risks

Article

Exposure as Duration and Distance in Telematics Motor Insurance Using Generalized Additive Models

Jean-Philippe Boucher [1], Steven Côté [1] and Montserrat Guillen [2,*]

[1] Département de Mathématiques, Université du Québec à Montréal (UQAM); Montréal, QC H3C 3P8, Canada; boucher.jean-philippe@uqam.ca (J.-P.B.); cote.steven@courrier.uqam.ca (S.C.)

[2] Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, 08007 Barcelona, Spain

* Correspondence: mguillen@ub.edu; Tel.: +34-934-037-039

Received: 16 July 2017; Accepted: 22 September 2017; Published: 25 September 2017

Abstract: In Pay-As-You-Drive (PAYD) automobile insurance, the premium is fixed based on the distance traveled, while in usage-based insurance (UBI) the driving patterns of the policyholder are also considered. In those schemes, drivers who drive more pay a higher premium compared to those with the same characteristics who drive only occasionally, because the former are more exposed to the risk of accident. In this paper, we analyze the simultaneous effect of the distance traveled and exposure time on the risk of accident by using Generalized Additive Models (GAM). We carry out an empirical application and show that the expected number of claims (1) stabilizes once a certain number of accumulated distance-driven is reached and (2) it is not proportional to the duration of the contract, which is in contradiction to insurance practice. Finally, we propose to use a rating system that takes into account simultaneously exposure time and distance traveled in the premium calculation. We think that this is the trend the automobile insurance market is going to follow with the eruption of telematics data.

Keywords: pricing; statistical inference; GAM models; pay-as-you-drive; usage-based insurance; GPS-based insurance; accident risk

1. Introduction

Automobile insurance premiums are normally calculated for one-year contracts, although the payment can be split and shorter coverage periods are possible. Therefore, for two insurees with identical characteristics, a one-year contract is normally considered twice as dangerous as a 6-month contract, basically because only the exposure period is taken into account when estimating the number of claims. Technically, in order to understand this approach we should consider the Poisson distribution, which is used for modeling the frequency of insurance claims (see Boucher 2013; Boucher et al. 2009; Boucher and Guillen 2009). Indeed, if we assume that the waiting time between two claims follows an exponential distribution of parameter λ, then the number of claims that would have occurred up to time t follows a Poisson distribution mean λt.

The question we ask ourselves now is whether the true risk exposure must be calculated solely on the basis of the duration of the contract in order to fix the premium, or whether other elements should be taken into account, in addition to the classical risk factors regarding the driver and the vehicle, which are traditionally considered in motor insurance ratemaking. Clearly, vehicle usage is the obvious risk factor explaining the risk of an accident. In fact, an insuree driving very frequently is expected to have more car accidents compared to an insuree using his car only occasionally. A way to quantify vehicle usage is to measure distance traveled per year, in kilometers or miles. However, there is possibly a learning curve that makes frequent drivers less prone to suffer accidents than occasional drivers due to the development of skills and the increase of expertise.

Several studies in the past showed that there exists a significant relationship between the number of kilometers and the risk of being involved in an automobile accident (Litman 2005; Langford et al. 2008; Boucher et al. 2013; Lemaire et al. 2016). This issue is addressed in this paper, but our main contribution is to analyze the simultaneous impact of the distance traveled and exposure time on the risk of claims.

1.1. Automobile Insurance Pricing Based on Vehicle Usage

Before studying the relationship between risk exposure and vehicle usage, it is appropriate to summarize what has been previously proposed in the literature. Vickrey (1968) is a first important contribution in this regard. In fact, the author argues against the classical pricing system of automobile insurance, saying that the premium calculation ignores the intensity of vehicle usage, which is a factor closely related to the risk of an accident. This can cause unfair insurance premiums charged to policyholders, because those who do not use the car very often have no discount compared to those who make a more intensive use of it.

Regarding the introduction of the distance traveled in the rating system, the contribution of Lourens et al. (1999) concludes that regardless of the age of the insuree, the higher the distance traveled, the higher the number of occasions when the insuree is involved in a car accident. Litman (2005) reaches a similar conclusion and establishes a positive and a non-linear relationship between the number of accidents and the distance traveled during the year. This author also notes that regardless of whose fault the accident is, the frequency of accidents tends to increase when the annual distance traveled increases. Bordoff and Noel (2008) corroborates the non-proportional relationship advocated by Litman (2005) between the distance traveled and the number of accidents per year. These authors argue that the non-proportional relationship between distance and accident frequency is observed only when the analysis is performed with aggregate data. However, they think that when a driver is analyzed individually, this relationship is governed by proportionality. For example, this means that if the distance traveled by the driver is reduced by 10%, then the risk of accidents is likely to be reduced by 10% also.

Traditionally, insurers did not have precise data on vehicle usage of their policyholders, but this situation is beginning to change due to the application of telematics technology in insurance. To compensate the lack of usage information, insurers consider the estimation on the distance traveled per year declared by the insuree at the beginning of each contractual period. However, in most cases, this information is inaccurate. For example, according to Butler et al. (1988) an American insurer found that between 60% and 70% of the insured vehicles that were supposed to travel less than 12,000 km per year (according to the estimation provided by the driver) are actually travelling approximately 20,000 km. In such circumstances, and in the absence of a reliable verification mechanism available for insurers, the most advantageous strategy for the policyholders is to lie about their average yearly distance driven, if no odometer periodic check is performed. If only the estimated number of kilometers provided by the driver is taken into account, it is difficult for insurers to show that such an approach is inaccurate. However, there is a way to control that the declared distance traveled coincides with the reality. Indeed, when the insuree declares a claim, the company can verify the odometer and detect any inconsistency, which may involve a penalty in future premiums. In this sense, there is a deterrent against not disclosing correctly the distance traveled, but it seems that this effect may have not been sufficient to make the data completely reliable in the past.

It seems difficult to design a fair pricing method based on the use of the vehicle if the insurer cannot measure the distance traveled accurately. When the premium calculation accounts for the use of the vehicle, insurance is known as Pay-As-You-Drive Insurance (PAYD). Among other possibilities Vickrey (1968) proposed to integrate insurance in the price of gasoline or through the tires, so tire dealers are associated with an insurance company. He was a precursor who helped to launch the debate on PAYD insurance products.

1.2. Pay-As-You-Drive Pricing Systems

A possible price structure for PAYD insurance, according to Litman (2011), is to consider the distance traveled as a risk classification criterion in the calculation of insurance premiums. This approach is called Millage-Rate-Factor (MRF). Some companies offer a discount at the end of the year when the distance traveled by the insurer is below a threshold value. The main criticism of the MRF is that the structure depends on the estimation of the distance traveled declared by the insuree at the time of subscription. As already mentioned above, these estimates are generally lower than they should, so the actuarial accuracy of the calculated premium is affected. To improve the MRF, Litman (2011) suggests the introduction of a control procedure in order to adjust premiums during the year and/or at the end of the year.

Litman (2011) also examines the Per Mile Premium (PMP) price structure. Under this approach, the traditional unit of exposure (given by the duration of the contract, usually one year) is changed and a unit of distance is considered. Therefore, the insuree pays a price per unit of distance traveled. Other elements are also considered in premium calculation, so that drivers with a higher risk of accident pay more per unit of distance traveled than those with a lower risk. Litman (2005) observes different accident rates depending on the range of distance traveled, and argues that the price per kilometer should decrease as the distance traveled increases, so he justifies the application of decreasing PMP price structures.

With the advent of new technologies, it is now possible to install a GPS (Global Positioning System) in the vehicle or to use the data generated by mobile phones by using integrated GPS systems. These systems can provide the exact distance traveled by the insuree. This gives rise to a new form of PAYD insurance, offered by many insurance companies nowadays, which is already called by Litman (2011) the GPS-based pricing. This price structure should allow more accurate insurance premiums, as they vary depending directly on the distance traveled.

However, besides the distance, there are other telemetry data that can also be collected through this new technology which could be considered when calculating PAYD insurance prices: speed, time and place where driving is performed, the number of acceleration and hard braking (Paefgen et al. 2013, 2014). For example, to drive a lot at night usually means to pay a higher premium compared to those driving only during the day. Similarly, to drive in urban areas also means to pay a higher premium compared to those driving in rural areas. As telematics is still quite new, there are no historical data in insurance companies on GPS-based premiums that would validate the predictive models used for their calculation. These models are in some way subjective or based on generally accepted ideas. Although Jun et al. (2007) were able to show the potential of GPS-based data and Ayuso et al. (2017) review some ratemaking possibilities using real data, not many complete and advanced statistical studies on the topic are available, except for Verbelen et al. (2017) and Henckaerts et al. (2017).

It is important to remark that both the PMP and GPS-based pricing system are options that can be chosen by clients of insurance companies that offer them. Although several studies have already addressed the advantages of offering PAYD insurance (Buxbaum 2006; Bordoff and Noel 2008), many policyholders are reluctant to this type of products. According to Litman (2011), approximately between 25% and 50% of the insurees would be willing to choose a PMP premium option if the insurance company would offer it, with a percentage that is expected to increase over the years. In contrast, a GPS-based premium only draws from 2% to 5% of the policies in force. This low percentage is explained largely by the feeling of invasion of privacy that makes the GPS-based systems to be viewed with suspicion. In this sense, Iqbal and Lim (2006) propose a privacy preserving GPS-based PAYD insurance scheme that eliminates the intrusive side thereof. However, the low interest in GPS-based premiums leads to problems when performing the statistical analysis, as it is not clear that policyholders who accept this system have the same behavior as those who do not choose it.

1.3. Potential New Research Field

The installation of GPS devices in automobiles creates an enormous potential for research. This advanced technology helps to understand the risk of car accidents, since specific data on a multitude of risk factors are collected. In this paper, the impact of the distance traveled and exposure time is studied. To do this, we use Generalized Additive Models, often named by their acronym GAM. There are several papers in the actuarial literature focused on risk exposure. Boucher et al. (2013) showed that the inclusion of the exact distance traveled in a Poisson Generalized Linear Models (GLM) model is able to explain the relationship between annual distance and frequency of claims. Meanwhile, using a Weibull regression model, Ayuso et al. (2014) modeled the time and the number of kilometers traveled before a first communication of a claim by young policyholders who have purchased PAYD insurance. These authors observed significant differences by gender, so that men have a riskier driving pattern compared to women. In addition, they also observed differences between novice drivers and those more experienced when comparing their particular driving patterns and the effect on their risk of accident, concluding that young drivers are a heterogeneous risk group. The same authors (Ayuso et al. 2016a) conclude that the observed differences in gender are largely attributable to the intensity of vehicle usage: while gender has a significant effect in explaining the time elapsed until the first accident, it has no longer effect when the average daily distance traveled is introduced into the model. Therefore, the authors conclude that discrimination based on gender is not necessary if telematics provide enough information on driving patterns (see Ayuso et al. 2016b). This same conclusion is also obtained by Verbelen et al. (2017) using a Belgian database. Baecke and Bocca (2017) show that including standard telematics variables significantly improves the risk assessment of customers. As a result, insurers are better able to tailor their products to the customers' risk profile. Tselentis et al. (2017) review the existing literature on usage-based insurance (UBI) schemes and research gaps are identified, showing that there is a multiplicity and diversity of several research studies accumulated in modern literature examining the correlation between PAYD (based on driver's travel behavior and exposure) and pay how-you drive (PHYD, based on driving behavior) schemes and crash risk in order to determine accident risk.

The aim of this paper is to contribute to the actuarial literature by analyzing the simultaneous effect of distance and exposure time on the risk of accident, and to design a price structure for PAYD insurance based on this analysis and additional factors. The paper is organized as follows. In Section 2, a summary of the GAM approach is summarized and the data are presented. In Section 3, GAM is applied to the insurance database to analyze the simultaneous effect of duration and distance in the estimation of the number of insurance claims and, based on these results we propose a rating system for PAYD contracts. The last section concludes the work.

2. Data and Models

Generalized Additive Models were first proposed by Hastie and Tibshirani (1986) and they are an extension of GLM. The difference is at the level of the linear predictor. For the GLM, and for an individual i, i = 1, . . . , n, the mean μ_i of a response variable Y_i which has a probability distribution that belongs to the linear exponential family is related to the linear predictor $X_i'\beta$ through the link function $g(\cdot)$, so that $g(\mu_i) = X_i'\beta$. In that way, the expectation of the response variable Y_i is explained as a function of the explanatory variables $x_j, j = 1, \ldots, K$, where K is the total number of risk factors considered in the model.

On the other hand, GAM allows a higher flexibility compared to the GLM. The predictor remains in a linear form, but the variable of interest is explained by using functions of the explanatory variables. These functions are a way to deal with the nonlinear relationship that may exist between the dependent and independent variables. In that way, we have a less restrictive relationship between the expectation of the response variable Y_i and the explanatory variables. An example of GAM could be as follows:

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i})$$

where f_j are functions of the variables x_j, $j = 1$, 2 and 3 (three explanatory variables, for example). The right side of the equation is called nonparametric component. Of course, the functions that are considered in the model must be estimated. It is important to remark that it is possible to include terms in a GAM that are fully parametric. This also includes the intercept term of the model playing the same role as the constant in the equation of a classical linear regression.

GAMs are suitable to study the behavior of the factors that influence the expected value of a response variable. They are especially useful when it is suspected that the relationship is not linear. On the other hand, since this is a nonparametric approach, GAMs let the data help choosing the functional forms (what is known as letting the data speak) and therefore allow going beyond the typical parametric relationship of a GLM. However, GAMs are more complex and more difficult to interpret than GLMs. The classical references on these models are obviously Hastie and Tibshirani (1990) and Hastie and Tibshirani (1990), where most of the explanations on the technical tools used in this paper can be found. Of course, it is also possible to use existing R packages, or SAS procedures for the application of GAM to large databases. In the following sections, the data set used in our empirical application is presented and explained in detail.

2.1. Description of the Database

The database used was supplied by an insurer operating in the Spanish market. It consists of 71,489 PAYD automobile insurance policies and all of them refer to the year 2011. This product was mainly sold to young drivers for the purpose of installing a GPS that would contact the company and indicate the location of the car in case of an emergency. This data set has been extensively studied in Ayuso et al. (2014) among others.

In our database, each accident is classified according to the nature of the damage and the liability in the production of the damage. Thus, there are several categories of claims. We consider here claims where the insuree is at fault and there are property damages (*nb*1) and claims where the insuree is not at fault and there are property damages (*nb*2).

For each contract, there is a significant amount of information available, obviously including the distance traveled (*km*), given that a GPS device was installed in the car. This variable is absolutely essential for this study. We also consider the observation duration (*d*), or exposure to risk, as the fraction of the year when the policy was in force. The database also contains information about the profile of people who purchased a PAYD coverage offered by the Spanish insurance company. The most relevant variables are: insuree age (*age*), age of the vehicle (*ageveh*), insuree sex (*gender*) and parking type (*stn*).

Table 1 shows descriptive statistics for the quantitative variables and Table 2 presents some descriptive statistics the gender composition and the location where the car is parked overnight. By looking at Tables 1 and 2, we see that policyholders are mostly young adults. The average age is 26 years and the maximum age is 37. Regarding the age of the vehicle, the average is almost 8 years, so this means that the portfolio has a relatively mature fleet. The fact that the portfolio is composed predominantly of young drivers may explain why the vehicles have a higher age compared to the driving experience to the insuree (youngsters generally have fewer financial resources than older people and may have access to secondhand vehicles). The data also show a very similar percentage of men and women. Finally, the vast majority of policyholders have access to private parking.

Table 1. Descriptive statistics for variables *km*, *d*, *age*, *ageveh*, *nb*1 and *nb*2.

Variable	Average	Standard Deviation	Minimum	Maximum	k − Percentile		
					$k = 25$	$k = 50$	$k = 75$
km	4889.98	3978.50	0.005	50,035.56	1836.25	4018.32	6949.45
d	0.706	0.305	0.003	1	0.468	0.795	1
age	25.97	3.17	18.00	37.00	23.00	26.00	29.00
ageveh	7.91	4.55	0.00	34.00	4.00	7.00	11.00
*nb*1	0.075	0.277	0	4	0	0	0
*nb*2	0.076	0.281	0	4	0	0	0

Table 2. Relative frequency of policies by gender and type of parking.

Statistic	Variable *Gender*		Variable *stn*	
	Men	Women	Outside	Private Garage
Percentage (%)	53.7%	46.3%	23.3%	76.7%

Our data belong to a group of drivers that may not exactly be representative of the general population of drivers. Indeed, they are young drivers. Authors studying the driving population in Spain report the average age to be older than the average age of our sample, which is 25–97. Official figures on the age of citizens who have a driving license in Spain indicate that the average is 48.63 years. Alcañiz et al. (2014) analyze a sample of random drivers who were stopped at sobriety checkpoints and they report similar results for Catalonia (Spain). In addition, there is a possibility that opting for PAYD and installing a telematics device may influence driving patterns. This phenomenon has been noted by some authors who claim that once the driver knows that there is a telematics recording, then the driver actually pays more attention to careful driving. The literature reports evidence of drivers modifying their driving patterns. Bolderdijk et al. (2011) observed a significant impact on the reduction in speed violations among young drivers with a PAYD policy. Thus, the direct comparability of our results to the general population of driver and extrapolation of these results to all the drivers may not be valid.

Table 3 illustrates the distribution of the number of claims reported in 2011 for property damage accidents (variables *nb*1 and *nb*2).

Table 3. Absolute and relative frequencies of claims where there are property damages.

Number of Accidents	Claim at Fault *nb*1		Claim Not at Fault *nb*2	
	Frequency of insurees	Percentage (%)	Frequency of insurees	Percentage (%)
0	66,372	92.842	66,371	92.841
1	4882	6.829	4824	6.748
2	225	0.315	283	0.396
3	9	0.013	10	0.014
4	1	0.001	1	0.001
Total	71,489	100	71,489	100

It is also interesting to see graphically how the number of claims depends on the two main explanatory variables of interest: the distance traveled and exposure duration. Figure 1 shows the distribution of the kilometers traveled among all PAYD insurance policies in our sample in force in 2011. Intervals of 50 km are used to represent the histogram. In Figure 1 we observe that the frequencies of policies decrease as the distance traveled increases. Surprisingly, we observe that the first bands correspond to a very low distance traveled. Remember that all contracts are considered here, especially those covering the whole year 2011, but also those that are in force only a few days. In Table 1 we see that 75% of drivers have traveled less than 6950 km during the exposure period. The average distance is 4890 km. In comparison, considering only insurees covered during all the year, the average distance is 7160 km.

Regarding the exposure time, which we find in variable duration (d), we also observe some variability, as shown in Figure 2. The histogram is built by using intervals of 0.05 years, equivalent to about 18 days. We observe that more than 25,000 insurees are covered for the entire year (vertical bar on the right), which corresponds to 35% in 2011. The average duration of the portfolio is 0.706 years, i.e., about 258 days.

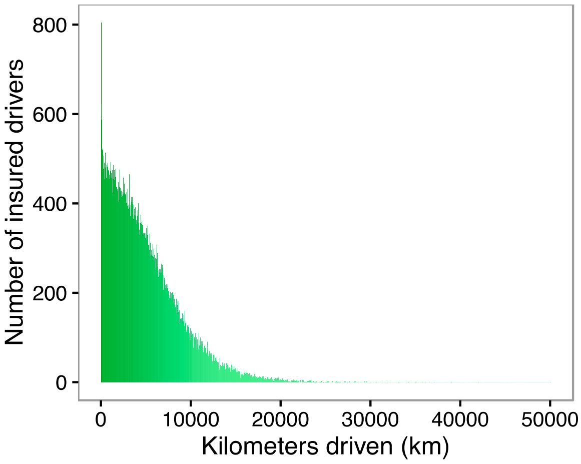


Figure 1. Distribution for the distance traveled for PAYD contracts in force in 2011.

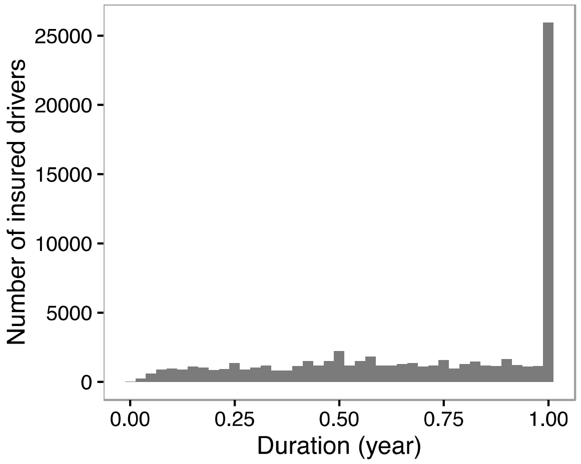


Figure 2. Distribution of the duration of PAYD contracts in 2011.

Figures 3 and 4 show the average number of claims for different groups of insurees. To calculate these averages, policyholders have been grouped by considering intervals of 500 km in the distance they travel. In the case of Figure 4, insurees are grouped by considering exposure time intervals of 0.1 years, equivalent to 37 days approximately. In Figure 3 we observe that while there is a positive relationship between the number of claims and distance, it seems that this relationship is not linear. The effect of distance on the propensity to report a claim appears to decrease when the distance increases. However, we must be careful with this interpretation, because it is only a descriptive analysis. As shown in Figure 3, there is a large concentration of policyholders who have traveled less than 10,000 km during the exposure period. Therefore, there are differences in the density of observations that can be found at each point of the graph shown in Figure 3. Regarding Figure 4, in this case we seem to perceive a linear relationship between the number of claims and the duration of exposure. Again, note that each point on the graph does not always represent the same number of observations.

A possible reason to observe increasing risk but less impact with larger distances or durations is that drivers may be too cautious at the beginning of the contract, as they may feel more surveilled with the telematics mechanism. This effect may vanish over time and gaining more confidence may possibly induce more risky driving. There is an additional effect for novel drivers like many of those in this sample. As they learn to drive they acquire more experience and they learn, therefore more exposure to risk implies that there is more risk of an accident due to more distance driven, but less risk due to the driver being more experienced. These two factors compensate.

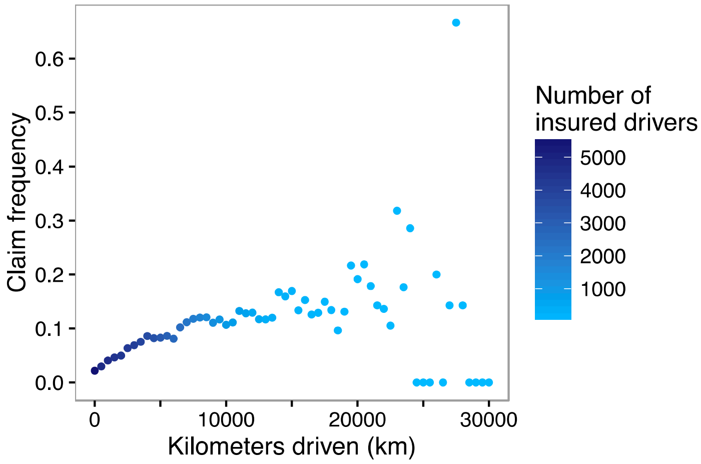


Figure 3. Distribution of the observed number of claims vs. distance traveled.

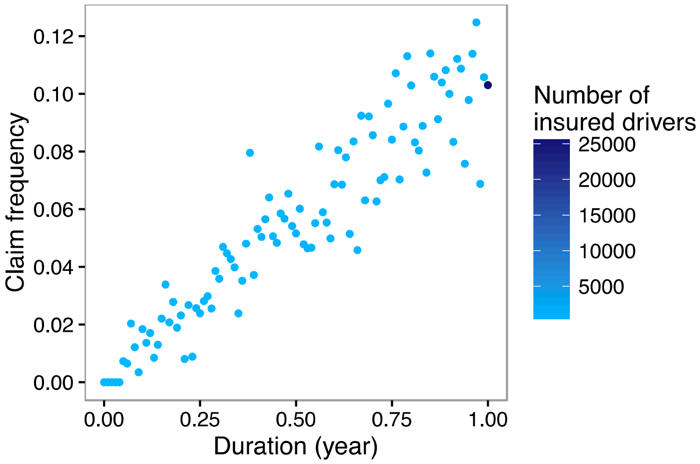


Figure 4. Distribution of the observed number of claims vs. exposure duration.

2.2. *Modeling with GAM*

The number of claims is analyzed by using different GAMs. The GAM approach allows a better understanding of the impact of the distance traveled and the exposure duration on the number of claims. Our data are separated into two different databases: training and validation data. The first database is used to carry out the estimation, while the validation database is used in to assess the predictive performance of the model in a different sample than the one used in the estimation. The separation of the data in these two samples is done totally at random: 5000 observations are assigned to the validation sample, while the rest are used to build the different models.

2.2.1. Model with Independent Cubic Splines

First, the number of claims is modeled by using a GAM where independent cubic splines are adjusted for the number of kilometers (km) and the duration of the insurance contract (*d*). Let us assume that the number of claims (*nb2*) reported by the insuree *i*, follows a Poisson distribution with expected value equal to μ_i. A logarithmic function is used to link the expectation of the response variable with the linear predictor. Therefore, the GAM approach described herein can be formulated by the following equation

$$\log(\mu_i) = \beta_0 + f_1(km_i) + f_2(d_i) \tag{1}$$

where β_0 is the independent term in the model. Functions f_1 and f_2 are cubic splines, which are defined as univariate smoothing functions, with the following linear form

$$f(x) = \sum_{k=1}^{q} b_k(x)\beta_k$$

where β_k is the vector of parameters to be estimated and $b_k(x)$, $k = 1, \ldots, q$ are functions created by a cubic spline basis with dimension given by q. The details of the expression of the functions , $k = 1, \ldots, q$ can be found in Wood (2006, Section 4.1.2). By doing this, the techniques used to estimate a GLM could also be used to estimate a GAM.

For the moment, there are no regressors capturing possible differences between the characteristics of the insured risk, such as age, sex of the driver, etc. That is, we are not going to use these variables in this first model. This allows us to make a more interesting graphical analysis as all insurees are observed simultaneously, because we believe that it is important to perform an initial analysis of the general profile. However, as we will see in the next section, we are going to incorporate these regressors for pricing, and then we will see that, in general, the results of this section are still valid when the heterogeneity of the insurees is considered.

For the parametrization defined by Equation (1), now this model will be simply called model 1, we select 7 and 3 nodes for f_1 and f_2, respectively. The choice of the number of nodes is done manually and depends on the desired degree of flexibility. We note that the number of knots is rather low, so we expect smoother shapes, but usually more knots are selected using quantiles of the observed values. However, this is an important stage in the modeling process, since very few nodes produce an adjustment that cannot capture important trends in the data, while too many nodes can lead to an over-fitting. It seems that there is no consensus among the scientific community regarding the determination of the optimal number of nodes. Therefore, the choice of the number of nodes is an important part of the modeling process and may depend on the proficiency in the application. Table 4 shows the results for model 1.

Table 4. Results for model 1.

Parametric Part	**Estimate**	***t* Value**	***p*-Value**
β_0	−2.735	−160.17	<0.001
Non-Parametric Part	**EDF** [1]	***F* Value**	***p*-Value**
$f_1(km)$	4.30	55.61	<0.001
$f_2(d)$	1.95	81.53	<0.001
GCV [2]	0.38412		

[1] Effective degrees of freedom; [2] Generalized cross validation.

The concept of degrees of freedom, which is generally used in statistics, is adapted in the context of GAM and is known as the effective degrees of freedom (EDF). If the smoothing parameters are zero, then the number of degrees of freedom for a smoothing function is simply the number of parameters to estimate minus one (due to the constraint that the function should add up to 1 for any given observation). If the smoothing parameters are not zero, the number of degrees of freedom is necessarily reduced and then the concept of effective degrees of freedom is considered in order to quantify the flexibility of a smoothing function or a general model. The F value shown in Table 4 is the result of a statistical test similar to the Wald test to verify the significance of nonparametric terms. Finally, the generalized cross validation (GCV) is a method associated to the minimization of a score in order to find the smoothing parameters. For more details we refer to Wood (2006), Sections 3.2.3 and 4.8.5.

The low p-values in Table 4 show that both the parametric and the nonparametric effects in model 1 are significant. The GCV score is equal to 0.38412, but does not let us conclude anything at the moment, since it must be compared to another model. The GCV score is a statistical measure that only makes sense when several models are compared.

Figure 5 visualizes the adjusted smoothing functions obtained after estimating model 1. The black curves in both illustrations correspond to the predicted values for each function. The grey areas correspond to the 95% confidence intervals for the predictions. At the bottom of the graph, the density of observations that were used in the model is shown. Note also, that the effects are presented on the scale of the linear predictor instead of on the scale of the response level.

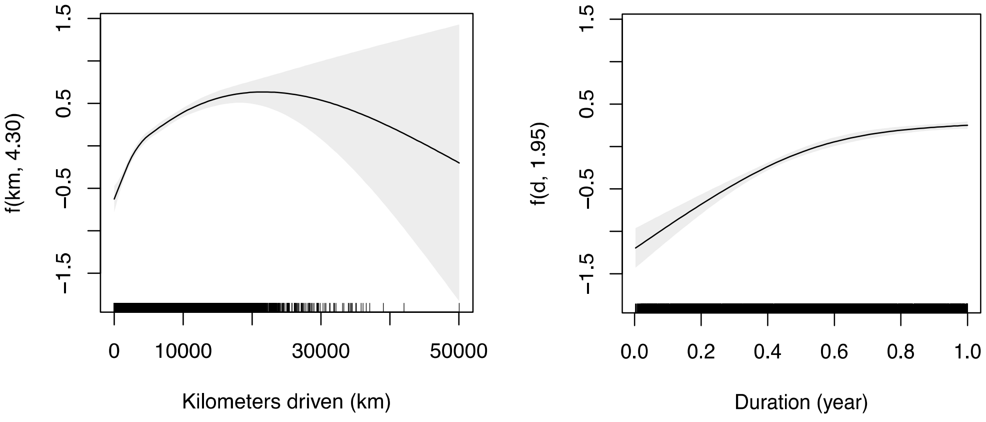


Figure 5. Model 1, $\hat{f}_1(km)$ on the left and $\hat{f}_2(d)$ on the right for claims not at fault with property damage ($nb2$).

In Figure 5 we see that $f_1(km)$ increases very fast for the first 10,000 km. Subsequently, the increase continues but not so sharply and stabilizes at 20,000 km. Finally, we observe a decreasing pattern but the confidence in the predicted values is very low because there are only a few contracts with such a large number of kilometers.

In addition, Figure 5 shows the lack of proportionality between the log-accident rates and the number of kilometers. The fact that the slope of the curve gradually decreases as the number of kilometers increase, lets us confirm that there are factors that mitigate the risk of accident of more experienced drivers who use more the car, compared to those who drive only occasionally. An intuitive interpretation of this result is that people who drive regularly tend to develop better reflexes and driving skills. Moreover, a large proportion of the distance traveled of the big drivers is done on the road where accidents are less frequent than in urban areas. It would be interesting to carry out further analysis and research that could confirm why we observe this result. The empirical study of Verbelen et al. (2017) shows that the proportionality assumption for both exposure duration and distance traveled is too restrictive and that the best model is obtained by simultaneously modeling the effect of the exposure duration and distance traveled using additive splines. By also taking the composition of the distance traveled on different road types, time slots and week vs. weekend into account, these authors obtain a more linear effect for the distance traveled compared to Figures 5 and 6. We also note that there are several differences between the approach by these authors and ours. While they have a more general setting with respect to the inclusion of more risk factors into the model, such as the type of road, day of the week and time slot, they somehow evolve from pure distance driven towards to driving habits, i.e., where and when is the vehicle driven. We find this point essential in analyzing the differences between these approaches and moreover, this has an impact on the implementation of telematics pricing. This is discussed in the last section.

Regarding the second graph in Figure 5, we observe that the exposure time basically has a linear effect on the liner predictor during the first six months. Thereafter, exposure time still has a positive effect on the risk of accident but it is less pronounced, and finally it has almost no impact after 10 months of observation. These findings may contradict what is applied nowadays in the vast majority of insurance companies, i.e., assuming that the number of accidents is directly proportional to the exposure duration. Therefore, we conclude that it is wrong to say that an insuree covered for a year is twice risky compared to one a covered only for 6 months, if the rest of risk factors are the same. It seems that this statement is not correct. It is interesting to see that this result confirms the conclusions obtained by Boucher and Denuit (2007), at least for this sample and given that the drivers know that they are being monitored.

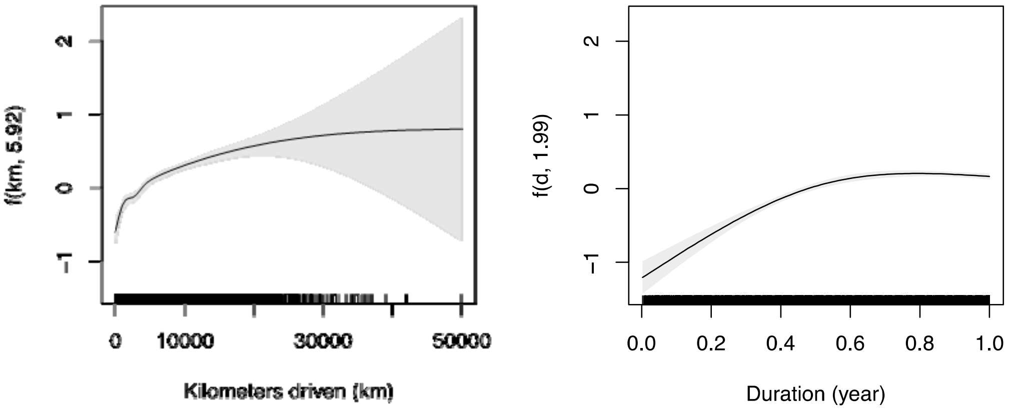


Figure 6. Model 1, $\hat{f}_1(km)$ on the left and $\hat{f}_2(d)$ on the right for claims at fault with property damage $(nb1)$.

Finally, for comparison, Figure 6 shows also the adjusted smoothing functions for the claims at fault. Therefore, the conclusions regarding the impact of the duration and distance traveled also seem to be valid for accidents at fault with property damage $(nb1)$. Once a certain number of traveled kilometers is reached, i.e., from approximately 20,000 km on, we observe a certain difference between the trend of $nb1$ and $nb2$. However, the corresponding confidence intervals do not seem to confirm a statistically significant difference.

2.2.2. Tensor Product Smoothing Model

For model 1, the distance traveled and exposure duration were introduced as explanatory variables by using cubic smoothing splines. Firstly, they were parameterized completely independently, and later they were estimated.

Now, we are going to show how the addition of an interaction term between the distance traveled and the exposure time changes the results obtained from model 1. We will use the same modeling procedure, i.e., GAM with cubic splines. The difference is that instead of using two separate cubic splines to include the distance traveled and exposure time in the model, we will use a smoothing tensor product base. Broadly speaking, the idea is to introduce a smoothing function for the interaction. It is then possible to evaluate the improvement of the GCV score.

Similarly to model 1, only the distance traveled and exposure duration are used, as these are the variables explaining the number of claims $(nb2)$ of each insuree. We will denote the number of claims by $\mathrm{N_i}$ and assume that it follows a Poisson distribution. A multiplicative link function is used to relate the expectation with the linear predictor. As mentioned above, the tension product smoothing base is used to define a two-variable function, which is introduced in the GAM. Specifically, the model is formulated as:

$$\log(\mu_i) = \beta_0 + f(km_i, d_i) \tag{2}$$

where β_0 corresponds to the independent term of the model and $f(km_i, d_i)$ is a smoothing function which depends on the distance (*km*) and the duration (*d*). The tensor product smoothing base is implemented both in R and SAS software. However, it is also possible to apply the technique by using the procedure described by Wood (2006, Section 4.1.8).

Similarly to the model with independent cubic splines, the function is also parameterized by using seven nodes for the distribution of the distance traveled and 3 nodes for the exposure duration. The model defined by Equation (2) will be called model 2. Table 5 shows the estimation results.

Table 5. Results for model 2.

Parametric Part	**Estimate**	***t* Value**	***p*-Value**
β_0	−2.740	−157.3	<0.001
Non-parametric Part	**EDF** [1]	***F* Value**	***p*-Value**
$f(km, d)$	13.69	62.56	0.001
GCV [2]	0.38403		

[1] Effective degrees of freedom; [2] Generalized cross validation.

The low *p*-values found in Table 5 show that the nonparametric part of the model 2 is important to explain the number of claims of the policyholders. The value of the GCV score, which is equal to 0.38403, suggests a slight improvement of model 2 with respect to model 1, which provided a GCV score equal to 0.38412. In other words, the added flexibility provided by the tensor product smoothing base results in a slight improvement in the adjustment of the frequency of claims. There are 13.69 effective degrees of freedom in model 2, against 6.25 (4.30 + 1.95) for model 1.

Since function $f(km, d)$ in model 2 is not expressed in terms of *km* or d separately, it is not possible to analyze the impact of the distance traveled and exposure time independently as we did in model 1. However, Figure 9 shows the surface derived from the predictions produced by the estimation of model 2 for every possible pair (km, d). It is interesting to see that the distance traveled has a significant impact for the first 10,000 or 15,000 km in the estimation of the function of model 2. Subsequently, the impact gradually fades. Regarding the duration, we see that the impact seems to be fairly constant on the predicted surface.

Obviously, when the distance traveled is large, since there are few observations at this level, it is expected that the predictions become much more volatile. In the next section, we compare models 1 and 2 and discuss the differences between them.

2.3. Comparative Analysis

Since the same basis of cubic splines was used, it is possible to directly compare models 1 and 2. In this situation, Equation (1) is strictly included (strictly nested) in Equation (2). In other words, the function $f(km, d)$ of model 2 could theoretically contain any possible value of the sum defined by $f_1(km)$ and $f_2(d)$ in model 1. The performance of the two models was very similar according to the GCV score. We also compare the surfaces generated from the two models.

Figures 7–10 present prediction surfaces and Figures 11 and 12 also present these surfaces together with standard deviations. We observe that Figures 7–10 are very similar. This was probably expected because of the similar GCV scores of the two estimated models. However, one can see that the surface of model 1 is more smoothed than the one corresponding to model 2, due to the additional added tensor product I the latter model. The number of effective degrees of freedom for both models supports this: 6.25 for model 1 versus 13.69 for model 2.

As no other explanatory variables than the annual distance traveled and duration of exposure are considered in model 2, the surface of the predictions of the frequency of claims, in fact, turns out to be exactly the estimated function $f(km,d)$ of Equation (2), but shifted $\exp(-2.7402) = 0.0646$ units up.

In Figure 8 we see that flexibility becomes apparent in the predicted claim frequencies for insurees travelling between 0 and 10,000 km per year. The first thousands of kilometers seem to have a higher impact on the prediction of the expected number of claims compared to what we see in Figure 8. Then, we see that around a driving distance of 5000 km per year, there is a change in the surface, indicating a change in the trend of the expected number of claims. It would be interesting to estimate the same model with data from the same insurance company but from a different calendar year. This may contribute to explain the reason of this change. Probably there is a logical reason behind this phenomenon, but also a slight over parameterization or over-adjustment problem could be taking place in this model.

Regarding the difference in the results provided by the estimated models, Figures 11 and 12 present the red surface results from adding one standard deviation to the predicted values, while the green surface results from subtracting one standard deviation from the predicted values. Again, the difference between the results obtained by the two models is due to the greater flexibility of model 2 in which we have added a dependence relationship between km and d. In fact, 9 parameters are estimated for model 1, while 21 parameters are needed in the case of model 2.

Different cross-sectional perspectives are considered in order to easy visualize the three-dimensional graphs. For each model, one of the two variables (distance traveled or exposure duration) is fixed at different levels, while the other variable is left free. Therefore, it is possible to analyze the number of claims from different perspectives. This is available in the supplementary material.

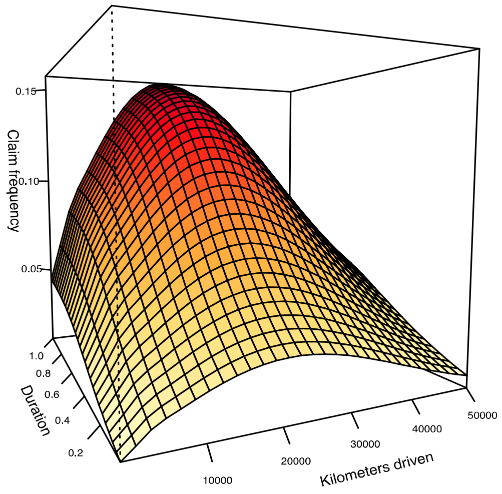


Figure 7. Model 1—Predictions surface (angle 1).

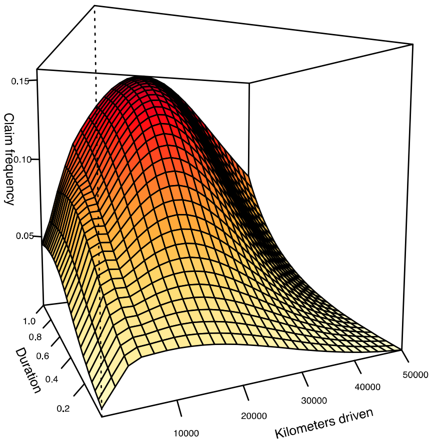


Figure 8. Model 2—Predictions surface (angle 1).

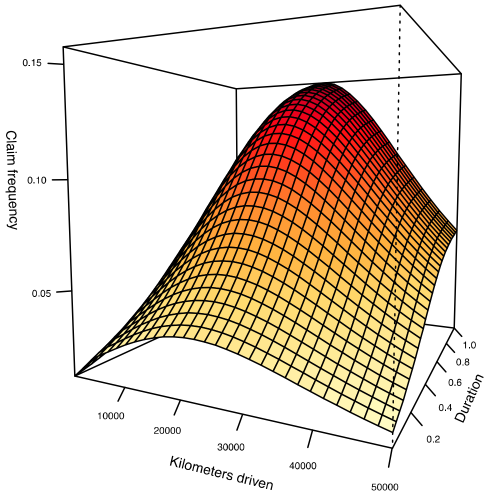


Figure 9. Model 1—Predictions surface (angle 2).

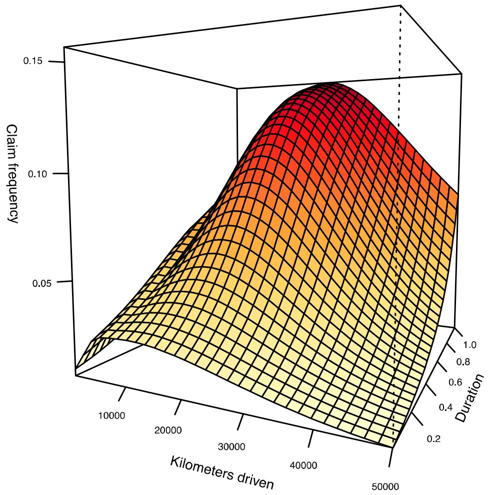


Figure 10. Model 2—Predictions surface (angle 2).

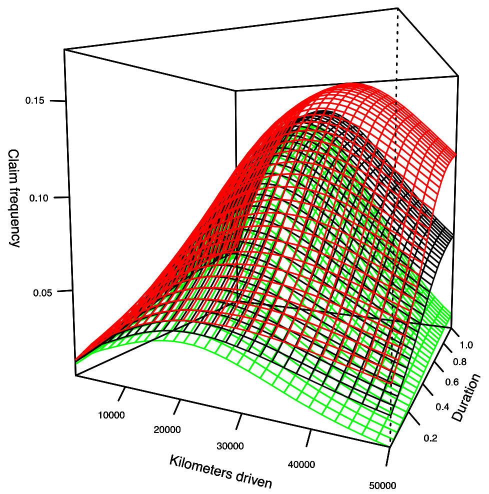


Figure 11. Model 1—Predictions ± 1 std. dev.

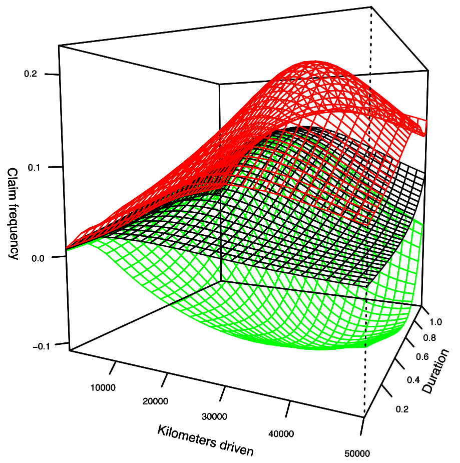


Figure 12. Model 2—Predictions ± 1 std. dev.

3. Pricing Application

The purpose of this section is to compare the conventional methodology used in practice by insurance companies to model the frequency of claims reported by policyholders and the models presented in the previous section. The starting point considered by actuaries in modeling the frequency of automobile insurance claims is a GLM that assumes that the number of reported claims follows a Poisson distribution.

As in models 1 and 2, we will only consider the distance traveled (*km*) and exposure duration (*d*) to explain the number of claims (*nb*2) reported to the company by the insuree. In order to generate a pricing system based on the results obtained by the GAM, the distance traveled will be categorized into five classes, including a reference category. The choice of the classes is purely illustrative. Therefore, four regressors related to the distance traveled will be included in the model. Exposure duration is considered to be an offset variable. The goal here is to see if it is possible to easily replicate the previous results obtained with the GAM, but using a Poisson GLM model, which is nowadays the standard model used in insurance companies. Note that Henckaerts et al. (2017) work on a similar question, i.e., they start from GAMs with smooth effects and transform these models into GLMs with categorical effects that satisfy the practical needs of an insurance company.

Let us assume that the number of reported claims follows a Poisson distribution. A multiplicative link is used to specify the relationship between the linear predictor and the expectation of the response variable. Specifically, the model is represented by the following equation:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \log(d_i) \quad (3)$$

where β_0 is the constant term in the model. This model will be called model 3. The binary variables used in the model are described in Table 6. The reference category corresponds to a distance traveled ranging from 1000 and 5000 km. Table 7 shows the results of the model. In particular, we observe that all parameters in the model are significant. In addition, the estimate of β_1 is the only one which is negative apart from the constant term, which means that on average, only insurees traveling less than 1000 km have a lower premium than the one corresponding to the reference group.

Figure 13 shows the surface of the predictions generated by model 3. Note that this parametric model hardly approximates to the surface illustrated in the previous figures.

Table 6. Binary variables used for the segmentation of the distance traveled.

Variable	Description
x_1	It takes value 1 if km $\leq$ 1000
x_2	It takes value 1 if 5000 < km $\leq$ 10,000
x_3	It takes value 1 if 10,000 << km $\leq$ 15,000
x_4	It takes value 1 if 15,000 < km $\leq$ 20,000
x_5	It takes value 1 if km > 20,000

Table 7. Results of the estimation of model 3.

	Estimate	Standard Error	*t* Value	*p*-Value
β_0	−2.3568	0.0242	−97.33	<0.001
β_1	−0.2201	0.0727	−3.03	<0.001
β_2	0.1989	0.0338	5.89	<0.001
β_3	0.3426	0.0463	7.40	<0.001
β_4	0.4734	0.0837	5.66	<0.001

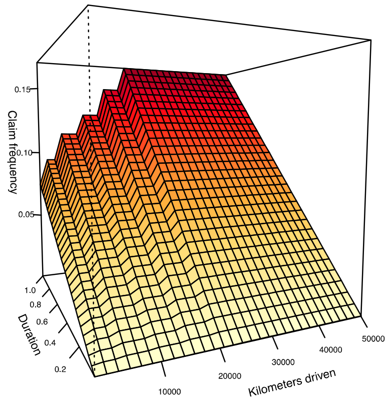


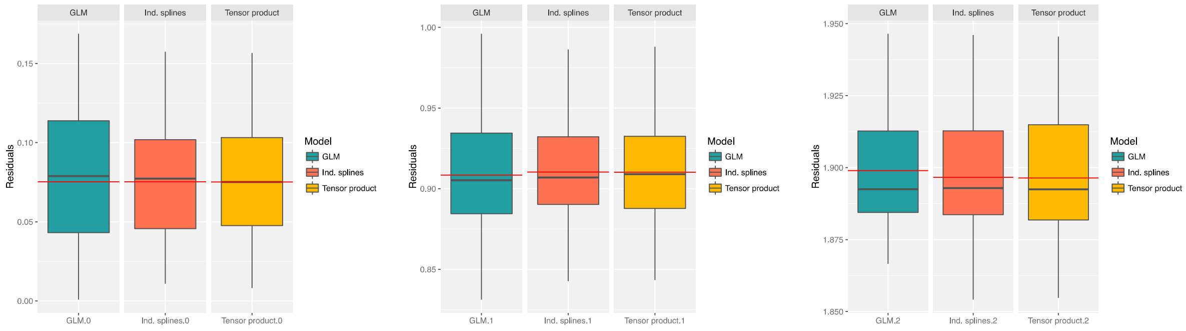
Figure 13. Model 3 (GLM). Predictions surface for claims not at fault with property damage.

3.1. Comparison of Prices Based a Conventional GLM versus GAM

While the GAM approach seems to offer a better fit to our PAYD insurance data compared to GLM, it is possible to evaluate the actual performance of a model (and thus avoid the risk of over adjustment) by using a validation dataset. For this reason, 5000 observations have not yet been included in any of the models. Each of these observations of the validation dataset has been used to produce a prediction in each of the three fitted models. Next, the corresponding residual is calculated, which is the absolute difference between the expected and observed number of claims.

Figure 14 shows boxplots of the distribution of residuals for the three models. In each figure we see from left to right, respectively, the results for models 3 (GLM Poisson), model 1 (GAM with independent cubic splines) and model 2 (GAM with a tensor product base), respectively.

The residuals illustrate how the three models have a similar predictive power. Despite the fact that the GLM is by definition less flexible compared to GAM, the GLM has a good predictive power. It may even be the best model when the average residuals for policyholders who have had one accident (see horizontal red lines in Figure 14) are analyzed. Regarding insurees who do not have any accident, i.e., the vast majority of the 5000 observations, the average residuals are practically the same for all models. However, we see that the two GAM models produce predictions that are less volatile than those of GLM.



0 claims (4.644 observations) 1 claim (332 observations) 2 claims (24 observations)

Figure 14. Residuals for insurees with 0, 1 and 2 claims (in each part, left to right GLM, GAM with independent cubic splines and GAM with a tensor product base).

3.2. Risk Characteristics Segmentation

So far, no other explanatory variables, apart from the distance traveled and exposure duration, were considered. It may be interesting to compare the three models when other variables are added to the linear predictor. Namely, we consider the age of the driver, age of the vehicle, sex and type of parking. All these variables were included in a traditional GLM, but we only found that one of them was important to explain the number of claims: the age of the driver.

Models 1, 2, and 3 have been modified to include the variable age (age), by using different categories, as shown in Table 8. Note a smooth age affect could be considered instead.

Table 8. Binary variables used for the segmentation of the age.

Variable	Description
x_6	It takes the value 1 if age ≤ 25
x_7	It takes the value 1 if $25 <$ age ≤ 30

It is assumed that the reference category is an insuree older than 30 years. In summary, the following equations are used to define the three new adjusted models that will be compared below:

$$\log(\mu_i) = \beta_0 + \beta_6 x_{6i} + \beta_7 x_{7i} + f_1(km_i) + f_2(d_i)$$
$$\log(\mu_i) = \beta_0 + \beta_6 x_{6i} + \beta_7 x_{7i} + f(km_i, d_i)$$
$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \log(d_i)$$

Two regressors were added (x_6 and x_7) for the variable age in each of the equations. Regarding binary variables x_j, $j = 1, 2, \ldots, 5$, we consider the same definitions as before. Table 9 shows the parameter estimates and their standard errors for each model.

Table 9. Results of the estimation of the parameters associated with age, standard errors in parenthesis (all p-values < 0.001) for claims not at fault with property damage.

	Model 1 (GAM)	**Model 2 (Extended GAM)**	**Model 3 (GLM)**
β_6	0.3917 (0.0597)	0.3906 (0.0598)	0.4260 (0.0592)
β_7	0.1655 (0.0602)	0.1647 (0.0603)	0.1942 (0.0597)
GCV [1]	0.38295	0.38285	0.38499

[1] Generalized cross validation.

Again, the numerical results obtained for both GAMs are very similar. All models have an estimated value of β_6 and β_7 greater than 0, which intuitively makes sense since they refer to the classes where younger insurees are. This means that drivers of 25 years old or less have a higher risk than those aged between 26 and 30, both inclusive. As the estimated parameter values are positive, it follows that insurees over 30 are those with the lowest risk of accident. It is important to note that these arguments are made only on the basis of age, assuming that the distance traveled and exposure duration is fixed.

The values obtained for GCV are 0.38285, 0.38277 and 0.38499 for models 1, 2 and 3, respectively and they show that extended model 2 has the best performance. Interestingly, the estimated values of β_6 and β_7 are greater for the GLM compared to GAM. Therefore, although it seems that according to GCV the models are very similar, there are differences between the GAM and GLM estimates for these coefficients. There are therefore differences in the treatment of the distance traveled and exposure duration of the two approaches.

Table 10 shows, for different age groups, the premiums generated by the three models for five (km; d) profiles. We observe that the premiums produced by GAM models (model 1 and model 2) are similar, while the GLM (model 3) results are somewhat different.

Table 10. Estimated premiums for the three models by age groups.

	Premium for Insurees Aged 25 Years or Less		
(*km*;*d*)	**Model 1 (GAM)**	**Model 2 (Extended GAM)**	**Model 3 (GLM)**
(3500;0.35)	0.0519	0.0571	0.0379
(4500;0.50)	0.0748	0.0778	0.0542
(9000;0.65)	0.1147	0.1058	0.0860
(15,500;0.90)	0.1654	0.1638	0.1594
(19,000;1.00)	0.1803	0.1791	0.1771
	Premium for Insurees Aged from 26 to 30 Years		
(*km*;*d*)	**Model 1 (GAM)**	**Model 2 (Extended GAM)**	**Model 3 (GLM)**
(3500; 0.35)	0.0414	0.0456	0.0301
(4500; 0.50)	0.0596	0.0621	0.0430
(9000; 0.65)	0.0915	0.0844	0.0682
(15,500; 0.90)	0.1319	0.1306	0.1264
(19,000; 1.00)	0.1438	0.1429	0.1405
	Premium for Insurees Older than 30 Years (Baseline Category)		
(*km*;*d*)	**Model 1 (GAM)**	**Model 2 (Extended GAM)**	**Model 3 (GLM)**
(3500; 0.35)	0.0351	0.0386	0.0248
(4500; 0.50)	0.0505	0.0527	0.0354
(9000; 0.65)	0.0775	0.0716	0.0562
(15,500; 0.90)	0.1118	0.1108	0.1041
(19,000; 1.00)	0.1219	0.1212	0.1157

3.3. A Simple Price Structure for PAYD

A great advantage of the GLM with multiplicative link is that it is easy of use in practice. Furthermore, the interpretation of the model results is very intuitive. In this section, we calculate a reference premium and apply relativities to adjust up or down the premium based on the risk characteristics of the insuree. Although it is not as simple as in the case of a GLM, it is possible to replicate a classical system of relativities with the results of a GAM with cubic splines.

To show how to build a simple table of rates with the results of the GAM, we consider the case of model 1 (note that subscript i has been omitted):

$$\log(\mu) = \beta_0 + f_1(km) + f_2(d)$$

where μ is the average number of reported claims and β_0 is the constant term of the model. Since the model has already been estimated, estimates of β_0, and $f_2(d)$ are available. This is the three-step procedure to develop of a price structure with the results of model 1:

1. The reference premium is equal to $\exp(\hat{\beta}_0) = \exp(-2.27352)$;
2. The relativities for distance traveled equal $km_j, j = 1, \ldots, m$ to are $\exp(\hat{f}_1(km_j))$;
3. The relativities for chosen values of the exposure duration $d_k, k = 1, \ldots, h$ are $\exp(\hat{f}_2(d_k))$.

We assume that we want to apply price segments as a function of the distance traveled and exposure duration. For the distance traveled we use intervals of 500 km and for the exposure time intervals of 0.05 years. These values are arbitrary and changeable, but obviously, the finer the segmentation, the better the predictions obtained directly with the GAM approach. Therefore, as the highest value registered for the distance traveled is 50,035 km, there will be m = 102 relativities associated with variable km. Similarly, since the maximum duration of exposure in 2011 is 1 (the whole year), there will be h = 20 relativities associated with the variable d. In total, there will be 2040 relativities associated with each possible combination (km; d).

Table 11 shows the price structure for the five profiles introduced in the previous section. The premium is the estimated frequency and is calculated by multiplying the total relativity (relativity km by relativity d) by the reference premium exp ($-$2.7352).

Table 11. A simple PAYD price structure based on model 1.

(*km*;*d*)	**Relativity for *km* (*a*)**	**Relativity for *d* (*b*)**	**Total Relativity (*c*) = (*a*) × (*b*)**	**Premium exp(−2.7352) × (*c*)**
(3500; 0.35)	0.9975	0.7144	0.7127	0.0462
(4500; 0.50)	1.0944	0.9341	1.0223	0.0663
(9000; 0.65)	1.4154	1.1059	1.5652	0.1016
(15,500; 0.90)	1.7713	1.2540	2.2213	0.1441
(19,000; 1.00)	1.8665	1.2851	2.3988	0.1556

Figure 15 shows the price structure that we have just created for model 1, which replicates Figure 3 very accurately. The same methodology can be applied to the GAM with a tensor product base. The only difference is that we will get directly relativities for profiles (km; d), which are equal to $\exp(\hat{f}(km, d))$ The reference premium can be calculated as exp($-$2.7401). Table 12 shows the results for the different profiles. We can also see that the premiums calculated are very similar to those in Table 12. The prediction surface is also presented in Figure 15 (right). This method manages to replicate the results of GAM models very accurately (see also Figure 9), while adopting a standard rate structure with relativities in that case.

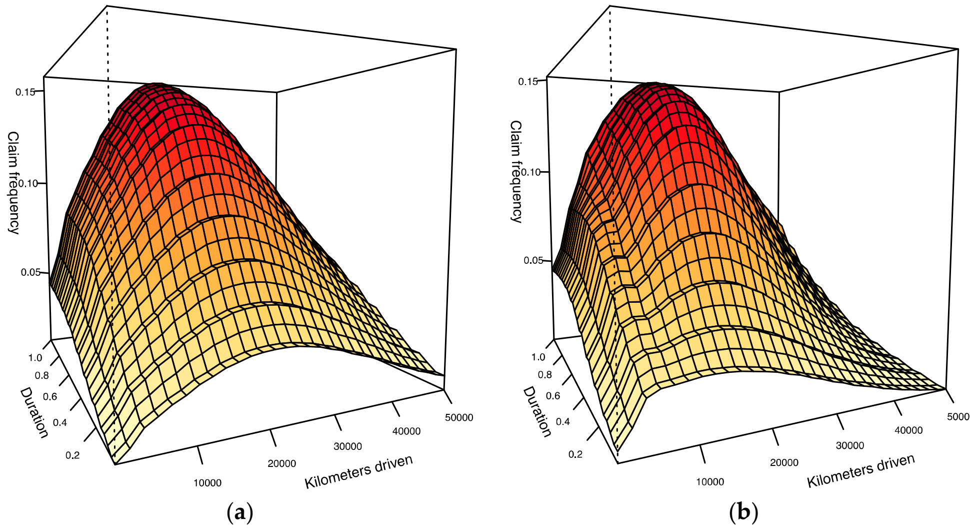


Figure 15. PAYD price structure—surface predictions for models 1 (**a**) and 2 (**b**).

Table 12. PAYD price structure based on model 2.

(*km*;*d*)	**Relativity for (*km*;*d*)**	**Premium** [1]
(3500; 0.35)	0.7904	0.0510
(4500; 0.50)	1.0691	0.0690
(9000; 0.65)	14565	0.0940
(15,500; 0.90)	2.2103	0.1427
(19,000; 1.00)	2.3428	0.1513

[1] Premium is obtained as exp(−2.7401) × Relativity(*km*;*d*).

4. Discussion and Conclusions

In the coming years, we will probably see a revolution in the general non-life insurance world (i.e., car insurance or household insurance). The new technology is going to transform our daily life, and probably change the face of insurance in the medium term. In recent years, a new type of automobile insurance called PAYD has been developed thanks to the recent possibility of installing a GPS in the car, allowing the calculation of the premium based on the information obtained through this device. These GPS devices transmit a lot of data and this represents a challenge for insurance companies who need to manage and effectively use this technology.

The aim of this paper is to provide some answers to the question of whether it is possible to change the way of calculating car insurance premiums. A database of an insurer operating in Spain, which collected usage-based data, has been used. Among other information, we have been able to use the exact number of kilometers (collected by GPS) for each insuree. The GAM approach has been used to measure the impact of the distance traveled and exposure duration on the risk of claims in automobile insurance.

Initially, the GAM approach based on independent cubic splines highlighted the existence of a non-proportional relationship between the number of claims and the traveled kilometers. It was also noted that the expected number of claims appears to stabilize after a certain distance traveled, suggesting that people who drive a lot tend to develop better driving skills. In addition, regarding the exposure time, the model results are in contradiction to what is usually done in practice, which is to assume that the number of claims is proportional to the duration of the contract.

Secondly, the GAM using a tensor product base (which includes an interaction between the distance traveled and exposure duration) has proved to be performing statistically slightly better than the GAM with independent cubic splines. However, the improvement in predictions is very small and the addition of the interaction to the model produces a much higher standard deviation of the adjustment. The GAM using a tensor product base lead us to the same conclusions and interpretations discussed in the previous paragraph, because of the similarity of the results obtained with the two GAM approaches.

We also compare a pricing model widely used in practice with the two GAM approaches. The analysis performed on the validation dataset regarding residuals shows that the overall performance of the three models is similar. In contrast, the analysis of the contribution of the age of the driver in the three models clearly indicates that the traditional model treats some profiles of individuals in a different way compared to the GAM approach.

GAMs are often difficult to interpret but in practice they offer more flexibility than other alternatives. Future work could determine the actual value (benefits, improved customer satisfaction, etc.) to implement pricing systems based on GAM.

Finally, to highlight some of the benefits of GAM, it should be noted that we have proposed a simple PAYD price structure based on the results of such models. In that sense, we have considered a conventional price structure in which a reference premium is multiplied by relativities that have been obtained from the combination of the effect of the distance traveled and exposure duration. The contribution of GAM could be relevant in future research where more telemetric information could be introduced in the pricing system, such as sudden accelerations or braking, without including necessarily the moment and location of driving, which is, for many drivers, a privacy concern. The dependence between different types of claims could also be studied, in order to add the contracts of the same insuree. In such situations, the Generalized Additive Models for the Location, Scale and Shape (GAMLSS) proposed by Rigby and Stasinopoulos (2005) could be considered.

Our aim is to start finding ways to correct the premium dynamically based on telematics information, where the premium should already have been calculated knowing the characteristics of the driver. Therefore, indirectly, driving styles would already have been captured in the original premium. We argue that PAYD should evolve towards pricing parallel to the traditional products, with some bonuses or rebates once telematics data are available. It is still a bit premature to propose an online system where drivers would pay by trip, or at the end of the day for the traveled distance, however exposure to risk will certainly play a central role in the future of PAYD. In other words, our contribution concentrates on the importance of kilometers driven and also on the periodicity of premium checking and rebalancing.

Acknowledgments: Jean-Philippe Boucher acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada. Montserrat Guillen appreciates the recognition of ICREA Academia and ECO2016-76203-C02-2-P. We thank Ana M. Pérez-Marín for support and many interesting inputs in this work.

Author Contributions: J.-P.B. and S.C. estimated the models; M.G. prepared the initial data; all authors wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

Alcañiz, Manuela, Montserrat Guillen, Miguel Santolino, Daniel Sánchez-Moscona, Oscar Llatje, and Lluís Ramon. 2014. Prevalence of alcohol-impaired drivers based on random breath tests in a roadside survey in Catalonia (Spain). *Accident Analysis and Prevention* 65: 131–41. [CrossRef] [PubMed]

Ayuso, Mercedes, Montserrat Guillen, and Ana María Pérez-Marín. 2014. Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention* 73: 125–31. [CrossRef]

Ayuso, Mercedes, Montserrat Guillen, and Ana María Pérez-Marín. 2016a. Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C* 68: 160–67. [CrossRef]

Ayuso, Mercedes, Montserrat Guillen, and Ana María Pérez-Marín. 2016b. Telematics and gender discrimination: some usage-based evidence on whether men's risk of accident differs from women's. *Risks* 4: 10. [CrossRef]

Ayuso, Mercedes, Montserrat Guillen, and Jens Perch Nielsen. 2017. Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation*, under revision. [CrossRef]

Baecke, Philippe, and Lorenzo Bocca. 2017. The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems* 98: 69–79. [CrossRef]

Bolderdijk, Jan Willem, Jasper Knockaert, Linda Steg, and Erik T. Verhoef. 2011. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. *Accident Analysis and Prevention* 43: 1181–86. [CrossRef] [PubMed]

Bordoff, Jason, and Pascal J. Noel. 2008. Pay-As-You-Drive Auto Insurance: A Simple Way to Reduce Driving-Related Harms and Increase Equity. Hamilton Project Discussion Paper. Washington, DC, USA: The Brookings Institution.

Boucher, Jean-Philippe. 2013. Regression with Count Dependent Variables. In *Predictive Modeling Applications in Actuarial Science*. Edited by Edward W. Frees, Richard Derrig and Glenn Meyers. Cambridge: Cambridge University Press. ISBN 978-1107029873.

Boucher, Jean-Philippe, and Michel Denuit. 2007. Duration Dependence Models for Claim Counts. *Blätter der Deutsche Gesellschaft fur Versicherungsmathematik (German Actuarial Bulletin)* 28: 29–45. [CrossRef]

Boucher, Jean-Philippe, and Montserrat Guillen. 2009. A survey on models for panel count data with applications to insurance. *RACSAM, Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas* 103: 277–94. [CrossRef]

Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillen. 2009. Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance* 76: 821–64. [CrossRef]

Boucher, Jean-Philippe, Ana Maria Pérez-Marín, and Miguel Santolino. 2013. Pay-as-you-drive insurance: The effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles* 19: 135–54.

Butler, Patrick, Twiss Butler, and Laurie L. Williams. 1988. *Sex-Divided Mileage, Accident, and Insurance Cost Data Show That Auto Insurers Overcharge Most Women*. Kansas City: National Association of Insurance Commissioners.

Buxbaum, Jeffrey N. 2006. *Mileage-Based User Fee Demonstration Project: Pay-As-You-Drive Experimental Findings*. Technical Report; St. Paul: Minnesota Department of Transportation.

Hastie, Trevor, and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science* 1: 297–310. [CrossRef]

Hastie, Trevor, and Robert Tibshirani. 1990. *Generalized Additive Models*. Boca Raton: CRC Press, vol. 43. ISBN 978-0412343902.

Henckaerts, Roel, Katrien Antonio, Maxime Clijsters, and Roel Verbelen. 2017. *A Data Driven Binning Strategy for the Construction of Insurance Tariff Classes (No. 583471)*. Leuven: Department of Decision Sciences and Information Management, Faculty of Economics and Business, KU Leuven.

Iqbal, Muhammad Usman, and Samsung Lim. 2006. A Privacy Preserving GPS-Based Pay-As-You-Drive Insurance Scheme. Paper presented at Symposium on GPS/GNSS (IGNSS2006), Queensland, Australia, July 17–21.

Jun, Jungwook, Jennifer Harper Ogle, and Randall L Guensler. 2007. Relationships between crash involvement and temporal-spatial driving behavior activity patterns: Use of data for vehicles with global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board* 2019: 246–55. [CrossRef]

Langford, Jim, Sjaanie Koppel, Dennis McCarthy, and Sivaramakrishnan Srinivasan. 2008. In defense of the 'low-mileage bias'. *Accident Analysis and Prevention* 40: 1996–99. [CrossRef]

Lemaire, Jean, Sojung Carol Park, and Kili C. Wang. 2016. The use of annual mileage as a rating variable. *ASTIN Bulletin* 46: 39–69. [CrossRef]

Litman, Todd. 2005. Pay-as-you-drive pricing and insurance regulatory objectives. *Journal of Insurance Regulation* 23: 35–53.

Litman, Todd. 2011. *Pay-As-You-Drive Insurance: Recommendations for Implementation*. Victoria: Victoria Transport Policy Institute, Available online: www.vtpi.org (accessed on 18 January 2017).

Lourens, Peter F., Jan A. M. M. Vissers, and Maaike Jessurun. 1999. Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. *Accident Analysis & Prevention* 31: 593–97. [CrossRef]

Paefgen, Johannes, Thorsten Staake, and Frédéric Thiesse. 2013. Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems* 56: 192–201. [CrossRef]

Paefgen, Johannes, Thorsten Staake, and Elgar Fleisch. 2014. Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice* 61: 27–40. [CrossRef]

Rigby, Robert A., and D. Mikis Stasinopoulos. 2005. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54: 507–54. [CrossRef]

Tselentis, Dimitrios I., George Yannis, and Eleni Vlahogianni. 2017. Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis and Prevention* 98: 139–48. [CrossRef] [PubMed]

Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. 2017. *Unravelling the Predictive Power of Telematics Data in Car Insurance Pricing*. KBI Research Report—KBI1624; Leuven: KU Leuven, Available online: https://ssrn.com/abstract=2872112 or http://dx.doi.org/10.2139/ssrn.2872112; (accessed on 18 January 2017).

Vickrey, William. 1968. Automobile accidents, tort law, externalities, and insurance: An economist's critique. *Law and Contemporary Problems* 33: 464–87. [CrossRef]

Wood, Simon N. 2006. Generalized Additive Models: An Introduction with R. Boca Raton: Chapman & Hall/CRC Texts in Statistical Science, Abingdon: Taylor & Francis.