TU Delft

Delft University of Technology

On Developing a Driver Identification Methodology Using In-Vehicle Data Recorders

Moreira-Matias, Luis; Farah, Haneen

DOI
10.1109/TITS.2016.2639361

Publication date
2017

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Intelligent Transportation Systems

Citation (APA)
Moreira-Matias, L., & Farah, H. (2017). On Developing a Driver Identification Methodology Using In-Vehicle Data Recorders. *IEEE Transactions on Intelligent Transportation Systems*, *18*(9), 2387-2396. Article 7819486. https://doi.org/10.1109/TITS.2016.2639361

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

On Developing a Driver Identification Methodology Using In-Vehicle Data Recorders

Haneen Farah, Luís Moreira-Matias[1]

***Abstract*— Recently, multiple cutting edge technologies to facilitate data collection processes have emerged. One of the most prominent ones is the In-Vehicle Data Recorder (IVDR). Various identification technologies were employed to relate the IVDR's data to multiple drivers sharing the same vehicle. Irrespective to the level of sophistication, all of these technologies still have considerable limitations on identifying drivers' identity.**

The purpose of this study is to propose a methodology which can identify the driver of a given trip using historical trip-based data. To do so, an off-the-shelf Machine Learning (ML) framework is proposed. The main goal is to take advantage of inexpensive data – such as driver-labelled trip data - to build a pattern-based algorithm able to identify the trip's driver category when its true identity is unknown. The proposed framework includes feature evaluation and category identification. Our ultimate goal is to provide an inexpensive alternative to existing IVDR technologies which can serve as their complement and/or validation purposes.

Experiments conducted using four different types of induction learners over a real-world case study from Israel uncover the potential of this idea: decision trees obtained a promising range of accuracies on this task (i.e. 75% to 100%).

***Index Terms*— Identification methods, in-vehicle data recorders, data entropy, feature selection, classification, supervised learning.**

I. Introduction

In the last decade, significant advances have been made in measuring and communication technologies. Such advances led to a considerable growth in the development and use of Intelligent Transportation Systems. One of the widely used technologies regarding observing driver behavior is In-Vehicle Data Recorder (IVDR). IVDR is a system able to measure vehicle's movement, driver control, and vehicle's performance. Early usage of these systems was targeted towards fuel efficiency and vehicle location tracking purposes. Recently, it has been also proposed for driver behavior monitoring and traffic safety purposes [1]. IVDR can record detailed information on driving performance and thus assist in developing intelligent systems adapted to each driver's unique driving characteristics. Several researchers applied classification and identification methods and developed algorithms on driver behavioral characteristics to detect abnormal driving behaviors for automotive control applications [2-4].

In many cases, the same vehicle is shared by multiple drivers (two or more). Thus, one of the challenges that researchers face when using IVDR devices in shared vehicles is the **driver's identification**. Several identification technologies exist. However, the use of these technologies does not solve this issue adequately (i.e. expensive and/or inaccurate). Consequently, it is critical to develop affordable methodologies able to deal with such information loss.

The availability of Global Positioning System (GPS) data faced an explosive grow. This data is available everywhere and widely used among transportation industry. Recently, Wallace *et al.* [5] used GPS and OBDII logs (on-board diagnostics) from a preliminary sample of 100 trips and 4 drivers to test the potential of time of day, road choice, velocity and acceleration data to provide attributes to distinguish between drivers of a shared vehicle. Hence, the sample size is reduced to generalize significant conclusions on this topic. At the best of our knowledge, this is the only research work proposed using this approach.

The main purpose of this study is to develop a methodology which can identify the driver for a given trip of interest using historical trip-based data. This data is not more than a high level aggregation of Floating Car Data (FCD) collected through a user-identified device, such as a registered GPS antenna and/or smartphone. Per opposition to most of existing IVDR and/or computer vision techniques to perform driver identification, we aim to leverage in simple things such as the **daily seasonality** inherent to the human behavioral routines (e.g. wake up, go to school and get back home for lunch). Moreover, the collection of this type of data is easier due to the amount of devices that already exist in our surroundings with capabilities of storing and/or broadcasting this type of data. Such availability makes the information about the driver's identification easier and cheaper to get than for any other data source (e.g. video cameras) or abovementioned IVDR technologies. Throughout this simple idea, we intend to boost the existing technologies with a knowledge discovery framework.

To do so, an off-the-shelf Machine Learning (ML) framework is proposed. The main goal is to take advantage of the driver-labelled trip data to build a pattern-based algorithm able to identify the trip's driver where its true identity is unknown. Data collected from a particular case study from Israel ([6]) is used to validate the applicability of the proposed methodology to this task. The contributions of this study are twofold: (1) the suggestion and exploitation of supervised learning approach over trip-based data (easier to collect and to process) to serve as complement to existing IVDR technologies through an exhaustive comparison of different

[1] Haneen Farah, Assistant Professor at the Department of Transport and Planning, Delft University of Technology, e-mail: h.farah@tudelft.nl

Luís Moreira-Matias, Research Scientist at NEC Laboratories Europe (Heidelberg, Germany), e-mail: luis.matias@neclab.eu

types of induction learners; (2) a simple State-of-the-Art entropy-based framework to describe the explanatory power of multiple features regarding the driver identification from trip-based data.

This paper is structured as follows: next section presents a comprehensive literature review about the topic. Section III presents the research methodology and discusses ML techniques including feature evaluation and category identification. This is followed by a description of the real-world case study used to evaluate the applicability of this method. Section IV presents the results including: (1) the commonalities between the trip's data on identified and unidentified trips; (2) the importance of each data feature and (3) category identification accuracy. Section V introduces a brief discussion on the obtained results, followed by section VI which concludes the paper.

II. Literature Review

Earlier research have used IVDR as a measurement tool to observe drivers' naturalistic driving behavior, such as the "100 cars naturalistic study" [7, 8], DriveAtlanta [9], and PROLOGUE [10]. Later, this tool was also used for intervention purposes; it supported reducing risky behaviors by providing feedback to drivers or to those who are responsible for their driving (e.g. parents, fleet managers), [11-13].

IVDRs are widely applied. Yet, one of their drawbacks is the driver identification. For example, in Farah et al. [6, 13] all members of a participating family were requested to identify themselves at the beginning of each trip using Dallas keys (personal magnetic identification keys). However, when analyzing the trips' dataset it was found that 22% of these trips were unidentified.

The following sub-section provides an overview of the main identification technologies, as well as a comparison of their related weaknesses and strengths.

A. Identification Technologies

There are several identification technologies existing today that can be classified into two main categories: Physical systems and Sensing systems.

Physical Systems:

- **Dallas keys/iButtons** are personal magnetic identification keys (chip-based data carrier) which were used by [10] and [6, 13] for driver identification.
- **iRFID (Radio Frequency IDentification)** technology is based on the use of radio waves to read and capture information stored on tags attached to persons, vehicles and other objects [14]. It has a similar concept to a barcode. However, unlike the barcode, the tag does not necessarily need to be in direct line-of-sight of the reader. This technology enables remote and automated data gathering using a wireless communication [15].

Sensing Systems:

- **In-vehicle video cameras** [16]. The most sophisticated technologies on this research line are the Apple iPhoto and the Google Picasa [17, 18], which are based on face detection and identification. In the context of driving, this technology takes a single snapshot of the driver's face at the trip's start to identify him/her.
- **Biometric fingerprint systems** [19]. In the context of driving, upon vehicle startup, drivers need to verify their identity by a pre-authorized fingerprint.
- **Voice recognition and iris technology** [20-22] can be also used for driver's identification. These are two highly unique features in the human body (even identical twins present differences with respect to these features). The voice recognition is done through an in-vehicle microphone combined with a biometric speech identification software. Iris technology relies on two basic types of eye scans: iris scanning and retinal scanning.

B. Shortcomings of existing identification technologies

Each of the previously mentioned technology has its own strengths and weaknesses. Table I summarizes those points for each of them.

Reference [23] considered several driver identification methods including key fobs or entry codes [24, 25]. However, these still require driver activation. Other options which were considered included the use of wearable devices [26] or applications downloaded onto mobile phones [27]. Nevertheless, these devices are not convenient as drivers will need to carry them on personally.

ML and Computer Vision approaches [28, 29] have also been explored to this end. This type of methods aims to somehow identify the driver based on the images captured by a high-resolution video camera. Indeed, the richness of the features provided by multiple images per second represents a *perfect gem* to be explored by many of the cutting edge classification algorithms (e.g. Convolution Neural Networks [30]). Yet, they require a considerable large amount of samples to provide a reasonable output. On the other hand, more traditional approaches such as Support Vector Machines (SVM) require a complex and expensive future engineering process, which is not affordable in many cases, e.g. [32]. Contrary to video cameras, GPS antennas are widely spread among the automotive industry. Consequently, the accessibility of such individual probe car traces is higher than those camera footages – and therefore, to be explored by a knowledge discovery framework.

C. The main contributions of this paper

The analysis performed throughout the previous section uncovered that the state-of-the-art technologies for driver's identification still have multiple limitations to overcome. They are related with the relationship between their identification accuracy and their cost – which are still far unbalanced in

most of the cases. The main motivation for this paper is to provide a way of generalizing driver's behavior regarding their mobility seasonality. We propose to leverage on inexpensive FCD acquired from each individual to understand who is driving through a ML pipeline. Such data driven methodology is introduced throughout the next section.

TABLE I
COMPARATIVE TABLE BETWEEN EXISTING IDENTIFICATION (IVDR) TECHNOLOGIES

Technology	Strengths	Weaknesses
	Physical Systems	
Dallas keys/ iButtons	- Relatively have a low cost; - Easy to implement [17];	- Requires drivers' activation; - Can be transferable among drivers, i.e. a driver can use another driver's Dallas key [1];
iRFID (Radio Frequency IDentification)	- Relatively have a low cost; - Easy to implement;	- Requires to attach the tag directly to the driver and not to the vehicle; - Drivers need to remember to wear it with them when driving;
	Sensing Systems	
In-vehicle video cameras	- It does not require drivers' activation;	- Costs more than the physical systems; - Camera's lens can be highly sensitive to illumination conditions and driver orientation;
Biometric fingerprint systems	- Easy to use; - Have mono identification (i.e. unique for every person). - Relatively more accurate and reliable; - Moderately cheap;	- Highly intrusive [31];
Voice recognition	- Quite accurate;	- Can be easily bypassed by using a pre-recorded voice of another driver; - Relatively expensive [1];
Iris technology	- Quite accurate; - Not that easily bypassed;	- Drivers may need to remove eyewear; - Scans may not work with people with cataract or glaucoma [1]; - Relatively expensive [1];

III. METHODOLOGY

The regularities of the human behavior have been providing important advances in many transportation-based research topics. Some successful examples on applying such insights are the passenger demand prediction problem [33] or the bus schedule planning [34]. In this particular application, the authors intend to take advantage on such type of seasonal patterns and trends to address the driver identification failures presented by most of the state of the art methodologies.

Although being mostly *invisible* to human eye, such patterns are contained by the trip-based data. Consequently, data driven methodologies can solve such information loss in an effective and inexpensive way.

Let $T = \{t_1, t_2, \ldots, t_n\}$ define a series of n trips t_i (which can also be denoted as *data samples*). Each trip t_i can be expressed as a pair $t_i = (X_i, Y_i)$ where $X_i = \{x_{1,i}, x_{2,i}, \ldots, x_{a,i}\}$ stands for a set of $'a'$ variables which describe a given trip and $Y_i \in L$ denotes the driver category form a set of $c = |L|$ possible categories (which correspond to driver's category in each family from the present context). From now on, $X = \{x_1, x_2, \ldots, x_a\}$ is denoted as the *Features* - which have values X_i for each sample i - while the driver category Y_i will be denominated as the *Target Value*.

Theoretically, there is a function ω capable of determining the Driver Category (i.e. Target Value) based on the feature values X_i. It can be expressed as $\omega(X_i) = Y_i$. However, the scope of this study is on exploring the different driving behaviors on each family. Therefore, a *Statistical Independence* is assumed to be in place between the trips of each family. Consequently, the previous equation can be extended as

$$\omega(X_i, F_i) = Y_i \quad (1)$$

where F_i stands for the family of the trip i. To introduce such independence, the original dataset T was split into f datasets as $T = \cup_{j=1}^{f} T_j$ where f stands for the total number of families included in the study and T_j contains only the trips performed by someone from family j.

Hereby, a series of methods is proposed to extract Driver Information from a real world trip dataset T. Then, ML techniques are proposed to automatically explore the acquired data in order not only to 1) approximate the real function ω as much as possible, but also 2) to explain which are the relationships expressed between the feature values X_i and the target values Y_i and also within the feature set F. Concretely, these problems can be divided on the following two tasks: (i) Features Evaluation - to determine which are the most informative features to determine the trip's driver in each family dataset; (ii) Category identification - to infer a function to determine the trip's driver in each family dataset by employing Supervised ML techniques over the dataset.

The remaining of this section is structured as follows:

firstly, the learning techniques for the above mentioned tasks are described in subsections *A* and *B.* Finally, the case study is detailed in the subsection *C.*

A. Features evaluation

Empirically, it is possible to conclude that the variable's importance to determine the driver category on each family depends on the *degree of randomness* of the target variable (i.e. the trip's driver) on each dataset of trips for a given family i, i.e. T_i. In Information Theory, such quantity is also known as **Entropy**. This can easily be expressed by the following example: in family A, you have always the same driver...while in family B, you have always different drivers on each trip. While it is easy to predict the driver for the trips by family A - as it is always the same driver (i.e. *low* entropy), it may not be that easy for family B (i.e. *high* entropy).

Information Gain (IG) is a metric commonly used in many Information Theory problems. One of its most popular applications is on selecting the attribute on which a split criteria must be set regarding the building of a decision node on Decision Trees (i.e. see, for instance, C4.5 algorithm in [35]). Hereby, IG was used to evaluate the feature relevance on inferring the trip's driver on each family.

B. Category Identification

Departing from the features discovered to be *relevant* to infer the trip's driver on each family using the abovementioned Information Gain, the dataset need to be used in order to extract the **dependencies** in place between the Y and X. By doing so, it would be possible to predict which will be the trip's driver on future trips - where the driver is unidentified. The process of assigning output values given a set of input ones based on a given training dataset is known as **Supervised Learning**. As Y is a categorical variable (i.e. the category of the driver inside a given family), this comprises a **Classification** problem. Formally, the target is to **Generalize** the behavior and the characteristics of a Population (i.e. family) using only some data samples extracted from it. The aim is to build a function $\overline{\omega}(X, F) \sim \omega(X, F)$ based on the labelled trips dataset of each family. This function is named as the *predictive model*. Obviously, the generalization comprises a given *error* ϵ - since each dataset represents only a part of the entire population. Based on ϵ, it is possible to express the previous relationship as follows:

$$\omega(X, F) = \overline{\omega}(X, F) + \epsilon \quad (2)$$

The task behind a classification problem is to select a **learning method** able to approximate $\omega(X, F)$ by minimizing the value of ϵ. Hereby, we focus on four types of algorithms: 1) Decision Trees using the **C4.5 algorithm** [36] and **Logistic Model Trees (LMT)** [37], 2) Bayesian Learning using **Naive Bayes (NB)** [38], 3) Kernel-based using *Support Vector Machines (SVM)* [39] and 4) Instance Based using **k-Nearest Neighbors (kNN)** [40].

In this paper, it is proposed to explore these five state-of-the-art algorithms on this task. By doing so, we expect to provide insights on which is the best one to handle these type of problems and/or datasets. The different learning algorithms are evaluated by their **accuracy** using the **10-fold cross validation** method on each family dataset.

To evaluate the applicability of the above mentioned method, a comprehensive data set on both identified and unidentified trips was collected from a real world case study. This study is described in the following section.

C. Case Study: "The First Year Study"

In the *"The First Year Study"* the IVDR system that was used [6, 13] was the GreenRoad technology. It is a g-force based system which tracks all trips made by the vehicle and records the following information:

- Trip start and end times;
- Driver identification using Dallas keys;
- Vehicle location;
- Events of excessive manoeuvres defined by patterns of g-forces measured in the vehicle.

Data from the IVDR was collected throughout one year.

Participants & recruitment process

A rolling recruitment procedure was carried out between July 2009 and November 2010. The data collection process was already in place throughout this period. 242 families started participation in the experiment. However, just 217 completed the one year period. This represents a drop rate of 10.3%. Participants were informed at the beginning of the study about a monetary compensation for their participation of 1000 NIS (approximately $250) which they received at the end of the study. This was used as an incentive for participation.

The recruitment's process, the characteristics of the drivers and their families is described in detail in [6, 13]. Implications derived from the participation of the driver's parents in this study is also reported in the abovementioned studies.

Data collection

The data collected in the experiment in terms of number of trips and driving hours by each family member, and for the unidentified trips are described in Table II.

From analysing Table II, it is possible to conclude that roughly 22% of the trips are unidentified, while 21% of the total driving time also follows the same pattern. Such ratio constitutes a significant portion of the total data which should not be discarded at any case to carry out any data driven analysis, regardless its goal. When further analysing this ratio per family it was found that many families present highly unbalanced identification ratios. In particular, ~62% of the families had an identification ratio above 0.8, ~25% had an identification ratio between 0.6-0.8. However, there are some families, ~8%, that had relatively similar number of identified and unidentified trips (between 0.4-0.6), and others ~5% that had an identification ratio between 0-0.4.

The set of variables which describe a given trip $X_i = \{x_{1,i}, x_{2,i}, \ldots, x_{a,i}\}$ are detailed in Table III. Four types of trips

TABLE II
NUMBER OF TRIPS AND DRIVING HOURS BY CATEGORY OF DRIVER

Category	No. of Trips	Driving Hours (h)
Young driver	108191	34074
Father	78963	31325
Mother	102120	33701
Other family members	20070	7187
Unidentified	87181	28489
Total	396525	134776

were defined: HH (home to home), which are trips that start and end in the area around home; HO (home to other), which are trips that start at home area toward a more distant location; OH (other to home), which are trips that start from a distant location toward the home area, and OO (other to other), are trips that start and end from locations distant from the home area. In this study, a home area is defined as the length of the radius of a circular area around the location of its exact address. This definition has an impact on the trip's classification as (HH, HO, OH, and OO). An algorithm was developed in [41] to define a specific radius of a circular area around the home location of each participating driver. The average radius across families in the dataset was 1034 meters (SD= 346). Further details regarding this algorithm can be found in [41].

TABLE III
SET OF VARIABLES THAT DESCRIBE A GIVEN TRIP

Variable	Type	Domain
Weekday	Categorical	{SUN, MON, …, SAT}
Departure time	Categorical	{00-3, 3-6, 6-9, 9-12, 12-15, 15-18, 18-21, 21-00}
Trip duration (min)	Categorical	{0-5, 5-15, 15-30, 30-60, 60-120, 120-180, >180}
Trip aggressiveness level	Categorical	{Moderate, Intermediate, High}
Solo or accompanied	Categorical	{Solo, Accompanied}
Number of events (IVDR)	Ordinal	{count number}
Cluster ID of trip origin	Nominal	{1,2,3,…,263}
Cluster ID of trip destination	Nominal	{1,2,3,…,263}
Cluster ID of home	Nominal	{1,2,3,…,80}
Trip type	Categorical	HH (home to home); HO (home to other) OH (other to home); OO (other to other)
Previous category	Categorical	{father, mother, young driver, other}

IV. RESULTS

This section presents the experimental results. First, the results of the comparison between the identified and unidentified trips are introduced followed by the results of the feature evaluation and category identification.

A. Comparison between identified and unidentified trips

There is no information on the true driver-based label of the trips contained by the unidentified dataset. Such unawareness is a limitation to apply any learning method since the performance of any supervised learning method depends on the dataset on which is being applied to. Therefore, a training set with identified trips is required. One of the most important constrains when it comes to using a training set is to guarantee that the samples on both the training and the test sets belong to the very same population [42].

In this particular problem, it is important to show that the identified trips have similar distribution to the unidentified trips, in terms of the variables that describe a trip. Therefore, the probability distributions of the identified trips and the unidentified trips were compared using Kolmogorov-Smirnov (K-S) test. The results are depicted in Table IV. It illustrates that the distributions of the identified trips and the unidentified trips for all the tested features do not differ significantly at the 95% confidence level, except for the Cluster ID of the trip origin and trip destination.

TABLE IV
KOLMOGOROV-SMIRNOV TEST RESULTS

Variable of Interest	Sig.
Weekday	0.938
Departure time	0.964
Trip duration (min)	0.998
Trip aggressiveness level	0.996
Solo or accompanied	0.964
Number of events (IVDR)	0.931
Cluster ID of trip origin	<0.001
Cluster ID of trip destination	<0.001
Trip type	0.508
Previous category	0.560

The significantly different distributions in the trip origin and destination, between the identified and unidentified trips, can be explained from the large domain range of those variables (263 categories, see Table III). Consequently, such large range coupled with an unbalanced ratio between the identified (78%) and unidentified (22%) trips could point this result as an expected one. However, we argue that since there was no significant difference in the probability distributions of the identified and unidentified trips in all the remaining features, it can be assumed that the two samples do not differ significantly. In other words, the training set can be used to train pattern-based models to predict the driver category of this test data set. Following this conclusion, the next section presents the results of the feature relevance analysis.

B. Experimental Setup

All the experiments and analysis were conducted using the R Software [43]. The information gain was computed using the functions within the R package [infotheo]. The classification algorithms employed were the 1) C4.5, 2) LMT, 3) NB, 4) SVM and 5) kNN using the implementations supplied by the R packages (1,2,3) [RWeka] [44], 4) [e1071] [45] and 5) [class] [46], respectively. The parameter setting for these algorithms followed their default values. The two mandatory parameters were the number of neighbors *k* in the kNN (i.e. empirically set to 2) and the kernel employed on the SVM algorithm (i.e. empirically defined to be linear).

The tenfold cross validation process was manually implemented (without using a pre-defined coding package) by dividing each family trip dataset into ten folds. Each fold contains roughly 10% of the total number of trip samples available for each family. The sample selection to compose each individual followed is completely random. By doing so, we end up training 10 different models to be evaluated in 10 different test sets. The main goal in employing such process is to assess the true generalization error achieved by each individual method.

C. Feature Evaluation

This section presents an analysis of variables' importance to determine the driver category on all families. Table V illustrate descriptive statistics of the Average Information Gain by each feature (i.e. AIG). Based on it, it is possible to conclude that the most informative trip's feature is the **Previous Category (PC)** (i.e. AIG=0.49). It corresponds to the driver's category {father, mother, young driver, other} in the previous trip. In other words, if for example the driver in the previous trip was the father, it is most likely that the driver in the current subsequent trip is also the father. The second two most informative features are the **trip's origin/destination** (both AIG=0.18), followed closely by the trip's **departure time** (AIG=0.16).

One of the most well-known methods to estimate the probability distribution of a given Population whenever it is *unknown* is to compute sample-based Probability Density estimations. Such estimations provide fair approximations of the Probability Density Function (*p.d.f.*) of a given random variable. In this particular study, the sample-based *p.d.f.* can be used to approximate the probability of the IG of a certain feature to fall into a given range of values by calculating the area under such *p.d.f.* (i.e. the integral within such range).

Fig. 1 presents a sample-based *p.d.f.* for the four most important features - which was computed using the Kernel Density Estimation (KDE) [47]. To do it so, the well-known Gaussian kernel was employed along with the Silverman's rule of thumb [48] to determine the correspondent bandwidth. Fig. 1 clearly illustrates that PC is the most informative feature to determine the category of a trip. To demonstrate such a concept, we can infer the validity of the following equation:

$$p(IG(PLC) > 0.5) \geq [p(IG(OrigClsID) > 0.5) + p(IG(DestClsID) > 0.5) + p(IG(TOD) > 0.5)] \quad (3)$$

by just performing an empirical analysis on the Figure's insights. Such relationship clearly uncovers the informative power of PLC regarding the remaining features.

D. Driver Identification

In this study, the abovementioned five state-of-the-art algorithms were explored to carry out the identification task. By doing this, we expect to provide insights on which is the most adequate algorithm to handle these type of problems and/or datasets – while, at the same time, we want to uncover the power of the relevant information hidden on this type of data relevant for this problem.

Table VI presents the average weighted accuracy for all families for each one of the methods. Fig. 2 presents the corresponding confusion matrices. These results illustrate that the C4.5 presents the best (averaged) accuracy on this task – followed closely by LMT.

To give a better view of the performance of these methods, the accuracy's p.d.f. of each one of the classification methods employed was computed and illustrated in Fig. 3.

TABLE V
DESCRIPTIVE STATISTICS OF THE AVERAGED INFORMATION GAIN (AIG) OF EACH FEATURE

	Information Gain (IG)					
Variable	Min	Max	Mean	Std.	1st qrt.	3rd qrt.
Weekday	0.00	0.39	0.07	0.06	0.02	0.09
Departure time	0.00	0.62	**0.16**	0.11	0.07	0.24
Trip duration (min)	0.00	0.19	0.04	0.04	0.02	0.06
Trip aggressiveness level	0.00	0.24	0.04	0.05	0.01	0.04
Solo or accompanied	0.00	0.59	0.03	0.06	0.00	0.03
Number of events (IVDR)	0.00	0.24	0.04	0.05	0.01	0.05
Cluster ID of trip origin	0.00	0.49	**0.18**	0.11	0.10	0.26
Cluster ID of trip destination	0.00	0.48	**0.18**	0.11	0.10	0.26
Cluster ID of home	0.00	0.13	0.00	0.02	0.00	0.00
Trip type	0.00	0.18	0.03	0.03	0.01	0.04
Previous category	0.00	1.22	**0.49**	0.23	0.33	0.63

Based on Table VI, it is possible to confirm that the LMT and the C4.5 are the methods with the highest performance. The accuracy of their outputs range between 0.75 and 1.00. The LMT accuracy's p.d.f. (presented in Fig. 3) is quite similar to the one computed on the C4.5 outputs (i.e. both are right-shifted). On the other hand, the NB and the SVM p.d.f. have lower peaks and minor areas regarding the high-accuracy space (i.e. > 0.75).

The p.d.f. of the accuracies by the NB method and the SVM method have lower peaks and are less shifted to the right compared to the LMT and C4.5 methods. The method with

worst performance was kNN, which its accuracy's values are mostly concentrated within 0.65-0.75.

The high peaks close to zero accuracy in Fig. 3 result from the fact that many times, the learning functions are unable to fit an objective function $\overline{\omega}(X, F)$ to their training data. It happens because some families had relatively few number of trips from which is not possible to *generalize* the population's behavior.

Nevertheless from the results exhibited in Fig. 3, it is not possible to conclude which is the most adequate method to use for every situations on this particular task. Even if Fig. 3 suggests

TABLE VI
AVERAGE WEIGHTED ACCURACY PER FAMILY

Method	Average Weighted Accuracy
Naïve Bayes (NB)	0.671
Logistic Model Tree (LMT)	**0.716**
k-Nearest Neighbors (kNN)	0.580
Decision Tree (C4.5 algorithm)	**0.717**
Support Vector Machines (SVM)	0.651

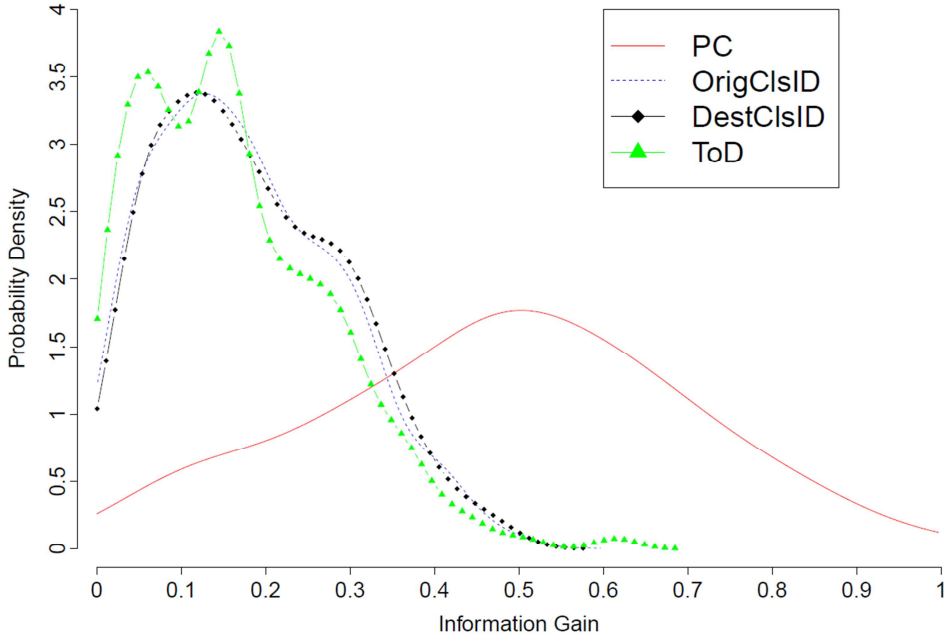


Fig. 1. Probability density functions of the four most informative features to determine the driver category.

C4.5

	F	M	Y	O
F	0.2925	0.0175	0.0272	0.0051
M	0.0159	0.2343	0.0104	0.0028
Y	0.0212	0.0101	0.3090	0.0020
O	0.0083	0.0048	0.0035	0.0353

kNN

	F	M	Y	O
F	0.2052	0.0498	0.0704	0.0170
M	0.0495	0.1751	0.0301	0.0087
Y	0.0673	0.0277	0.2394	0.0078
O	0.0175	0.0092	0.0085	0.0167

LMT

	F	M	Y	O
F	0.2923	0.0173	0.0275	0.0052
M	0.0159	0.2341	0.0106	0.0029
Y	0.0214	0.0102	0.3085	0.0022
O	0.0080	0.0048	0.0036	0.0355

NB

	F	M	Y	O
F	0.2576	0.0312	0.0443	0.0092
M	0.0295	0.1979	0.0294	0.0067
Y	0.0374	0.0206	0.2798	0.0045
O	0.0116	0.0070	0.0083	0.0251

SVM

	F	M	Y	O
F	0.2543	0.0370	0.0480	0.0029
M	0.0270	0.2142	0.0202	0.0021
Y	0.0354	0.0215	0.2827	0.0027
O	0.0116	0.0095	0.0081	0.0227

Fig. 2. Confusion Matrices of each method using relative frequencies. Legend: F- Father, M-Mother, Y-Young Driver, O-others.

that $p(\mathrm{ACC}_{\mathrm{C4.5}} > \mathrm{ACC}_{\mathrm{LMT}}) > 0.5$ (by analyzing the areas below the lines), it is not possible to compute the significance level of such hypothesis using the current results.

V. DISCUSSION

The informative feature set uncovered by the IG is not surprising. They suggest that the families behavior regarding the use of their vehicle follow some **seasonalities** very close to the regularities of the human behavior. Such regularities are usually based on some commuter trips where the pair origin/destination is the same on a daily basis given some specific time of the day [49]. Some examples could be commuter trips from home to a work place, a specific trip to school or grocery performed by a given family member. In many cases, it is reasonable to assume that such feature values (i.e. departure time and trip's origin and destination) can be known even before the trip starts – which highlight the possibility of performing an **apriori driver identification**.

This insight is key to illustrate the contribution of our predictive model: more than completing the missing identification data produced by other identification methods based on the driver behavior (such as the face recognition ones), it can **improve their accuracy**. The idea is to reduce uncertainty around the outputs of behavior-based

identification methods by providing an **apriori** likelihood of who might be driving the car even before its engine starts. However, the validity of such hypothesis requires a real world validation to be claimed as so. The apriori driver identification is problematic when driving behavioral characteristics (speed, acceleration) are combined with trip-based information to identify the driver category of a shared vehicle as done recently by [5]. However, such information of driver behavior can be used after a trip is initiated to enhance the predictability power. For example, a driver is first identified based on the trip information as done in this study, and after the trip is initiated the identification is re-examined and validated based on the added attributes of the unique driving behavior characteristics of the specific driver.

The PC (i.e. previous category) is, by far, the most informative feature on the present dataset – which also follows the abovementioned regularities of the families' behavior. However, it also uncovers strong dependencies between trips regarding its driver. Such dependencies between sequences of trips are common in many transportation problems (see, for instance, the Bus Bunching issues in [50]. A good insight on its impact is depicted in Table VII, where the results of testing

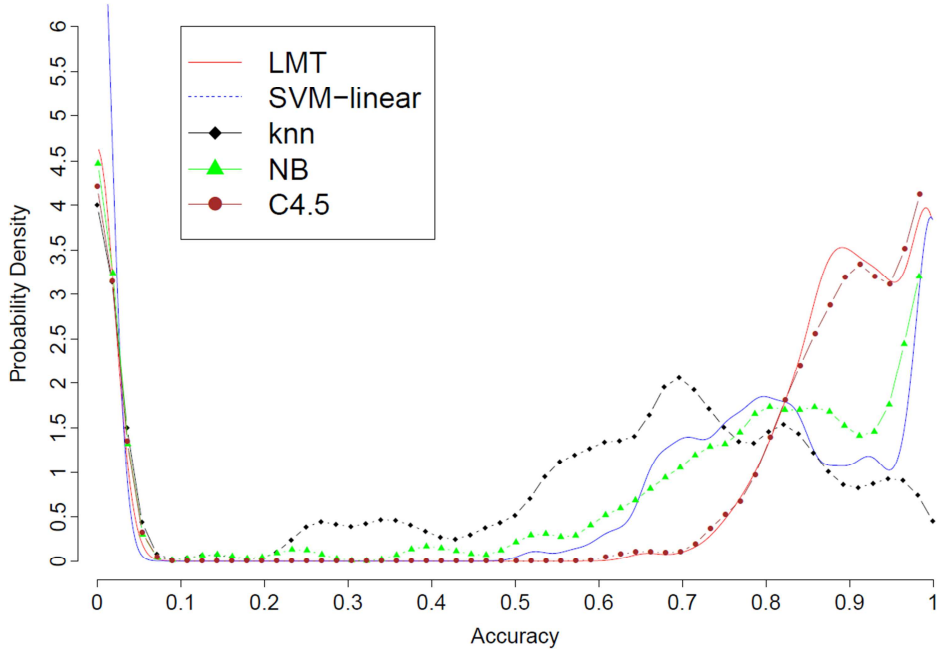


Fig. 3. Probability density functions of the accuracy of the five learning methods used throughout all the families.

the induction learners on same dataset but excluding the PC are displayed.

IG is a simple metric which provides insights on the contribution of an individual feature on reducing the entropy about a given target variable (i.e. driver identification). However, it disregards the dependencies that might be in place between features (e.g. the trip's origin/destination may also have some dependency on the departure time). Consequently, it cannot be faced as a pure rank of the most informative features since it only addresses the relationship between each individual feature and its target variable. Such characteristic comprise one of the main limitation of this method. Fig. 5 displays a pairwise scatterplot to compare the interdependences among the present feature space. On a first glance, the features seem quite uncorrelated among each other. However, the authors want to highlight that such analysis should be conducted in any dataset (prior to the application of any supervised learning task as we are performing hereby).

The LMT and C4.5 are the methods which present the best accuracy on the present dataset. It is well-known that SVM is mainly a binary classifier and therefore, the quality of its results may degrade along with the incensement of the cardinality's (number of possible output labels) of the target variable. The NB also assumes the abovementioned statistical independence between the features – which may explain its lower accuracy.

However, the authors want to highlight that all methods but kNN present a good predictive capacity (i.e. ACC > 0.65). The low performance of kNN may be explained on its sensitivity to parameter changes (such as the number of neighbors k).

The errors uncovered by Fig. 2 also indicate that the errors may be due to high-variance of the methods output (e.g. overfitting of majority class performed by kNN). One known

TABLE VII
AVERAGE WEIGHTED ACCURACY WITHOUT PC FEATURE

Method	Average Weighted Accuracy
Naïve Bayes (NB)	0.637
Logistic Model Tree (LMT)	0.691
k-Nearest Neighbors (kNN)	0.554
Decision Tree (C4.5 algorithm)	**0.694**
Support Vector Machines (SVM)	0.533

technique to reduce such variance-type of error is ensemble learning. To test such hypothesis in the present context, we tested a known boosting-based algorithm – Adaboost [51].

The obtained results (using bootstrapping and a reduced number of trials, i.e.10) are promising: they pointed a superior accuracy of 73.72%. However, it is important to note that this comes in exchange of combining multiple models– thus providing no fair comparison to the previous ones.

Ideally, a study to determine the best algorithm on a given supervised learning task should contain two steps that were not followed on this study: a hyperparameter tuning and a statistical test (such as the Friedman rank test, as proposed by [52] to evaluate the significance of its results. The hyperparameter tuning stage serves to determine which is the best parameter setting to employ with each algorithm. A common technique to address such issue is a grid search over a cross validation procedure conducted with the training set. A more sophisticated version of this is the sequential Monte Carlo method [53] – where multiple possible combinations of parameters are tested on a validation data set prior to the test data (see, for instance, [42]). Regarding a supervised learning context, the aim of employing a Friedman test is to attest if a given method is significantly better than other given a specific predictive task [52]. Such steps were not included in this study because we are not focused on algorithmic details (such as preprocessing/postprocessing tasks and/or significance tests) but on the concept of reusing trip labelled data as additional source of information regarding the driving identification.

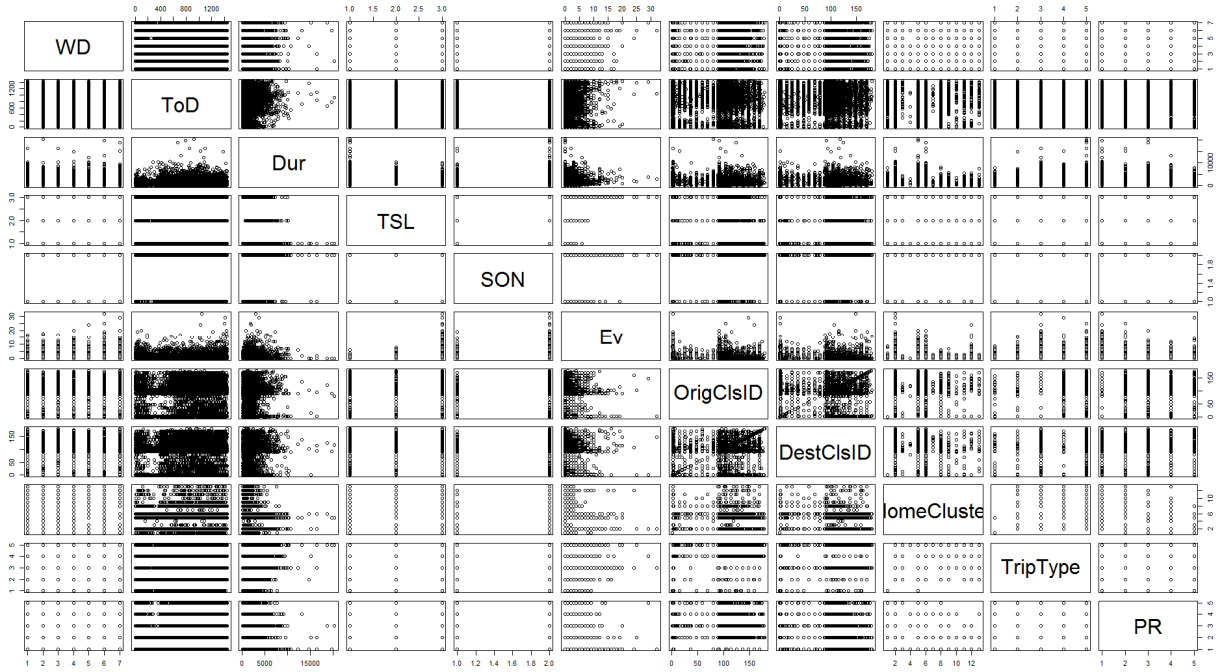


Fig. 5. Scatterplot of the Pairwise combinations of all available feature pairs.

VI. Conclusions

This study aimed at developing and testing an identification methodology when using trip-based data. The main motivation to do it so is to complement the deficiencies of the existing IVDR technologies. For this purpose, a ML framework, including feature evaluation and category identification, was proposed to take advantage of the underlying patterns of the human behavior. Trip-based data collected in Israel was used to test the usefulness of the proposed methodology.

The results of this paper provide a first glimpse on which are the most promising ML techniques for the applications of driver's identification, as well as which travel features are deemed to be informative – and thus relevant - for predictive analytics. A high accuracy was achieved in predicting the driver category using basic trip information. Therefore, the authors believe that this methodology is worth to be further investigated in future studies when using IVDR or similar identification devices. It should be noted, however, that the proposed methodology is recommended to be used as a support method to the different identification technologies (e.g. Dallas Key, Face Identification) and not as a standalone methodology for driver identification. Furthermore, the assumption, in the proposed methodology, is that the training data-set is reliable and trustworthy to a certain extent. Therefore, further research is still needed.

In order to carry out such future work, possible directions are proposed as follows: (1) testing the information gained by other trip features that were not included in this study, such as route choice, type of roads (rural, urban, suburban), and purpose of the trip; (2) testing unsupervised learning approaches to derive the number of driver categories when it is not possible to know it apriori; (3) testing the usefulness and accuracy of the methodology on a larger domain of the drivers' category. In this study four categories were included {father, mother, young driver, and other}, future study can

include siblings, grandparents, etc.; (4) testing the appropriateness of this methodology as a complement and a validation technique to the deficiencies of various identification technologies, such as the iPhoto. In particular, how well this method can work when advanced technologies fails to identify the driver category; (5) investigating the relationships between family characteristics and the importance (IG) of trips' features. This will provide insights for which families the information gained by a specific trip feature will be high and for which it will be low. This can improve the predictive power; (6) testing the proposed methodology on larger datasets, and (7) explore additional ensemble learning approaches (such as gradient boosting [54]).

ACKNOWLEDGMENTS

The authors thank Dr. Tsippy Lotan from Or Yarok Association for providing the data of "The First Year Study". The authors acknowledge Prof. Hans Van Lint from the Department of Transport and Planning, TU Delft for providing the opportunity to bridge the collaboration built to carry out this work.

REFERENCES

[1] N. Lerner, J. Jenness, J. Singer, S. Klauer, S. Lee, M. Donath, *et al.*, "An Exploration of Vehicle-Based Monitoring of Novice Teen Drivers: Final," *NHTSA, report, DOT HS*, vol. 811, p. 333, 2010.

[2] N. Lin, C. Zong, M. Tomizuka, P. Song, Z. Zhang, and G. Li, "An Overview on Study of Identification of Driver Behavior Characteristics for Automotive Control," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[3] A. Silver and L. Lewis, "Automatic identification of a vehicle driver based on driving behavior," ed: Google Patents, 2014.

[4] K. J. Sanchez, A. S. Chan, M. R. Baker, M. Zettinger, B. Fields, and J. A. Nepomuceno, "Systems and methods to identify and profile a vehicle operator," ed: Google Patents, 2014.

[5] B. Wallace, R. Goubran, F. Knoefel, S. Marshall, and M. Porter, "Measuring variation in driving habits between drivers," in *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, 2014, pp. 1-6.

[6] H. Farah, O. Musicant, Y. Shimshoni, T. Toledo, E. Grimberg, H. Omer, *et al.*, "The first year of driving–can IVDR and parental involvement make it safer," in *Proceedings of the Transportation Research Board Annual Conference*, 2013.

[7] T. A. Dingus, S. Klauer, V. Neale, A. Petersen, S. Lee, J. Sudweeks, *et al.*, "The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment," 2006.

[8] V. Neale, S. Klauer, R. Knipling, T. Dingus, G. Holbrook, and A. Petersen, "The 100 car naturalistic driving study, Phase I-experimental design," 2002.

[9] J. H. Ogle, "Quantitative assessment of driver speeding behavior using instrumented vehicles," 2005.

[10] T. Lotan, G. Albert, T. Ben-Bassat, and D. Ganor, "Potential benefits of in-vehicle systems for understanding driver behaviour," 2010.

[11] C. G. Prato, T. Toledo, T. Lotan, and O. Taubman-Ben-Ari, "Modeling the behavior of novice young drivers during the first year after licensure," *Accident Analysis & Prevention*, vol. 42, pp. 480-486, 2010.

[12] C. M. Farmer, B. B. Kirley, and A. T. McCartt, "Effects of in-vehicle monitoring on the driving behavior of teenagers," *Journal of Safety Research*, vol. 41, pp. 39-45, 2010.

[13] H. Farah, O. Musicant, Y. Shimshoni, T. Toledo, E. Grimberg, H. Omer, *et al.*, "Can providing feedback on driving behavior and training on parental vigilant care affect male teen drivers and their parents?," *Accident Analysis and Prevention*, vol. 69, pp. 62-70, Aug 2014.

[14] C. M. Roberts, "Radio frequency identification (RFID)," *Computers & Security*, vol. 25, pp. 18-26, 2006.

[15] Y. Xiao, S. Yu, K. Wu, Q. Ni, C. Janecek, and J. Nordstad, "Radio frequency identification: technologies, applications, and research issues," *Wireless Communications and Mobile Computing*, vol. 7, pp. 457-472, 2007.

[16] B. G. Simons-Morton, M. C. Ouimet, Z. Zhang, S. E. Klauer, S. E. Lee, J. Wang, *et al.*, "Crash and risky driving involvement among novice adolescent drivers and their parents," *American journal of public health*, vol. 101, p. 2362, 2011.

[17] S. C. Hoi, D. Wang, I. Y. Cheng, E. W. Lin, J. Zhu, Y. He, *et al.*, "Fans: face annotation by searching large-scale web facial images," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 317-320.

[18] D. Wang, S. C. Hoi, P. Wu, J. Zhu, Y. He, and C. Miao, "Learning to name faces: a multimodal learning scheme for search-based face annotation," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 443-452.

[19] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*: Springer Science & Business Media, 2009.

[20] R. L. Klevans and R. D. Rodman, *Voice recognition*: Artech House, Inc., 1997.

[21] R. P. Wildes, "Iris recognition: an emerging biometric technology," *Proceedings of the IEEE*, vol. 85, pp. 1348-1363, 1997.

[22] L. Masek, "Recognition of human iris patterns for biometric identification," Master's thesis, University of Western Australia, 2003.

[23] J. Thompson, M. Baldock, J. Mathias, and L. Wundersitz, "The benefits of measuring driving exposure using objective GPS-based methods and subjective self-report methods concurrently," in *Australasian Road Safety Research, Policy & Education Conference (2013: Brisba, Australiae)*, 2013.

[24] R. A. Blanchard, A. M. Myers, and M. M. Porter, "Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers," *Accident Analysis & Prevention*, vol. 42, pp. 523-529, 2010.

[25] S. C. Marshall, K. G. Wilson, F. J. Molnar, M. Man-Son-Hing, I. Stiell, and M. M. Porter, "Measurement of driving patterns of older adults using data logging devices with and without global positioning system capability," *Traffic injury prevention*, vol. 8, pp. 260-266, 2007.

[26] P. Stopher, C. FitzGerald, and J. Zhang, "Search for a global positioning system device to measure person travel," *Transportation Research Part C: Emerging Technologies*, vol. 16, pp. 350-369, 2008.

[27] C. Inbakaran and A. Kroen, "Travel Surveys–Review of international survey methods," in *Australasian Transport Research Forum*, 2011.

[28] J. Thornton, M. Savvides, and B. V. K. V. Kumar, "A Bayesian Approach to Deformed Pattern Matching of Iris Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 596-606, 2007.

[29] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver Inattention Monitoring System for Intelligent Vehicles: A Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 596-614, 2011.

[30] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view Face Detection Using Deep Convolutional Neural Networks," presented at the Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 2015.

[31] S. Bhosale and B. Sawant, "Security in e-banking via card less biometric atms," *International Journal of Advanced Technology & Engineering Research*, vol. 2, pp. 457-462, 2012.

[32] C. Zhang, M. Patel, S. Buthpitiya, K. Lyons, B. Harrison, and G. D. Abowd, "Driver Classification Based on Driving Behaviors," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 2016, pp. 80-84.

[33] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting Taxi–Passenger Demand Using Streaming Data," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, pp. 1393-1402, 2013.

[34] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, and J. F. de Sousa, "Validating the coverage of bus schedules: A Machine Learning approach," *Information Sciences*, vol. 293, pp. 299-313, 2015.

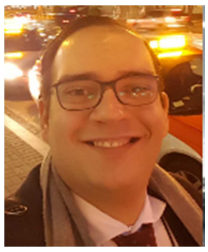
[35] R. Nunes, L. Moreira-Matias, and M. Ferreira, "Using exit time predictions to optimize self automated parking lots," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, 2014, pp. 302-307.

[36] J. R. Quinlan, *C4. 5: programs for machine learning*: Elsevier, 2014.

[37] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, pp. 161-205, 2005.
[38] P. E. Duda and O. Richard, "Hart, pattern classification and scene analysis," ed: John Wiley and Sons, New York, 1973.
[39] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273-297, 1995.
[40] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, pp. 175-185, 1992.
[41] O. Musicant and Y. Benjamini, "Driving patterns of novice drivers–a temporal spatial perspective," in *Transportation Research Board 91st Annual Meeting*, 2012.
[42] L. Moreira-Matias, J. Mendes-Moreira, J. Gama, and P. Brazdil, "Text categorization using an ensemble classifier based on a mean co-association matrix," in *Machine Learning and Data Mining in Pattern Recognition*, ed: Springer, 2012, pp. 525-539.
[43] R. C. Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012," ed: ISBN 3-900051-07-0, 2014.
[44] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.
[45] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien," 2015.
[46] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*: Springer Science & Business Media, 2013.
[47] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065-1076, 1962.
[48] B. W. Silverman, *Density estimation for statistics and data analysis* vol. 26: CRC press, 1986.
[49] L. Moreira-Matias, J. M. Mendes-Moreira, M. Ferreira, J. Gama, and L. Damas, "An online learning framework for predicting the taxi stand's profitability," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, 2014, pp. 2009-2014.
[50] L. Moreira-Matias, J. Gama, J. Mendes-Moreira, and J. F. de Sousa, "An Incremental Probabilistic Model to Predict Bus Bunching in Real-Time," in *Advances in Intelligent Data Analysis XIII*, ed: Springer, 2014, pp. 227-238.
[51] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119-139, 8// 1997.
[52] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the fbietkan statistic," *Communications in Statistics-Theory and Methods*, vol. 9, pp. 571-595, 1980.
[53] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, pp. 899-924, 2007.
[54] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.

Haneen Farah is an Assistant Professor at the Department of Transport and Planning, Delft University of Technology. Her research focuses on traffic and road safety, driver behaviour modelling, road geometric design, cooperative systems and automated vehicles. In her research she combines elements from the transport systems analysis, behavioural and human factors sciences and econometrics. She developed models for crash prediction, risk analysis, driver behaviour, and impact of in-vehicle data recorders on driver behaviour and safety. She received her Ph.D. degree from the Technion - Israel Institute of Technology in 2009. Between 2009 and 2011 she was a post-doctoral researcher at the Department of Transport Sciences, KTH - Royal Institute of Technology, Stockholm, Sweden. During that period she focused on important problems in the area of emerging technologies and on advancing the state of the art in modelling driver behavior. She has experience in several European projects and COST Actions, such as COOEPRS, MULTITUDE, TEMPUS, HFAuto.

She has more than 25 scientific papers published in leading refereed international journals, such as Accident Analysis & Prevention, Transportation Research Part B, C, and F.



Luis Moreira-Matias received his Ms.c. degree in Informatics Engineering and Ph.d. degree in Machine Learning from the University of Porto, in 2009 and 2015, respectively. During his studies, he won an International Data Mining competition held during a Research Summer School at TU Dortmund (2012). Luis served as invited reviewer and/or in the Program Committee of multiple high-impact research journals/venues, such as ECML/PKDD, IEEE ITSC, TRB, ACM SIGKDD, IEEE TKDE, Elsevier's ESWA or IBERAMIA, among others.

Currently, he works as a Research Scientist at NEC Laboratories Europe (Heidelberg, Germany), integrated in the Intelligent Transportation Systems group. His research interests include Machine Learning, Intelligent Public Transports and Big Data Analytics applied to improve Urban Mobility in general. He authored 25+ publications in world-leading venues on related topics.