Usage-based vehicle insurance: Driving style factors of accident probability and severity

Ivan Stankevich\textsuperscript{a,b}, Konstantin Korishchenko\textsuperscript{c}, Nikolay Pilnik\textsuperscript{a,b,d}, and Daria Petrova\textsuperscript{e}

\textsuperscript{a}Department of Applied Economics, HSE University, Moskva, Russian Federation; \textsuperscript{b}Lebedev Physical Institute of the Russian Academy of Sciences, Moskva, Russian Federation; \textsuperscript{c}Russian Presidential Academy of National Economy and Public Administration, Moskva, Russian Federation; \textsuperscript{d}Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moskva, Russian Federation; \textsuperscript{e}Meta Telematics, Moscow, Russian Federation

ABSTRACT
This paper introduces an approach to telematics device data application in automotive insurance. We conduct a comparative analysis of different types of devices that collect information on vehicle utilization and driving style of its driver and describe advantages and disadvantages of these devices. The possible formats of telematics data are described and methods of their processing to a format convenient for modeling are proposed. We also introduce an approach to classify the accidents’ strength. Using all the available information, we estimate accident probability models for different types of accidents and identify an optimal set of factors for each of the models. We assess the quality of resulting models using both in-sample and out-of-sample estimates.

KEYWORDS
Usage-based insurance; driving style; road accidents; accident prediction

1. Introduction
The Internet of Things is currently rising not only in the consumer sphere but also in application to different business processes. Combined with modern data analysis and data mining techniques, it can significantly enhance the performance of various business automation and decision support systems, including, among other applications, the use of online individual-level data in insurance (usage-based insurance [UBI]). Our study is focused on one of these applications: the use of telematics data in vehicle insurance.

At the moment, the general approach to vehicle insurance is the use of commonly available factors, such as vehicle characteristics and general information about the driver (e.g., age, gender, driving experience, accident history). One of the ways to enrich this data is to add driving style
indicators that cannot be measured directly but that can be estimated using the data on how and when the driver uses his or her car.

The application of telematics data to vehicle insurance are presented, for example, in Husnjak, Peraković, Forenbacher, and Mumdziev (2015), but a smaller set of factors is used here than the one proposed in our paper. A more thorough analysis is performed in Baecke and Bocca (2017), with a considerably wide set of models (including not only econometrics models but also machine learning approaches) and accuracy comparison of models with different sets of indicators. Bian, Yang, Zhao, and Liang (2018) provide a more descriptive overview of usage-based systems in vehicle insurance.

Another very close and significantly better-researched problem is accident analysis. The current research in accident analysis and modeling is mostly centered on road- or city-level data, where aspects of road design and traffic organization are discussed. Another significant research area uses accident-level data, where accident participants’ characteristics as well as some external factors, such as weather conditions, are considered. The first group is represented, for example, by Quddus (2008), which uses time series econometrics models to predict the number of accidents in London and Great Britain; Anderson (2007), Mohaymany et al. (2013), and Bîl, Andrașik, and Janoșka (2013) identify the spots with high accident probability on road maps. Erdogan, Yilmaz, Baybura, and Gullu (2008) consider climatic, geographic, and day-of-week factors. Olszewski, Szagała, Wolański, and Zielińska (2015) also discuss the influence of time period, area, road types, and lighting conditions on the occurrence of accidents with pedestrians.

These studies give some insights into external factors influencing road accidents, such as weather conditions and road design, but they do not give any information on how driver-related factors influence accident probability. Studies based on accident-level data are more fruitful in this sense. Tesema, Abraham, and Grosan (2005) analyze the influence of driver (age, gender, driving experience) and vehicle characteristics (age and type) as well as common external factors, such as weather and lighting conditions on accident severity. A similar approach is adopted in Kashani and Mohaymany (2011) and Kashani, Shariat-Mohaymany, and Ranjbari (2012).

Studies discussing the influence of personal characteristics and driving behavior are scarcer and are mostly based on survey data. Among the papers that analyze driving style factors we can cite Bagdadi and Várhelyi (2011), who study the relationship of jerky driving with self-reported accidents, and Mousavi, Zhang, Parr, Pande, and Wolshon (2019), who consider vehicle movement on certain highways. A review study by Sagberg, Selpi, Piccinini, and Engström (2015) cites only three papers that investigate jerky driving and four papers in which driving style is measured with
lateral acceleration behavior. Hill, Horswill, Whiting, and Watson (2019) analyze the relationship between heavy braking and hazard perception (that can be linked with accident probability). Naude et al. (2019) provide a thorough analysis of acceleration data in relation to incidents (critical driving situations), not registered accidents as in our case.

There are some psychological studies on the connection between character traits and driving behavior. For example, French, West, Elander, and Wilding (1993) analyze the relationship between decision-making characteristics of a person and his or her driving behavior (both are based on specialized questionnaires) using a sample of 711 persons of different ages and genders. Elliott, Armitage, and Baughan (2003) study the influence of sociodemographic characteristics and personal traits on the intention to violate the speed limits. Taubman-Ben-Ari, Mikulincer, and Gillath (2004) use specialized questionnaires to construct driving style indicators and study the relationship between these indicators and personality traits. But the data used in these studies are quite aggregated, are based on subjective indicators, and may correspond poorly with actual driving behavior.

More detailed data partly similar to what we use are considered in Park, Kim, Kho, and Park (2017). Some indicators of driver’s behavior, such as alcohol use, speeding, and signal violation are used to model the severity of an accident. Kosuge, Okamura, Kihira, Nakano, and Fujita (2017) predict the crash involvement of elderly drivers using driver characteristics based on survey data.

In contrast to the most of the existing research, we:

1. Use a combined set of driving style factors that can characterize jerky driving: frequencies of rapid accelerations, heavy braking, and sharp turns.
2. Utilize the data provided by telematics data operator, so the data set is much larger and the accident data are obtained from insurance companies (no self-reporting leads to high reliability).
3. We estimate separate models for accidents of different severity.

The structure of the rest of the paper is as follows. Section 2 provides an overview of UBI systems and describes the model of telematics devices application in insurance industry. We describe the possible options for the target functions of the insurance company depending on the need for customer feedback. We also show that accounting for accidents can have a multilevel structure, which takes into account not only the fact of the accident but also the degree of its severity. Section 3 describes the data collected by telematics devices and proposes an approach to calculation of a limited and interpretable set of indicators that can be used for scoring
based on the original raw data. Section 4 presents some modeling results of car accident probability based on these indicators.

**Usage-based insurance and telematics**

Based on PTOLEMUS research (PTOLEMUS Consulting Group, 2016), UBI is becoming mainstream in auto insurance worldwide, especially in the US, Italy, UK, and so on. The US will become the leading UBI market in the world. In Europe, growth will be driven by Italy, but the UK, Germany, and France will see UBI subscriptions take off in the next 5 years. New major markets will emerge, including China and Russia. By 2020, nearly 100 million vehicles globally will be insured with telematics policies. This will grow to nearly 50% of the world’s vehicles by 2030, generating more than €250 billion in premiums for insurers. In the US and Europe, most carmakers will have adopted UBI by 2020. The most successful model will use a central data hub provided by telematics service providers connected to insurance companies. On-Board Diagnostics (OBD) dongles will become the leading UBI device, reaching all the continents. Aftermarket black boxes will continue to grow, specifically in high-premium markets and for high-value cars. Smartphone-based Pay How You Drive (PHYD) programs are set to become one of the key devices to collect driving data, with or without a Bluetooth beacon. However, they will not replace dongles or black boxes, which will still make up the majority of devices used in 2020.

Telematics data can be generated from a number of different sources:

- The car network, in particular the CAN-bus, part of which is accessed by a dongle through the OBD port
- Sensors embedded in the vehicle, generally including motion and location sensors but also more commonly magnetometers (which can help determine the vehicle direction)
- Mobile devices, notably smartphones, which include all the sensors but are not physically linked to the vehicle.

Telematics devices, notably black boxes, are often fitted with one or sometimes two accelerometers (generally with three axes). While fixed-device measurements are generally more reliable, they have now to compete with smartphone-based sensors that includes magnetometers (compasses) and microphones that can help detect the type of road, occupancy, and the type of car driven.

For the UBI service provider, the choice of data is great but these data are only useful until the most predictive criteria are identified. The data set selected will therefore change with experience and the understanding of
what really identifies risky behavior and which behavior is most predictive of losses.

Nevertheless, the availability of telematics data does not mean that it will be used in insurance companies automatically. The aim of this paper is to propose an approach of estimation the riskiness of client using the telematics data. As we will see later, it is impossible to obtain a model of sufficient accuracy without the use of specialized data. In our case, these are data on the use of accelerations of various types by the driver.

3. Methodology and data description

3.1. Types of telematics devices

UBI On-Board Units (OBUs) collect telematics data using GNSS positioning (GPS/GLONASS, etc.) and communicate to a server via GSM/GPRS (for beacons via SMS) modem with embedded global M2M e-SIM (2G/3G). The server receiver gets and parses (MXP, XML, JSON) raw data by User Datagram Protocol (UDP), providing OBU diagnostics and managing and updating device firmware (Over-the-Air (OTA)). Then filtered and aggregated telematics data are used in web/mobile products for business matters.

When the ignition is ON, the device constantly fixes the position and sends the data to the server with frequency depending on the driver’s behavior. If the vehicle speed and acceleration levels are steady, the lowest possible frequency is 1 data package per 5 seconds. If the driver goes over the set acceleration threshold, the device starts to analyze the quantity of threshold excesses and transmit them to the server with related positions and speed. In the case of a crash, the device collects all the accelerations and speed and positions every 100 ms.

These data sets can represent around 2–45 Mb of data per month per customer, depending on the list of indicators collected, the frequency, and the number and length of annual trips. Properly collecting the relevant data is crucial to obtain the accurate scoring.

The telematics-enabled motor insurer has access to:

- Vehicle data, such as mileage, driving times, driver behaviors, thefts, accidents, and types of roads.
- Driver data, if there is only a single driver insured; experienced insurers will also be able to identify a new driver based on his or her driving style.
- Environment data, notably the type of roads (motorways, urban roads, etc.), traffic jams, weather, urban/rural environment, dangerous spots.
3.2. Original data format

Originally, the data are received as a set of data packages from the telematics device that can contain different indicators. Some of them provide purely technical information about the device functioning, such as voltage and different test results. More informative packages contain information about the ignition being turned on and off, vehicle coordinates, speed, and accelerations.

Packages are sent not on regular time intervals but when some predefined events take place. This approach seems natural for ignition-related events. Acceleration-related events take place when the acceleration of one of the axes is higher than some threshold value in order to save only the most informative data. Speed-related events take place when the vehicle drives with speed above some threshold. This approach allows for more compact yet information-preserving data storage, but it makes the data processing more complicated because of the need to account for differing data frequency. The number of packages obtained from a vehicle with the current settings is around several thousand per day.

Time-aggregated data are more convenient for the purposes of driving behavior description. In this case, all the package data are turned to data of certain frequency (hourly, daily, weekly, monthly, and so on), and indicators used are turned to mean of cumulative values during the corresponding time period. Our research shows that aggregation to the hourly data is the most convenient. So, instead of data on separate events we use cumulative indicators, such as total mileage during the hour, mean speed and mean frequencies of acceleration of different intensities.

This approach allows us to move from several thousand observations per car per day to 24 observations per day (see Figure 1 for sample data). As the vehicle is typically not used 24 hours per day, some observations are zero (Figure 2). The situation also changes because of morning and evening rush hour traffic congestion and the effect of weekends and holidays.

3.3. Data description

The data set employed in our study consists of 5,050 observations, and each observation contains time-averaged data for a specific driver for the entire observation period. So, we have information about 5,050 unique drivers, most of which are available for more than a year.
3.4. Data aggregation and feature engineering

On the one hand, the obtained data set describes the driver’s behavior during the day in a detailed and very informative way. On the other hand, this information is not structured enough to be used in scoring algorithms. To solve this contradiction, we need to determine the format of data aggregation. In our case, we need to make two transformations. First, we move from hourly intervals to longer time intervals. Most modern insurance companies sell annual insurance contracts. Therefore, by default, they are most interested in long-term customer behavior. In this work, all data on drivers are translated into averaged values over the entire observation period for each driver. This approach allows comparison of drivers who are observed for different amounts of time. However, because the data themselves are limited to a maximum depth of several years, there are no significant changes in the long-term driving style during this time.
Nevertheless, it is important to note that recently (for example, against the background of the development of car sharing and short-term insurance), interest in telematics data with a high frequency has been growing. The data we use can be presented in a weekly format and even at the level of individual days and trips. Second, we introduce a wider range of indicators that characterize the mileage, speed, and accelerations to preserve the information on driving style differences between business days and weekends and different times of day. The actual set of indicators is determined by the data available from telematics devices (acceleration levels at different axes and GPS coordinates) and preprocessing (data filtration and aggregation) conducted in data provider computer system.

Mileage indicators (Table 1), even though specific for each driver, mostly characterize the utilization of the vehicle, not the driving style itself. In the medium-run perspective (several months to several years), a driver usually cannot influence the distance between the main points he or she is located during the day (mostly home and work). Hence, there is always some minimum daily mileage. The situation is the same with the timing of his or her mileage that is determined by working hours. Both mileage and its structure do influence accident probability, but one should understand the mechanism behind this influence.

Figures 3 and 4 show the dynamics of weekly mileage and average morning rush hour weekly mileage averaged for all cars in the data set. We can clearly see the seasonal wave with summer period and New Year week (official a holiday week in Russia).

Speed indicators (Table 2), especially average speed indicators, also characterize utilization of the vehicle and external conditions of her trips rather than the driving style. Mean speed and mean speed at different times of

| Indicator       | Description                                                      |
|-----------------|------------------------------------------------------------------|
| mileage         | Total mileage during the observation period (in km)              |
| trips_day       | Average number of trips per day                                  |
| below_10_pr     | Share of trips shorter than 10 km                                |
| below_30_pr     | Share of trips shorter than 30 km                                |
| over_200        | Share of trips longer than 200 km                                |
| over_400        | Share of trips longer than 400 km                                |
| d_total_m       | Average mileage                                                  |
| avg_trip_mil    | Average trip mileage                                             |
| avg_trip_dur    | Average trip duration                                            |
| d_business_m    | Average mileage on business days                                 |
| d_day_m         | Average daytime mileage (7 am–7 pm)                             |
| d_evening_jam_m| Average evening rush hour mileage (6 pm–8 pm)                   |
| d_morning_jam_m| Average morning rush hour mileage (8 am–10 am)                  |
| d_holi_m        | Average holiday mileage                                          |
| d_night_m       | Average nighttime mileage (0 am–6 am)                            |
| day_m_pr        | Ratio of daytime mileage to total mileage                        |
| ej_m_pr         | Ratio of evening rush hour mileage to total mileage             |
Table 2. Speed indicators.

| Indicator       | Description                                      |
|-----------------|--------------------------------------------------|
| avg_sp          | Average speed                                    |
| max_sp          | Maximum speed (over the whole observation period)|
| max_ej_sp       | Maximum speed during evening rush hour           |
| max.mj_sp       | Maximum speed during morning rush hour           |
| max_n_sp        | Maximum night speed                              |
| m_pr_below_20   | Share of mileage with speed lower than 20 kph     |
| m_pr_below_60   | Share of mileage with speed lower than 60 kph     |
| m_pr_over_100   | Share of mileage with speed over 100 kph          |
| m_pr_over_130   | Share of mileage with speed over 130 kph          |

Figure 3. Average weekly mileage, km. Year and week number on horizontal axis.

Figure 4. Average weekly morning rush hour mileage, km. Year and week number on horizontal axis.
day and weekdays (such as mean speed at night, mean speed at weekends, etc) is mostly determined not by the driving style but by the situation on roads and traffic congestions: mean speed is lower during the rush hour and at business days. At certain moments of time, the driver can choose less congested roads, but generally even in this case he or she is restricted by the characteristics of place in which one lives. Road congestion influences accident probability, but it is not determined by the driver’s behavior.

Figures 5 and 6 show the dynamics of average speed and average evening rush hour speed averages for all the cars in the data set. The seasonal wave is also clear here. The accelerations data are more complicated: Simple calculation of the number of accelerations above a certain threshold does not contain any information about direction and strength of these accelerations. To preserve this information in the final data set, we consider the following indicators separately:

- Accelerations (the positive component of acceleration along the longitudinal axis)
- Decelerations (the negative component of acceleration along the longitudinal axis)
- Side accelerations (absolute value of acceleration along the transverse axis)

Acceleration, deceleration, and side acceleration indicators are indicators of jerky driving, often studied in accident modeling literature (see, e.g., Naude et al. (2019) and Hill et al. (2019)), but in contrast to many existing studies, we analyze frequencies of linear accelerations and side accelerations
separately and combine them with heavy breaking data (deceleration rates). Moreover, our data allow us to analyze the influence of accelerations of different intensity separately. In contrast to the previous groups of indicators, acceleration indicators mostly characterize the driving style of a particular driver. This information allows us to obtain a description of both medium-run and long-run driving style and to track the changes in behavior inside a week. The availability of acceleration indicators is the key feature of our data that is possible to be obtained using telematics devices and that is often not used in accident modeling studies. As shown below, this unique information can significantly improve the quality of resulting models.

For every indicator we introduce three levels of strength and calculate the frequency of corresponding accelerations, decelerations, and side accelerations per 100 kilometers to eliminate the effect of mileage. The resulting set of indicators is shown in Table 3.

The acceleration thresholds that are used to divide the data into accelerations and decelerations of different levels are standard for all the telematics devices installed and cannot be changed by analysts. Nevertheless, as we will see later, such a separation allows us to obtain models of sufficient quality (Figures 7 and 8).

### 3.5. Accident rate in the data

To perform a more thorough analysis of factors influencing the accident probability, it was decided to separate the accidents into three groups accordingly to the ratio of losses over the observation period to the insurance sum. The loss data as well as the insurance sum data were obtained
Table 3. Acceleration indicators.

| Indicator | Description |
|-----------|-------------|
| a1        | Frequency per 100 km of level 1 acceleration (0.3–0.4 G) events |
| a2        | Frequency per 100 km of level 2 acceleration (0.4–0.5 G) event |
| a3        | Frequency per 100 km of level 3 acceleration (0.5+ G) events |
| d1        | Frequency per 100 km of level 1 deceleration (0.2–0.3 G) events |
| d2        | Frequency per 100 km of level 2 deceleration (0.3–0.4 G) events |
| d3        | Frequency per 100 km of level 3 deceleration (0.5+ G) events |
| s1        | Frequency per 100 km of level 1 side acceleration (0.3–0.4 G) events |
| s2        | Frequency per 100 km of level 2 side acceleration (0.4–0.6 G) events |
| s3        | Frequency per 100 km of level 3 side acceleration (0.6+ G) events |

Figure 7. Average frequency of level 1 accelerations, events per 100 km. Year and week number on horizontal axis.

Figure 8. Average frequency of level 2 side accelerations, events per 100 km. Year and week number on horizontal axis.
from insurance companies of the drivers under investigation. It may not
include some minor accidents, in which the driver decided not to repair
the car or make the repairs himself or herself, but it cannot create a signifi-
cant bias. The ratio was calculated by simple division of total losses by cor-
responding insurance sums.

Analysis of Figure 9 allows us to form three groups (vertical lines on the
plot are borders of the groups). It was decided to treat accidents with loss
to insurance ratio smaller than 5% as small, from 5% to 20% as medium-
sized, and as greater than 20% as strong. The separation of accidents into
groups was done expertly based on the distribution of loss to insurance
sum ratio (Figure 9) in a way that accounts for the properties of data: a
single peak in the 0% to 5% group, a decline in frequency for 5% to 20%
groups, and a roughly constant frequency for groups with loss to insurance
sum ratio of 20% and higher.

In some cases, the same vehicle has more than one accident; hence, there
can be a bias of some sort (for example, we cannot distinguish between
two small accidents and one medium), but the number of observations
with more than one accident is small and this factor will not influence the
results significantly. This approach gives us 386 observations with weak
accidents, 278 with medium accidents, and 102 observations with strong
accidents. We also have 30 observations where insurance company indi-
cates that a vehicle was in an accident but with zero losses; we treat them
as observations without accident (there probably was an accident, but a
minor one and no repairs were needed, so we can neglect it).
Some descriptive statistics are presented in Table 4. It can be noted that the means of some variables differ between the two groups, especially mileage and average number of trips per day. Drivers that were involved in accidents tend to have higher mileages and number of trips per day. The same is true for all other indicators of mileage. Accident-free drivers are also characterized by lower maximum speeds and shares of mileage with high (over 100 and over 130 kph) speed. Typical loss size is about 10% of typical insurance sum.

The correlation matrix of the variables used in model selection is presented on Figures 10 through 12. We see that high maximum speeds are associated with each other and with high mileage. A high frequency of 0 to
Figure 10. Correlation matrix for mileage indicators.

Figure 11. Correlation matrix for speed indicators.
20 kph speed limit violations is, not surprisingly, also associated with maximum speeds as well as high average speed. Accelerations and side accelerations are generally independent of other variables and each other. Mileage indicators, in contrast, are highly correlated.

### 3.7. Modeling approach and accuracy evaluation

We use logistic regression as the main accident probability model. Here we follow the argumentation of Baecke and Bocca (2017) that points out the importance of interpretability and necessity of “white box” models as a standard regulatory requirement.

Event probability in logistic regression is calculated as

$$P(Y_i|X_i) = \frac{1}{1 + \exp^{\beta X_i}}$$  \hspace{1cm} (1)

where $X_i$ – vector of factors for i-th observation, including intercept; $\beta$ – vector of model parameters, estimated usually with maximum likelihood.

To measure the quality of resulting models we use in- and out-of-sample area under the receiver operator characteristic curve (ROC AUC) that can be interpreted as the probability that a randomly chosen positive instance (driver that has an accident) has probability higher then randomly chosen negative instance (driver without accidents). We also calculate more common in econometrics literature McFadden R-squared measure that is...
\[ R_{McFadden}^2 = 1 - \frac{\log L(\text{model})}{\log L(\text{constant})}, \]
where \( \log L(\text{model}) \) and \( \log L(\text{constant}) \) are logarithms of likelihood functions for the model under consideration and a model that gives a constant forecast.

4. Results and discussion

4.1. A general approach

Before we proceed to accident probability models estimation and analysis, it is also important to describe our approach to data types when solving scoring problems. We have already mentioned the two approaches to data processing that differ in time interval that is used for data averaging and aggregation. Each of these approaches solves its own problem.

Lifetime data–based scoring (or long-term data–based scoring) is used mostly to assess and filter the clients of an insuring company that uses the scoring system. There is no way to analyze short-term behavior of particular drivers under this approach, but it is possible to make a general assessment of his or her driving style and skills.

Short-term scoring that uses weekly or monthly data (work with daily data is much more complicated as it requires to deal with intra-week heterogeneity) can be used for the purposes of monitoring and correction of current behavior of the driver. Under this approach, it is possible to make a monitoring and individual recommendation system. In this paper, we focus on the first approach.

In our sample, we have 4,284 observations without accidents, 597 observations with one accident, and only 199 observations with more than one accident. Due to the small number of observations with more than one accident, it was decided to restrict the model scope to simple logic with the fact of accident (whether the vehicle was in any accidents) as a dependent variable. As a reminder, all the accidents discussed here are accidents in which the driver of the vehicle under consideration was at fault for the accident.

Four models were estimated: one for all accidents, and three models for accidents of different strength. In all the models we left only statistically

| Table 5. The influence of individual factors on the probability of an accident. |
|-----------------------------------------------|
| All accidents | Weak accidents | Medium accidents | Strong accidents |
| Mileage       | +             | +                | +               |
| a1            | +             | +                | +               |
| a2            | −             | −                | −               |
| avg_sp        | −             | −                | −               |
| max_n_sp      | +             | −                | +               |
| max_mj_sp     | +             | +                | +               |
| s1            | −             | +                | −               |
| d_night_m     | −             | +                | −               |
| max_ej_sp     | +             | +                | +               |
significant variables. As we are analyzing the binary outcome variable, we focus on logistic regressions. Table 5 contains the lists of significant factors for every accident type and the direction of their influence.

As was anticipated, one of the main factors influencing the probability of an accident is mileage. The higher the mileage, the higher the probability of accidents. This rule holds for all models except the model for strong accidents; here only driving style and car utilization factors matter. Frequency of 0.2 to 0.3 G accelerations also increases the accident probability, again, everywhere except the model for strong accidents. It was seen that 0.2 to 0.3 G is the level achieved when the car starts to move, whether at traffic lights or in congestion. Hence, this variable can be viewed as an indicator of frequent use of a car in the city and on congested roads, where the probability of small accidents is higher (at least in Russia). The same is true for day_m_pr (ratio of daytime mileage to total mileage): It reduces the probability of strong accidents (strong accidents more often happen at nighttime; this idea is confirmed by positive influence of average nighttime mileage on medium-sized accidents). The last car utilization indicator is average speed: The higher the average speed, the less time the vehicle spends in traffic congestions, which reduces the probability of accidents in general and weak accidents.

**Table 6.** The results of the estimation of the probability of accident models.

| Dependent variable | All accidents | Weak accidents | Medium accidents | Strong accidents |
|--------------------|---------------|----------------|------------------|-----------------|
|                    | DTP_f         |                |                  |                 |
| mileage            | 0.0000***     | 0.0000***      | 0.0000***        | 0.007**         |
| (0.000)            |               |                |                  | (0.003)         |
| max_ej_sp          | 0.010***      | 0.007***       | 0.006**          | 0.022***        |
| (0.003)            |               |                | (0.003)          | (0.006)         |
| a1                 | -0.029**      |                |                  | -0.119***       |
| (0.013)            |               |                |                  | (0.042)         |
| max_mj_sp          | 0.004***      | 0.009***       |                  |                 |
| (0.001)            |               |                |                  |                 |
| s1                 |                | -0.047***      |                  | 0.017***        |
|                    |                | (0.013)        |                  | (0.005)         |
| avg_sp             | -0.020***     | -0.021***      |                  |                 |
| (0.006)            | (0.007)       |                |                  |                 |
| max_n_sp           | 0.005***      |                | 0.004**          | 0.005*          |
| (0.001)            |               |                | (0.002)          | (0.003)         |
| d_night_m          |                |                | 0.004*           |                 |
|                    |                |                | (0.002)          |                 |
| Constant           | -2.880***     | -3.352***      | -3.863***        | -5.641***       |
| (0.196)            | (0.249)       | (0.229)        | (0.388)          |                 |
| Observations       | 5,050         | 5,050          | 5,050            | 5,050           |
| Log likelihood     | -2.080        | -1.303         | -1.038           | -480            |
| Akaike information criterion | 4,174.6 | 2,618.3 | 2,087.4 | 972.7 |

*Note. Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001.*
Moving on to driving style indicators, we see that maximum speed during morning rush hour, night, evening rush hour increases the probability of weak, medium and strong accidents respectively (for first two cases the influence is also present in general model). The mechanism is quite clear: The higher the maximum speed, the less careful and more prone to accidents the driver is. The difference of influences of different types of maximum speeds is less clear and may be due to imperfections of data. Finally, 0.2 to 0.3 G side accelerations (usually sharp changes of lane) reduce the probability of weak accidents and increase the probability of strong ones. Weaving is typical for careless drivers, and the model estimates confirm that they are more prone to get into more serious accidents. Now we move to a detailed analysis of models separately and describe a portrait of a typical driver who gets into an accident of a certain type.

4.2. Detailed modeling results

A typical driver with high accident probability (Table 6):

- Has high mileage
- Makes a lot of level 1 accelerations
- Drives with low average speed
- Drives fast at night and in mornings

A typical driver with high weak accident probability:

- Has high mileage
- Drives with low average speed
- Drives fast in the mornings
- Makes a lot of level 1 accelerations but few level 1 side accelerations

A typical driver with high medium accident probability:

- Has high mileage and high night mileage
- Makes a lot of level 1 accelerations
- Drives fast at night

A typical driver with high strong accident probability:

- Drives fast in the evening
- Drives a lot at night
- Makes a lot of level 1 accelerations and level 1 side accelerations
- Makes few type 2 accelerations
4.3. Accident probability models quality assessment

To estimate the forecasting power of the model we use one of the standard indicators: the ROC AUC. Table 7 contains both the in-sample and out-of-sample results for the four models. For out-of-sample calculation, we used a randomly chosen 10% of observations: The model was estimated on 90% of the sample, and the forecast was built for the other 10% of observations.

We can see that in-sample and out-of-sample estimates are very close. It may indicate that the models do not overfit the data and have a reasonable forecasting power.

Another important issue is the increase of models’ quality due to acceleration information introduction to the models. To measure it, we estimated the same models without acceleration indicators (with mileage and speed indicators as main variables). Table 8 compares McFadden $R^2$ for models with and without accelerations.

We can see that introduction of acceleration indicators (“jerky driving” indicators) increases the quality of models, especially for weak and strong accidents.

5. Conclusions

This paper contributes to automotive insurance literature by offering a methodology of telematics data application from data collection and preprocessing to accident probability model estimation. It also contributes to accident research literature by providing some insights into driving style factors of accident probability and severity and highlighting the importance of jerky driving indicators in accident prediction.

We present an application methodology for telematics device data to vehicle insurance. The paper discusses the general approach to data collection and preprocessing and identifies the most significant factors that influence the accident probability using real-world data. Both vehicle utilization and driving style indicators influence the accident probability, but the set of significant variables depends on the strength of the accident examined.
Utilization of accelerations data can greatly improve the quality of accident probability predictions models, especially for weak and strong accidents. The results obtained in this paper can be valuable both for insurance companies as a guideline for telematics data application in automotive insurance and for authorities to prioritize the road policy measures development.

**ORCID**

Ivan Stankevich [http://orcid.org/0000-0001-7391-4055](http://orcid.org/0000-0001-7391-4055)

**References**

Anderson, T. (2007). Comparison of spatial methods for measuring road accident 'hotspots': A case study of London. *Journal of Maps, 3*(1), 55–63. doi:10.1080/jom.2007.9710827

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems, 98*, 69–79. doi:10.1016/j.dss.2017.04.009

Bagdadi, O., & Várhelyyi, A. (2011). Jerky driving—An indicator of accident proneness? *Accident Analysis & Prevention, 43*(4), 1359–1363. doi:10.1016/j.aap.2011.02.009

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

Bil, M., Andrásik, R., & Janoška, Z. (2013). Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. *Accident; Analysis and Prevention, 55*, 265–273. doi:10.1016/j.aap.2013.03.003

Durduran, S. S. (2010). A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform. *Expert Systems with Applications, 37*(12), 7729–7736. doi:10.1016/j.eswa.2010.04.068

Elliott, M. A., Armitage, C. J., & Baughan, C. J. (2003). Drivers’ compliance with speed limits: An application of the theory of planned behavior. *Journal of Applied Psychology, 88*(5), 964–972. doi:10.1037/0021-9010.88.5.964

Erdogan, S., Yılmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: City of Afyonkarahisar. *Accident; Analysis and Prevention, 40*(1), 174–181. doi:10.1016/j.aap.2007.05.004

French, D. J., West, R. J., Elander, J., & Wilding, J. M. (1993). Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics, 36*(6), 627–644. doi:10.1080/00140139308967925

Hill, A., Horwill, M. S., Whiting, J., & Watson, M. O. (2019). Computer-based hazard perception test scores are associated with the frequency of heavy braking in everyday driving. *Accident; Analysis and Prevention, 122*, 207–214. doi:10.1016/j.aap.2018.08.030

Husnjak, S., Peraković, D.,Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering, 100*, 816–825. doi:10.1016/j.proeng.2015.01.436

Kashani, A. T., & Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science, 49*(10), 1314–1320. doi:10.1016/j.ssci.2011.04.019
Kashani, A. T., Shariat-Mohaymany, A., & Ranjbari, A. (2012). Analysis of factors associated with traffic injury severity on rural roads in Iran. *Journal of Injury & Violence Research, 4*(1), 36–41. doi:10.5249/jivr.v4i1.67

Kosuge, R., Okamura, K., Kihira, M., Nakano, Y., & Fujita, G. (2017). Predictors of driving outcomes including both crash involvement and driving cessation in a prospective study of Japanese older drivers. *Accident Analysis & Prevention, 106*, 131–140. doi:10.1016/j.aap.2017.05.019

Mohaymany, A. S., Shahri, M., & Mirbagheri, B. (2013). GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. *Geo-spatial Information Science, 16*(2), 113–119.

Mousavi, S. M., Zhang, Z., Parr, S. A., Pande, A., & Wolshon, B. (2019). Identifying high crash risk highway segments using jerk-cluster analysis. In *International Conference on Transportation and Development 2019: Smarter and Safer Mobility and Cities* (pp. 112-123). Reston, VA: American Society of Civil Engineers.

Naude, C., Serre, T., Dubois-Lounis, M., Fournier, J. Y., Lechner, D., Guilbot, M., & Ledoux, V. (2019). Acquisition and analysis of road incidents based on vehicle dynamics. *Accident; Analysis and Prevention, 130*, 117–124. doi:10.1016/j.aap.2017.02.021

Olszewski, P., Szagal, P., Wolański, M., & Zielinska, A. (2015). Pedestrian fatality risk in accidents at unsignalized zebra crosswalks in Poland. *Accident; Analysis and Prevention, 84*, 83–91. doi:10.1016/j.aap.2015.08.008

Park, H. C., Kim, D. K., Kho, S. Y., & Park, P. Y. (2017). Cross-classified multilevel models for severity of commercial motor vehicle crashes considering heterogeneity among companies and regions. *Accident; Analysis and Prevention, 106*, 305–314. doi:10.1016/j.aap.2017.06.009

PTOLEMUS Consulting Group. (2016). *Usage-based insurance global study*. 2016 ed. Brussels: Ptolemus.

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident; Analysis and Prevention, 40*(5), 1732–1741. doi:10.1016/j.aap.2008.06.011

Sagberg, F., Selpi, B., Piccinini, G. F., & Engström, J. (2015). A review of research on driving styles and road safety. *Human Factors, 57*(7), 1248–1275.

Taubman-Ben-Ari, O., Mikulincer, M., & Gillath, O. (2004). The multidimensional driving style inventory—Scale construct and validation. *Accident; Analysis and Prevention, 36*(3), 323–332. doi:10.1016/S0001-4575(03)00010-1

Tesema, T. B., Abraham, A., & Grosan, C. (2005). Rule mining and classification of road traffic accidents using adaptive regression trees. *International Journal of Simulation, 6*(10), 80–94.