Machine learning approaches exploring the optimal number of driver profiles based on naturalistic driving data

Tselentis, Dimitrios I.; Papadimitriou, Eleonora

DOI
10.1016/j.trip.2023.100900

Publication date
2023

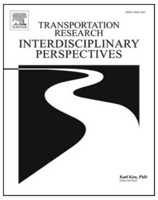
Document Version
Final published version

Published in
Transportation Research Interdisciplinary Perspectives

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



TRANSPORTATION RESEARCH INTERDISCIPLINARY PERSPECTIVES

Machine learning approaches exploring the optimal number of driver profiles based on naturalistic driving data

Dimitrios I. Tselentis *, Eleonora Papadimitriou

Faculty of Technology, Policy, and Management, Safety and Security Science Section, Delft University of Technology, Jaffalaan 5, Delft 2628 BX, The Netherlands

ARTICLE INFO

Keywords:
Driving behaviour
Driving profiles
Optimal driver profile recognition
Machine learning
Unsupervised learning
Naturalistic driving data

ABSTRACT

Driver behavior analytics is an important concept that plays a significant role in the understanding of road crashes. This paper investigates the optimal number of driver profiles to understand the most important characteristics that differentiate drivers and extract useful insights on the value of using different clustering approaches in profile recognition. To this end, two Machine Learning clustering algorithms, the K-Means and OPTICS algorithms, are applied on driving data from a large naturalistic experiment using almost 18 K trips recorded from 130 drivers. The results revealed 3 profiles, the less risky drivers, the modest drivers and the more aggressive drivers. Clustering was based on 3 important driving behavior characteristics, namely the number of speeding, headway and harsh events per 100 km. The less risky drivers profile was revealed by both algorithms, whereas drivers of higher aggressiveness are distinguished by K-Means based on the driving feature that dominates the rest. The OPTICS algorithm showed that many drivers, especially the aggressive ones, present unique behavior that cannot be grouped together with other drivers. The interpretability of driver profiles resulting from the application of these unsupervised learning techniques is worsened as the number of clusters increases. The association between driver profiles and individual characteristics leads to the conclusion that aggressiveness is mainly driven by personality traits and less by specific characteristics such as gender, age or past accident history. The results of this study can be potentially used to develop profile-specific applications that provide feedback to drivers and reduce their crash risk.

Introduction

Driver behavior analytics is an emerging concept with several important applications during the past decades. As we have entered into the Big Data era, new data collection schemes and advanced modelling techniques related to Machine Learning (ML) and Artificial Intelligence (AI) are available. These create considerable opportunities for large-scale collection of new data such as driver physiological indicators, trip driving time and conditions, congestion, road surface and environment conditions, detailed weather and spatial information (Weidner et al., 2017; Ellison et al., 2015), which can be used for the analysis of driving behavior.

According to (Nilsson, 1982), AI is a subpart of computer science, concerned with how to give computers the sophistication to act intelligently, and to do so in increasingly wider realms. It is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as reasoning, visual perception, speech recognition, automated learning and scheduling, robotics, decision-making and translation between languages. AI leverages computers to mimic the problem-solving and decision-making capabilities of the human mind. ML is a branch of AI that develops algorithms imitating human way of learning and gradually improve prediction accuracy.

The feasibility and benefit of the highly accurate identification of driver profiles and driving styles based on metrics collected from inertial sensors (speed, acceleration, braking, steering etc.) across time and space is shown in literature (Weidner et al., 2017; Ellison et al., 2015; Tselentis et al., 2019; Tselentis et al., 2021; Papadimitriou et al., 2019; Mantouka et al., 2019; Mantouka and Vlahogianni, 2022). Nonetheless, it is also recently indicated that changes in driving behavior may sometimes be quick in time (Tselentis et al., 2021). In order to ensure that these behavioral shifts are correctly captured over time and that their safety implications are adequately understood, drivers should be continuously monitored at a high resolution. Moreover, there are several ML algorithms that have been used in these questions, but little research has been done to comparatively assess the advantages and limitations of different methods in driver behavior and profiling analysis.

* Corresponding author.
E-mail address: D.Tselentis@tudelft.nl (D.I. Tselentis).

https://doi.org/10.1016/j.trip.2023.100900
Received 28 February 2023; Received in revised form 2 August 2023; Accepted 3 August 2023

It is clarified that this study explores the topic of driver profile recognition using the definition of driver profiles as provided by (Tselentis and Papadimitriou, 2023), which is a "group of drivers having similar driving behaviour and characteristics". This is clarified because it was noticed that there is ambiguity in the way the terms 'driver profiles" and "driving patterns" are used in literature. The difference between these two is that driving pattern is defined as "a driving behaviour characteristic, such as a driving manoeuver like a harsh braking event, which is occurring repetitively either by the same driver or by different drivers in a population. Hence, in our study driving pattern is a more microscopic aspect than personalized driving behaviour". Therefore, the level of analysis is the most significant difference between these two, where in the field of driver profile recognition the analysis is focusing on the macroscopic characteristics of driving behaviour.

Literature review and objectives

K-Means is one of the best known clustering methodologies that belongs to unsupervised learning techniques and aims to group the data into a number of clusters K previously specified by the researcher. It has a wide range of application in road safety studies for discovering profiles and patterns of all road users including pedestrians and cyclists (Vogel et al., 2014; Kim and Yamashita, 2007). Several driver profiling studies have used the K-Means algorithm to identify the existing profiles (Tselentis et al., 2021; Mantouka et al., 2019; Warren et al., 2019). This algorithm can cluster several subjects into groups with similar behaviour based on multiple features. Its simplicity has made it popular among the developed clustering algorithms.

Density-based clustering algorithms, such as DBSCAN, have been used in the past in this scientific field. For instance (Li et al., 2016) clustered driver physiological data into "Normal", "Event" and "Noise" clusters using an application of the DBSCAN algorithm on real-world data. Other studies exploited density-based clustering to classify vehicle approaching patterns and analyse driver behaviour (Wen et al., 2021). Nonetheless, no studies were found that have exploited any density-based clustering algorithm for driver profile recognition.

Another clustering method that belongs to the density-based clustering algorithms is the OPTICS (Ordering points to identify the clustering structure) algorithm that has been widely applied to fields related to spatial data clustering (Ankerst et al., 1999; Agrawal et al., 2016; Pei et al., 2009; Deng et al., 2015; Duan et al., 2007; Malzer and Baum, 2020). It has also been used in previous studies to understand the regular travel behaviour of private vehicles, identify the driving destination, (Liu et al., 2021; Levin and Håkansson, 2015). Literature review also revealed one study related to road safety, where the authors used OPTICS to analyse the road traffic crashes through the identification of the road crash locations (Islam et al., 2021).

Neural Networks (NN) have also been used in the past to classify driving behaviour in different profiles as normal or aggressive (Savelonas et al., 2020; Saleh et al., 2017). These classification methodologies were based on Recurrent Neural Networks (RNN), long-short-term memory (LSTM) RNN and Gated recurrent units (GRUs) and results showed a high accuracy precision. In most cases, studies that use RNN also make use of time-series data and make predictions of the driver profiles by observing a sequence of driving time, called time-slice or driving pulse (Tselentis and Papadimitriou, 2023). This task is also called time-series classification. Data in these studies were collected through naturalistic driving experiments and recorded through sensors installed in the vehicle and smartphone devices, such as GPS, accelerometer, gyroscope and compass.

Based on the literature review conducted, the K-means algorithm is most commonly used, among the Artificial Intelligence methodologies that are used for driver profile identification. This is followed by NN-based models, the extended use of which also appears in recent studies (Mukherjee et al., 2021; Savelonas et al., 2020; Saleh et al., 2017). Methodologies based on statistics and optimization are also utilized, whereas PCA is employed in studies to reduce dimensionality of the datasets used (Fugiglando et al., 2018; Constantinescu et al., 2010). Nonetheless, it is found that NN-based methodologies are usually used in supervised learning approaches when the driver classes are known and the scope is to assign drivers that have newly appeared to those classes. Moreover, they are usually found to be using time-series data and not trip or driver-level indicators that will be used in this study.

In terms of the driving metrics used in driver profiling studies, the review revealed that speed and positive acceleration are the two driving metrics that were mostly collected and used. These two metrics are followed by negative acceleration (braking), timestamp, driver distraction that is usually measured through mobile phone usage or eye-tracking, and GPS coordinates (Tselentis et al., 2019; Tselentis et al., 2021; Mantouka et al., 2019; Warren et al., 2019; Bergasa et al., 2019; Fugiglando et al., 2018; Saleh et al., 2017).

The collection of these driving metrics mainly takes place through naturalistic driving experiments that were record data either using sensors installed inside instrumented vehicles or mobile phone sensors (Tselentis et al., 2019; Tselentis et al., 2021; Papadimitriou et al., 2019; Mantouka et al., 2019; Warren et al., 2019; Bergasa et al., 2019; Saleh et al., 2017; Ellison et al., 2015). Since these experiments are naturalistic, the driver sample usually ranges between 6 and 300 and the recording duration from a few minutes (one trip) up to one year. In terms of the recording frequency, the most commonly used is 1 Hz, since driver profiling focuses on the macroscopic driving behaviour and therefore does not consider microscopic changes that need a higher frequency to be captured. It can be inferred that this is an acceptable frequency that is balancing both noisy data collection with lack of sufficient information (Tselentis and Papadimitriou, 2023). Therefore, it can be considered adequate for performing macroscopic analyses such as driver profile recognition.

Finally, in terms of the driver profiles discovered in literature, the most commonly identified driver profiles are related to aggressiveness. There is certainly a relationship between this and what mentioned above, that most studies use the driving metrics of speed and acceleration for driver profile recognition. The number of profiles ranges from 2 to 6, with 3 and 4 being the most frequent (Fugiglando et al., 2018; Liao et al., 2022; Nouh et al., 2021). Moreover, the profile of "normal" or "typical" drivers is also found to be a common driver profile by several studies (Tselentis et al., 2021; Saleh et al., 2017). Other groups also discovered are those of drowsy, calm, cautious and conservative drivers (Warren et al., 2019; Saleh et al., 2017). There are several other driver group characterizations in literature, e.g. in terms of driving efficiency, or consistency and stability of their temporal behavioural characteristics (Tselentis et al., 2021; Tselentis et al., 2019). Finally, it was found that some studies did not discuss the driver profiles discovered and focused mainly on the methodology used.

Summarizing the above, driver profiling studies usually apply clustering methodologies, such as the K-Means algorithm that has proved to provide relatively good results, on driving data collected through naturalistic driving experiments. Nonetheless, an important gap found in existing research is the absence of a robust methodology for the identification of driver profiles in terms of safety (Tselentis and Papadimitriou, 2023). Moreover, to the best of our knowledge, the OPTICS algorithm has not been used in the driving behavior analysis field. The objective of this study is the development and comparison of two different methodological approaches for driver profile recognition, based on a commonly used clustering algorithm such as K-Means and a less used algorithm such as the OPTICS algorithm. It aims to shed light on the different driver profiles that exist in terms of safety, taking into account several driving characteristics recorded during a large-scale real-world naturalistic experiment across Europe. It will also provide a description of each driver profile discovered, including the recognition of risky behaviors such as aggressiveness.

This research is based on the Rhapsody H2020 research project (Rhapsody, 2021–2023) that is developing new algorithms and models using high-resolution, large-scale data collected through the naturalistic driving experiments of the i-Dreams H2020 research project (i-Dreams, 2019–2022). Those data consist of both driving data such as harsh maneuvers and distance to the vehicle in front as well as physiological indicators of the driver recorded in real-time and data coming from questionnaires. Since there are no labelled data and there is no prior knowledge on what the different profiles are, an unsupervised learning methodology will be employed in order to discover the existing driver profiles. This approach will explore both machine learning approaches, namely K-Means and OPTICS clustering algorithms.

Data collection, cleansing and pre-processing

Naturalistic driving data were collected from the experimental data of the i-Dreams H2020 project (i-Dreams, 2019–2022) using an Application Programming Interface (API) service developed by the project partners. Almost 29 k trips were collected that took place between September 2021 and June 2022 by 130 passenger car drivers in Belgium (54 drivers), Germany (25 drivers) and the UK (51 drivers). Analytical information per trip was collected and data were aggregated to produce key trip performance indicators, which were the number of harsh acceleration, braking and cornering events, the level of headway distance and the level of speed limit violation. These indicators were initially recorded on a trip level and ultimately aggregated on a driver level during data pre-processing.

For the purposes of this study, all harsh events recorded (acceleration, braking and cornering) were aggregated into one indicator named harsh events. The harsh acceleration and braking events are defined as harsh manoeuvers or vehicle movements related to a significant increase or reduction of longitudinal speed respectively. As for the harsh cornering events, those are defined as sudden right or left turn of the vehicle. It is highlighted at this point that the data provider used a classified algorithm to classify events as harsh or not. The exact details of this algorithm cannot be disclosed due to confidentiality reasons; in general, it uses data from several sensors such as the GPS and the accelerometer to detect the events taking place and their intensity level. This algorithm is trained using Machine Learning techniques and calibrated through annotated field experiments.

Headway distance events are recorded when the following vehicle is within a close distance from the leading vehicle and finally, speed limit violation events are those indicating the speed limit exceedance. A short description and the descriptive statistics of min, max, mean, median and standard deviation are provided for each indicator in Table 1. There are more indicators collected during the i-Dreams experiments but only these are currently available in a robust and validated way. As more indicators will become available in the future, they will be included in this research.

During data cleansing, trips with distance less than 300 m as well as those with duration less than 90 s were eliminated. In order to be included in the final dataset, drivers should have travelled at least a total number of 20 trips and 200 km (Stavrakaki et al., 2020). The sample was checked for outlier driving behavior but no driver was found with such behavior. Outliers were detected using the boxplot definition of outliers that exist above maximum, i.e. higher than the Q3 + 1.5 * IQR (Inter-Quartile Range). The same was done also for the outliers below minimum value, i.e. Q1 – 1.5 * IQR. The final dataset that was analyzed, included 27,919 trips from 130 drivers. The driving metrics used in the analysis were estimated as the total number of events occurred per 100 km travelled i.e. the number of harsh events per 100 km, the number of headway events per 100 km and the number of speeding violations per 100 km. It should be noted that after data were cleansed and before applying the clustering algorithm, clustering features were standardized by subtracting the mean and dividing by the standard deviation to scale data values. Data collection, cleansing, pre-processing and analysis were implemented using Python 3.7 and the Python packages of requests, pandas, numpy and sklearn.

Methodological approach

As described above in brief, this research makes use of two clustering algorithms, the K-Means and the OPTICS algorithm, to identify the existing driver profiles. Apart from the description of these algorithms, this section provides also a short description of the elbow method that is used for the determination of the number of clusters that should be considered in this analysis as well as a description of the Silhouette analysis used to evaluate the performance of the resulting clusters that are identified by the two algorithms. It is clarified at this point that the number of clusters resulting from the elbow method will be used as an indication of the number of clusters that will be presented herein for both clustering methods. Similarly to that, the Silhouette analysis will also be used to assess the clustering performance for all clusters resulting from both methods.

K-means algorithm

Clustering allows finding and analyzing the groups that were formed naturally, instead of defining groups prior to looking at the data. K-Means clustering is a type of unsupervised learning, which aims to find the optimum way to group given data, with the number of groups represented by the variable k that is given as input. This data grouping is based on the feature similarity of the observations. The centroid of each cluster is a collection of feature values that defines resulting groups, based on which the average behavior of the resulting groups is interpreted (Hartigan and Wong, 1979). K-Means is an iterative algorithm that starts with randomly selecting K points as the initial centroids and assigning each observation to each cluster based on their distance to each centroid. Once the full sample is assigned to K clusters for the first time, the centroids are recalculated. The process of measuring the distances between each observation and each centroid is repeated using the new centroid position and the full sample is re-assigned. This procedure is repeated again several times until either the centroids are changing or the maximum number of algorithmic iterations is reached.

OPTICS algorithm

OPTICS, which stands for ordering points to identify the clustering structure, is a density-based clustering algorithm, which was proposed by (Ankerst et al., 1999). The concept of density-based clustering is that the neighborhood of a given radius has to contain at least a minimum number of objects for each object of a cluster. In other words, a point p is a core point of a cluster if at least the minimum number of points are found within its neighborhood (including point p itself). These two parameters (maximum radius and minimum number of points that form a cluster) are the minimum required inputs for a density-based clustering

Table 1
Indicators used in this study and their statistical metrics per driver.

Driving performance indicator	Description	Min	Max	Mean	Median	St. Dev.
Harsh events	Number of harsh acceleration, braking and cornering maneuvers	597	17,171	5,587	4,552	5,466
Headway distance	Number of headway distance events	330	14,427	4,535	4,094	4,054
Speed limit violation	Number of speed limit violations	306	11,609	3,503	2,814	2,848

to be performed. OPTICS is used for finding density-based clusters in spatial data and works similarly to the DBSCAN algorithm (Ester et al., 1996). Nonetheless, OPTICS is capable of detecting meaningful clusters in data of varying density, which is DBSCAN's major weakness. This is achieved by ordering the points to be clustered in such a way that closest points become neighbors in the ordering. The distance between points and therefore, the density threshold for considering both points to be in the same cluster is represented through a dendrogram that is created.

Elbow method

Since clustering is an unsupervised learning process, the number of clusters should be decided beforehand. This can be determined by running the K-Means algorithm for k times and comparing the results. One of the most commonly used metrics for comparing results across different values of k is the mean distance between cluster centroids and the data points assigned to each one of them, which is decreased while the number of clusters is increased. When this metric is plotted as a function of the number of clusters k, k can be used to estimate the "elbow point", which is the point where the rate of this metric's decrease sharply shifts.

Silhouette analysis

The separation distance between the clusters resulting from a clustering algorithm can be studied through Silhouette analysis. The silhouette index provides a measurement of how close each point within a cluster is to points in the neighboring clusters. This measure is in the range of −1 to 1 and it is helpful for the assessment of the total number of clusters. The silhouette index can be displayed in a silhouette plot. Silhouette indices closer to + 1 provide an indication that neighboring clusters are well-separated and distanced from each other. Values closer to 0 show that clustered points are very close to the decision boundary between two neighboring clusters. On the other hand, values close to −1 are clearly showing that these clustered points have most probably not been assigned to the correct cluster.

Results and main contributions

Number of clusters

Fig. 1 illustrates the results of the elbow method, which is used to indicate the optimal number of clusters, which represents the number of driver profiles. This optimal number is found to be in the range of 3 to 5, since this is the point where a sharp shift in the clustering score is observed. As previously described, this score is calculated using the sum of distances between each observation and the centroid of the cluster at which it is assigned.

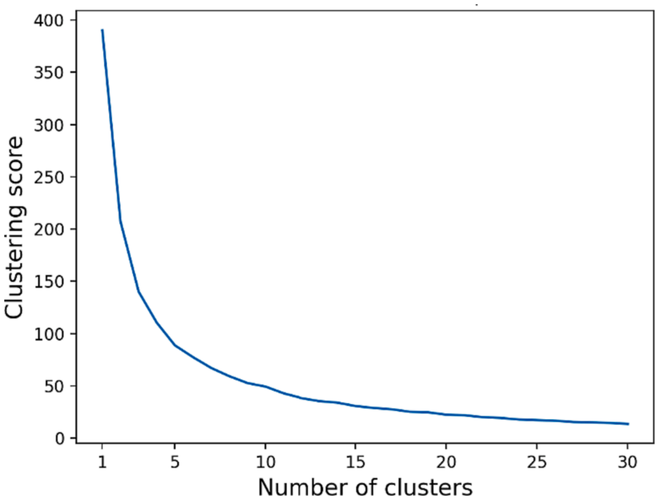


Fig. 1. Elbow graph showing the clustering score for different numbers of driver clusters.

This range is rational also from the perspective of producing explainable results, considering that the driver sample size could not justify a much higher number of clusters since each cluster would not have a sufficient number of drivers to draw statistically significant results. This study will investigate the full range of 3 to 6 clusters, which will also test the robustness of the two algorithms.

Results of the K-means algorithm

Table 2 illustrates the results of the K-Means algorithm. For each clustering performed, between 3 and 6 driver clusters, this table shows the number of drivers and their number of trips as well as the average trip distance and number of events per 100 km for the 3 clustering features, harsh events, headway and speeding. These metrics are provided for each clustering group formed. The silhouette plots together with the detailed 3D illustrations of each of the clusters 3 to 6 are presented in Fig. 2.

Regarding the 3-cluster results, the algorithm revealed the three driver profiles of less risky, modest and most aggressive drivers. Less risky drivers appear to have significantly lower number of all types of events, headway, speeding and harsh events. This becomes apparent also from the fact that cluster 1 is very well separated from the other, which is shown in Fig. 2. On the other hand, clusters 0 and 2 have equally high number of speeding violations per 100 km but have a noticeable difference in headway events and especially in harsh events. The silhouette score of 0.409 shows a medium to low performance of the clustering algorithm in terms of how well the three clusters are separated.

The results of the 4-clusters shown in Fig. 3 run show that the less risky driver cluster (cluster No 1) remains almost the same, which indicates that it is better separated from the rest. This is also shown in Fig. 3. The main part of the modest drivers also remains the same in this run, which is confirmed by the fact that the metrics of cluster No 3 of this run are exactly the same with those of cluster No 0 of the 3-cluster results. Moreover, the more aggressive part of the modest drivers is moved towards the cluster of the most aggressive drivers that is now split into 2 parts, those performing either more harsh events (cluster number No 2) or speeding violations (cluster number No 0). In this run, the silhouette score is 0.384 indicating that drivers are slightly worse clustered compared to the 3-cluster run.

In the 5-cluster run shown in Fig. 4, the less risky drivers cluster remains again almost the same as in the previous runs. On the other hand, the cluster of the modest drivers are split into two in this run, the less and the more aggressive part. The more aggressive part of the modest drivers has incorporated also a part of the initially (3-cluster run shown in Fig. 2) defined aggressive drivers. The higher aggressiveness of this part of the modest drivers is mainly depicted in the number of harsh acceleration events. The split of the most aggressive drivers cluster remains but leads to significantly smaller clusters. Again, their main difference is that half of them are performing a high number of harsh events whereas the other half is performing a higher number of speeding violations. Overall, the most aggressive driver cluster appears to be cluster 1 where apart from the number of speeding events, the number of headway and harsh events are higher compared to the rest. The silhouette score is also found here to be reduced to 0.356 compared to the previous two runs.

As for the 6-cluster results shown in Fig. 5, the cluster of the less risky drivers remains again almost the same as in the previous runs. The modest drivers cluster remains also split into two parts, the less and more aggressive modest drivers. The less aggressive part is almost the same as in the previous clustering, whereas the more aggressive part includes also some drivers that were initially defined as aggressive.

Table 2
Clustering results of the K-Means algorithm.

Clustering #	Driver cluster ID	# of drivers	# of trips	Average trip distance	Harsh events/ 100 km	Headway events/ 100 km	Speeding violations/ 100 km
3	0	71	6,216	14	24	19	15
3	1	30	3,906	14	18	2	4
3	2	29	7,797	10	37	23	16
4	0	32	8,001	11	29	22	18
4	1	29	3,777	14	18	1	4
4	2	9	2,227	9	53	23	12
4	3	60	3,914	15	24	19	14
5	0	42	8,654	16	23	20	13
5	1	15	3,605	11	32	26	21
5	2	29	3,777	14	18	1	4
5	3	8	2,010	9	53	22	11
5	4	36	9,873	12	27	18	15
6	0	17	5,162	9	33	18	12
6	1	28	3,501	14	18	0	3
6	2	29	7,387	12	24	19	16
6	3	5	1,297	9	58	21	11
6	4	40	8,135	17	24	21	13
6	5	11	2,437	13	31	29	27

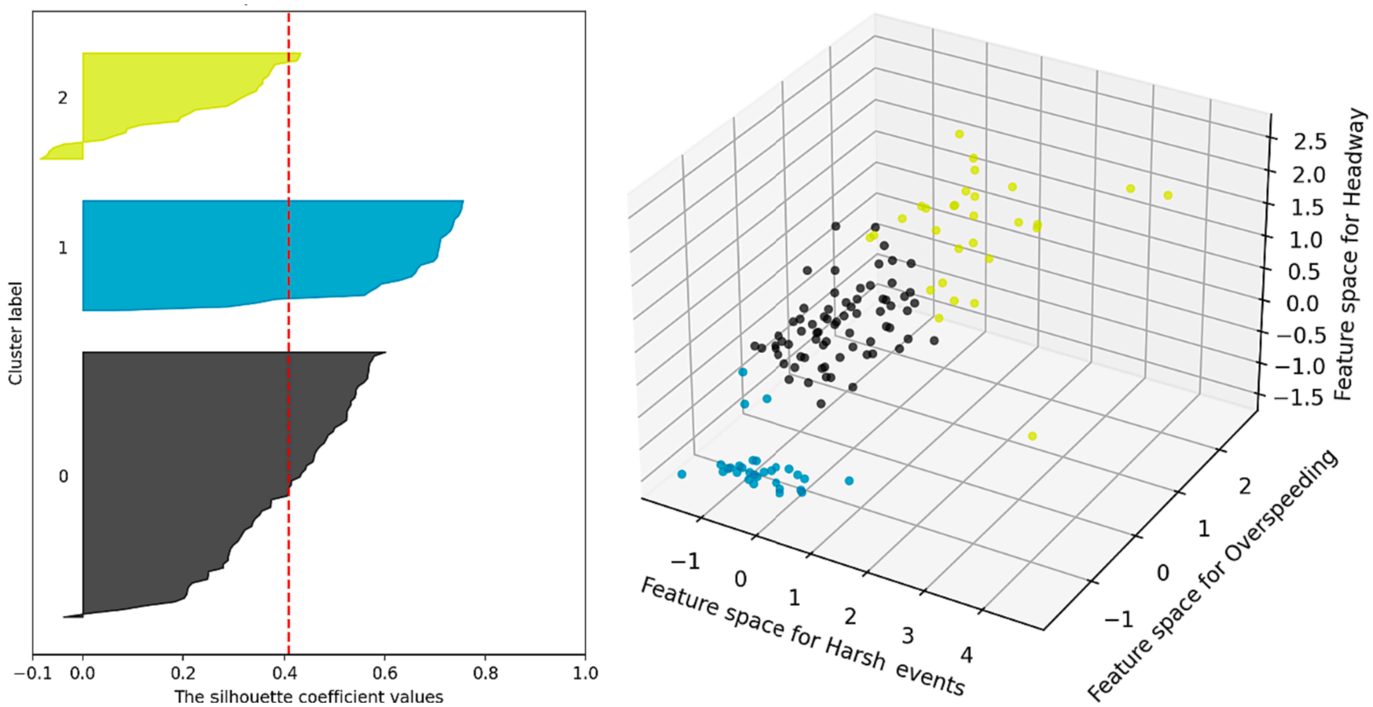


Fig. 2. Results of the K-Means 3-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

Regarding the most aggressive drivers, they are now split into 3 but in significantly smaller clusters. As also indicated in Table 2, the first group of aggressive drivers includes those performing a huge number of harsh events (cluster No 3). Apart from the number of speeding events, drivers of this cluster present a number of headway and harsh events that is close to the average. The other 2 clusters (clusters No 0 and 5) with a large number of harsh events but differ in speeding violation and headways. In fact, cluster No 0 presents normal number of speeding violation and headways events, being closer to the profile of modest drivers. Similarly to the above, the silhouette score is reduced to 0.344 now that the number of clusters is increased.

Results of the OPTICS algorithm

As explained above, density-based clustering assigns points to clusters only when a minimum number of points is found within the neighborhood of a point. This is why in each clustering presented in Table 3, there is also a row dedicated to those drivers that were not assigned to any cluster. As the number of clusters grows, the number of drivers not clustered is reduced due to the fact that more drivers are assigned to at least one cluster.

Considering that i) the number of clusters when using the OPTICS algorithm is a result of parameter tuning and that ii) the purpose of this study is to compare and explain the performance of 2 clustering algorithms that are performing on a totally different basis, the 2 main parameters of OPTICS were fine-tuned so that they provide the same number of clusters 3 to 6. Table 3 provides the OPTICS parameters used for each clustering.

Table 4 presents the results of the OPTICS algorithm for 3 to 6 clusters. Similarly to the results of the K-Means algorithm, this table shows the number of drivers and their number of trips as well as the average trip distance and number of events per 100 km for the 3 clustering features used in each clustering performed (3 and 6 driver cluster). These metrics are provided for each clustering group formed. The

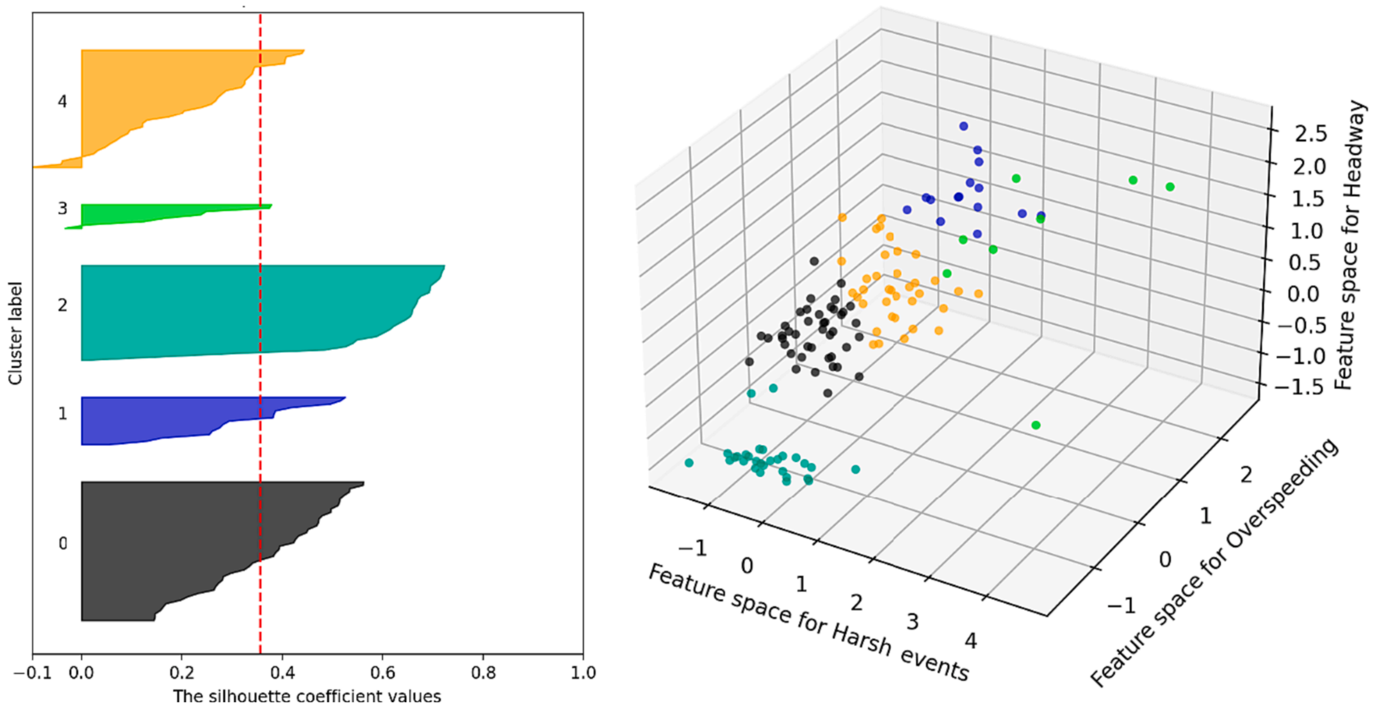


Fig. 3. Results of the K-Means 4-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

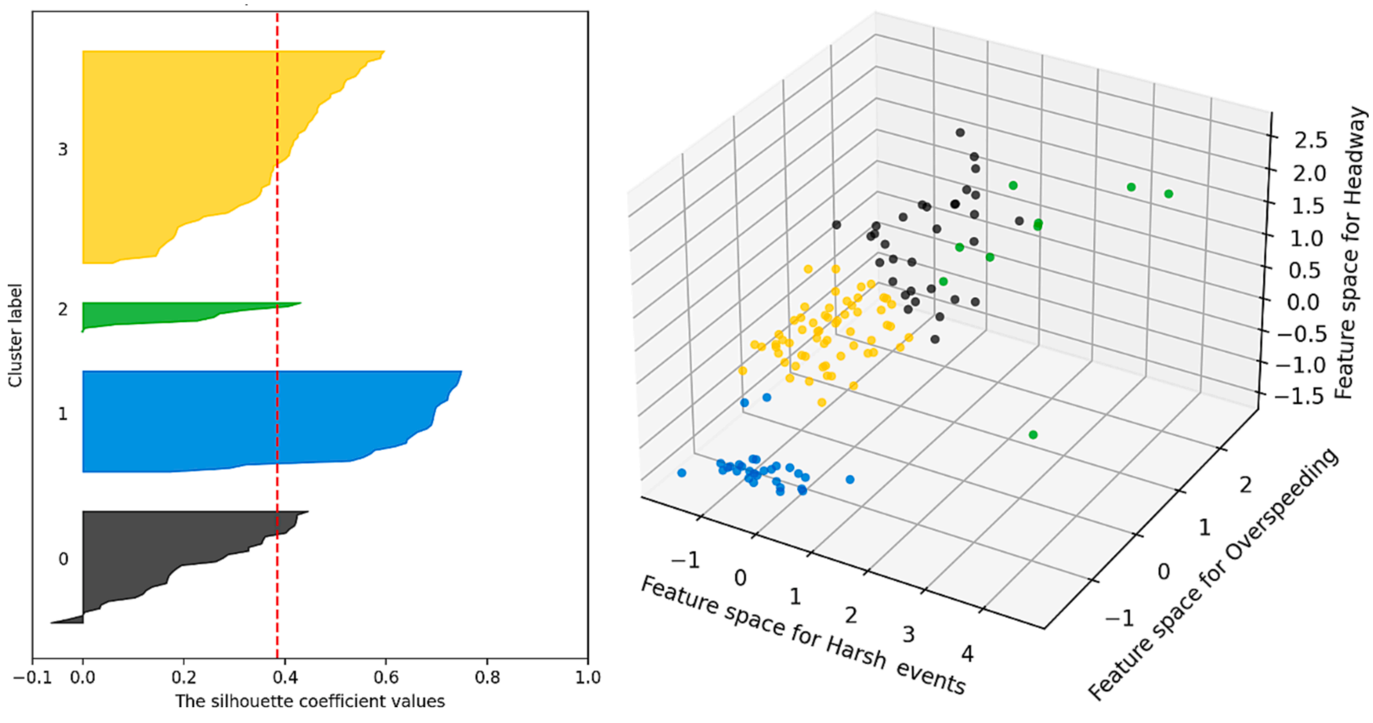


Fig. 4. Results of the K-Means 5-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

silhouette plots together with the detailed 3D illustrations of each of the clustering performed are presented below.

The 3-cluster run of the OPTICS algorithm shown in Fig. 6 revealed 2 clusters of less risky drivers, which are both part of the less risky drivers cluster of the K-Means algorithm. This is confirmed both by Table 4 and Fig. 6. Both these clusters have 0 headway events per 100 km and there are only slight differences between them in terms of speeding and harsh events. This means that less risky drivers consistently present driving risk metrics that are close to 0 and therefore, a low variability in their behavior. This conclusion is based on the fact that the OPTICS algorithm detects high-density areas i.e., areas with drivers that present very similar behavior. The 3rd cluster represents modest drivers and it is part of the modest drivers cluster discovered also by the K-Means algorithm, including slightly more aggressive drivers. It is also observed that the 85% of drivers is not assigned to any cluster. Those drivers present a very high variability in terms of behavioral characteristics with none of

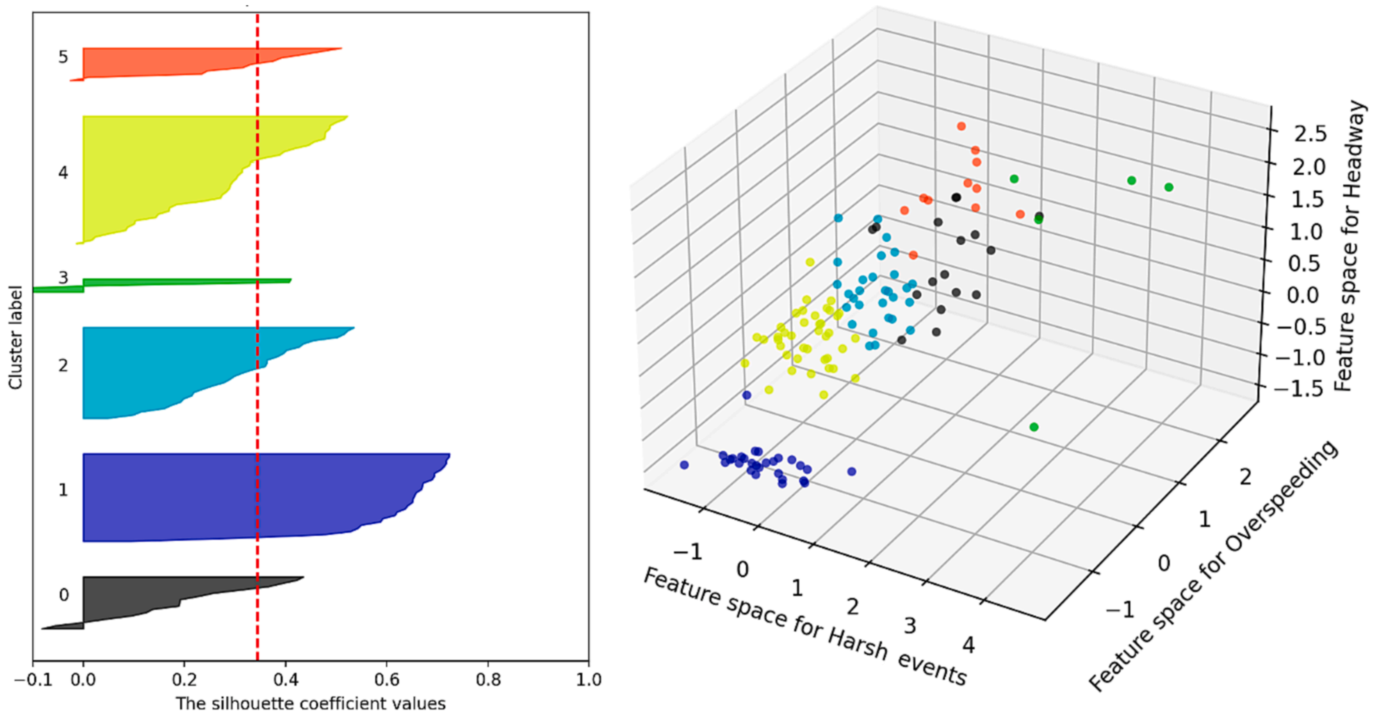


Fig. 5. Results of the K-Means 6-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

Table 3
Parameters used for each clustering of the OPTICS algorithm.

# clusters	Minimum number of points	Maximum radius
3	5	0.3
4	6	0.7
5	5	0.4
6	5	0.7

them having at least 4 drivers with similar characteristics to form a common group. This can be interpreted as that the OPTICS algorithm creates small and compact profiles with very common behaviors and it does not necessarily assign every driver to a specific group if no other driver is observed to have similar behavior. The significantly higher silhouetted index of 0.683 is measured for this clustering, showing that despite the fact that not all drivers are clustered, the clusters formed are compact and well-separated from the rest that were discovered.

The less aggressive cluster of the OPTICS algorithm's 4-cluster run shown in Fig. 7 is almost the same as the respective one of the K-Means algorithm, which is confirmed by both Table 4 and Fig. 7. In this run, 3 clusters of modest drivers are discovered (clusters No 1 to 3), with cluster No 3 being the most aggressive compared to the rest. Its slightly higher aggressiveness is observed especially in the number of headway and harsh events and not in overspeeding. The percentage of drivers not belonging to any cluster is significantly reduced to 65% displaying a sensitivity of the algorithm to the maximum radius and the minimum number of points in a cluster considered. The silhouette index is significantly reduced to 0.598, which indicates that the clusters of

Table 4
Clustering results of the OPTICS algorithm.

Clustering #	Driver cluster ID	# of drivers	# of trips	Average trip distance	Harsh events/ 100 km	Headway events/ 100 km	Speeding violations/ 100 km
3	Not clustered	111	25,002	13	27	18	14
3	0	7	470	24	18	0	5
3	1	7	847	15	21	0	3
3	2	5	1,600	16	33	28	14
4	Not clustered	84	19,602	12	28	20	15
4	0	27	3,337	14	19	0	3
4	1	6	1,685	9	27	14	13
4	2	7	1,512	19	22	23	13
4	3	6	1,783	17	34	29	14
5	Not clustered	90	19,925	12	27	18	14
5	0	7	470	24	18	0	5
5	1	7	847	15	21	0	3
5	2	11	2,786	10	26	15	14
5	3	7	1,950	19	37	31	15
5	4	8	1,941	18	21	21	12
6	Not clustered	86	19,087	12	28	19	14
6	0	7	470	24	18	0	5
6	1	7	847	15	21	0	3
6	2	11	2,786	10	26	15	14
6	3	8	1,941	18	21	21	12
6	4	6	1,783	17	34	29	14
6	5	5	1,005	10	15	7	10

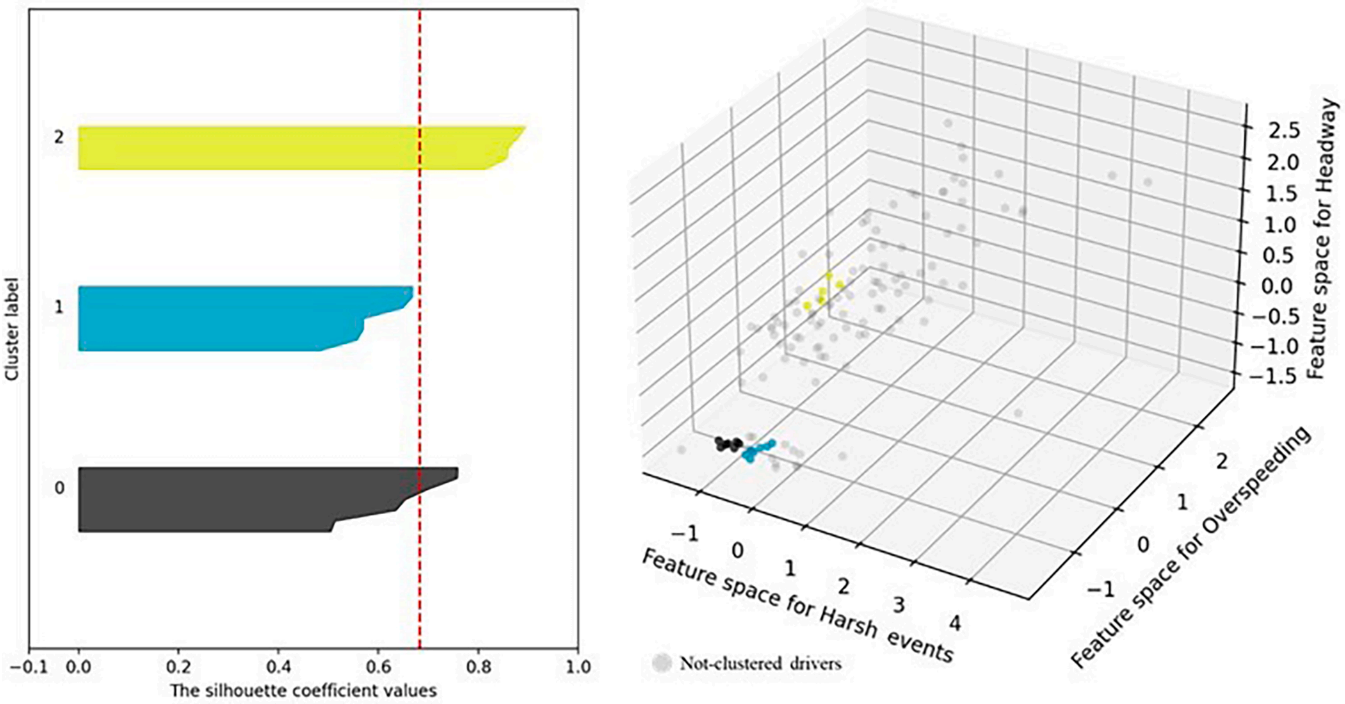


Fig. 6. Results of the OPTICS 3-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

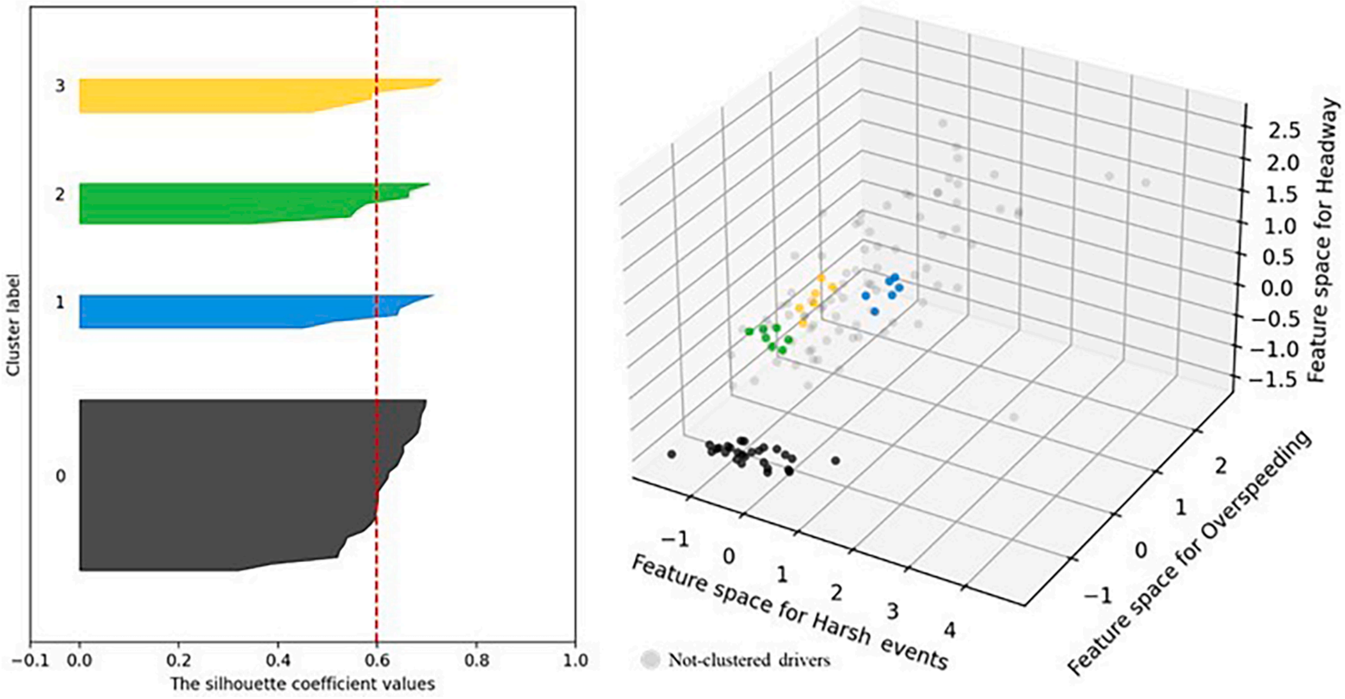


Fig. 7. Results of the OPTICS 4-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

modest drivers may have some similarities to each other.

In the 5-cluster run of the OPTICS algorithm shown in Fig. 8, the cluster of less risky drivers is split again into two parts, showing that these two parts do not significantly differ. The other 3 clusters represent also the modest drivers with cluster No 3 having included a few more aggressive drivers in terms of all three driving features considered. The driver percentage not clustered is almost not altered and the silhouette index is reduced to 0.561.

Finally, the 6-cluster run of the OPTICS algorithm shown in Fig. 9, revealed a new cluster of modest drivers that is less aggressive than the rest (cluster No 5). It is apparent from Fig. 9 that this cluster less dense than the rest meaning that the distance of each driver from the average driver behavior of this cluster is higher than it is in the other clusters. The results for the rest of the driver clusters are very similar to those of the 5-cluster run. This stands also for the silhouette index that was 0.546 in this run.

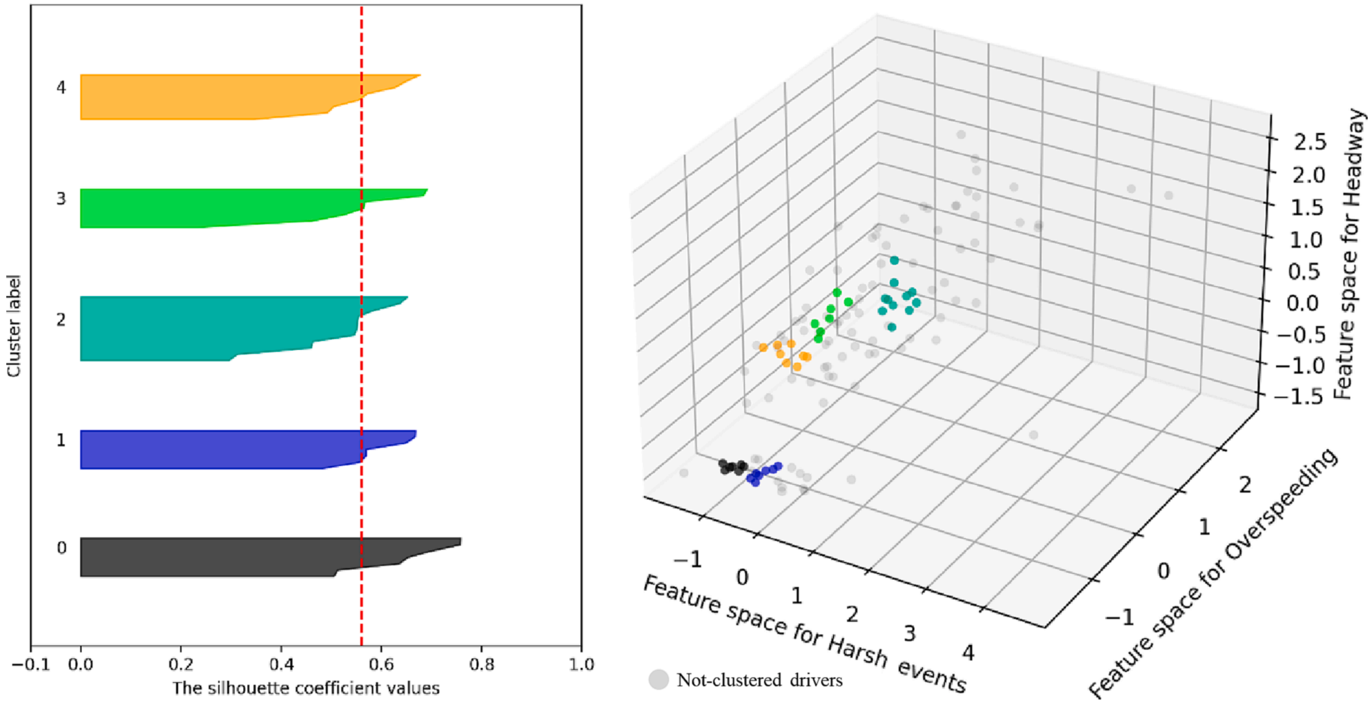


Fig. 8. Results of the OPTICS 5-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

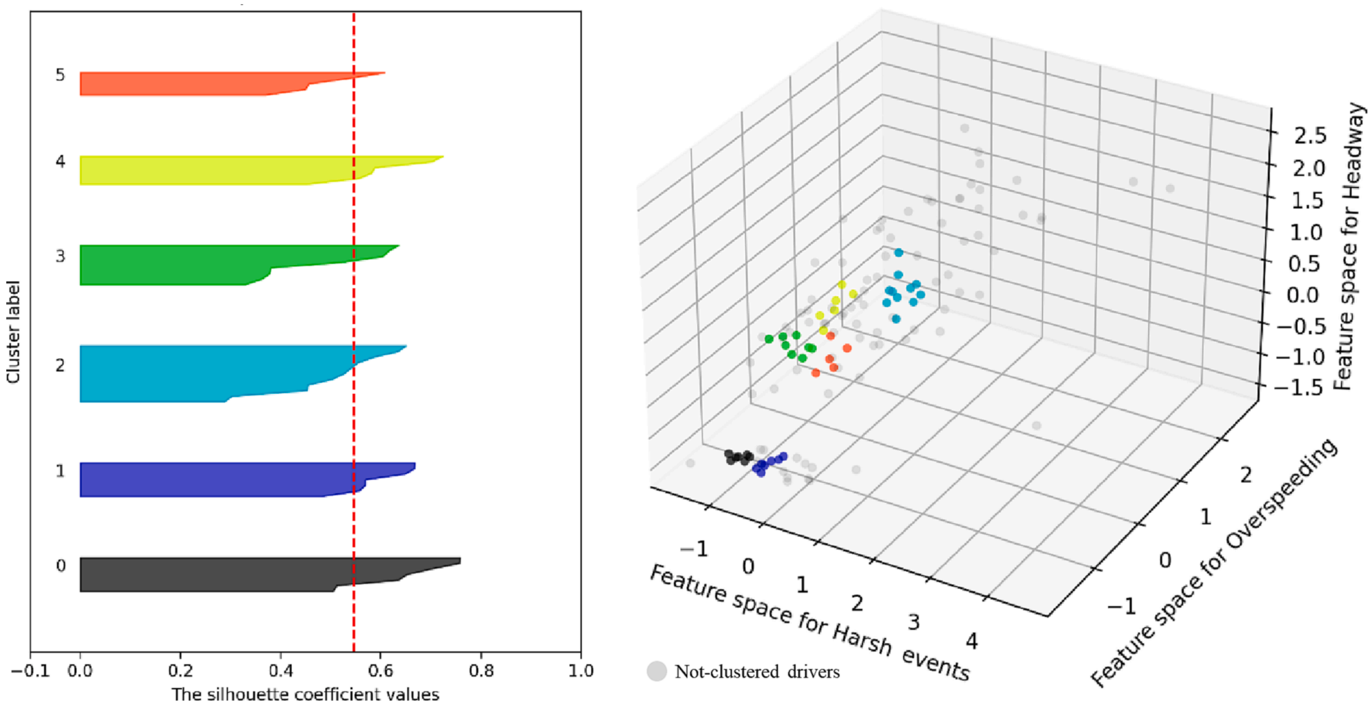


Fig. 9. Results of the OPTICS 6-cluster clustering. (a) Silhouette plot for all clusters; (b) 3D illustration of the 3 normalized features used in clustering.

Discussion

Main findings

One significant finding of this analysis is that the less risky/aggressive driver profiles are more consistently identified by the clustering algorithms. The less risky drivers were identified by both algorithms in all runs performed, which shows that this profile is very dominant, most probably because being consistently less risky can only be achieved by performing a low number of events in all risk indicators considered. On the other hand, the instability in the algorithmic detection of more aggressive driver clusters shows a difficulty in their identification which can be interpreted as that extreme behaviors are much more diverse in terms of their potential measurement characteristics. In other words, this means that despite the fact that more aggressive drivers exist, they present unique behavior and therefore, cannot form a specific group. This was confirmed also by the OPTICS algorithm that did not reveal any cluster of highly aggressive drivers showing that their behavior can be extremely variable.

Moreover, drivers of higher aggressiveness are distinguished based on their dominant driving feature, which is also an indication that the K-Means clustering algorithm is more sensitive to the higher values of aggressiveness indicators used and less to the lower ones. This is also related to the observation mentioned above that it is more rare to observe drivers that present safe behavior in terms of all driving safety aspects and more frequent for drivers to be aggressive at least to one of those aspects. This is also an indication that this driver group is heterogeneous, which may be considered as one of the reasons of not having one single group of aggressive drivers in different clustering findings.

Regarding the optimal number of driver profiles, it was highlighted that the silhouette index is reducing while the number of clusters is increasing and that profiles present less significant differences because they are as well-separated as in cluster runs with lower number of clusters. Based on the interpretation of the physical meaning of the clusters, increasing the number of clusters more than 4 does not provide more meaningful results. This might be an indication that the optimal number of clusters is 3 to 4, at least based on this sample. The results of 3 and 4 clusters are also confirmed by (Sanjurjo-De-No et al., 2020; Fugiglando et al., 2018; Liao et al., 2022; Nouh et al., 2021; Chronis et al., 2021), who found that similar results were produced by the K-Means and hierarchical clustering algorithms. A few other studies though support the existence of 5 or 6 driver profiles or states (Weidner et al., 2017; Mantouka et al, 2019; Warren et al, 2019; Constantinescu et al., 2010; Payyanadan and Angell, 2022).

The main driver profiles discovered found are those of less risky, modest drivers and more aggressive drivers. Results indicate that modest drivers can be possibly further divided into two sub-profiles of lower and higher aggressiveness, which is still lower than that of the more aggressive drivers. Finally, more aggressive drivers exist but probably need more data to draw conclusions on which are the groups that present consistently common characteristics compared to the rest. Based on the above, it is suggested to choose a combination of these types of clustering to identify different levels of profiles and their underlying sub-profiles. The results found herein in terms of the types of driver profiles discovered are also similar to those found in the past literature (Tselentis et al., (2019, 2021, 2023), Bergasa et al., 2019; Saleh et al., 2017; Liao et al., 2022; Nouh et al., 2021; Abdulwahid et al., 2022), which are categorizing drivers in terms of driving risk, normality and aggressiveness. There are also studies utilizing different driving characteristics, such as drowsiness and mobile phone usage, and therefore revealing different types of profiles. For instance, these profiles could be defined by stress, resilience, distraction, velocity etc. (Chronis et al., 2021; Weidner et al., 2017).

In terms of the clustering algorithms tested, the K-Means algorithm assigned all drivers to the derived clusters, whereas the OPTICS algorithm created mini-profiles and not necessarily assigned everyone to one cluster/group. It was found that approximately 65% of drivers was not assigned to clusters displaying thus a unique and highly variable behavior of most drivers without significant similarities with the rest of the drivers' sample. This leads to the general remark that drivers present significant diversity in terms of driving characteristics and only a minority presents common characteristics that could be grouped.

At this point, it is highlighted that one of the goals of this study was to present two different clustering approaches for the purpose of driver profile recognition. The fact that the OPTICS algorithm did not assign all drivers to a profile/cluster may have a twofold interpretation, i) the OPTICS algorithm should always be coupled with another algorithm and ii) the rest of the drivers not assigned to a cluster/profile present a diverse behaviour and therefore cannot be considered part of any driver group or profile. An alternative interpretation would be to exploit the results of this algorithm to identify dense clusters areas and utilize this as an indication of the actual number of driver profiles.

As for the above suggestion to couple the OPTICS algorithm with another method, a suggestion would be to couple it with an optimization method that will assign the not clustered drivers to an existing cluster, by finding the optimal fit to a profile/ cluster in terms of the driving characteristics considered. Building a similarity index that will be calculated between each combination of driver profile and driver not clustered to measure could be helpful towards this direction. The algorithm would aim to minimize the total sum of all similarity indices and constraints such as a maximum number of drivers or a specific value range for a driving metric in each profile could be considered.

Finally, regarding the important features used in this analysis, both clustering algorithms showed sensitivity in all 3 features considered, which shows that all of them are significant in driver profiling.

Association between profiles and driver characteristics

Further analysis was performed on the association between driver profiles discovered by KMeans and the driver characteristics of each cluster. The results of the KMeans algorithm and not those of the OPTICS algorithm were analysed because the OPTICS algorithm assigns the minority of the drivers to clusters and the rest are categorized as outliers. It is highlighted that the interpretation of this association is based on descriptive statistics and not on statistical tests. This was not done in this study since, because on the one hand it was not in our main scope, and on the other hand the incompleteness of driver data (as shown in Table 5) would result in a total sample that would not be adequate for statistical comparisons.

The driver characteristics collected through questionnaires administered during the project were gender, age, income and accident involvement within the last 3 years. The association of these characteristics was investigated for all clustering tests from 3 to 6 and results were found to be similar. For this reason, the results for 3 clusters will be indicatively shown here, but it is highlighted that similar findings were discovered regardless of the number of clusters. Table 5 shows the distribution of age groups across clusters for KMeans clustering of 3 clusters. Almost half of the drivers' sample belongs to the 30–60 age group whereas 9% has not declared their age group. The 26% is younger than 30 years old and the rest are over 60 years old. It becomes apparent that Cluster 1 of the less aggressive drivers has the highest % of younger drivers. Moreover, the majority of older drivers are concentrated in Cluster 0, which is the cluster of the moderate drivers.

Regarding drivers' income, it was found that high income drivers (>5K€/month) are distributed across all clusters. It was also discovered that Cluster 2 (more aggressive drivers) includes a slightly higher number of lower income drivers (<2K€/month).

Significant differences in the drivers' distribution across clusters were not found in terms of drivers' gender. Cluster 0 follows a gender distribution that is very similar to the overall sample distribution, whereas Cluster 1 includes slightly more male drivers and Cluster 2 slightly more female drivers.

Finally, regarding accident involvement in the past 3 years, no specific cluster of drivers was found to have a significantly higher involvement than the rest. The 87% of drivers were not involved in any accident within the last 3 years and the 13% were involved in at least 1. A similar distribution is followed by Cluster 0 and 1, with Cluster 2 (more aggressive drivers) showing a slight difference in accident involvement that was 19%.

It should be highlighted that this study used a relatively small sample size and large, heterogeneous age groups. Moreover, when comparing clusters in terms of a certain variable e.g. gender, the confounding effect other variables such as age and income, was not adjusted. Therefore, a larger and more homogeneous sample should be collected to draw statistically significant results and understand whether significant differences exist between drivers with different characteristics. It is also important to adjust for the effect of other variables when comparing clusters in terms of one variable. Based on the findings of this research, it can be concluded that aggressiveness is mainly related to personality

Table 5
Distribution of age groups across clusters for KMeans clustering of 3 clusters.

Age group	Modest (Cluster 0)	Less risky (Cluster 1)	Most Aggressive (Cluster 2)	Total (% Total)
<30	12 (35% of group)	16 (47% of group)	6 (18% of group)	34 (26%)
30–60	36 (56% of group)	10 (16% of group)	18 (28% of group)	64 (49%)
>60	16 (80% of group)	2 (10% of group)	2 (10% of group)	20 (16%)
Not declared	7 (58% of group)	2 (17% of group)	3 (25% of group)	12 (9%)

traits of each individual driver rather than to drivers' demographics such as the gender or the income group. It is likely that a correlation exists between previous accident involvement and aggressive driving, but the available dataset is not sufficient to draw conclusions. Nonetheless, this could also be attributed to the lack of enough variability in age and traffic accidents (such as narrow age range, etc.) or to the non-sensitivity of this research method for discovering different driver categories because of high homogeneity of study population to this section.

Conclusions and future work

This paper applied an interdisciplinary research approach based on Machine Learning (ML) and driver behavior / road safety disciplines and employed two ML clustering algorithms (K-Means and OPTICS) to identify driver profiles. Different clustering approaches seem to provide better insights on the optimal number of driver profiles and how these could be defined. The profiles that exist in the examined sample are i) less risky drivers, ii) modest drivers and iii) aggressive drivers. It was also found that less risky drivers can be clearly defined by both clustering algorithms, while more aggressive drivers present a more diverse behavior. Moreover, both algorithms produced more robust results when the number of clusters was reduced. Finally, no significant association between driver characteristics and clustering was found other than that of drivers' age. The less aggressive cluster has the highest percentage of younger drivers and the majority of older drivers are concentrated in the cluster of moderate drivers. There is an indication that more aggressive drivers have a higher accident involvement.

The implementation of this research contributes to the discovery of existing driver profiles. It was observed that the K-Means algorithm identifies the main driver profiles but also presents some disadvantages such as that it necessarily assigns all drivers to clusters. On the other hand, the OPTICS algorithm identified driver profiles with very similar characteristics but presented some difficulties in the process of identifying more aggressive profiles with higher variability. It is suggested for future research to investigate whether combining the results of these two types of clustering algorithms could lead to improved overall results. Especially in order to deep dive into more aggressive driver profiles, it is recommended to use a larger sample of drivers, which will provide a clearer picture for all profiles as well. Incorporating more driving behavior characteristics as clustering features such as mobile phone usage and drowsiness level while driving, and analyzing profiles separately in different road types and time periods would also be very insightful in terms of the number and types of profiles. Finally, future studies should define and optimize parameter tuning (minimum number of points and maximum radius per cluster) for the OPTICS algorithm.

The characteristics of each driver and driver profile identified can potentially be used to develop applications that can support drivers and reduce crash risk. According to the results of the OPTICS algorithm, many drivers present unique behavior that cannot be grouped together with other drivers into profiles. Therefore, those drivers not belonging to profiles should be treated individually when developing mechanisms for feedback and interventions. For instance, in a case where a driver is being monitored and classified as aggressive driver with high number of harsh events, a feedback in the form of a message could be provided through a Smartphone application suggesting that the number of should be decreased to become less risky. Another example in the same direction could be real-time warnings provided by the vehicle's ADAS to a driver of the same aggressive profile, when a harsh event is foreseen based on the driver's macroscopic behavior.

The information on driver profiles could be used to predict the future state of a driver profile. Having quantified the probability of involvement into a road crash for each driver profile, the results of this study could be exploited from a traffic manager to quantify the total risk of a traffic network on the basis of the measured indicators, and their future changes. It would be important to investigate driver profiles in separate road types as well as to examine whether driver profiles differ from country to country. Each driver profile should be further investigated to understand which specific driving patterns are associated with each of them.

CRediT authorship contribution statement

Dimitrios I. Tselentis: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Eleonora Papadimitriou:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101027222.

The authors would like to thank the iDreams consortium, for providing all necessary data to accomplish this study, and the Technical University of Munich in particular for their valuable support in the data processing. The iDreams research project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814761.

An earlier version of this study was accepted for presentation at the World Conference on Transport Research 2023.

References

"i-Dreams: Safety Tolerance zone calculation and interventions for diver – vehicle – environment interactions under challenging conditions" of the Horizon 2020 framework programme on transport research of the European Commission (2019-2022).

"Rhapsody: Recognition of HumAn PatternS of Optimal Driving for safetY of conventional and autonomous vehicles" of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie scheme (2021-2023).

Abdulwahid, S.N., Mahmoud, M.A., Ibrahim, N., Zaidan, B.B., Ameen, H.A., 2022. Modeling Motorcyclists' Aggressive Driving Behavior Using Computational and

Statistical Analysis of Real-Time Driving Data to Improve Road Safety and Reduce Accidents. Int. J. Environ. Res. Public Health 19 (13), 7704.

Agrawal, K.P., Garg, S., Sharma, S., Patel, P., 2016. Development and validation of OPTICS based spatio-temporal clustering technique. Inf. Sci. 369, 388–401.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: Ordering Points To Identify the Clustering Structure. SIGMOD Rec. 28 (2), 49–60.

Bergasa, L. M., Araluce, J., Romera, E., Barea, R., López-Guilén, E., del Egido, J., & Hernanz-Mayoral, C. A. (2019, October). Naturalistic Driving Study for Older Drivers based on the DriveSafe App. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 1574-1579). IEEE.Deng, Z., Hu, Y., Zhu, M., Huang, X., & Du, B. (2015). A scalable and fast OPTICS for clustering trajectory big data. Cluster Computing, 18(2), 549-562.

Chronis, C., Sardianos, C., Varlamis, I., Michail, D., Tserpes, K., 2021. November). A driving profile recommender system for autonomous driving using sensor data and reinforcement learning. In: In 25th Pan-Hellenic Conference on Informatics, pp. 33–38.

Constantinescu, Z., Marinoiu, C., Vladoiu, M., 2010. Driving style analysis using data mining techniques. Int. J. Comput. Commun. Control 5 (5), 654–663.

Deng, Z.e., Hu, Y., Zhu, M., Huang, X., Du, B.o., 2015. A scalable and fast OPTICS for clustering trajectory big data. Clust. Comput. 18 (2), 549–562.

Duan, L., Xu, L., Guo, F., Lee, J., Yan, B., 2007. A local-density based spatial clustering algorithm with noise. Inf. Syst. 32 (7), 978–986.

Ellison, A.B., Greaves, S.P., Bliemer, M.C., 2015. Driver behaviour profiles for road safety analysis. Accid. Anal. Prev. 76, 118–132.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Fugiglando, U., Massaro, E., Santi, P., Milardo, S., Abida, K., Stahlmann, R., Netter, F., Ratti, C., 2018. Driving behavior analysis through CAN bus data in an uncontrolled environment. IEEE Trans. Intell. Transp. Syst. 20 (2), 737–748.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 28 (1), 100–108.

Islam, M.R., Jenny, I.J., Nayon, M., Islam, M.R., Amiruzzaman, M., Abdullah-Al-Wadud, M., 2021. In: August). Clustering algorithms to analyze the road traffic crashes. IEEE, pp. 1–6.

Kim, K., Yamashita, E.Y., 2007. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transpo. 41 (1), 69–89.

Levin, C., & Håkansson, C. (2015). Clustering driver's destinations-using internal evaluation to adaptively set parameters.

Li, N., Misu, T., Miranda, A., 2016. In: November). Driver behavior event detection for manual annotation by clustering of the driver physiological signals. IEEE, pp. 2583–2588.

Liao, X., Mehrotra, S., Ho, S., 2022. Yuki Gorospe, Xingwei Wu, and Teruhisa Mistu. "Driver Profile Modeling Based on Driving Style, Personality Traits, and Mood States.". In: In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 709–716.

Liu, C., Cai, J., Wang, D., Tang, J., Wang, L., Chen, H., Xiao, Z., 2021. Understanding the regular travel behavior of private vehicles: an empirical evaluation and a semi-supervised model. IEEE Sens. J. 21 (17), 19078–19090.

Malzer, C., Baum, M., 2020. In: September). A hybrid approach to hierarchical density-based cluster selection. IEEE, pp. 223–228.

Mantouka, E.G., Vlahogianni, E.I., 2022. Deep reinforcement learning for personalized driving recommendations to mitigate aggressiveness and riskiness: modeling and impact assessment. Transportation Res. Part C: Emerging Technol. 142, 103770.

Mantouka, E.G., Barmpounakis, E.N., Vlahogianni, E.I., 2019. Identifying driving safety profiles from smartphone data using unsupervised learning. Saf. Sci. 119, 84–90.

Mukherjee, S., Wallace, A.M., Wang, S., 2021. Predicting Vehicle Behavior Using Automotive Radar and Recurrent Neural Networks. IEEE Open Journal of Intelligent Transportation Systems 2, 254–268.

Nilsson, N.J. (Ed.), 1982. Principles of Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nouh, R., Singh, M., Singh, D., 2021. SafeDrive: Hybrid recommendation system architecture for early safety predication using Internet of Vehicles. Sensors 21 (11), 3893.

Papadimitriou, E., Argyropoulou, A., Tselentis, D.I., Yannis, G., 2019. Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. Saf. Sci. 119, 91–97.

Payyanadan, R.P., Angell, L.S., 2022. A Framework for Building Comprehensive Driver Profiles. Information 13 (2), 61.

Pei, T., Jasra, A., Hand, D.J., Zhu, A.X., Zhou, C., 2009. DECODE: a new method for discovering clusters of different densities in spatial data. Data Min. Knowl. Disc. 18 (3), 337–369.

Saleh, K., Hossny, M., Nahavandi, S., 2017. In: October). Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. IEEE, pp. 1–6.

Sanjurjo-De-No, A., Arenas-Ramírez, B., Mira, J., Aparicio-Izquierdo, F., 2020. Driver pattern identification in road crashes in spain. IEEE Access 8, 182014–182025.

Savelonas, M., Mantzekis, D., Labiris, N., Tsakiri, A., Karkanis, S., Spyrou, E., 2020. In: October). Hybrid Time-series Representation for the Classification of Driving Behaviour. IEEE, pp. 1–6.

Stavrakaki, A.M., Tselentis, D.I., Barmpounakis, E., Vlahogianni, E.I., Yannis, G., 2020. Estimating the necessary amount of driving data for assessing driving behavior. Sensors 20 (9), 2600.

Tselentis, D.I., Papadimitriou, E., 2023. Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions. IEEE Open. J. Intell. Transp. Syst. 4, 83–100.

Tselentis, D., Vlahogianni, E., Yannis, G., 2019. Driving safety efficiency benchmarking using smartphone data. Transp. Res. C 109, 343–357.

Tselentis, D.I., Vlahogianni, E.I., Yannis, G., 2021. Temporal analysis of driving efficiency using smartphone data. Accid. Anal. Prev. 154, 106081.

Vogel, M., Hamon, R., Lozenguez, G., Merchez, L., Abry, P., Barnier, J., Borgnat, P., Flandrin, P., Mallon, I., Robardet, C., 2014. From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset. J. Transp. Geogr. 41, 280–291.

Warren, J., Lipkowitz, J., Sokolov, V., 2019. Clusters of driving behavior from observational smartphone data. IEEE Intell. Transp. Syst. Mag. 11 (3), 171–180.

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

Wen, X., Fu, L., Fu, T., Pan, X., Zhong, M., 2021. In: October). Driver Behavior Analysis at Unsignalized Intersections with or without Stop Sign Using Trajectory Data. IEEE, pp. 1470–1475.