Percentile charts for speeding based on telematics information

Montserrat Guillen[*]

Dept. Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain; mguillen@ub.edu ([*]corresponding)

Ana M. Pérez-Marín

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

Manuela Alcañiz

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

Abstract

Reference charts are widely used as a graphical tool for assessing and monitoring children's growth given gender and age. Here, we propose a similar approach to the assessment of driving risk. Based on telematics data, and using quantile regression models, our methodology estimates the percentiles of the distance driven at speeds above the legal limit depending on drivers' characteristics and the journeys made. We refer to the resulting graphs as *percentile charts* for speeding and illustrate their use for a sample of drivers with Pay-How-You-Drive insurance policies. We find that percentiles of distance driven at excessive speeds depend mainly on total distance driven, the percentage of driving in urban areas and the driver's gender. However, the impact on the estimated percentile for these covariates is not constant. We conclude that the heterogeneity in the risk of driving long distances above the speed limit can be easily represented using reference charts and that, conversely, individual drivers can be scored by calculating an estimated percentile for their specific case. The dynamics of this risk score can be assessed by recording drivers as they accumulate driving experience and cover more kilometres. Our methodology should be useful for accident prevention and, in the context of Manage-How-You-Drive insurance, reference charts can provide real-time alerts and enhance recommendations for ensuring safety.

Keywords: motor insurance; speed; telematics; quantile regression; reference curves; risk score.

1. Introduction

Growth reference charts are used worldwide to provide a simple graphical tool for monitoring the evolution in children's height and weight. As such, they enable doctors and parents to track a child's estimated percentile path and observe his or her position with respect to that of their corresponding reference population of either boys or girls. Here, we seek to design a similar tool for assessing driving risk, based on the distance driven above the posted speed limit as an indicator of peril. A driver's risk evolution is then analysed with respect to total distance driven and other circumstances that need to be taken into consideration, primarily driving zone. The tool developed is both highly informative and simple, and can be directly used to communicate driving risk.

Recent research on road traffic safety specifically related to speeding highlights that speed increases both the risk and severity of an accident (see Dissanayake and Lu, 2002; Ossiander and Cummings, 2002; Jun et al., 2007, 2011; Vernon et al., 2004 and Viallon and Laumon, 2013). Viallon and Laumon (2013) analysed the effectiveness of the speed regulation policies introduced during the period 2001-2010 in France with respect to high-level speeding. The authors built a model which relates the number of fatal crashes to speed, and highlighted the effectiveness of speed regulation policies. More recently, Arvin et al. (2019) investigated the relationship between pre-crash driving instability and crash intensity. They found that higher instability in driving increases the probability of a severe crash, and that the speed prior a crash is highly correlated with its intensity. Bogstrand et al. (2015) analysed fatal road traffic crashes in Norway during 2005-2010, and found statistically significant associations between impairment by alcohol or amphetamines and driving unbelted or speeding. The authors also found that excessive speeding is one of the main reasons for traffic crashes. Additionally, speeding and being unbelted are the main reasons for a fatal outcome. More recently, Høye (2020) intestigated fatal car crashes in Norway as well, but from 2005 to 2015. They found that individuals who drive under the influence of alcohol and/or drugs are more often male, unbelted and unlicensed. Moreover, they normally drive old cars, and are involved in single-vehicle crashes at night, in the weekend and on low-volume roads.

There are also evidences that drivers are not homogeneous with respect to their level of risk and driving style. Specifically, men present riskier driving patterns, driving more kilometres per day, during the night and at speeds above the limit than women (Ayuso et al., 2014, 2016a, 2016b). All these factors have been shown to be associated with a greater number of accidents (Gao et al., 2019a; Gao and Wüthrich, 2019; Guillen et al., 2019). Moreover, Paefgen et al. (2014) report that the risk of accident is higher on urban roads, during weekends, at nightfall and at low- (0–30 km/h) or high-range speeds (90–120 km/h). Indeed, Pérez-Marín and Guillen (2019) concluded that if excess speeds could be eliminated, the expected number of accident claims would be reduced by half. Interestingly, Pérez-Marín et al. (2019a) showed that young drivers tend to reduce posted speed limit violations after an accident, probably because they are more aware of the risk.

Speed and driving distance have been exhaustively analysed in transport research (see, for example, Hewson, 2008 or Plötz et al., 2017). Moreover, analyses of speeding in traffic safety research have focused not only on the average speed, but also on its quantiles. Specifically, Hewson (2008) explored the benefits of using quantile regression to evaluate whether or not an intervention is able to significantly modify the 85th percentile speed. Recently, Pérez-Marín et al. (2019b) applied quantile regression to an analysis of the effects of telematics information (location and time of driving and the total distance driven) on a range of percentiles of the distance driven at speeds above the limit by using a sample of drivers covered by a Pay-How-You-Drive (PHYD) insurance policy. In PHYD policies, the premium is calculated based on the customer’s driving pattern (such as speeding, harsh acceleration, sudden braking or hard cornering). Based on these patterns, a driver’s risk score can be obtained and used to calculate his or her premium (see a survey in Arumugam and Bhargavi, 2019).

In this paper, we propose a methodology for displaying percentiles that allows us to quantify a driver’s risk score. To do so, we use a graphical representation of the percentiles of distance driven at speeds above the limit, depending on specific driver characteristics and on the sort of trips they make. Employing charts similar to the well-known reference curves for child growth, we develop a new methodology in the context of speeding that should prove useful when a large number of covariates can influence a driver’s behaviour on the road and, hence, their risk profile.

Specifically, we call our graphs *percentile charts* for speeding, as they provide each driver with their corresponding percentile of distance driven at speeds above the legal limits, given all available information on that driver. This proves to be a straightforward risk score for the driver. We take the article by Perez-Marín et al. (2019) as our starting point, and use the methodology proposed by Wei et al. (2006) in the context of growth charts (based on quantile regression) to produce percentile charts for speeding. We use the same data as presented in Perez-Marín et al. (2019b) and explore alternative model formulations in the context of generalized linear models (GLMs) and quantile regression. In particular, we investigate in-depth the relationship between distance driven at speeds above the legal limits (the dependent variable in our regression models) and total distance driven. We conclude that their relationship is not linear, but exponential. This exponential relationship determines the shape of the percentile charts for speeding. As a result, we also observe that our methodology substantially improves the initial results obtained in Perez-Marín et al. (2019b).

The rest of this paper is organized as follows. In section 2, the quantile regression model and the database used in our study are presented. In section 3, the main results of the regression models are summarized and the percentile charts are provided. In section 4, the main results are discussed, and finally in section 5 the main conclusions of the paper are presented.

2. Material and Methods

2.1. Methods

Percentile charts for speeding are obtained by means of quantile regression, where each curve corresponds to a percentile level. A web application is easily designed, so that when a user enters his or her covariate information and observed mileage above the speed limit, a graph is displayed, locating the specific driver on the chart. Quantile regression is specially recommended when the response variable is asymmetric conditional on the explanatory variables. OLS and Gaussian assumptions would produce biased estimates of extreme quantiles (see Khattak et al , 2016, for a comparison in a case study on transportation). Eide and Showalter (1998) presented one of the earliest comparisons between OLS and quantiles regression in the field of education. They found a differential effect of school size at different points of the students' test score gain conditional distribution, meaning that school size has increasing return of scale on the gain in students' scores (see also Castellano and Ho, 2013). Similar examples of quantile regression versus OLS can be found in Bel et al. (2015) and O'Garra and Mourato (2006) in environmental economics.

In this paper, we also fit a GLM model prior to quantile regression; specifically, we fit a gamma model because the dependent variable, which is mileage above the speed limit, is expected to be asymmetric. That is, while a large number of drivers can be expected not to exceed the speed limit over a certain number of kilometres, only a few are expected to exceed the limit over a high percentage of the distance driven.

The τ-quantile of a continuous random variable Y is the value c_τ for which $P(Y \leq c_\tau) = \tau$. In the financial and actuarial industries, the τ-quantile, or the percentile at the level τ, is known as the value-at-risk at level τ. Quantile regression is used in order to estimate conditional quantiles, as the model assumes that the $c_\tau(Y)$ depends on certain explanatory variables. Specifically,

$$c_\tau(Y_i \,|X_{1i}, \dots, X_{ki}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \dots + \beta_k^\tau X_{ki}, \tag{1}$$

where Y_i is the dependent variable for the i-th individual, with $i=1,\dots,n$, and X_{ji} are the observations of the explanatory variables, with $j=1,\dots,k$. It can be proved (Koenker and Bassett, 1978) that

$$\widehat{\beta^\tau} = \underset{b}{\operatorname{argmin}} \left[\sum_{Y_i \geq X_i' b} \tau |Y_i - X_i' \mathrm{b}| + \sum_{Y_i < X_i' b} (1-\tau)|Y_i - X_i' \mathrm{b}| \right]. \tag{2}$$

The objective function (2) corresponds to the sum of n components, called $\rho_\tau(Y_i - X_i'\mathrm{b})$ that are expressed as follows:

$$\rho_\tau(Y_i - X_i'\mathrm{b}) = \tau(Y_i - X_i'\mathrm{b}) I_{\{Y_i \geq X_i' b\}} + (\tau - 1)(Y_i - X_i'\mathrm{b}) I_{\{Y_i < X_i' b\}} =$$
$$= (Y_i - X_i'\mathrm{b})(\tau - I_{\{Y_i < X_i' b\}}), \tag{3}$$

where $I_{\{\cdot\}}$ is an indicator function equal to 1 if the condition in the subindex is fulfilled, and 0 otherwise. A quantile regression model can be easily fitted, for example in R, by using the function *qr* of the *quantreg* R package (Koenker et al., 2018).

Koenker and Machado (1999) proposed an expression to measure the goodness-of-fit of the quantile regression based on a comparison of the values of the objective functions of the estimated model and of the constrained model that only includes an intercept term. Specifically, let

$$\hat{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(Y_i - X_i' \widehat{\beta^\tau}) \tag{4}$$

be the value of the objective function of the estimated model and

$$\tilde{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(Y_i - \beta_0^\tau) \tag{5}$$

be the value of the objective function of the constrained model that only includes the intercept term. Then, the goodness-of-fit measure proposed by Koenker and Machado (1999) is

$$R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau) \tag{6}$$

which is similar to the R^2 in the multiple linear regression model. Additional details of quantile regression implementation in R can be found in Uribe and Guillen (2020).

2.2. Data

The dataset used in this article is the same as that employed in Pérez-Marín et al. (2019b). The data contain the complete portfolio of 9,585 drivers aged 35 years or less with a PHYD policy in a Spanish motor insurance company. All insureds covered in 2010 were included. The description of the variables is presented in Table 1. We know the gender (variable *Gender*) and age of the driver at the beginning of 2010 (variable *Age*). Additionally, we also know the total number of kilometres driven during 2010 (*Km*), the number of kilometres driven at speeds above the posted limit (*Tolerkm*, which is our dependent variable), the percentage of kilometres driven on urban roads (*Urban*) and, finally, the percentage of kilometres driven at night (*Night*). An urban road is a segment of a road within the boundaries of a built-up area, which is an area with entries and exits especially sign-posted as such and where speed is limited by law at least to 50 Km/h. In order to fit the gamma model, note that 29 observations with zero kilometres driven at speeds above the posted limit – 0.3% of the size of the dataset – had to be removed from the original data base.

Table 1. Description of variables used in the insurance dataset

Variable	Description
Tolerkm	Number of kilometres driven at speeds above the posted limit during 2010
Km	Total number of kilometres driven during 2010*
Urban	% of kilometres driven on urban roads during 2010*
Night	% of kilometres driven at night (between midnight and 6 am) during 2010
Age	Age of the driver at the beginning of 2010
Gender	1 = male, 0 = female

*Power transformations were used in the gamma model, $Km\_tg = Km^{0.1}$ and $Urban\_tg = Urban^{0.7}$, and in the quantile regression models, , $Km\_tqr = Km^{1.7}$ and $Urban\_tqr = Urban^{0.1}$

As shown in Table 2, *Tolerkm* presents a positive asymmetry (skewness coefficient = 3.64), with a long tail. The sample comprises 49% women and 51% men. The average age of drivers is 24.78 years. The average number of kilometres driven during the observed year was 13,099.91 (standard deviation of 7,698.98). On average, drivers travelled 26.2% of kilometres on urban roads, 7.02% of kilometres at night and 1,402.44 kilometres at speeds above the limit.

Table 2. Descriptive statistics of the insurance data set

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	St. Dev.	Skewness
Tolerkm	0.03	285.78	692.92	1,402.44	1,710.44	23,500.19	1,996.90	3.64
Km	27.79	7,575.15	11,719.83	13,099.91	17,350.12	57,756.98	7,698.98	1.08
Urban	0.00	15.59	23.36	26.20	34.25	96.41	14.04	0.99
Night	0.00	2.49	5.32	7.02	9.85	78.56	6.12	1.67
Age	18.11	22.66	24.63	24.78	26.88	35.00	2.82	0.11

3. Results

The aim of the paper is to fit the conditional percentile of *Tolerkm*. In order to build the model, we need to know *Tolerkm* and some explanatory variables (age, gender, total distance travelled and percentage of urban and nigh ttime driving).

Firstly, we employed a gamma regression model[1] to fit the conditional mean of *Tolerkm* and tried different transformations of *Km* and *Urban* (including logarithmic and power transformations) to minimize the Akaike information criterion (AIC). We also tried to transform the other two continuous explanatory variables (*Night* and *Age*), but this had almost no impact on the AIC score. Power transformations were more effective in order to reduce AIC compared to logarithmic transformations. We sought many power transformations on *Km* and *Urban*, specifically, Km^i and $Urban^j$ where i = 0.05 to 0.5 increasing by 0.05, and j = 0.1 to 1 increasing by 0.1, and finally the combination that reduced the AIC the most (equal to 149,299.5) was $Km\_tg = Km^{0.1}$ and $Urban\_tg = Urban^{0.7}$ for the gamma model.

The parameter estimates of the corresponding gamma regression model are shown in Table 3. Coefficient estimates with a p-value lower than 1% correspond to gender, the transformed total number of kilometres driven (*Km_tg*) and the transformed percentage of kilometres driven in urban areas (*Urban_tg*). Age effect is only significant at the 10% level (p-value=0.0807), probably because the insurance policies were sold exclusively to young drivers. Likewise, the positive effect of percentage of kilometres driven at night (*Night*) is only significant at the 10% level (p-value=0.0987), which would indicate that drivers with a higher percentage of night time driving tend to have an average excess speed distance greater than those with a lower percentage of night time driving. *Km_tg* has a positive parameter estimate, indicating that an increase in the total number of kilometres driven contributes to increasing the expected number of kilometres driven at speeds above the posted limits. In contrast, *Urban_tg* presents the opposite effect: the higher the percentage of kilometres driven on urban roads, the lower the expected number of kilometres driven at speeds above the posted limit. Finally, gender (baseline reference: female) has a positive parameter estimate, indicating that men seem to drive more kilometres at speeds above the posted limit than women.

Table 3. Results of the gamma regression model for the insurance data set. Dependent variable is the number of kilometres driven above posted speed limits

	Parameter estimate (p-value)
Intercept	-5.126659 (<0.0001)
Km_tg	4.966361 (<0.0001)
Urban_tg	-0.065209 (<0.0001)
Night	0.002475 (0.0987)

[1] We also used lognormal (but it provided a higher AIC score) and inverse Gaussian regressions (but it was eventually discarded because of convergence problems in the algorithm).

Age	-0.005587
	(0.0807)
Gender	0.207654
	(<0.0001)

For the quantile regression model we proceed in the same way as in the gamma model, and searched different transformations (including logarithmic and power transformations) and finally those that reduced the AIC the most were *Km_tqr* = $Km^{1.7}$ and *Urban_tqr* = $Urban^{0.1}$.

The parameter estimates and goodness-of-fit of the quantile regression models at different levels ($\tau = 0.5, 0.75, 0.90, 0.95, 0.975, 0.99$) are shown in Table 4. We see that *Km_tqr* has a significant effect, with a positive parameter estimate, for all levels of the quantile. This means that, for a specific quantile, increasing the total number of kilometres driven increases the quantile of the number of kilometres driven at speeds above the posted limits, *ceteris paribus*. In contrast, while *Urban_tqr* also has a significant effect, it has a negative parameter estimate. Thus, as the percentage of kilometres driven in urban areas increases, the quantile of the number of kilometres driven at speeds above the limits decreases. *Night* has a significant effect only when estimating the median of the kilometres driven at speeds above the limits, but for other levels of the quantile, it has no significant effect. In the case of the median, the parameter estimate is positive, indicating that increasing the percentage of kilometres driven at night increases the median kilometres driven at speeds above the limits. *Age* has a significant effect only when estimating the quantiles at the 95th and 97.5th levels. In both cases, the corresponding parameters are positive; thus, increasing the driver's age also increases the corresponding percentiles of the distance driven at speeds above the limits. Finally, gender (baseline reference: female) has a significant parameter for all levels of the quantiles up to the 95th. The coefficient is positive; thus, men have higher percentile values of distance driven at speeds above the limits than women. In the case of the goodness-of-fit criterion, it is apparent that the contribution explaining the quantiles of the model with covariates vs. the model without increases with the increase in percentile level, reaching 61.22% at the 99th level. Additionally, in Figure A1 in the Appendix we also provide the marginal effect (estimated parameter) of each explanatory variable in the quantile regression models, as a function of the level of the estimated quantile, showing that the impact of covariates on different percentile levels is not always constant, which highlights the great utility of reference charts as graphical tools.

Figure 1 shows the percentile charts for speeding for males and females, respectively, together with the sample data. The plots show *Tolerkm* vs. *Km*, and additionally the grey lines represent the estimated quantiles at different levels. The red line represents the conditional of *Tolerkm* estimated using the gamma regression model in Table 3. In Figure 1, the values of *Urban*, *Night* and *Age* have been fixed at the mean values in the sample for men and women, respectively. Table 5 provides various examples of percentiles obtained when using the speed reference curves in Figure 1. For example, if a male driver drives 10,000 km per year and 2,000 km are above the limit, he is in the 90th percentile curve. On the other hand, the same driver is in the 54th percentile curve if he drives 20,000 km per year, and finally, he is in the 29th percentile curve if he drives 30,000 km per year. The corresponding percentiles for women are also shown in Table 5, and are very similar if just a little higher, indicating that women seem to drive at speeds above the posted limit speed less than men. For example, if a woman drives 10,000 km per year and 2,000 of them at speeds above the limits, she is in the 92th percentile curve.

Table 4. Parameter estimates of the quantile regression model for different percentiles of mileage above the speed limit

	50th percentile (p-value)	75th percentile (p-value)	90th percentile (p-value)	95th percentile (p-value)	97.5th percentile (p-value)	99th percentile (p-value)
Intercept	1077.1401	3359.33609	6581.18447	8506.75790	10173.74522	11288.44387
	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)
Km_tqr	0.00008	0.00013	0.00020	0.00024	0.00028	0.00032

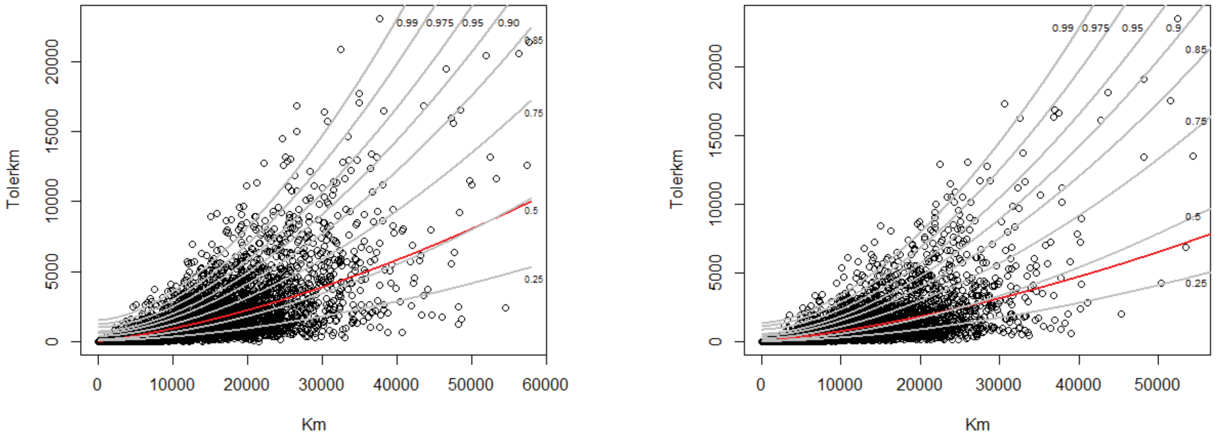
	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)
Urban_tqr	-739.33632	-2285.83883	-4529.95580	-5938.64364	-7176.65122	-7960.61577
	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)	(<0.00001)
Night	2.36200	1.08333	-0.94645	4.77960	9.13793	-0.57912
	(0.00224)	(0.43992)	(0.64086)	(0.18409)	(0.26052)	(0.95843)
Age	-1.80286	-1.22676	6.18758	17.61657	28.01504	37.93405
	(0.20585)	(0.70373)	(0.25224)	(0.02532)	(0.00894)	(0.09753)
Gender	104.81103	167.33072	167.22760	140.47488	101.91653	189.44456
	(<0.00001)	(<0.00001)	(<0.00001)	(0.00360)	(0.16106)	(0.23436)
Goodness of fit (%)	23.43	33.59	44.56	50.89	55.55	61.22

As discussed above, the speeding reference curves shown in Figure 1 have been obtained by assuming that the other explanatory variables (*Urban*, *Night* and *Age*) are equal to the corresponding sample means for men and women, respectively. In Figure 2 we show how these reference curves for the 95th percentile change for different values of *Urban* (which is the most relevant explanatory variable, apart from *Km*, for explaining *Tolerkm*). Specifically, for men we considered values of *Urban* equal to 8.70, 23.45 and 52.47% (5th, 50th and 95th percentiles of *Urban* in the male sample, respectively), and we refer to these values as low, median and high levels of urban driving. Similarly, for women we considered values of *Urban* equal to 8.37, 23.01 and 53.81% (5th, 50th and 95th percentiles of *Urban* in the female sample, respectively), and similarly we refer to them as low, median and high levels of urban driving. The corresponding reference speed curves for the 95th percentile of *Tolerkm* are represented in Figure 2 for men and women, respectively, where the red lines are used to represent the corresponding curves for the average values obtained with the gamma regression model (Table 3). We observe that, as the percentage of urban driving increases, all curves move downwards, as *Urban_tqr* has a negative coefficient. Specifically, in Table 6 we show some examples of the 95th percentile of *Tolerkm* for certain values of *Urban* and *Km*. For a male driver driving 10,000 km per year the 95th percentile of *Tolerkm* is equal to 1,753.22 km if he has a high percentage of urban driving, 2,466.03 km if he has a median percentage and 3,234.56 km if he has a low percentage. When the distance driven by the male driver increases to 20,000 km per year, then the 95th percentile of *Tolerkm* is equal to 5,120.66, 5,803.48 and 6,572.01km for high, median and low percentages of urban driving, respectively. Table 6 also shows the results corresponding to women drivers, and we observe that they are slightly lower than those for male drivers.

Table 5. Percentiles obtained using the percentile charts for examples of speeding (*Tolerkm*) and total distance driven (*Km*) by gender.

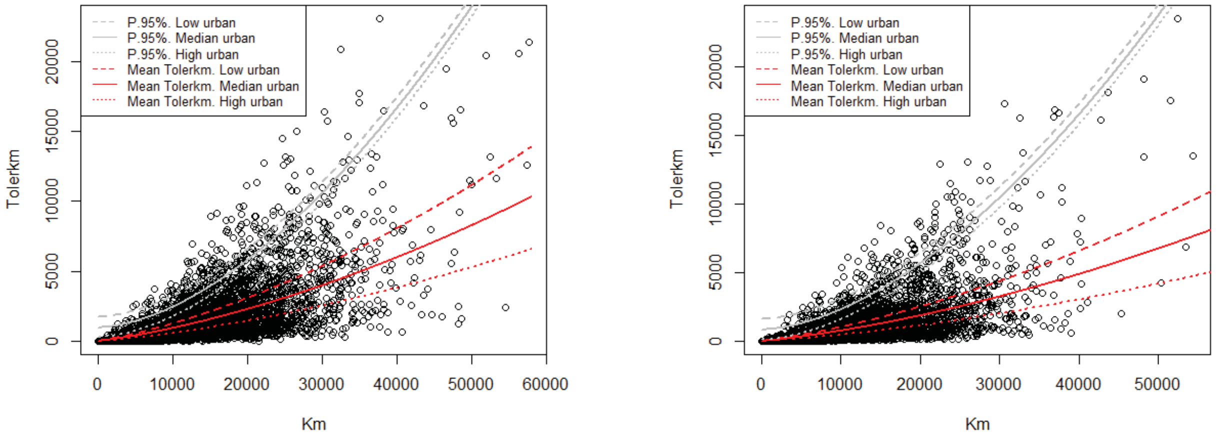
			Km		
			10,000	20,000	30,000
Tolerkm	Men	2,000	90th	54th	29th
		3,000	98th	74th	44th
	Women	2,000	92th	57th	30th
		3,000	98th	75th	45th

Other explanatory variables (Urban, Night and Age) are equal to their sample means for men and women, respectively.



Male drivers Female drivers

Figure 1. Percentile chart for speeding for male drivers (left) and female drivers (right). *Tolerkm* vs. *Km*, where grey lines represent the estimated quantiles at different levels. The red line represents the mean of *Tolerkm* estimated using the gamma regression model in Table 3. Age, urban and night driving are fixed at the mean level by gender.



Male drivers Female drivers

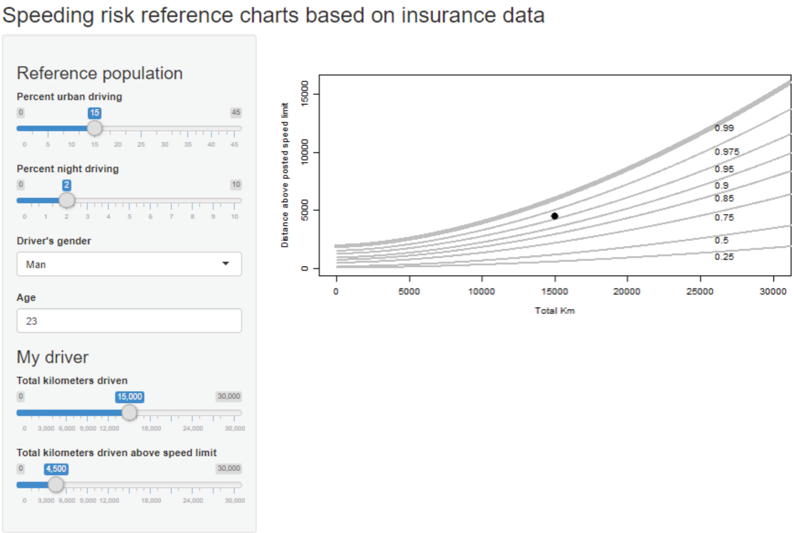
Figure 2. Percentile chart for speeding for male drivers (left) and female drivers (right) at the 95th level. *Tolerkm* vs. *Km*, where grey lines represent the estimated 95th percentile and red lines represent the mean of *Tolerkm* estimated using the gamma regression model, for different values of *Urban* (dashed = low level of urban driving, solid = median level, and dotted = high level). Age and night driving are fixed at the sample mean level by gender.

Table 6. Estimated *Tolerkm* for the 95th percentile chart for different values of *Urban* and *Km* for men and women. Age and night driving are fixed at the sample mean level.

		Km		
		10,000	20,000	30,000
Men	Low *Urban*	3,234.56	6,572.01	11,356.31
	Median *Urban*	2,466.03	5,803.48	10,587.78
	High *Urban*	1,753.22	5,120.66	9,904.96
Women	Low *Urban*	3,110.06	6,447.51	11,231.81

Median *Urban*	2,328.48	5,665.93	10,450.22
High *Urban*	1,607.98	4,945.43	9,729.72

An interactive graphical tool that displays the evolution of a driver's speeding percentile as a function of total distance driven, night-time driving, gender and principal driving zone can be seen in Figure 3 and it can also be accessed online[2] .



Speeding risk reference charts based on insurance data

Reference population

Percent urban driving: 15

Percent night driving: 2

Driver's gender: Man

Age: 23

My driver

Total kilometers driven: 15,000

Total kilometers driven above speed limit: 4,500

Figure 3. Example of interactive speeding percentile chart that locates a particular driver (black dot), given total distance driven, total speeding kilometres and all other reference characteristics stated in the left panel.

4. Discussion

We have found that the most relevant variables explaining the number of kilometres driven at speeds above the limits are: total distance driven, percentage of urban driving and gender. This result is in line with those obtained previously by Perez-Marín et al. (2019b), but with a clear improvement in the methodology, more refined results and providing a tool enhancing their applicability (the *percentile charts for speeding*). Moreover, we also conclude that men have riskier driving patterns compared to women, which is in line with previous research (Høye, 2020; Ayuso et al., 2014; 2016a; 2016b). In most of the models for these data, we found that age and night time driving do not have a significant impact, although they are relevant factors explaining the risk of an accident (see Paefgen et al, 2014, Ayuso et al., 2014 and 2016b). In both cases, this appears to be due to the lack of variability in the PHYD policies, which in our sample were sold exclusively to young drivers. Note that, night time diving could have been included as a categorical variable (indicating whether or not the drivers use the car during the night). Nevertheless, we have percent driving at night (percentage of kilometres travelled at night) and this is why we use this covariate instead of a qualitative indicator.

We analysed the relationship between the distance driven at speeds above the limits and the total distance driven, and found this relationship not to be linear, but rather exponential. This means that as the total distance driven increases, the number of kilometres driven at speeds above the limits also increases, but at an ever-increasing rate. This might be due to the driving experience gained or to an excess of confidence on the part of the driver. A nonlinear effect was also found between the driving experience (measured by the distance travelled) and the risk of accident (see Boucher et al. 2013). In that case, the authors found that the risk associated to a higher number of driven kilometres is fully balanced by the larger experience of the driver and

[2] https://riskcenter.shinyapps.io/speeding_risk_reference_chart/

the other safety factors, and, under certain conditions, the expected frequency of claims is not increased by the number of driven kilometres.

The exponential relationship introduced in the covariates of quantile regressions determines the shape of the reference risk curves for driving at excess speeds. Such models allow the factors associated with higher quantile values to be identified and, hence, for risky drivers to be detected. Our results contribute to calculating the percentile risk score for each driver by controlling for their specific characteristics (and not for the whole population of drivers). Based on these quantile regression models, risk reference curves have been obtained. These graphical tools provide, for each driver, the corresponding percentile of the distance driven at speeds above the limits (which constitutes that driver's risk score), as a function of the total distance driven. Moreover, these curves can be easily obtained for particular types of driver, depending on their characteristics (gender, percentage of urban driving, etc.).

We consider this methodology of risk quantification to be very useful in application with Manage-How-You-Drive (MHYD) insurance products, where the premium is calculated using the same procedure as that used in PHYD insurance, but, in addition, drivers are provided with real-time alerts and recommendations for guaranteeing their safety (Arumugam and Bhargavi, 2019). As such, MHYD insurance improves both customer service and protection in the sector. In this context, the methodology presented here is able to deliver valuable graphical information in terms of preventive early warnings. Estimating just how a driver ranks with respect to distance driven above the posted speed limit is personalized information that should constitute interesting feedback for policy holders (Pérez-Marín et al., 2019b). However, it is necessary to remark that if the result of this study is provided to the driver, drivers could drive at speeds above the legal limit with confidence before they reach a certain speeding mileage. So, we advocate that communication should warn that speeding is always dangerous and should be avoided in any circumstances. Here, it should be stressed that excess speed is perhaps the only feature a driver can easily modify, given that other factors, such as percentage of urban driving, are largely determined by external circumstances and drivers are essentially unable to change them. Indeed, Pérez-Marín et al. (2019a) report that young drivers have a tendency to reduce speed limit violations after an accident, probably because of their greater awareness of the associated risks. As speed is the main cause of severe accidents, those who present lower risk scores (a lower percentile on the risk reference curve) should probably have lower insurance premiums. As to how this ranking should be translated into an insurance price is a question we leave for further research, but there is no doubt that direct bonuses rewarding careful drivers could easily be introduced.

One limitation of the analysis reported here, and which should be pointed out, is that the degree to which drivers exceeded the posted speed limit was not recorded and, therefore, we do not know the severity of the speed violation.

5. Conclusions

In this paper, we have presented the design for a prototype graphical tool that displays the evolution of a driver's speeding risk percentile as a function of total distance driven, night time driving, gender and principal driving zone. We conclude that the effect of these covariates on the estimated percentile is not constant. The heterogeneity in the risk of driving above the speed limit can be easily represented using our graphical tool. Moreover, our interface produces a personalized percentile that provides immediate feedback to the user. By measuring a driver's current speeding based on telemetry, that driver can see their evolution over distance driven and they can be provided with a score that is based on their peers' driving records. We firmly believe such reference charts are set to become a standard in the visualization of driving risk.

References

Ayuso, M., Guillen, M. and 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: 125–31. DOI: 10.1016/j.aap.2014.08.017.

Ayuso, M., Guillen, M. and Pérez-Marín, A.M. 2016a. Telematics and gender discrimination: some usage-based evidence on whether men's risk of accident differs from women's. Risks 4:2: 10. DOI: 10.3390/risks4020010.

Ayuso, M., Guillen, M. and Pérez-Marín, A.M. 2016b. Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. Transportation Research Part C Emerging Technologies 68: 160–7. DOI: 10.1016/j.trc.2016.04.004.

Arumugam, S. and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data, Journal of Big Data, 6, 86, 1-21. DOI: 10.1186/s40537-019-0249-5.

Arvin, R., Kamrani, M., and Khattak, A. J. 2019. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. Accident Analysis & Prevention, 132, 105226. DOI: 10.1016/j.aap.2019.07.002.

Bel, G., Bolancé, C., Guillén, M., and Rosell, J. 2015. The environmental effects of changing speed limits: A quantile regression approach. Transportation Research Part D: Transport and Environment, 36, 76-85. DOI: 10.1016/j.trd.2015.02.003.

Bogstrand, S. T., Larsson, M., Holtan, A., Staff, T., Vindenes, V., and Gjerde, H. 2015. Associations between driving under the influence of alcohol or drugs, speeding and seatbelt use among fatally injured car drivers in Norway. Accident Analysis & Prevention, 78, 14-19. DOI: 10.1016/j.aap.2014.12.025.

Boucher, J. P, Pérez-Marín, A. M. and Santolino, M. 2013. Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. Anales del Instituto de Actuarios Españoles, 3ª Época 19: 135-54.

Castellano, K. E., and Ho, A. D. 2013. Contrasting OLS and quantile regression approaches to student "growth" percentiles. Journal of Educational and Behavioral Statistics, 38(2), 190-215. DOI: 10.3102/1076998611435413.

Dissanayake, S. and Lu, J.J. 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accident Analysis and Prevention 34: 5: 609–18. DOI: 10.1016/S0001-4575(01)00060-4.

Eide, E., and Showalter, M. H. 1998. The effect of school quality on student performance: A quantile regression approach. Economics letters, 58(3), 345-350. DOI: 10.1016/S0165-1765(97)00286-3.

Gao, G., Meng, S. and Wüthrich, M.V. 2019. Claims frequency modeling using telematics car driving data. Scandinavian Actuarial Journal 2019: 2: 143-62. DOI: 10.1080/03461238.2018.1523068.

Gao, G. and Wüthrich, M.V. 2019. Convolutional neural network classification of telematics car driving data. Risks 7: 1: 6. DOI: 10.3390/risks7010006.

Guillen, M., Nielsen, J.P., Ayuso, M. and Pérez-Marín, A.M. 2019. The use of telematics devices to improve automobile insurance rates. Risk Analysis, 39: 3: 662-72. DOI: 10.1111/risa.13172.

Hewson, P.J. 2008. Quantile regression provides a fuller analysis of speed data. Accident Analysis and Prevention 40: 502–10. DOI: 10.1016/j.aap.2007.08.007.

Høye, A. 2020. Speeding and impaired driving in fatal crashes—Results from in-depth investigations. Traffic injury prevention, 1-6. DOI: 10.1080/15389588.2020.1775822.

Jun, J., Ogle, J. and Guensler, R. 2007. Relationships between crash involvement and temporal-spatial driving behavior activity patterns: use of data for vehicles with global positioning systems. Transportation Research Record 2019: 246–55. DOI: 10.3141/2019-29.

Jun, J., Guensler, R. and Ogle, J. 2011. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: application of in-vehicle monitoring technology. Transportation Research Part C Emerging Technologies 19: 4: 569–78. DOI: 10.1016/j.trc.2010.09.005.

Khattak, A. J., Liu, J., Wali, B., Li, X. and Ng, M. 2016. Modeling traffic incident duration using quantile regression. Transportation Research Record, 2554(1), 139-148. DOI: 10.3141/2554-15.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* 46, 1, 33-50.

Koenker, R. and Machado, J.A.F. 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94: 448, 1296-310. DOI: 10.1080/01621459.1999.10473882.

Koenker, R., Portnoy, S., Ng, P.T., Zeisleis, A., Grosjean, P. and Ripley, B.D. 2018. Package ‘quantreg’. R Package Version 5.38, https://cran.r-project.org/web/ packages/quantreg/quantreg.pdf.

O’Garra, T. and Mourato, S. 2007. Public preferences for hydrogen buses: comparing interval data, OLS and quantile regression approaches. Environmental and Resource Economics, 36(4), 389-411. DOI: DOI: 10.1007/s10640-006-9024-0.

Ossiander, E.M. and Cummings, P. 2002. Freeway speed limits and traffic fatalities in Washington State. Accident Analysis and Prevention 34: 13–8. DOI: 10.1016/S0001-4575(00)00098-1.

Paefgen, J., Staake, T. and Fleisch, E. 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. DOI: 10.1016/j.tra.2013.11.010.

Pérez-Marín, A.M., Ayuso, M. and Guillen, M. 2019a. Do young insured drivers slow down after suffering an accident? Transportation Research Part F: Traffic Psychology and Behaviour 62: 690-99. DOI: 10.1016/j.trf.2019.02.021.

Pérez-Marín, A.M., Guillen, M., Alcañiz, M. and Bermúdez, L. 2019b. Quantile regression with telematics information to assess the risk of driving above the posted speed limit, Risks 2019, 7(3), 80. DOI: 10.3390/risks7030080.

Pérez-Marín, A.M. and Guillen, M. 2019. Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. Accident Analysis and Prevention 123: 99–106. DOI: 10.1016/j.aap.2018.11.005.

Plötz, P., Jakobsson, N. and Frances Sprei, S. (2017). On the distribution of individual daily driving distances. Transportation Research Part B 101, 213–227. DOI: 10.1016/j.trb.2017.04.008.

Uribe, J. and Guillen, M. (2020) *Quantile Regression for Cross-Sectional and Time Series Data:Applications in Energy Markets Using R*. SpringerBriefs in Finance. Springer. DOI: 10.1007/978-3-030-44504-1.

Wei, Y., Pere, A., Koenker, R. and He, S. (2006). Quantile regression methods for reference growth charts, Statistics in Medicine 25, 8, 1369-1382. DOI: 10.1002/sim.2271.

Vernon, D., Cook, L.J., Peterson, K.J., and Dean, J.M. 2004. Effect of the repeal of the national maximum speed limit law on occurrence of crashes, injury crashes, and fatal crashes on Utah highways. Accident Analysis and Prevention 36: 223–9. DOI: 10.1016/S0001-4575(02)00151-3.

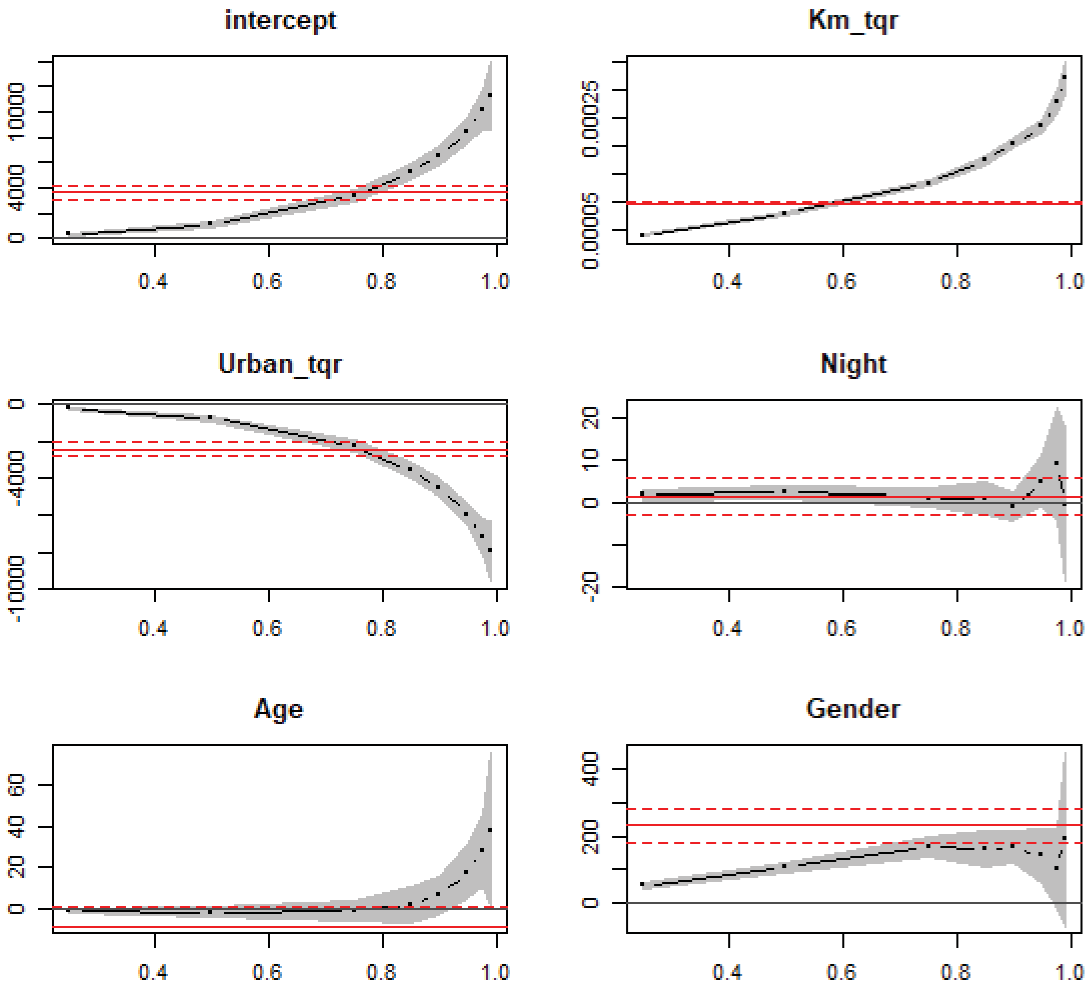
Viallon, V., and Laumon, B. 2013. Fractions of fatal crashes attributable to speeding: Evolution for the period 2001–2010 in France. Accident Analysis & Prevention, 52, 250-256. DOI: DOI: 10.1016/j.aap.2012.12.024.

Acknowledgements

The authors thank all reviewers by their comments. We sincerely think that the quality of the manuscript has improved with the suggestions and recommendation received by all reviewers. The authors also thank Fundación BBVA Big Data grants, ICREA Academia and research project PID2019-105986GB-C21, Ministerio de Ciencia, Innovación y Universidades.

Appendix

Figure A1. Parameter estimates of quantile regression for total kilometres driven above the speed limit at different levels of the quantile. Confidence intervals at a 5% level of significance are shown as shaded bands. The horizontal red line represents the corresponding parameter estimate in a classical linear regression model.


intercept
Km_tqr
Urban_tqr
Night
Age
Gender