Identifying Heavy Good Vehicle Driving Styles in the United Kingdom

Article *in* IEEE Transactions on Intelligent Transportation Systems · October 2018

DOI: 10.1109/TITS.2018.2875343

CITATIONS
2

READS
216

9 authors, including:

Grazziela P. Figueredo
University of Nottingham
71 PUBLICATIONS **171** CITATIONS
SEE PROFILE

Utkarsh Agrawal
University of Nottingham
13 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Jimiama Mafeni Mase
University of Nottingham
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Daniele Soria
University of Kent
68 PUBLICATIONS **806** CITATIONS
SEE PROFILE

Identifying Heavy Goods Vehicle Driving Styles in the United Kingdom

Grazziela P. Figueredo[1,2], Utkarsh Agrawal[1], Jimiama M. Mase[1], Mohammad Mesgarpour[3], Christian Wagner[1], Daniele Soria[4], Jonathan M. Garibaldi[1,2], Peer-Olaf Siebers[1], Robert I. John[1]
1. School of Computer Science, The University of Nottingham, NG8 1BB, UK
2. The Advanced Data Analysis Centre, The University of Nottingham, NG8 1BB, UK
3. Microlise, Farrington Way, Eastwood, Nottingham NG16 3AG, UK
4. Department of Computer Science, University of Westminster, W1W 6UW
grazziela.figueredo@nottingham.ac.uk

***Abstract*—Although driving behaviour has been largely studied amongst private motor vehicles drivers, the literature addressing heavy goods vehicle (HGV) drivers is scarce. Identifying the existing groups of driving stereotypes and their proportions enables researchers, companies and policy makers to establish group-specific strategies to improve safety and economy. In addition, insights into driving styles can assist predicting drivers' reactions and therefore enable the modelling of interactions between vehicles and the possible obstacles encountered on a journey. Consequently, there are also contributions to the research and development of autonomous vehicles and smart roads. In this study our interest lies in investigating driving behaviour within the HGV community in the United Kingdom (UK). We conduct the analysis of a telematics dataset containing incident information on 21,193 HGV drivers across the UK. We are interested in answering two research questions: (i) What groups of behaviour are we able to uncover? (ii) How do these groups complement current findings in the literature? To answer these questions we apply a two-stage data analysis methodology involving consensus clustering and ensemble classification to the dataset. Through the analysis, eight patterns of behaviour are uncovered. It is also observed that although our findings have similarities to those from previous work on driving behaviour, further knowledge is obtained, such as extra patterns and driving traits arising from vehicle and road characteristics.**

***Keywords*—Driver Profiling, Driving Pattern, Driving Habit, Driver Behaviour, Clustering Analysis, Ensemble Clustering, Ensemble Classification, Big Data Analysis**

I. Introduction

A large proportion of private and public sectors in the United Kingdom (UK) rely on heavy goods vehicles (HGVs) transport for procurement and delivery of goods and services. Statistics for the UK between October 2014 and September 2015 show that HGVs delivered 1.63 billion tonnes of freight within the UK, and 8.5 million tonnes were imported and exported [1]. Due to the importance of HGVs in the country's economy, there are great efforts being employed to reduce their incident numbers [2], [3]. Such incidents, which incur in significant losses, are caused by the characteristics of the vehicle, road, weather conditions, defiance of traffic rules, company policies and mostly by driving behaviour. In order to mitigate these mishaps and their consequences, companies and researchers have invested time and resources to improve the technology to manage HGV fleets. Fleet management includes a range of activities, such as vehicle maintenance, telematics, driver monitoring, health and safety inspections, etc. Most of these practices target the drivers' welfare, as they are the main actors responsible for the vehicles within the fleet and the cargo being transported. Currently, telematics is one of the leading technologies in fleet management, where data acquisition is performed by tracking and diagnostics devices. Telematics enable capturing large amounts of driving and vehicle data; this information assists companies to identify, understand and define strategies for incident prevention. Actions such as driving performance scoring, risk assessment of drivers, education and alerts are now widely adopted. To define and to follow such procedures, an in depth understanding of driving behaviour and their responses to the environment is necessary.

In this work we investigate driving behaviour in the UK by analysing a large telematics dataset. Our review of the literature indicates that most studies in the area are limited by the number of drivers, which is in most cases fewer than 100. Furthermore, the driving data employed previously is collected for a small number of journeys per driver; and to the best of our knowledge, there is very little literature on HGV driver profiling for the UK. We address these gaps by analysing a much larger dataset, in which hundreds of thousands of HGV driving incidents in most roads in the UK for the year of 2015 are considered. In total, 21,193 HGV drivers are investigated. We want to answer the following research questions: (i) What are the existing patterns of behaviour within the UK HGV driving community? (ii) How do these patterns complement the current knowledge in the literature? To answer these questions we apply a two-phase approach to the dataset, in which eight patterns of behaviour are identified and further validated by experts in industry. In the following sections we provide the literature review (Section II), introduce the methodology employed to analyse the data (Section III), followed by the results and discussion (sections IV and V), conclusions and opportunities for further exploratory studies with the dataset (Section VI).

II. Related Work

Researches in driving behaviour, vehicle monitoring, driving style prediction and driver modelling are on the rise, as there is a global demand for intelligent solutions to improve driving economy and road safety. The challenge lies in determining solutions capable of addressing social, financial and geographical requirements within a problem context. To achieve

this, the investigation of actions and policies for education, law enforcement and improvements in infrastructure are of essence. Equally important is the understanding and prediction of the drivers responses to these interventions. In this review we focus on the work carried out to understand, model and predict driver behaviour. Current literature targeting solely HGV drivers is scarce. In addition, there is not much literature targeting UK HGV drivers. We therefore include in our review studies concerning the analysis of specific driving behaviour for both commercial and non-commercial drivers, obtained through the analysis of telematics, GPS and other mobile devices data. Studies on general driving style for both small vehicles and HGV drivers obtained through self-assessment questionnaires [4], [5] is far more common. However, due to the differences in the nature of the data collected, we do not include this research in our review. Our objective is therefore to identify the research gaps in telematics data findings – which justify the need for our contribution – and to establish a baseline for comparison with our methodology and results.

A relevant work in the area of driving behaviour was conducted by Constantinescu *et al.* [6], in which six driving profiles were identified from telematics data. Data was obtained in a five-day experiment, involving 23 drivers and two control drivers with up to nine journeys each. Controls were employed to introduce extreme behaviours (very aggressive and slow, non-aggressive, economical). 200 journeys were analysed, and the features considered were the percentage of time above the speed limit (60kph), the mean and standard deviation (std.) of speed, the mean and std. of acceleration, the mean and std. of braking and the total energy required to increase the speed. Hierarchical clustering (HC) identified 6 clusters and Principal Component Analysis (PCA) assisted in their interpretation and in the detection of four factors determining behaviour: aggressiveness, speed, acceleration and braking. For each factor, the authors identified the range values and how drivers within the clusters behaved. Ranges were converted into labels characterising drivers' stereotypes, (Table I). Although the findings were an important contribution to understand driving behaviour, this work is limited to the number of drivers and journeys investigated. More data needs to be collected to further validate the clusters detected. In addition, control subjects forced the analysis to establish pre-defined clusters with extreme behaviour. These artificial clusters also need to be further contrasted with more data instances to assess their likelihood of occurrence in real-world scenarios.

TABLE I: Driving behaviour clusters from Constantinescu *et al.* [6]

Group	Aggressiveness	Speed	Acceleration	Braking
1	Moderately low	Low-Moderate	Moderate	Smooth-Moderate
2	Very low	Low-Moderate	Low-Moderate	Smooth-Moderate
3	Moderately high	Moderate	Moderate	Sudden
4	Neutral	Moderate	High	Moderate
5	Neutral	Moderate-High	Low-Moderate	Moderate-Sudden
6	High	High	High	Sudden

Kalsoom and Halim [7] applied K-means (KM) and HC to a dataset obtained via simulation of 30 drivers. Data collected included the number of left and right turns, left and right indicators, brakes, horns and gear change with speed. Three driving styles were established, slow, normal and fast drivers. This work is also limited to amount of drivers studied. Less clusters of driver behaviour were detected, which might suggest that either the simulation experiments or the cluster analysis did not capture all patterns of behaviour found in Constantinescu *et al.* [6].

Castignani *et al.* [8] collected data from motion sensors and GPS, which is later analysed to determine driving events. Fuzzy logic was employed to detect harsh braking, harsh acceleration, over speeding and aggressive steering. The events obtained were subsequently merged with time and weather information to determine their risk level and to score the driver. Experiments employed the same vehicle over a predefined path in Luxembourg. Journeys occurred at daytime, with different weather and predefined behaviour. For the first lap the driver was supposed to drive calmly, observing the speed limit and avoiding abrupt manoeuvres; in the second lap the drivers behaved aggressively. Results showed consistency in the number of incidents obtained during calm and aggressive driving. One of the limitations of the study is that it tested the efficiency of the fuzzy system over predefined driving behaviour. In real-world scenarios, however, variation of driving styles and unplanned actions are likely to occur.

Halim *et al.* [9] employed KM, fuzzy c-means and Model-Based Clustering (MBC) to determine four driving profiles from 50 drivers. Data was acquired with a hardware consisting of vehicle functionalities and a simulator software with a virtual driving environment. Each subject drove the car in three scenarios (high, average and low traffic) for 15 minutes per scenario. The variables considered were number of left indicator, left turns, right indicator, right turns, brake use, horn use, reverse gear use, average gear, maximum gear, average speed, maximum speed and gender. KM results were compared with those from fuzzy c-means and MBC to establish the optimal number of clusters. PCA determined the best features (average and maximum speed, brakes and horns were selected). Profiles 1 and 4 comprised of cautious drivers with high average speed but appropriate number of brakes. Profile 2 characterised slow and sluggish drivers with high frequency of braking and indicators. The third profile included safe drivers with moderate speed, number of use of indicators and brakes.

Ellison *et al.* [10] proposed a risk index framework for scoring drivers based on their behaviour, personality, perceptions, demographics, vehicle characteristics, weather conditions and time of the day to assess the risk of crashes. 8 million GPS data observations with 1Hz frequency from 106 drivers in Sydney over several weeks were considered. Results indicated that over 90% of drivers exhibit more variability in speeding, acceleration and braking behaviour between different road environments when compared to the same road. The author's results suggest the potential for using more disaggregate data but also the necessity to control for temporal and spatial factors when studying driver behaviour. These conclusions reiterate the importance of our study, where a larger data sample should provide us with better insights into HGV driving behaviour. Similarly, Saiprasert *et al.* [11] calculated a drivers safety index based on their driving events and categorize the driver into four safety level profiles (very safe, safe, aggressive, very aggressive). The calculation is based on an equation in which the events (harsh acceleration, harsh braking, harsh turning, sudden lane changing and over speed events) are weighted according to the characteristics of the road. Data collection was

carried out with 20 drivers in two of Thailands main highways, with 30 journeys per route. Results showed that the majority of the drivers are either safe or aggressive for both routes.

From the literature it is possible to identify three clear gaps in the current research: (i) the number of drivers investigated is very limited across the studies; (ii) the experiments are mostly conducted within a small number of routes and journeys; and (iii) to the best of our knowledge, there is very little literature regarding driver profiling based on telematics data for the UK. Our work therefore aims at contributing in filling these research gaps. The next sections introduce further details regarding the dataset studied and the methodology employed.

III. Materials and Methods

A. Data Collection and Pre-Processing

The dataset employed in this research is obtained in collaboration with a telematics provider company, namely, Microlise [12]. All data generated by the telematics systems are transmitted and collected from vehicles in real-time across the year [13]. The lorries are equipped with standard sensors as specified in the Society of Automotive Engineers (SAE) standard J1939-71. Data is gathered by a collection of sensors connected to multiple electronic control units by controller area network (CAN)-bus. The data is subsequently transferred to the telematics unit that populates Microlise's databases via 3G. Microlise's telematics solutions are implemented in around 25% of the HGV vehicles circulating in the UK.

For our analysis 21,193 anonymised drivers are considered. Further details on how the data has been merged is found in [13]. The requirement for a HGV driver to be included in the dataset is that they must have taken part in at least 10 journeys per quarter of the year. Each driver therefore has a minimum set of 40 journeys travelled on any road of the UK. The period considered is between the first of January until the thirty first of December of 2015. Driving style data includes measures such as engine revs (to identify green band driving and over-revs), fuel consumption, accelerator position, idling, use of cruise control, use of power take off, use of primary and secondary (engine) braking, harsh acceleration and harsh braking. Many of these features collected, however, are not present in all vehicles and are therefore not available to all drivers. To provide a consistent, standardised analysis, we select the following attributes, which are present in all vehicles: **(i) Driving time in seconds; (ii) Average daily distance in metres; (iii) Number of harsh braking events; (iv) Over speed duration in seconds; (v) Excessive throttle duration in seconds; and (vi) Number of over rev events**. In addition, the number of journeys per quarter is necessary to determine those drivers to be included.

This dataset has been previously studied in the Driver of the Year competition, promoted by Microlise. The driver ranking based on incidents for the competition was performed by Figueredo *et al.* [13]. The idea of determining drivers' profiles as a continuation of the previous data exploratory work came from the necessity of the company to better understand how HGV drivers in the UK behave in order to develop better policies, and also due to the scarcity of related literature. From the analysis for the competition, three clusters based on the total distance travelled (short, medium and long distance) were identified. We extended the authors' work by sub-clustering drivers based on their average daily distance travelled, as further explained next. The second step of the pipeline employed addresses the uncategorised data after the clustering exercise.

B. Sub-group Clustering

Tseng *et al.* [14] hypothesise that "the driving mileage is a determinant factor in driver speeding incidents". The goal of this work is to determine whether (and how) distance and time driven possibly affect the occurrence of incidents in the UK. The average distance driven per day is therefore adapted instead of total distance, as it seems better suited to the driving behaviour. KM is run on the data to identify the distance ranges. Once the distance ranges are identified, the next phase is to identify driving profiles within them. A two-stage framework is employed, as explained in the next subsection.

As mentioned previously, our analysis is conducted in two stages[1]. In the first stage of our pipeline, we adapt part of a consensus clustering framework proposed by Soria *et. al.* [15] to elucidate core groups in our dataset (Figure 1). Their framework comprises the following steps, as tailored to our problem:

STAGE 1

Data characterisation and pre-processing. If necessary the data is cleaned, normalised and inconsistencies are fixed. For our dataset we perform consensus clustering employing the attributes harsh braking events, over speed duration, number of excessive throttle events and number of over rev events. These features are normalised by the driver's total driving time in seconds. Figure 2 shows the exploratory analysis of the data. The data are skewed as drivers with high numbers of incidents are rarer (as expected). We have, however, not removed what may appear to be outliers in the box plots distributions shown below, as extreme behaviour and the sparsity of the data instances are important for driving characterisation. Further discussion on the need for keeping what appear to be outliers in the analysis is given in the next sections, as we conduct the analysis. No further transformation is adopted to the dataset as we want to observe how the grouping of the real-world behaviour occurs, i.e., to establish a clear separation between bad drivers and good drivers. We have, however, attempted to cluster the data after a Min-Max normalisation, but no relevant cluster is obtained from this exercise.

Clustering. We employ the following clustering techniques: the Hierarchical Clustering Approach [16] (HCA) for determining the candidate cluster centres for KM; and KM [17], [18] and Partitioning Around Medoids (PAM) [19], [20] for clustering. These techniques differ from each other as they have different measurements to define proximity of the data instances to establish the clusters. This means that they group the data considering different characteristics. They are chosen as they are among the most widely used.

[1] An in depth explanation of the framework is given in the supplementary material.

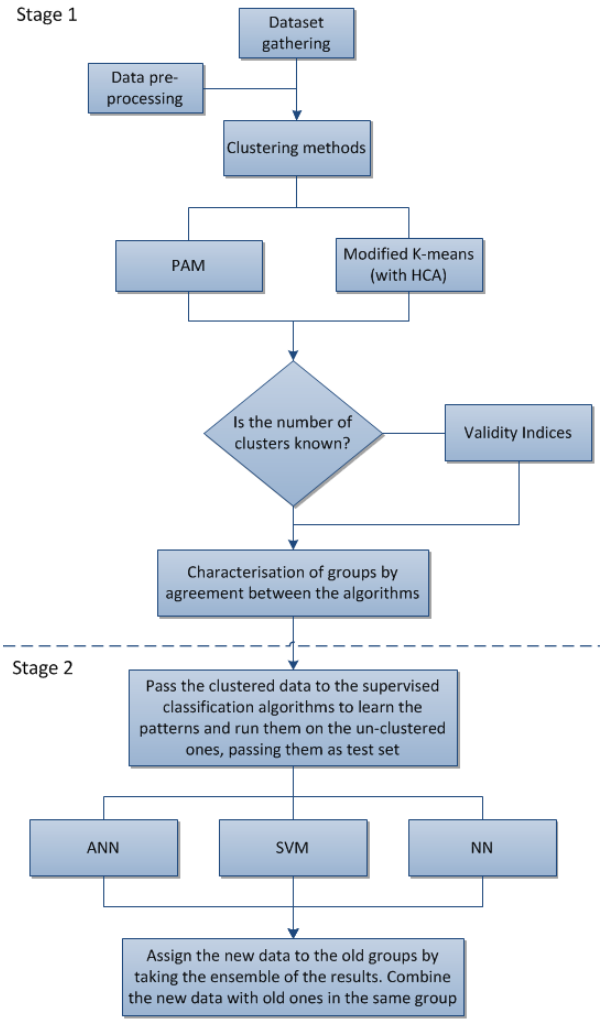


Fig. 1: Flowchart of the pipeline to determine the clusters in the drivers data set

Determining the number of clusters. In this step, validity indices are applied to the clustering results. As for our problem the number of clusters is unknown, in an attempt to determine an ideal value, the indices Calinski and Harabasz [21], Hartigan [22], Scott and Simons [23], Marriot [24], Trace W [25] and Trace $W^{-1}B$ [25] are applied to KM and PAM, for which the number of clusters and cluster centres are explicit parameters, over a range of number of clusters varying from 2 to 20. The cluster centres are defined by applying HCA to the data. The nodes in the hierarchy graph that split the data into the n desired clusters are used as parameters for KM. According to specific rules, validity indices indicate the appropriate number of groups to consider in the analysis. In the indices implementation, cluster stability is also assessed.

Consensus. It is achieved by aligning the clusters found by the different techniques. The agreement among solutions of the different clustering methods employed is evaluated using the Cohen's kappa coefficient k [26]. This coefficient is a statistical measure of agreement for categorical items.

Data Visualisation. Graphs such as box-plots and bi-plots

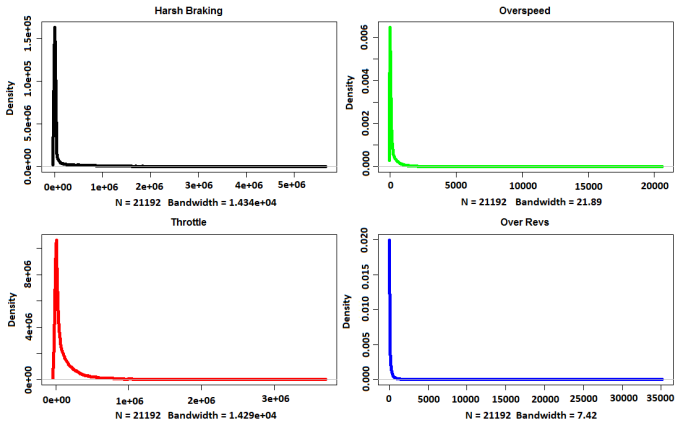


Fig. 2: Data set features density distribution

are employed for a general characterisation of the clusters obtained.

After employing the consensus clustering, those instances in which there is no consensus as to which cluster they should belong to are labelled as *unclassified*. To reduce the number of unclassified instances, we add a second stage, as defined in Agrawal *et. al* [27] to the framework:

STAGE 2

Ensemble Classification: Ensemble Classification aims at combining multiple classification algorithms to improve the predictive power of a system. The idea of the second stage is to use the clustered data from Stage 1 to train the ensemble of classifiers. The clustered data and the unclustered data in Stage 1 are therefore employed as training and test sets, respectively. We attempt to classify the previously unclustered data (test set) with the goal to assign them to one of the previously identified groups. Support Vector Machines (SVM) [28], Nearest Neighbour (NN) [29] and Multi-Layer Perceptron Neural Networks (MLP) [30] are employed as part of the ensemble to learn the patterns. Among a number of decision level fusion methods present in the literature, majority voting is chosen to be included in the framework, given their level of confidence [31].

To evaluate and validate the characteristics of the data assigned to clusters from Stage 1 after ensemble classification, visual tests, statistical tests and cluster quality assessment need to be performed. The distribution of each attribute in a cluster after Stage 1 is compared to that from Ensemble Classification from Stage 2 using boxplots, the Mann-Whitney-Wilcoxon non-parametric t-test at 0.05 level of significance. If a group of newly classified instances is statistically similar to any of the previous clusters after the tests, the drivers in the newly classified group are combined with the equivalent group of drivers previously clustered after Stage 1. Otherwise, the group of misclassified instances (instances classified as belonging to a certain cluster by the ensemble, but the statistical tests fail to confirm the classification) is checked for the possibility of being an extra profile. Lastly, the majority voting ensemble classification results are evaluated using the Davies-Bound (DB) internal cluster quality assessment index. The DB index

computes the ratio of intra- and inter-cluster distance of the drivers. Therefore, to study the quality of clusters, the DB index values of each subgroup of drivers assigned after ensemble classification are compared to the values of the DB index obtained if they had been assigned to each of the other clusters. These tests are employed as there is no ground truth for the problem.

To clarify this part of Stage 2, let us assume that after Stage 1 being applied to a 6600 instances data set, 3 clusters are found and 600 instances are unclassified. 6000 data points from clusters 1, 2 and 3 are therefore used to train the ensemble. The 600 unclassified instances are subsequently classified by models within the ensemble using majority vote. Let us assume that the result of the 600 instances classification is as follows: 100 instances are classified as belonging to cluster 1 (from Stage 1); 200 instances are classified as part of cluster 2; and 300 instances are classified as belonging to cluster 3. To ensure that this classification is correct, we subsequently resort to the box-plots, Mann-Whitney-Wilcoxson test and the DB cluster quality assessment. We therefore plot the distribution of the original cluster 1 (from Stage 1) and contrast with the distribution of the 100 instances newly classified; we also compare them using the statistical tests. If those tests show that there is no visual or statistically significant difference between cluster 1 and newly classified 100 instances, then these 100 instances are merged into cluster 1. Otherwise, they remain unclassified or, if the group is large enough, their instances are tested to assess if they represent a new cluster (or pattern of driving behaviour). The same process is repeated for the other clusters. After ensemble classification, more data are expected to be assigned to one of the groups, and the remaining data are still unlabelled or 'unclassified'. The clusters obtained from Stage 1 are combined with the newly classified data in the same groups to achieve the final clusters.

C. Experimental Settings for the Ensemble

Two instances of SVM classification algorithm are used in the ensemble, with radial basis function and sigma set at $0.5/n_f$ and $1/n_f$ respectively, where n_f is the number of features in the dataset. In addition, two MLP classification models are considered with 1 and 3 neighbours with Euclidean distance function. MLP is adopted with three hidden layers and a sigmoid activation function to model the weighted sum. Three instances of MLP are part of the ensemble: (i) 15 nodes for layer 1, 20 nodes for layer 2 and 15 nodes for layer 3 with 100 epochs; (ii) 10 nodes for layer 1, 15 nodes for layer 2 and 10 nodes for layer 3 with 100 epochs; and (iii) 10 nodes for layer 1, 20 nodes for layer 2 and 10 nodes for layer 3 with 100 epochs. All algorithms are implemented using 10-fold cross validation.

D. Driving Profile Label Algorithm

To establish driver profiles (driving stereotypes), we define an algorithm that compares the median values for each variable of each cluster (i.e., harsh braking, over speed, excessive throttle and over revs) with the quartiles in the box-plot for the entire group data. The categories defined for the attributes are **low**, **moderate**, **high** and **very high** occurrence of incidents, as detailed in Algorithm 1. Figure 3 shows the graphic representation of the determination of the labels.

Algorithm 1: Label behaviour

inputs: Box-plot for $variable_i$ (i:1..4) in cluster,
box-plot for all data for $variable_i$ for subgroup.

output: $label_i$ of behaviour for each incident variable

foreach *Incident variable* v_i, *i:1..4* **do**

- **if** $Median(variable_i) \leq Median(AllDataInSubgroupForVariable_i)$ **then**
  - $label_i \leftarrow$ "Low";
- **else if** $Median(variable_i)$ *IsBetween* $(Median(AllDataSubgroupForVariable_i)$ *and* $ThirdQuartile((AllDataSubgroupForVariable_i)))$ **then**
  - $label_i \leftarrow$ "Moderate";
- **else if** $Median(variable_i)$ *IsBetween* $ThirdQuartile(AllDataInSubgroupForVariable_i)$ *and* $Maximum(AllDataInSubgroupForVariable_i)$ **then**
  - $label_i \leftarrow$ "High";
- **else if** $Median(variable_i) > Maximum(AllDataSubgroupForVariable_i)$ **then**
  - $label_i \leftarrow$ "Very High";

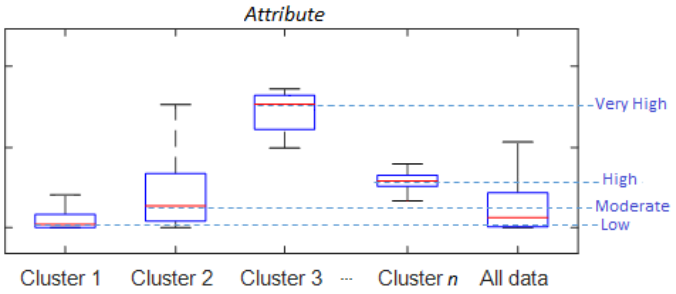


Fig. 3: Example of the comparison of each variable with the median values found in the cluster. The corresponding labels are on the right side of the figure.

Next we present the results of the framework applied to the HGV data. Initially, an attempt to cluster the entire dataset (not considering the data split per average distance) was carried out. Results revealed a small number of clusters, i.e. four. Based on the literature, however we hypothesise that more patterns can be uncovered.

IV. Results

In Figueredo *et al.* [13], three subgroups based on distance are identified (short, medium and long distance drivers). We adapt these groups to consider the average distance driven per day. KM groups the distance ranges and the proportions for the low, medium and high daily average distance groups, as shown in Table II. The two-stage framework is therefore applied to the data subsets within high, medium and low daily average distance travelled. For the next phase we perform consensus clustering within the subgroups found.

TABLE II: Three main subgroups based on average distance travelled

Groups	Average daily Km range	Number of drivers
1 - Low daily average	from 36Km to 220Km	5077
2 - Medium daily average	from 220Km to 350Km	8391
3 - High daily average	from 350Km to 750Km	7725

A. *Low Daily Average Distance Travelled Subgroup*

Validity indexes results indicate that three and four clusters are the most suitable numbers. Kappa index of 0.7725 on the consensus results confirms four clusters should be adopted (Kappa for 3 clusters is 0.3757). Stage 1 results for four clusters are shown in Table III (second column). 4,471 out of 5,077 drivers within this group are assigned to one of the identified clusters; 606 however are associated with no group. Stage 2 is therefore applied to the unclassified data; results are presented in the third column of Table III. It contains the number of unclassified drivers re-assigned to clusters, with p-value confidence interval greater than 0.05. These results were also verified by the DB index. The 'Total' column in Table III shows the final classification of drivers after combining the results of Stage 1 and 2. The number of unclassified drivers is reduced to 217. Stage 2 results show that 148 instances have been assigned to cluster 3 and another 241 instances form a new cluster. We achieved this outcome by employing the classification ensemble to the unclassified data and performing statistical comparison on the results. Table IV therefore shows the p-value comparison (for each feature) of the existing clusters found in Stage 1 against the new clusters formed after the ensemble classification (Stage 2). We employ unanimous and majority voting. For this case, both unanimous and majority voting show similar results. In order to achieve maximum classification, however, the drivers assigned to clusters in Stage 2 are obtained after majority voting (Table IV).

The instances assigned to cluster 3 (148 instances) have p-values greater than 0.05, except for overspeed; the statistical test results therefore confirm that those 148 instances belong to cluster 3 found previously in Stage 1. The drivers designated

TABLE III: Distribution of drivers among the groups for Low Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	3,107	-	3,107
2	177	-	177
3	432	148	580
4	755	-	755
New Cluster	**-**	**241**	241
Unclassified	606	217	217

TABLE IV: p-value of majority voting for Low Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Harsh Braking	0.08	0.004	**0.15**	0.73
Overspeed	$2.2x10^{-16}$	$2.2x10^{-16}$	**0.02**	0.5
Throttle	$2.2x10^{-16}$	0.008	**0.11**	$2.2x10^{-16}$
Over Revs	0.27	0.46	**0.22**	0.04
Number of drivers	**241**	60	148	157

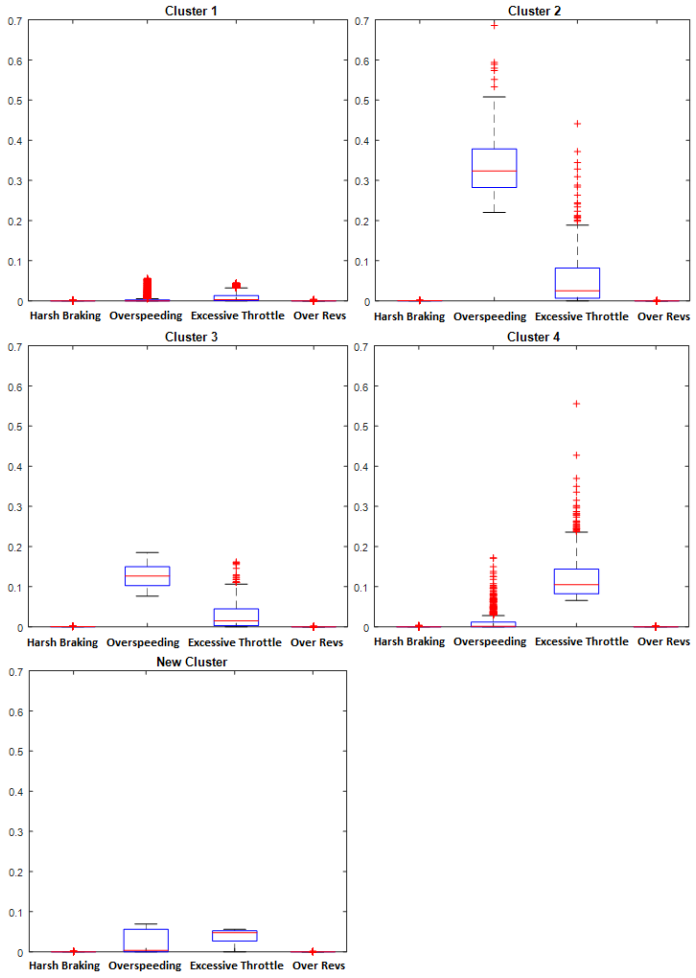


Fig. 4: Clusters within low daily average distance

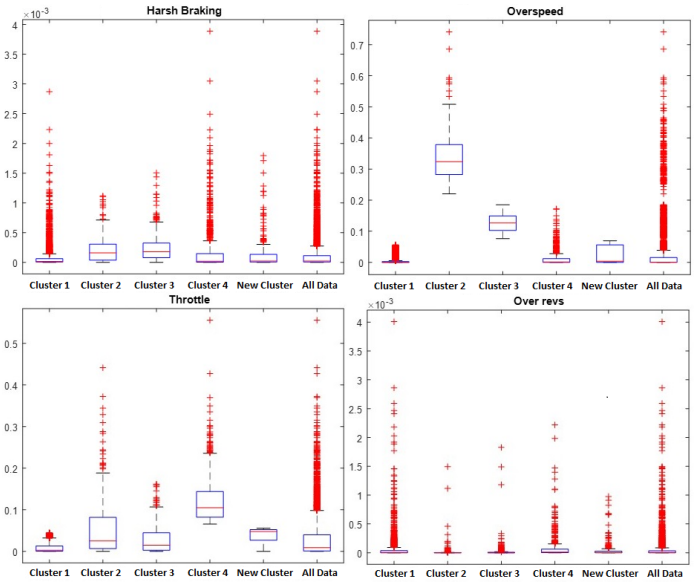


Fig. 5: Clusters within low daily average distance: comparing each variable with the median values found in the cluster.

to the other three groups by the ensemble classification had p-value smaller than 0.05 for most attributes, which indicates that they do not belong to the existing clusters and they are misclassified. In the bottom of the Table IV for the majority vote, it is possible to observe therefore that after Stage 2 there are 241 drivers misclassified as being from cluster 1; 60 instances misclassified as belonging to cluster 2 and 157 misclassified as being part of cluster 4. We check whether these groups of instances can possibly form a new pattern of behaviour. Algorithm 1 is therefore run on the driver groups of 241 drivers from Stage 2, as they form a large group. Results are compared with the current profiles to check if there are any differences in behaviour. As a result a new driving profile is uncovered. The remaining non-assigned instances are discarded (continue being labelled as 'unclassified'), as they are low in number for a new profile.

After Stage 2, we obtain the patterns of behaviour of Table V. Figure 4 also shows the cluster characterisation using box plots. Figure 5 shows the values for each variable in the clusters against the median of all low average daily distance drivers. From the figures, it is possible to observe that in the first cluster, the number of incidents are low. The second cluster comprises of drivers with the very high over speeding and high throttle values. Cluster 3 is characterised by high over speed and throttle, although these values are lower than in cluster 2. Cluster 4 has overall high throttle. The new cluster from Stage 2 has moderate values for harsh braking, overspeed and over revs, but has high values for throttle and low over revs. It can be observed that the algorithm proposed produces the same pattern for the second and third rows of the table. Although in Figure 5 there is a distinct difference between the values of overspeed between these groups, we give them the same label, as for both cases the median is above the maximum value from the box-plot of the entire dataset. In the future, however, if experts identify the need of distinguishing between both over speed behaviours, another label for speed (i.e., "extremely high") can be added to the algorithm. Lastly, it is important to notice that the outliers in the two figures represent the extreme driving behaviours within the identified clusters, shown in detail in the discussion section.

TABLE V: Profiles for Low Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throtle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	High	High	Moderate	Low	2
3	High	High	Moderate	Low	2
4	Low	Moderate	High	Moderate	3
New Cluster	Low	Moderate	High	Low	4

Table V shows the driving profiles after employing Algorithm 1 on the low daily average distance instances. It is possible to identify four patterns of behaviour: very safe drivers, who comply with traffic rules and have mostly moderate amounts of over revs; unsafe drivers, with high values of overspeed and harsh braking; groups 3 and 4 comprise mildly unsafe drivers, with distinct characteristics between them. Group 3 presents low harsh braking, moderate overspeed and over revs and high throttle. Group 4 has moderate values for harsh braking and overspeed, high throttle and low over revs.

TABLE VI: Distribution of drivers among the groups for Medium Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	4,863	-	4,863
2	892	288	1,180
3	1748	-	1,748
New Cluster	-	502	502
Unclassified	888	98	

B. Medium Daily Average Distance Travelled Subgroup

Validity indexes suggest the data should be divided either in three (second best number in certain cases), four or six clusters; however, three clusters produce a higher kappa (0.8026) and a smaller number of unclassified instances (888). 7,503 out of 8,391 drivers are distributed in three groups after Stage 1 (Table VI, second column). Stage 2 adds 288 drivers to the second cluster and detects a new cluster with 502 drivers (Table VI, third column). The final number of unclassified instances is reduced to 98. The p-value comparison of clusters detected after Stage 1 with those from Stage 2 is shown in Table VII. This classification was also confirmed by DB index.

Figures 6 and 7 show the cluster results for the medium daily average distance dataset. The first cluster comprises a pattern of behaviour identified previously, as it contains the group of safe drivers, with low values for all incidents. Cluster two differs from the first cluster mostly by the number of excessive throttle, which is much higher. The graph for cluster three presents high values for overspeed and throttle, and the fourth cluster appears to have lower values for harsh braking and over revs and higher values for overspeed and throttle. Table VIII shows the profiles detected after employing Algorithm 1 on medium daily average distance drivers. Among the four driving profiles, two are also present in the low average group, as indicated in the last column of the table.

C. High Daily Average Distance Travelled Subgroup

For the last subgroup, although there is more consensus (less unclassified instances) with three clusters (905 unclassified and kappa equals 0.7658), the kappa coefficient favours four clusters (0.7776). In addition, more patterns of behaviour are uncovered when four clusters are considered. After Stage 1, 6,810 out of 7,725 drivers are distributed in four clusters (Table IX, column 2). The third column shows Stage 2 results,

TABLE VII: p-value comparison of ensemble clustering with ensemble classification for Medium Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3
Harsh Braking	$6.1\text{x}10^{-6}$	**0.4**	0.07
Overspeed	$2.2\text{x}10^{-16}$	**0.75**	$2.2\text{x}10^{-16}$
Throttle	$2.2\text{x}10^{-16}$	$\mathbf{2.2\text{x}10^{-16}}$	0.02
Over Revs	$7.2\text{x}10^{-4}$	**0.41**	$8.1\text{x}10^{-5}$
Number of drivers	502	288	98

TABLE VIII: Profiles for Medium Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throtle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	Low	Low	High	Moderate	5
3	High	High	Moderate	Low	2
New Cluster	Low	Moderate	High	Low	4

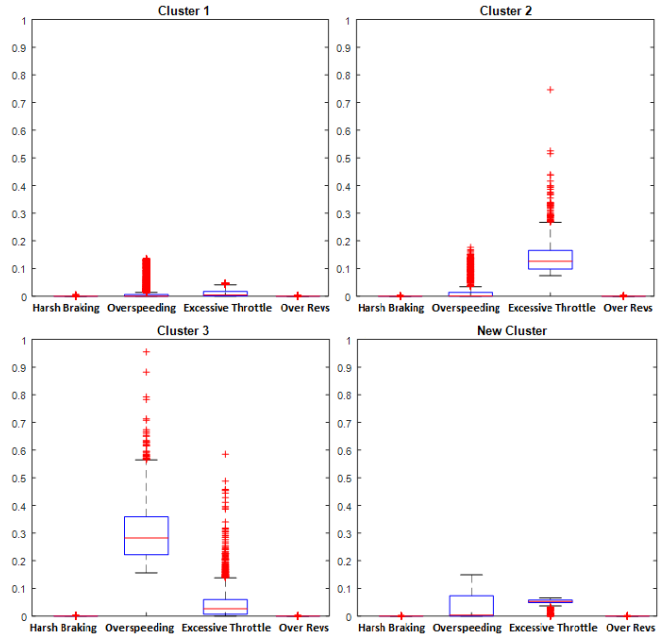


Fig. 6: Clusters within medium daily average distance

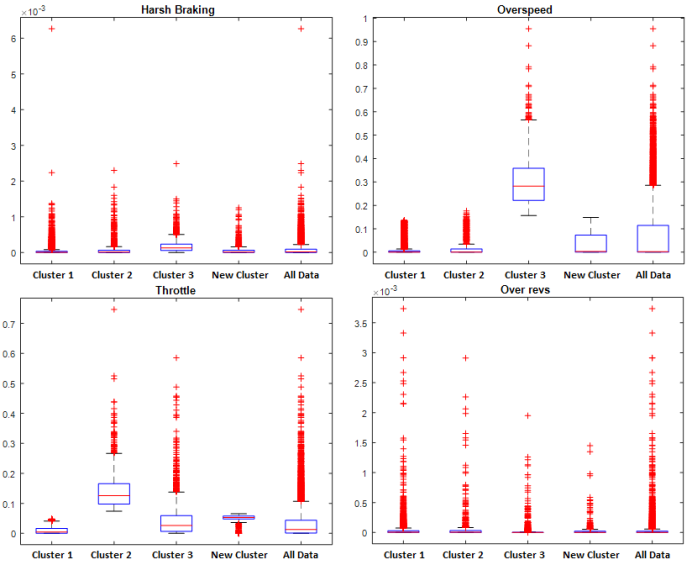


Fig. 7: Clusters within medium daily average distance: comparing each variable with the median values found in the cluster.

in which 31 instances are assigned to cluster 2; 254 drivers are added to cluster 3 and 59 are associated to the fourth cluster. A new cluster is identified and contains 569 drivers. The number of unclassified drivers is reduced to 2. The p-value comparison of clusters detected after Stage 1 with those from Stage 2 is shown in Table X, once again verified by DB index. Figures 8 and 9 characterise the clusters identified. In the table there is a cluster with low values of incidents (cluster one); cluster two is characterised by high harsh braking and overspeed and moderate throttle, whereas cluster three has drivers with high values for harsh braking and throttle. The fourth cluster is mostly characterised by high overspeed. The new cluster has

TABLE IX: Distribution of drivers among the groups for High Daily Average Distance

Cluster	Stage 1	Stage 2	Total
1	4,657	-	4,657
2	1,007	59	1,066
3	604	254	858
4	542	31	573
New Cluster	-	569	569
Unclassified	915	2	

TABLE X: p-value comparison of ensemble clustering with ensemble classification for High Daily Average Distance

Driving Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Harsh Braking	$8.7x10^{-8}$	**0.71**	**0.87**	**0.08**
Overspeed	$1.1x10^{-13}$	$\mathbf{7.1x10^{-12}}$	**0.06**	$\mathbf{2.2x10^{-16}}$
Throttle	$2.2x10^{-16}$	**0.7**	$\mathbf{2.2x10^{-16}}$	**0.52**
Over Revs	$3x10^{-15}$	**0.71**	**0.48**	**0.18**
Number of drivers	569	59	254	31

high throttle but presents low values for all the other variables. The complete characterisation of the stereotypes is found in Table XI. In the table, three new driving styles are found. These patterns are not identified for both low and medium daily average distance driven.

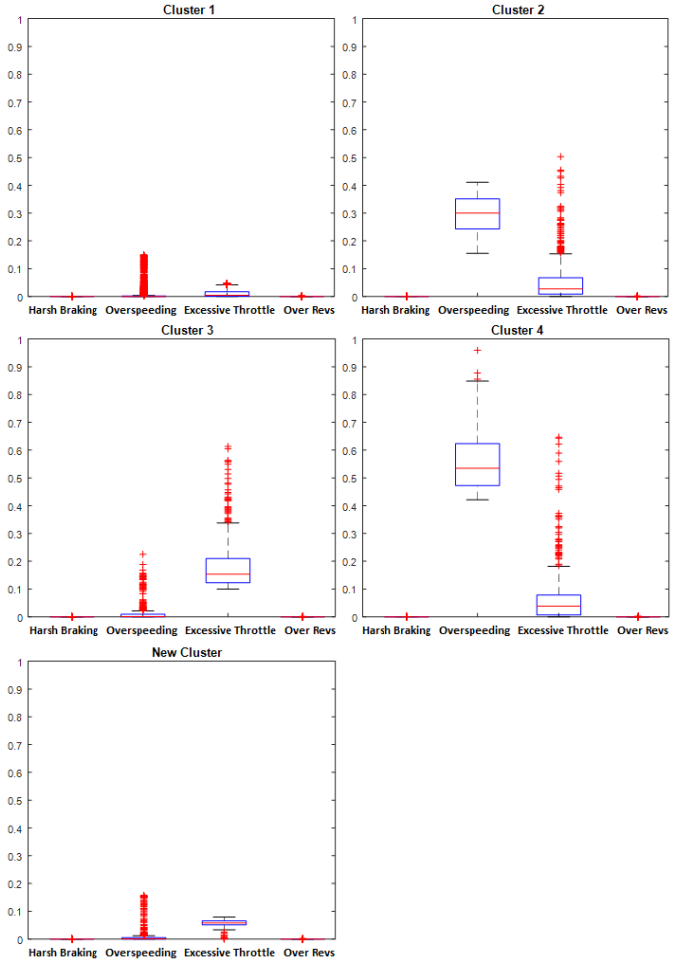


Fig. 8: Clusters within high daily average distance

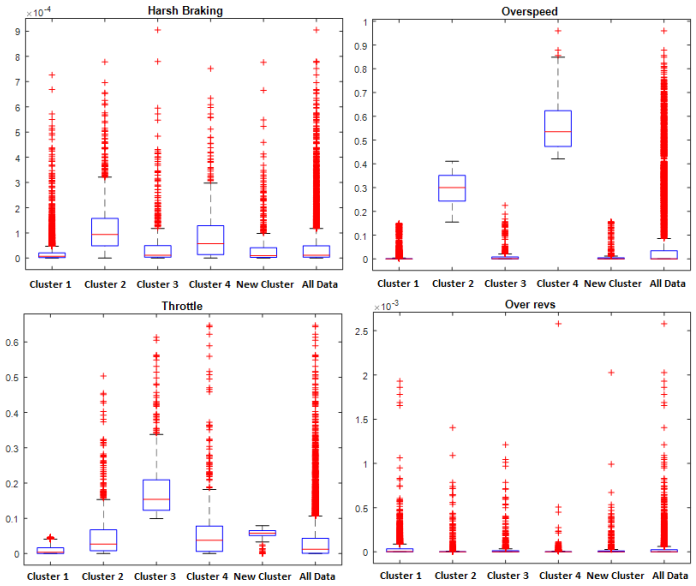


Fig. 9: Clusters within high daily average distance: comparing each variable with the median values found in the cluster.

TABLE XI: Profiles for High Daily Average Distance

Cluster	Harsh Braking	Overspeed	Throttle	Over Revs	Profile
1	Low	Low	Low	Moderate	1
2	High	High	Moderate	Low	2
3	Moderate	Moderate	High	Low	6
4	Moderate	High	High	Low	7
New Cluster	Low	Low	High	Low	8

V. Discussion

A. Profiles Identified

Eight driving profiles were uncovered, six obtained in Stage 1 and another two in Stage 2. The final set of driving profiles is displayed in Table XII. In the table, we also show the number of drivers present in each group and their proportions in relation to the entire dataset after each stage. The sixth column shows the numbers after the first stage; the seventh column displays the final results, after employing Stage 2. The last column shows in which subgroups (low (L), medium (M) or high (H) average daily distance) the profiles are found. The first stereotype (Profile 1) comprises of safe drivers and is common to all the three subgroups and represents the largest proportion of drivers in the dataset (59.58%). Profile 2 (aggressive drivers) is also common to all the three subgroups, characterised by high harsh braking and over speeding, but moderate throttle. It is the second profile highest in proportion, with 16.72%. The remaining profiles seem to appear in lower proportions within the UK drivers. Profiles 3 and 4 have similar number of harsh braking, overspeed and excessive throttle but differ only in over revs events. Profiles 5 and 8 present low harsh braking, low over speeding and high throttle; however, they differ only with regards to the number of over revs. It is important to notice that profiles 5 and 8 comprise medium and high daily average distance drivers, respectively. Drivers in profiles 6 and 7 differ by the number of overspeed events. The high throttle for profiles 5 and 8, rather than indicating bad driving behaviour, might suggest peculiarities of roads (high inclination), or the vehicles or the nature of the cargo being transported. Further information therefore is necessary to verify this hypothesis. If further evidence is provided, profiles 1, 5 and 8 might indicate very similar or even the same behaviour, which increases the proportions of very safe drivers.

After Stage 2, only 317 drivers remained unclassified. Further investigation is necessary to identify the profile to which these drivers belong to. Figure 10 shows the 3-D scatter plot of the clusters in the three driving groups. These were obtained by taking the principal components on the three groups. Figure 10 (a), (b) and (c) show the drivers among 5 groups for low, 4 groups for medium and 5 groups for high average distances, respectively. Lastly, figure 10 (d) show all the profiles combined in one scatter plot (Note: This figure also show the scatter plot rotated by 180 degrees, to show the profiles on the reverse side of the image). The four figures represent the driving behaviours within the identified profiles among all the three subgroups. In the figures, the corresponding profiles in all graphs have the same colour. For instance, Profile 1, which is common to all subgroups of average distance is shown in blue. Points in black are the unclassified instances. Profiles that are not common for the subgroups are represented in different colours.

In the figure it is possible to observe why so many points that seem to be outliers appear in the box plots of the clusters. There is a number of instances that is more sparce compared to the other points in the clusters, even though these instances still belong to the same pattern of behaviour as the clusters they have been assigned to. These points, however, are not outliers. The consensus clustering method would have separated outliers into extra clusters. In addition, we have attempted to remove the outliers that appear in the box-plots, and as a consequence, new outliers appear, demonstrating that the data is spread within the clusters. We believe therefore, that within a specific stereotype, these instances represent the extreme values.

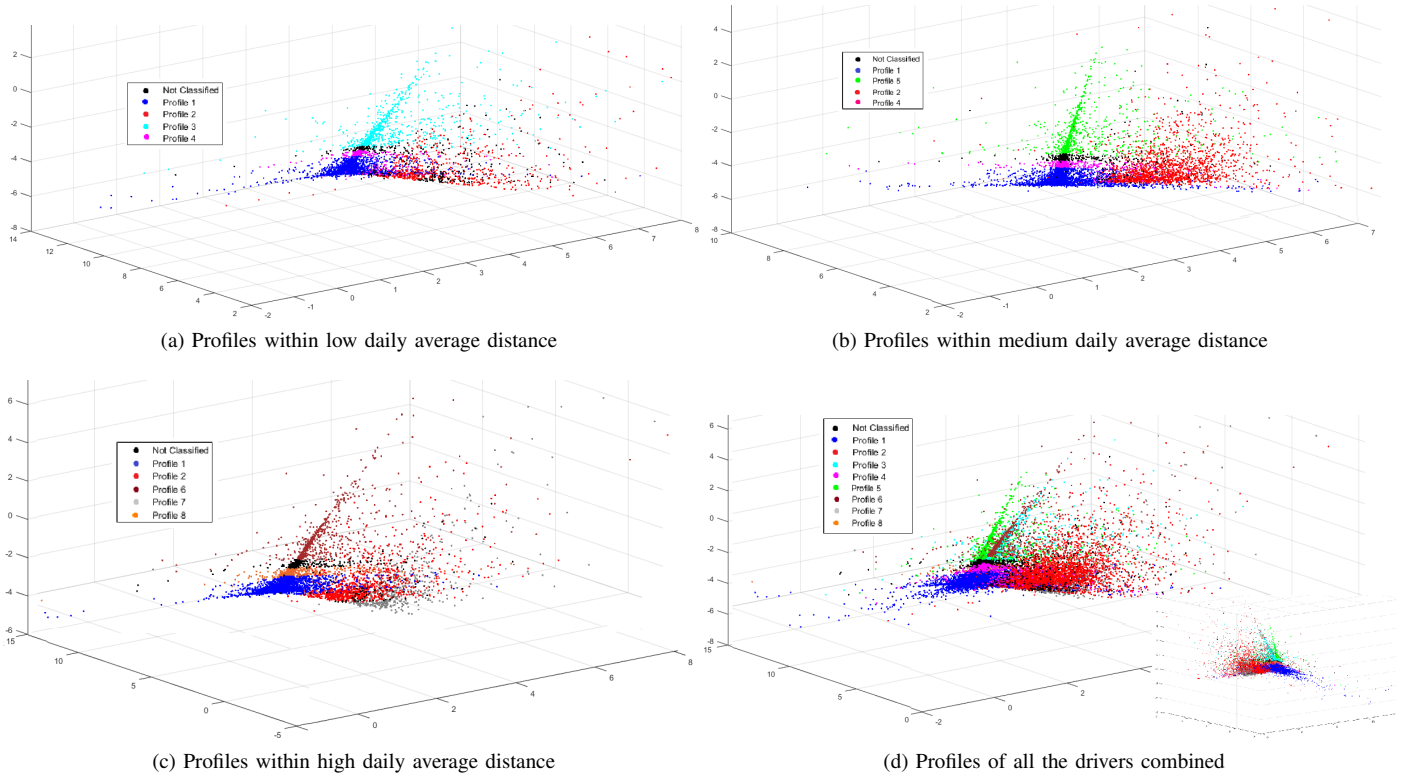
B. Comparison with the State-of-the-Art

Although there is no work in the literature with which we could directly compare our results (due to differences in the data collected, the types of drivers included and the methodology employed), we confirmed the existing research, as few stereotypes previously described were also found within our data. For instance, from the profiles identified by Constantinescu *et al.* [6] in Table I, we observed that their second profile is loosely similar (where the attributes considered in the data are the same) to our profiles 1, 5 and 8 of well-behaved drivers, under different average daily distances travelled. Similarly, our profiles 2 and 7 match their sixth cluster. The other groups found would require further investigation, as their criteria for classification of profiles and the attributes used for clustering are different. Other authors, such as Halim *et al.* [9] have also identified those types of behaviour for smaller vehicles.

Compared to the current methods found in the literature, our methodology has a few advantages. Our employed framework allows for more stable clusters, as results are obtained via consensus of multiple methods. The current literature mostly employs fewer clustering algorithms and does not consider consensus. In addition, in Stage 2 we further investigate the

TABLE XII: Driving Profiles

Profile Number	N. of Harsh Braking Events	Overspeed	Excessive Throttle	N. of Over Revs Events	N. of Drivers Classified after Stage 1	N. of Drivers classified after Stage 2	Sub-Group Label
1	Low	Low	Low	Moderate	12,627 (59.58%)	12,627 (59.58%)	L,M,H
2	High	High	Moderate	Low	3,364 (15.87%)	3,543 (16.72%)	L,M,H
3	Low	Moderate	High	Moderate	755 (3.56%)	755 (3.56%)	L
4	Low	Moderate	High	Low	–	743 (3.51%)	L,M
5	Low	Low	High	Moderate	892 (4.21%)	1180 (5.57%)	M
6	Moderate	Moderate	High	Low	604 (2.85%)	663 (3.13%)	H
7	Moderate	High	High	Low	542 (2.56%)	796 (3.76%)	H
8	Low	Low	High	Low	–	569 (2.69%)	H
Unclassified					2,409 (11.37%)	317 (1.5%)	



(a) Profiles within low daily average distance

(b) Profiles within medium daily average distance

(c) Profiles within high daily average distance

(d) Profiles of all the drivers combined

Fig. 10: Scatter plot of the drivers in their respective driving profiles for (a) low, (b) medium (c) high daily average distance and (d) the entire data set

data to identify behaviours not present in the clusters from Stage 1. This allows for a more detailed report on less frequent behaviour, which is yet statistically different from those groups detected in Stage 1. Regarding the data collection, we have advantage over current approaches, as we gathered real-world data during a longer period in a much larger area than current studies. We do not use artificial data (produced via simulation) and we do not interfere on driving behaviour during the data collection.

Our work, however is limited by the fact that there is no ground truth regarding HGV behaviour previously determined to validate our findings. In addition, there is still the need to further verify our results with experts and human factors researchers. Moreover, inputs related to road characteristics, weather conditions, traffic conditions, that are not included explicitly in the analysis, are themselves causes of incidents disassociated of driving behaviour. With our current data, therefore, it is not possible to distinguish output profiles caused by driving behaviour from those arising from other factors. We do believe, however, that by considering driving data across the whole year, under different conditions, and considering several roads and locations driven across the UK, the impact of those external factors is overall reduced.

C. Applications and Future Directions

As previously discussed in Section I, there are several possible applications of our findings. Informing about the patterns of behaviour and their proportions to stakeholders (managers, policy makers, vehicle manufacturers, etc.) will allow them to (i) understand what are the characteristics of UK HGV drivers and (ii) to investigate, together with human factors specialists, what mitigation actions could be taken to

educate drivers and to reduce the number of road incidents. In addition, in particular for telematics companies, such as Microlise, an implementation of a classification system able to predict the driver's behaviour could assist in the development of internal policies, methods for scoring, the creation of games and rewards for improved behaviour. One example of an existing action to reward driving behaviour is the Microlise Driver of the Year Competition [13]. This competition, however, focuses on past behaviour. By establishing an on-line, real-time classification approach, more effective and immediate reinforcement of good behaviour can be employed. Similarly, such classification can be used in inbuilt systems in smart cars to assess and trigger profile-specific actions for incident prevention. Further investigation regarding the impacts of the driving profiles on energy consumption can elucidate how to promote economy, especially regarding electric vehicles. The profiles found can also be translated into stereotypes defining agents behaviour in agent-based modelling simulation exercises. This allows for the creation of artificial laboratories to investigate traffic, incident prevention, responses to actions, etc. In addition, we are currently studying the correlation of these patterns of behaviour with incidents and accidents hot spots. We want to identify whether there are specific profiles responsible for high areas of incidents.

VI. Conclusions

In this work we report the identification of eight types of driving behaviour amongst UK HGV drivers. To achieve these results, we investigated a dataset containing incidents over the year of 2015 for 21,193 drivers. The incidents considered were harsh braking, over speeding, excessive throttle and over revving. To the best of our knowledge, this dataset was significantly larger and more diverse than those currently studied in the literature. This work is aimed at answering the following research questions: (i) what are the existing profiles within UK HGV drivers? And (ii) how do these patterns complement the current findings in the literature?

We employed a two-stage (Ensemble Clustering, as defined in Soria *et al.* [15], and Ensemble Classification, as defined in Agrawal *et al.* [27], [32]) pipeline to elucidate core groups from the data. Subsequently, 8 driving patterns were uncovered, not described previously. In Stage 1, 18,784 drivers out of 21,193 were distributed among six driving profiles. 2,409 (11.37%) drivers, however, remained unclassified. Stage 2 was therefore employed to tackle those instances. Two extra driving profiles were uncovered after Stage 2 and the number of unclassified drivers was reduced to 317 (1.5%). The use the 2-stage pipeline has lead to the formation of robust groups, which have been further validated with experts in the HGV industry.

Additionally, we also identified behaviour that occur possibly as a consequence of the vehicle driven and the road slope, such as profiles 5 and 8 in Table XII, in which the only event with high incidence was excessive throttle. Although there is no work in the literature with which we could directly compare our results, due to differences in the data collected and the methodology employed, we confirmed the existing research, as few stereotypes previously described were also found within our data. For further verification whether some types of behaviour arise caused by external, environmental factors, however, more observation and data collection regarding external factors would be necessary.

As we have worked with a real-world data set, current stereotypes of behaviour and their proportions within the UK were elucidated. We believe our findings will assist policy makers and stakeholders in the HGV industry to better understand their drivers and to define actions to improve road economy and safety. As future directions, we intend to merge the dataset from the year of 2015 with data from 2014 and 2016 to investigate whether stereotypes have changed over the years and whether their proportions change in different seasons. We also intend to include other types of incidents in the analysis, such as harsh cornering, which seem to be relevant for HGV and cargo safety.

VII. Acknowledgements

We thank Innovate UK for funding this project and Microlise for providing us the data.

References

[1] D. of Transport, "Road freight statistics," *Available at: https://www.gov.uk/government/collections/road-freight-domestic-and-international-statistics*, Last accessed 30 Sep 2016.

[2] I. Triguero, G. Figueredo, M. Mesgarpour, J. M. Garibaldi, and R. I. John, "Vehicle incident hot spots identification: An approach for big data," in *The 11th IEEE Int Conf On Big Data Sci and Eng*, 2017.

[3] G. P. Figueredo *et al.*, "Detecting danger in roads: An immune-inspired technique to identify heavy good vehicles incident hot spots," *IEEE Trans on Emmerging Topics in Comp Int*, 2017.

[4] van Huysduynen *et al.*, "Measuring driving styles: A validation of the multidimensional driving style inventory," in *Proc of the 7th Int Conf on Auto User Interf and Interactive Vehicular Apps*, 2015, pp. 257–264.

[5] F. Sagberg, Selpi, G. F. B. Piccinini, and J. Engstrm, "A review of research on driving styles and road safety." *Human Factors*, vol. 57, no. 7, pp. 1248–1275, 2015.

[6] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," *Int J of Comp Communications & Control*, vol. 5, no. 5, 2010.

[7] R. Kalsoom and Z. Halim, "Clustering the driving features based on data streams," in *Multi Topic Conference (INMIC), 2013 16th International*, Dec 2013, pp. 89–94.

[8] G. Castignani, T. Derrmann, R. Frank, and T. Engel, "Driver behavior profiling using smartphones: A low-cost platform for driver monitoring," *IEEE Int Transp Sys Magazine*, vol. 7, no. 1, pp. 91–102, 2015.

[9] Z. Halim, R. Kalsoom, and A. R. Baig, "Profiling drivers based on driver dependent vehicle driving features," *Appl Int*, vol. 44, no. 3, pp. 645–664, 2016.

[10] A. B. Ellison, S. Greaves, and R. Daniels, "Profiling drivers risky behaviour towards all road users," in *A safe system: expanding the reach: Australasian College of Road Safety national conference*, 2012.

[11] C. Saiprasert, S. Thajchayapong, T. Pholprasit, and C. Tanprasert, "Driver behaviour profiling using smartphone sensory data in a v2i environment," in *2014 Int Conf on Connected Vehicles and Expo (ICCVE)*, Nov 2014, pp. 552–557.

[12] Microlise, "Microlise (telematics, transport and fleet management solutions)," *Available at: http://www.microlise.com/*, Last accessed 08 Oct 2018.

[13] G. P. Figueredo, P. R. Quinlan, M. Mesgarpour, J. M. Garibaldi, and R. I. John, "A data analysis framework to rank HGV drivers," in *2015 IEEE 18th Int Conf on Intelligent Transp Sys*, 2015, pp. 2001–2006.

[14] C.-M. Tseng *et al.*, "A comprehensive analysis of factors leading to speeding offenses among large-truck drivers," *Trans Research Part F: Traffic Psychology and Behaviour*, vol. 38, pp. 171 – 181, 2016.

[15] D. Soria *et al.*, "A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients," *Comp in Bio and Med*, vol. 40, pp. 318–330, 2010.

[16] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, pp. 364–366, Jan. 1977.

[17] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[18] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 1, pp. 81–87, 1984.

[19] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids*, ser. Rep of the Fac of Mathematics and Informatics, 1987.

[20] X. Li, "K-means and k-medoids." in *Encyclopedia of Database Systems*, L. Liu and M. T. zsu, Eds. Springer US, 2009, pp. 1588–1589.

[21] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Com in Statistics-Simulation and Comp*, vol. 3, no. 1, pp. 1–27, 1974.

[22] J. A. Hartigan, *Clustering Algorithms*, 99th ed. New York, NY, USA: John Wiley & Sons, Inc., 1975.

[23] A. J. Scott and M. J. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, vol. 27, no. 2, pp. pp. 387–397, 1971.

[24] F. H. C. Marriott, "Practical problems in a method of cluster analysis," *Biometrics*, vol. 27, no. 3, pp. pp. 501–514, 1971.

[25] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *J of the Am Stat Assoc*, vol. 62, no. 320, pp. 1159–1178, 1967.

[26] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

[27] U. Agrawal, D. Soria, and C. Wagner, "Cancer subtype identification pipeline: A classifusion approach," in *IEEE World Cong on Comp Int (WCCI) - IEEE Cong on Ev Comp (CEC)*, 2016.

[28] A. J. Pinar, J. Rice, L. Hu, D. T. Anderson, and T. C. Havens, "Efficient multiple kernel classification using feature and decision level fusion," *IEEE Transactions on Fuzzy Systems*, 2016.

[29] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient knn classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.

[30] K. Gurney, *An introduction to neural networks*. CRC press, 2014.

[31] P. Thomas, M. Neves, T. Rocktäschel, and U. Leser, "Wbi-ddi: drug-drug interaction extraction using majority voting," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, 2013, pp. 628–635.

[32] U. Agrawal, D. Soria, C. Wagner, and J. Garibaldi, "Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles," *Artificial Intelligence in Medicine*, 2018.

AUTHORS' BIOGRAPHY

Dr Grazziela P Figueredo is a Senior Research Data Scientist at the Advanced Data Analysis Centre (ADAC) at The University of Nottingham. The focus of her research is the development and application of techniques for systems simulation and intelligent data analysis. She has been working with data analysis for a wide range of areas, including academic, medical and industrial clients. She specialised in providing consultancy within the University of Nottingham and externally, with the mission to enhance current research and business by providing state-of-the-art tools and the expertise to engineer and interpret data.

Utkarsh Agrawal received his Masters in Information Technology in 2015 form the Indian Institute of Information Technology Allahabad. He is currently a Ph.D. student at the The University of Nottingham. His research interests include data fusion, data analytics and applied machine learning in real world applications.

Jimiama M Mase Mafeni Mase Jimiama Mosima is a data analyst graduated from Nottingham University and is currently working at MnApro Cameroon, a company that distributes water purification tablets (Aquatabs) in Central Africa.

Dr Mohammad Mesgarpour Mohammad Mesgarpour received his PhD degree in Operational Research from the University of Southampton in 2012. He worked for the University of Nottingham as a KTP research associate for two years before joining Microlise in 2014 as a Technical Research Analyst. His main area of research is in the fields of Transport Management, Predictive Modelling, Data Analytics and Combinatorial Optimisation.

Dr Christian Wagner is an Associate Professor in Computer Science at the The University of Nottingham. His research focuses on modelling and handling of uncertain data decision support systems and data-driven policy design. He has published more than 80 peer-reviewed articles and is an Associate Editor of the IEEE Transactions on Fuzzy Systems journal. His current research projects focus on the development, adaptation, deployment and evaluation of artificial intelligence techniques in inter-disciplinary projects.

Dr Daniele Soria is a lecturer in Computer Science at the University of Westminster. His research interests include the development of decision support tools for real-world applications, with a particular focus on biomedical domains. Dr Soria has worked on several multi-disciplinary projects and generated impact when his work on refining the phenotypic characterisation of the breast cancer disease was the subject of extensive national media coverage.

Professor Jonathan M Garibaldi is Head of School of Computer Science at the The University of Nottingham. His main research interest is in developing intelligent techniques to model human reasoning in uncertain environments. Prof. Garibaldi has published over 200 articles on fuzzy systems and intelligent data analysis. He is the Editor-in-Chief of the IEEE Transactions on Fuzzy Systems.

Dr Peer-Olaf Siebers is an Assistant Professor in the School of Computer Science at the University of Nottingham. His main research interest is the application of data driven computer simulation to study human-centric complex adaptive systems. His current research focusses on Urban Sustainability. He is a COI in several related projects and a member of the University's "Sustainable and Resilient Cities" Research Priority Area management team.

Professor Robert I John received his Ph.D. degree in Fuzzy Logic from De Montfort University, Leicester, U.K., in 1979, 1981, and 2000, respectively. He worked in industry for 10 years as a mathematician and knowledge engineer developing knowledge based systems for British Gas and the financial services industry. Bob spent 24 years at De Montfort University. He has over 150 research publications of which about 50 are in international journals with over 6000 citations. Bob joined the The University of Nottingham in 2013 where he heads up the research group ASAP in the School of Computer Science.