Zhu, X., Hu, X. and Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making, Transportation Research Part A, 113, 243–258.

Osafune, T., Takahashi, T., Kiyama, N., Sobue, T., Yamaguchi, H. and Higashino, T. (2017). Analysis of Accident Risks from Driving Behaviors, International Journal of Intelligent Transportation Systems Research, 15, 3, 192-202.

Quddus, M. A., Noland, R. B. and Chin, H. C. (2002). An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research 33, 445– 462.

Sanders, R. L. (2015). Perceived traffic risk for cyclists: The impact of near miss and collision experiences, Accident Analysis and Prevention, 75, 26-34.

Stipancic, J., Miranda-Moreno, L. and Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. Accident Analysis and Prevention, 115, 160-169.

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unraveling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society: Series C (Applied Statistics). Accepted online https://doi.org/10.1111/rssc.12283.

Wahlström, J., Skog, I. and Händel, P. (2015). Detection of dangerous cornering in GNSS-data-driven insurance telematics. IEEE Trans. Intell. Transp. Syst. 16 (6), 3073–3083.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K. and Li, K. (2015). Driving risk assessment using near-crash database through data mining of tree-based model. Accident Analysis and Prevention, 84, 54–64.