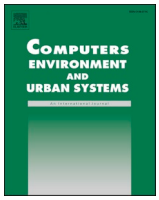
Computers, Environment and Urban Systems 93 (2022) 101760

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Predicting individuals' car accident risk by trajectory, driving events, and geographical context

Livio Brühwiler [a], Cheng Fu [a,*], Haosheng Huang [b], Leonardo Longhi [c], Robert Weibel [a]

[a] Department of Geography, University of Zurich, Zurich, Switzerland
[b] Department of Geography, Ghent University, Ghent, Belgium
[c] SolarEdge E-Mobility (SE-em) S.p.A., Montecastelli, Italy

ARTICLE INFO

Keywords:
Trajectory
Car accident prediction
Geographical context
Usage-based insurance
Machine learning

ABSTRACT

With the prevalence of GPS tracking technologiees, car insurance companies have started to adopt usage-based insurance policies, which adapt insurance premiums according to the customers' driving behavior. Although many risk models for assessing an individual driver's accident risk based on the history of driving trajectories, driving events, and exposure records exist, these models do not take the geographical context of the driven trajectories and driving events into account. This study explores the influence of enriching the existing purely driving-behavior-based feature set by multiple geographical context features for the task of differentiating between accident and accident-free drivers. Prediction performances of five machine learning classifiers—logistic regression, random forest, XGBoost, feed-forward neural networks (FFNN), and long short-term memory (LSTM) networks—were evaluated on the usage records of over 8000 vehicles in one year from Italy. The results show that the inclusion of geographical information such as weather, points of interest (POIs), and land use can increase the relative predictive performance in terms of AUC by up to 8%, among which land use is the most informative. For the data of this study, XGBoost generally yielded the best performance and made most use out of the geographical information, while logistic regression is only slightly outperformed by more complex models if the proposed geographical information is not available. LSTM did not outperform the other methods, possibly due to the small volume of training data available. The results outline the potential of including the geographical context in usage-based car insurance risk modeling to improve the accuracy, leading to fairer usage-based insurance policies.

1. Introduction

Every year around 1.35 million people die in road accidents, which today are the leading cause of death among children and young adults (WHO, 2020). A vast majority of these accidents are caused by human errors. For instance, in both the United States and Switzerland, around 94% of car accidents involve some form of human errors (National Highway Traffic Safety Administration, 2018; Swiss Federal Office of Statistics, 2018). Among the three main influencing factors of road safety—driver, road, and vehicle—driver is the most important but also the hardest to change/improve (Eboli, Mazzulla, & Pungillo, 2017; J. Wang, Wu, & Li, 2015). Unsafe drivers need to be identified before turning them into safer drivers. However, the conventional self-report is not an optimal method for driver's risk assessment as 80.3% of drivers self-describe their driving ability as above average (Nees, 2019).

The car insurance industry has been trying to classify drivers into different risk levels for years. Traditionally, car insurance companies only consider demographic factors such as gender, age, or vehicle model for rating risk levels (Lemaire, Park, & Wang, 2015). In recent years, with the prevalence of GPS tracking devices, insurers have started to adopt usage-based insurance (UBI) policies, including Pay-As-You-Drive (PAYD) and Pay-How-You-Drive (PHYD). PAYD takes the mileage or exposure of a driver into account; PHYD evaluates the profile of specific driving behaviors such as braking, acceleration, or speeding (Tselentis, Yannis, & Vlahogianni, 2016). Such UBI policies can provide a financial incentive for drivers to adopt a safer driving style and stop cautious drivers from subsidizing more risky or dangerous drivers (Tselentis et al., 2016). PHYD policies can reduce the gender and age bias of pure demography-based risk assessment models (Ayuso, Guillen, & Pérez-Marín, 2016; Ayuso, Guillén, & Pérez-Marín, 2014; Verbelen, Antonio,

* Corresponding author.
E-mail address: cheng.fu@geo.uzh.ch (C. Fu).

https://doi.org/10.1016/j.compenvurbsys.2022.101760
Received 31 August 2021; Received in revised form 7 January 2022; Accepted 17 January 2022
Available online 31 January 2022

& Claeskens, 2018) and also have a positive effect on reducing fuel consumption and noise emission when drivers adapt their driving behavior patterns to safer styles (Bordoff & Noel, 2008).

Although many insurance companies have already started to offer UBI policies, challenges remain for the car insurance industry to extract meaningful and explainable features from the raw trajectory data to reflect a driver's risk profile more accurately and therefore provide fairer policies. Currently, existing risk models reported in the literature still focus on modeling the drivers' own behavior without integrating the geographical and environmental context of the driven trajectories, such as weather conditions or land use. However, evidence has shown that geographical context, such as weather, significantly impact accident risk (Kantor & Stárek, 2014; Winlaw, Steiner, MacKay, & Hilal, 2019). The same driving behavior, such as braking, also corresponds to different risk levels in different circumstances, e.g., on the highway or in the urban context (Husnjak, Peraković, Forenbacher, & Mumdziev, 2015). Therefore, Husnjak et al. (2015) describe the inclusion of environmental factors as the most critical step in the further development of UBI.

This paper explores how integrating different geographical context sets, such as weather conditions, points-of-interests (POIs), and land use, can improve different machine learning-based PHYD risk assessment models on *differentiating accident and accident-free drivers*. We compare the influence of each geographical context set on both interpretability-oriented and performance-oriented machine learning models, and address the tradeoffs of the inclusion. Even though this study does not aim to propose any specific driver-rating methods for the car insurance industry, it aims to illustrate the value of including geographical and environmental features in assessing the risk level of individual drivers. Specifically, via a case study, this paper aims to answer the following research questions:

- RQ1: To what extent can the selected geographical context information improve the existing behavior-based risk prediction models in identifying drivers at car accident risk?
- RQ2: Which machine learning technique performs best for predicting car accident risk, and what is the tradeoff between interpretability and predictive performance?

Answers to these questions will help to better assess the risk level of individual drivers. The findings of this study, if communicated properly to the drivers, can ultimately help them to drive more carefully according to their circumstances for better road safety.

2. Related work

2.1. Car accident risk assessment modeling

2.1.1. Exposure-based Pay-As-You-Drive models

Pay-As-You-Drive (PAYD) policies are exposure-based, that is, they consider how much, when, and where someone drives (Baecke & Bocca, 2017). The premium is expected to be higher for drivers who travel longer distances (Denuit, Marchal, Pitrebois, & Walhin, 2007) because long-distance trips significantly increase the risk of accidents (J.-P. Boucher, Peârez-Marõân, & Santolino, 2013; Lemaire et al., 2015; Litman, 2005). Early ideas of implementing mileage into insurance pricing include pay-at-the-pump (Sugarman, 1994), where a surcharge is applied for each liter of petrol, and self-reported mileage estimates with occasional verification by the insurance company (Litman, 2011).

However, the positive relationship between total mileage and accident risks is not necessarily always a monotonically increasing linear function. Higher mileage may also result in a lower per-mile crash rate because high-mileage drivers are also more skilled and long-distance driving happens more on highways than inside urban areas (Guillen, Nielsen, Ayuso, & Pérez-Marín, 2019; Janke, 1991; Langford, Koppel, McCarthy, & Srinivasan, 2008; Litman, 2011; Paefgen, Staake, & Fleisch, 2014). With the advent of GPS technologies, the exposure can be further categorized by temporal details such as peak-hours driving or night driving, and geographical details such as driving by road types and urban-rural driving (Ayuso et al., 2014; Ayuso, Guillen, & Nielsen, 2019; Baecke & Bocca, 2017; Guillen et al., 2019; Paefgen et al., 2014; Paefgen, Staake, & Thiesse, 2013).

2.1.2. Behavior-based Pay-How-You-Drive models

Pay-How-You-Drive (PHYD) is an extension of PAYD, where in addition to the exposure features, driving behaviors such as speeding, braking, and acceleration are also considered (Tselentis et al., 2016). There is no clear conceptual boundary to distinguish PAYD and PHYD; different studies may categorize the same feature set into either type (Husnjak et al., 2015; Tselentis et al., 2016; Verbelen et al., 2018). The frequency of acceleration events (Af Wåhlberg, 2004), braking events (Stipancic, Miranda-Moreno, & Saunier, 2018), and speeding (Ma, Zhu, Hu, & Chiu, 2018) has been found to have a significant positive correlation with the occurrence of car accidents. The high frequency of such events is suggested to be an indicator of dangerous driving behavior (Musicant, Bar-Gera, & Schechtman, 2010).

Car accident risks vary temporally. Several aforementioned PAYD and PHYD studies have observed higher crash frequencies during peak hours and weekdays, probably attributable to higher traffic volumes. Furthermore, frequent night driving, especially on weekends and Friday evenings, has been attributed to higher accident probability due to bad visibility and other factors, such as intoxicated driving (Paefgen et al., 2014).

2.1.3. Impact of geographical context on car accident risks

2.1.3.1. Weather. In general, bad weather conditions have been shown to increase the frequency of car accidents (Peng, Jiang, Lu, & Zou, 2018) and the severity of the accidents (Fountas, Fonzone, Gharavi, & Rye, 2020). Rainfall (Andrey & Yagar, 1993; Bergel-Hayat, Debbarh, Antoniou, & Yannis, 2013; Caliendo, Guida, & Parisi, 2007), fog (Black & Mote, 2015; Eisenberg & Warner, 2005), and winter precipitation such as snowfall, freezing rain, and ice pellets (Eisenberg & Warner, 2005) were all found having a positive effect on higher crash frequency. Extreme temperature, either high (Bergel-Hayat et al., 2013; Wyon, Wyon, & Norin, 1996) or low (Malyshkina, Mannering, & Tarko, 2009), can also increase the car accident probability. The joint effect of bad weather and bad lighting conditions such as night driving in the rain can further amplify the probability of driving errors, hazardous driving, and resulting accidents (Fountas et al., 2020). However, exceptional cases also show that rainfall can lead to lower crash frequencies, perhaps due to adapted driving behavior or different exposure levels (Bergel-Hayat et al., 2013; Yannis & Karlaftis, 2011).

2.1.3.2. POIs. Different POIs are associated with different human behaviors. Very few studies have explored the relationship between POIs and car accidents, perhaps due to the connection not being very intuitive. Ng, Hung, and Wong (2002) discovered a significant positive relationship between cinemas, hospitals, markets, railway stations, and the number of accidents in Hong Kong. Jia, Khadka, and Kim (2018) found areas with a high density of banks, hospitals, and residential areas to have a higher crash frequency. Kufera et al. (2020) found a newly built casino to contribute to an increased crash frequency in nearby areas, especially on the weekends. Additionally, a higher number of pedestrian-vehicle collisions were observed to happen around certain POIs such as hotels, retails, restaurants, and transportation stations (Lee, Chae, Yoon, & Yang, 2018; Yao, Wang, Fang, & Wu, 2018).

2.1.3.3. Land use. Land use categories offer a more general functional description of places than POIs. Several studies observed a higher crash frequency in commercial and residential land use areas (Lym & Chen, 2020; Yang & Loo, 2016). More specifically, Yang and Loo (2016) found

commercial mixed with residential land use to be linked with the highest crash frequency. In contrast, Kim and Yamashita (2002) observed higher crash frequency in commercial than residential areas. The mixture of commercial and residential land use has also been shown to increase the frequency of vehicle-pedestrian collisions (Y. Wang & Kockelman, 2013; Wier, Weintraub, Humphreys, Seto, & Bhatia, 2009). Additionally, rural and agricultural areas have been shown to have lower crash frequencies than urban areas (Alkahtani, Abdel-Aty, & Lee, 2019; Kim & Yamashita, 2002).

2.1.3.4. Other factors. Several other environmental and anthropogenic factors that can have an impact on car accidents are described in the literature, which includes road types and intersections (Zhang, Xu, Cheng, Chen, & Zhao, 2018), traffic conditions at road and lane levels (de Medrano & Aznarte, 2021; Wang, Lin, Guo, & Wan, 2021), large events (Gutierrez-Osorio & Pedraza, 2020), air pollution (Wan, Li, Liu, & Li, 2020), and even the stock market (Giulietti, Tonin, & Vlassopoulos, 2020).

2.2. Machine learning in UBI modeling

Regression models such as generalized additive models (GAM) and logistic regression are heavily used in both PAYD and PHYD applications, either as the primary modeling tool (J. P. Boucher & Turcotte, 2020; Verbelen et al., 2018) or baseline models (Pesantez-Narvaez, Guillen, & Alcañiz, 2019). For achieving the best modeling performance, researchers usually use multiple machine learning models. For example, Bian, Yang, Zhao, and Liang (2018) proposed a bagging-based ensemble learning approach for a multilevel risk classification on a driver level and compared the proposed model with logistic regression and Naive Bayes. Huang and Meng (2019) applied several prediction models, including logistic regression, Poisson regression, random forest, XGBoost, support vector machine (SVM), and artificial neural network (ANN), to a PHYD task. Recently, deep learning models were also tried out, such as convolutional neural networks (CNN) for risk modeling (Yan, Wang, Liu, Liu, & Liu, 2020) and long short-term memory (LSTM) networks for driver profile modeling (Cura, Kucuk, Ergen, & Oksuzoglu, 2020). While the deep learning models have a powerful capacity for modeling complexity, the tradeoff is the loss of interpretability that the regression models have, as well as the need for large volumes of training data.

In Table 1, we summarize existing computational approaches proposed for UBI modeling, comparing the features used, predicting models employed, and the model output. We also position the contribution of this study in line with the progress of the existing studies. As can be seen from the table, existing UBI modeling did not take sufficient geographic context into account.

2.3. Research gaps

Many studies exist regarding various driver-centered PHYD and PAYD models, using a large variety of exposure-based and behavior-based driving information as well as several different machine learning techniques. However, there has been no consensus regarding which combinations of features and techniques are optimal. Simultaneously, the impact of geographical conditions such as weather, POIs, and land use on car accident frequency is well documented. However, existing work has mainly focused on a location or temporal perspective, rather than a driver's perspective. In other words, geographical information has not been integrated into models that assess the risk level of an individual driver's driving behaviors. Lastly, the previous studies mainly rely on data aggregated over several months. There exists a potential for trip-based models with finer temporal granularity, where the sequence and order of the driven kilometers are taken into account as well.

Table 1
Summary of previous PAYD and PHYD studies in reverse chronological order; D: Demographics; V: Vehicle Specific; E: Exposure; EV: Driving Events; S: Speeding, G: Geographic.

Study	Input features	Models	Model output
Boucher and Turcotte (2020)	E	Generalized additive models	Claim Number Prediction
Yan et al. (2020)	E, EV	Convolutional Neural Network, Support Vector Machine	Multilevel Risk Classification
Ayuso et al. (2019)	D, V, E, S	Poisson Regression	Claim Number Prediction
Guillen et al. (2019)	D, E, S	Zero Inflated Poisson Regression	Claim Number Prediction
Huang and Meng (2019)	D, E, EV, S, V	Logistic Regression, Poisson Regression, Support Vector Machine, Random Forest, Neural Network, and XGboost	Claim Number Prediction, Accident vs. Accident-Free Classification
Ma et al. (2018)	D, E, EV, S	Logistic Regression, Poisson Regression	Accident Number Prediction, Accident vs. Accident-Free Classification
Verbelen et al. (2018)	D, V, E	Generalized additive model	Claim Number Prediction
Baecke and Bocca (2017)	D, V, E	Logistic Regression, Random Forest, and Neural Network	Accident vs. Accident-Free Classification
Ayuso, Guillén, and Pérez Marín (2016)	D, V, E, S	Survival Analysis (Weibull Regression)	Distance Travelled to First accident
Ayuso et al. (2014)	D, E, S	Survival Analysis (Weibull Regression)	Distance and Time Travelled to First Accident
Paefgen et al. (2014)	E	Logistic Regression	Accident vs. Accident-Free Classification
Paefgen et al. (2013)	E	Logistic Regression, Neural Network, and Decision Tree	Accident vs. Accident-Free Classification
Af Wåhlberg (2004)	EV	Correlation Analysis	Effect of Acceleration and Braking on Bus Driver Accident Involvement
This Study	**E, EV, G**	**Logistic Regression, Random Forest, XGBoost, Feed-forward Neural Network, and LSTM**	**Accident vs Accident-Free Classification; Influence of Geographic Features**

3. Methodology

3.1. Data

3.1.1. Driving data

Two datasets are used in this study: a vehicle GPS trajectory dataset recording the driving trajectories and associated driving events, and a crash dataset. The private-owned vehicle GPS trajectory data set, with waypoints annotated with driving behavior and crash labels, was provided by the Track&Know Project[1] funded by the European Horizon 2020 program. The trajectory data were collected by black boxes mounted to the vehicles. The black boxes collect GPS records, engine status, heading, and acceleration of the vehicle. The data provider had preprocessed the raw records by downsampling the raw GPS waypoints to 2 km, unless a waypoint also involves an engine status change, i.e., starting and stopping (Fig. 1). Four driving behaviors, i.e., acceleration,

[1] www.trackandknowproject.eu

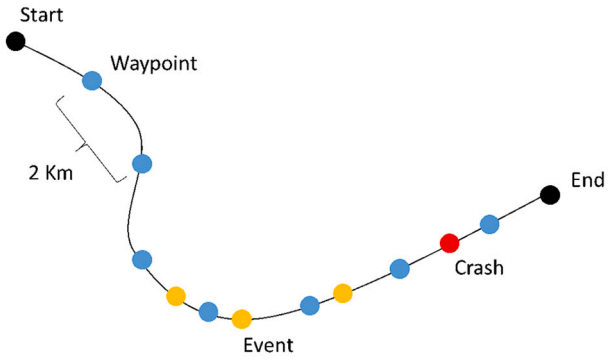


Fig. 1. An illustration of the downsampled raw GPS waypoints during a single trip. Blue points are downsampled raw GPS waypoints at a regular interval of 2 km. Orange points denote the GPS waypoints associated with one of the four driving behaviors, i.e., acceleration, braking, cornering, and quick lateral movement, which are recorded on top of the regular waypoints. Red points are the locations of a crash during a trip. Black points denote the start and end GPS waypoints of a trip. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

braking, cornering,[2] and quick lateral movement, are labeled by the data provider by modeling records from the accelerometer in the black box, and the associated GPS waypoint of recognized driving behavior is also provided, regardless of the 2-km downsampling strategy. The crash data set contains crash accidents recorded by the black box and validated by an automatic crash validation system in the black box or the crash assistance operation center of the data provider or both. In total, the data set contains trajectories of 12,145 randomly selected drivers (Table 2) in Rome and Tuscany, Italy, 2017. Rome is the capital city of Italy, with a complex urban transportation pattern. Tuscany is a region in central Italy, with a hilly rural landscape and cities such as Florence, Lucca, and Pisa.

It should be noted that the two data sets are associated with anonymized insurance policy IDs, each of which is assigned to a car. Each ID is recognized as a driver in this study, although a car can be shared by several drivers, such as family members in reality. In addition, the human-validated crash alarms only indicate that a vehicle was involved in a crash but do not indicate if the vehicle caused the accident.

3.1.2. Geographical context data

Freely accessible historical hourly weather forecast data were downloaded through the API of worldweatheronline,[3] which are used as a proxy for the actual weather records in many studies (Choi, Kim, Briceno, & Mavris, 2016; Elbeltagi, Zhang, Deng, Juma, & Wang, 2020). The underlying model data used by worldweatheronline is provided by the World Meteorological Organization and the NCEP global forecast system (Worldweatheronline, 2020). Furthermore, these forecast data follow the same terminology and classification methods for every region, making their output comparable, whereas local weather stations might have different formats of reporting their measurements and/or missing values. Eight weather-related features (Table 3) were collected with a 10*10 km spatial resolution.

OpenStreetMap (OSM) POI data were downloaded from geofabrik.[4] 25 POI types in the study area were collected (Table 4), which resulted in 458,134 point and polygon POIs.

OSM land use data were also downloaded from geofabrik. Eight original land use types were considered: *industrial*, *commercial*, *farm*, *grassland*, *park*, *residential*, *forest*, *retail*. For simplicity, *commercial* and *retail* were combined into *commercial*; and *farm*, *grassland,* and *park* were combined into *rural*.

Table 2
Overview of drivers and trips.

Driver type	Driver count	Trip count
Total drivers	12,145	7,675,648
Drivers with at least one accident	4097 (w/ 4319 accidents)	4,169,635
Drivers without accidents	8048	3,506,013

3.2. Defining the classification problem

There are two prediction strategies in regard to the driver-at-accident classification problem: The first strategy is to classify the driver into *accident* and *accident-free* categories according to their whole year of driving behavior regardless of when the accident happens. The other strategy is to take an early period of a year, e.g., January to June, for training to predict the possibility of having an accident in the later period, e.g., July to December. Both approaches have been used by previous studies: For example, Huang and Meng (2019) took the former strategy while Baecke and Bocca (2017) took the latter strategy.

The first strategy has the disadvantage of taking data after an accident into account: After an accident, a driver's driving behavior may change drastically, or a driver can be stopped from driving entirely by a severe crash (Mayou, Bryant, & Duthie, 1993). The second strategy, however, does not take seasonal variability into account, which is not optimal if weather is involved. Furthermore, the sample of drivers who had an accident would be further narrowed down to only those that had an accident between July and December of the year. Therefore, the former prediction strategy was chosen in this study. Ideally, more than one year of data would be available, so a prediction for the second year could be made based on the observations of the first year. To summarize, the classification problem can be formalized as follows:

Given a set of features derived from the driving behavior data over a whole year, a binary classification is performed to separate *accident drivers* from *accident-free drivers*. And in mathematical terms:

$$h : X \rightarrow y \in \{0, 1\}$$

where h is a classification function; X is a set of input vectors consisting of n driving behavior features $\{x_1, x_2, \ldots, x_n\}$; and y is a binary label, where that 0 stands for accident-free driver and 1 stands for accident driver.

3.3. Overview of the methodology

Fig. 2 shows the general workflow of this study. Firstly, raw data were cleaned and sampled. The raw positional and events data were enriched with geographical context data, i.e., weather, POI, and land

Table 3
Overview of weather-related features.

Feature	Unit
Temperature	°C
Humidity	%
Pressure	hPa
Precipitation	mm
Visibility	km
Weather Condition	Weather condition (e.g., sunny, rain, etc.)
Windspeed	km/h
Cloudcover	%

[2] Cornering is a harsh turning event identified by the black box on the vehicle when the overall acceleration is over 1.1 g during a turn.

[3] www.worldweatheronline.com

[4] www.geofabrik.de

Table 4
Overview of POI data.

Category	Included POI types
Commercial	Convenience Stores, Supermarkets, Pharmacies, Clothing Stores
Touristic	Attraction, Hostel, Hotel, Motel, Tourist Info
Nightlife	Restaurant, Pub, Cinema, Nightclub, Café, Bar, Fast Food
Public	Police station, School, Library, University, Kindergarten, Parks
Transportation	Bus stops, Railway stations, Taxi Stops

use. Secondly, enriched data were aggregated for each vehicle, and meaningful predictor variables (features) were computed. Aggregation took place at two temporal levels, *yearly* and *per-trip* (for LSTM only). Features then were grouped into different feature sets. As the trip-based aggregation suffers from data sparsity for features from driving events, the features in these two feature sets have slight differences, although they share the same principles of being generated.

Five machine learning classifiers were selected: logistic regression (LR), random forest (RF), XGBoost, feed-forward neural network (FFNN), and long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997). As introduced in the literature review, LR is commonly used in the insurance industry for its interpretability. RF, XGBoost, and FFNN have a powerful capacity for modeling complex data distributions. LSTM is a deep learning architecture built for modeling sequential, time series data, and therefore is a good candidate for modeling trip-based features. Their prediction performance on classifying drivers at risk using different feature sets were evaluated. If the trip-based aggregation combining LSTM could have a better prediction performance than the models from yearly aggregated features, insurance companies can use data from a much shorter observation period to customize policies. For model selection for a specific feature set, applicable hyperparameters were fine-tuned using another 5-fold cross-validation on the training folds in each model selection iteration to prevent information leaks (Cawley & Talbot, 2010). The overall performance of the best instance of each machine learning model were then evaluated by 5-fold cross-validation as a proxy for ground truth performance. Interpretation and comparison of models were then conducted to conclude the findings.

3.4. Data preprocessing

Drivers with low yearly mileage were firstly filtered out, leading to 3892 drivers with accident records and 4178 drivers with no accidents that remained. The details of this preprocessing step are given in Appendix A.1. The GPS waypoints of the remaining drivers were then enriched by the POIs and land use polygons. Again, the details can be found in Appendix A.2.

3.4.1. Feature engineering by yearly aggregation strategy

For each driver, features of their trajectories were aggregated from raw data using two strategies, i.e., by year and by trip, to support different classifiers. The yearly-aggregated features were fed to LR, RF, XGboost, and FFNN models, while the trip-aggregated features were fed to the LSTM models (which require sequential data as input). Eventually, 40 yearly-aggregated features for modeling driving risk were aggregated by drivers using both PAYD and PHYD strategies (Table 5). The detailed feature engineering strategies are described in Appendix A.3.

3.4.2. Feature engineering by trip-based aggregation strategy

LSTM networks model sequential data for prediction, which is different from the prediction procedure of the other four classifiers. Therefore, a different aggregation strategy had to be used. Firstly, each trajectory was segmented into trips. If the time difference between two consecutive waypoints was more than 15 min, the latter waypoint was taken as the start point of a new trip. Driving events were then assigned to trips by matching time stamps. After the segmentation, the enriched

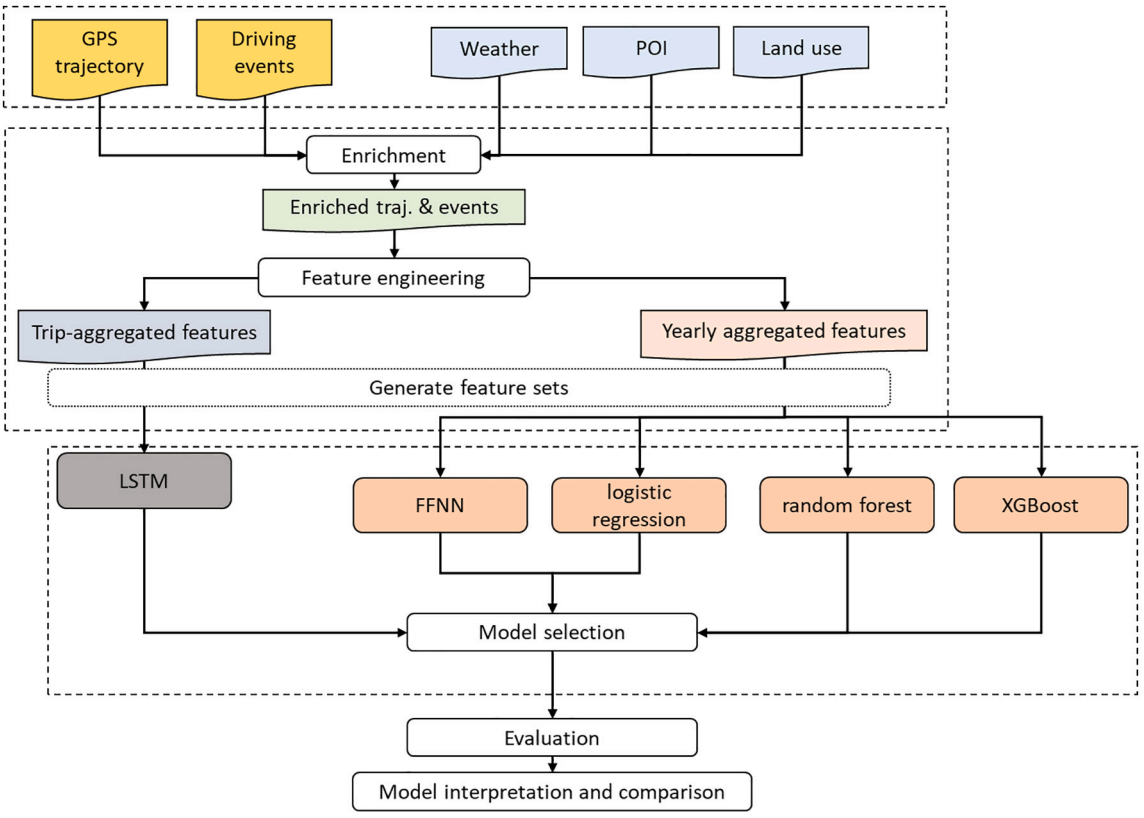


Fig. 2. Overall workflow.

Table 5
Derived yearly-aggregated features for LR, RF, XGBoost, and FFNN.

Source	Type	Feature name	Sample Median	
			Accident (n = 3892)	Accident-Free (n = 4178)
Trajectory	Basic exposure	Log of total yearly mileage (km)	4.09	3.97
		Median of trip mileage (km)	8.61	9.04
		Standard deviation of trip mileage (km)	13.97	15.48
		Median of trip duration (minutes)	25.08	25.12
		Mean of mean trip speed (km/h)	26.25	28.35
	Speed	Fraction of driving between 0 and 30 km/h	0.30	0.27
		Fraction of driving between 30 and 60 km/h	0.35	0.32
		Fraction of driving between 60 and 90 km/h	0.17	0.15
		Fraction of driving between 90 and 130 km/h	0.08	0.12
		Fraction of driving above 130 km/h	0.00	0.00
	Temporal	Fraction of night driving	0.10	0.09
		Fraction of rush-hour driving	0.33	0.32
		Fraction of weekend driving	0.28	0.29
	Weather	Fraction of driving in rain	0.07	0.7
		Fraction of driving in good weather	0.69	0.69
		Fraction of driving in overcast weather	0.23	0.23
		Fraction of driving in snow	0.00	0.00
		Fraction of driving in fog	0.01	0.01
		Fraction of driving above 25 °C	0.12	0.12
		Fraction of driving between 0 and 25 °C	0.87	0.87
		Fraction of driving below 0 °C	0.00	0.00
		Fraction of driving at night during rain	0.00	0.00
Driving Events	Basic exposure	Acceleration events per 1000 km	6.70	3.95
		Braking events per 1000 km	72.55	50.66
		Cornering events per 1000 km	294.35	219.59
		Quick lateral movement events per 1000 km	3.08	1.83
	Land use	Braking events in forest land use per 1000 km	0.47	0.65
		Cornering events in forest land use per 1000 km	6.34	4.40
		Cornering events in mixed forest-residential land use per 1000 km	0.72	0.43
		Accelerations in residential land use per 1000 km	1.67	0.82
		Braking events in residential land use per 1000 km	19.76	12.22
		Braking events in rural-residential land use per 1000 km	26.87	19.16
		Cornering events in residential land use per 1000 km	56.88	35.09
		Cornering events in mixed rural-residential land use per 1000 km	25.92	15.77
			26.87	19.16
	POI	Cornering events in rural land use per 1000 km	1.45	1.16
		Braking events in rural land use per 1000 km	0.08	0.07
		Fraction of driving events near commercial POI	0.11	0.10
		Fraction of driving events near nightlife POI	0.20	0.17
		Fraction of driving events near public POI	0.02	0.02
		Fraction of driving events near touristic POI	0.11	0.10
		Fraction of driving events near transportation POI		

waypoints were grouped by trips to derive selected features (Table 6). Only trips longer than 3 km were kept. The remaining trip sequences of each driver were padded to the maximum number of trips per driver as LSTM networks only accept inputs with the same size. Eventually, each driver has a sequence of trips as input for training and testing in LSTM.

3.5. Feature combinations

For assessing the impact of different feature groups, especially the geographical context information, six different feature combinations

Table 6
Derived trip-aggregated features for LSTM. * If either start or end time at night (23:00–7:00); ** If either start or end time in rush hour (7:00–9:00 and 16:00–19:00) in workdays; *** Predominant weather condition refers to the weather condition among the five conditions in which the majority of kilometers in the trip were driven, encoded as one-hot variables. **** Driving events were not distinguished as features for the conventional machine learning classifiers due to the data sparsity per trip.

Source	Type	Feature name
Trajectory	Basic exposure	Trip length (km)
		Trip duration (minutes)
		Trip start time
		Trip end time
		Average speed
	Temporal	Binary variable for night driving*
		Binary variable for rush-hour driving**
		Binary variable for weekend driving
	Weather	Predominant weather condition***
		Average Precipitation
		Average Temperature
Driving Events****	Basic exposure	Number of acceleration events
		Number of braking events
		Number of cornering events
		Number of quick lateral movement events
	Land use	Number of driving events in forest land use
		Number of driving events in commercial land use
		Number of driving events in residential land use
		Number of driving events in rural land use
		Number of driving events in industrial land use
	POI	Number of driving events near commercial POI
		Number of driving events near nightlife POI
		Number of driving events near public POI
		Number of driving events near touristic POI
		Number of driving events near transportation POI

were created based on the yearly-aggregated features (Table 7). Feature Set B serves as a baseline, including all exposure (also with speed and temporal) and event-related features, without the geographical context information. Compared to Feature Set B, Feature Set A only has pure exposure-based features without event-related features. Feature Set C, D, and E add weather, POI, and land use features to Feature Set B, respectively. Feature Set F includes all 40 available features. The LSTM models consume trip-aggregated feature sets but with the same set of themes. Therefore, we denoted the trip-aggregated feature sets for LSTM with an asterisk (*).

3.6. Machine learning models and evaluation

3.6.1. Model selection and evaluation

We used the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, and F1-Score for the assessment. As AUC is insensitive to class imbalance, which makes it suitable for imbalanced classification problems as this study, a stronger focus will be on the AUC for the interpretation of the results.

We then found the best setting for each machine learning model per feature set based on their performance. For LR, the best instance means we needed to find a subset of the input feature set, while for the other classifiers, the best instance means finding a set of hyperparameters to achieve the best classification results. The detailed strategies for each model are slightly different and are introduced in the next section. Once the best instance of each machine learning model was found, we applied 5-fold cross-validation, where each iteration uses 80% data for training and 20% for validation, and collected the average performance as a proxy for their performance on future data in real life.

3.6.2. Find model hyperparameter settings

We implemented logistic regression, random forest in sklearn (Pedregosa et al., 2011), and XGBoost using its Python API (Chen & Guestrin, 2016). The two neural networks, FFNN and LSTM, were implemented in TensorFlow (Abadi et al., 2016). The feature values for the four non-deep learning classifiers were rescaled between the first and third quartile using the *robust rescale strategy* in sklearn before being used for training the models of LR, RF, FFNN, and XGBoost. For LSTM, the values were rescaled to [0,1].

Since LR performs poorly for co-correlated features, the optimal set of features had to be determined. This was done using a stepwise cross-validation approach, where features were omitted in a stepwise fashion and 10-fold cross-validation performed to determine the AUC after each step (Baecke & Bocca, 2017; Huang & Meng, 2019; Paefgen et al., 2014). The feature set with the highest AUC was eventually chosen. This feature selection approach was independently performed on each of the six feature sets.

RF has many hyperparameters for fine-tuning to obtain the best model. For each feature set, the optimal hyperparameter set was achieved by grid searching the parameter combinations with 5-fold cross-validation (Appendix C). AUC was chosen as the scoring metric. All other hyperparameters were kept as default. Similar to RF, the same grid search procedure was applied to XGBoost to find the hyperparameter set that achieves the best classification accuracy (Appendix D).

For each grid-search step of RF and XGBoost, we applied a 5-fold-outer 5-fold-inner nested cross-validation approach (details described in Appendix E) to prevent information leakage, which can cause biased and over-optimistic performance measurements if the hyperparameters of a given model are optimized using a dataset that includes the validation data (Cawley & Talbot, 2010).

The FFNN architecture (Fig. 3.a) used in this study consisted of two dense layers with rectified linear unit activation functions (RELU) and 256 neurons for each layer. Furthermore, a dropout layer with 0.4 possibility was added after each hidden layer to prevent overfitting. The activation layer included a softmax activation function to return a probability between 0 and 1. Binary cross-entropy was chosen as the loss function. The batch size was set to 32. The models were trained for 50 epochs.

The LSTM architecture (Fig. 3.b) consisted of two stacked LSTM layers, which had 64 LSTM cells each. Furthermore, since LSTM models tend to overfit, three dropout layers with 0.4 possibility were added to combat this problem. As in FFNN, a softmax activation layer combining the binary cross-entropy was used as the output layer. The model was trained for 50 epochs, and the batch size was set to 100.

Since a systematic grid search procedure would be too expensive to find the best hyperparameter setting for a deep learning model considering the limited computational resources available, only a few different values for neuron numbers, batch size, number of epochs, and size of dropout layer were tried out manually for both neural networks.

Table 7
Feature sets as the combination of feature groups.

Feature Set	Exposure	Events	Weather	POI	Land Use
A, A*	X				
Baselines: B, B*	X	X			
C, C*	X	X	X		
D, D*	X	X		X	
E, E*	X	X			X
F, F*	X	X	X	X	X

4. Results

4.1. Model performance

4.1.1. Models for yearly aggregation strategy

The average performance figures of the models across all yearly-aggregated feature sets are presented in Table 8. In terms of the potential contribution of the geographical context to improving the prediction performance (RQ1), the following results can be highlighted:

1) The overall accuracy and AUC of all the ML models using the baseline feature set (Feature Set B) are both about 0.60.
2) More importantly, all ML models were improved when considering geographical data. Among all combinations, Feature Set F combining XGBoost performed best for all metrics (AUC: 0.71; Accuracy: 0.65; F1-Score: 0.644). The maximum gain of the model performance by introducing the three geographical feature sets can be about 0.05 in absolute terms, which translates to a relative increase of about 8%.

Regarding the contributions of the individual geographical feature

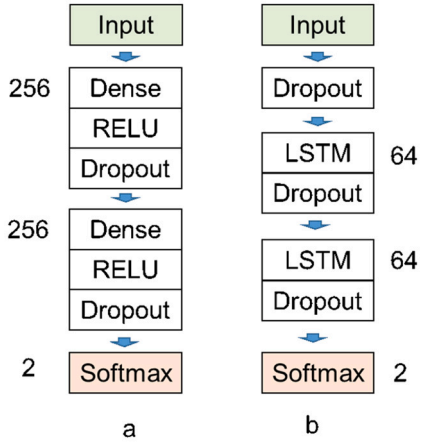


Fig. 3. The architecture of the feed-forward neural network (a) and the LSTM network (b). The numbered annotations indicate the number of neurons contained in the network layers.

Table 8
The average AUC, accuracy, and F1-score of the nested cross-validation per model and feature set combination for yearly-aggregated features.

Model	A	B (baseline)	C	D	E	F
AUC						
LR	0.643	0.652	0.654	0.653	0.665	0.664
FFNN	0.634	0.649	0.642	0.643	0.696	0.684
RF	0.637	0.655	0.662	0.656	0.700	0.699
XGBoost	0.641	0.655	0.655	0.655	0.709	0.714
Accuracy						
LR	0.591	0.596	0.602	0.597	0.609	0.609
FFNN	0.590	0.597	0.597	0.589	0.635	0.629
RF	0.594	0.603	0.610	0.605	0.643	0.637
XGBoost	0.590	0.600	0.606	0.606	0.648	0.650
F1-Score						
LR	0.604	0.601	0.606	0.602	0.612	0.611
FFNN	0.604	0.596	0.594	0.567	0.605	0.610
RF	0.607	0.607	0.614	0.609	0.639	0.634
XGBoost	0.626	0.620	0.616	0.623	0.642	0.644

groups (RQ1, Fig. 4), the following results can be observed

1) The land use-enriched feature group (Feature Set E) is outstanding compared to weather (Feature Set C) and POI (Feature Set D). For most classifiers, the results of Feature Set E are almost the same as Feature Set F for all three metrics (Fig. 4.a).
2) Weather can slightly increase the classification performances.
3) POI-related features contribute little and partly even negative effects on classifier performance, which is not as expected.

With regards to the comparison of ML models (RQ2), Table 8 further shows the following:

1) The tree-based classifiers, i.e., random forest and XGBoost, perform better than other classifiers in general, especially when the number of features is large. XGBoost performed the best for both baseline Feature Set B and Feature Set F, which considers all geographical features.
2) However, logistic regression (LR) models, which are relatively much simpler computationally, are only slightly worse than the other models.

4.1.2. Model for trip-based aggregation strategy

The average performance figures of LSTM models across all trip-aggregated feature sets are presented in Table 9.

For the potential contribution of the geographical context to improving the prediction performance (RQ1), a similar observation to the yearly aggregation strategy can be made comparing the results of Feature Set B* and Feature Set F*, although the improvement is not as big as with conventional machine learning models.

Regarding model comparison (RQ2), although the values of the F1-score for the LSTM models are also close to other model results in the yearly aggregation strategy, their performance in terms of AUC and accuracy are generally lower than the rest. The differences can be as big as 0.10 regarding the performance of XGBoost.

4.2. Model interpretation

We will now take a closer look at how the individual features contribute to the performance of the ML models. We do this separately for the logistic regression and tree models, respectively.

4.2.1. Feature coefficient analysis of logistic regression models

As a consequence of the feature selection, only a subset of the baseline features is selected for each feature set (Table 10). The exponentiated coefficients (EC) describe the multiplicative change in odds of being an accident driver over being an accident-free driver for each unit increase in the corresponding feature, assuming that all other features are being kept constant. Distance (*log of total yearly mileage*) is a strong

Table 9
The average AUC, accuracy, and F1-score of the nested cross-validation and feature set combination for trip-based aggregation for LSTM.

Model	A*	B* (baseline)	C*	D*	E*	F*
AUC						
LSTM	0.572	0.591	0.607	0.606	0.587	0.618
Accuracy						
LSTM	0.530	0.553	0.549	0.561	0.548	0.563
F1-Score						
LSTM	0.561	0.584	0.582	0.580	0.605	0.641

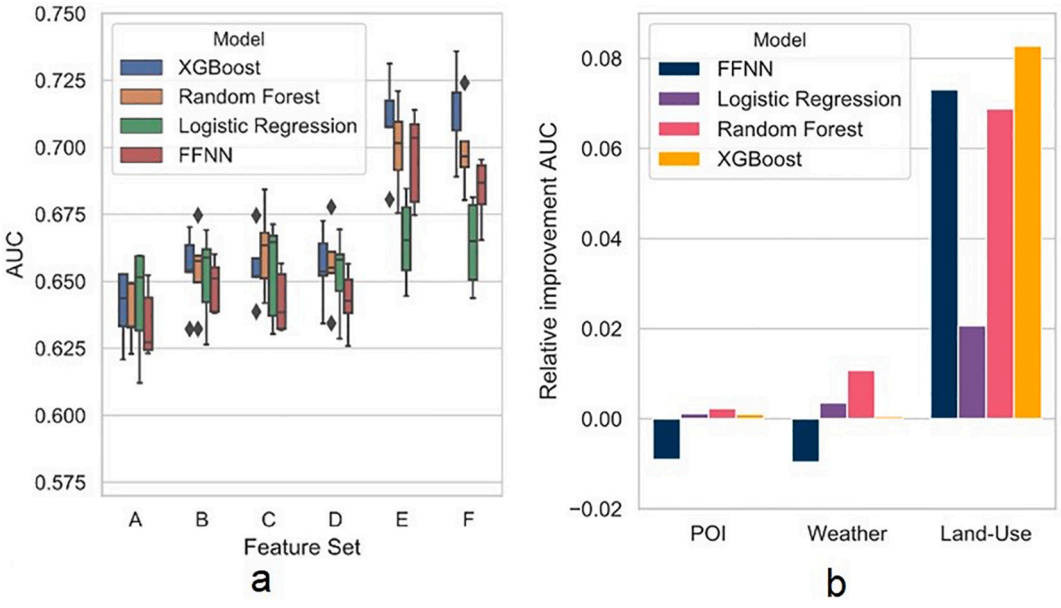


Fig. 4. a) AUC of cross-validation folds for feature set-classifier combinations. b) The relative improvement of each type of geographical information for classifiers in cross-validation.

common positive feature that contributes to an increase being an accident driver, while *median of trip distance* is a strong negative factor of risk. The number of cornering events (*cornering events per 1000 km*), mileage fraction of driving at night, and *median of trip duration* are the three common moderately positive features. However, most of the selected features still only have a moderate influence, as their exponentiated coefficients are close to 1.

After integrating geographically enriched features, the baseline features are still selected by the LR models, and their contributions to the risk are also consistent for most features. As we observed regarding the overall performance in Section 4.1, the contributions of other geographical contexts to the LR models in other scenarios are rather subtle. Particularly, the factions of breaking events in forest and residential areas and the faction of cornering events in rural areas are three negative features for accident risk. The two breaking-related features may be the indicators of a cautious driver. Other than that, the fractions of driving in rain, fog, or in warm weather (*fraction of driving above 25 °C*) have a negative effect on the risk increase. A possible reason is that drivers adapt to be more cautious in bad weather, or there might be simply fewer vehicles on the road in these situations, which was reported in previous studies as well (Section 2.1.3.1).

4.2.2. Feature importance analysis of the tree models

Table 11 shows that the feature importance scores (FI) of basic exposure features is consistent in both RF and XGBoost. For geographical context-related features, the land use-related features are important to predict accidents for both model types. For RF, the number of braking events in forests per 1000 km (*LU forest B*) exceeds the basic exposure features in terms of the importance when all features are used for prediction. Other land use-related event features also replace some basic exposure features in the top ranks. For XGBoost, the number of braking events in forests per 1000 km is not as important as in RF models but still ranks at the top. This observation also corresponds to the results of the logistic regression. Besides the land use-related features, temperature-related features also show relative importance in certain classifiers. For modeling drivers using XGBoost, the fraction of driving in high temperature (*High temp*) is also a relatively important indicator.

A further investigation on finding a minimal feature set for LR, RF, and XGBoost shows basic exposure features are essential for determining drivers at risk, but geographical features are critical for contributing to a higher prediction accuracy (Appendix G).

Table 10
Logistic regression coefficients of Feature Sets B and F. Features that are not significant in either case are not displayed. EC: exponentiated coefficient. EC > 1 indicates a positive effect of the feature value on the risk; EC = 1 indicates no influence. EC < 1 indicates a negative effect of the feature value on the risk. Symbol '-' indicates the feature is not selected in the logistic regression model.

Feature name	EC(B)	EC (F)
Log of total yearly mileage	1.674	1.660
Median of trip duration	1.146	1.201
Standard deviation of trip distance	–	0.901
Median of trip distance	0.848	0.868
Mean of mean trip speed	–	0.920
Fraction of driving between 60 and 90 km/h	1.144	1.079
Fraction of driving between 0 and 30 km/h	1.271	1.070
Fraction of driving between 90 and 130 km/h	0.786	0.835
Fraction of driving night driving	1.163	1.121
Fraction of driving above 25 °C	–	0.955
Fraction of driving in fog	–	0.939
Fraction of driving in rain	–	0.926
Braking events per 1000 km	–	1.264
Cornering events per 1000 km	1.184	1.143
Quick lateral movement events per 1000 km	–	1.048
Acceleration events per 1000 km	–	1.033
Braking events in forest land use per 1000 km	–	0.952
Cornering events in rural land use per 1000 km	–	0.930
Braking events in residential land use per 1000 km	–	0.888
Fraction of driving events near commercial POI	–	1.054
Percentage of driving events near transportation POI	–	0.962

Table 11
Top 15 important features of the RF and XGBoost models for Feature Set F. Features with a green background also rank in the top 15 important features of the corresponding RF or XGBoost models for the baseline Feature Set B. FI stands for Gini-based feature importance. To save space, feature aliases rather than the original feature names are listed. The pairs of feature aliases and their original names are listed in Appendix F.

RF		XGBoost	
Feature alias	FI	Feature alias	FI
LU forest B	0.054	0–30 km/h	0.073
Weighted distance	0.051	Mean trip speed	0.036
0–30 km/h	0.043	Weighted distance	0.039
Q	0.040	Q	0.031
90–130 km/h	0.037	LU forest B	0.036
30–60 km/h	0.031	LU forest C	0.030
Mean trip speed	0.033	B	0.030
B	0.033	30–60 km/h	0.024
C	0.026	LU rural residential B	0.025
Median trip distance	0.028	LU residential B	0.024
LU rural residential B	0.026	Median trip duration	0.024
LU residential B	0.027	High temp	0.024
Median trip duration	0.025	90–130 km/h	0.026
LU rural residential C	0.024	130 km/h+	0.022
LU forest C	0.025	LU rural residential C	0.022

4.3. Summary

From the overall model performance perspective, LR is similar to other classifiers if no geographical context information is involved in modeling. RF and XGBoost with Feature Set F that includes all possible features are better than any other ML model and feature set combinations. LSTM performs worst in all situations.

It can also be concluded that adding geographical context features can improve model performance for all ML models regarding all three metrics. Such performance increase can be as large as 5% in absolute terms, and 8% in relative terms, compared to the baseline.

Among the three types of geographical context features, land use-related features can help more for improving the overall performance. In the feature analysis, we observe that the fraction of braking in forests is an important indicator for accident prediction for LR, RF, and XGBoost. Other land use-related features and some weather-related features are also informative in certain situations. POI-related features, however, are not informative for any model in our study.

5. Discussion

5.1. Model comparison and impact of geographical information

The performance of all conventional machine learning models is close to previous studies, such as AUC = 0.62 reported by Baecke and Bocca (2017) and AUC = 0.61 by Huang and Meng (2019), although all studies have slightly different data and feature settings, rendering direct comparison difficult. Logistic regression, the current baseline method employed in PAYD and PHYD schemes performs, reasonably well and only gets slightly outperformed by more sophisticated classifiers. XGBoost generally yields the best predictive performance, paired with decent interpretability. However, due to the small performance difference from logistic regression, in practical situations it makes sense to use a logistic regression rather than more complex classifiers, especially for the interpretation of size and direction of the effects of the individual features. Particularly if the maximum possible performance is not the only aim and interpretability is taken into account, logistic regression should be the preferred model in practice. This confirms the findings of previous studies, namely Paefgen et al. (2013, 2014), Baecke and Bocca

(2017), and Huang and Meng (2019), who all report small benefits from the usage of more complex classifiers over logistic regression. Paefgen et al. (2014) specifically recommend the usage of the latter.

The LSTM network yields worse performance than the classifiers using the yearly-aggregated features. This might be due to the size of the training set, data aggregation procedure, trip definition, model architecture, or differences in the number and types of features used. Furthermore, it can be argued that car accident risk prediction also relies heavily on feature engineering and domain knowledge. Therefore, the LSTM model with its main strength in utilizing large volumes of raw time series data, might not be the optimal choice in our empirical case. The usage of more complex deep learning models in individual car accident risk prediction, however, is worth further investigation. The LSTM network showed significant improvement when including more features, which still reflects our main argument. Therefore, adding more geographic or other information might increase its performance further.

Enriching trajectories with geographical context information is able to increase the performance of all models. However, the contribution to such improvement is not equal across the different geographical information types. Land use seems to be the most impactful geographical feature for the Italian data used in this study, with weather and POIs only showing minor or no improvements over the baseline. This may relate to the complexity of the POI layout in big cities like Rome and Florence, where the enrichment of trajectories by POIs actually introduces more uncertainty. In addition, due to the lack of ground truth, our method to associate trajectories and POIs is purely based on spatial proximity, whereas a given associated pair of POI and waypoint may not actually have an interaction in reality. Road information, such as road types and road network attributes, may also improve model performance if map matching could be applied.

Apart from geographical features, several features that were deemed important in previous studies are confirmed. The total mileage remains an important factor among almost all models. Simultaneously driving at low speed and a high frequency of driving events resulted in higher accident risk, as expected. Furthermore, frequent trips that take a long time and night driving are linked to higher accident risk, possibly due to driver fatigue.

In short, for RQ1, it has been shown that the inclusion of geographical context features can further strengthen the prediction. Among the three types of geographical context information, land use data can improve the prediction more when used alone, and they are also stable and easy to access, compared to weather and POI data.

For RQ2, tree-based classifiers, such as random forest and XGBoost, perform best for the prediction task at hand, possibly owing to their robustness against overfitting and ability to model non-linear relationships. However, the gap of classification results between logistic regression and tree-based models is small, especially when geographical information is not included in the features. Replacing LR models with tree-based models may not be worthwhile in the insurance industry, considering the tradeoff between interpretability and predictive performance is still rather large, as highly interpretable models such as LR are required by law in many countries.

5.2. Influence of omitting low-mileage drivers

As introduced in Appendix A.1, drivers with less than 1500 km per year were excluded from building the models discussed so far. If these low-mileage drivers are also included, the overall model performance of all classifiers will increase significantly, up to 0.833 in terms of AUC (Table 12). However, the influence of geographical context features on the model performance is significantly reduced, as adding all features only leads to an increase of 0.027, compared to the baseline for XGBoost, due to the increased importance of mileage. A further sensitivity analysis using different minimal mileage thresholds also shows that prediction performance increases with larger thresholds (Appendix H).

The study hereby also showed that the chosen minimum driving distance matters. If low-mileage drivers are not removed, model performance increases significantly. However, the performance stays roughly the same between the baseline feature set and feature sets with geographical context features.

5.3. Limitations

There are a few limitations to this study that need to be pointed out, mainly regarding data availability. One of the biggest limitations stems from the fact that there was no information available on whether the driver was actually at fault for an accident. This is different from previous studies, which usually had information from the insurance company about whether an at-fault claim was made. Therefore, some safe drivers who do not exhibit any typical dangerous behavior are labeled as accident drivers through circumstances beyond their control. Logically, these drivers are very hard to classify as accident drivers since they do not exhibit any dangerous driving patterns.

Another data-related limitation of this study lies in the relatively low spatial resolution, which was resampled by the data provider to one waypoint per 2 km in driving. This makes potential map matching very difficult and inaccurate, especially for urban areas. Incorporating more information about the road network, such as various centrality measures or average traffic volumes, are further variables that could be included if higher resolution data was available.

Furthermore, it could be possible to achieve higher performance with more detailed geographical information. For example, real weather data might also yield better results than the historical forecast data. Also, the driving events could have been explored in more detail according to their acceleration values to distinguish between events with different levels of severity. In addition, the total driven time could have been used instead of distance as the main exposure factor.

From a modeling perspective, it is possible that different neural network architectures or parameters might yield better results. In addition, the model performance under different ratios of accident versus accident-free drivers and different classification thresholds could have been explored because this ratio can be highly imbalanced in other real-world applications. It also needs to be noted that the model performance figures of this study are not really comparable to other studies. This is due to different recording types, different frameworks for registering a claim or accident, different class balances, etc.

Lastly, it should be pointed out that, as with all person-related GPS tracking applications, privacy concerns exist within the context of behavior-based car insurance. However, the discussion of those goes beyond the scope of this study (and has also been addressed on the technical level by the waypoint downsampling to 2 km performed by the data provider).

Table 12
The average AUC, accuracy, and F1-score of the nested cross-validation per model and feature set combination with short-mileage drivers included.

Model	A	B (baseline)	C	D	E	F
AUC						
LR	0.796	0.780	0.801	0.801	0.807	0.807
FFNN	0.796	0.793	0.800	0.793	0.820	0.818
RF	0.797	0.805	0.808	0.807	0.825	0.824
XGBoost	0.795	0.805	0.804	0.806	0.831	0.833
Accuracy						
LR	0.751	0.752	0.753	0.752	0.755	0.755
FFNN	0.752	0.754	0.752	0.752	0.751	0.749
RF	0.751	0.757	0.757	0.755	0.758	0.756
XGBoost	0.753	0.755	0.755	0.754	0.759	0.763
F1-Score						
LR	0.786	0.786	0.786	0.786	0.788	0.788
FFNN	0.788	0.790	0.787	0.786	0.780	0.772
RF	0.787	0.791	0.793	0.790	0.792	0.791
XGBoost	0.790	0.792	0.791	0.792	0.786	0.790

6. Conclusions and future work

As the main contribution of this study, our results show that the inclusion of geographical information can improve the performance of machine learning models for modeling car accident risk, which has the potential to improve PHYD car insurance and its benefits. It should be noted that car accidents are to a large degree random events, which are very hard to predict. Even though the margins for improvement are thus generally narrow, a small improvement in accuracy can potentially translate into a big impact in an insurance context. Not to mention the value of human health that could potentially be improved through the incentive of safer and more ecological driving, which PHYD car insurance schemes provide. Given this, it is very important to include geographical context when modeling risk drivers in UBI policies.

Further research about addressing the above mentioned limitations and including more detailed geographical information into car accident risk modeling are recommended. Regional differences between Rome and Tuscany may also suggest localized influence for certain factors. It would further be beneficial to have demographic data of the drivers, such as age, gender, and driving experience, to allow comparison or combination with traditional car insurance risk models, which however is more sensitive to privacy issues.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 780754.

Disclosure statement

No potential conflict of interest was reported by the authors.

Author statement

Livio Brühwiler: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft Preparation, Visualization.

Cheng Fu: Conceptualization, Methodology, Writing - Original Draft Preparation, Writing - Review & Editing, Visualization, Supervision.

Haosheng Huang: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Leonardo Longhi: Conceptualization, Resources, Data Curation, Project administration.

Robert Weibel: Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Declration of Competing Interest

None.

Acknowledgments

The authors also appreciate the comments of the anonymous reviewers which helped improve the paper.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2022.101760.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *, Vol. 101. Proceedings of the 12th USENIX conference on operating systems design and implementation* (pp. 265–283). USA: USENIX Association. https://doi.org/10.5555/3026877.3026899.

Af Wåhlberg, A. E. (2004). The stability of driver acceleration behavior, and a replication of its relation to bus accidents. *Accident Analysis and Prevention, 36*(1), 83–92. https://doi.org/10.1016/S0001-4575(02)00130-6

Alkahtani, K. F., Abdel-Aty, M., & Lee, J. (2019). A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. *International Journal of Sustainable Transportation, 13*(4), 255–267. https://doi.org/10.1080/15568318.2018.1463417

Andrey, J., & Yagar, S. (1993). A temporal analysis of rain-related crash risk. *Accident Analysis and Prevention, 25*(4), 465–472. https://doi.org/10.1016/0001-4575(93)90076-9

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation, 46*(3), 735–752. https://doi.org/10.1007/s11116-018-9890-7

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. *Transportation Res. Part C: Emerg. Technol., 68*, 160–167. https://doi.org/10.1016/j.trc.2016.04.004

Ayuso, M., Guillen, M., & Pérez-Marín, A. (2016). Telematics and gender discrimination: Some usage-based evidence on whether Men's risk of accidents differs from Women's. *Risks, 4*(2), 10. https://doi.org/10.3390/risks4020010

Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention, 73*, 125–131. https://doi.org/10.1016/j.aap.2014.08.017

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

Bergel-Hayat, R., Debbarh, M., Antoniou, C., & Yannis, G. (2013). Explaining the road accident risk: Weather effects. *Accident Analysis and Prevention, 60*, 456–465. https://doi.org/10.1016/j.aap.2013.03.006

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. https://doi.org/10.1016/j.tra.2017.10.018

Black, A. W., & Mote, T. L. (2015). Effects of winter precipitation on automobile collisions, injuries, and fatalities in the United States. *Journal of Transport Geography, 48*, 165–175. https://doi.org/10.1016/j.jtrangeo.2015.09.007

Bordoff, J. E., & Noel, P. (2008). *Pay-As-You-drive auto insurance: A simple way to reduce driving-related harms and increase equity. the journal of risk and insurance (Vol. 37). Washington, D.C.* Brookings Institution. Retrieved from http://www.brookings.edu/~/media/Files/rc/papers/2008/07_payd_bordoffnoel/07_payd_bordoffnoel.pdf.

Boucher, J.-P., Peârez-Marôân, A. M., & Santolino, M. (2013). Pay-as-you-drive insurance: The effect of the kilometers on the risk of accident. *Anales Del Instituto De Actuarios EspanÄoles, 19*, 135–154.

Boucher, J. P., & Turcotte, R. (2020). A longitudinal analysis of the impact of distance driven on the probability of car accidents. *Risks, 8*(3), 1–19. https://doi.org/10.3390/risks8030091

Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis and Prevention, 39*(4), 657–670. https://doi.org/10.1016/j.aap.2006.10.012

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107. https://doi.org/10.5555/1756006.1859921

Chen, T., & Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (Vol. Vols. 13-17-Augu, pp. 785–794)*. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2016). Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1–6). IEEE. https://doi.org/10.1109/DASC.2016.7777956.

Cura, A., Kucuk, H., Ergen, E., & Oksuzoglu, I. B. (2020). Driver profiling using long short term memory (LSTM) and convolutional neural network (CNN) methods. *IEEE Transactions on Intelligent Transportation Systems, 1–11*. https://doi.org/10.1109/TITS.2020.2995722

Denuit, M., Marchal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial modelling of claim counts. actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems.* Chichester, UK: John Wiley & Sons, Ltd.. https://doi.org/10.1002/9780470517420

Eboli, L., Mazzulla, G., & Pungillo, G. (2017). How to define the accident risk level of car drivers by combining objective and subjective measures of driving style. *Transportation Research Part F: Traffic Psychology and Behaviour, 49*, 29–38. https://doi.org/10.1016/j.trf.2017.06.004

Eisenberg, D., & Warner, K. E. (2005). Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. *American Journal of Public Health, 95*(1), 120–124. https://doi.org/10.2105/AJPH.2004.048926

Elbeltagi, A., Zhang, L., Deng, J., Juma, A., & Wang, K. (2020). Modeling monthly crop coefficients of maize based on limited meteorological data: A case study in Nile Delta, Egypt. *Computers and Electronics in Agriculture, 173*(April), Article 105368. https://doi.org/10.1016/j.compag.2020.105368

Fountas, G., Fonzone, A., Gharavi, N., & Rye, T. (2020). The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. *Analytic Methods in Accident Research, 27*, Article 100124. https://doi.org/10.1016/j.amar.2020.100124

Giulietti, C., Tonin, M., & Vlassopoulos, M. (2020). When the market drives you crazy: Stock market returns and fatal car accidents. *Journal of Health Economics, 70*, Article 102245. https://doi.org/10.1016/j.jhealeco.2019.102245

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis, 39*(3), 662–672. https://doi.org/10.1111/risa.13172

Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. *J. Traffic and Transport. Eng. (English Edition)*.. https://doi.org/10.1016/j.jtte.2020.05.002

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, Y., & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems, 127*(September), Article 113156. https://doi.org/10.1016/j.dss.2019.113156

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. In *, Vol. 100*. *Procedia engineering* (pp. 816–825). Elsevier Ltd.. https://doi.org/10.1016/j.proeng.2015.01.436

Janke, M. K. (1991). Accidents, mileage, and the exaggeration of risk. *Accident Analysis and Prevention, 23*(2–3), 183–188. https://doi.org/10.1016/0001-4575(91)90048-A

Jia, R., Khadka, A., & Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis and Prevention, 121*, 223–230. https://doi.org/10.1016/j.aap.2018.09.018

Kantor, S., & Stárek, T. (2014). Design of Algorithms for payment telematics systems evaluating Driver's driving style. *Transactions on Transport Sciences, 7*(1), 9–16. https://doi.org/10.2478/v10158-012-0049-5

Kim, K., & Yamashita, E. (2002). Motor vehicle crashes and land use empirical analysis from Hawaii. *Transportation Research Record, 1784*, 73–79. https://doi.org/10.3141/1784-10

Kufera, J. A., Al-Hadidi, A., Knopp, D. G., Dezman, Z. D. W., Kerns, T. J., Okedele, O. E., & Tracy, J. K. (2020). The impact of a new casino on the motor vehicle crash patterns in suburban Maryland. *Accident Analysis & Prevention, 142*, Article 105554. https://doi.org/10.1016/j.aap.2020.105554

Langford, J., Koppel, S., McCarthy, D., & Srinivasan, S. (2008). In defence of the "low-mileage bias.". *Accident Analysis and Prevention, 40*(6), 1996–1999. https://doi.org/10.1016/j.aap.2008.08.027

Lee, J., Chae, J., Yoon, T., & Yang, H. (2018). Traffic accident severity analysis with rain-related factors using structural equation modeling – A case study of Seoul City. *Accident Analysis and Prevention, 112*, 1–10. https://doi.org/10.1016/j.aap.2017.12.013

Lemaire, J., Park, S. C., & Wang, K. C. (2015). The use of annual mileage as a rating variable. *ASTIN Bulletin, 46*(1), 39–69. https://doi.org/10.1017/asb.2015.25

Litman, T. (2005). Pay-as-you-drive pricing and insurance regulatory objectives. *Journal of Insurance Regulation, 23*(3), 35.

Litman, T. (2011). Distance-based vehicle insurance feasibility, costs and benefits: comprehensive technical report. Retrieved from https://trid.trb.org/view/1549618.

Lym, Y., & Chen, Z. (2020). Does space influence on the frequency and severity of the distraction-affected vehicle crashes? An empirical evidence from the Central Ohio. *Accident Analysis and Prevention, 144*, Article 105606. https://doi.org/10.1016/j.aap.2020.105606

Ma, Y.-L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice, 113*, 243–258. https://doi.org/10.1016/j.tra.2018.04.013

Malyshkina, N. V., Mannering, F. L., & Tarko, A. P. (2009). Markov switching negative binomial models: An application to vehicle accident frequencies. *Accident Analysis & Prevention, 41*(2), 217–226. https://doi.org/10.1016/j.aap.2008.11.001

Mayou, R., Bryant, B., & Duthie, R. (1993). Psychiatric consequences of road traffic accidents. *British Medical Journal, 307*(6905), 647–651. https://doi.org/10.1136/bmj.307.6905.647

de Medrano, R., & Aznarte, J. L. (2021). A new Spatio-temporal neural network approach for traffic accident forecasting. *Applied Artificial Intelligence, 35*(10), 782–801. https://doi.org/10.1080/08839514.2021.1935588

Musicant, O., Bar-Gera, H., & Schechtman, E. (2010). Electronic records of undesirable driving events. *Transportation Research Part F: Traffic Psychology and Behaviour, 13*(2), 71–79. https://doi.org/10.1016/j.trf.2009.11.001

National Highway Traffic Safety Administration. (2018). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey, TRAFFIC SAFETY FACTS. Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506 Accessed: 28 January 2021.

Nees, M. A. (2019). Safer than the average human driver (who is less safe than me)? Examining a popular safety benchmark for self-driving cars. *Journal of Safety Research, 69*, 61–68. https://doi.org/10.1016/j.jsr.2019.02.002

Ng, K. S., Hung, W. T., & Wong, W. G. (2002). An algorithm for assessing the risk of traffic accident. *Journal of Safety Research, 33*(3), 387–410. https://doi.org/10.1016/S0022-4375(02)00033-6

Paefgen, J., Staake, T., & 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. https://doi.org/10.1016/j.tra.2013.11.010

Paefgen, J., Staake, T., & Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems, 56*(1), 192–201. https://doi.org/10.1016/j.dss.2013.06.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peng, Y., Jiang, Y., Lu, J., & Zou, Y. (2018). Examining the effect of adverse weather on road transportation using weather and traffic sensors. *PLoS One, 13*(10), Article e0205409. https://doi.org/10.1371/journal.pone.0205409

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks, 7*(2), 70. https://doi.org/10.3390/risks7020070

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. *Accident Analysis and Prevention, 115*, 160–169. https://doi.org/10.1016/j.aap.2018.03.005

Sugarman, S. D. (1994). "Pay at the pump" auto insurance: The vehicle injury plan (VIP) for better compensation, fairer funding, and greater safety. *Journal of Policy Analysis and Management, 13*(2), 363. https://doi.org/10.2307/3325018

Swiss Federal Office of Statistics. (2018). Strassenverkehrsunfälle | Bundesamt Für Statistik. https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/unfaelle-umweltauswirkungen/verkehrsunfaelle/strassenverkehr.html Accessed January 5, 2021.

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2016). Innovative insurance schemes: Pay as/how you drive. *Transportation Research Procedia, 14*, 362–371. https://doi.org/10.1016/j.trpro.2016.05.088

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 67*(5), 1275–1304. https://doi.org/10.1111/rssc.12283

Wan, Y., Li, Y., Liu, C., & Li, Z. (2020). Is traffic accident related to air pollution? A case report from an island of Taihu Lake, China. *Atmospheric Pollution Research, 11*(5), 1028–1033. https://doi.org/10.1016/j.apr.2020.02.018

Wang, B., Lin, Y., Guo, S., & Wan, H. (2021). GSNet : Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(5), 4402–4409.

Wang, J., Wu, J., & Li, Y. (2015). The driving safety field based on driver–vehicle–road interactions. *IEEE Transactions on Intelligent Transportation Systems, 16*(4), 2203–2214. https://doi.org/10.1109/TITS.2015.2401837

Wang, Y., & Kockelman, K. M. (2013). A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis and Prevention, 60*, 71–84. https://doi.org/10.1016/j.aap.2013.07.030

WHO. (2020). *Road traffic injuries*. Retrieved March 22, 2020, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis and Prevention, 41*(1), 137–145. https://doi.org/10.1016/j.aap.2008.10.001

Winlaw, M., Steiner, S. H., MacKay, R. J., & Hilal, A. R. (2019). Using telematics data to find risky driver behaviour. *Accident Analysis and Prevention, 131*, 131–136. https://doi.org/10.1016/j.aap.2019.06.003

Worldweatheronline. (2020). *Weather API | JSON | World Weather Online*. Retrieved from https://www.worldweatheronline.com/developer/.

Wyon, D. P., Wyon, I., & Norin, F. (1996). Effects of moderate heat stress on driver vigilance in a moving vehicle. *Ergonomics, 39*(1), 61–75. https://doi.org/10.1080/00140139608964434

Yan, C., Wang, X., Liu, X., Liu, W., & Liu, J. (2020). Research on the UBI Car insurance rate determination model based on the CNN-HVSVM algorithm. *IEEE Access, 8*, 160762–160773. https://doi.org/10.1109/ACCESS.2020.3021062

Yang, B. Z., & Loo, B. P. (2016). Land use and traffic collisions: A link-attribute analysis using empirical Bayes method. *Accident Analysis & Prevention, 95*, 236–249. https://doi.org/10.1016/j.aap.2016.07.002

Yannis, G., & Karlaftis, M. G. (2011). Weather effects on daily traffic accidents and fatalities: A time series count data approach. In *Proceedings of the 89th annual meeting of the transportation research board (p. 10)*.

Yao, S., Wang, J., Fang, L., & Wu, J. (2018). Identification of vehicle-pedestrian collision hotspots at the Micro-level using network kernel density estimation and random forests: A case study in Shanghai, China. *Sustainability, 10*(12), 4762. https://doi.org/10.3390/su10124762

Zhang, H., Xu, L., Cheng, X., Chen, W., & Zhao, X. (2018). Big data research on driving behavior model and auto insurance pricing factors based on UBI. In S. Sun, N. Chen, & T. Tian (Eds.), *International conference on signal and information processing, networking and computers* (pp. 404–411). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-7521-6_49.