Accident Analysis and Prevention 131 (2019) 131–136

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Using telematics data to find risky driver behaviour[☆]

Manda Winlaw[a,*], Stefan H. Steiner[a], R. Jock MacKay[a], Allaa R. Hilal[b]

[a] Department of Statistics and Actuarial Science, University of Waterloo, Canada
[b] Shopify, Canada

ARTICLE INFO

Keywords:
Pay-how-you-drive
Driving behaviour
Crash risk
Case-control study
Logistic regression

ABSTRACT

Usage-based insurance schemes provide new opportunities for insurers to accurately price and manage risk. These schemes have the potential to better identify risky drivers which not only allows insurance companies to better price their products but it allows drivers to modify their behaviour to make roads safer and driving more efficient. However, for Usage-based insurance products, we need to better understand how driver behaviours influence the risk of a crash or an insurance claim. In this article, we present our analysis of automotive telematics data from over 28 million trips. We use a case control methodology to study the relationship between crash drivers and crash-free drivers and introduce an innovative method for determining control (crash-free) drivers. We fit a logistic regression model to our data and found that speeding was the most important driver behaviour linking driver behaviour to crash risk.

1. Introduction

Traditional car insurance policies are based on driver demographics such as age, income, gender and location and on vehicle characteristics such as make, model and vehicle age (Weidner et al., 2017; Tselentis et al., 2017). Often annual distance driven is a factor in these policies, however, it is often an estimate provided by the driver and as noted in previous studies (White, 1976; Butler et al., 1988), reported values are usually lower than actual values. Unfortunately, this form of insurance has several drawbacks. Most notably, the cross subsidies phenomenon whereby safer drivers with fewer annual kilometres driven subsidize the insurance costs of more risky drivers with more annual kilometres driven (Tselentis et al., 2017). One of the implications of this policy is to increase social inequality since lower income individuals tend to drive fewer annual kilometres (Litman, 2002). In addition, these policies do not encourage drivers to modify their behaviour leading to more accidents, congestion and pollution.

Modern vehicles are equipped, or can be retrofitted, with a set of sensors that can infer information about a vehicle's state and its surrounding environment. The data collected from these sensors, known as telematics data, can be used to better assess a driver's risk and allow insurance companies to better individualize policies and mitigate risk. These policies are often broadly referred to as Usage-based insurance (UBI) schemes. Tselentis et al. (2017) provide a review of current UBI research. They divide UBI into two broad categories: Pay-As-You-Drive (PAYD) insurance and Pay-How-You-Drive (PHYD) insurance. PAYD insurance is based on what they refer to as travel behaviour, defined as a driver's strategic choices concerning which type of road to use, what time of day to drive and how much to drive. PHYD insurance, on the other hand, is based on what the authors refer to as driver behaviour. Driver behaviour is defined as a driver's operational choices in handling their vehicle and includes behaviour such as speeding, harsh braking or hard acceleration. Often times PHYD insurance is an extension of PAYD insurance and will include both driver and travel behaviour. In this paper, we focus on evaluating the impact of driver behaviour on the risk of a crash.

Much of the research on UBI policies has focused on travel behaviour (Tselentis et al., 2017) and in particular the relationship between distance and accident risk. The interest in distance is partially due to a number of non-telematics based studies showing a strong relationship between these two variates (Litman, 2005, 2011) and as mentioned previously the inaccurate reporting of distance driven by drivers. However, several recent studies have incorporated driver behaviour in their analysis. Ayuso et al. (2016) use survival analysis methods to show differences between the driving and travel behaviour of men and women. Jin et al. (2018) examine the impact of route familiarity in determining accident risk and incorporate both travel and driving behaviour variates in their analysis. Paefgen et al. (2013) compare different models for classifying drivers according to their accident risk using both travel and driver behaviours. They note that while neural networks perform best, logistic regression is best suited for use in actuarial models given its ease of interpretability. In Paefgen et al. (2014),

[☆] This work was supported by Mitacs through the Mitacs Accelerate program.
[*] Corresponding author.

https://doi.org/10.1016/j.aap.2019.06.003
Received 27 March 2019; Received in revised form 8 June 2019; Accepted 8 June 2019
Available online 25 June 2019

Paefgen, Staake and Fleisch examine the nonlinear relationship between mileage and accident risk while also incorporating driving behaviour in their model. Jun et al. (2011) focus on different measures of speed and their relationship to accident risk. They find that for different travel behaviours different measures of speed are relevant. Other studies have explored ways of classifying different drivers based on their driving behaviour without directly tying these behaviours to accidents (Weidner et al., 2017; Vaiana et al., 2014; Ferreira Júnior et al., 2017; Joubert et al., 2016). While these behaviours cannot be directly tied to accident risk they do highlight the heterogeneity apparent in driving styles.

In this paper, we are interested in studying the relationship between driving behaviour and the risk of a crash. Using a telematics dataset from a PHYD insurer we conduct a case-control study to compare driver behaviour between crash-involved drivers and crash-free drivers. We introduce a novel approach for selecting the controls. For each crash-involved driver, we select controls (i.e. crash-free drivers) so that their travel behaviour and location closely matches that of the crash-involved driver. The matching eliminates confounding due to travel behaviour and location and leads to more precise estimation of the effects of the driver behaviour factors on crash risk. Once we have our controls we use driver behaviour variates from these drivers along with our crash drivers and evaluate the impact of these variates using a logistic regression model with LASSO regularization. We find that speeding (defined relative to the speed limit) is the most important driver behaviour in our model.

The remainder of the paper is organized as follows. Section 2 gives a description of the data set including the trip-based travel and driver behaviour variates. Section 3 describes the case/control methodology we develop to reduce the impact of travel behaviour variates in determining the relationship between crash risk and driver behaviour. In Section 4 we include a description of the logistic regression model we employ to study the relationship between crash risk and driver behaviour as well as our results. Section 5 includes a discussion of the results and some of the limitations due to the data as well as concluding thoughts.

2. Data set

The data in our study is a sample from vehicles enrolled in a PHYD insurance program. There are 28,170,535 trips and 29,416 drivers. The trips occurred between March 2014 and April 2016, however, 99.94% of the trips occurred between July 2015 and April 2016. Within this time frame, drivers both enrolled and dropped-out of the program thus not all drivers have trips spanning the entire period. The median enrolment length is 9 months. Each trip is recorded using an OBD-II device (Comparing Mobile Apps, 2017), which collects Global Positioning System (GPS) data and speed data at a frequency of 1 Hz (i.e. collected once per second). Each GPS record includes latitude, longitude, HDOP (horizontal dilution of precision), a measure used to determine GPS accuracy, course over ground (COG) (i.e. the direction of the vehicle), and vehicle speed using the Doppler shift of satellite signals. The vehicle speed is also recorded directly from the vehicle and for all our analysis this is the measure of speed we use and not the GPS speed. The vehicle speed is a more robust measure and not subject to error due to drift, tunnels, etc. Note that each GPS and speed record in a trip has an associated timestamp. From the GPS and speed data we can determine a number of features about the trip. Sections 2.1 and 2.2 describe the trip-based travel behaviour and driver behaviour variates, respectively. These trip-based variates are then aggregated to build driver-based variates. The driver-based travel behaviour variates are used in our case/control methodology and the driver-based driver behaviour variates are used in our logistic regression model. Note that given the nature of the data there are sometimes anomalies in the trip data and we remove these trips from our analysis which reduces the number of trips to 28,104,042 and the number of drivers by one. As well, although we refer to the driver-based aggregates as belonging to one driver they in fact belong to one vehicle. The OBD-II device is attached to a particular vehicle and records all trips taken by that vehicle regardless of driver.

In addition to calculating the different variates, the trip data can also be used to detect crashes. In fact, the 28 crashes in our dataset are found using a proprietary trip-based algorithm. The algorithm is a function of extreme driver behaviour and in addition to relying on GPS and speed data, it is also based on accelerometer data, (i.e. acceleration in the x, y and z directions). Although the algorithm we use is proprietary, see Syedul Amin et al. (2014) for a description of how sensor data including accelerometer data can be used to detect a crash. In our analysis, accelerometer data is recorded by the OBD-II at 1 Hz and while used in the determination of crashes it is not used elsewhere in our analysis. Once the crashes have been detected using extreme values they are verified using location information and subsequent trip information including evidence of towing, vehicle change, etc. Given the nature of the algorithm and the limitations of the data frequency, it only detects high impact or severe crashes. An additional draw back of this methodology is that it does not provide at-fault information about the crash. Thus, for our crash drivers we do not know if the driver was at-fault. Once we have determined the crash-involved drivers we want to find control drivers with similar travel behaviour variates. Before describing the case/control methodology, we first examine the possible trip-based travel and driver behaviour variates. For a given driver, we can aggregate their trip-based variates in a number of ways to get driver-based variates. This then allows us to perform comparisons across drivers both to find controls and in determining the relationship between driver behaviour and the risk of a crash. Note that for potential control drivers we use the entire history of their trips to determine their driver aggregates however for crash-involved drivers we only use the trips prior to and not including their crash.

2.1. Travel behaviour variates

As mentioned previously, we define travel behaviour as a driver's choices concerning which type of road to use, what time of day to drive and how much to drive. Based on the speed and GPS data and their associated timestamps, we can determine several travel behaviour variates for each trip: the duration, the distance, the time of day and the day of the week. We can also use the location data to determine the road type. This information is gathered from the Open Source Routing Machine (OSRM) (OSRM, 2018). The OSRM is a routing engine that uses OpenStreetMap (OpenStreetMap, 2018) data to find road information from the GPS coordinates including road type and the posted speed limit, which we use in Section 2.2 to define a measure of driver behaviour. For each driver, we can use these trip-based travel behaviour variates to build driver-based travel behaviour variates and then use these aggregates to find our controls. We discuss the exact trip-based travel variates we use for aggregation in Section 3.

2.2. Driver behaviour variates

We are interested in studying the impact of a driver's operational choices in handling their vehicle on their crash risk. To do this, we define four measures of driver performance/behaviour for each trip. These four measures, which we refer to as penalties, are based on acceleration, braking, speeding and cornering. Note that while acceleration can often be used to describe both positive and negative acceleration (i.e. change in speed over time), for the purposes of our exposition, acceleration refers to non-negative acceleration while negative acceleration is referred to as braking. The penalties for each of the four categories are calculated using a similar methodology. To highlight this methodology, let us focus on the acceleration penalty and describe how it is determined. To calculate the acceleration penalty, we use the speed data collected from the trip to calculate an acceleration vector. Then using the acceleration vector we can map each value to a particular penalty and sum up all the penalties to get an acceleration penalty for the entire trip. Higher penalties imply worse driver behaviour. Note, however, that the penalty is normalized by the duration of the trip, thus a high acceleration value in a shorter trip will imply a larger penalty than in a longer trip. More formally, let a_t represent trip acceleration at time t where $t \in A$ and A is a subset of $\{1, 2, \ldots, n\}$ where n is the duration of the trip and A contains all timestamps where the change in speed is non-negative. Then the acceleration penalty is defined as

$$p_a = \frac{1}{n} \sum_{t \in A} f(a_t) \tag{1}$$

where f is a proprietary function. We can follow the same process to calculate the braking penalty. To calculate the speed penalty, we use the vehicle speed data and the speed limit data from OpenStreetMap (OpenStreetMap, 2018), to determine the amount of speeding at each point in time and then map each value to a penalty and aggregate for a trip-wise speed penalty. The cornering penalty is calculated in a similar manner to the previously described penalties using the COG to calculate a cornering vector. Note that for the acceleration, braking and cornering penalties, to account for different driving behaviour on different types of roads, a different penalty function is applied depending on if the driver is on a regular road or if they are on a ramp or roundabout. For the speed penalty, a different penalty function is applied for different speed limits. The penalty function f is different between speed limits 60 km/h and below and speed limits above 60 km/h. The four penalties we have just defined can be used as measures of driver behaviour and aggregated in various ways across all recorded trips for a given driver to define a driver's driving behaviour.

3. Case/control methodology

To investigate which driver behaviors are correlated with crashes we employ a case-control methodology. Case–control studies (Porta, 2008) are widely used in epidemiology to study rare diseases due to their relative cost-effectiveness. In a classic case–control study, we match each diseased individual with one or more non-diseased individuals (controls) who are otherwise similar (e.g. same age, gender, etc.). We then compare the cases and controls to identify exposure variates that are associated with the disease. Note that because case–-control studies are observational we need to be careful in drawing causal conclusions.

In our example, we define our cases as all 28 drivers that were identified as being involved in a crash. We then find control drivers (vehicles) for each case that closely matched the crash drivers in terms of the driver similarity measure as defined below.

To determine control drivers for each crash-involved driver we must first choose aggregate travel behaviour variates. From our dataset, there are many different ways to aggregate the trip-based travel features. We choose aggregates that are less sensitive to outliers. This includes the median number of trips per week, the median trip distance and the median trip duration. As well, we include the percentage of trips taken during the week (i.e. Monday–Friday) and the percentage of trips taken during the day (i.e. 6:00 am–5:59 pm) where a trip is categorized using the start time. In addition to the traditional travel behaviour variates, we are also interested in finding control drivers that are geographically close to our crash-involved drivers. To do this we use a driver's home and work locations. We do not know the actual home or work locations of each driver, so these locations are identified using a clustering algorithm on the trip end location values (i.e. latitude and longitude) for each trip in the dataset, for a particular driver. We assume the home location is the center of the largest cluster where the clusters are weighted according to garaging time which we define in Section 3.2. Similarly, we can use the same clustering algorithm to identify the work location. The work location is defined as the center of the second largest cluster. Section 3.2 includes a detailed description of the algorithm used to determine the home and work locations. Including the home and work locations, there are seven variates used to find control drivers, summarized in Table 1. Note that for each crash-involved driver there are 29,387 possible control drivers. However, we exclude some drivers from the set of possible control drivers if they do not have enough data. We exclude drivers from the potential control set if their enrolment length (the difference between the date of their first trip and their last) is less than 30 days. We also want to ensure that the drivers have a sufficient number of trips in the dataset, so we also impose a restriction on the minimum number of trips. Drivers must have at least 60 trips in their datasets. Given some anomalies in the data we also require that the distance to work is greater than 0 and less than 200 km. As well, we restrict the number of kilometres driven per year to be greater than 0, the median trip distance must be greater than 1 km and the median trip duration must be longer than 60 seconds. The drivers that do not satisfy these requirement are removed from the set of potential control drivers. This reduces the number of possible control drivers by approximately 3000 to 26,054. We next define a similarity measure using these variates to compare drivers.

Table 1
Driver travel variates.

Driver travel variates
Home location
Work location
Median number of trips per week
Median trip distance
Median trip duration
Percentage of weekday trips
Percentage of daytime trips

3.1. Driver similarity measure

One challenge with defining a similarity measure using the travel variates we have chosen is the inclusion of location data represented by latitude and longitude. Typically, when defining a distance measure we standardize the variates (e.g. for each variate subtract the mean and divide by the standard deviation) so they all have approximately the same importance in any distance/similarity measure. However, this is not possible for location data. In fact, when working with location data we must perform separate calculations for each crash driver. We must first calculate the distance between the crash-involved driver and all potential control drivers, then normalize, and finally define a distance measure on this vector. More formally, let

$$\mathbf{x}_c = ((x_{\mathrm{lat}}, x_{\mathrm{lon}})_{c1}, (x_{\mathrm{lat}}, x_{\mathrm{lon}})_{c2}, x_{c3}, x_{c4}, \ldots, x_{\mathrm{cl}}) = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, x_{i3}, \ldots, x_{\mathrm{il}}) \tag{2}$$

be the feature vector of a crash driver where $l = 7$ is the number of features and $\mathbf{x}_{c1} = (x_{\mathrm{lat}}, x_{\mathrm{lon}})_{c1}$ represents the home location and $\mathbf{x}_{c2} = (x_{\mathrm{lat}}, x_{\mathrm{lon}})_{c2}$ represents the work location. Let

$$\mathbf{y}_i = ((y_{\mathrm{lat}}, y_{\mathrm{lon}})_{i1}, (y_{\mathrm{lat}}, y_{\mathrm{lon}})_{i2}, y_{i3}, y_{i4}, \ldots, y_{\mathrm{il}}) = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, y_{i3}, \ldots, y_{\mathrm{il}}) \tag{3}$$

represent the feature vector for a potential control driver where $i = 1, \ldots, n$, and n is the number of potential control drivers. For each potential control driver, $i = 1, \ldots, n$, we calculate the following distance vector $\mathbf{d}_i$,

$$\begin{aligned} \mathbf{d}_i &= \mathbf{x}_c - \mathbf{y}_i \\ &= (d_h(\mathbf{x}_{c1}, \mathbf{y}_{i1}), d_h(\mathbf{x}_{c2}, \mathbf{y}_{i2}), |x_{c3} - y_{i3}|, \ldots, |x_{\mathrm{cl}} - y_{\mathrm{il}}|) \\ &= (d_{i1}, d_{i2}, \ldots, d_{\mathrm{il}}), \end{aligned} \tag{4}$$

where d_h represents the haversine distance function (Van Brummelen, 2013) and $d_{ij} \geq 0 \; \forall \, i, j$. Note that the distances across features may have widely different ranges. Suppose, we ignore this problem, then we could choose control drivers using the distance vector $\mathbf{d}_i$. We would do this by finding the drivers that have the smallest overall "distance" to the crash driver in terms of $\mathbf{d}_i$ (i.e. we find the drivers that are the most similar to the crash driver). To do this we define a similarity measure as follows:

$$s_i = \sum_{k=1}^{l} d_{\text{ik}}, \tag{5}$$

where $s_i \geq 0$ and a small similarity value (i.e. a value close to 0) suggests a close relationship between the control driver and the crash driver. We can use the similarity measure to choose control drivers. For example, suppose we want 5 control drivers, then we choose as control drivers, the potential control drivers with the 5 smallest similarity values. As mentioned above, no standardization of the features has occurred. Thus, for a given $i = 1, \ldots, n$, each distance value, d_{ij}, $j = 1, \ldots, l$, in the similarity function, s_i, may have wildly different ranges and thus different features may have varying importance in determining the control drivers. Since we want the different features to have approximately the same weight we need to standardize the distance values of each feature so they are on approximately the same scale. We propose using a multivariate approach. Let D be the matrix composed of distance vectors (i.e. $D = [\mathbf{d}_1; \mathbf{d}_2; \ldots; \mathbf{d}_n] \in \mathbb{R}^{n \times l}$) then,

$$D^* = (D - \mathbf{1}\boldsymbol{\mu}^T)\Sigma^{-1/2}, \tag{6}$$

where $\Sigma \in \mathbb{R}^{l \times l}$ is the covariance matrix of D, $\boldsymbol{\mu} \in \mathbb{R}^l$ is the mean of D and $\mathbf{1} \in \mathbb{R}^n$ is a column of all ones. Let $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{il})$ be the ith row of D and $\mathbf{d}_i^*$ be the ith row of D^*. Thus, $\mathbf{d}_i^*$ represents the standardized distance vector for potential control driver i, $i = 1, \ldots, n$, where the distances are now on approximately the same scale. Using this standardized distance vector we can now calculate a similarity metric for each potential control driver as follows:

$$s_i^* = \sum d_{\text{ik}}^*, \tag{7}$$

and choose as control drivers the drivers with the smallest standardized similarity values. Note the similarity value is not longer restricted to nonnegative values. The same procedure is repeated for each of the crash drivers to choose control drivers for each of the crash drivers.

In calculating the matrix D^* from D we include all of the potential control drivers. However, the calculation of the mean and covariance matrix may be affected by outliers which in turn will affect the normalization. To reduce this effect we perform some additional filtering on potential controls before we normalized the data (i.e. we reduce the size of D before transforming to D^*). We remove potential controls from D where the distance between their home and the crash-invovled driver home is greater than 40 km. We also remove all potential control drivers with a distance between work locations (potential control vs. crash-involved) greater than 50 km. To further reduce the effect of outliers we also remove all potential control drivers with any one of their feature values in the top 25% of the distribution for that feature. Once we have reduced the size of D we then normalize the features, calculate the similarity value and choose the 20 drivers with the smallest similarity values as controls or if there are less than 20 drivers after filtering we choose all of the drivers as control drivers. In fact, only 1 driver had fewer than 20 controls due to filtering. Thus, the data set we used in our logistic regression model consists of 576 drivers, including crash-involved drivers.

The choice of 20 controls is somewhat arbitrary. In typical case/control studies, the number of controls per case is often up to 5. In our example, we selected 20 controls for each case (crash driver) because the cost and effort associated with finding and assessing each control driver was negligible. We avoided selecting even more cases because we wanted to ensure that the controls were similar to the cases.

3.2. Home and work determination

In determining control drivers for each crash driver we are interested in finding drivers who drive in the same geographical area. To capture this similarity we can use the home and work locations of each driver. Unfortunately, we do not have the actual home and work locations for each driver but as mentioned previously we used a clustering algorithm to find them. Using the end trip locations (latitude, longitude) as inputs we use the DBSCAN clustering algorithm (Ester et al., 1996) to cluster a driver's end trip locations. DBSCAN is a clustering algorithm designed to cluster together points that are geographically close to one another. We use the end trip location instead of start trip location since the GPS reading is more accurate at the end of the trip vs. the beginning of the trip. Once we have the location clusters we can then label one cluster as home and another as work. Our labeling is based on garaging time. The garaging time of a particular location is the amount of time spent at that location. We determine the garaging time of a particular location (i.e. trip end point) by looking at the time difference between consecutive trips (i.e. the garaging time of location x is the time difference between the trip that ends at x and the next trip that starts at x). For each point in a cluster we calculate the garaging time and the garaging time of the cluster is simply the sum of all the cluster point garaging times. For each driver, the home location is the defined as the center of the cluster with the largest garaging time and the work location is defined as the center of the cluster with the second largest garaging time. Note, that we do not in fact care if these points are a driver's actual home and work. We are more interested in using them as proxies for where the drivers are driving. However, we do want to avoid capturing two home locations (i.e. the driver moved, so we have two home locations with a lot of garaging time). So, to avoid this we only use trips from the most recent 30 days of their history. As well, we exclude garaging times greater than a day. This is to avoid assigning high garaging time values to places like airports.

4. Risk analysis

4.1. Model data set

Given the 576 case and control drivers described in Section 3.1 we want to fit a model using these drivers and their associated driver behaviour to identify risky driving behaviours. However, we must first define aggregate driver behaviour variates for each driver. As described in Section 2.2 the trip-level driver behaviour variates are four penalties: acceleration, braking, speeding and cornering. For each of these four penalties we calculate the mean, median, standard deviation, interquartile range (difference between the 75th and 25th percentiles) and the 90th percentile value for each driver using their trip-based variates. As with the aggregate travel behaviour variates, we calculate the aggregate travel behaviour variates using all available trips for non-crash drivers and all trips prior to the crash for crash drivers. We include the 90th percentile penalty value as one of our aggregates since a large penalty indicates very undesirable behaviour, and thus the 90th percentile penalty value captures information about a driver's most undesirable driving behaviour. In addition, since we worried about the effect of many short trips, we also calculated what we refer to as the overall penalty. To determine this aggregate we assumed all the individual historical trips were combined into a single long trip. Thus, the overall penalty is determined over the total elapsed time rather than the average by trip. In total, for each driver there are 24 travel behaviour variates (4 penalties with 6 aggregates each).

4.2. Logistic regression with LASSO regularization

For our analysis, we fit a logistic mixed effect regression model with the response variate

$$Y = \begin{cases} 1 & \text{for crash driver (case)} \\ 0 & \text{otherwise (for all controls).} \end{cases}$$

We considered the 24 fixed effects explanatory variates as described above, denoted as $x_{speed\_m\ ean}, \ldots, x_{cornering\_o\ verall}$. We also included a single random effect, denoted R, defined by each crash driver and their associated control drivers. The model with $P(Y_i = 1) = p_i$ is thus

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{\mathrm{speed\_m\ ean}} + \cdots + \beta_{24} x_{\mathrm{cornering\_o\ verall}} + R_i, \tag{8}$$

where $R_i \sim G(0, \sigma_R)$ represents the random effect for the group.

To fit the model in (8) we used the glmmLasso function in R (https://cran.r-project.org/web/packages/glmmLasso/index.html). This function allowed us to fit the mixed effect logistic regression model with LASSO regularization (Tibshirani, 1994). With the LASSO approach, we estimate the parameters by maximizing $L + \lambda\Sigma_i|\beta_i|$, where L is the likelihood and λ is a regularization parameter. We added the LASSO regularization to help us identify the important explanatory variates and build a parsimonious model (since it tends to force coefficients to zero).

To find a recommended model we used 5 fold cross validation to choose the best value for the LASSO penalty parameter. Doing this with the available data and looking at $\lambda = (0, 1, 2, \ldots, 20)$ we found that $\lambda = 17$ provided the model with the smallest sum of the deviance residuals. The model with $\lambda = 17$ has only two fixed effect explanatory variates and is summarized in Table 2 (where we suppress all the fixed effect terms that had a coefficient value of zero).

In the model from Table 2, the variate $x_{accel\_s\ td}$ is not statistically significant at the 5% level, so we drop it from the model. The model with just the variate $x_{speed\_o\ verall}$ is shown in Table 3.

Note that the random effect term for the group is non-zero. However, it is not clear if the random effect is important or not. To assess this we compare the model from Table 3 with another model without the random effect using AIC (Akaike, 1974).

We found that the model without the random effect had AIC = 217.9, while the same model with the random effect had AIC = 221.2. Thus, since smaller AIC is better, the best model appears to be one without the random effect. Fitting this final simple model with only a single fixed effect we get the results shown in Table 4.

4.2.1. Interpretation of the final model

Since the drivers included in the regression analysis were chosen using the case/control method we cannot interpret the model intercept nor can we make a prediction about how likely it is that a particular driver will be in a crash. However, we can interpret the coefficient for $x_{speed\_o\ verall}$ in terms of how much changes in this explanatory variate changes the risk of a crash.

In general, we can compare the odds ratio for two different values of $x_{speed\_o\ verall}$. We have

$$\begin{aligned}\frac{p_1/(1-p_1)}{p_2/(1-p_2)} &= \frac{\exp(-3.77 + 1.09 x_{\mathrm{speed\_o\ verall},1})}{\exp(-3.77 + 1.09 x_{\mathrm{speed\_o\ verall},2})} \\ &= \exp(1.09(x_{\mathrm{speed\_o\ verall},1} - x_{\mathrm{speed\_o\ verall},2}))\end{aligned} \tag{9}$$

In this case, because we believe the risk of a crash should be close to zero, $\frac{p_1/(1-p_1)}{p_2/(1-p_2)} \approx \frac{p_1}{p_2}$. To interpret this we look at the distribution of values for the overall speed penalty $x_{speed\_o\ verall}$ across the 29,381 drivers in our data set. The distribution of $x_{speed\_o\ verall}$ across the population of drivers is numerically summarized in Table 5.

Table 2
LASSO model fit.

Fixed effects	Coefficients			
	Estimate	StdErr	z.value	p.value
(Intercept)	−4.03383	0.22857	−17.6482	< 2.2e−16
speed_overall	1.13240	0.34729	3.2607	0.001111
accel_std	7.81802	4.53851	1.7226	0.084962
Random effects (StdDev):				
Group 0.2642461				

Table 3
LASSO model fit (excluding $x_{accel\_s\ td}$).

Fixed effects	Coefficients			
	Estimate	StdErr	z value	p value
(Intercept)	−3.79177	0.22302	−17.0019	< 2.2e−16
speed_overall	1.13331	0.33835	3.3495	0.0008095
Random effects (StdDev)				
Group 0.2672874				

Table 4
Final model.

Fixed effects	Coefficients			
	Estimate	StdErr	z Value	p Value
(Intercept)	−3.76848	0.21717	−17.3527	< 2.2e−16
speed_overall	1.09341	0.33172	3.2962	0.0009799

Table 5
Population summary statistics for $x_{speed\_o\ verall}$.

Mean	0.575390
Std	0.485216
Min	0.000000
25%	0.200564
50%	0.441413
75%	0.826429
Max	3.254108

The difference between the 25th percentile and 50th percentile of $x_{speed\_o\ verall}$ is 0.44–0.20 = 0.24. Using our model results, this difference of 0.24 corresponds to approximately a 30% increase in the risk of a crash, while similarly going from the 50th percentile driver to the 75th percentile driver corresponds to roughly a 53% increase in the risk of a crash. These relatively large differences are important for automotive insurers to consider when building or refining Usage-based insurance programs.

5. Discussion and conclusions

Using automotive telematics data from over 28 million trips, we showed that the overall speeding penalty was the only significant driver behaviour variate linked to the risk of a crash. The effect was large with a 75th percentile driver having over a 50% greater chance of a crash than a 50th percentile driver. However, given the small number of identified crashes (28), we suggest these results need validation using other data.

In addition to the need for validation, we could extend or improve the approach we followed in a number of ways. Through the direct engagement with an insurance company, we could have avoided the use of the crash detection algorithm and instead linked directly to insurance claims to find crashes. This would have also provided at-fault information about the crashes. In our crash detection approach, we were conservative to avoid falsely identifying accidents. With insurance claims information we would have been able to include more crashes and this would have increased the precision of the results. Alternatively, we could have included probable crashes in our analysis (rather than only very likely crashes). This would have also increased the number of available crashes at the risk of contaminating the results if the probable crashes were not actual crashes. We could have tried to consider this risk in the analysis by weighting observations using our confidence that they represent actual crashes.

In our analysis, we used the acceleration, braking, cornering and speeding penalties as previously defined, i.e. with a predefined function in (1). It may be possible to use the raw telematics data directly in the analysis. In this way, we could optimize the penalty functions to improve their connection with the risk of a crash.

We believe another way to potentially improve the analysis is incorporate weather information. Clearly aggressive driving is riskier in poor weather conditions. A challenge here is obtaining precise local weather and road surface condition information that translates well to poor driving conditions.

Finally, it would be interesting to see if we could duplicate these results with sample data obtained from embedded automotive telematics and mobile (rather than OBD) telematics data. Mobile phone based telematics data is less expensive to collect, but introduces additional measurement and data quality challenges.

Conflict of interest

None.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (December (6)), 716–723.

Ayuso, M., Guillén, M., Pérez Marín, A.M., 2016. Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. Transport. Res. Part C: Emerg. Technol. 68, 160–167.

Butler, P.M., Butler, T., Williams, L.L., 1988. Sex-divided mileage, accident, and insurance cost data show that auto insurers overcharge most women. J. Insur. Regul. 6, 243–282.

Comparing Mobile Apps, 2017. Hybrid bluetooth and OBD-II: data collection options. technical report. Intell. Mechatron. Syst.

Ester, M., Kriegel, H.-P., Sander, Jörg, Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD.

Ferreira Júnior, J., Carvalho, E., Ferreira, B.V., de Souza, C., Suhara, Y., Pentland, A., Pessin, G., 2017. Driver behavior profiling: an investigation with different smartphone sensors and machine learning. PLOS ONE 12 (4), 1–16.

Jin, W., Deng, Y., Jiang, H., Xie, Q., Shen, W., Han, W., 2018. Latent class analysis of accident risks in usage-based insurance: evidence from Beijing. Accid. Anal. Prev. 115, 79–88.

Joubert, J.W., de Beer, D., de Koker, N., 2016. Combining accelerometer data and contextual variables to evaluate the risk of driver behaviour. Transport. Res. Part F: Traffic Psychol. Behav. 41, 80–96.

Jun, J., Guensler, R., Ogle, J., 2011. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: application of in-vehicle monitoring technology. Transport. Res. Part C: Emerg. Technol. 19 (4), 569–578.

Litman, T., 2002. Evaluating transportation equity. World Transp. Policy Pract. 8 (2), 50–65.

Litman, T., 2005. Pay-as-you-drive pricing and insurance regulatory objectives. J. Insur. Regul. 23.

Litman, T., 2011. Distance-Based Vehicle Insurance Feasibility, Costs and Benefits. Technical Report. Victoria Transport Policy Institute.

OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/Main_Page (accessed 12.04.18).

OSRM. http://project-osrm.org (accessed 12.04.18).

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

Paefgen, J., Staake, T., Fleisch, E., 2014. Multivariate exposure modelling of accident risk: insights from pay-as-you-drive insurance data. Transport. Res. Part A: Policy Pract. 61, 27–40.

Porta, M. (Ed.), 2008. A Dictionary of Epidemiology, 5th ed. Oxford University Press, New York.

Syedul Amin, M., Mamun Ibne Reaz, M., Sobhan Bhuiyan, M.A., Sheikh Nasir, S., 2014. Kalman filtered GPS accelerometer based accident detection and location system: a low-cost approach. Curr. Sci. 106 (11), 1548–1554.

Tibshirani, R., 1994. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. 58 (Series B), 267–288.

Tselentis, D.I., Yannis, G., Vlahogianni, E.I., 2017. Innovative motor insurance schemes: a review of current practices and emerging challenges. Accid. Anal. Prev. 98, 139–148.

Vaiana, R., Iuele, T., Astarita, V., Caruso, M.V., Tassitani, A., Zaffino, C., Giofré, V., 2014. Driving behavior and traffic safety: an acceleration-based safety evaluation procedure for smartphones. Mod. Appl. Sci. 8 (1), 88–96.

Van Brummelen, G., 2013. Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry. Princeton University Press.

Weidner, W., Transchel, F.W.G., Weidner, R., 2017. Telematic driving profile classification in car insurance pricing. Ann. Actuar. Sci. 11 (2), 213–236.

White, S.B., 1976. On the use of annual vehicle miles of travel estimates from vehicle owners. Accid. Anal. Prev. 8 (4), 257–261.