The use of telematics devices to improve automobile insurance rates

Montserrat Guillen[1*], Jens Perch Nielsen[2], Mercedes Ayuso[1], and Ana M. Pérez-Marín[1]

[1]Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain

[2]Cass Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK

[*]Address for correspondence to Montserrat Guillen, Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain

Abstract

Most automobile insurance databases contain a large number of policyholders with zero claims. This high frequency of zeros may reflect the fact that some insureds make little use of their vehicle, or that they do not wish to make a claim for small accidents in order to avoid an increase in their premium, but it might also be because of good driving. We analyse information on exposure to risk and driving habits using telematics data from a Pay-as-you-Drive sample of insureds. We include distance travelled per year as part of an offset in a zero- inflated Poisson model to predict the excess of zeros. We show the existence of a learning effect for large values of distance travelled, so that longer driving should result in higher premium, but there should be a discount for drivers that accumulate longer distances over time due to the increased proportion of zero claims. We confirm that speed limit violations and driving in urban areas increase the expected number of accident claims. We discuss how telematics information can be used to design better insurance and to improve traffic safety.

KEY WORDS: Usage based insurance; pay-as-you-drive; mileage.

Summary for Social Media:

Many automobile insurance companies offer the possibility to monitor driving habits and distance driven by means of telematics devices installed in the vehicles. This provides a novel source of data that can be analysed to calculate personalised tariffs. For instance, drivers who accumulate a lot of miles should be charged more for their insurance coverage than those who make little use of their car. However, it can also be argued that drivers with more miles have better driving skills than those who hardly use their vehicle, meaning that the price per mile should decrease with distance driven. The statistical analysis of a real data set by means of simple machine learning techniques shows the existence of a learning effect for large values of distance travelled, so that longer driving should result in higher premium, but there should be a discount for drivers that accumulate longer distances over time due to the increased proportion of zero claims. We confirm that speed limit violations and driving in urban areas increase the expected number of accident claims. We discuss how telematics information can be used to design better insurance and to improve traffic safety.

1. INTRODUCTION AND MOTIVATION

According to the World Health Organization (WHO, 2017), road traffic injuries are responsible for more than 1.2 million deaths every year. Indeed, they are the leading cause of mortality among those aged between 15 and 29, at a cost to governments of approximately 3% of their GDP. This situation is exacerbated if we contemplate the fact that from the beginning of 2013 until the end of 2015, there was a 16% increase in the number of vehicles on the world's roads.

Automobile insurance is compulsory in almost all countries and, recently, many insurance companies have begun to collect telematics data about drivers' exposure to traffic (i.e. distance driven and vehicle location) and their driving behaviour (excess speed and aggressiveness). This information can improve the insurance ratemaking process and also allows conclusions to be drawn about how to make driving safer (Ayuso, Guillen, & Nielsen, 2018, Lemaire, Park, & Wang, 2016, Paefgen, Staake, & Fleisch, 2014, Ferreira & Minikel, 2013, Paefgen, Staake, & Thiesse, 2013, Langford, Koppel, McCarthy, & Srinivasan, 2008, Sivak et al., 2007, Litman, 2005 and Edlin, 2003). New automobile insurance products (known by the acronyms PAYD, *pay-as-you-drive*, or PHYD, *pay-how-you-drive*) necessitate the introduction of a GPS device in the insured vehicle to record and store relevant information about variables that change over time, including, for example, the number of kilometres driven per day by the insured, the percentage of kilometres driven above the speed limit, and the percentage of kilometres driven at night, among others. This development represents a remarkable advance, given that, previously, automobile insurance companies could only use variables related to certain fixed characteristics of the insured (for example age,

gender, or number of years since the driver's license was issued) and the vehicle (age of the automobile, engine power, etc.).

Most automobile insurance databases contain many policyholders with zero claims. This high frequency of 'zeros' may be due to the presence of insureds that have no wish to claim for small accidents in order to avoid a premium increase or, alternatively, it might be due to the relative lack of use they make of their vehicles. If the vehicle is parked in a garage, it is not exposed to the risk of accident. Here, we analyse distance driven as a measure of exposure to risk and examine its role in the probability of an insured having zero claims. We show how to differentiate those drivers that almost never use their vehicles (and so have little exposure to the risk of an accident) from those that are good drivers, i.e. those who, despite recording high mileages, are not involved in any accidents. In what follows we refer to accidents as opposed to claims, even though we are aware that some accidents are not reported to the insurance company. Indeed, a detailed discussion of the difference between the number of accidents and the number of claims has previously been reported by Boucher, Denuit, and Guillen (2009).

We discover a positive relationship between the distance driven and the number of excess zeros observed in the number of claims. We argue that this is due to a learning effect, where good drivers are more frequent than expected among those that drive long distances. The overall effect of the driving distance variable is positive, however, even if it is true that longer driving should obviously result in higher premium, there is a discount due to the increased proportion of zeros in the frequencies, due to a learning effect. The overall effect is still an increase in the premium, however not as much as we would expect without the learning effect.

Our research is innovative because (1) we introduce telematics covariates while dealing with the excess of zeros and (2) we discuss the implications for new insurance products and traffic safety that are obtained on the basis of distance driven. Additional variables may be measured to assess the quality of drivers and in future work these new telematics signals could be much more sophisticated than distance driven.

Various studies have explored the potential of telematics when applied to risks of road accidents, beginning in 1968 with a preliminary analysis by Vickrey (1968). More recently, several papers have examined the impact of new technologies on road safety and how driving habits can be measured (Shafique & Hato, 2015, Xu et al., 2015, Ellison, Bliemer, & Greaves, 2015, Ayuso, Guillen, & Pérez-Marín, 2014, Underwood, 2013, Jun, Guensler, & Ogle, 2011, Elias, Toledo, & Shiftan, 2010 and Ayuso, Guillen and Alcañiz, 2010), while others have focused specifically on mileage and new risk factors that might be included in the ratemaking process, see Ayuso, Guillen, and Nielsen (2018) for an extended review. Recently, it has been proven that including standard telematics variables significantly improves risk assessment of insureds, therefore insurers should be able to tailor their products to the customers' risk profile (Baecke & Bocca, 2017). The objective for the insurance industry is to penalize high risk drivers with higher premiums by taking into consideration factors related to dangerous driving, including, for example, exceeding the speed limits or not respecting safety distances. We show that having information about the annual distance driven by the insured improves the ratemaking process considerably not only because it is a measure of exposure to risk, but because of the crucial role it plays in the analysis of the absence of claims, i.e. the probability of not claiming or, in other words, the probability of zero claims. See the following papers on the relevance of including distance driven as a traffic risk factor (Segui-Gomez et al., 2011 and Mercer, 1989).

In terms of methodology, Poisson regression models have traditionally been used to predict the number of automobile claims in insurance. The Poisson regression model is a special case of the generalized linear model class and serves as a benchmark model (Gourieroux, Monfort, & Trognon, 1984a and 1984b). However, various corrections have to be made when assuming that the probability of zero is larger than the probability under the Poisson assumption – a so-called excess of zeros. Various papers suggest that this excess is caused by asymmetrical information with an insured preferring not to declare a claim so as to avoid certain deductibles or the application of a bonus-malus system (Chiappori & Salanié, 2000 and Dionne & Vanasse, 1992). In this paper, we wish to differentiate those drivers that have no claims because they rarely use their vehicles during the year (in the extreme case, making no use of the vehicle at all) from those that have no claims despite being frequent drivers. To do this, we propose using a zero-inflated Poisson (ZIP) model corrected by distance (kilometres driven per year by the driver). While various studies have used ZIP models (Cameron & Trivedi, 2013, Winkelmann, 2003 and Lambert, 1992) and applied them to the context of automobile insurance (Sarul & Sahin, 2015, Boucher, Denuit, & Guillen, 2007 and Lord, Washington, & Ivan, 2005), none of these contributions has analysed the role of exposure to risk in terms of distance driven.

From an empirical point of view, we draw on a real automobile claims database for a sample of insureds. This includes individual details about annual mileage travelled and other aspects of driving behaviour, which enable us to study the effects of various indicators on the probability of making a claim. We highlight the implications of this for the design of new insurance ratemaking processes.

The rest of the paper is structured as follows. In section 2, we present the methodology used when including distance as an offset variable in the ZIP model. The database and

some descriptive results are presented in section 3 and our main results obtained with the models specified are analysed in section 4. Finally, a discussion and the main conclusions drawn from this research are presented in section 5.

2. METHODOLOGY

A Poisson regression with an offset variable is the logical way to include an exposure to the risk variable in our model. Here, therefore, we opt to use a Poisson model with offset and a two-step procedure aimed at introducing telematics data, which serves as a correction to the classical model.

Zero-inflated Poisson (ZIP) regression is a model for count data with an excess of zeros. It assumes that with probability p the only possible observation is 0, and with probability $1–p$, a Poisson (λ) random variable is observed. For example, in a different context, the same model can be used in quality control. Thus, when a manufacturing system is properly aligned, defects are nearly impossible, and the p is large. However, when the machine is misaligned, defects may occur according to a Poisson (λ) distribution. This same principle is also plausible in motor insurance when modelling the number of accidents per year. Some drivers hardly use their vehicle or use it very rarely, so for them the probability of not being involved in an accident should be large.

Both the probability of no accidents and the mean number of defects λ in the imperfect state (when people use their cars) may depend on covariates that are defined for each individual. Here, we have not included subscript i to refer to the i-th observation in a

sample of size n, to make notations easier. Sometimes p and λ are unrelated; but on other occasions p is a simple function of λ, such as $p = 1/(1 + \lambda^T)$ for an unknown constant T. In either case, ZIP regression models are easy to fit. Maximum likelihood estimates (MLE) are approximately normal in large samples, and confidence intervals can be constructed by inverting likelihood ratio tests or using the approximate normality of the MLE. The estimation can be performed with standard statistical software, such as R or SAS, but the interpretation of the results of a ZIP regression model is not straightforward. For example, Lambert (1992) reports that in an experiment involving soldering defects on printed wiring boards, two sets of conditions resulted in roughly the same mean number of defects; however, the perfect state was more likely under one set of conditions and the mean number of defects in the imperfect state was smaller under the other set. In other words, ZIP regression can show not only which conditions give the lower mean number of defects but also why the means are lower.

Notice that formally we introduce an extended model of zero claims in insurance using distance driven as the exposure to risk variable. However, while this simple model extension primarily improves understanding of zero claims, it may have another important effect. When factors other than just mileage are included in the model, then essentially the extension suggested here also serves as a bias correction. With the data provided herein, the adjustment via our extended model improved considerably when mileage was included, and only marginally when further variables were included. Finally, therefore, we opted only to include mileage in the extension of the model, thus facilitating a straightforward interpretation. In this way, the excess zeros in our extended model are simply interpreted as a function of miles driven.

In the zero part of the model, we have only a Bernoulli variable that distinguishes between the zero event (no claim) versus the non-zero event (at least one claim), so the

expectation for this binary response random variable is exactly the probability of excess zero claims, which should be limited to the [0,1] interval. For this reason, we have no offset in this part and the parameter of the log-distance is not necessarily equal to one.

Below we first introduce the simple Poisson model with and without exposure as it has traditionally been presented. Exposure, in our study, is equivalent to miles driven per year.

2.1. The Poisson model

Let us assume that given x_i, the dependent variable Y_i follows a Poisson distribution with parameter λ_i, which is a function of the linear combination of parameters and regressors, $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k\, x_{ik}$. Indeed,

$$E(Y_i|x_i) = \lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k\, x_{ik}). \quad (1)$$

The unknown parameters to be estimated are $(\beta_0, \ldots, \beta_k)$.

2.2. The Poisson model with exposure

When exposure to risk is introduced, then an offset is included in the model. Let us call T_i the exposure factor for policyholder i ($i=1,\ldots,n$), in our case $T_i=\ln(D_i)$, where D_i indicates distance travelled. Then the model can incorporate this factor as follows:

$$E(Y_i|x_i, T_i) = D_i \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k\, x_{ik}) = \exp(T_i)\lambda_i. \quad (2)$$

Under this model, the probability of zero using the Poisson distribution is calculated as follows, $\mathrm{P}(Y_i = 0) = \exp(-D_i\lambda_i)$, so it depends on the distance and, since λ_i is always positive by definition, then the probability of zero claims declines naturally as distance driven increases.

We are now ready to extend the traditional Poisson regression models above to include excess zeros via ZIP models. This extension is also introduced with and without exposure.

2.3. The zero-inflated Poisson model

In the ZIP model, the probability of zero is specified as follows:

$$\mathrm{P}(Y_i = 0) = p_i + (1 - p_i)\mathrm{P}(Y_i^* = 0), \quad (3)$$

where p_i is the probability of the perfect, zero defect state and (1-p_i) is the probability of the complementary state. The new Y^* variable follows a Poisson distribution with parameter $\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k\, x_{ik})$ and captures the claims distribution that is not contaminated by the excess of zeros. Note that p_i may depend on some covariates.

Under this model, the probability of suffering k accidents, when k is bigger than or equal to one is:

$$\mathrm{P}(Y_i = \mathrm{k}) = (1 - p_i)\mathrm{P}(Y_i^* = \mathrm{k}), \quad \mathrm{k}>0.$$

2.4. The zero-inflated Poisson model with exposure

Here we assume that p_i is the probability of an excess of zeros for the i-th observation and it is specified as a logistic regression model such that

$$p_i = \frac{\exp(\alpha_0 + \alpha_1 \ln(D_i))}{1 + \exp(\alpha_0 + \alpha_1 \ln(D_i))} \tag{4}$$

The Poisson model for Y^* is specified as follows, with an exposure, $E(Y_i^*|x_i, T_i) = D_i \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k\, x_{ik}) = D_i\lambda_i = \exp(\ln(D_i))\lambda_i = \exp(T_i)\lambda_i$. Then,

$$\mathrm{P}(Y_i = 0) = \frac{\exp(\alpha_0 + \alpha_1 \ln(D_i))}{1 + \exp(\alpha_0 + \alpha_1 \ln(D_i))} + \frac{1}{1 + \exp(\alpha_0 + \alpha_1 \ln(D_i))} \exp(-D_i\lambda_i)$$

$$P(Y_i = \mathrm{k}) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1 \ln(D_i))} (\exp(-D_i\lambda_i)) {D_i}^k {\lambda_i}^k / k!$$

Using the definition of the expectation of a discrete random variable, the expectation of the Poisson part is

$$E(Y_i|x_i, T_i) = (1 - p_i)E({Y^*}_i|x_i, T_i) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1 \ln(D_i))} D_i\lambda_i = {D^*}_i\lambda_i \tag{5}$$

where ${D^*}_i = \frac{D_i}{1+\exp(\alpha_0+\alpha_1 \ln(D_i))}$ is a transformation of the original exposure D_i. So, when we include zero-inflation there is a transformation of the exposure in the Poisson model. Let us study the transformation. If $\alpha_1 > 1$, when D_i is large then ${D^*}_i$ tends to zero, but when $\alpha_1 < 1$ then ${D^*}_i$ increases when D_i increases. On the other hand, when D_i tends to zero, ${D^*}_i$ tends to zero.

If we examine the logistic regression part (4), we observe that p_i can be understood again as a transformation of the exposure into the [0,1] interval, which tends to zero when D_i tends to zero if α_1 is positive. Moreover, the derivative of (5) with respect to D_i shows how much the expected claims would change as a function of D_i and indicates that if α_1 is significantly different from zero, then the relationship is not linear. Since insurance premiums are based on expected number of claims, this is an important result as it potentially shows that insurance prices should not necessarily be linearly proportional to distance driven.

3. DATA

We use information on the risk exposure and number of claims for 25,014 insureds with car insurance coverage throughout 2011, that is, individuals exposed to the risk for a full year. Note that in our case these data concern drivers up to a maximum age of 37, given that the insurance product was sold primarily to young drivers. Our aim is to discriminate between good and bad drivers in this portfolio segment and to identify the influence of driving short distances (Ayuso, Guillen, & Pérez-Marín, 2014). Claim frequencies are presented in Table I, with an expected value of 0.23 claims per person. Table I has information on the frequency of all reported claims. The sum of reported claims that were not at fault is 3,108, while the sum of claims at fault is 2,652. Overall 5,760 claims were reported. Descriptive statistics for the risk exposure indicator (kilometres per year) are presented in Table II, where we analyse drivers with and without claims separately. The rest of the indicators, both those derived from traditional ratemaking factors and those obtained from telematic devices, are presented in Table III,

where we also present the definitions of these variables and their main descriptive statistics.

Table I. Frequency of claims per driver (n=25,014) in the Spanish insurance dataset (all claims, at fault, and not at fault)

	Absolute frequency per driver		
Number of claims	All claims	Claims at fault	Claims not at fault
0	20,608	22,837	22,432
1	3,310	1,750	2,111
2	889	385	424
3	165	37	40
4	34	4	6
5	7	1	1
6	1	0	0

One insured driver had 6 claims, 2 were at fault and 4 where not at fault.

Table II. Descriptive statistics for the risk exposure indicator (total kilometres travelled per year in 000s)

	All Sample n = 25,014	Drivers with no claims n = 20,608 (82.4%)	Drivers with claims n = 4,406 (17.6%)
Mean	7.16	6.99	7.96
1st Quartile	4.14	4.00	4.87
Median	6.46	6.28	7.22
3rd Quartile	9.40	9.22	10.30
Standard Deviation	4.19	4.14	4.35

The results presented in Table II in relation to the annual distance travelled by the insured drivers reveal differences between those with no claims and those with claims. If we focus on the 25% of drivers that travelled the smallest distance over the year (1st quartile), we observe that the insureds that claim at least one accident drove more kilometres per year than those with no claims – the respective quartile values being 4.87 vs. 4.00. A similar pattern of behaviour is observed for the second (median) and third

quartiles with those making claims driving larger distances than those with no claims. This result was as expected and is a clear indication of a relationship between claims and distance driven.

The Mann Whitney test is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample is less than or greater than a randomly selected value from a second sample. The Mann-Whitney test shows that the differences in the mean for the exposure risk regressor (Table II), as well as for the other classical and telematic regressors (Table III) are statistically significant in the cases of drivers with no claims and drivers with claims, with the exception of vehicle age (p-value=0.331) and the percentage of kilometres driven over the speed limit squared (p-value=0.9293). Note that the normality hypothesis of these variables is rejected when using the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare the statistical distribution of two samples. From a univariate point of view, drivers that made a claim for at least one accident are, on average, younger than those that made no claim and have held their driving licence for fewer years. A similar conclusion can be drawn in the case of ownership of a powerful vehicle, where those insureds making at least one claim present a higher value than those making no claims. Unexpectedly, in the case of cars parked overnight in a garage, the percentage value is higher among those who made at least one claim than it is among those who made no claim. We would expect such cars to be safer, but it appears that this variable may be closely related to car type, with powerful, more expensive cars being kept in garages. As for the new driving behaviour indicators derived from telematics, driving at night and driving in urban areas present larger mean values in the claims group than in the no claims group.

Table III. Explanatory variables* included in the models and descriptive statistics

	Description	All sample		Drivers with no claims		Drivers with claims	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Traditional ratemaking factors							
Age	Age of the insured driver (in years)	27.57	3.09	27.65	3.09	27.18	3.10
Age^2	Age squared of the insured driver						
Male (%)	Sex of the insured driver (1 if male, 0 female)	48.91	-	48.61	-	50.32	-
Age Driving Licence	Experience of the insured driver	7.17	3.05	7.27	3.07	6.73	2.94
Vehicle age	Age of the insured vehicle	8.75	4.17	8.76	4.19	8.69	4.11
Power	Power of the insured vehicle	97.22	27.77	96.98	27.83	98.36	27.46
Parking (%)	1 if the vehicle is parked in a garage over night, 0 otherwise	77.38	-	77.21	-	78.17	-
New telematic ratemaking factors							
Km per year at night (%)	Percentage of kilometres travelled at night during the year	6.91	6.35	6.85	6.32	7.16	6.49
Km per year at night $(\%)^2$	Percentage of kilometres travelled at night squared						
Km per year over speed limit (%)	Percentage of kilometres during the year above the speed limits	6.33	6.83	6.28	6.87	6.60	6.59
Km per year over speed limit $(\%)^2$	Percentage of kilometres during the year above the speed limits squared						
Urban km per year (%)	Percentage of kilometres travelled in urban areas during the year	25.87	14.36	25.51	14.31	27.56	14.47

* In addition to risk exposure (km per year in 000s)

4. RESULTS

Tables IV and V present the zero-inflated Poisson models including exposure to risk (kilometres driven per year) as the offset variable in the models as discussed in section 2. Fig. 1 gives an overview of the estimated models.

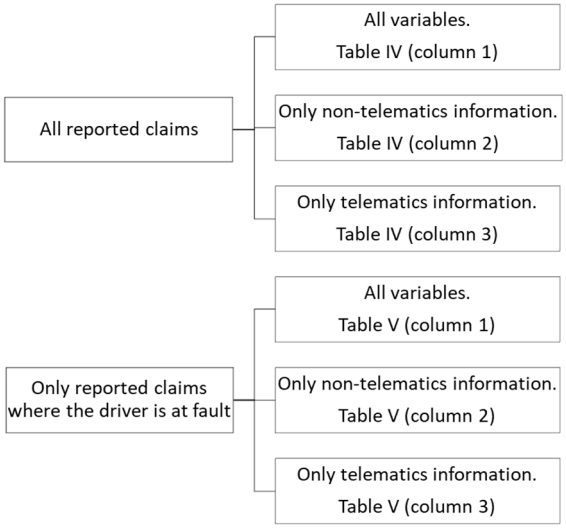

Fig. 1. Summary of the estimated models

Traditional software programs facilitate the maximum likelihood estimation of these models, their results being obtained using SAS, PROC GENMOD. To compare the models, we use the Akaike Information Criterion (AIC), calculated as twice the number of parameters in the model minus twice the value of the log-likelihood in the maximum. The best model is the one that presents the smallest AIC value.[1]

Table IV highlights a clear improvement in the results when considering all the model regressors (the lowest AIC value being obtained for the first specification). These results seem to validate the conclusions drawn in previous studies (Ayuso, Guillen, & Nielsen, 2018, Lemaire, Park, & Wang, 2016 and Ferreira & Minikel, 2013), in which the relevance of the new indicators related to distance travelled and driving habits is highlighted, but where they are used in conjunction with the classical regressors. Individual significance is observed for a large number of parameters, including those of the logit model in its zero-inflation part. On first inspection, the positive sign of the parameter associated with the log-distance in the logistics part might seem surprising and it could be interpreted erroneously. This value (0.404) in the first column does not mean that the greater the distance driven, the greater the probability of the insured having zero claims. Rather it means that the greater the distance driven, the greater the proportion of excess zero claims, indicating a deviation from the Poisson distribution that can be captured by the ZIP model.

[1] The AIC penalizes the number of parameters less strongly than the Bayesian information criterion (BIC), which is calculated on the basis of the logarithm of the number of observations as opposed to multiplying the number of parameters by two, as with the AIC.

Table IV. Zero-inflated Poisson model with offsets (Log of km per year in 000s). All types of claims.

	All variables		**(Only significant)**		**Non-telematics**		**Telematics**	
	Coefficient	(p-value)	Coefficient	(p-value)	Coefficient	(p-value)	Coefficient	(p-value)
Poisson part								
Intercept	-2.148	0.045	-3.396	<.001	-0.829	0.440	-3.461	<.001
Age	-0.094	0.232			-0.123	0.121		
Age^2	0.002	0.221			0.002	0.131		
Male	-0.068	0.029	-0.074	0.017	-0.011	0.719		
Age Driving Licence	-0.059	<.001	-0.056	<.001	-0.067	<.001		
Vehicle Age	0.014	<.001	0.014	<.001	0.017	<.001		
Power	0.003	<.001	0.003	<.001	0.001	0.017		
Parking	0.029	0.420			0.032	0.381		
Log of km per year (thousands) - offset	1.000	--	1.000	--	1.000	--	1.000	--
Km per year at night (%)	-0.004	0.312					-0.001	0.771
Km per year at night $(\%)^2$	0.0001	0.467					0.000	0.931
Km per year over speed limit (%)	0.019	0.001	0.019	0.001			0.018	0.001
Km per year over speed limit $(\%)^2$	-0.001	0.001	-0.001	0.001			-0.001	0.003
Urban km per year (%)	0.026	<.001	0.026	<.001			0.027	<.001
Zero-inflation part								
Intercept (Logit)	-0.847	<.001	-0.857	<.001	-1.639	<.001	-0.795	<.001
Log of km per year (thousands) (Logit)	0.404	<.001	0.410	<.001	0.824	<.001	0.406	<.001
AIC	28,877.112		28,870.556		29,427.423		29,005.172	
BIC	28,999.019		28,951.828		29,508.694		29,070.189	

In the case of the classical variables, all the parameters for gender, driving experience, vehicle age and the power of the vehicle are statistically significant. Thus, we find an increasing expectation in the number of claims for women drivers as opposed to men, inexperienced drivers as opposed to experienced, and owners of old and powerful vehicles as opposed to owners of newer and less powerful cars. As for the new telematic regressors, two – the percentage of kilometres per year driven over the speed limit and the percentage of urban kilometres driven per year – are significant in explaining the expected number of claims. Thus, the number of claims increases as these two regressors increase. No significance is observed in the case of night driving. In column

2, we present the estimation results of the reduced model when removing the covariates with insignificant coefficients in the full model. Finally, if we compare the results of the third and fourth specifications (columns 3 and 4, respectively), the best results are obtained for the model that only includes variables related to driving habits (telematics), as indicated by its lower AIC value.

Our model predicts the highest number of expected claims for younger women, with little driving experience, driving old and powerful vehicles, driving in urban zones, and exceeding the speed limit. Note that this result is in line with the results reported by Mercer (1989).

Previous research (Mercer, 1987) has shown that it may be interesting to include Age and Gender interaction in the model. The results for all the models, which are available from the authors, show that this interaction is not significant. In practice, Gender cannot be used for pricing insurance in the EU, but it can certainly be used for risk evaluation and it can help to understand male/female differences with implications on traffic safety. Our conclusion for this sample is that there is no interaction between Age and Gender. There are potentially two reasons for that. (1) The sample consists of drivers aged less than 37 years, so Age may not have enough range to show a significantly different effect by Gender . (2) As found by other authors, the influence of Gender is masked by the fact that men on average drive significantly longer distances than women. The relationship between distance driven and Gender was discovered by independent researchers in different EU countries considering average daily distance in a Spanish dataset (Ayuso, Guillen, & Pérez-Marín, 2016), or using average trip distance for a Belgian sample (Verbelen, Antonio, & Claeskens, 2018) or even taking both average trip distance and total distance in another European portfolio sample (Wüthrich,

2017). They all concluded that Gender differences in the risk of accidents are, to a large extent, attributable to the fact that men drive longer average distances than women.

Similar results are obtained when only claims at fault are considered in Table V, with the exception that the age of the driver is now significant while gender is not. Here, again, a better goodness of fit is obtained for the specification that includes all variables (both telematic and non-telematic) and the model that includes only the telematics variables (the lowest AIC value being obtained for served column 1). As in Table IV, a lower AIC is obtained for the specification using only telematic variables as opposed to that using only classical variables (columns 2 and 3, respectively).

Table V. Zero-inflated Poisson model with offsets (Log of km per year in 000s). Claims for which the policyholder was at fault

	All variables		**Non-telematics**		**Telematics**	
	Coefficient	(p-value)	Coefficient	(p-value)	Coefficient	(p-value)
Poisson part						
Intercept	-0.697	0.653	0.278	0.857	-3.892	<.001
Age	-0.224	0.050	-0.224	0.049		
Age^2	0.004	0.039	0.004	0.045		
Male	0.000	0.998	0.076	0.093		
Age Driving License	-0.083	<.001	-0.088	<.001		
Vehicle Age	0.013	0.015	0.016	0.004		
Power	0.001	0.163	0.001	0.351		
Parking	-0.035	0.497	-0.025	0.637		
Log of km per year (thousands) - offset	1.000	--	1.000	--	1.000	--
Km per year at night (%)	0.0052	0.386			0.010	0.083
Km per year at night $(\%)^2$	-0.0001	0.685			-0.0002	0.272
Km per year over speed limit (%)	0.035	<.001			0.031	0.000
Km per year over speed limit $(\%)^2$	-0.001	0.001			-0.001	0.001
Urban km per year (%)	0.024	<.001			0.026	<.0001
Zero-inflation part						
Intercept (Logit)	-0.228	0.151	-0.765	<.001	-0.140	0.358
Log of km per year (thousands) (Logit)	0.442	<.001	0.743	<.001	0.441	<.001
AIC	16,912.217		17,125.313		17,004.642	
BIC	17,034.124		17,206.584		17,069.659	

The age of at-fault drivers is inversely related to the expected number of claims, that is, a higher number of accidents are expected among younger drivers. However, the significance of the age squared parameter indicates a non-linear relationship between the two variables. Inexperienced drivers (measured in terms of the number of years in which they have been in possession of a driving licence) and drivers of old vehicles show a higher expected number of claims than that recorded by their more experienced counterparts and drivers of newer vehicles. In common with the result in Table IV, the percentage of kilometres per year driven over the speed limit, and additionally here the percentage of kilometres driven at night, have an impact on the expected number of claims in which the driver is at fault. The percentage of kilometres driven at night is significant at the 10% level when we only consider the telematic variables but the AIC value for this model is lower than that obtained for the first model.

Results for the models on the not at fault claims indicate similar conclusions. We have not discussed the not at fault cases because in insurance premium calculation only claims at fault are of main interest. Claims at fault indicate that the driver has caused an accident, while not at fault means that the accident was due to someone else. If the accident is caused by someone else, then the insured driver should not pay a higher insurance premium compared to someone who did not report a claim.

Comparisons with the classical Poisson model with offsets (without considering zero inflation), both for the total sample and for claims where the policyholder is at-fault, are not included here, but they do not enable us to see the impact of distance on the excess of zeros. These results are available on request from the authors. The goodness of fit results are always better in the zero-inflated models because they take into account

differences between false zeros (non-risk exposure) and true zeros (risk exposure and zero claims).

In a similar context, it has been shown that prediction models for hurricane power outage can be improved by a new two-step outage prediction model and the inclusion of additional environmental variables that increase the overall accuracy (McRoberts, Quiring, & Guikema, 2016). Our model also improves the classical approach by introducing telematics information into the prediction of the number of claims and this can be done in a two stage model approach (Ayuso, Guillen, & Nielsen, 2018).

In addition to the results presented in Tables IV and V, we have performed a hold-out analysis, and we have tested the models against test sets which were not used in the training process. We have chosen a 70% training sample, versus a 30% hold-out sample. In all cases we have confirmed the conclusions on the significance of the parameter that we had in the initial analysis. The Chi squared test of differences between observed and fitted frequencies was equal to 946.7 for the whole sample. The hold out analysis indicates very similar values (1,041.3 with 6 degrees of freedom in the training sample and 1,005.9 with 6 degrees of freedom in the test sample for the model of all claims and all variables). We find analogous results for other predictive performance measures at policyholder level, such as the Gini index (Frees, Meyers, & Cummings, 2011), which is equal to 82.4% in the whole sample while it equals 82.5% and 82.1% in the training and test samples, respectively.

In order to evaluate the variable importance, we have estimated the models using standardized covariates, so that we can compare the coefficients. This analysis reveals that the most important factor that determines the risk of a crash is the percentage in urban driving, followed by the age of the driver's license. The third factor is the

percentage of speed limit violation. The least relevant factors are the age of the vehicle, gender of the driver, percent of night distance driven and parking in a garage.

5. CONCLUSIONS

We have shown that the part of the zero accident frequency not explained by traditional insurance risk factors increases with the distance driven by the policyholder. This means that when considering policyholders with the same characteristics but with different exposures to risk in terms of distance driven per year, we can conclude that those with a greater exposure present a larger proportion of excess zero claims than those with less exposure. This can be understood as an indication of a learning effect, or in terms of distance driven, that even if exposure to risk increases with distance driven, the probability of not making a claim also increases compared to that of drivers in the group that drive a shorter distance. This finding is evidence of the fact that good drivers – if we identify them with those reporting no claims – are more frequent than expected among the group of drivers that drive long distances than among those that drive shorter distances, all other things being equal.

This conclusion has a direct impact on the future design of PAYD insurance products, insofar as the premium paid should not be strictly proportional to the distance driven. Moreover, the premium should take into account the learning effect analysed here. One possible solution would be to make the marginal increase in the insurance price per kilometre driven dependent on the accumulated distance. Here, we have shown that this relationship is not linearly dependent, as we report that the zero-inflation part plays a

significant role. Taking the derivative of (5) makes this non-linearity immediately apparent.

The probability of excess zeros increases with distance. The coefficient for the logarithm of the number of kilometres driven per year in the logit model (which predicts zero inflation) is positive, i.e. the probability of observing false zeros increases with increasing distance. Moreover, we have shown that the ZIP model gives better results in terms of goodness of fit than those obtained with the classical Poisson model (non-zero-inflated Poisson model).

Here, therefore, we have shown both the significance of the impact of the distance variable coefficient and the positive relationship between traffic violations involving excess speed and urban driving with the expected number of claims. These results are in line with reports issued by official traffic institutions where it is argued that speed limit violations should be considered in the design of insurance premiums so that safer driving is rewarded (Ayuso, Guillen, & Alcañiz, 2010).

Previous traffic studies published in Risk Analysis (Segui-Gomez et al., 2011 and Mercer, 1989) have stressed the desirability of including risk exposure in terms of distance driven. We have shown that indeed vehicle telemetry, and the collection of information using GPS-based technology such as percentages of kilometres driven at night, over the speed limit, and in urban zones, among others can be included in the ratemaking process thus improving the results obtained when just using classical driver variables, such as age and gender. This opens the question whether pay-as-you drive should also consider a different price per mile depending on the time of the day and the location.

Our study shows that ZIP models with mileage as their offset variable can improve the definition of drivers' risk profiles and provide valuable policy guidelines that might be implemented to improve driving behaviour. Furthermore, the higher premium associated with a higher percentage of kilometres driven in an urban area (as a consequence of a higher expected number of claims) could discourage the use of private vehicles in cities, as called for by various European institutions (not least to reduce levels of pollution). Clearly, similar conclusions can be drawn in terms of traffic violations, with an increase in the premium for drivers with a tendency to exceed the speed limit.

ACKNOWLEDGMENTS

The study was supported by ICREA Academia, the Spanish Ministry of Economy and Competitiveness, and the ERDF under grants ECO2015-66314-R and ECO2016-76203-C2-2-P. We thank three anonymous reviewers and the editor for insightful suggestions.

REFERENCES

Ayuso, M., Guillen, M., & Alcañiz, M. (2010). The impact of traffic violations on the estimated cost of traffic accidents with victims. *Accident Analysis and Prevention*, 42(2), 709-717.

Ayuso, M., Guillen M., & Nielsen, J. P. (2018). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*. Accepted, online https://doi.org/10.1007/s11116-018-9890-7.

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis and Prevention*, 73, 125-131.

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

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

Boucher, J. P., Denuit, M., & Guillen, M. (2007). Risk classification for claim counts: a comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, 11(4), 110-131.

Boucher, J. P., Denuit, M., & 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.

Cameron, A. C., & Trivedi , P. K. (2013). *Regression analysis of count data*. Cambridge, USA: University Press, 2nd Ed.

Chiappori, P. A., & Salanié, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy*, 108(1), 56-78.

Dionne, G., & Vanasse, C. (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics*, 7(2), 149-165.

Edlin, A. S. (2003). Per-mile premiums for auto insurance. In R. Arnott, B. Greenwald, R. Kanbur & B. Nalebuff B (Eds.), *Economics for an imperfect world: essays in honor of Joseph E. Stiglitz* (p.53). Cambridge, MA: MIT Press.

Elias, W., Toledo, T., & Shiftan. Y. (2010). The effect of daily-activity patterns on crash involvement. *Accident Analysis and Prevention*, 42(6), 1682-1688.

Ellison, A. B., Bliemer, M. C. J., & Greaves, S.P. (2015). Evaluating changes in driver behaviour: a risk profiling approach. *Accident Analysis and Prevention*, 75, 298-309.

Ferreira Jr, J., & Minikel, E. (2013). Measuring per mile risk for Pay-As-You-Drive auto insurance. *Transportation Research Record: Journal of the Transportation Research Board*, 2297(1), 97-103.

Frees, E. W., Meyers, G., & Cummings, D. A. (2011). Summarizing insurance scores using the Gini index. *Journal of the American Statistical Association*, 106(495), 1085-1098.

Gourieroux, C., Monfort, A., & Trognon, A. (1984a). Pseudo maximum likelihood methods: Theory. *Econometrica*, 681-700.

Gourieroux, C., Monfort, A., & Trognon, A. (1984b). Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica*, 701-720.

Jun, J., Guensler, R., & 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: Emerging Technologies,* 19(4), 569-578.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34(1), 1-14.

Langford, J., Koppel, S., McCarthy, D., & Srinivasan, S. (2008). In defence of the 'low-mileage bias'. *Accident Analysis and Prevention*, 40(6), 1996-1999.

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin*, 46(1), 39-69.

Litman, T. (2005). Pay-As-You-Drive pricing and insurance regulatory objectives. *Journal of Insurance Regulation*, 23(3), 35-53.

Lord, D., Washington, S., & Ivan, J. (2005). Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, 37, 35-46.

McRoberts, D. B., Quiring, S. M., & Guikema, S. D. (2016). Improving hurricane power outage prediction models through the inclusion of local environmental factors. *Risk Analysis*. First published: 25 October 2016, https://doi.org/10.1111/risa.12728.

Mercer, G. W. (1987). Influences on passenger vehicle casualty accident frequency and severity: unemployment, driver gender, driver age, drinking driving and restraint device use. *Accident Analysis and Prevention*, 19, 231-236.

Mercer, G. W. (1989). Traffic accidents and convictions: group totals versus rate per kilometer driven. *Risk Analysis*, 9(1), 71-77.

Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modelling of accident risk: insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61, 27-40.

Paefgen, J., Staake, T., & Thiesse, F. (2013) Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach. *Decision Support Systems*, 56, 192-201.

Sarul, L. S., & Sahin, S. (2015). An application of claim frequency data using zero inflated and hurdle models in general insurance. *Journal of Business, Economics and Finance*, 4(4), 732-743.

Segui-Gomez, M., Lopez-Valdes, F. J., Guillen-Grima, F., Smyth, E., Llorca, J. &, Irala, J. (2011). Exposure to traffic and risk of hospitalization due to injuries. *Risk Analysis*, 31(3), 466-474.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, 42(1), 163-188.

Sivak, M., Luoma, J., Flannagan, M. J., Bingham, C. R., Eby, D. W., & Shope, J. T. (2007). Traffic safety in the U.S.: re-examining major opportunities. *Journal of Safety Research*, 38(3), 337-355.

Underwood, G. (2013). On-road behaviour of younger and older novices during the first six months of driving. *Accident Analysis and Prevention*, 58, 235-243.

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.

Vickrey, W. (1968). Auto accidents, tort law, externalities and insurance: an economist's critique. *Law and Contemporary Problems*, 33(3), 464-487.

Winkelmann, R. (2003). *Econometric analysis of count data*. Berlin, Germany: Springer-Verlag, 4th Ed.

World Health Organization (2017). "10 facts on global road safety. Updated July 2017" available at http://www.who.int/features/factfiles/roadsafety/en/.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1), 89-108.

Xu, Y., Shaw, S. L., Zhao, Z., Yin, L., Fang, Z., &Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: a home based approach. *Transportation*, 42(4), 625-646.