Noname manuscript No.
(will be inserted by the editor)

A "Pay How You Drive" Car Insurance Approach through Cluster Analysis

Maria Francesca Carfora · Fabio Martinelli · Francesco Mercaldo · Vittoria Nardone · Albina Orlando · Antonella Santone · Gigliola Vaglini

the date of receipt and acceptance should be inserted later

Abstract As discussed in the recent literature, several innovative car insurance concepts are proposed in order to gain advantages both for insurance companies and for drivers. In this context, the "pay how you drive" paradigm is emerging, but it is not thoroughly discussed and much less implemented. In this paper we propose an approach in order to identify the driver behaviour exploring the usage of unsupervised machine learning techniques. A real world case study is performed to evaluate the effectiveness of the proposed solution. Furthermore, we discuss how the proposed model can be adopted as risk indicator for car insurance companies.

M. F. Carfora
Istituto per le Applicazioni del Calcolo "M. Picone", Consiglio Nazionale delle Ricerche, Napoli, Italy
E-mail: f.carfora@iac.cnr.it

F. Martinelli
Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Pisa, Italy
E-mail: fabio.martinelli@iit.cnr.it

F. Mercaldo
Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Pisa, Italy
E-mail: francesco.mercaldo@iit.cnr.it

V. Nardone
Department of Engineering, University of Sannio, Benevento, Italy
E-mail: vnardone@unisannio.it

A. Orlando
Istituto per le Applicazioni del Calcolo "M. Picone", Consiglio Nazionale delle Ricerche, Napoli, Italy
E-mail: a.orlando@iac.cnr.it

A. Santone
Department of Bioscience and Territory, University of Molise, Pesche (IS), Italy
E-mail: antonella.santone@unimol.it

G. Vaglini
Department of Information Engineering, University of Pisa, Pisa, Italy
E-mail: gigliola.vaglini@unipi.it

Keywords insurance, risk analysis, OBD, CAN, cluster analysis, machine learning

1 Introduction

As of 2015, there were over 263 million registered vehicles on the roads in the United States. Of those millions of registered vehicles, each year there are also millions of vehicle crashes. In 2015, there were 32,166 fatalities, 1,715,000 injuries and 4,548,000 car crashes which involved property damage. Of these fatalities, there are far more driver deaths, than passenger, pedestrian or motorcyclist deaths[1]. Therefore the statistics indicate the importance of automobile insurance and in most cases, auto insurance is required by law. Car insurance is really important because not only it covers any physical damage that may occur in an accident, but also any damage or injury that might be caused because of a vehicular accident or which may be done upon oneself or ones vehicle by another vehicle or accident, as a falling tree for example [1].

The insurance industry is a key component of the economy by virtue of the amount of premiums it collects, the scale of its investments and, more fundamentally, the essential social and economic role it plays by covering personal and business risks.

Auto insurance markets are changing rapidly. As technology has evolved and as the price of data has fallen, rates can now be produced through millions of variables in a multivariate analysis. Through telematics, risks can be rated on an individual basis; an insurer can now identify, measure and rate a particular person's driving ability. The Usage Based Insurance (UBI) concept was introduced into the personal motor insurance market over a decade ago. It consists of two typical models: "pay as you drive" (PAYD) and "pay how you drive" (PHYD). Premiums are based upon time of usage, distance driven, driving behavior and places driven to. In particular, in PHYD insurance premium is calculated based on how the vehicle is driven, in PAYD scheme insurance premium is calculated dynamically, according to the amount driven. PHYD is the more mature of the two offerings, giving more detailed data to insurers and costumers [2][2,3,4].

This represents a different approach with respect to traditional insurance, which attempts to differentiate and reward "safe" drivers, giving them lower premiums and/or a no-claims bonus. However, conventional differentiation is a reflection of historic rather than present patterns of behaviour. This means that it may take a long time before safer (or more reckless) patterns of driving and changes in lifestyle feed through into premiums.

UBI programs offer many advantages to insurers, consumers and society. Linking insurance premiums more closely to actual individual vehicle or fleet

[1] https://www.statista.com/topics/3087/car-insurance-in-the-united-states/

[2] http://www.ey.com/Publication/vwLUAssets/ey-introducing-pay-how-you-drive-insurance/$FILE/ey-introducing-pay-how-you-drive-insurance.pdf

performance allows insurers to price premiums more accurately. [5]. This increases affordability for lower-risk drivers, many of whom are also lower-income drivers. It also gives consumers the ability to control their premium costs by encouraging them to reduce miles driven and adopt safer driving habits. Fewer miles and safer driving also aid in reducing accidents, congestion, and vehicle emissions, which benefits society [3].

Starting from these considerations, in this paper we propose an approach able to characterize the driver behaviour using a set of features gathered from the vehicle CAN bus.

As a matter of fact, as demonstrated in the current literature, drivers typically exhibit different driving style on different kind of roads [6,7,8,9]. Basing on this evidence, the proposed method considers the unsupervised machine learning i.e., the machine learning task of inferring a function to describe hidden structure from unlabeled data, to discriminate between urban and highway roads. In order to perform this task, we consider cluster analysis in order to group the feature extracted from the driver under analysis: the main assumption that will be verified in the experiment is that CAN bus features gathered from the highway path exhibits different values from the ones gathered from urban road (and for this reason grouped in different clusters). Furthermore, on the basis of the cluster analysis results, we compute an aggressiveness index of the driver under analysis in order to propose a "pay how you drive" possible risk assessment calculation.

We evaluate the proposed approach on a real-world dataset gathered from a vehicle running through several (urban and highway) roads.

The reminder of the paper is organized as follows: Section 2 discusses the current literature, Section 3 introduces the method, Section 4 illustrates the results of the cluster analysis based experiment, Section 5 describe a possible risk index computation. Finally, conclusions and future works are given in Section 6.

2 Related Work

In the following section we review the current literature related to the driving style recognition. We discuss the approaches that involve driving data analysis, i.e., methods connecting driver behavior with car-related feature. These methods are different from the ones that identify the driver behavior by the usage of smartphone sensors, video motion or questionnaire. Furthermore, we discuss also current literature about risk assessment.

[3] http://www.naic.org/cipr_topics/topic_usage_based_insurance.htm

2.1 Approaches using car-related features

In the past, the automotive real-world data retrieving was limited due to the difficulty to equip the sensors in cars, since the introduction of CAN this limit is overcome.

Authors in [10] propose a driver identification method that is based on the driving behavior signals that are observed while the driver is following another vehicle. They analyze signals, as accelerator pedal, brake pedal, vehicle velocity, and distance from the vehicle in front, were measured using a driving simulator. The identification rates were 81% for twelve drivers using a driving simulator and 73% for thirty drivers.

Data from the accelerator and the steering wheel were analyzed by researchers in [11]. Observing the considered features, they employ hidden Markov model (HMM) to model the driver characteristics. They build two models for each driver, one trained from accelerator data and one learned from steering wheel angel data. The models can be used to identify different drivers with an accuracy equal to 85%.

Researchers in [12] classify a set of features extracted from the powertrain signals of the vehicle, showing that their classifier is able to classify the human driving style based on the power demands placed on the vehicle powertrain with an overall accuracy of 77%.

Van Ly et alius [13] explore the possibility of using the inertial sensors of the vehicle from the CAN bus to build a profile of the driver observing braking and turning events to characterize an individual compared to acceleration events.

Researchers in [14,15] model gas and brake pedal operation patterns with Gaussian mixture model (GMM). They achieve an identification rate of 89.6% for a driving simulator and 76.8% for a field test with 276 drivers, resulting in 61% and 55% error reduction, respectively, over a driver model based on raw pedal operation signals without spectral analysis.

Driver behavior is described and modeled in [16] using data from steering wheel angle, brake status, acceleration status, and vehicle speed through Hidden Markov Models (HMMs) and GMMs employed to capture the sequence of driving characteristics acquired from the CAN bus information. They obtain 69% accuracy for action classification, and 25% accuracy for driver identification.

In reference [17] the features extracted from the accelerator and brake pedal pressure are used as inputs to a fuzzy neural network (FNN) system to ascertain the identity of the driver. Two fuzzy neural networks, namely, the evolving fuzzy neural network (EFuNN) and the adaptive network-based fuzzy inference system (ANFIS), are used to demonstrate the viability of the two proposed feature extraction techniques.

A hidden-Markov-model-(HMM)-based similarity measure is proposed in [18] in order to model driver human behavior. They employ a simulated driving environment to test the effectiveness of the proposed solution.

Authors in [19] propose a method based on driving pattern of the car. They consider mechanical feature from the CAN vehicle evaluating them with four

different classification algorithms, obtaining respectively an accuracy equal to 0.939 with Decision Tree, equal to 0.844 with KNN, equal to 0.961 with RandomForest and equal to 0.747 using MLP algorithm.

Differently from the discussed current literature, we propose a method to assess two different aggressiveness indexes. To this aim, the method is focused on the road identification issue, discriminating between urban and highway roads. Finally, a driver-related risk index is estimated. In addition, relating to road identification task, we highlight that the classification does not require previous knowledge about the type of road, since a cluster analysis algorithm is considered.

2.2 Risk Assessment

Risk assessment, also called underwriting, is exploited by insurers for evaluating and assessing the risks associated with an insurance policy. It is also useful in the premium calculation for an insured. A detailed analysis of risk factors and rating factors for general insurance is discussed in [20]. In [21], the estimation of risk premium for traditional individual car models is discussed. The authors exploit cluster analysis with the aim to identify groups of car exhibiting similar technical attributes. Credibility theory is used to combine estimates of risk premium from individual car model claim statistics and technical assessment. Thr Usage-based motor insurance (UBI) concept is introduced in [3], where the existing literature is critically reviewed and research gaps are identified. Findings show that there is a multiplicity and diversity of several research studies accumulated in modern literature examining the correlation between Pay-as-you-drive (based on driver's exposure) and Pay-how-you-drive (based on driving behavior) schemes and traffic risk in order to determine accident risk. Moreover, it seems that UBI implementation would eliminate the cross-subsidies phenomenon, which implies less insurance costs for goods and less exposed drivers. An approach to risk assessment is provided in [5]: the authors present a platform able to acquire data from the vehicle under analysis to a framework as part of a Pay-As-You-Drive system. Their main aim is to monitor vehicle usage and, exploiting these information, assess the associated risk with the aim to set the appropriate insurance premium. To determine vehicle usage, the system analyses the driver's respect for speed limits, driving style (aggressive or non-aggressive), mobile telephone use and the number of vehicle passengers. This work is the only one in the current state of the art that propose a risk formula based on features like the number of kilometers travelled by the vehicle. Differently from the proposed risk assessment formula, we take into account the driver aggressiveness as a parameters to determine the insurance premium.

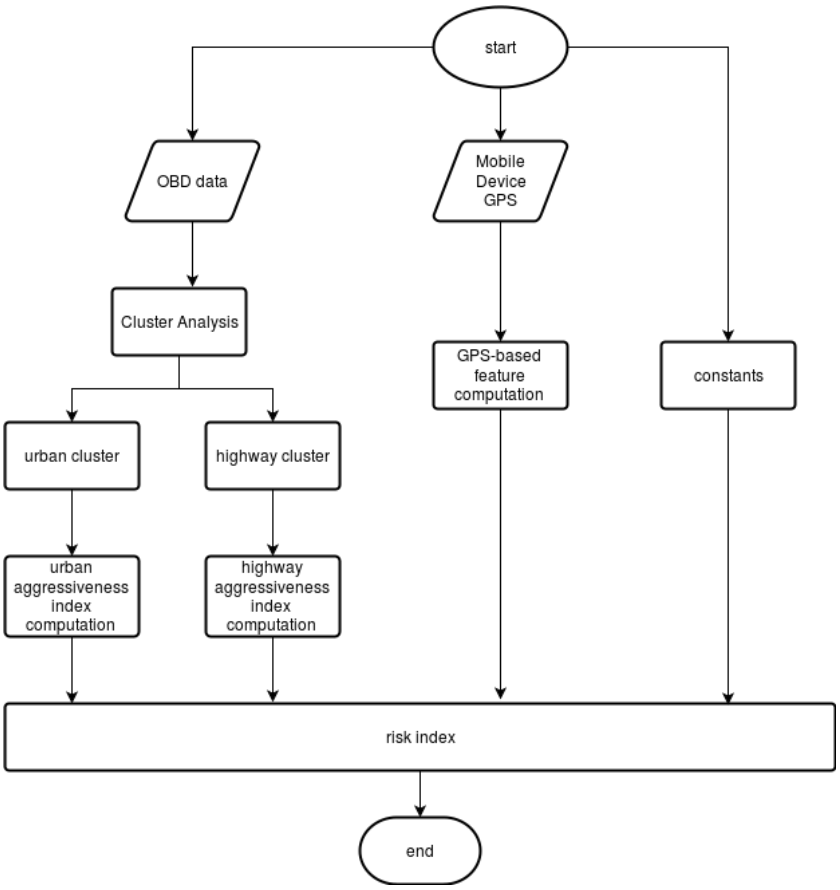

Fig. 1: Flow diagram of the proposed approach for risk index computation.

3 The Method

In following section we describe the considered approach in order to evaluate the driving style (in terms of aggressiveness) from a set of features extracted from the in-vehicle CAN data and from GPS sensor.

Figure 1 depicts the flow diagram of the proposed approach for the risk index computation.

As Figure 1 shows, the cluster analysis process is concerned to OBD data in order to label the gathered data as belonging to the urban or to the highway roads. The OBD is a standard available on all cars (European and American) and from since 1996 is mandatory [22]. In our analysis we considered the feature set (belonging to OBD and to GPS) shown in Table 1.

Feature	Description	Info	OBD	GPS
F1	Engine RPM	Revolutions Per Minute	X	
F2	Mass Air Flow	expressed in g/s	X	
F3	Instantaneous Fuel Consumption	expressed in liters/100 km	X	
F4	Boost pressure estimation	expressed in KPa/Bar/Kg	X	
F5	Acceleration	expressed as g (gravity)	X	
F6	Engine power	expressed in KW	X	
F7	Engine torque	expressed in NM/Kg	X	
F8	Altitude	expressed in degree		X
F9	Longitude	expressed in degree		X
F10	Time	expressed in hh:mm:ss		X

Table 1: Features involved in the study.

We considered features gathered from different sources: the first one is represented by the OBD (i.e., F1, F2, F3, F4, F5, F6 and F7) while the second one is computed by the user device GPS sensor (i.e., F8, F9 and F10).

The GPS sensor features are considered in order to add meta information useful to have the confirmation about the kind of road (i.e., urban or highway). It is identified by the cluster analysis using the F8 and F9 features. Moreover, the GPS sensor features are used to know whether the route was taken in the daytime or at night (using the F10 feature): in Figure 3 this task is represented by the GPS-based feature computation block.

As stated into the introduction, in order to characterize the driver style in terms of aggressiveness, we resort to an unsupervised machine learning approach i.e., cluster analysis.

The reason why we consider unsupervised machine learning algorithms is that our aim is understand whether the considered features exhibits different values in urban and highway paths. Differently from the supervised machine learning algorithms, where to each trained instances there is the target label, cluster analysis algorithms splitting the data in several clusters without no a priori knowledge.

The cluster analysis itself is not one specific algorithm, but the general task to be solved: it can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them [23]. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions.

In this paper we consider following unsupervised classification algorithms: k-means algorithm [24], one of the simplest unsupervised learning algorithms that solve the well known clustering problem [25], Cobweb [26,27], Canopy [28] and FarthestFirst[29] ones.

The k-means procedure follows a simple and easy way to classify a given data set through a certain number of clusters that are fixed a priori. Let us assume k clusters; with particular regard to the designed approach we consider k=2. The main idea is to define 2 centroids, one for each cluster.

These centroids should be placed in a cunning way because different location causes different results [30]. Therefore, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done [31]. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid [24]. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more.

We consider the k-means implementation in the Weka data mining toolkit [4] i.e., SimpleKMeans. This implementation can use either the Euclidean distance (as default) or the Manhattan distance. In this study we set the SimpleKMeans algorithm with the Euclidean distance, maximum iterations number equal to 500 and maximum of generated clusters equal to 2. Since the features given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm (this is one way of distinguishing unsupervised learning from supervised learning): for this reason we consider the incorrectly clustered instances number and percentage in order to evaluated the goodness of the proposed method (i.e., to evaluate whether the first cluster contains the majority of urban while the second one contains the majority of highway ones).

Once the k-means algorithm are evaluated, in order to distinguish between features gathered while the driver is traveling on urban roads and features gathered while the driver is traveling on highway ones, we discuss an approach to use this information providing an aggressiveness index for the PHYD car insurance.

Cobweb is an incremental system for hierarchical conceptual clustering, basically it incrementally organizes observations into a classification tree. In this tree each node represents a class and is labeled by a probabilistic concept able to summarizes the attribute-value distributions of objects classified under the node [26]. This classification tree can be used to predict missing attributes or the class of a new object.

Canopy clustering algorithms requires the specification of distance thresholds, its applicability for high-dimensional data is limited by the curse of dimensionality [28].

FathersFirst traversal of a bounded metric space is a sequence of points in the space, where the first point is selected arbitrarily and each successive point is as far as possible from the set of previously-selected points. The same concept can also be applied to a finite set of geometric points, by restricting the selected points to belong to the set or equivalently by considering the finite metric space generated by these points [29]. The farthest-first traversal of a

[4] `https://www.cs.waikato.ac.nz/ml/weka/`

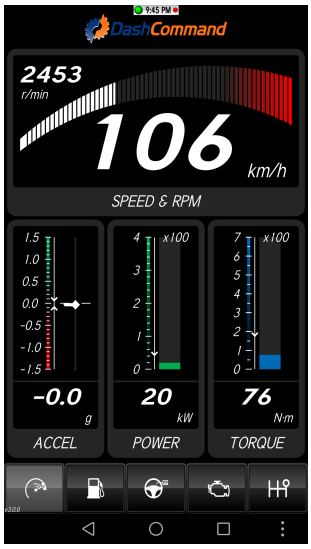

Fig. 2: The DashCommand (OBD ELM App) app while is running on the highway track.

finite point set may be computed by a greedy algorithm that maintains the distance of each point from the previously selected points.

Also with regards to Cobweb, Canopy and Farthestfirst algorithms we consider their implementation in the Weka machine learning tool suite.

4 Experimental Evaluation

In this section we discuss the experiment we performed by means of cluster analysis, in order to classify between urban and highway paths.

The evaluation consists of two stages: (i) a comparison of descriptive statistics of the populations of features and (ii) an unsupervised classification analysis aimed to assess whether the urban and highway features are grouped in different clusters.

We realize a real-world dataset, gathering data from the in-vehicle CAN bus. The vehicle involved in the experiment is a Fiat Punto Evo 1.3 Diesel with 75 horsepowers and with one driver.

In order to collect data, the DashCommand (OBD ELM App)[5] application and Mini Bluetooth ELM327 OBD 2 Scanner were used.

OBD is available on modern car to produce the self-diagnostic report by monitoring vehicle system in terms of measurement and vehicle failure [22].

[5] `https://play.google.com/store/apps/details?id=com.palmerperformance.DashCommand&hl=it`

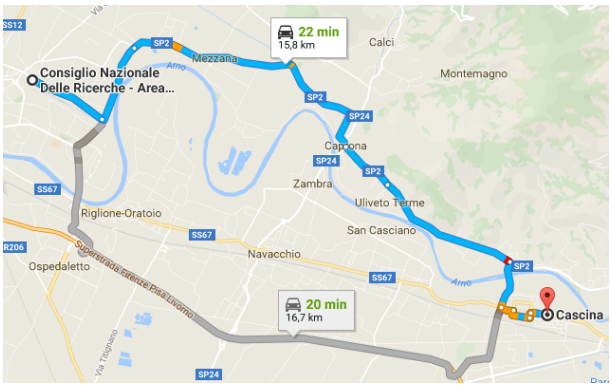

Fig. 3: The urban path considered in the study highlighted in blue: it consists of 22 Km from the Istituto di Informatica e Telematica in Pisa to Cascina, in the center of Italy.

The data are recorded every 1 second during driving using the DashCommand application by an Android smartphone (i.e., a Huawei p8 lite 2017 with Android 7.0 Nougat onboard) fixed in the car by a car support.

In order to label the track using the "urban" or the "highway" label, we developed a Java script able to generate an address from a latitude and longitude through the reverse geocoding Java wrapper [6] able to query the Nominatim search engine for OpenStreetMap data[7].

We collected data from the vehicle in an urban an a highway area in Italy, in Figure 3 the urban path considered: it consists of 22 Km from the Istituto di Informatica e Telematica in Pisa to Cascina, in the center of Italy. The highway path 4 is related to the main Italian highway (the A1, Autostrada del Sole) between the Center and the South of Italy and it consists of 234 Km. In order to balance to traveled kilometers between the urban and the highway paths, we have considered 10 urban paths (i.e., ten different routes of the urban path of 22 Km) and one highway path: in this way we have a dataset composed of 220 Km of urban path and 234 Km of highway path for a total equal to 454 Km.

We represent two scatterplots with the aim to give statistical evidence that considered feature population exhibits different trend between the urban and the highway ones. Similar considerations can be addressed for the other considered features.

Figure 5 shows the scatterplot related to the Engine RPM (i.e., the F1 feature) and Boost pressure estimation (i.e. the F4 feature): the Engine RPM feature is represented on the X axis while the Boost pressure estimation one on the Y axis.

[6] `https://www.daniel-braun.com/technik/reverse-geocoding-library-for-java/`
[7] `http://nominatim.openstreetmap.org/`

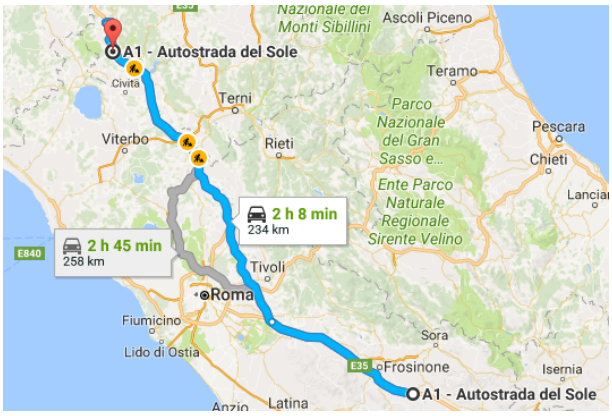

Fig. 4: The highway path considered in the study highlighted in blue: it is related to the main Italian highway between the Center and the South of Italy and it consists of 234 Km

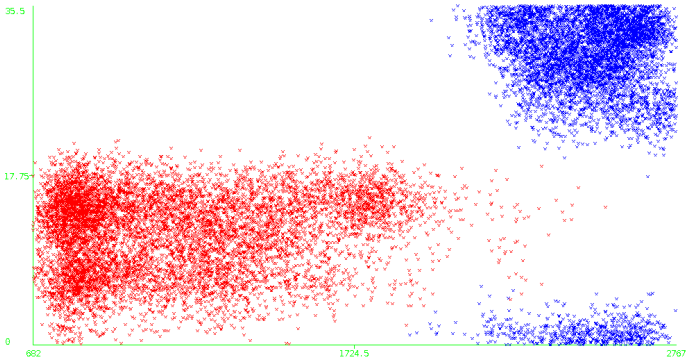

Fig. 5: Scatterplot related to the F1 feature and the F4 feature (the red distribution is related to the urban path, while the blue distribution in related to the highway path).

The red distribution is related to the urban path, while the blue one is related to the highway path: from the scatterplot it is clear the division between the red points, mostly allocated on the center-low left side of the scatterplot, and the blue one, mostly allocated on the high and low right side of the scatterplot.

Figure 6 shows the scatterplot related to the F1 feature and the F7 one i.e., the Engine torque: the F1 feature is represented on the X axis while the F7 one on the Y axis.

In the scatterplot in Figure 6 the red distribution is related to the urban path, while the blue one is related to the highway path (as in Figure 5). In

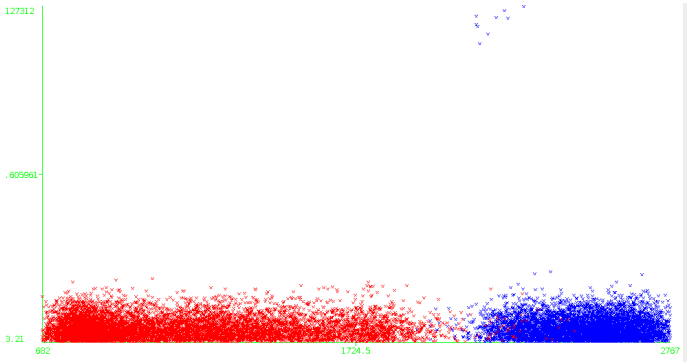

Fig. 6: Scatterplot related to the F1 feature and the F7 feature (the red distribution is related to the urban path, while the blue distribution in related to the highway path).

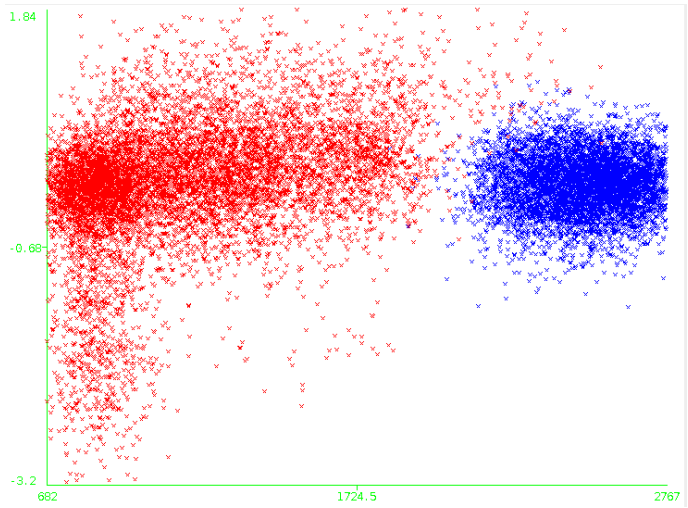

Fig. 7: Scatterplot related to the F1 feature and the F5 feature (the red distribution is related to the urban path, while the blue distribution in related to the highway path).

this case, both the red and the blue distributions are allocated in the down side of the graph, however we can distinguish them clearly: the red points are in the left and middle part of the scatterplot, while the blue points are most allocated in the left side.

In the scatterplot in Figure 7 the red distribution is related to the urban path, while the blue one is related to the highway path. In this case, both the red and the blue distributions are allocated on the left and on the right side

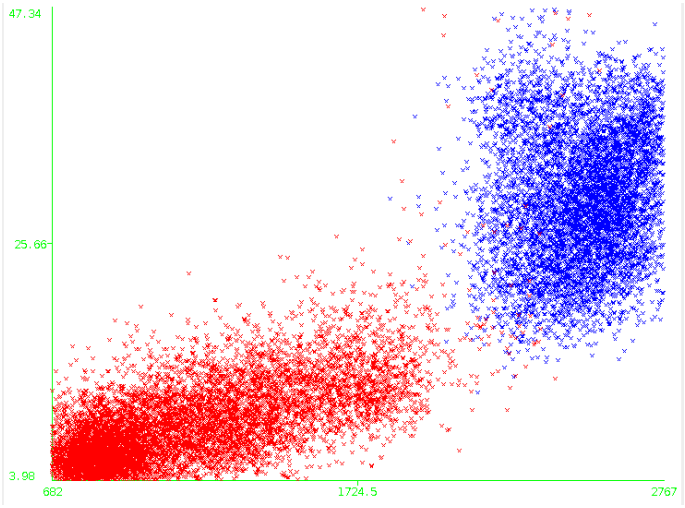

Fig. 8: Scatterplot related to the F1 feature and the F2 feature (the red distribution is related to the urban path, while the blue distribution in related to the highway path).

of the graph: the red points are in the left part of the scatterplot, while the blue points are most allocated in the right side.

In the scatterplot in Figure 8 the red distribution is related to the urban path, while the blue one is related to the highway path. In this case, both the red and the blue distributions are allocated on the left and on the right side of the graph: the red points are in the down left part of the scatterplot, while the blue points are most allocated in the high right side.

From the considerations related to scatterplots in Figures 5 , 6, 7 and 8 we state that the features under analysis can be useful to discriminate between urban and highway paths and, consequently, they can represent good candidates for the cluster analysis phase.

Relating the unsupervised classification, we compute the incorrectly clustered instances number and the percentage in three different scenarios (i.e., we perform three different clustering experiments) with following instances:

- C1: instances related only to the urban path;
- C2: instances related only to the highway path;
- C3: instances related to the urban and highway path (i.e., the full dataset);

We consider three difference instance set (i.e., C1, C2 and C3) with the aim to demonstrate that the more appropriate clusters are obtained using the C3 instances (related to the urban and highway path).

Table 2 shows the results of the C1, C2 and C3 unsupervised classifications.

As shown in Table 2, the C1 experiment (with only urban path instances) obtains an Incorrectly clustered instances value equals to 5551 (i.e., 63% of the

Algorithm	Exp.	ICI	%	time
SimpleKMeans	C1	5551	63.4545%	0.06
	C2	8735	83.5437%	0.09
	C3	444	2.4636%	0.06
Coweb	C1	6527	69.5443%	0.08
	C2	9974	84.3256%	0.18
	C3	845	3.9865%	0.07
Canopy	C1	8064	83.4956%	0.06
	C2	11486	92.0032%	0.10
	C3	1042	5.0002%	0.05
Farthestfirst	C1	5930	64.5673%	0.07
	C2	9042	81.9358%	0.08
	C3	678	2.9983%	0.06

Table 2: Results of the C1, C2 and C3 experiments.

instances considered) with the SimpleKMeans algorithm, to 6527 (i.e., 69% of the instances considered) with the Coweb algorithm, to 8064 (i.e., 83% of the instances considered) with the Canopy algorithm and to 5930 (i.e., 64% of the instances considered) with the Farthestfirst algorithm; the C2 experiment (with only highway path instances) obtains an Incorrectly clustered instances (ICI) value equals to 8735 (i.e., 83% of the instances considered) with the SimpleKMeans algorithm, to 9974 (i.e., 84% of the instances considered) with the Coweb algorithm, to 11486 (i.e., 92% of the instances considered) with the Canopy algorithm and to 9042 (i.e., 81% of the instances considered) with the Farthestfirst algorithm, while the C3 experiment (with both urban and highway paths instances) gives an Incorrectly clustered instances value that equals to 444 with a percentage of incorrectly clustered instances of 2% with the SimpleKMeans algorithm, to 845 (i.e., 3% of the instances considered) with the Coweb algorithm, to 1042 (i.e., 5% of the instances considered) with the Canopy algorithm and to 678 (i.e., 2.9% of the instances considered) with the Farthestfirst algorithm.

We obtain better results when using the SimpleKMeans algorithm: considering the full dataset only the 2% if instances are misclassified.

These results demonstrate that the adoption of the unsupervised machine learning techniques is promising: as a matter of fact, considering the different driving styles that should be adopted in urban and highway roads, we can consider the Incorrectly clustered instances value as an estimator of the driving style. In case this value is low, the driver exhibits a different driving style between urban and highway paths and this is the result of the different driving style that should be adopted on different roads. On the other hand, whether the Incorrectly clustered instances value exhibits an high value (for instance, in the C1 and C2 experiment), as we demonstrated, the cluster analysis is not able to correctly define the clusters (C1 and C2 experiment), and this is symptomatic that the driver under analysis exhibits a driving style pretty similar in urban

and highway roads and the feature set considered is representative of the kind of traveled roads.

Once the clusters with regards to the urban and to the highway path are obtained, in order to compute the two driver aggressiveness indexes, we consider the acceleration feature (i.e., F5) variation: this is the reason why we resort to the *standard deviation* statistical dispersion index i.e., an estimate of the variability of a data population or a random variable (in this case the variable is represented by the F5 feature).

Considering u_i the value of the i-th urban path occurrence of the F5 feature, N_u the total number of urban path occurrences of the F5 feature (with $1 \leq i \leq N_u$) we define the driver aggressiveness index σ_{urban} in urban path as follows:

$\sigma_{urban} = \sqrt{\frac{\sum_{i=1}^{N_u}(u_i - \bar{x}_{urban})^2}{N_u}}$

where $\bar{x}_{urban}$ represents the arithmetic mean of F5 feature urban path distribution and it is defined as:

$\bar{x}_{urban} = \frac{1}{N_u}\sum_{i=1}^{N_u} u_i$

Relating to the driver aggressiveness index $\sigma_{highway}$ in highway path, considering h_k the value of the k-th highway path occurrence of the F5 feature, N_h the total number of highway path occurrences of the F5 feature (with $1 \leq i \leq N_h$), we define the $\sigma_{highway}$ index as follows:

$\sigma_{highway} = \sqrt{\frac{\sum_{i=1}^{N_h}(h_k - \bar{x}_{highway})^2}{N_k}}$

where $\bar{x}_{highway}$ represents the arithmetic mean of F5 feature highway path distribution and it is defined as:

$\bar{x}_{highway} = \frac{1}{N_h}\sum_{i=1}^{N_k} u_k$

The estimated values of the driver aggressiveness indexes are the following: $\sigma_{urban} = 6.4734$ and $\sigma_{highway} = 2.4519$.

From these results, we deduce that the driver under analysis exhibits a more aggressive driving style in the urban path (with $\sigma_{urban} = 6.4734$) than in the highway one (i.e., $\sigma_{urban} = 2.4519$).

We consider this behaviour as normal: typically urban roads require more accelerations and decelerations if compared to the highway ones.

The opposite behavior would be considered highly aggressive.

5 A Risk Assessment Calculation

Behavioural aspects of driving, should be incorporated in insurance models in order to contribute towards current trends of personalized vehicle insurance.[3]. In line with this observation, in the following we discuss a possible risk assessment calculation taking into account several parameters:

- driver aggressiveness index in urban path (i.e., σ_{urban});
- driver aggressiveness index in highway path (i.e., $\sigma_{highway}$);
- time bands (day/night): identification of two time bands, each of which is assigned an appropriate penalty (this information is acquired using the GPS sensor);

– road traveled (in Km);
– history: number of times that the driver in previous analysis was considered as aggressive / not aggressive, we consider the history of the previous urban and highway aggressiveness index (with an appropriate penalty defined by constant values).

A possible risk index (i.e., RI) calculation is shown below:

$$RI = (KM*K_1)+(\frac{\sum_{i=1}^{t} i*\sigma_{urban_i}}{\sum_{i=1}^{t} i})+K_4*(\frac{\sum_{i=1}^{t} i*\sigma_{highway_i}}{\sum_{i=1}^{t} i})+(\frac{K_2*\%day+K_3*\%night}{100})$$

where: KM represents the road traveled expressed in KM, $K_1 = 0.001$ (constant value), σ_{urban_i} represents the i-th urban aggressiveness index value, $K_4 = 2$ (constant value),$\sigma_{highway_i}$ represents the i-th highway aggressiveness index value, $K_2 = 40$ (constant value) [3], %day is the percentage of time the vehicle is used during the daytime,$K_3 = 60$ (constant value) [3] and %night is the percentage of time the vehicle is used at night and t is the length of the historical series of the aggressiveness indexes. We set different values for $K_2 = 40$ and $K_3 = 60$ because we consider the time in which the vehicle is used at night potentially more dangerous than the time in which the vehicle is used during day, for this the reason the same kilometers traveled during night have a greater impact than the ones traveled during the day time. The reason why we set $K_1 = 0.001$ is that we do not want that the kilometers traveled can be decisive for the risk index calculation (for this reason we multiply the kilometers traveled with 0.001).

In order to assign a greater weight to the most recent aggressiveness indexes (both the urban and the highway one), we consider the last i-th value as the most recent aggressiveness index computed.

In the following, we present an example of RI calculation for three different drivers: A, B and C. Table 3 shows the results where t=5 for **Driver A** and **Driver C** and t=3 for **Driver B**.

Variables	**Driver A**	**Driver B**	**Driver C**
KM	248	480	384
$5\sigma_{urban_5}$	5.34	n.a.	5.98
$4\sigma_{urban_4}$	4.23	n.a.	4.93
$3\sigma_{urban_3}$	2.34	6.43	4.76
$2\sigma_{urban_2}$	4.21	5.45	6.08
σ_{urban_1}	4.28	5.38	5.38
$5\sigma_{highway_5}$	2.26	n.a.	3.89
$4\sigma_{highway_4}$	1.89	n.a.	3.63
$3\sigma_{highway_3}$	2.45	2.17	2.48
$2\sigma_{highway_2}$	3.07	3.09	2.85
$\sigma_{highway_1}$	3.52	2.58	2.44
%day	31%	76%	42%
%night	69%	24%	58%
RI	*63.05*	*56.29*	*64.01*

Table 3: Risk index computation for three different drivers.

As shown in Table 3, from the risk index computation of the A, B and C drivers we obtained that driver with the lower risk index is the B driver (but we highlight that the B driver exhibits 3 aggressiveness index values (i.e., $1 \leq i \leq 3$)). Relating to the A and C drivers (both of them with 5 aggressiveness index values (i.e., $1 \leq i \leq 5$))), we obtain that the A driver presents a lower risk index if compared with the C one: as a matter of fact the RI for the A driver is equal to 63.01, while the IR related to the C driver is equal to 64.01.

6 Conclusion and Future Work

"Pay How You Drive" insurance scheme presents many potentials and appears to have many benefits. In line with the need to develop new methodologies which take into account several parameters to evaluate driving behaviour, we propose an approach assessing driver's aggressiveness through cluster analysis and unsupervised machine learning techniques.

Basing on the evidence that drivers exhibit different driving styles on different kind of roads (urban or highway), we propose an approach to compute the driver aggressiveness. We identify the kind of road traveled through unsupervised machine learning in order to assess the driver aggressiveness on urban and highway paths. Than we propose a driver related risk index. In order to verify the cluster analysis method discerning between urban and highway data, we use a set of features extracted from the CAN bus of real-world car while traveling in different roads (i.e., urban and highway) in the center and south of Italy. As future work we plan to adopt formal verification techniques aimed to identify whether a driver can be classified in several predefined categories (for instance: the young driver, the ruthless driver, the cautious driver) in order to propose a risk index considering the category to which a driver belongs. In addition we explore whether deep learning algorithms can be helpful to obtain better performances in driver aggressiveness computation.

Acknowledgment

This work has been partially supported by H2020 EU-funded projects NeCS and C3ISP and EIT-Digital Project HII and PRIN "Governing Adaptive and Unplanned Systems of Systems" and the EU project CyberSure 734815.

7 Compliance with Ethical Standards

Maria Francesca Carfora declares that she has no conflict of interest
Fabio Martinelli declares that he has no conflict of interest
Francesco Mercaldo declares that he has no conflict of interest
Vittoria Nardone declares that she has no conflict of interest
Albina Orlando declares that she has no conflict of interest

Antonella Santone declares that she has no conflict of interest

Gigliola Vaglini declares that she has no conflict of interest

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

This article does not contain any studies with animals performed by any of the authors.

Informed consent was obtained from all individual participants included in the study.

References

1. A. Marotta, F. Martinelli, S. Nanni, A. Orlando, and A. Yautsiukhin, "Cyber-insurance survey," *Computer Science Review*, 2017.
2. P. Desyllas and M. Sako, "Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance," *Research Policy*, vol. 42, no. 1, pp. 101–116, 2013.
3. 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.
4. S. Kantor and T. Stárek, "Design of algorithms for payment telematics systems evaluating driver's driving style," *Transactions on Transport Sciences*, vol. 7, no. 1, p. 9, 2014.
5. L. Boquete, J. M. Rodríguez-Ascariz, R. Barea, J. Cantos, J. M. Miguel-Jiménez, and S. Ortega, "Data acquisition, analysis and transmission platform for a pay-as-you-drive system," *Sensors*, vol. 10, no. 6, pp. 5395–5408, 2010.
6. A. Mehar, S. Chandra, and S. Velmurugan, "Speed and acceleration characteristics of different types of vehicles on multi-lane highways," *European Transport*, vol. 55, pp. 1825–3997, 2013.
7. J. Wang, K. Dixon, H. Li, and J. Ogle, "Normal acceleration behavior of passenger vehicles starting from rest at all-way stop-controlled intersections," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1883, pp. 158–166, 2004.
8. F. Martinelli, F. Mercaldo, A. Orlando, V. Nardone, A. Santone, and A. K. Sangaiah, "Human behavior characterization for driving style recognition in vehicle system," *Computers & Electrical Engineering*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045790617329531
9. 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.
10. T. Wakita, K. Ozawa, C. Miyajima, K. Igarashi, I. Katunobu, K. Takeda, and F. Itakura, "Driver identification using driving behavior signals," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 3, pp. 1188–1194, 2006.
11. X. Zhang, X. Zhao, and J. Rong, "A study of individual characteristics of driving behavior based on hidden markov model," *Sensors & Transducers*, vol. 167, no. 3, p. 194, 2014.
12. G. Kedar-Dongarkar and M. Das, "Driver classification for optimization of energy usage in a vehicle," *Procedia Computer Science*, vol. 8, pp. 388–393, 2012.
13. M. Van Ly, S. Martin, and M. M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013, pp. 1040–1045.
14. C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 427–437, 2007.
15. Y. Nishiwaki, K. Ozawa, T. Wakita, C. Miyajima, K. Itou, and K. Takeda, "Driver identification based on spectral analysis of driving behavioral signals," in *Advances for In-Vehicle and Mobile Systems*. Springer, 2007, pp. 25–34.

16. S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. H. Hansen, "Analysis and classification of driver behavior using in-vehicle can-bus information," in *Biennial Workshop on DSP for In-Vehicle and Mobile Systems*, 2007, pp. 17–19.
17. X. Meng, K. K. Lee, and Y. Xu, "Human driving behavior recognition based on hidden markov models," in *Robotics and Biomimetics, 2006. ROBIO'06. IEEE International Conference on.* IEEE, 2006, pp. 274–279.
18. M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2016.
19. B. I. Kwak, J. Woo, and H. K. Kim, "Know your master: Driver profiling-based anti-theft method," in *PST 2016*, 2016.
20. P. Booth, S. Haberman, R. Chadburn, D. James, Z. Khorasanee, R. H. Plumb, and B. Rickayzen, *Modern actuarial theory and practice.* Chapman and Hall/CRC, 2004.
21. M. Campbell, "An integrated system for estimating the risk premium of individual car models in motor insurance," *ASTIN Bulletin: The Journal of the IAA*, vol. 16, no. 2, pp. 165–183, 1986.
22. F. Martinelli, F. Mercaldo, V. Nardone, and A. Santone, "Car hacking identification through fuzzy logic algorithms," in *Fuzzy Systems (FUZZ-IEEE), IEEE International Conference on. IEEE.* IEEE, 2017.
23. L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons, 2009, vol. 344.
24. J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.
25. S. Har-Peled and A. Kushal, "Smaller coresets for k-median and k-means clustering," *Discrete & Computational Geometry*, vol. 37, no. 1, pp. 3–19, 2007.
26. D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning*, vol. 2, no. 2, pp. 139–172, 1987.
27. J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial intelligence*, vol. 40, no. 1-3, pp. 11–61, 1989.
28. A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2000, pp. 169–178.
29. D. S. Hochbaum and D. B. Shmoys, "A best possible heuristic for the k-center problem," *Mathematics of operations research*, vol. 10, no. 2, pp. 180–184, 1985.
30. A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
31. D. Arthur, B. Manthey, and H. Röglin, "k-means has polynomial smoothed complexity," in *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on.* IEEE, 2009, pp. 405–414.