The Impact of Driving Styles on Fuel Consumption: A Data Warehouse and Data Mining based Discovery Process

João C. Ferreira, José Almeida, Alberto Rodrigues da Silva

***Abstract*—This paper discusses the results of an applied research on the eco-driving domain based on a huge dataset produced from a fleet of Lisbon's public transportation buses, for a 3-year period. This dataset is based on events automatically extracted from the CAN bus and enriched with GPS coordinates, weather conditions and road information. We apply online analytical processing (OLAP) and knowledge discovery (KD) techniques to deal with the high volume of this dataset and determine the major factors that influence the average fuel consumption and then to classify the drivers involved according to their driving efficiency and, consequently, we identify the most appropriate driving practices and styles. Our findings show that introducing simple practices – such as optimal clutch, engine rotation and engine running in idle – can reduce fuel consumption in average from 3 to 5 l/100km, meaning that saving of 30l per bus on one day. These findings have been strongly considered in the drivers' training sessions.**

Index Terms—Driver Profile, Eco-Driving, Fuel Efficiency, Data Warehouse, Knowledge Discover, Public Transportation.

I. INTRODUCTION

Evaluating driver performance and promoting energy-efficient driving has received scarce attention from the research community. This is due to the difficulty of objectively evaluating human driver performance. The driver controls the speed, acceleration, braking, engine rotation speed, the gear engaged [1, 2], and the position of the vehicle on the street in an environment characterized by certain traffic conditions, itinerary, load, etc. Different driving styles result in different fuel consumption levels, thus related to driving efficiency. Different external conditions result in different levels of consumption. For example, levels of fuel consumption in a public bus are strictly linked to the number of stops made per itinerary unit. The number of stops is a parameter that is not controllable by the driver, but on traffic, the bus route, or the number of passengers.

One of the most efficient approaches to evaluate driver performance is to register a set of events (parameters) read from the CAN bus [3], which stores messages from all driving events on an on-board recorder, from where data is retrieved and stored in a database for subsequent analysis. Due to analytical needs and the huge data generated from this approach, this is a field where the application of online analytical processing (OLAP) and knowledge discovery (KD) techniques [4,5] is needed. KD in databases has been attracting a significant amount of interest from both research and industry. There is an increasing need for approaches and tools to assist humans in extracting useful information from the fast growing volumes of digital data. KD techniques are being applied with success into big data scenarios and in different application demands [6].

This paper discusses the results of an applied research on the eco-driving domain that used the data automatically produced from the Lisbon's fleet of public transport buses, for a 3-year period, and involving 1,041 drivers with 745 different buses and 73 carrier bus routes. We apply OLAP and KD techniques to deal with the high volume of data and extract the associated knowledge, i.e. the impact factors (IFs) that most influence the average fuel consumption (AFC) in liters per 100 km.

This paper is organized in eight sections. Section 1 presents the motivation and the goals of the research. Section 2 presents eco-driving initiatives and discusses related work. Section 3 presents the working methodology and the four main processes involved. Section 4 describes the pre-processing process that merges data from the CAN events' dataset with GPS coordinates, weather conditions and road-related information. Section 5 describes the OLAP process, namely by introducing data cube-based operations such as roll-up, drill-down, slicing and dicing. Section 6 describes the data mining process with the tools and algorithms used to identify the impact factors (IFs) that most influence AFC classes. Section 7 discusses the results and highlights the lessons learned from this research. Finally, section 8 presents the main conclusions.

João C. Ferreira (jferreira@deetc.isel.ipl.pt) and José Almeida (jadcda@gmail.com) belong to Group of Energy and Power Electronics, ALGORITMI Research Centre, University of Minho, Guimarães, Portugal and ADEETC at ISEL, Lisbon, Portugal; and Alberto Rodrigues da Silva (alberto.silva@tecnico.ulisboa.pt) is at Instituto Superior Técnico – Universidade de Lisboa & INESC-ID, Lisbon, Portugal.

II. Related Work

Driving style is seen as "the attitude, orientation and way of thinking for daily driving", and is usually captured by questionnaires and surveys [9-10]. Recent works use a virtual driving simulator to collect realistic driving data from human drivers and to model human driving actions [11], or to classify driving actions into styles by combining objective rank methods [12]. However, this driver behavior can be analyzed through real driving parameters obtained from vehicle interfaces such as the CAN bus or driving data obtained from mobile device sensors [13]. The availability of such data offers new opportunities due to the interpretation of raw data from real human drivers [14]. Also, it is used in the application of intelligent algorithms to identify changes in driver behavior [15]. The main problem concerning the collection of data from CAN buses is that handling a huge dataset of events is not an easy task and requires proper techniques and tools.

TABLE 1: ECO-DRIVING RELATED PROJECTS

Author (year) [ref]	Description	Achievements claimed
Rolim et at. (2014) [17]	Monitor driver behavior (20 drivers) with lessons towards fuel savings	Around 4,8% reduction in fuel consumption
Barth (2009) [18]	Investigates the impact of providing real time eco-driving advice to the drivers based on real-time traffic speed, density and flow	Reduction in fuel consumption of 10-20% can be achieved without a significant increase in travel time
Boriboonsomsin (2010) [21]	Investigate how real-time feedback affects driving behavior	20 sample drivers show a reduction of 6% in fuel consumption on city streets and 1% on highways
Wahlberg (2007) [22]	Monitored fuel consumption reduction in buses during the 12 months after training	2% reduction in fuel consumption
Zarkadoula et al. (2007) [20]	Training monitoring driving experience for a period of two months	Mentioned fuel savings on buses of 4.35%
Ecomove (2013) [23]	FP7 European project, for testing and evaluating a series of “green” technologies and applications that aim to reduce fuel consumption	Reduced fuel consumption and CO2 emissions in road transport by 20%
Raghu K. Ganti, (2010) [24]	GreenGPS gives drivers the most fuel efficient route for their vehicle as opposed to the shortest or fastest route.	10% reduction in fuel consumption
Bart Beusen (2009) [19]	Eco-driving, on 10 drivers, based on the evaluation of individual driving style analysis for 10 months (real data)	Initial (first four months) the average fuel consumption after the course fell by 5.8% and then became stable. Some tended to fall back into their original driving habits

This is due to the difficulty of objectively evaluating driver's performance. This topic also relates to previous work in eco-driving, which has primarily taken the form of advice to drivers based on the several studies performed. In an attempt to reduce fuel consumption and CO_2 emissions, the analysis of drivers’ behavior has been applied [18-20] and several companies have long before recognized the value of training their drivers to that purpose.

As identified in Table 1, numerous studies have started to address these issues by monitoring drivers’ behavior before and after they were given eco-driving sessions and presenting drivers with driver-specific information and the best practices.

Additionally, a list of practical recommendations and advice for eco-driving that may reduce fuel consumption [25], is also summarized in Table 2.

Moreover, the identification of driving actions that influence consumption is yet more important in the cases of bus fleets due to high consumption and the number of hours that drivers perform on a daily basis. As we discuss in this paper, investing in eco-driving education and in the promotion of good driving styles may result in relevant savings for transport and logistics companies.

TABLE 2: RECOMMENDATIONS FOR ECO-DRIVING

Recommendation	Description	Explanation
Engine rounds per minute (RPM).	Drive in the highest possible gear at the lowest possible RPM.	Fuel consumption is lower at low RPM due to friction.
Maintain a steady speed	Avoid constantly braking and accelerating	Fuel is primarily consumed when accelerating
Eliminate idling	It is more fuel efficient to switch off the engine than leave the engine running	An average modern vehicle uses around 0.9-1.3 liters per hour (l=h) during idling [26]
Braking	Slow down by using the engine brake or the neutral gear instead of the actual brakes.	Modern vehicles use no fuel when using the engine brake, i.e. the vehicle is in gear and the accelerator is released
Acceleration	High accelerations consume much more fuel	Recommendation of acceleration in low gears with the RPM below 2,000 for diesel and with the throttle at half position
Speed	Drive at or below the speed limit	Fuel consumption increases at higher speeds.
Weight and air resistance	Minimize extra weight and air resistance	Both increase the load on the engine, thereby increasing fuel consumption.
Approach curves	Performed at the correct speed and in the highest possible gear	This reduces the need to accelerate after the curve
Tire pressure	Incorrect tyre pressure increases the rolling resistance	
Vehicle consuming accessories.	Air-conditioning, heating and other accessories that consume fuel	

III. WORK METHODOLOGY

The work methodology of the research presented in this paper involves four main processes as suggested in Figure 1.

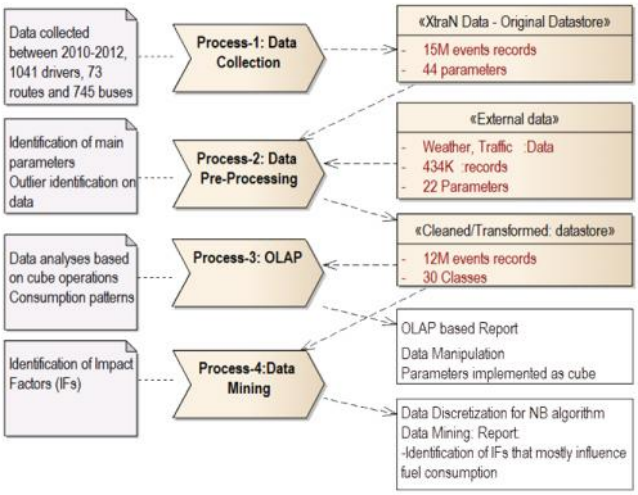

Fig.1. Work methodology conducted in this research.

Process-1: The data collection process occurred between 2010 and 2012 (from Jan 1st 2010 until Dec 31st, 2012, with 1096 consecutive days). This process collected driving event data from Lisbon's bus fleet and stored this information in a SQL database. This dataset has the following details:

1. During this 3-year period, around 1,500 different registered, professional company drivers were considered, with an average age of 39, 10 years of driving experience(on average) and 95% of the driver's population are male. From this data, we used 1,041 different registered drivers corresponding to those that were completely involved during this 3-year period. These drivers have an average driving time of 6,000 hours, distributed throughout 700 days of driving, on average. No Eco-Driving training was offered to these drivers during this period. The first training only occurred in January 2013, for a 6-month period, and was only given to the drivers who were identified with the worst fuel efficiency consumption in our study.
2. From 152 different routes, we only considered the city routes for our study during the day period (4am to 10pm), so this corresponds to 73 routes with an extension of 667 km. This produces 7.2 million route events over 19.7 thousand monitored hours with 70 million km over the 3-year period;
3. All the 745 buses analyzed in this study had diesel engines and manual transmissions with an average age of 7.6 years, distributed through 44 different models, mainly from Volvo (models B7L, B7R LE, B10L GNC and B7R LE MK3), MAN (models 18,310 HOCL-NL GNC, 18,310 HOCL-NL and 18,280 LOH 02) and Mercedes-Benz (models O405, O530 and OC500). All buses belong to the same company.

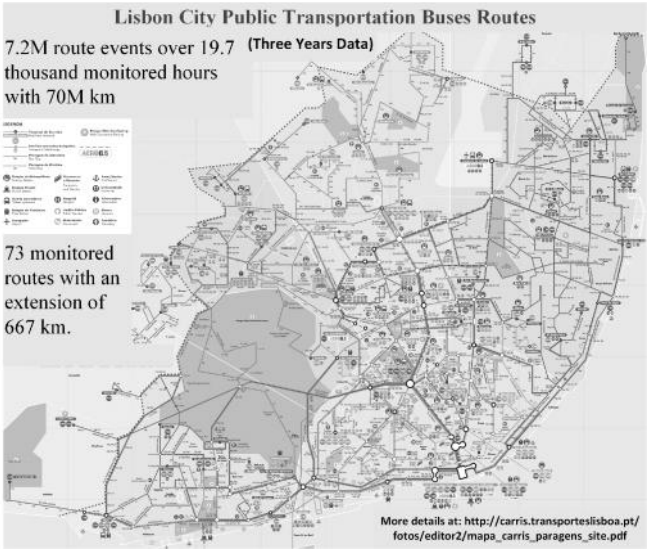

Fig. 2. Visualization of the City of Lisbon's map with its 73 bus routes.

This data acquisition process was performed by the XtraN (www.tecmic.pt/por/xtran) commercial product by Tecmic (www.tecmic.pt). This product allows us to measure 44 event parameters, such as engine, ignition on/off, acceleration, braking and clutch use, fuel consumption, engine RPM's and date. This data acquisition process uses Tecmic's 25 years of experience in this application domain. XtraN performs a continuous monitoring process with trigger events, like start/stop route carrier and a driver swap. Figure 2 shows the City of Lisbon's Public Transportation routes used in our study. For each bus route, the XtraN collected events in a pre-defined sample rate (about 30 seconds). This data was initially stored locally in XtraN, then transmitted to a cloud server database and subsequently stored in a SQL database containing 15M (million) events records.

Process-2: The data pre-processing process involves the identification data of the used, data manipulation towards a pre-defined class, performed by Tecmic experts as well as the identification of outliers.

Inconsistent data was also removed in an outlier's implementation approach [7]. This outlier process reduces the data in 20%, so after Process-2, we have 12M (completed) event records. Since we worked with 44 parameters, each outlier's identification (in any of these 44 parameters) originated the removal of the complete data of that event.

Secondly, we added external data, as suggested in Figure 3, to represent traffic information (available from a public web service) and weather conditions (available at the weather underground service www.wunderground.com). A unique point was assumed as the representative of the city's region, and information was used from one weather station in Lisbon, located near the Lisbon airport). An example of collected variables were: main sea level pressure, temperature, visibility (related to fog information), wind speed and rainfall. Weather information was used to check the influence on driving fuel consumption, namely the influence of rain and temperature, among other parameters. The weather-related data had an hourly periodicity, so the 3-year period meant 19,728 hours

with 21 parameters, in a grand total of 434,016 records. Data and time were used to merge the data in a process that we describe further in Section 4.

Process-3: The OLAP process (detailed in Section 4) performed data analysis and helped to identify the variables with more impact in what concerned fuel savings.

Process-4: Finally, a data mining process was used to determine the impact factors (IFs) that most influenced fuel savings (this process is also further detailed in Section 5).

We designed and built a decision support system that allows slicing and dicing by any selected dimension. Then, we applied Data Mining techniques to find the hidden patterns to allow the assessment of drivers, vehicles, routes, periods of the day and meteorological conditions in the selection of the most efficient entities in any given scenario.

This system was developed with the help of Microsoft SQL Server 2008R2 with SSIS (SQL Server Integration Services), SSAS (Microsoft SQL Server Analysis Services) and SEMMA (Sample, Explore, Modify, Model and Assess).

IV. Data Pre-Processing Process

Process-2 (see Figure 1) was supported by an SQL SSIS package to load a star-shaped data repository. Event records were analyzed to identify error measurements and inconsistent data based on outlier identification. By using the SQL tool and the Remove Outliers wizard, we can either display a graph, a line or a bar chart, to help you understand the distribution of all values.

The developed Data Warehouse follows the "star shape" design principle proposed by [27]. The analysis was conducted per Driver, Route, Bus Engine, Date and Time. Historical meteorological data was used as a characteristic of the day as well as a single meteorological station. The Weather Underground site was designated to represent the geographical area covered.

The 66 parameters collected were grouped by their origin into pre-defined classes (31 in total, see Figure 3 in bold). The data collected from CAN buses was reduced and transformed into 12 pre-defined classes (See Figure 3, classes under the issue "Transformed"). This process was performed by Tecmic experts, due to their deep knowledge of the data acquisition process. The information regarding engine RPMs was divided into three classes (green, yellow and red) and based on the engine type. The acceleration and braking events also allowed the measurement of the intensity (based on two pre-defined angular information concerning the acceleration and braking on the vehicle's pedal), thus representing excessive braking and acceleration events when the angular information of braking and acceleration was above a pre-defined threshold. From these operational variables, some ratios were derived as cubic calculated members to obtain the ratio of the sums, instead of the common trap of obtaining the sum of ratios, as data is being rolled up, down or sliced.

This dataset was also enriched with more data, directly captured from XtraN, namely, date (week, semester and year), route identification, driver identification, vehicle identification and GPS data. From all this data, we calculated, per bus route, the following parameters: average speed, distance and accumulated altimetry.

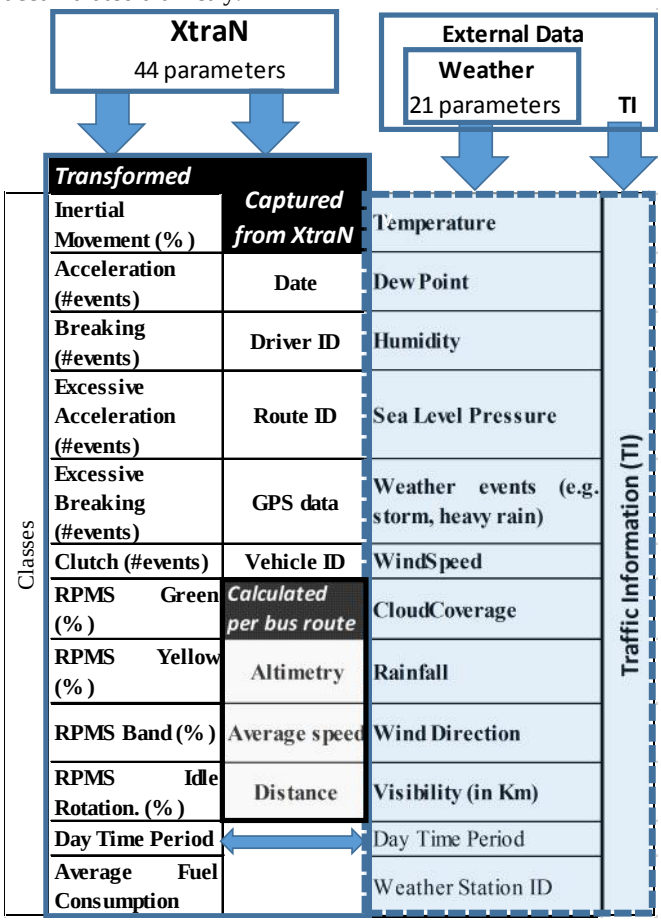

Fig.3. Data transformation: the major parameters involved.

Additional external meteorological data and traffic data, (consisting of 22 variables), were transformed into pre-defined classes. This generated dataset was huge, so intensive human analysis was complex, therefore we followed the "Online Analytical Processing (OLAP)" approach, which is named after a set of principles proposed by Codd [8].

We decided to apply KD and OLAP approaches in order to, technically, be able to answer the following questions: Who were the most efficient drivers per route and per bus? What are the IFs that most influence fuel consumption? To which drivers should we give specific training? What routes should we most care about? What are the best driving styles?

V. The OLAP Process

OLAP tools are designed to simplify and support interactive data analysis, but the goal of the KD approach is to automate this process, as much as possible, towards the identification of patterns. Pattern identification is based on fitting existing data to a model or commonly make any high-level description of a set of data. The KD process comprises many activities, namely data simplification from the outliers' identification, pre-processing, the search for patterns, knowledge identification,

and refinement. All these activities should be repeated in several iterations. This means that KD is ahead of what is currently supported by most standard database systems.

The platform used in this research has several OLAP features available to help analyze multidimensional data interactively, from multiple perspectives. In general, OLAP involves three analytical operations: (i) consolidation or roll-up; (ii) drill-down; and (iii) slicing and dicing. There are several analyzes to perform and we have the complete 130-page report available on request [7].

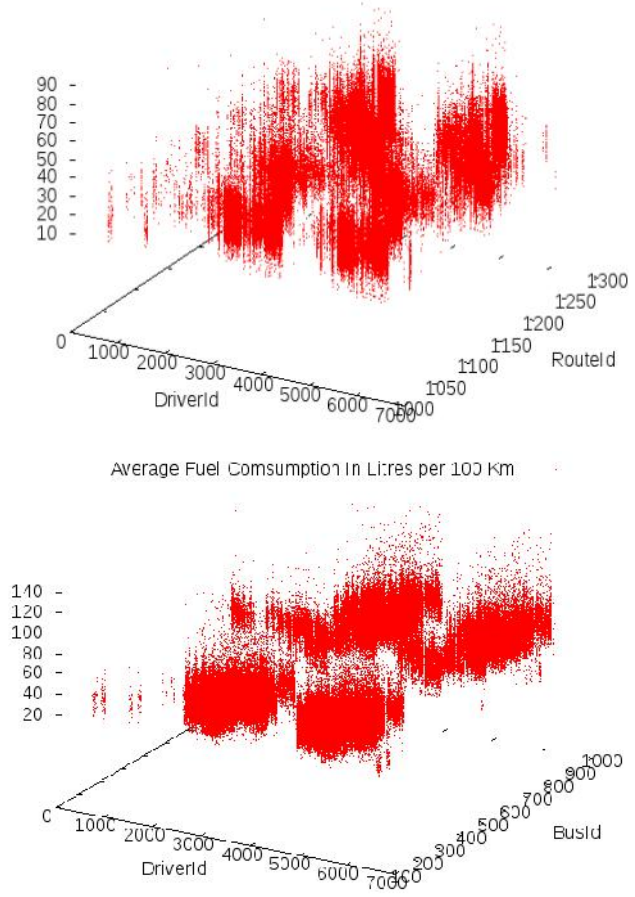

Fig. 4. Average fuel consumption (liters per 100 km) per Driver and Route (top, identified as (a)) and per Driver and Bus (bottom, identified as (b)). The vertical axis represents the consumption in liters per 100km of each event.

Slicing and dicing with the multidimensional cube (as a result of Process-2), supports the most common questions, such as: "What are the most efficient vehicles operating on a given route?" This can be verified by the combination of both graphics represented in Figure 4. For example, the average consumption of bus 886 on route 1021 is 37.2 liters per 100km whereas it is 54.2 liters per 100km on route 1024. This is one of the biggest and most difficult buses to operate in an inner city route with narrow streets. Figure 4a shows the consumption based on drivers and route taken and Figure 4b shows the consumption based on drivers and buses driven. As suggested in Figure 4a, it is necessary to check the diversity of the consumption in drivers using the same route or bus. For example, on the route identified as 1024, the average consumption goes from 37.2 l/100km (liters per 100km) to 61.8 l/100km amongst different drivers. This OLAP analysis allows multiple perspectives in multidimensional data arrays, where we can aggregate data in one or more dimensions interactively. An example of this approach is the identification of above-average consumption from major drivers on an inner city route. The comparison with weather variables makes it possible to verify that this high consumption was on rainy days, especially with heavy rain. From this route, it was possible to identify problems with storm water runoff, for example, flood situations. Figure 4 shows a diversity of consumption levels that have an impact on drivers, routes and buses, but do not identify their causes. For that reason, we will apply a data mining process, described in the next section.

VI. The Data Mining Process

The major objective with the data mining process was the analysis of fuel consumption per driver. The approach was complex due to the 30 dimensions involved. In the literature, we find several data mining approaches and algorithms [5], but because we had training data, we used the Naive Bayes (NB) approach supported by the Microsoft platform (SQL Server 2008R2). This approach has a better performance with discrete data [5,28]. For this reason we performed a discretization process based on the following:

1. Heuristics were applied based on pre-defined criteria. One example is the time variable, for example, 8.35am or 9.45am has no particular meaning. When we want to extract knowledge, it is better to associate 8.35am to a discrete period of time, e.g. the "morning rush hour" or 10.45am to "day time". So the time associated with the DTP class (day time period) is divided into five subclasses: DTP_1 week (periods from 5am to 8am, 10am to 5pm and 8pm to 11pm); DTP_2 morning rush hour (8am to 10am); DTP_3 afternoon rush hour (5pm to 8pm); DTP_4 weekend and bank holidays, DTP_5 night period (11pm until 5am). Another example is the data variable that can be classified in winter or summer period, weekday or weekend, and also, the school and holiday period.
2. Equal area clustering based on the method of equal areas [7,27]. The main idea is to divide the data population in subclasses with approximately the same number of events. An example of this is the average fuel consumption (AFC), by imposing it onto the DISCRETIZATIONED SSAS type in the training dataset, with the input of five subclasses. This discretization process is based on an interactive maximized expectation algorithm, in order to divide training data into groups of similar population size. The output of this process applied to AFC is shown in Figure 5 with the creation of five subclasses. This technique was selected due to the presence of pronounced peaks, and because this method selects ranges of buckets to contain equal quantities of cases. A result of this process is the subclass division in Figure 5, where the disparity of consumption events from 20 to 110 liters per 100km was divided into five subclasses.

3. The division of data population based on percentage. We decided to divide each class into five subclasses, namely: on average, below average, above average, extreme above average and extreme below average. For example, the acceleration events (class Ac), captured from XtraN data, may have events ranging from 100 to 3,500 (these numbers mean the number of times in 100km that the accelerator pedal was used). The first subclass, AC_1, ranges from 50 to 510, so the upper limit corresponds to 15% of 3400 (maximum value less minimum value). The second subclass, AC_2, goes from 510 to 1,190 (so the upper limit corresponds to 35% of 3,400). The third subclass, AC_3, goes from 1,190 to 2,210 (so the upper limit corresponds to 65% of 3,400). The fourth subclass, AC_4, goes from 2,210 to 2,890 (so the upper limit corresponds to 85% of 3,400). And the last subclass, AC_5, goes from 2,890 to 3,400.

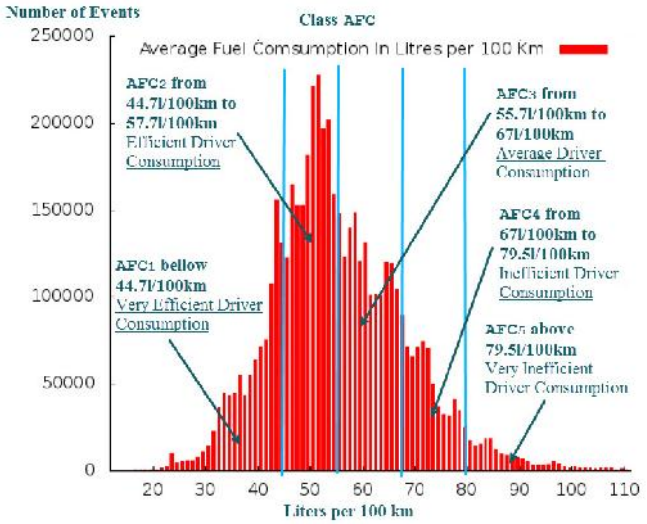

Fig. 5. AFC subclasses created from the discretization process.

Table 3 shows the result of this discretization process for the main event classes. Since it is not possible to show the representation of all data, we highlight a few major findings: the number of excessive braking (EBr) events per 100km occurs more than excessive acceleration (EAc) (approximately twice as much) but the number of Ac events tend to occur more than Br events per 100km. These events – acceleration, braking and also clutching – are related to traffic conditions and weather conditions (especially heavy rain). Traffic situations, on average, also increase these events.

After the discretization process and taking into account the data available, we use Naive Bayes' algorithm to identify the main IFs that may have a major influence on fuel consumption. This process is performed based on the estimation of probabilities, $P(AFC_j|C_k)$, which means the probable fuel class determination based on C_k measurements performed using the Bayes theorem, where j represents the AFC five subclasses and k the number of event classes:

$$P(AFC_j|C_k)=\frac{P(C_k|AFC_j)P(AFC_j)}{P(C_k)}\propto P(C_k|AFC_j)P(AFC_j)=P(C_k|AFC_j) \quad (1)$$

$P(AFC_j)$ is the same for all AFC subclasses because of the discretization based on equal areas. C_k is based on the measurements of Table 3, including weather, traffic and altimetry information (the parameter Km_Slope_acc_time is the total amount of seconds whilst accelerating with an altimeter change).

Considering the lower consumption subclass AFC_1 (below 44.7 liters per 100 Km), the IF identification towards fuel efficiency for the driver's case is based on the highest probabilities of:

$$\sum_{K=1}^{30}P(C_k \mid AFC_1)=1 \text{ and } P(C_k \mid AFC_1)=\frac{n^o \textit{ events of } C_K \text{ for } AFC_1 \text{ in TS}}{n^o \textit{ events of } AFC_1 \text{ in TS}} \quad (2)$$

where TS means training set (around 2.2M events in each AFC class). Based on a collection of 12M records, we estimate all these probabilities (count operation based on data training set). IFs were determined based on the top-10 probabilities for the AFC_1.

Taking into account Formula (2), a diversity of probabilities can be calculated using the 3-year dataset. Table 4 shows the highest values for these probabilities, for two lower fuel consumption subclasses. This process allows the identification of the IFs that most influence the reduction of fuel consumption, namely: (1) lowest clutch use; (2) maximize the time using inertial movement; (3) maximize the time of rotation in idle; (4) minimize the time of excessive engine rotation, avoid yellow and red band (this is already avoided); (5) avoid daytime periods of traffic.

TABLE 3: DISCRETIZATION INTERVALS OF THE MAIN EVENT CLASSES

Event Classes		**Subclasses**				
	Abr	**1**	**2**	**3**	**4**	**5**
Inertial Movement (%)	***IM***	<3	3 to 5	5-8	8-11.5	>11.5
Acceleration (#events)	***Ac***	<510	510-1,190	1,190-2,210	2,210-2,890	>2,890
Braking (#events)	***Br***	<500	500-1,155	1,155-2,145	2,145-2,805	>2,805
Excessive Acceleration (#events)	***EAc***	<38	38-88	88-163	163-213	>213
Excessive Braking (#events)	***EBr***	<50	50-116	116-217	217-283	>283
Clutch (#events)	***Cl***	<276	277-604	604-976	977-1,223	> 1,223
RPMS Green band (%)	***RG***	<4	4-15	15-50	50-80	>80
RPMS Yellow band (%)	***RY***	<0.5	0.5-2	2-4.5	4.5-9	>9
RPMS Red Band (%)	***RR***	<0.1	0.1-0.2	0.2-0.5	0.5-1	>1
RPMS Idle Rotation. (%)	***RI***	<24.2	24.2-31.1	31.1-36.8	36.8-43.2	>43.2
Day Time Period	***DTP***	Weekday	Morning rush hour	Afternoon rush hour	Weekend and Holidays	Night

TABLE 4: TOP NB PROBABILITY SUBCLASSES FOR AFC_1 AND AFC_2.

NB Probabilities	Sub-Classes (X)									
	Cl_1	RY_1	Cl_2	IM_5	RI_4	RI_5	DP_5	DP_4	RY_2	IM_4
$P(X\|AFC_1)$	24.5%	2.1%	15.1%	10.5%	8%	7%	3.6%	3.3%	2.5%	0.1%
$P(X\|AFC_2)$	12%	24%	5%	2%	12%	11%	2%	1.5%	10%	9%

Weather parameters were also considered and we concluded that extreme weather has impact on fuel consumption. This effect does not appear as the major consumption parameter listed in Table 5, due to good weather conditions in Lisbon. Examples of this can be seen on temperature and visibility distance distribution events in Figure 6. Extreme weather events are very few, but that affects driver and engine behavior. The main parameters on extreme weather, are rain (mainly heavy rain that appears in 40 days of the 3-year period, representing less than 0.1%), high temperatures (above 30ºC) and some fog in weather events. These events represent less than 1% of weather events and are mainly the high temperature during some of the days in the summer period. We found a correlation between hot days, with the class engine rotation in idle (RI) thus increasing times, because the drivers do not tend to turn off the engine between route carriers, mainly due to the use of air conditioning.

Only taking weather variables into consideration, the NB probabilities are represented in Table 5 only taking into account the effect of weather and distributed by their impact on the five *AFC's*. The weather data was captured on an hourly basis on most of the available parameters. The majority of these parameters were divided into three subclasses, for example, the temperature (T) was divided into T_1 for temperatures below 4ºC, T_2 for the temperatures in the range of 4ºC to 30ºC and T_3 for temperatures above 30ºC. The same for points of pressure, humidity or dew. The rain was divided into four subclasses: R_1 - no rain, R_2 - drizzle (drops with diameters of less than 0.02 inches falling closely together), R_3 - light rain (less than 0.1 inches in an hour) R_4 - moderate rain (more than 0.1 and less than 0.3 inches in an hour) and R_5 - heavy rain when it's above 0.3 inches in an hour. Visibility was mainly satisfactory during the 3-year data: we only had bad visibility for 20 days, totaling 80 hours. This visibility parameter is divided into two subclasses: fog and good visibility. The AFC_5 the R_5 (heavy rain) and T_3 (hot temperature) being the most influential. The AFC_4, R_4 rain and T_2 temperature subclasses are also the main influential weather parameters. From the data available, there was also a correlation between R_3 events and traffic. In an R_5 day's events, traffic news increases and also its severity. There is also a correlation between the R_5 and R_4 events with traffic information and the RI class (the time percentage of an engine in idle increases).

In Figure 7, the majority of events above 40% were traffic situations and on the left side of Figure 7, some routes show a higher percentage of RI, which means a route with more traffic and could also imply that the buses were overcrowded with passengers (longer times at bus stops also increased this RI time percentage). If the traffic is reduced, the increased time on the RI class could mean longer times at bus stations, in the loading and unloading process, due to the high flow of passengers.

TABLE 5: NB PROBABILITIES FOR ALL CONSUMPTION CLASSES TAKING INTO ACCOUNT ONLY WEATHER PARAMETERS

		All	AFC_1	AFC_2	AFC_3	AFC_4	AFC_5
Probability	$P(T_2\|AFC_x)$	70%	90%	22%	79%	42%	1%
	$P(R_4\|AFC_x)$	19%	1%	42%	12%	44%	1%
	$P(R_5\|AFC_x)$	5%	0%	1%	7%	1%	39%
	$P(Fog\|AFC_x)$	2%	1%	9%	0%	1%	1%
	$P(T_3\|AFC_x)$	1%	5%	2%	1%	1%	43%
	Other Weather classes	3%	3%	24%	2%	6%	16%

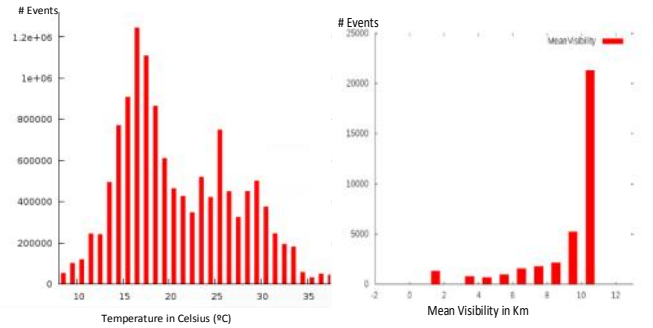

Fig. 6. Temperature and visibility weather parameter distribution over the 3-year period.

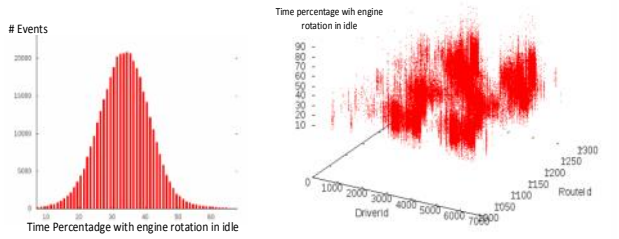

Fig. 7. Rotation with engine rotation in idle distribution events (left) and distribution events per driver and route (right).

TABLE 6: CONFUSION MATRIX WITH THE PRECISION MEASUREMENT

Actual Class	Predicted Class				
	AFC_1	AFC_2	AFC_3	AFC_4	AFC_5
AFC_1	90%	9%	1%		
AFC_2	9%	75%	9%	6%	1%
AFC_3	1%	9%	70%	13%	7%
AFC_4		6%	13%	64%	17%
AFC_5		1%	7%	17%	75%

The results of this approach are discussed in the next session (Lessons Learned). This study focused its analysis on the driver's behavior, but the same might be applied to other types of analysis, such as focused on route type or on bus type.

NB also allows the prediction mode that is used for the estimation of driving actions in fuel consumption. We used 70% of the data for training (around 8M event records) and the others to evaluate the results. Table 6 shows the confusion matrix of the results in terms of precision, with the predicted class in columns. Precision is lower at high AFC subclasses

and most of this wrong classification is due to measurements near the subclass' limits. AFC_1 has a 90% correct prediction and 9% of errors are due to values near the class limits. This precision worsens in high AFC subclasses mainly because 92.3% of the driver population belong to the first three subclasses and middle subclasses have more neighbors. Since the main goal of this research was the identification of IFs, we did not explore this prediction aspect.

VII. LESSONS LEARNED

Taking into account the impact of the driving style of bus fuel consumption, the data from our study shows a huge disparity among drivers' consumptions as seen in Figure 5. Table 7 shows the average annual consumption among all driving events and the results show a continuous improvement. These improvements were due to the introduction of our study as a monitoring activity with the identification of IFs.

TABLE 7: AVERAGE FUEL CONSUMPTION PER YEAR IN LITERS PER 100 KM

AFC 2010	AFC 2011	AFC 2012	AFC 2013	AFC 2014
56.6	**54.4**	**52.6**	**50.1**	**47.2**
Note: Data from 2013 was not shown in our study; Data from 2014 only included data from the first semester.				

Considering the AFC_1 subclass, the IFs on fuel efficiency for the driver's case are shown in Table 4 and Table 10, where the top-10 IFs are highlighted:

IF-1: Use of clutch events;
IF-2: Observation of optimal engine rotation;
IF-3: Minimum engine idling;
IF-4: Maximize the inertial movement;
IF-5: Day period associated with traffic;
IF-6: Use of excessive acceleration events;
IF-7: Use of excessive braking events;
IF-8: Use of acceleration events;
IF-9: Use of braking events; and
IF-10: Severe weather conditions.

That means that more efficient drivers tend to use the clutch less times, thus enjoying more inertia, and the percentage using the engine to idle and less time in the range "Yellow and Green" rotations.

It is similarly trivial to determine which driver styles to promote in order to maximize fuel efficiency. For example, to educate a driver to move from the AFC_2 into the AFC_1 class: emphasis should be placed on the importance of optimal clutch, engine rotation and avoiding engine running in idle. Table 8 shows the IFs to change from AFC_2 to AFC_1. The most important IF is the clutch use per 100km, the AFC_1 driver uses the clutch, on average, 276 less times in 100 km (belonging to subclass Cl_1) and the AFC_2 drivers' use of the clutch parameter is in the interval of 276 to 604 times (belonging to subclass Cl_2). This also depends on the route driven, but in the same route it is possible to see considerable differences among the drivers. The negative values in Table 8 identify the probability of change from AFC_1 to AFC_2.

TABLE 8: IF PROBABILITIES TO CHANGE FROM AFC_2 TO AFC_1

Sub-Classes (X)									
Cl_1	RY_1	Cl_2	IM_5	RI_4	RI_5	DP_5	DP_4	RY_2	IM_4
50%	22%	-40%	13%	-10%	-7%	5%	4%	-6%	-4%

This research also permits the identification of the main IFs dependency on a scale from the strongest to the weakest. Traffic DTP (day time periods) is also an important IF, identified as rush hours in the morning and afternoon period. Traffic is also related to weather conditions - mainly rain and bad visibility events. R_4 and R_5 and fog event subclasses are present in more than 70% of traffic events. Severe weather conditions show an impact on AFC, but due to a reduced number of these events in the Lisbon area, the overall view of this effect is neglected. However, we simulate the effect of a higher percentage of these events (more than five times) and so the effect of (heavy) rain appears in the top-5 IFs.

Table 9 shows the number of drivers in each class (the average consumption during the 3-year period), but only 1,041 drivers were considered, corresponding to a complete 2-years of data registers of day time drivers that performed all 73 routes. Since each driver has different consumption patterns, the class in which he belongs is determined by the AFC class in which he has more events.

TABLE 9: DISTRIBUTION OF DRIVER POPULATION PER AFC CLASSES

	All	AFC_1	AFC_2	AFC_3	AFC_4	AFC_5
Population	1,041	368	394	202	43	34

As a relevant result of our research, we identify the following lessons learned and recommendations regarding eco-driving practices:

Use the clutch moderately (IF-1): This practice is related with the change of speed (braking and acceleration) only available in manual transmission engines (MTE). Most eco-driving studies are based on an automatic transmission engine and do not take this MTE parameter into account. This class of events averaged in the range of 300 to 650 events per 100 km, with the maximum value of 1,840 events per 100 km.

Keep a stable engine round per minute (IF-2): We added few red class events (drivers did not force the engine, perhaps due to their knowledge of this monitoring process), but the yellow band class was detected in all drivers, on average of 1 to 2% of the time and from an NB probability (see Table 4).

Maintain a steady speed (IF-3): This is related to inertial movement. This can also be related with Ac and Br that we accounted for separately. This class of events averaged in the range of 4% to 8% of time, with a maximum value of 14%.

Eliminate idling (IF-4): Original data showed this was the main IF because drivers did not turn the engine off between driving periods (mainly for air-conditioning use). The idling time is directly related to the traffic as well as the number of passengers at a bus stop. This class of events averaged in the range of 25% to 35% of time, with a maximum value of 80%.

TABLE 10: NORMALIZED PROBABILITY VALUES OF THE FIVE TOP IFS DIVIDED BY THE FIVE PRE-DEFINED AFC CLASSES

		AFC_1	AFC_2	AFC_3	AFC_4	AFC_5
IF-1: Clutch Class #events per 100km	Cl_1 <276	87%	47%	30%	29%	20%
	Cl_2 [276-604]	12%	46%	50%	32%	30%
	Cl_3 [604-976]	1%	5%	15%	17%	29%
	Cl_4 [976-1223]	0%	2%	4%	13%	11%
	Cl_5 >1223	0%	1%	2%	9%	10%
IF-2: RPMS Yellow Band in time %	RY_1 <0.5%	77%	50%	30%	20%	5%
	RY_2 [0.5-2]%	17%	34%	35%	25%	43%
	RY_3 [2-4.5]%	1%	9%	13%	11%	16%
	RY_4 [4.5-9]%	5%	5%	12%	27%	16%
	RY_5 >9%	1%	2%	10%	17%	20%
IF-3: RPMS Idle Rotation in time %	RI_1 <24.2%	10%	15%	18%	44%	70%
	RI_2 [24.2-31.1]%	15%	27%	22%	28%	17%
	RI_3 [31.1-36.8]%	24%	23%	29%	14%	9%
	RI_4 [36.8-43.2]%	28%	22%	21%	9%	3%
	RI_5 >43.2%	23%	13%	10%	5%	1%
IF-4: Inerial Movement in time %	IM_1 <3 %	1%	4%	11%	20%	42%
	IM_2 [3-5]%	9%	13%	38%	42%	32%
	IM_3 [5-8]%	37%	30%	20%	22%	19%
	IM_4 [8-11,5]%	27%	29%	18%	11%	5%
	IM_5 >11,5%	26%	24%	13%	5%	2%
IF-5: Day Time Period Class	DTP_1	21%	31%	36%	23%	19%
	DTP_2	10%	14%	16%	25%	29%
	DTP_3	9%	14%	17%	26%	29%
	DTP_4	31%	20%	15%	13%	12%
	DTP_5	29%	21%	16%	13%	11%

Minimize the use of braking (EBr is the IF-7 and Br is the IF-9): We defined two classes of braking: Br and EBr when the intensity of braking is above a certain value. This EBr appears, on average, for all drivers in 120 to 160 events per 100km with a maximum of 333 per 100km. Br events, on average, range from 750 to 1,250 events per 100km with a maximum of 3300 per 100km.

Minimize the use of acceleration (EAc is the IF-6 and Ac the IF-8): We defined two classes of acceleration: Ac and EA when the intensity of braking is above a certain value. The EAc appears, on average, for all drivers from 50 to 100 events per 100km (less than EBr) with a maximum of 250 events per 100km. *Ac* events, on average, range from 1,100 to 1,900 events per 100km with a maximum value of 3,400 events per 100km (more than Br).

Other conditions and practices that might well influence the eco-driving style: Severe weather conditions show an evidence on the impact of fuel consumption, mainly regarding temperature, which is related to the use of air-conditioning and heating. Rain also has an important impact due to its direct correlation with traffic. In addition, the weight of a vehicle, air resistance, approach to curves, tire pressure or the vehicle's deterioration status directly influences the level of fuel consumption.

Table 10 shows the probability distribution of the top-5 IFs for all the AFC subclasses. For example, for the ACF_1 subclass it is important for the IF-1 (Clutch) to be in subclass one, so drivers should avoid the use of unnecessary clutching, thus enjoying more inertia (IF-4), the percentage of the engine use to idle (IF-3) and less time in the range "Yellow" rotations (IF-2).

VIII. CONCLUSION

We are aware that road conditions, the number of passengers on the bus, and especially traffic can influence fuel consumption. The drivers' state of mind could also be another important parameter to be considered in this process, but we did not consider it in this research.

Implementing a data warehouse was fundamental to help business experts better understand, decide and act upon the daily running of a bus fleet. As a by-product of the ETL data warehouse activities, the application of data mining models to describe data and detect IFs on fuel efficiency allows for a better insight from the collected data.

The acquisition of knowledge in this study was driver-centric. That reinforces the importance of the driver's continuous education and motivation to long-term fuel efficiency savings, contributing to better adoption and retention of eco-driving behaviors.

Weather conditions influence fuel consumption in two ways. First, on hot days, driver behavior depends on the use of the air conditioning that subsequently shows an increase of engine idle events (RI). Secondly, on days of heavy rain, the traffic increases and so, there is a correlation of the traffic and heavy rain events. On the other hand, Lisbon's good weather does not allow a significant impact on fuel consumption in what concerns weather conditions.

The use of data collected from a fleet of buses can be applied to other scenarios and others types of vehicles, like privately owned cars, provided similar datasets and similar data analysis approaches are used.

The predictive accuracy and capability of the models built is still to be optimized and compared with other algorithms, to be applied to the same mining structures, to improve prediction accuracy, such as the classification and regression trees, KNN Clustering and Neural Networks, in order to overcome the known frailties of linear classifiers.

Overall, our findings show that adopting appropriate driving styles can reduce fuel consumption on an average between 3 to 5 liters per 100 km. This can save 15 to 30 liters per bus in just a working day. Taking into account fuel prices, these savings represent 20€ to 40€ a day, per bus. Considering the working days in a year and with around 1,500 involved drivers, this may impact significant savings that can go up to 1.5M€ per year.

Another important output of the current study is the possibility to collect real-time data. Based on the events collected in real-time and available in a central database, it is

possible to apply our methodology (based on the discussed processes) to dynamically analyze and identify major driving parameters that have an impact on the reduction of fuel consumption. However, that approach should be slightly different because it should allow the monitoring of a huge set of driving parameters and simultaneously identify the most influential ones.

Acknowledgements. This work was partially supported by national funds through FCT (Fundação para a Ciência e Tecnologia) with references PEst-UID/CEC/00319/2013, UID/CEC/50021/2013 and EXCL/EEI- ESS/0257/2012 (DataStorm).

REFERENCES

[1] M. Chan, A. Herrera and B. Andre´, "Detection of changes in driving behavior using unsupervised learning," *IEEE International Conference on Humans, Information and Technology*, Vol. 2, pp. 1979–1982, 1994.

[2] U. Reiter, "Modeling the driving behavior influenced by information technologies," *In Highway Capacity and Level of Service (Ed. Brannolte)*, pp. 309–320, 1991.

[3] M. Natale, H. Zeng, P. Giusto and A. Ghosal, "Understanding and Using the Controller Area Network Communication Protocol: Theory and Practice," *Springer New York*, ISBN 978-1-4614-0313-5, January 2012.

[4] J. Gray, A. Bosworth, A. Layman and H. Priahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals". *Proc. 12th International Conference on Data Engineering. IEEE*. pp. 152–159, 1995.

[5] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," *AAAI/MIT Press*, 1996.

[6] G. Mariscal, O. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowledge Engineering Review* 25.2 (2010): 137.

[7] J. Almeida and J. Ferreira, "BUS Public Transportation System Fuel Efficiency Patterns," in *proceedings of the 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013)*, 2013, Malaysia.

[8] E. F. Codd, "Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate," *E. F. Codd and Associates*, 1993.

[9] M. Ishibashi, M. Okuwa, S. Doi and M. Akamatsu, "Indices for Characterizing Driving Style and their Relevance to Car Following Behavior," *SICE Annual Conf.*, pp. 1132-1137, 2007.

[10] O. Taubman-Ben-Ari, M. Mikulincer and O. Gillath, "The multidimensional driving style inventory-scale construct and validation," *Accident Analysis and Prevention*, Vol. 36, pp. 323-332, 2004.

[11] H. Hiromitsu, Y. Nakajima and T. Ishida, "Agent Modeling with Individual Human Behaviors," *Proc. of 8th Int'l. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pp. 1369-1470, 2009.

[12] A. Augustynowicz, "Preliminary Classification of Driving Style with Objective Rank method," *International Journal of Automotive Technology*, Vol. 10, No. 5, pp. 607-610, 2009.

[13] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya and M.C. González, "Safe Driving Using Mobile Phones," Intelligent Transportation Systems, IEEE Transactions on , vol.13, no.3, pp.1462,1468, Sept. 2012.

[14] M. Rigolli and M. Brady, "Towards a Behavioral Traffic Monitoring System," *International Conference on Autonomous Agents, Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 449-454, 2005.

[15] M. Chan, A. Herrera and B. Andre, "Detection of changes in driving behavior using unsupervised learning," *IEEE International Conference on Humans, Information and Technology*, Vol. 2, pp. 1979–1982, 1994.

[16] IEA, "Transport Energy Efficiency Implementation of IEA Recommendations, since 2009 and next steps". Available at http://www.iea.org/publications/freepublications/publication/transport_energy_efficiency.pdf, 2010.

[17] C. C. Rolim, P. C. Baptista, G. O. Duarte and T. L. Farias, "Impacts of On-board Devices and Training on Light Duty Vehicle Driving Behavior," *Procedia-Social and Behavioral Sciences*, 111, 711-720, 2014.

[18] M. Barth, K. Boriboonsomsin, "Energy and emissions impacts of a freeway-based dynamic eco-driving system," *Original Research Article Transportation Research Part D: Transport and Environment*, Volume 14, Issue 6, pp 400-410, August 2009.

[19] B. Beusen, S. Broekx, T. Denys, C. Beckx, B. Degraeuwe, M. Gijsbers, K. Scheepers, L. Govaerts, R. Torfs, L. I. Panis, "Using on-board logging devices to study the longer-term impact of an eco-driving course," *Transportation Research Part D: Transport and Environment*, Volume 14, Issue 7, pp 514-520, October 2009.

[20] M. Zarkadoula, G. Zoidis and E. Tritopoulou, "Training urban bus drivers to promote smart driving: A note on a Greek eco-driving pilot program," *Transportation Research Part D: Transport and Environment*, Volume 12, Issue 6, pp 449-451, August 2007.

[21] K. Boriboonsomsin and M. Barth, "Eco-driving: Pilot evaluation of driving behavior changes among U.S. drivers," *University of California, Riverside*, 2010.

[22] A.E.Wahlberg, "Long-term effects of training in economical driving: fuel consumption, accidents, driver acceleration behavior and technical feedback," *International Journal of Industrial Ergonomics* 37, pp 333–343, 2007.

[23] EcoMove Project web site - www.ecomoveproject.eu

[24] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher. "Greengps: a participatory sensing fuel-efficient maps application," in *Proceedings of the 8th international conference on Mobile systems, applications, and services, MobiSys '10*, pp 151–164, New York, NY, USA, ACM, 2010.

[25] Treatise, "Ecodriving - the smart driving style," *Utrecht for the EC TREATISE project*, September 2005.

[26] K. Jakobsen, S. Mouritsen, and T. Kristian, "Evaluating eco-driving advice using GPS/CANBus data," *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013.

[27] J. MacLennan, B. Crivat, Z. Tang, "Data mining with Microsoft SQL server 2008-2009", ISBN 978-0-470-27774-4

[28] S. Sumathi and S.N. Sivanandam, "Introduction to Data Mining and Its Applications", Springer, ISBN 9783540343509, 2006.

João C. Ferreira is Professor at the Polytechnic Institute of Lisbon (IPL/ISEL) and Consultant with different companies and institutions. He graduated in Physics at the Technical University of Lisbon (UTL/IST), Portugal, received an MSC in Telecommunication and a PhD degree in Computer Science Engineering from UTL/IST. His professional and research interests are in retrieval, geographic and multimedia retrieval, Electric Vehicle, Intelligent Systems, intelligent transportation (ITS) and sustainable mobility systems. He is the author of over 130 scientific papers of international conferences and workshops in different areas of computer science.

José de Almeida was a student of the Master Degree in Informatics Engineering at the Instituto Superior de Engenharia de Lisboa (IPL/ISEL).

Alberto Rodrigues da Silva is Associate Professor at the Instituto Superior Técnico (Universidade de Lisboa), senior researcher at INESC-ID Lisboa, and partner of the SIQuant company. He has research interests in the following areas: Information Systems, Modeling and Metamodeling, Model Driven Engineering, and Requirement Engineering. He is the author of 5 technical books and over 200 peer-reviewed scientific documents. He is a member of the ACM, PMI and the Portuguese Society of Chartered Engineers.