A Data-Driven Approach for Customized Pay-As-You-Drive Insurance Premiums

Rahmat Ullah
CEMET, University of South Wales
Pontypridd, United Kingdom
Rahmat.ullah@southwales.ac.uk

Ikram Asghar
CEMET, University of South Wales
Pontypridd, United Kingdom
ikram.asghar@southwales.ac.uk

Mark G Griffiths
CEMET, University of South Wales
Pontypridd, United Kingdom
Mark.griffiths@southwales.ac.uk

Seamus Ballard-Smith
CEMET, University of South Wales
Pontypridd, United Kingdom
Seamus.ballard-smith@southwales.ac.uk

Nestor Alonso
Driverly Company Limited
Cardiff, United Kingdom
nestor@driverly-insurance.co.uk

***Abstract*—Insurance companies have recently started to adopt usage-based insurance policies to provide insurance premiums according to the customer's driving behavior. Risk models are proposed based on the driver's behavior, driving history, and telematics data. These models do not consider exogenous factors, such as the geographical context, weather information, and driving events. This paper presents a data-driven approach for personalized premiums for car insurance by integrating endogenous and exogenous factors to calculate journey risk. The proposed algorithm uses accident, casualty, and vehicle data sets to calculate a baseline risk value for a given route. A final risk value is calculated by adding weather and journey-specific risk factors. The algorithm creates Voronoi regions to associate weather observations with geographical locations and analyses accident data to assign risk values to individual junctions and roads. The overall journey risk is calculated based on weather and accident risks, refined using metrics for the time of route or accident and journey purpose. The algorithm provides a simple and flexible way of calculating journey risk, considering multiple factors that can be used to provide truly personalized insurance premiums. The proposed approach has the potential to benefit insurance providers, regulators, and drivers by improving the accuracy of risk assessment and promoting safer journeys.**

***Index Terms*—Exogenous factors, usage-based insurance, personalized insurance premiums, risk model, geographical context**

I. Introduction

Risk transfer is a significant factor to consider within the insurance industry. The car insurance industry has been trying to classify drivers based on their driving style and into different risk levels. It is crucial that the insurance price accurately reflect the policyholder's actual risk to avoid financial instability and fair insurance premiums. Historically, traditional risk factors for auto insurance have included personal qualities, claim history, and the vehicle's physical attributes [1]. It has been observed, however, that these indicators do not always have a direct causal relationship with the actual driving risk, resulting in mistakes in the premium calculation [2].

European Regional Development Fund and Welsh Government.

Vehicle insurance companies have begun to innovate to provide personalized insurance to their customers. With the advent of technology, these companies use many factors to understand individuals' driving habits and tailor their insurance policies accordingly. One such factor that has recently garnered significant attention is using a driver's mobile phone. The data collected from the driver's mobile phone is analyzed to understand the driving style better. This information identifies areas where drivers can improve their driving behavior, leading to safer roads for everyone. This approach to personalizing insurance premiums also lowers premiums for drivers with safe and responsible driving habits.

The development of in-vehicle networks and telematics technologies have increased the appeal of usage-based insurance (UBI) products for insurers and consumers. Initially, UBI products were designed to price insurance based solely on the usage of each vehicle. However, it has now expanded to encompass a general type of insurance, where the price is determined by the daily driving habits of each driver, as collected through telematics data [3]. Compared to traditional risk factors, the stronger correlation between driving behavior variables and traffic accidents is leveraged to improve the accuracy of insurance pricing. The dynamic premiums of UBI products also incentivize policyholders to reduce their expenses by driving less or improving their driving habits. The literature has highlighted the positive impact of UBI in addressing social issues such as traffic congestion, energy consumption, and air pollution [4].

As of 2021, 32 million cars were registered in the United Kingdom (UK). There has been an increase of 39.6% over the past 25 years. A total of 2.1 million vehicles were registered for the first time in 2020 [5]. The total number of fatal road accidents in the UK was 1,460 and 1,608 in 2020 and 2021, respectively. There were a total of 26,701 seriously injured in these two years [6]. The insurance sector is essential to the economy because it collects premiums, makes investments, and takes on personal and commercial risks. Due to technological improvements and data analysis, the vehicle insurance business is fast changing, giving rise to usage-based insurance (UBI) with two models: "pay as you drive" and "pay how you drive" [7]. Premiums are calculated based on usage time, driving distance, driving behavior, and location. UBI represents a shift from traditional insurance that rewards safe drivers with lower premiums and a no-claims bonus based on historical behavior patterns.

The driving environment also plays a crucial role in affecting drivers' behavior. Several studies have explored this relationship and the impact of various road conditions and geographical factors on driving habits [8]. In addition to the driving environment, the road's infrastructure and characteristics also affect the driver's behavior. Adverse weather conditions can substantially impact driver behavior, posing a risk to the driver and other vehicles on the road. Research has shown that bad weather conditions can increase accident frequency and severity [9]. Rainfall, fog, and winter precipitation, such as snowfall and ice pellets, have all been found to have a direct relationship with higher crash frequency. Extreme high or low temperatures can also increase the probability of car accidents. When bad weather combines with poor lighting conditions, such as night driving in the rain, the risk of driving errors, hazardous driving, and resulting accidents are further amplified [10]. However, there are also exceptional cases where rainfall can lead to lower crash frequencies, possibly due to adapted driving behavior or different exposure levels.

This work presents a novel data-driven approach for generating personalized insurance premiums for drivers in the UK. A data-driven approach is proposed that integrates various factors, including driving behavior data, real-time weather conditions, and the characteristics of major roads, to generate a risk profile for each driver. The proposed approach differs from traditional insurance models that rely on driving style, telemetries data, and generalized risk assessments to determine insurance premiums.

II. Related Work

Pay-as-you-drive (PAYD) policy premiums are based on a driver's driving. The more drivers drive, the higher their premium is expected because they are exposed to a higher risk of accidents. In addition to distance, PAYD policies may also consider the time of day that a driver is on the road and the location of their driving [11]. This allows insurers to more accurately assess the risk associated with insuring a particular driver and to charge premiums that reflect this risk. By considering these exposure-based factors, PAYD policies can potentially incentivize safer driving behaviors, as drivers driving less or driving during safer times of the day may pay lower premiums.

Various methods are proposed based on driving behavior and usage of the car. Several methods have been proposed for driving style analysis to evaluate the driving styles. A Driving simulator was used to measure different factors, such as acceleration, braking, vehicle velocity, and distance from the car in front [12]. The results indicated that the identification rate was 81% for 12 drivers and 73% for 30 drivers. Similarly, different sensors were used to build the driving profile by observing braking and turning events [11]. A new approach to evaluate driver aggressiveness in "pay-how-you-drive" insurance schemes was proposed using cluster analysis and unsupervised machine learning techniques [7]. The study assesses the driver's aggressiveness based on the type of road travelled (urban or highway) and proposes a driver-related risk index. The method uses features extracted from different vehicles driving on urban and highway roads in Italy.

A set of evaluation indices was proposed in [13], later visualized using a hexagonal eye diagram to display the driver's driving style in China [14]. The diving styles of young and middle-aged drivers were studied, and speed limit devices were explored for car rental companies [15]. Another study classified driving styles into low-risk, moderate-risk, and high-risk categories, using operational driving pictures and artificial intelligence algorithms [16]. In terms of driving risk evaluation, a novel framework was proposed for reflecting driving risk levels based on multiple-kernel learning and achieved good accuracy with laboratory driving data [17].

The potential benefits of incorporating geographical context features in usage-based car insurance policies have been explored in [18]. The study evaluates the predictive performances of five machine learning classifiers on the usage records of over 8000 vehicles in Italy. The results indicated that including geographical information significantly improves predictive performance, with XGBoost yielding the best results. These findings suggest that incorporating geographical context in risk modelling can improve the accuracy and fairness of usage-based car insurance policies. A comprehensive review of the driving style evaluation methods and the design of commercial vehicle UBI products can be found in [19].

The literature review above shows that even if there are many related works on driving style and driving risk analysis, there is a lack of a data-driven method that includes weather and road profiling for insurance purposes.

III. Methodology

This section presents a comprehensive overview of the proposed data-driven approach that vehicle insurance companies could use to provide truly personalised insurance premiums.

To calculate the risk profile, two main dimensions are considered.

A. Endogenous Dimension

It refers to the internal factors that contribute to the risk associated with a particular route. This dimension is divided into two sub-dimensions: the driver and the vehicle.

The Driver sub-dimension encompasses various factors related to the driving style. These factors include the driver's age, driving proficiency, years of experience, and prior driving history, including any accidents or convictions. Monitoring a customer's driving style through their mobile phone is a prevalent solution in the current market. Acceleration and deceleration events and g-loads are recorded to categorize the individual's driving style. It is posited that those who exhibit dangerous driving behavior, such as excessive speeding or abrupt braking and turning, should incur higher insurance premiums than those who exhibit safer driving habits. It is important to note that a single instance of dangerous driving behavior is not considered a defining factor of a driver's overall style, as external circumstances may occasionally necessitate such actions. Instead, the cumulative evaluation of the individual's driving behaviour over time determines their driving style and corresponding insurance premiums.

The Vehicle sub-dimension encompasses factors related to the physical attributes and maintenance of the vehicle. These factors include the age of the vehicle, its value, engine power, and the level of maintenance it has received. For this dimension, the data or metrics for assessment are easy to create. In addition to the driving style of the individual, the type of vehicle being used also plays a role in determining the inherent risks associated with any journey. This is determined through the analysis of historical data, including the frequency of collisions involving certain types, makes, and models of vehicles, and MOT data, which reflects the vehicle's current condition. A vehicle with a higher frequency of collisions and a state of disrepair presents a greater inherent risk than a vehicle with a lower frequency of collisions and in a well-maintained condition. This information is factored into determining personalized insurance premiums, allowing insurance companies to offer their customers more accurate and relevant coverage.

B. Exogenous Dimension

It refers to external factors that impact the risk associated with a particular route. These factors are not dependent on the driver or the vehicle but relate to external conditions and circumstances. Some exogenous factors that can impact the risk associated with a route include weather conditions, records of accidents along the route, traffic density, and visibility. Considering these external factors when evaluating the risk associated with a given route is crucial. All these factors are associated with the route of the journey.

The choice of the route taken during a journey can have a significant impact on the inherent risks associated with the trip. Different roads and routes may have varying levels of traffic, road conditions, and environmental factors that can affect driver behavior and increase the likelihood of accidents. In addition, specific hot spots or locations along the road network may have a higher incidence of collisions or accidents, which can significantly increase the overall risk of a journey.

Weather conditions also have a significant impact on journey risk and are taken into consideration when determining personalized insurance premiums. For instance, bad weather conditions such as rainfall, fog, snowfall, freezing rain, and ice pellets have increased the frequency and severity of car accidents. Extreme high or low temperatures can also increase the probability of accidents. Additionally, driving in bad lighting conditions, such as, at night in the rain, can further amplify the likelihood of driving errors, hazardous driving, and resulting in accidents.

C. Data Used

To provide accurate risk evaluation and coverage to customers, it is essential to consider data from multiple sources, including identifying hotspots and weather conditions along the route. For this purpose, data for both endogenous and exogenous dimensions is considered. A simple scale matrix was used to calculate a risk score concerning endogenous factors such as age and experience, while exogenous factors were evaluated by accessing multiple APIs. Three major APIs were utilized to obtain relevant data, including the Met Office for weather updates, the MOT database for vehicle data, and the ONS for a map of the UK coastline. Three different datasets were used, which contained information about road accidents in the UK over the last five years, focusing specifically on accidents, people, and vehicles.

Figure 1 shows the accident data used in the algorithm development. As seen in the figure, the majority of accidents involve male drivers. However, It is important to mention that this data was normalized based on the population.

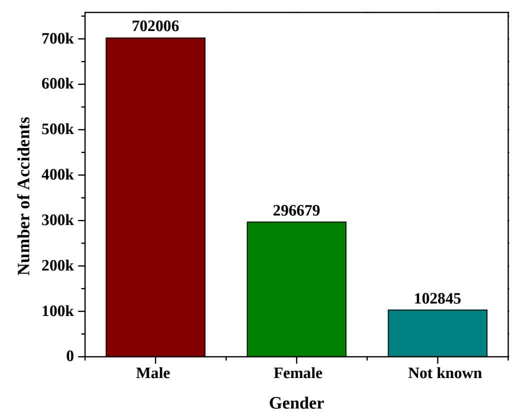

Fig. 1: Male/female driver split of the accident data.

Different metrics were used to calculate a risk score for the development of the algorithm. The metrics include driver age, driver Index of Multiple Deprivation decile, and the vehicle driven. Risk values were defined in a range for each of these metrics. For instance, it was assumed that a driver's likelihood of being involved in an accident is lower on a fair day with clear visibility compared to a day with heavy hail and low visibility. Similarly, the likelihood of accidents occurring is lower during off-peak hours when traffic density is lower. For the vehicle driven, the MOT API was used to retrieve data. The data retrieved include the type of car, test date, results, and comments.

IV. PROPOSED ALGORITHM

The proposed algorithm integrates endogenous and exogenous factors to provide a data-driven approach for personalized premiums. Endogenous factors, previously discussed, are considered in the calculation of risk values. An algorithm is developed to evaluate the risk associated with a route taken through the road network and return a risk value that can be compared against a baseline. The risk value does not require units or to be derived from specific probabilities, compare journeys, or communicate relative risk levels.

The three main data sets, accident, casualty, and vehicle, were integrated into a single data set for ease of use. The risk values for a specific route were calculated by retrieving data from OpenStreetMaps, the Met Office and data from ONS in the form of a perimeter map. The initial stage in constructing a visualization of the UK's road network is to acquire the geometry of the country. The UK government provides various map data options; only the boundary was utilized for this project. The geometry of the UK coastline was obtained through a GET request to a data source hosted by the UK government.

Similar to the UK map data, the UK weather data is readily available from the MET office and was obtained from weather stations across the UK, providing hourly readings over the past 24 hours. An API key was used to connect to a user account to access the data, which includes a list of all data from all weather stations. To attribute the weather data to all locations in the UK, the multipolygon was split into sub-regions associated with points within that polygon, creating Voronoi regions. The weather data is then processed by assigning a risk factor to the weather type category assigned by the Met Office. To create the Voronoi regions, the boundary of the UK was first represented as a GeoDataFrame using the shapely multipolygon. The latitude and longitude values of the weather stations were then converted to GeoDataFrame points. The boundary and weather station points were converted into a standard coordinate reference system. The resulting Voronoi regions are displayed in Figure 2.

Figure 2 displays the perimeter map of the UK coastline, the weather stations (blue dots), and the generated Voronoi regions (green lines). Areas outlined in red are regions that do not contain a station or were not considered Voronoi regions for other reasons. For example, a sample road network of the Cardiff Voronoi region was created. The accident data was interrogated to assign risk to individual junctions (nodes) and roads (edges), and accidents were assigned to their nearest point, as shown in Figure. 3.

Furthermore, a random route in a road network was generated with specific start and end points. The NetworkX library in Python was used to find the shortest path between the start and endpoints. The generated route was represented as a list of nodes and edges, as shown in Figure 4. The accident count data is then retrieved using the nodes and edges in the generated route. The implementation provides a simple and efficient way of generating a random route in a given road network and an example of analyzing accident count data for a specific route.

Finally, the weather risk is added to calculate the final risk value. The algorithm calculates the overall journey risk based on weather and accident risks. The weather risk is computed from weather observation data, while the accident risk is computed as the average of incidents in the past five years. Additionally, the algorithm implements two metrics for classifying risk factors: time of route/accident and journey purpose. These metrics are used to refine the overall journey risk calculation further. Overall, the algorithm provides a concise and flexible way of calculating journey risk that could be used to provide truly personalized insurance premiums.

Fig. 2: Voronoi regions (outlined in green) with corresponding weather stations (blue dots) as they are found in the UK.

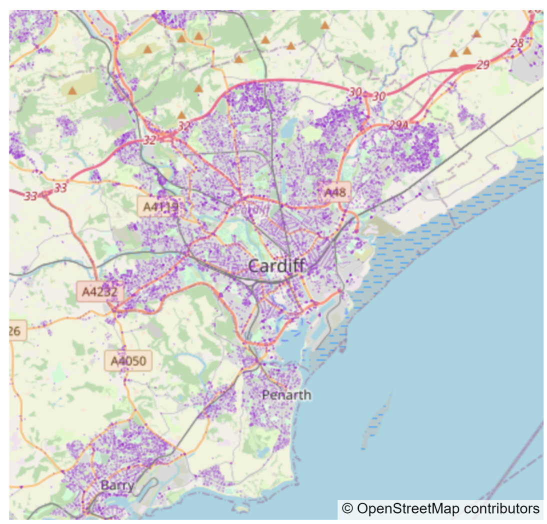

Fig. 3: The roads (edges) plotted as lines for the Cardiff region.

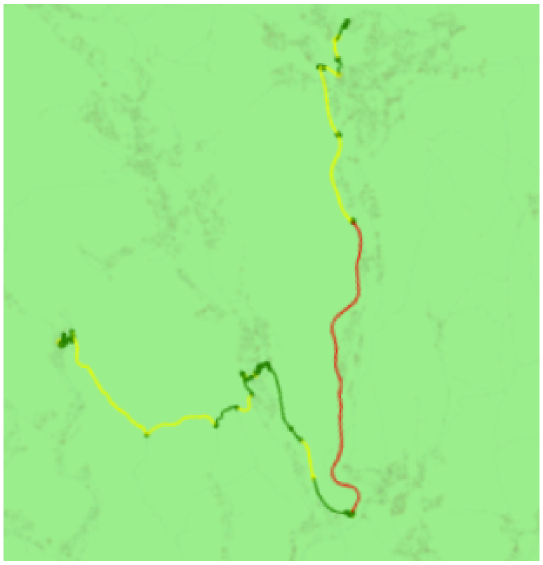

Fig. 4: A random route in the Cardiff region. The red colour represents the route with the most accidents in the past five years, followed by yellow and green.

V. CONCLUSION AND FUTURE WORK

This paper introduces an innovative approach for calculating personalized car insurance premiums based on journey risk by integrating endogenous and exogenous factors. The proposed algorithm creates Voronoi regions and analyzes accident data to provide a more accurate risk assessment than traditional methods. Using accident, casualty, vehicle, and weather data sets provides a comprehensive understanding of the risk associated with a specific route. The algorithm can refine risk values using various metrics for the time of route/accident and journey purpose, making it a valuable tool for decision-makers. Ultimately, the proposed approach can promote safe journeys and contribute to effective risk management, thereby reducing the number of road accidents.

Further research is recommended to explore the practical implications of the proposed algorithm, including its scalability and impact on insurance pricing. The regulatory and legal challenges related to privacy concerns and compliance with insurance regulations will also be addressed. Additionally, challenges related to customer acceptance and adoption of usage-based insurance policies before using this approach in real-world scenarios. The proposed approach has been evaluated on the road network in the Cardiff region. A database of the UK roads network will be developed to store relevant data against each node and edge, including hourly weather updates, to improve the algorithm's accuracy. Additional data for the road network, such as the number of lanes and speed limits, will also be obtained to explore further avenues for investigation and risk calculation.

ACKNOWLEDGMENT

The authors would like to acknowledge the European Regional Development Fund (ERDF) and the Welsh Government for funding this study.

REFERENCES

[1] M. Ayuso, M. Guillen, and J. P. Nielsen, "Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data," *Transportation*, vol. 46, pp. 735–752, 2019.

[2] Y. Huang and S. Meng, "Automobile insurance classification ratemaking based on telematics driving data," *Decision Support Systems*, vol. 127, p. 113156, 2019.

[3] S. Alfiero, E. Battisti, and E. Hadjielias, "Black box technology, usage-based insurance, and prediction of purchase behavior: Evidence from the auto insurance sector," *Technological Forecasting and Social Change*, vol. 183, p. 121896, 2022.

[4] O. Ghaffarpasand, M. Burke, L. K. Osei, H. Ursell, S. Chapman, and F. D. Pope, "Vehicle telematics for safer, cleaner and more sustainable urban transport: A review," *Sustainability*, vol. 14, no. 24, p. 16386, 2022.

[5] "Vehicles statistics - gov.uk," https://www.gov.uk/government/collections/vehicles-statistics, Dec 2022, (Accessed on 03/24/2023).

[6] "Uk collision and casualty statistics — brake," https://www.brake.org.uk/get-involved/take-action/mybrake/knowledge-centre/uk-road-safety, (Accessed on 03/24/2023).

[7] M. F. Carfora, F. Martinelli, F. Mercaldo, V. Nardone, A. Orlando, A. Santone, and G. Vaglini, "A "pay-how-you-drive" car insurance approach through cluster analysis," *Soft Computing*, vol. 23, pp. 2863–2875, 2019.

[8] A. Doshi and M. M. Trivedi, "Examining the impact of driving style on the predictability and responsiveness of the driver: Real-world and simulator analysis," in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 232–237.

[9] Y. Peng, Y. Jiang, J. Lu, and Y. Zou, "Examining the effect of adverse weather on road transportation using weather and traffic sensors," *PLoS one*, vol. 13, no. 10, p. e0205409, 2018.

[10] G. Fountas, A. Fonzone, N. Gharavi, and T. Rye, "The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents," *Analytic methods in accident research*, vol. 27, p. 100124, 2020.

[11] D. I. Tselentis, G. Yannis, and E. I. Vlahogianni, "Innovative insurance schemes: pay as/how you drive," *Transportation Research Procedia*, vol. 14, pp. 362–371, 2016.

[12] M. L. Bernardi, M. Cimitile, F. Martinelli, and F. Mercaldo, "Driver and path detection through time-series classification," *Journal of Advanced Transportation*, vol. 2018, 2018.

[13] W. Zheng, W. Nai, F. Zhang, W. Qin, and D. Dong, "A novel set of driving style risk evaluation index system for ubi-based differentiated commercial vehicle insurance in china," in *CICTP 2015*, 2015, pp. 2510–2524.

[14] W. Nai, Y. Chen, Y. Yu, F. Zhang, D. Dong, and W. Zheng, "Effective presenting method for different driving styles based on hexagonal eye diagram applied in pay-how-you-drive vehicle insurance," in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*. IEEE, 2016, pp. 1–6.

[15] H. Kim, D. Yoon, H. Shin, and C. H. Park, "Driving characteristics analysis of young and middle-aged drivers," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2016, pp. 864–867.

[16] G. Li, F. Zhu, X. Qu, B. Cheng, S. Li, and P. Green, "Driving style classification based on driving operational pictures," *IEEE Access*, vol. 7, pp. 90 180–90 189, 2019.

[17] J.-L. Yin and B.-H. Chen, "An advanced driver risk measurement system for usage-based insurance on big driving data," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 4, pp. 585–594, 2018.

[18] L. Brühwiler, C. Fu, H. Huang, L. Longhi, and R. Weibel, "Predicting individuals' car accident risk by trajectory, driving events, and geographical context," *Computers, Environment and Urban Systems*, vol. 93, p. 101760, 2022.

[19] W. Nai, Z. Yang, Y. Wei, J. Sang, J. Wang, Z. Wang, and P. Mo, "A comprehensive review of driving style evaluation approaches and product designs applied to vehicle usage-based insurance," *Sustainability*, vol. 14, no. 13, p. 7705, 2022.