Driver profile classification on heavy trucks based on telemetry

(extended abstract of the MSc dissertation)

Daniel Janeiro Alves
Instituto Superior Tecnico
daniel.alves@tecnico.ulisboa.pt

Abstract—This work proposes a method to create profiles for heavy-duty truck drivers describing their driving efficiency. It works with time series (TS) data about trips collected using telemetry technology. It also investigates the current approaches in the creation of driving profiles and grouping of time series data. Some example analysis are made using the proposed method with many validation metrics.

The proposed solution (used in the analysis) groups the drivers based on the distributions of the main features. It is described the method to classify a new driver based on the profiles identified earlier.

This method gives information about the weakness and strengths of drivers. The logistics’ companies can use this information to better train their drivers, allocate them to trips smartly, measure and reward their progress.

I. INTRODUCTION

A. Motivation

We live in a society drastically becoming more industrialized and the climate effects caused by that change have been accentuating. Hence, its combat is becoming increasingly more important, not just for everyday people and governments, but also for the industries. One of the industries that contributes heavily to these effects is the automobile one, where the companies try to minimize their emissions using technological advancements. In the logistics’ industry these advancements only help so far, so the companies introduce smart routing and try to educate their drivers with eco-driving techniques.

Eco-driving is defined as the set of techniques that, when applied by the driver, promote fuel savings and reductions in pollute emissions. These reductions are expected at a relatively low cost and without compromising road safety nor travel time.

What most companies do regarding eco-driving is simply promoting training sessions that teach good practices to their drivers. This strategy shows immediate benefits, however these benefits fade over time when the training stops. Studies show reductions in fuel consumption ranging from 10% to 20% in manual transmission vehicles just after the training, but in the long-term these values go down to 5% to 10% [16, 2].

In order to overcome this issue, we propose a study of drivers’ behaviors along with the creation of driving profiles that best describe them. Having these profiles, we aim to help finding the behavior changes, in each profile, that would best improve their performance, both in terms of fuel consumption and pollute emissions. In order to assure that drivers keep improving, the profiles found can be used to inform the driver about his/her weaknesses and how to overcome them.

B. Problem description

Our problem consists in analyzing multivariate time series (MTS) data. The dataset was acquired using telemetry technology, from heavy-duty truck fleets performing their routine tasks. Our data is composed of information about the vehicle such as the model, the weight of both the vehicle and the cargo and data about the state of the vehicle like the position of the pedals, instantaneous speed and GPS position, including the altitude.

We characterize this as a multivariate time series clustering problem, where our clusters must be able to separate different driving styles. Since we are working with real-life data, we also need to make sure the clusters are grouping the driver style and not some other external factors, like the road, the weather or time of the day.

Knowing that the style of the driver influences its consumption and security, what is the best way to group driver styles? And, how do we make sure our groups depend more on the drivers’ style than on external factors?

C. Contributions

To accomplish this work, we developed a method for creating profiles that describe driving behaviors on the different features we found important. Using these profiles, we analyzed the parameters that most influence the vehicle’s fuel consumption and therefore emissions. Consequently, the contributions of this work are subdivided into:

1) Method to create driving profiles using clustering algorithms.
2) Method to classify new drivers into the profiles identified.
3) Metric to evaluate the clustering and classification methods.
4) Examples of the method created being used.
5) Identification of the parameters that most influence the fuel consumption, by the analysis of the driving profiles previously created.

II. RELATED WORK

With the objective of better understanding the behavior of drivers in traffic, the research on the profiling of driving styles began.
Constantinescu et al. [3] proposed a method for analyzing the risk-proneness of driving styles, using principal component analysis and hierarchical algorithms. To monitor the vehicle’s position, it uses the Gipix system, with one-second acquisition interval. The Gipix system provides real time data for the instantaneous speed (v), altitude (h), acceleration (a) and GPS error (e). The authors worked with simulated data, meaning every driver had the same route.

To perform the statistical analysis, the following parameters were extracted from the data: $V_{60}$ to represent the percentage of time above 60 km/h, $V_{mn}$ and $V_{sd}$ to represent the mean value and standard deviation for the speed, $A_{+mn}$, $A_{+sd}$, $Br_{mn}$, $Br_{sd}$ to represent the means and standard deviation for the positive acceleration and braking and W, to represent the sum of all positive kinetic energy to increase vehicle speed over time. Using hierarchical clustering and analyzing the distance between the clusters in the dendrogram, the authors concluded that the best way to group the data was using 6 clusters. To analyze the resulting clusters, they used PCA (Principal Component Analysis) [14, 9] to obtain the two principal components that explained at least 80% of the variance of the data. Since the resulting principal components could not differentiate the use of the braking and accelerating pedals, they converted them into the rotated components. Through the analysis of the rotated components, they concluded that the drivers in each cluster had similar average velocity and similar accelerating and braking patterns, from very aggressive with a tendency of high speed combined with sudden braking, to non-aggressive with tendency to lower speeds.

Analyzing this method, it is relevant to notice that some external factors are missing like weather, road type and time of the day. All these factors affect the driver’s behavior and are relevant for the creation of a driving profile.

I. Obuhuma et al. [8] explored the use of probabilistic methodologies to profile drivers. This work used on-board GPS data gathered from one trip. The discrete data collected was converted into time slices of differences in speed, altitude, time and angle of direction. A Bayesian Network [10] and the Dynamic Bayesian Network [5] were used to classify each time slice into behavior and environment profiles. Behavior profiles are: normal braking, harsh braking, normal acceleration, harsh acceleration, normal cornering and harsh cornering, while the environment profiles are: meander, straight, up-hill and down-hill. The authors concluded the best way to set the thresholds for harsh and normal acceleration and braking was to use the Stopping Sight Distance, given by the sum of reaction/perception distance with the braking distance, meaning this threshold is different at different speeds and different vehicles. The final model had 64 states a driver could be in, where each state had its own probabilities for current state profile. The final classification of the driver is the average of the profiles he was in, during the trip.

This profile classification takes the sharpness of corners into consideration, which is a relevant parameter many times overlooked and also distinguishes different driving styles for the same trip. However, this approach has weak separation between positive and negative acceleration in the states, meaning the driver has the same probability of either braking or accelerating in each state. Moreover, the method was only tested using one trip, which is not enough to validate it.

Fugiglando et al. [6] had a similar problem to ours of creating driver profiles based on data from an uncontrolled environment. Over a span of 55 days they worked with 64 different drivers that were not instructed to follow any particular route or behavior while driving, resulting in 1987 driving sessions of 64 minutes on average. From the data acquired they used 8 variables: (1) brake pedal pressure, (2) acceleration pedal position, (3) engine speed, (4) vehicle speed, (5) steering wheel angle, (6) steering wheel momentum, (7) frontal acceleration and (8) lateral acceleration. Each of these variables was converted into 7 features: (1) values of the signal, (2) discrete first derivative of the signal, (3) time interval between two singular points (local minimum or maximum), (4) values of the local maxima (5) moving mean, (6) moving median and (7) moving standard deviation.

For each variable, they built a frequency histogram that represents how its values are distributed in the session. Having $w_{i,k,u}$ as the value of the variable $k$ for the user $u$ at the time $i$, they build a vector $W^k$ that represents every value of the variable $k$ of the set of drivers,

$$W^k = \bigcup_{u \in U} \bigcup_{i} \{w_{i,k,u}\},$$

where $U$ is the set of drivers. To create the histogram, they take the maximum and minimum values of $W^k$ and divide it into 10 equal size bins $b^k_1, ..., b^k_{10}$. Having the bins defined, each bar of the histogram can be computed as the number of points, whose values reside in the bins’ intervals. Finally, the bins are normalized to be the percentage of points inside.

To group the histograms of each variable they use the $k$–means algorithm, which requires the number of clusters as argument, consequently, they proposed a method using V-measure score [12] to obtain it. The v-measure score takes a list of real labels and uses it to evaluate the quality of another labeling method. It is a combination of the completeness and homogeneity scores. The completeness score evaluates if every object of a class is grouped together and the homogeneity score evaluates how many classes are in each cluster. Hence, if we define the number of clusters as 1, we get completeness 1 and homogeneity 0. If we define the number of clusters as the number of samples, we have completeness 0 and homogeneity 1. To get the number of clusters $k$ of a variable, they followed the following method: for each $k$ between 2 and 10 and for each driver, they create a validation vector using 70% of its data (of the variable being tested) selected randomly and a training vector with the remain 30%. Each driver gets 2 histograms (per variable), one created using the training vector and another using the validation one. For each number of clusters, two $k$–means are performed, one using the validation histograms and another
using the training histograms. Each $k$ gets a score given by the V-measure between the clusters produced by the two methods. The $k$ that produced the highest V-measure score is chosen.

The creation of the histograms in this work looks interesting. It allows for a more detailed analysis of the driver than other works and still solves the high dimensionality problem. In terms of the validation method they used, we think it is evaluating more the consistency of the driver or the environment than the quality of the clusters. In our work we, propose the use of the fuel consumption to help validate the clusters’ quality.

Feng et al. [4] proposed a new method to classify driving styles, dividing trips into segments (accelerating, braking, maintaining and stopped) using a dynamic sliding window. They worked with 3 drivers and on a total of 12 trips. The trips were in similar weather condition and time of the day to minimize potential variance caused by these external factors. PCA was used to extract the prominent factors and support vector clustering was used to classify the driving style during the trip. They used four statistical features: mean, standard deviation, maximum and minimum values on four input signals: vehicle speed, engine speed, pedal position and headway distance. Headway distance is the distance to the next car obtained with a continental radar and a monocular camera.

Inside each segment, the data is classified into 3 clusters, aggressive, normal and defensive. Regarding fuel consumption, the expected conclusions were made, the drivers with the highest percentage of defensive driving had the lowest fuel consumption on average and vice-versa, with an interesting finding. When a defensive driver drives in an aggressive way it consumes, on average, more fuel than an aggressive driver.

In this method, it is relevant to look at the segmentation made in the pre-processing phase. This allows the separation of styles during the trip, it does not assume the driver’s style is the same during the entire trip. However, this method was only used on a small population of 3 drivers, we think this is not enough to assess its validity.

III. DATA ANALYSIS & PREPROCESSING

A. Data Overview

The data we worked with is from a period of 15 days, where 11 vehicles were used for different time periods. These 11 vehicles were from two models, 8 were the iveco stralis 460 and the other 3 were the iveco eurocargo 190. During the day, the vehicles can change driver more than once, resulting in 30 different drivers for these 11 vehicles, 2 of which drove both models, 22 drove only the first model and the other 6 drove only the second model.

As our data consisted of real-life trips and routes, throughout the analysis we need to make sure that our method grouped based on the driving behavior of the drivers and not on some other external/hidden factors. For this reason, we only analyzed the data from the first vehicle model, the iveco stralis 460, because it was the vehicle we had more data.

B. Feature Selection

In our problem, we had a lot of information about the automobile field, making easier the selection of the most important features. The features eligible for our analysis are the ones that the driver somehow has direct control over. This qualifies variables like the position of the pedals and the speed of the vehicle disqualifying features such as engine temperature.

The calculation of their correlation with the fuel consumption is used to confirm the quality of the variables by evaluating known correlations in the automobile field. High absolute correlation values mean that changes in that variable have a direct influence on fuel consumption. At first, we tried simply calculating the correlation using every pair (feature, fuel rate) available, without any data cleaning or data organization (i.e. calculating the correlation using every blue point in Figure 1). This did not show great results in some of the variables, like the engine speed, that showed a correlation of 0.55. However, when we plotted the engine speed over fuel rate (blue dots in Figure 1), we can clearly see a dense area that represents the important influence that the speed of the engine has on the fuel consumption. The low value in correlation is due to the calculation not taking into account many factors, for instance: what gear the vehicle is in and the road slope. To compensate that, we divided the X axis into $n$ equal intervals and calculated the average fuel rate in each interval. Then, we simply calculated the correlation of the center of the intervals and the average fuel rate values (red lines and marks in Figure 1). Figure 1: Scatter plot of Engine speed over Fuel rate in blue, with the mean values of engine speed spread over 50 intervals in red.

Table I shows the correlation values obtained using the first method (Simple Correlation) and the second method, using the mean of the intervals (Mean Correlation). It can be observed that when using the second method the obvious correlations stand out.
C. Feature Engineering

The set of features previously identified can be engineered in different ways. In this section, we describe the different feature engineering approaches investigated, the corresponding motivation and the results obtained.

One of the main ways of creating a new feature is by differentiating one. The new feature created captures the rate of change of the previous one. Using this method, we create the vehicle acceleration and the rate of change of the acceleration pedal position variables. These new variables capture the ability of the driver to maintain a constant velocity and how aggressively he/she presses/releases the pedal. This information is relevant to identify the aggressiveness of the driver. Aggressive drivers tend to waste fuel by pressing the throttle harder than required for the wanted acceleration.

The slope of the road is important when dealing with heavy vehicles. Therefore, we create the road slope variable using the GPS position of the vehicle. The calculation (as shown in Equation 2) is done by dividing the change in altitude between each two points by the distance the vehicle travelled in between those points. The mentioned distance is calculated using the python library GeoPy. This library calculates the geodesic distance in meters between two coordinates.

\[
slope_t = \frac{alt_{t-1} - alt_t}{\text{geodist}(\text{coords}_{t-1}, \text{coords}_t)}.
\]  

(2)

When evaluating/describing a driver, it is important to consider how he/she behaves on turns. For this reason, we created the turning angle variable. First, we create the direction vectors (arrows in Figure 2) by subtracting the coordinates at time \( t + 1 \) by the coordinates at time \( t \). The angle of the road is then computed as the angle between the current direction vector and the next one, as shown below,

\[
\text{Angle}_t = \arccos\left(\frac{u_t \cdot u_{t+1}}{|u_t||u_{t+1}|}\right),
\]

(3)

where \( u_t \) is the direction vector at time \( t \).

One important factor to isolate, when working with driving data, is the type of road the vehicle is at any moment. Is it a highway? Is it a residential area? If the driver is in a residential area, the vehicle cannot be travelling above a certain velocity. This affects not only the velocity, but also the amount of actions the driver must perform with the pedals. This means that we cannot compare two drivers if they were in different road types. Hence, we created the road type variable. To create this variable, we used the Open Street Map’s API and query it for the road type of each coordinate. Then, we grouped the results into three groups:

- 1 - Represents the highways and trunk highways.
- 2 - Represents primary and secondary roads.
- 3 - Represents more urban and residential roads.

The classifications can have errors when different road types that are close to each other, since the GPS error can make the API return the wrong road.

The results of this evaluation are shown in Table II. It is clear that, similarly to section above, the mean correlation is a more robust method to filter the information of the data. Three of the four variables created confirm the expected high correlation values, indicating that a good control over these variables is important to achieve a good fuel consumption.

D. Feature Preprocessing

The preprocessing step described here is responsible for dealing with the noise inherent to sensor readings, missing data and outliers. We approach these three issues differently for different features.

Outliers in our project are divided into two types: (1) situation outliers and (2) normal outliers. The situation outliers are the values that make sense in some situations but not in others. For instance, if the vehicle is travelling at 50km/h and a value of 90 shows up, it is considered an outlier. On the other hand, if the same value shows up and

Table I: Table of correlation values of the features we started with.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Simple Correlation</th>
<th>Mean Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceleration Pedal</td>
<td>0.853</td>
<td>0.980</td>
</tr>
<tr>
<td>Braking Pedal</td>
<td>-0.119</td>
<td>-0.160</td>
</tr>
<tr>
<td>Engine Speed</td>
<td>0.565</td>
<td>0.938</td>
</tr>
<tr>
<td>Vehicle Speed</td>
<td>0.452</td>
<td>0.832</td>
</tr>
</tbody>
</table>

Figure 2: The blue scatter plot represents the latitude and longitude values at each second, the arrows represent the direction vector computed using them.

Table II: Table of correlation values of the features created.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Simple Correlation</th>
<th>Mean Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>d_Vehicle Speed</td>
<td>0.510</td>
<td>0.955</td>
</tr>
<tr>
<td>d_Acceleration Pedal</td>
<td>0.178</td>
<td>0.927</td>
</tr>
<tr>
<td>Road Slope</td>
<td>0.576</td>
<td>0.989</td>
</tr>
<tr>
<td>Turning Angle</td>
<td>-0.078</td>
<td>-0.945</td>
</tr>
</tbody>
</table>
the vehicle is travelling at 88km/h the value is considered acceptable. The second type of outliers are the ones that constitute a wrong value independently of the scenario. For example, values too high or impossible for the road slope or vehicle speed.

To deal with the first type of outliers, we use the median filter [11, 7] of window size 3. To deal with the second type, we assume that the marginal values of a feature are more likely to be an outlier [15]. Consequently, we only use 98% of our dataset, using the data from the percentile 1 onward until the percentile 99. These two methods of detecting and replacing outliers cannot be used for the two pedal position variables, because abrupt changes and values of 0% and 100% are expected in them.

Dealing with the missing data in our work was a simple process and not a mandatory one as it does not change much the distribution of the features. Most of the missing data in our dataset occurred when the vehicle engine was already turned off and some sensors readings continued, meaning these portions of missing data have no information. When only one value is missing, we deal with it using interpolation. When two or more sequent values are missing, we do not assume any values for them, by doing this we might have less data but we increase the quality of it.

Finishing with the denoising portion of this step, most of our work here is using the moving average method. Using the moving average of 3 elements, every value of the TS is replaced by the mean between its value, the previous and the next ones. This method smoothes out abrupt changes in the data, for this reason, we do not use any denoising method for the pedals.

GPS data has error associated with it. This error is mitigated at high speeds. If the vehicle travelled 25 m in 1 second (90 km/h) between the two time points and the GPS missed the second position by 1 m, this merely represents a relative error of 4%. On the other hand, if the same error happens at 15 km/h, the vehicle has just travelled around 4 m in one second and the 1 m error now represents a relative error of 25%. Hence, for the road slope and turning angle variables, we only consider the values where the vehicle speed is above 15 km/h. To summarize, we use the moving average with 3 elements for every variable except the pedal ones, and we remove the road slope and turning angle variables for the vehicle speed values of 15 km/h and below.

IV. PIPELINE

The method we propose to profile the drivers, can take different variables as input. In Chapter V, we exemplify the use of our method in different analyses, each one having different conditions. In this chapter, we describe the pipeline of our method, the evaluation process and the parameter selection done using the evaluation results.

A. Cluster Creation

Before we implement any clustering method, we first need to define what would be the shape of our objects. In an ideal environment, these objects would be somewhat equal sized TS representing each driver. Since, in our work, the amount of data each driver had was very different, we divided the data of each driver into equal sized segments. To select the size of the segments, we need to consider that the bigger the segment, the more we can ensure it represents the profile of the driver and not simply how the environment is changing. For instance, in one minute segments the driver can be climbing the road in one and in another, turning on a flat road. Small segments can be used to identify changes in behavior during a trip, only if the environment between every segment is very similar.

In each segment, we only consider the parts of the TS where the cruise control is off and the vehicle is moving. The cruise control is a system that tries to maintain the vehicle at a predefined speed. We disregard this part of the data because no driver inputs are made here other than controlling the steering wheel. We also disregard the times when the vehicle stops, so we do not have segments appearing to have lower consumptions just because the vehicle was in motion for a shorter period in those segments than in others. These two criteria together removed around 40% of data.

Most of the works [3, 4] simply turn their time series objects into their maximum and minimum values, mean and standard deviation. While this captures some information about the TS, we believe that, if we capture the whole distribution of values of a segment, we can have more and easier to interpret information, similar to what was done by Fugiglando et al. [6]. To do this, for each feature, we average the range of each segment. The average is calculated after the outlier removal step talked in Section III-D, where we remove the data after the 99 percentile and before the 1 percentile. The new calculated range is divided into 10 equal bins. Several numbers of bins were tested and not much difference was detected between them. The only noticeable difference was between odd and even numbers. We chose an even number to be able to differentiate the center values in the differentiated variables, one value for the positive and one for the negative. Among the even values, we opted for 10 as the final size for the histograms to be easier to interpret. The distribution of values of each variable is represented as the percentage of time in each of the feature intervals. An example of this is shown in Figure 3, where the histogram for the acceleration pedal position of one of the segments is depicted.

Having the histograms for every feature, we were faced with two options. Should we cluster the feature histograms together, producing only one set of clusters per segment? Or, should we cluster them separately, resulting in a set of clusters per feature? We opted for the second option, because clustering the features separately allows for more detail in the profiles produced. Having a set of clusters per feature also provides the ability to evaluate if the feature is influencing the fuel consumption. We can analyze if the standard deviation of the fuel consumption is bigger before the creation of the groups or inside them.

Having the histograms for every segment for every fea-
tecture, we proceed to use the k-Means algorithm to find the cluster centroids that represent each feature. In Figure 4, we show the 4 clusters found for the engine speed variable using the k-means algorithm. In figure, we can see the cluster 4 (red lines) that represents drivers who drive most of the time with their engine rotations at 1300 RPM and higher (the peak in this cluster is around 1300 RPM is above 50% of time). Every other cluster represents drivers who gradually drive with lower engine revolutions, culminating in the cluster 2, that represents drivers who drive with the lowest engine speed. We represent the centroids with points of a line plot and not as a bar plot to better visualize every centroid of the features at the same time.

With the centroids of every feature obtained, we can classify new drivers according to the same principle. After preprocessing the driver, we segment its data, then we compute the histograms that represent the distribution of each of the six features we use. Finally, we match each feature to one of the pre-found clusters using a 1-Nearest Neighbour classifier. This classification is also evaluated.

We have several evaluation methods focusing on different aspects of cluster and classification quality. This section serves as support for a better understanding of the results of the evaluations in the analysis.

**B. Clustering Quality**

We wanted our clusters to be grouping the driver based on their different styles and not based on the environments they were in. To verify this quality, we used a score of completeness [13]. The completeness score is used in semi supervised learning methods to evaluate the clusters’ quality, when there is some way of labeling the objects. This method evaluates if the objects labeled as being the same were then placed in the same cluster (i.e. every object of a class must be in the same cluster). The method works by evaluating how each class is distributed among the clusters, evaluating one class at the time using the conditional entropy of the distribution proposed. Let us assume that $N$ is the number of driver segments, where each segment belongs to a driver $D = \{d_1, \ldots, d_n\}$ and we grouped them into a set of clusters, $K = \{k_1, \ldots, k_m\}$, resulting in the matching table $A$, where $a_{ij}$ represents the number of drivers $d_i$ that were placed into the cluster $k_j$. The completeness score will be given as,

$$ c = \begin{cases} 
1 & \text{if } m = 0 \\ 
1 - \frac{H(K|D)}{H(K)} & \text{else (4)}
\end{cases} $$

where,

\[
H(K|D) = - \sum_{d=1}^{D} \sum_{k=1}^{K} \frac{a_{dk}}{N} \log \frac{a_{dk}}{\sum_{k=1}^{K} a_{dk}}
\]

\[
H(K) = - \sum_{k=1}^{K} \frac{\sum_{d=1}^{D} a_{dk}}{n} \log \frac{\sum_{d=1}^{D} a_{dk}}{n}
\]

If our segments are big enough, we can assume that the different external factors are averaged out and consequently the fuel consumption is varying based on the driving style of the driver. This means that drivers with the same driving style and level of aggressiveness have similar consumptions. If this assumption holds, we can use fuel consumption has a confirmation of cluster quality. To help the assumption, we can restrict the environment of the analysis, for instance, using only data from one type of road at a time.

There are two possibilities of using fuel consumption as a clustering quality metric: (1) determine if the segments grouped together have similar fuel consumption and (2) determine if the fuel consumption differs between the groups.

The first mentioned metric represents the within cluster fuel consumption variation or simply WCV. The fuel variance of a cluster is given by the sum of the distances between the fuel consumption of each object and the average fuel consumption of the cluster. The WCV returns the sum of those distances normalized,

\[
WCV = \frac{1}{k} \sum_{i=1}^{k} \sqrt{\frac{1}{|K_i|} \sum_{x_j \in K_i} (x_j - \mu_i)^2}, \tag{6}
\]
where $K_i$ refers to the set of elements in cluster $i$, $k$ is the number of clusters, $\mu_i$ is the mean of the fuel consumption of every segment in the cluster $i$ and $x_j$ is the average fuel consumption of the segment $j$ that belongs to the cluster $i$ being computed. This method is used to analyze if a variable is influencing the fuel rate. If we group more the drivers who use the variable similarly (increase the number of clusters), we expect the WCV to get lower. Using this method, lower scores represent better cluster quality.

In the second metric, we check if there is a significant distance between fuel average of every cluster. This distance means that being classified into one of the clusters really affects fuel performance. To depict the distance in the consumption between clusters, we use the Between Cluster Fuel Consumption Variation or simply BCV. This method is the opposite of the previous one, giving the fuel variance between the clusters. It is given by the sum of distances between the fuel consumption of each cluster $c_i$ and the fuel consumption of the dataset, normalized. If we use the same notation used in 6, just adding $F$ as the mean of fuel consumption the dataset, we get,

$$BCV = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (\mu_i - F)^2}. \quad (7)$$

As this method represents the distance between the clusters, higher scores represent better cluster quality. However, this is not completely linear, as different styles can have similar consumption, leading, wrongly, to a low score using this method.

The three methods described above are used to evaluate different aspects of the clustering process. To evaluate the classification quality, we check if our classification method places the drivers it has never seen before in the same clusters as the clustering method did. We assume that the clusters where the segments are placed are correct and use that information as the label. This way, we can train our model on a portion of the dataset (training set) and use supervised learning metrics to evaluate it with the remaining data (testing set). This way, we can test how robust the classification is when classifying new drivers that it has never seen before.

We do not simply divide the dataset at random, as it is common in validation metrics, because if the training and testing sets contain segments of the same driver, the output of the evaluation could be invalid. Therefore, we use a leave-one-out cross validation method. First, we train our model with every segment, registering the prototypes created by every cluster (the centroid produced by the k-means) and the clusters each segment was placed in. At each step of the process, we remove every segment of a driver from our dataset, training our model with the remaining data. Then, we use a $1-NN$ classifier to place each of the segments of driver under test (the one removed) in one of the training clusters. This validation process comes with one issue that needs to be tackled. The clusters produced by the training model are not exactly the same as the ones produced by the original model (the model trained with every segment). In every iteration of the process, the original model clusters remain the same, however the training model ones change based on the driver removed. How do we match the clusters from the original model to the clusters in the training model? We compute the distance matrix between the centroids of the original and training models, matching the combination of centroids that minimizes the sum of distances. Table III is an example of the distance matrix calculated when matching the centroids represent one of the folds of the method.

The distance between the centroids in the original and training models increases when evaluating drivers with more segments. These drivers had more influence in the original clusters, deviating the training model more from the original one. To evaluate this inaccuracy, we compute the $mae$ of the distances in every cluster match. Looking into the example in Table III the error would be calculated as $mae = \frac{(0.167 + 0.0333 + 0.01164)}{3}$, resulting in an error of 0.07065%. The error component of the model is given by the mean of the $mae$ of every validation fold. This error component not only evaluates the assumption that the clusters of the training model match the clusters of the original model, it also evaluates how robust the method is to the addition of new drivers. Lower values indicate the method is more robust. After going through every driver, we get a labeled set and a prediction set for the model. This allows for the computation of the accuracy of the model by simply computing the number of correct classifications over the total number of segments.

### C. Selection of the number of clusters

We compute many score metrics for each feature, allowing a flexible evaluation of the clusters. We could use only one metric or a combination of metrics. Using more than one metric would require the selection of weights and scaling of the metrics, which would be more parameters in the method. To avoid that, we use only one metric, having merely to select the most relevant one for the problem under study. The completeness metric evaluates how much the clusters represent the same behavior in each trip. It is not useful if the clusters are not good. The accuracy metric evaluates the classification, it does not evaluate the clusters directly. Since our problem is the creation of the clusters, the selection was narrowed down to the two variances. Between them, we opted for the WCV, because, as we said in the previous section, two distinct
behaviors can have similar fuel consumption, resulting in a
bad BCV score, even if they were well grouped.

Since the WCV tends to decrease as the number of
clusters increases, we looked for the elbow in the WCV over
number of clusters plot, defining it as the optimum value for
the number of clusters of that feature. To find elbow, we add,
to the said plot, a diagonal line from the first point to the
last (orange line in Figure 5). Then, for each X (number of
clusters), we find a perpendicular line to the one we created
that passes through its WCV value (green dashed lines in
Figure 5). The elbow is the number of clusters that gives the
longest of these last lines, which is 3 for the acceleration
pedal feature in Figure 5.

This is the method we use throughout our work, however,
the same process of finding the elbow can be used on another
metric. This choice depends on what one values the most.

V. ANALYSIS

The method we propose can take as input different por-
tions of the data. In this chapter we exemplify our method if
the following analyses: (1) one analysis using every portion
of the trips and (2) three analyses, each using only the data
of one road type.

We analyze the clusters produced by our methods and
the most common profiles. The profile of the driver is
represented by the clusters he/she was placed in each of
the features. We can also analyze subprofiles, which are
combinations of n features and they can make easier to
identify patterns of the drivers. For example, a combination
of acceleration and acceleration pedal can show the ability
of the driver in using external factors like the slope of the
road to change the vehicle speed.

A. Overall Analysis

The first analysis is an overall one. We neither divide
the data by road type nor road slope, meaning we need
large segments to accommodate that. Larger segments can
help reduce the external differences between them (i.e. in
an 1 hour segment, the driver is more likely to have been
in different types of road than in a 20 minutes segment).
However, this can still have problems if we are comparing
drivers that always drive in different environments. For
example, one driver only drives in residential areas and
another in highways.

In this analysis, we used one hour segments, resulting
in 92 trips of 16 different drivers, which represents 8 less
drivers than the 24 we started with. These 8 drivers drove
less than one hour without cruise control, excluding them
from further analysis.

The resulting clusters produced the results shown in Table
IV. The acceleration and engine speed clusters have the
highest completeness score, meaning the drivers maintain
similar behaviors in these features in every trip. In terms
of WCV, the acceleration pedal and vehicle speed have the
lowest value, meaning fuel consumption is being influenced
by them. The classification algorithm had an accuracy of 1
when predicting the brake pedal and acceleration clusters.
This indicates both, classification quality and a lack of
segments in the analysis, because even the best classificator
should be converging to one, not exactly one.

When we analyzed the centroids that represent the clus-
ters, we note that the centroids of the vehicle speed appear
to be clustering different environments and not different
driver styles (Figure 6). A closer analysis revealed that every
segment in the cluster 0 (Blue in Figure 6) belongs to
segments of the same driver. The driver in question, drove
most of the time in urban roads. This is caused by the
some drivers having more segments than others and some
always driving in different environments than others. The
next analysis we made, isolates the type of road the driver
was in, hoping to solve this heterogeneity.

B. Road Type Analysis

When comparing trips, it is important to isolate the type
of road in which the driver was in. Highways allow for much
higher speeds and little use of the pedals than urban roads.
For this reason, this analysis can be viewed as three smaller
analyses, one for each of the three road types found before.
In each analysis, we removed the data from the other two
road types.

Considering that we are using a portion of the data at a
time, we need to reduce the size of the segments so we can
have a considerable number of segments. Our criteria were:
(1) ensure the size of the segments remained the same among
the three analyses, varying the number of segments between
them and (2) minimum number of segments in an analysis

Figure 5: Representation of the method to find the elbow of
the WCV.

Figure 6: Vehicle speed centroids of the Overall analysis
Table IV: Evaluation of the overall analysis.

<table>
<thead>
<tr>
<th>n_clusters</th>
<th>Feature</th>
<th>Completeness</th>
<th>Accuracy</th>
<th>Error</th>
<th>WCV</th>
<th>BCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Acceleration</td>
<td>0.6928</td>
<td>1.0</td>
<td>0.0195</td>
<td>3.5112</td>
<td>2.2826</td>
</tr>
<tr>
<td>3</td>
<td>AccelPedal</td>
<td>0.5201</td>
<td>0.913</td>
<td>0.0643</td>
<td>3.6309</td>
<td>3.9347</td>
</tr>
<tr>
<td>2</td>
<td>BrakePedal</td>
<td>0.4313</td>
<td>1.0</td>
<td>0.0032</td>
<td>3.6292</td>
<td>2.3834</td>
</tr>
<tr>
<td>3</td>
<td>d_AccelPedal</td>
<td>0.6329</td>
<td>0.8326</td>
<td>0.0392</td>
<td>3.2792</td>
<td>2.3858</td>
</tr>
<tr>
<td>4</td>
<td>EngSpeed</td>
<td>0.7084</td>
<td>0.8759</td>
<td>0.0708</td>
<td>3.1273</td>
<td>3.1269</td>
</tr>
<tr>
<td>5</td>
<td>VehSpeed</td>
<td>0.5168</td>
<td>0.9674</td>
<td>0.0308</td>
<td>2.9984</td>
<td>2.8535</td>
</tr>
</tbody>
</table>

Figure 7: Vehicle speed centroids of the highway analysis
to be 50. The size found was 20 minutes, representing 1/3 of the size of the segments in the overall analysis, making comparisons between the two less accurate.

The highway analysis had the most segments (150), making it the most relevant to analyse here. Additional information can be found in the work of Alves [1]. The cluster produced by the highway analysis are shown in Table V.

The clusters produced with this analysis have completeness lower than in the previous one as expected, since the size of the segments is much lower. The accuracy, on the other hand, kept the high scores (above 0.8). While comparisons with the overall analysis should be made with caution, we note that some patterns are repeated. For instance: the engine speed still has the highest completeness score and the vehicle speed still has the lowest WCV.

This analysis appears to have solved the problem found with the vehicle speed centroids. The centroids seem to show the speed profile of the driver and not the environment (in Figure 7 the centroids have their peaks close to each other).

The most common profiles and subprofiles in the analysis are important to analyse, because they relate to the most common behaviors drivers have. New drivers will land in these profiles many times. In Figure VI the most common subprofiles using the acceleration pedal, engine speed and vehicle speed are depicted. The first profile refers to a more defensive type of driver, with less use of the acceleration pedal (the last value in the blue centroid in Figure 8 represents the pedal in the 100% position), lower engine rotations (the blue centroid in Figure 9 is more distributed than the other two) and lower vehicle speeds (the peak in the blue centroid is farther back than in the others). The other profile is the extreme opposite, referring to more aggressive drivers.

VI. Conclusions

When conducting this work, it became obvious the importance of the environment when comparing two drivers. A driver who follows the recommended behaviors can have higher fuel consumption than another who does not, if the first one is driving with a heavier vehicle or faced more traffic than the other. For this reason, large segments are important, because the larger the segment the more these external factors are averaged out. Furthermore, an analysis
of the centroids with a labeling step is very important to evaluate the drivers without having to rely on the fuel consumption alone.

This method allows a company to have more knowledge of its drivers, making it useful in many situations. Having the current profile of the driver, a company can have a more direct approach in terms of an advising system, both while and after driving. Having the history of profiles the driver had over time, allows the company to check his/her improvement rate. The profile of the driver after a trip can be turned into a quality score which can be used as a monetary reward to encourage improvement. A company can also use the combination of the profiles of its drivers to make a smart allocation of drivers to different routes.

A. Future Work

The next steps for this work should be the addition of new variables both in order to mitigate some environment differences inside the analysis and to increase the information given by the analysis. To improve the homogeneity of the data inside the analysis, some variables could be added, such has: (a) the traffic information, allowing the analysis to separate different traffic scenarios and (b) the meteorological conditions, distinguishing between sunny and rainy days. The road type information can also see improvements, for example, using the speed limit and the number of lanes of the road.

To augment the information given by the analysis, some variables could be created, like: the distance to the vehicle ahead, adding the information of the ability of the driver to maintain a big enough distance to be able to brake smoothly, the steering wheel angle and its derivative, adding the information of the ability of the driver variables could be created, like: the distance to the vehicle ahead, adding the information of the ability of the driver

<table>
<thead>
<tr>
<th>n_clusters</th>
<th>Feature</th>
<th>Completeness</th>
<th>Accuracy</th>
<th>Error</th>
<th>WCV</th>
<th>BCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Acceleration</td>
<td>0.475</td>
<td>1.0</td>
<td>0.023</td>
<td>4.196</td>
<td>1.558</td>
</tr>
<tr>
<td>3</td>
<td>AccelPedal</td>
<td>0.249</td>
<td>0.972</td>
<td>0.045</td>
<td>3.659</td>
<td>2.912</td>
</tr>
<tr>
<td>2</td>
<td>BrakePedal</td>
<td>0.276</td>
<td>0.98</td>
<td>0.002</td>
<td>4.334</td>
<td>0.714</td>
</tr>
<tr>
<td>3</td>
<td>d_AccelPedal</td>
<td>0.370</td>
<td>0.801</td>
<td>0.058</td>
<td>3.31</td>
<td>2.79</td>
</tr>
<tr>
<td>3</td>
<td>EngSpeed</td>
<td>0.609</td>
<td>0.911</td>
<td>0.087</td>
<td>4.063</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>VehSpeed</td>
<td>0.31</td>
<td>0.839</td>
<td>0.149</td>
<td>4.009</td>
<td>1.755</td>
</tr>
</tbody>
</table>

Table V: Evaluation of the highway analysis.

<table>
<thead>
<tr>
<th>AccelPedal</th>
<th>EngSpeed</th>
<th>VehSpeed</th>
<th># App</th>
<th>Fuel_m</th>
<th>Fuel_s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>20.15</td>
<td>3.24</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>17</td>
<td>27.61</td>
<td>3.31</td>
</tr>
</tbody>
</table>

Table VI: The most common profiles of the road type 1 analysis.

REFERENCES