An Unsupervised Approach for Driving Behavior Analysis of Professional Truck Drivers

Sebastiano Milardo[1(✉)], Punit Rathore[1], Paolo Santi[1], Richard Buteau[2], and Carlo Ratti[1]

[1] Senseable City Lab, Massachusetts Institute of Technology, Cambridge 02139, USA
{milardo,prathore,psanti}@mit.edu
[2] Dover Corporation, Illinois 60515, USA
rbuteau@dovercorp.com

Abstract. Modern vehicles can generate up to several Gigabytes of data per day which are mostly used only for aspects directly related to the proper functioning of the vehicle itself. However, these data have an enormous value as they can be collected and analyzed to better understand additional aspects of the driving experience, such as classifying the driver's behavior and driving style.

In this paper, we present a simple yet novel unsupervised methodology that is able to classify the behavior of a driver in a certain geographical area on the basis of the data collected from all the drivers in the same area. The proposed methodology has been tested on two different datasets involving professional truck drivers and it has been verified using human labelled ground truth data. The results obtained demonstrate the feasibility of the proposed solution. To our knowledge, this is the first study to classify driving behaviours of professional truck drivers and validate their performance on such large-scale data with actual safety scores.

Keywords: Driving behaviour classification · Driving style recognition

1 Introduction

In recent years, driver behavior monitoring and characterization have evolved tremendously due to their importance in traffic safety. The emergence of smart and connected vehicles has amplified the need to understand and characterize driver's individual behaviors as driving behavior recognition is now seen as a fundamental step in the design of Advanced Driver Assistance System (ADAS).

Human driving behavior is a complicated concept and the common association of "driving behavior" with "driving style" complicates its definition even further. Specifically, driving behavior focuses solely on drivers' instantaneous decisions made in response to external environmental conditions such as road type, surrounding traffic, weather, etc. These instantaneous driving decisions

A. L. Martins et al. (Eds.): INTSYS 2021, LNICST 426, pp. 44–56, 2022.
https://doi.org/10.1007/978-3-030-97603-3_4

result from a complex fusion of different factors. Whereas, driving style concerns the way a driver chooses to drive i.e. individual driving habits. This article focuses on driving behaviour analysis.

Modern vehicles can collect massive amounts of data about the vehicle itself, the driver, and the surrounding environment [17]. Effective analysis of these data provides us opportunities to efficiently model driving behavior.

Many studies have focused on classifying driving behavior for road safety. Most of them are developed on the basis of classification models and/or extracted features. These classification techniques are usually implemented using either traditional machine learning algorithms on extracted feature sets, statistical models or a set of rules. In most studies, driving data were collected via controlled/simulated-based experiments or naturalistic driving studies with a few drivers. Moreover, many studies validate their classification models to classify individual driving events (e.g. harsh turn, harsh braking, etc.) rather than overall driving behavior of a driver (e.g. safe driver, harsh driver) as it is usually hard and time-consuming to manually evaluate actual driving behavior of a driver.

Although driving behavior classification from naturalistic driving data is well studied, their characterization is still an open problem which leads to a broader question of "what are the indicators that allow us to characterize a driving behavior as positive or negative?" In this sense, many studies answer this question through the absolute metrics of evaluation as contextualizing driving events is a difficult task. For example, excessive speed is generally associated with car accidents therefore it is logical to classify all speeding events as negative. Contrarily, a harsh braking can be correlated both to an aggressive driving style or a spontaneous reaction to avoid an unpredictable hindrance. Therefore, a context-aware study of driving behavior that goes beyond individual events is required.

In this paper, we move the focus from discrete evaluations, where every driving event is evaluated independently, to global evaluations of driving behavior, assuming that the (average) driving behavior of most drivers on a road segment under the same environmental conditions remains similar. There exists many studies on driving behavior analysis for taxi/car drivers; however, very limited work focused on truck drivers' behavior analysis through real-world driving data. Since most truck drivers usually drive for many consecutive hours in different traffic and weather conditions, unsafe driving behaviors are more likely to occur. Therefore, timely detection of a driver's driving style is fundamental. Developing a simple, explainable yet accurate approach for driving behavior classification is our primary objective here. The major contributions of this research are as follows:

- We present a simple unsupervised approach to classify driving behaviors of professional truck drivers by analyzing Controller Area Network (CAN) bus signals, such as speed, lateral (right and left movements) and longitudinal accelerations (acceleration and braking).
- We capture the behavioral evolution of drivers from their instantaneous response to overall driving habits. We categorize instantaneous driving behavior into four categories (scores): poor (0), average (1), good (2), and excellent (3). Then, the overall score for a driver as a driving habit is computed by con-

tinuously accumulating instantaneous classifications gathered from his/her previous trips.
– We validate the detected events (e.g. aggressive acceleration, harsh braking etc.) and overall driving behavior score for each driver (an indicator of driving habit) against the actual harsh driving events and actual driver safety scores, respectively. These events and scores were provided by a group of domain experts based on their routine observation for each driver over a period of 3 months. To our knowledge, this is one of the few datasets that provides realistic feedback about the driving behavior and driving events for validation.

Very few studies have focused on developing scoring functions that accurately reflect the actual driving behavior. Indeed, the choice of scoring function is very subjective due to the lack of large-scale and reliable datasets with actual behavior information. To the best of our knowledge, this is the first attempt to understand and classify driving behavior of professional truck drivers and validate their performance on the basis of large-scale data and actual safety scores. With a simple and explainable framework, our model not only contributes to academic research, but it can be of high relevance for the automotive industry.

The rest of the paper is organized as follows: Sect. 2 summarizes the literature on driving behavior analysis, particularly on unsupervised techniques. Section 3 describes the characteristics of the data collected. The proposed methodology is reported in Sect. 4, and the validation with the provided ground-truth data is detailed in Sect. 5. Finally conclusions are drawn in Sect. 6.

2 Related Work

A comprehensive survey on driving behavior analysis is provided in [4,12,19]. Most techniques on driving behavior classification can be divided in two categories: unsupervised and supervised. In the first type of approach, driving behavior is classified through statistical analysis or using unsupervised machine learning (ML) techniques, without the knowledge of actual classification. Whereas, in the second case, knowledge of actual driving behavior classification is used for training the underlying ML model. Since our presented method in this work is unsupervised, we restrict our discussion to the first type of approach, which we deem pertinent to this article.

Among unsupervised approaches, k-means, Gaussian mixture model (GMM), and Bayesian learning techniques have been used extensively for driving behavior recognition. Constantinescu *et al.* [5] approached two types of methods: hierarchical cluster analysis and principal components analysis (PCA). First, they employed Ward's hierarchical clustering method with Euclidean distance measure to identify groups of drivers based on similarities in the driving features. Then, they used PCA to project the original dataset onto a lower-dimensional

space by extracting principal components (PCs). Further, they analyze the correlations between PCs and driving variables to estimate the significant PCs and plot data onto them to visualize clusters of different driving behaviors.

In [14], driver characteristics in car-following and pedal operational patterns were modeled using GMM and spectral analysis, respectively. The GMM model achieved 69% identification rate, while pedal spectral analysis achieved a classification rate of 89.6% in a simulated driving environment and 76.8% in a field test. Castigani *et al.* [3] presented a driving profiling platform, SensorFleet, to detect risky driving events from smartphone data using an adaptive profiling mechanism. Specifically, a fuzzy logic algorithm was implemented to compute the scores for different drivers using context information like road topology and weather. However, it was based on a 20 min calibration phase on a pre-defined path, which may not reflect naturalistic driving behavior.

Bender *et al.* [1] presented a Bayesian method for segmenting the naturalistic driving data into high-level driving behaviors. This study considered inertial data collected from a 13 minute drive by a single driver. Brambilla *et al.* [2] employed three unsupervised approaches: DP-means clustering, Hidden Markov Models (HMMs), and behavioral Topic Extraction, to detect different behavior along each trip and subsequently classify drivers based on their driving behavior profiles. Fugiglando *et al.* [8,9] used the k-means clustering algorithm to identify groups of similar drivers using CAN bus data based on the driving behaviors. A study on the stability of these groups was reported, however, no semantic explanation for the different resulting classes was provided.

Mudgal *et al.* [15] employed a hierarchical Bayesian regression technique to model instantaneous driving behavior at roundabouts. Similarly, McCall *et al.* [13] utilized Bayesian learning to analyze the driving behavior for braking assistance and collision avoidance. Experiments in these studies were either conducted in a simulated environment or they considered data from few drivers. Wang *et al.* [18] proposed a framework for driving style classification by using primitive driving patterns with the Bayesian nonparametric approaches. The features used in [18] were the vehicle longitudinal acceleration, speed, and the distance from the preceding vehicle.

Although extensive work has been done on driving behaviour analysis of car drivers, very limited studies have analyzed the driving behaviour of professional truck drivers provided that trucks have different vehicle and driving dynamics compared to cars and motorbikes. Linkov *et al.* [11] presented a study on the correlation between professional drivers' driving behavior and their personality traits using a truck simulator. This study [11] classifies the drivers based on their mean speed and mean lateral position from the center of the lane. However, the primary focus of this study [11] was on fuel efficiency [10] rather than driving behavior, which was considered as an auxiliary variable. In another similar study, Ferreira *et al.* [7] collected data from professional bus drivers in Lisbon, and applied Naive Bayes classifier to optimize fuel consumption and provide suggestions. Some suggestions such as "Minimize the use of acceleration" and "Minimize the use of braking" are generally related to both efficient fuel consumption and good driving behaviors.

3 Dataset

To classify the behavior of a driver using a statistical approach, it is fundamental to collect a meaningful and extended dataset. Fortunately, modern vehicles produce a huge amount of data that is rarely used outside of the activities needed by the vehicle itself. These data can be easily collected through the use of vehicular networks such as the CAN-Bus network. This standard, developed by Bosch, replaces the traditional point-to-point connections with a Bus topology in which a shared transmission medium is used to connect the electronic control units of the vehicle. This particular topology allows us to collect transmitted messages without interfering with the normal functioning of the vehicle. There are three main CAN-Bus standards used in modern vehicles:

- J1939: for heavy duty vehicles
- OpenCan, for robots and automation
- OBD2, for general vehicles

In this paper, as we focus mainly on trucks, we have leveraged the J1939 standard [17]. As different vehicles can produce different messages, we have decided to limit our analysis to the smallest set of common signals available from all the trucks in our dataset. We called this subset the *heartbeat* dataset.

This dataset contains information related to the GPS position of the vehicle, its speed, and the accelerations collected by the accelerometer mounted inside the cabin. The data is collected with a sampling rate 1 Hz. However, while the GPS is just sampled every second, the values reported for the acceleration are the averages of all the values measured during a second by an accelerometer working at a higher sampling rate (500 Hz).

The heartbeat data has been collected by professional truck drivers around the city of Trenton, NJ, USA during two different time intervals. The first dataset contains 54 million data points collected by 41 trucks and 34 drivers, from April 2019 to the end of June 2019, while the second dataset is made of 18 million data points collected during September 2020 from 37 vehicles and 22 drivers. In the remainder of this paper, we will refer to these datasets as *2019 dataset* and *2020 dataset* respectively. To validate the proposed solution we have used two different approaches:

3.1 2019 Dataset Ground Truth

The 2019 dataset includes a synthetic score for each driver. This score has been generated by domain experts based on values not contained in the analyzed dataset, but based on

- Number of tickets.
- Number of accidents.
- Direct observation of the drivers' behavior.

The scores assigned to the drivers are divided into 6 categories from A to F where A represents the best driving behavior and F the worst. In our dataset, there are no drivers with a score equal to E and there are only two drivers with a score equal to F. The distribution of the scores is reported in Fig. 1.

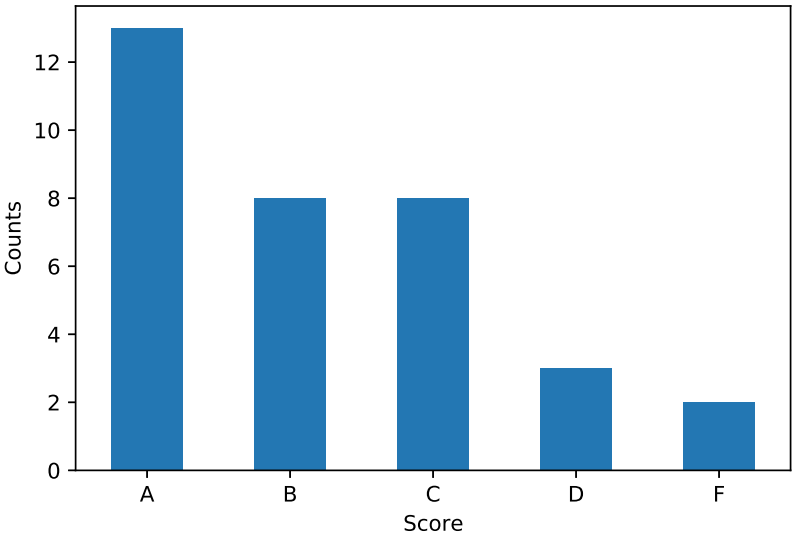

Fig. 1. Distribution of the Ground truth scores for the 2019 dataset.

3.2 2020 Dataset Ground Truth

The 2020 dataset provides a list of so-called *coachable events*. The activity of the drivers participating in the collection of the 2020 dataset was constantly monitored through cameras. When a trigger was fired (lateral acceleration >0.4 g or frontal acceleration >0.3 g), the video signal together with the accelerometric information were recorded and sent to the internal review process for manual verification by a supervisor. We call the data collected within a one minute window around the trigger event a *coachable event*. A summary of the list of coachable events is reported in Table 1.

Table 1. Coachable events available for 2020 Dataset

	Number of detected events
Harsh braking events	3638
Harsh turning events	436
Harsh acceleration events	274

4 Proposed Solution

The proposed approach is an evolution of a standard methodology used for the identification of harsh events based on thresholds [6]. Specifically, in [6], fixed predefined thresholds are used to classify each acceleration as safe or not. The main drawback of this approach, which is extremely simple to implement and deploy on embedded devices, is that it lacks additional contextual information that can be obtained from the vehicle itself.

For example, Fig. 2 shows that, given the dynamics of a truck, as the vehicle speed increases, accelerations that are very likely at lower speeds turn out to be extremely rare at higher speeds. Such correlation is not captured by a simple approach based on fixed thresholds. Additionally, by including spatial information and by leveraging the knowledge generated by a large fleet of vehicles, it is possible to improve even further the characterization process of the recorded accelerations.

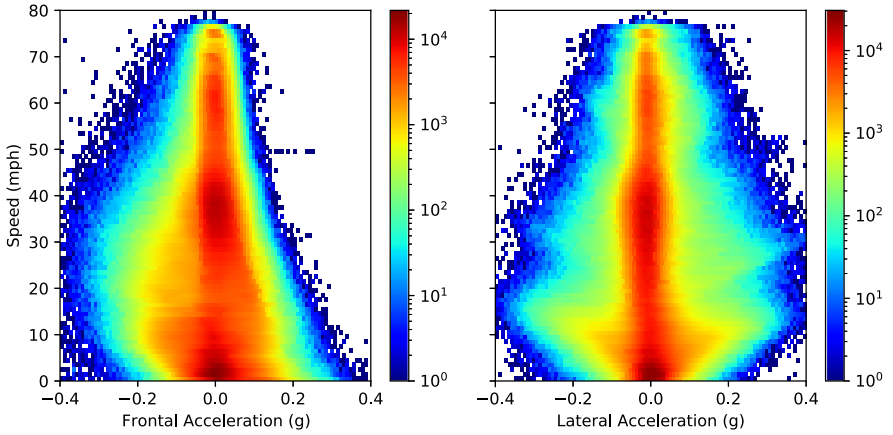

Fig. 2. Dataset 2019: Speed vs Frontal and Lateral accelerations

In fact, a lateral acceleration that can be considered acceptable on a curve may not be necessarily acceptable on a straight road segment and can therefore be considered an alarm or an event that should be investigated. In order to introduce crowd collected spatial information into this analysis, we developed a novel algorithm which, through the aggregation of information from different trucks, allows the creation of a geo-referenced model to classify the behavior of a driver.

The proposed solution is based on three different steps.

- Pre-processing
- Data Aggregation
- Scoring

4.1 Pre-processing

First, the collected data is filtered to remove errors or out of range values. For example, the GPS position can be inaccurate at the beginning of a recording or if there are not enough satellites. During pre-processing, the intervals when the vehicle is not moving are also removed from the dataset. Then, the dataset is segmented based on the path followed by the truck. To achieve this result, we used a geohashing algorithm.

Geohashes are a system used to encode a geographical position into a string. Each geohash algorithm ensures that:

- If two geohashes share a common prefix then the two points will be close.
- The longer the length of the prefix shared between the two points, the closer they will be.

In our proposed approach, we decided to use a geohash algorithm that aggregates GPS data points with a precision of 10m. This approach allows us to aggregate all the data points in our dataset on a $10m \times 10m$ grid, in a simple and efficient way. However, this simple aggregation method is optimal only on one-way road segments. In fact, more complex scenarios can result in the aggregation of data points collected by trucks moving on different paths.

An example is shown in Fig. 3. The upper two images show the 2D histogram of the frontal and lateral acceleration collected inside cell number 24745222599 for vehicles coming from cell number 24745222598 and moving towards cell number 24745222600. The bottom two images show the data points collected for vehicles moving in the opposite direction. As we will discuss it in detail in the subsequent paragraphs, we classify a driver based on how different his/her driving behaviour is from average behaviour of all the drivers in a cell.

In such a scenario, the average aggregated value for each driving attribute (such as speed, acceleration) will be different based on the routes followed by drivers as we mentioned above, and consequently, this may lead to conflicting results in our solution. Therefore, we compressed the path followed by a truck to a list of geohashes by removing consecutive duplicate geohashes and then associating a new feature to each data point in the dataset by concatenating the geohash at the current position with the previous and the next geohash in the path followed by the truck. We called this new feature *cell-ID*.

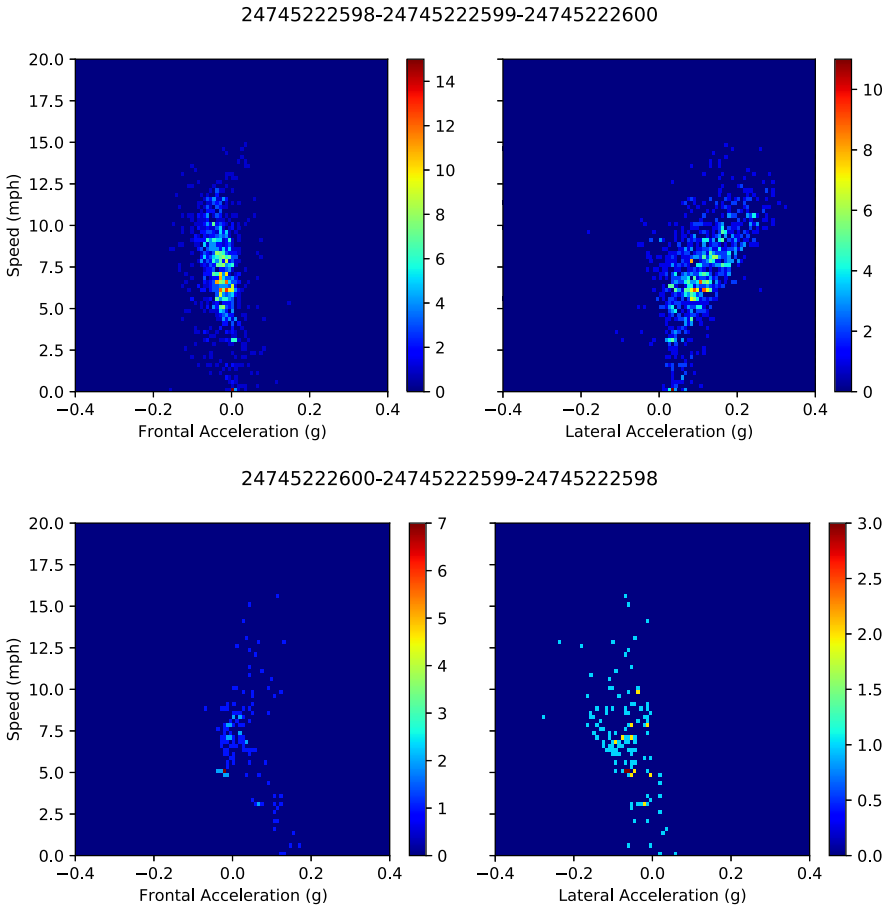

Fig. 3. Data points collected inside the same cell (24745222599) can show different characteristics based on the trajectory of the vehicle.

4.2 Data Aggregation

Next, we aggregate all the data points for each cell ID. This is done independently for the positive frontal acceleration, negative frontal acceleration, positive lateral acceleration, negative lateral acceleration, and speed signals. Specifically, the signals obtained from the frontal and lateral accelerations are divided according to speed in different bins. In particular, we created 8 bins equally distributed between 0 and 80 mph. This binned data is then used to calculate the mean and the standard deviation of each signal for each bin. At the end, the output of this step is a data structure containing the mean and standard deviation values of each signal for each bin for each cell ID. We call this data structure as *cell descriptor*.

4.3 Scoring

Once this data structure has been created, a model is trained on a given dataset that is used to classify the behavior of a driver within a given geographical area. In particular, in order to achieve a score that is representative of the driver's behavior and at the same time easy to interpret, we decided to transform each acceleration and speed value into a score between 0 and 3, following the *68-95-99.7 rule* on standard deviation, where 0 is assigned to values within one standard deviation from the mean of the data collected in the selected bin, and 3 for values outside the $[\mu - 3\sigma, \mu + 3\sigma]$ interval. Then, for each data point, we computed a maximum score that is equal to the maximum value among the scores collected for the signals analyzed. Finally, a trip (or a day) score of a driver is the weighted average of all the maximum scores obtained during the trip (or day). The weight used is a measure of the confidence for the specific score and it is equal to the number of points in the bin used to compute the score.

5 Evaluation

In order to verify the validity of the results obtained, we compared the scores calculated using the proposed method on the 2019 dataset with the scores assigned by the domain experts. For each hour in the dataset we computed the score and we normalized it between 0 and 5, where 0 corresponds to F and 5 to A. Then we matched these results with the ground truth by rounding the obtained score to the closest integer. The results are summarized in Fig. 4. As it is possible to notice, most of the drivers are scored correctly and the error is generally limited to the previous or next class. Of course there can be exceptions as the behavior of a driver can change over time but in general there is a good correlation between the computed and the provided ground truth score (MSE = 1.69).

However, to better understand the variability in the behavior of each driver, and to estimate the trade-off between the amount of data collected and the accuracy of the proposed solution we analyzed how the score changes when we analyze different amounts of data for a selected driver. Specifically, for each driver in the data set we randomly selected $10, 20, ...500$ hours of driving, we repeated this process 1000 times and we computed the average score and the standard deviation. As it is possible to notice from Fig. 5, the standard deviation for all drivers in the dataset follows a power law, with exponent in the -0.3 ± 0.01 range. In general, after 200 hours of driving the score computed can be considered reliable.

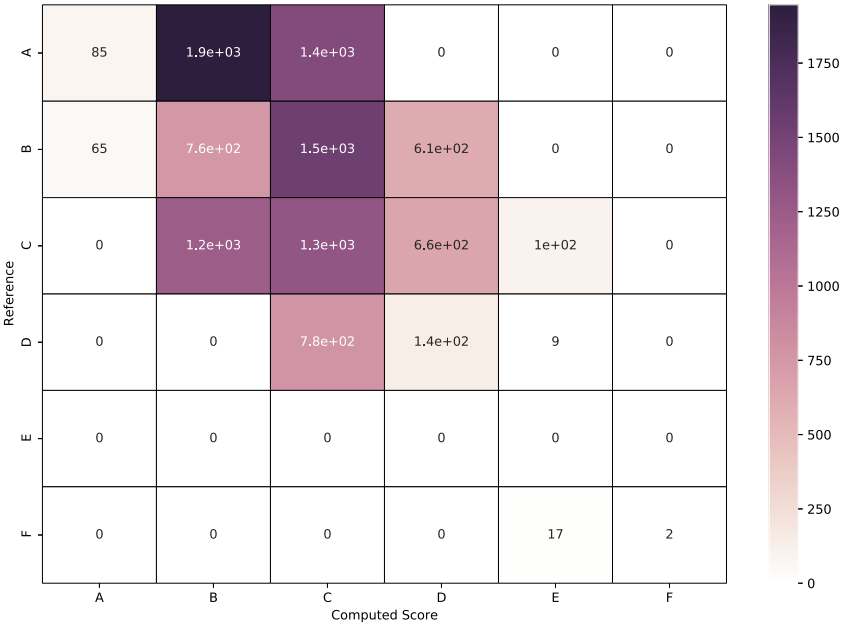

Fig. 4. The confusion matrix resulting from the comparison between the scores computed with the proposed solution and the reference scores for the 2019 dataset.

A further aspect that we addressed is the generalizability of the proposed solution. In fact, once the model is created, we verified that this model can be used to classify a different dataset collected over the same spatial region. Therefore, to verify this hypothesis, we classified the 2020 dataset by using the cell-descriptor generated using the 2019 dataset.

As the 2020 dataset does not contain a synthetic score (as ground truth) for each driver, we compared the proposed solution with the coachable events by computing how many coachable events are correctly detected. Specifically, all 4348 coachable events (shown in Table I) were correctly classified as harsh events (score ≥ 2), while the number of false positives was equal to 648, thus resulting in a precision of 0.87, a recall of 1 and an $F1$ score equals to 0.93.

Finally, it is important to emphasize that the proposed approach heavily relies on the availability of data for a specific area. As reported in [16], few vehicles moving randomly in an urban area can cover most of it in a relatively small amount of time but some areas might remain uncovered for longer periods. However, in the case of the data analyzed in this paper, the trajectories followed by trucks are not random, and the paths followed by the drivers allowed us to easily collect enough data from multiple drivers for each location in our dataset.

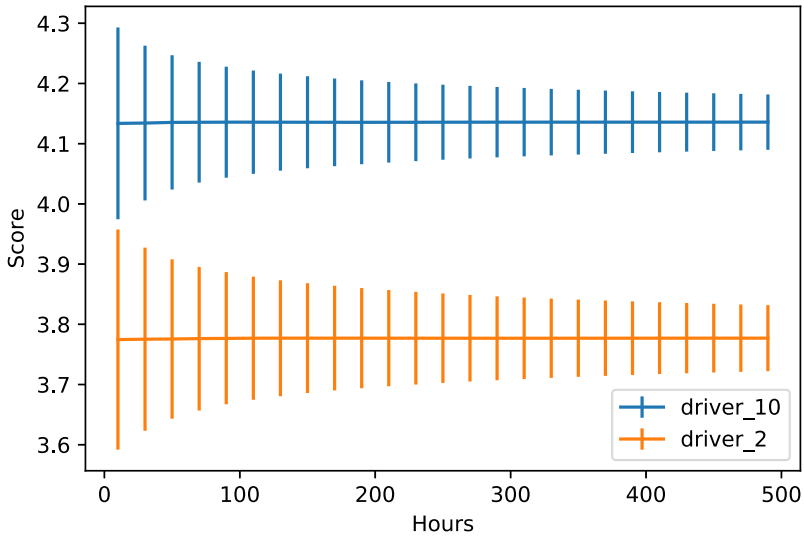

Fig. 5. The average and standard deviation of the computed scores derived by selecting different amounts of data for two sample drivers.

6 Conclusions

In this paper, we have presented a simple yet novel solution to classify the driving behavior of professional truck drivers using an unsupervised approach. We verified through the use of coachable events and synthetic scores that our methodology is generalizable and is able to detect and classify both instantaneous and global behaviors. Additionally, by aggregating the data collected from the vehicles using geographical information, the proposed approach allows not only to classify and model driving behavior but at the same time it can be used to map and underline dangerous or challenging road segments.

As shown in this paper, while the data collected allow us to detect events and label the behavior of drivers, our future works will focus on obtaining higher resolution data and additional signals, such as the steering wheel angle or the pedal position, in order to characterize and aggregate micro-events (e.g. minor steering wheel corrections) in order to focus on additional dangerous behaviors, such as distracted driving,

References

1. Bender, A., Agamennoni, G., Ward, J.R., Worrall, S., Nebot, E.M.: An unsupervised approach for inferring driver behavior from naturalistic driving data. IEEE Trans. Intell. Transp. Syst. **16**(6), 3325–3336 (2015)
2. Brambilla, M., Mascetti, P., Mauri, A.: Comparison of different driving style analysis approaches based on trip segmentation over GPS information. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 3784–3791. IEEE (2017)

3. Castignani, G., Derrmann, T., Frank, R., Engel, T.: Driver behavior profiling using smartphones: a low-cost platform for driver monitoring. IEEE Intell. Transp. Syst. Mag. **7**(1), 91–102 (2015)
4. Chan, T.K., Chin, C.S., Chen, H., Zhong, X.: A comprehensive review of driver behavior analysis utilizing smartphones. IEEE Trans. Intell. Transp. Syst. (2019)
5. Constantinescu, Z., Marinoiu, C., Vladoiu, M.: Driving style analysis using data mining techniques. Int. J. Comput. Commun. Control **5**(5), 654–663 (2010)
6. Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., González, M.C.: Safe driving using mobile phones. IEEE Trans. Intell. Transp. Syst. **13**(3), 1462–1468 (2012). https://doi.org/10.1109/TITS.2012.2187640
7. Ferreira, J.C., de Almeida, J., da Silva, A.R.: The impact of driving styles on fuel consumption: a data-warehouse-and-data-mining-based discovery process. IEEE Trans. Intell. Transp. Syst. **16**(5), 2653–2662 (2015). https://doi.org/10.1109/TITS.2015.2414663
8. Fugiglando, U., et al.: Driving behavior analysis through CAN bus data in an uncontrolled environment. IEEE Trans. Intell. Transp. Syst. (2019). https://doi.org/10.1109/TITS.2018.2836308
9. Fugiglando, U., Santi, P., Milardo, S., Abida, K., Ratti, C.: Characterizing the "driver DNA" through can bus data analysis. In: Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services, pp. 37–41. CarSys 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3131944.3133939
10. Hlasny, T., Fanti, M.P., Mangini, A.M., Rotunno, G., Turchiano, B.: Optimal fuel consumption for heavy trucks: a review. In: 2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), pp. 80–85, September 2017. https://doi.org/10.1109/SOLI.2017.8120974
11. Linkov, V., Zaoral, A., Řezáč, P., Pai, C.W.: Personality and professional drivers' driving behavior. Transp. Res. Part F Traffic Psychol. Behav. **60**, 105–110 (2019)
12. Marina Martinez, C., Heucke, M., Wang, F., Gao, B., Cao, D.: Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey. IEEE Trans. Intell. Transp. Syst. **19**(3), 666–676 (2018). https://doi.org/10.1109/TITS.2017.2706978
13. McCall, J.C., Trivedi, M.M.: Driver behavior and situation aware brake assistance for intelligent vehicles. Proc. IEEE **95**(2), 374–387 (2007)
14. Miyajima, C., et al.: Driver modeling based on driving behavior and its evaluation in driver identification. Proc. IEEE **95**(2), 427–437 (2007)
15. Mudgal, A., Hallmark, S., Carriquiry, A., Gkritza, K.: Driving behavior at a roundabout: a hierarchical Bayesian regression analysis. Transp. Res. Part D: Transp. Environ. **26**, 20–26 (2014)
16. O'Keeffe, K., Santi, P., Wang, B., Ratti, C.: Urban sensing as a random search process. Phys. A Statist. Mech. Appl. **562**, 125307 (2021)
17. Simma Software Inc: Understanding SAE j1939 (2021). http://www.simmasoftware.com/j1939-presentation.pdf
18. Wang, W., Xi, J., Zhao, D.: Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches. IEEE Trans. Intell. Transp. Syst. **20**(8), 2986–2998 (2019). https://doi.org/10.1109/TITS.2018.2870525
19. Wang, W., Xi, J., Chen, H.: Modeling and recognizing driver behavior based on driving data: a survey. Mathematical Problems in Engineering 2014 (2014)