Driving Behavior Identification and Real-World Fuel Consumption Estimation With Crowdsensing Data

Aurélie Pirayre, Pierre Michel, Sol Selene Rodriguez, and Alexandre Chasse

***Abstract*— It is now well established that car driving behavior impacts gas pollutant emission, fuel consumption and safety. With the rise of low-cost and easily accessible mobility data (measured by GPS, accelerometer, gyroscope, etc), through OBD or smartphones devices for instance, the community is increasingly interested in characterizing these driving behaviors. In this context, we used the IFPEN's application named GECO AIR to access GPS data on which we derived statistical descriptors to perform unsupervised classifications and highlight different trip- and driver-related behaviors (dynamic, slow, traffic jam, etc.), on different road types (urban, highway, etc). Using Markov chain, we then generate for each characterized behavior a representative velocity profile and combine them according to a given use to estimate the associated real-world fuel consumption. Promising performances have been obtained with more than half of the recorded trips for which the real-world fuel consumption is estimated with less than 10% of error. The added value of the proposed work is the capability to estimate fuel consumption in a posterior way without any GPS records, as information obtained through a questionnaire for instance could be sufficient.**

***Index Terms*— Smartphone sensor, GPS, unsupervised classification, driving behavior, real-world fuel consumption estimation.**

I. Introduction

DUE to well-known environmental problems, reducing carbon footprint is essential and a more reliable knowledge of the fuel consumption (e.g. CO_2 emission) becomes a decisive criterion for consumers in the choice of a vehicle. However, whereas the official fuel consumption is available for the consumer, this consumption is measured in standard laboratory conditions and underestimates several factors leading to a significant gap between the real fuel consumption and the official one, around 30% to 40% for a recent vehicle [1]. The factors explaining such a gap can be related to three mains categories: the vehicle, the use case and the driver behavior [2].

On the one hand, the vehicle affects the fuel consumption according to its characteristics, mainly related to i) car body: mass, extra-mass, segment, dimensions, and ii) car engine: fuel type, maximum power, displacement, Euro emissions standard. Additional characteristics linked to after-treatment heating and electric auxiliaries can also be taken into account. Due to the physical understanding and knowledge of the fuel consumption through an engine, a lot of work has been done to deal with the vehicle factor [3], [4]. For instance, quasi-static approaches, parametrized by vehicle characteristics, are useful to estimate the fuel consumption [5] from speed and slope traces.

On the other hand, as knowledge related to both use case and driver behavior is more difficult to capture, the study of these two factors is recent and has been growing since the high-throughput acquisition of mobility data through sensors such as accelerometer, gyroscope, GPS (Global Positioning System), OBD (On-Board Diagnostics), etc. Trip-related behavior, also named use case in this work, is linked to the road type such as urban extra-urban, highway, etc., and road conditions: traffic, weather, slope, etc. Such conditions are important to take into account as the impact of the driving behavior is significant [6], [7]. The last factor, driver-related, is directly linked to the driving style, or in other words, *"the way a driver chooses to drive"* [8]. This driver behavior is thus related to the aggressiveness, speed and/or safe level of the driver.

Combining vehicle characteristics as well as trip- and driver-related behaviors can allow to estimate the real fuel consumption. But the identification of the driving behavior cannot be restricted to the help of fuel consumption estimation. Indeed, detecting driving behaviors also plays an important role to help the driver to adopt *eco-driving* behavior i.e. driving behavior minimizing the air pollutants [9]. Recommendation, driving assistance systems, maintaining safe and sustainable transport and autonomous vehicles also benefit from driving behavior characterization [10].

However, as previously mentioned, detecting and identifying driving behavior require relevant data related to speed, acceleration, deceleration, *etc*. The rise of connected objects now makes it possible to obtain such information. One interesting device coming from the vehicle itself is the OBD, which allows to get information regarding engine speed, vehicle speed, pedal position, spark advance, airflow rate, coolant temperature, etc. Unfortunately, despite the richness of these data, they are still only accessible through manufacturers' devices, which can be an obstacle to its accessibility. Relevant mobility data can also be captured through motion sensors such as accelerometer, gyroscope or magnetometer (Inertial Measurement Unit - IMU) and/or location sensor such as GPS. These different types of sensors are nowadays found in smartphones and make them a very useful and powerful tool to collect and analyze mobility data [11].

In the literature a wide research field based on car-following models offers to reproduce measured acceleration dynam-

Manuscript received 29 April 2021; revised 10 November 2021 and 11 February 2022; accepted 12 April 2022. Date of publication 23 May 2022; date of current version 11 October 2022. The Associate Editor for this article was G. Wu. *(Corresponding author: Aurélie Pirayre.)*

The authors are with IFP Energies nouvelles, 92852 Rueil-Malmaison, France (e-mail: aurelie.pirayre@ifpen.fr).

Digital Object Identifier 10.1109/TITS.2022.3169534

ics [12] by modeling the inter-vehicle gaps with some refinements as for example drivers pedals management [13] or some specific actions modeling as for example lane changes [14]. This descriptive approach is useful in a context of traffic simulations to understand drivers decisions involvements in traffic dynamic and transportation systems. The approach presented in this paper is a data-oriented approach considering that driving behaviors are intrinsically contained into speed and acceleration data.

This work is dedicated to the analysis of smartphone GPS data to identify trip- and driver-related behaviors in order to estimate real-world fuel consumption. The advantage of this work is that the different behaviors are translated into representative velocity profiles allowing us to estimate a real-world fuel consumption only based on semantic information given by the user. For this purpose, driving behaviors are firstly identified for each road type (*i.e.* 30km/h zone, urban, extra-urban and highway) through an unsupervised classification of statistical features derived from velocity and acceleration GPS signals (see Section III). A representative velocity trace is then generated for each behavior of each road type and aggregate to represent a real use in order to estimate a real-world fuel consumption (see Section IV). Results regarding the behavior identification as well as the fuel consumption estimation are given in Section V before drawing some conclusions and perspectives (see Section VI).

II. Related Works

Due to the rise of mobile crowdsensing technique, lot of works have been proposed during the past decade regarding driving behavior identification. We refer to [10], [11], [15] for extended reviews of driving behavior analysis using smartphone sensor's data and devote this section to a selected overview of the main approaches done in this field, in terms of used methods and purpose of the detection.

On the one hand, a large panel of approaches are developed for driving behavior characterization to evaluate the riskiness of the driver. In [16], authors use smartphone-based data including accelerometer, gyroscope and magnetometer (IMU) and GPS to detect some maneuvers using an endpoint detection. Events are then qualify to aggressive driving behavior or not using a Dynamic Time Warping (DTW) based approach. In the same vein, authors in [17] developed a Maximum Likelihood classifier to distinguish aggressive to safe maneuvers obtained through an endpoint detection on GPS and IMU data. Authors in [18] used a similar approach only using IMU data, to detect maneuvers. They then used a DTW-based algorithm followed by a Bayes classifier to identify risky and safe events. In the study of [19], authors developed a mobile application for risky driving behavior detection based on IMU and GPS sensors. Fuzzy rules are defined to detect events and quantification of the riskiness of the event is performed using additional information such as weather and time of the day. Popular neural network can also be used to classify driving behaviors using smartphone sensor data (accelerometer, gyroscope and magnetometer) as in [20]. Indeed, driving behaviors are defined through maneuvers type that are detected using a multi-layer perceptron neural network on attributes derived from smartphone senor data. Qualification of the safe or aggressive maneuvers is then done using a fuzzy inference based on the knowledge of aggressive and safe patterns. Authors in [21] use smartphone sensors and ensemble learning (K-NN, SVM and MLP) to detect and classify type of maneuvers and congestion level. They then use fuzzy inferences systems to finally distinguish safe to dangerous maneuvers. Not based on maneuvers or events, authors in [22] designed a sensing platform based on smartphone, OBD2 and IMU sensors. This sensing platform, thinking to give real-time feedback to the user, thus collects data from which they define some features used to predict aggressive style driving using naive Bayes classifier. In [23], authors enhance Driver Assistance System (DAS) by detecting infrastructure (intersection or segment) with neural network to refine an unsupervised driver classification (cautious, aggressive or normal).

One the other hand, other works are dedicated to driving behavior characterization and their links to fuel consumption. Authors in [24] developed a mobile application which help the user to decrease its fuel consumption by detecting driving pattern and give real-time feedback to the driver. Data coming from CAN bus and OBD2, for which some features are extracted such as average, extrema, percentage of time that the vehicle is stopped. From these features, rule-based approaches are defined to detect road types and evaluate fuel consumption. These two criteria are then used through a fuzzy logic scheme to orientate the user through the best driving pattern will reduce its fuel consumption. In [25], authors developed an Android application which uses OBD2-based data and neural network to classify the drivers according to their driving styles as well as the road type (city, suburban, highway). In complement, they show a correlation between the driving style and the fuel consumption and greenhouse gas emissions. A correlation between driving behavior using smartphone data and fuel consumption using OBD data is shown in [26]. A feature selection gives seven attributes derived from smartphone's data and related to statistic on the speed, acceleration and deceleration to directly predict the fuel consumption. Fuel consumption prediction using three machine learning algorithms: Neural network, Support Vector Regression and random forest (RF), where RF seems to get the best accuracy.

Although these methods are conceptually close to ours, the main differences are: i) a larger variety of data from various sources such as ODB, CAN, GPS, *etc.* and/or ii) the fuel consumption information in a supervised way. Based on this analysis, a fair numerical benchmark seems to not be practicable and pertinent.

The originality of our proposed approach is threefold:

- model construction only depends on velocity and acceleration traces from GPS. In practice, velocity with a sufficient sampling from other types of acquisition device could be used.
- fuel consumption is estimated through an unsupervised data-driven approach. While the validation could be more difficult in practice, this prons allows to use the proposed methodology in a real world without having any manu-

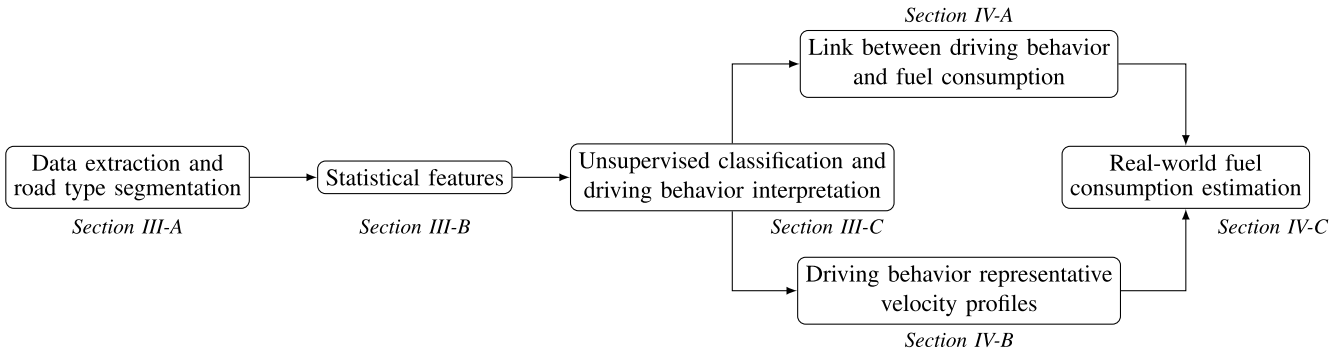


Fig. 1. Complete scheme of the proposed methodology for driving behavior identification from smartphone data and real-world fuel consumption estimation.

facturer data related to the fuel consumption, data that is often complicated to obtain.

- *a posteriori* prediction of the fuel consumption based only on origin-destination information and macro behavior. This added-value allows to predict more realistic usage-based fuel consumption for a new driver, without use any recorded data regarding this new driver.

III. Driving Behavior Identification

An overview of the proposed pipeline for driving behavior identification from smartphone data to estimate real-world fuel consumption is given in Fig. 1.

A. Data Presentation and Road Type Segmentation

In order to identify driving behaviors to estimate real-world fuel consumption, this work is based on two kinds of data: i) GPS data reaped from the IFPEN's application named GECO AIR and ii) road infrastructure data coming from the HERE mapmaker.

1) Smartphone GPS Data: The GECO AIR application (www.gecoair.fr) was firstly dedicated to fuel consumption and pollutant emission estimation [9]. From recent years, some improvements were performed to better understand the mobility and to offer to the user more service such as personalized ecological impacts advices to improve their mobility with car, walk, bike, public transport, etc. All models used in GECO AIR use as inputs GPS data reaped by the application such as longitude, latitude, velocity, heading, elevation, etc.

The application was used by more than 30 000 users since 2017, and gathered 75 million kilometers mainly covering the French territory. However, we can observe a light bias in the representativeness as the majority of trips are more related to urban than rural areas. Despite this bias, which will be considered in perspective, we decide to extract a subset of data for which the departure or the arrival is located in one of the two French departments: *Nord* (59) and *Pas-de-Calais* (62). Using this filter and other additional ones related to the quality and the validity of the signals, extracted data contain 73 261 trips that cover around 1 560 000 kilometers. We notice that multiple trips may be associated to the same driver. However, aggregating data at the driver scale tends to characterized the driving behavior of the driver itself while our goal is slightly different as it tends to extract the main representative driving behaviors in general. Thus we consider trips independently by neglecting underlying drivers. For each trip of this data subset, we have a set of time series sampled every second regarding longitude, latitude, cumulative distance, velocity, acceleration and time base. We would like to mention that data are anonymized and GDPR compliant. First, the driver is anonymized as it is stored in our database through an ID and no additional personal information are used. Secondly, the longitude and latitude signals are used at the recording for the map-matching in order to obtain infrastructure information. We then use the aggregated infrastructure information only and never use the original geolocation information. As we only deal with speed, speed limit, acceleration, we are not able to trace the identification of the person from whom the data originated.

2) Here Data: As previously introduced, our goal is to identify various behavior at a road type scale (30km/h zone, urban, extra-urban and highway, denoted by ZONE_{30}, URBAN, χ-URBAN and HIGHWAY, respectively). To perform this segmentation we used road infrastructure data coming from the cartographer HERE. We firstly map-match the GPS longitude and latitude into the HERE frame thank to their proprietary solution based on Hidden Markov Model [27]. We thus obtain, for each trip, a list of HERE links corresponding to road sections with constant attributes. These road sections are characterized by their matched longitude and latitude, but also by some additional information related to the infrastructure of the road. Among the information we may obtain, we specifically focus on the associated speed limit. Hence, for each of the 67 689 selected trips, we are able to reconstruct a one-second time series trace corresponding to the speed limit encountered at each second. As detailed in the next section, these speed limit traces will serve us to segment trips according to different road types.

3) Velocity and Acceleration Data Segmentation: In order to obtain accurate driving behaviors, we decide to segment the data to deal with different road types. We chose to define four road types, named ZONE_{30}, URBAN, χ-URBAN and HIGHWAY, and defining the set $\mathcal{R}$. Each road type is characterized by specific speed limits as detailed in Table. I.

TABLE I
SPEED LIMIT SEGMENTATION DEFINING ROAD TYPES

Road type	Speed limit (v_{lim})	Number of sub-trips (T)
ZONE$_{30}$	30	30403
URBAN	50	62362
χ-URBAN	{70, 80, 90}	50125
HIGHWAY	{110, 130}	22262

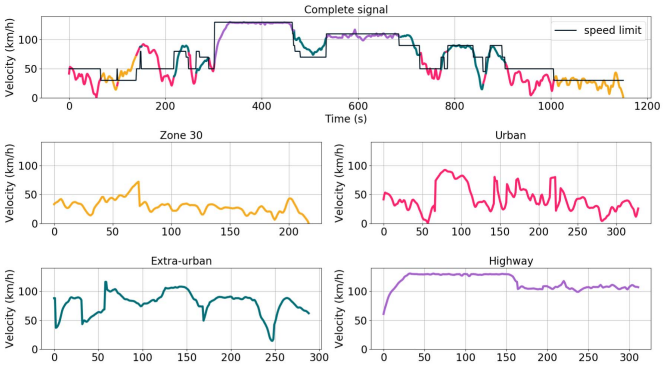


Fig. 2. Illustration on the speed signal v of the data reorganization used to obtained sub-traces according to road types.

The initial velocity v and acceleration a traces are segmented into sub-traces according to a cutting based on encountered speed limit v_{lim}. Hence, for each road type $r \in \{\text{ZONE}_{30}, \text{URBAN}, \chi\text{-URBAN}, \text{HIGHWAY}\}$, we might obtain the sub-traces $v^{(r)}$ and $a^{(r)}$ by concatenating sections of v and a, respectively, if the speed limit matches with the corresponding road type speed limit (cf. Fig. 2). From now, features computation and classification are performed for each road type independently. Hence, presented work in the two next subsections is identically usable for each road type.

B. Statistical Features for Classification

We assume that driving behaviors can mainly be captured through information derived from velocity and acceleration. Hence, for each road type $r \in \mathcal{R}$, we define a set of 18 interpretable features (variables) used as input for the unsupervised classification. For some of them, we were motivated by some metrics given in [28]. We deal with mixed variables as 16 of the 18 features are numerical while the two others are categorical. We now expose the different features while the mathematical definitions are given in Appendix A.

1) Numerical Features: Numerical variables are based on three types of signal: the velocity $v^{(r)}$, the acceleration $a^{(r)}$, and the 2D probability density of velocity and acceleration $\delta^{(r)}(v^{(r)}, a^{(r)})$. We consider global statistics on the whole reconstructed signals (referred as *signal*) and aggregated statistics, through average or median, computed from each section used to reconstruct the signals (referred as *section signals*). Note that as we compute the same features per road type, we now voluntary omit the superscript referring the road type r to lightened notations to come: for instance, $v^{(r)}_{\text{max}}$ will be simply denoted by v_{max}.

Specifically, from the velocity signal, we define: the average speed v_{mean}, the maximal speed v_{max} and the percentage of null speed v_{null}. Based on the velocity section signals, we define $\bar{v}_{\text{max}}$ as the average of the maximal speed encountered on each concatenated sections We also compute $\bar{v}_{\text{diff}}$ as the average of the difference, computed on each section, between the speed limit and the maximal speed. Note that our motivation to use percentage of null velocity or differences between the maximal and the limit speed is driven by their capability to specifically catch traffic jam conditions.

Regarding features related to acceleration, we define the average and maximal acceleration, denoted by a_{mean} and a_{max}, respectively. Two metrics, denoted by H and G and based on the sparsity metric Hoyer and Gini index respectively [29] are also computed from the acceleration to identify the dynamical behavior. The first one is a normalized version of the ℓ_2/ℓ_1 measure while the second is derived from the Lorentz curve. In complement to these two dynamics oriented features, we compute the relative positive acceleration RPA and its unweighted version RPA_{unweight}. We also compute A_{km}, the number of significant acceleration per kilometers.

Finally, based on both velocity and acceleration signal, we firstly compute a 2D probability density $\delta(v, a)$ using the bi-variate kernel density estimator developed in [30] (see Fig. 4a). The 2D probability density is a matrix where columns correspond to the velocity binning while rows correspond to the acceleration one. Thus, the vectorization across the velocity dimension is done by stacking the columns of the matrix on top of another. As a result, the obtained 1D signal is composed on acceleration distribution peaks regularly spaced over the velocity binning (see Fig. 4b). In order to numerically characterize the density, the sixteen highest peaks are detected and characterized through four criteria. For a given criteria, 16 values are thus obtained, one per considered peak. The median over these 16 values is then computed and use as feature to evaluate the considered criterion at the entire 1D probability scale. Based on this methodology, we define the following criteria for each selected peak p: the basal width of the peak b_p translated into the reached maximal acceleration $a_p = a(b_p)$, the width of the peak at a probability of 1e-5, denoted by ω_p. Based on these two measures, we define d_p the ration of the basal width b_p and ω_p, which reflect the dynamics of the acceleration by evaluating the proportion of null acceleration *i.e.* d_p equals 1 means a relatively smooth acceleration (without null acceleration) while a d_p close to 0 indicates a very dynamic acceleration (with numerous phases with a null acceleration). Finally, we also define r_p the ratio between the area of the peak, denoted s_p, and its basal width.

An illustration of first order descriptors according to different segments of a trip is given in Fig. 3. This illustration helps us to evaluate how the feature well captured the change in behaviors and road type effect.

2) Categorical Features: Two categorical features are defined. The first one, denoted by L_t, is related to the departure time, where the day is divided into four phases: from 6h to 10h (morning), from 10h to 16h (afternoon), from 16h to 20h (evening), and from 20h to 6h (night). The feature thus corresponds to the phase label encompassing the departure

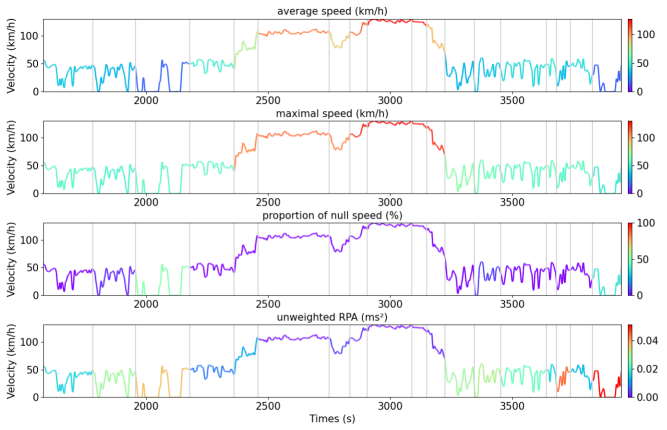


Fig. 3. Extract of a velocity profile of a real trip included multiple road types and driving behaviors. Profiles are colored in function of the studied feature (v_{mean}, v_{max}, v_{null}, and RPA_{unweight}-from top to bottom).

time: $L_t \in \{\text{MORNING, AFTERNOON, EVENING, NIGHT}\}$. The second categorical feature, denoted by L_d, corresponds to the label resulting from a classification performed on the 2D probability densities. To perform this classification, we adapted the K-means algorithm [31], [32]. Indeed, K-means is dedicated to classify vectors through the use of euclidean distance. As you data to classify corresponds to densities, we used a dedicated distance that quantify the difference between two probability distribution [33]. We thus replaced the euclidean distance used to evaluate the distance to the barycenter by a metric derived from the Kullback-Leibler divergence ($\mathcal{KL}$). As this divergence is not symmetric, the use of this metric can be prejudicial to the algorithm. To overcome this drawback, we used as distance the symmetric version defining as: $d(P, Q) = \mathcal{KL}(P, Q)/2 + \mathcal{KL}(Q, P)/2$, where P and Q are the two densities to compare. Hence, the proposed procedure to classify the set of densities in K clusters is sum up as follows:

1) Initialize cluster centroids$\{\Delta_1, \Delta_2, \ldots, \Delta_K\}$
2) Repeat until convergence {
   - For every sub-trip τ, set
   $$L_d^{(\tau)} = \arg\min_k d(\delta_\tau, \Delta_k).$$
   - For every centroids k, set
   $$\Delta_k = \frac{\sum_\tau \mathbb{1}(L_d^\tau = k)\delta_\tau}{\sum_\tau \mathbb{1}(L_d^\tau = k)}.$$

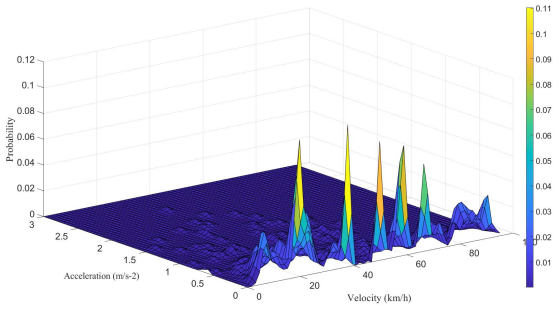
   }

As a results, naming the obtained classes by CLUSTER_i for the cluster i, the feature L_d takes its value in $\{\text{CLUSTER}_1, \ldots, \text{CLUSTER}_K\}$.

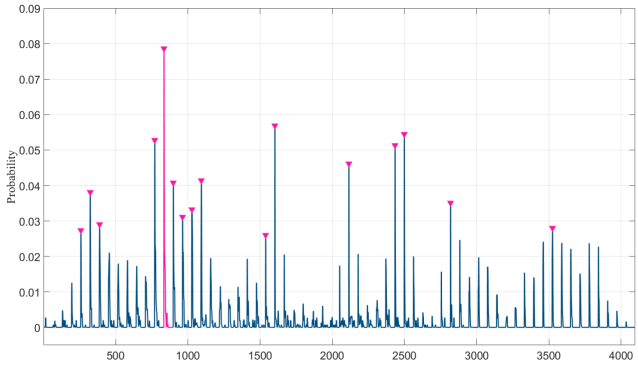
Based on this set of mixed features, we then performed an unsupervised classification to identify driving behaviors.

C. Unsupervised Classification for Driving Behavior Identification

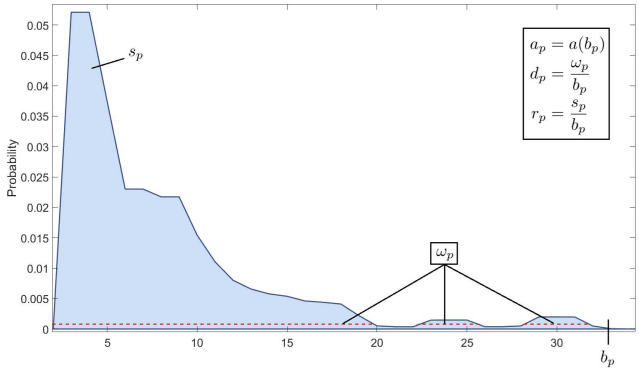
For each road type $r \in \mathcal{R}$, features are computed for each trip $\tau = \{1, \ldots, T\}$ of the corresponding road type



(a) 2D probability density $\delta(v, a)$. Velocity v and acceleration a signals are binned in their respective dimension (2D histogram) before computed a smooth density using bivariate kernel density estimator. The color of the density is correlated to the probability.



(b) 1D vectorized density obtained by stacking the columns of the matrix $\delta(v, a)$ on top of another. The sixteen highest detected peaks are marked with a pink triangle.



(c) Features description for the 4th selected peak (highlighted in pink in the figure above). b_p is the basal width of the peak and a_p refers to the corresponding binned acceleration. d_p is the ratio between ω_p, the width of the peak above a probability of 1e-5 and b_p. r_p is the ratio between s_p, the area of the peak, and b_p.

Fig. 4. Features description from 2D probability density $\delta(v, a)$.

and gathered in a matrix $\mathbf{F}^{(r)}$ of size $T \times A$, where $A = 18$ is the number of attributes. This matrix is then used as input for the classification. As our matrix of features is composed of mixed attributes i.e. numerical and categorical, we choose to perform the classification with the K-prototype algorithm [34]. This algorithm integrates the K-means and K-modes algorithms, which are used to specifically classify numerical and categorical variables, respectively. Note that we

have deliberately chosen to deal with mixed data to avoid any pre-processing or encoding of the data.

For a given cluster, the representation vector can classically be defined as the center-like of this cluster. In the context of mix numerical and categorical features, this center is called a prototype. Let Q_k be such a prototype of the cluster k. Let $\mathbf{U} \in \mathbb{R}^{T\times K}$ be the partition matrix encoding the affiliations to each datum at a cluster. For $\tau \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$, it is defined as follows:

$$u_{\tau,k} \in \{0, 1\}, \quad \text{and} \quad \sum_{k=1}^{K} u_{\tau,k} = 1$$

The objective function to be minimized for grouping the mixed data in $\mathbf{F}$ into K clusters is given by:

$$\sum_{k=1}^{K}\sum_{\tau=1}^{T} u_{\tau,k}\, d(F_\tau, Q_k),$$

where d is the dissimilarity measure taking into account both numerical and categorical data. Let us assume that the data in $\mathbf{F}$ are organized such that the first P columns of $\mathbf{F}$ gather the numerical data while the remaining others ones gather categorical data. In such a case, the dissimilarity measure d can be defined as follows:

$$d(F_\tau, Q_k) = \underbrace{\sum_{p=1}^{P}(f_{\tau,p} - q_{k,p})^2}_{\substack{\text{Euclidean cost for}\\ \text{numerical data}}} + \gamma \underbrace{\sum_{a=p+1}^{A} \mathbb{1}(f_{\tau,a} \neq q_{k,a})}_{\substack{\text{Dissimilarity cost}\\ \text{for categorical data}}},$$

where $\mathbb{1}(\cdot)$ denotes the characteristic function that equal to 1 if its argument is verified and 0 otherwise, and γ is a parameter controlling the influence of the categorical data with respect to the numerical ones. An iterative procedure is then used to solve this optimization process and obtained the final classification.

An expert interpretation of the resulting clusters is then performed in order to attribute a representative driving behavior at each classes. Results regarding both the classification and the cluster interpretations are given in Section V-A. We present in the next section how to add value to this classification results by using these clusters to generate driving cycles representing driving behaviors and estimate real-world fuel consumption.

IV. Real-World Fuel Consumption Estimation

Our aim is to estimate real-world fuel consumption corresponding to a specific driver use case. For this purpose, we assume that the global driving behavior is a composition of individual driving behaviors previously identified across multiple road types. Hence, we will detail in this section, our following proposition: we first validate our assumption based on dependence between driving behavior and fuel consumption, then we generate for each clusters (i.e. driving behavior) a representative driving cycle and then, aggregate their information according to the user to obtain the real-world fuel consumption.

A. Fuel Consumption Dependance to Driving Cycle Behavior

A common approach to model the fuel consumption is to model the longitudinal speed dynamic with a quasi-static model [3]. From a speed cycle and thanks to a forces balance the energy demand a the wheel is deduced. Then by modeling the powertrain gears and losses the energy demand at the engine is calculated. And finally the fuel consumption is computed with a fuel consumption map. The model used in this work is presented in [5] where the model is validated with experimental data. The model is sufficiently accurate to model Euro 4 to Euro 6 gasoline and diesel conventional vehicles i.e. the real-world fuel consumption for a given trip represented by a speed cycle is precisely calculated ensuring to model the corresponding fuel consumption.

Thanks to this model, we are able to estimate the real fuel consumption, in L/100km, from the speed cycle: for each complete speed profile (before the velocity segmentation), we computed the instantaneous fuel consumption. On this vector of instantaneous consumption, we then applied the same methodology as for the velocity consumption to obtain the instantaneous consumption per road type $r \in \mathcal{R}$. This procedure allows us to compute, for each sub-trips τ, the consumption per identified road type r and denoted by $c_\tau^{(r)}$.

As mentioned in Section III-C, sub-trips for a given road type are classified. Thus, based on the classification results, we obtained a set of consumption per class $\mathcal{C}_k^{(r)} = \left(c_\tau^{(r)}\right)_{\tau=\{1,\ldots,T_k\}}$, where T_k is the number of sub-trips in the cluster k. Each generated cluster is supposed to be dependent on the driving behavior. In order to validate our assumption that driving behavior affects the consumption, a statistical analysis was performed to identify whether the clusters show discriminant consumption. This statistical analysis was performed through multiple comparisons using a pairwise Welch's t-test [35] and the Benjamini-Hochberg procedure for the p-value adjustment [36].

B. Driving Cycle Reconstruction

As introduced, our goal is to generate a representative driving cycle, for each cluster for which we identified a driving behavior. A label l_τ is associated to each sub-trip τ. This label equals the cluster number obtained by the K-prototype algorithm. Then, for a given cluster k, we gathered in a set $\mathcal{D}_k$, all the real velocity v_τ and acceleration a_τ profiles corresponding to the sub-trips set for which the label l_τ equals k: $\mathcal{D}_k = \{v_\tau, a_\tau\}_{\tau=\{1,\ldots,T\}} | l_\tau = k,\ \forall k \in \{1, \ldots, K\}$. From this set $\mathcal{D}_k$ of real velocity and acceleration profile associated to a driving behavior, we used the methodology developed in [37] to generate a representative driving cycle.

The method is based on the discrete-time stochastic process of the Markov chains, thus driven by the postulate that the future state only depends on the current state (no memory effect). Designed for driving cycle, the proposed Markov chains allows thus to determine the probability to obtain the next acceleration (a_{t+1}) knowing the current velocity and acceleration (v_t, a_t): $a_{t+1} \sim P(a_{t+1} \,|\, a_t, v_t)$. The probability density, assumed Gaussian, can be fitted using velocity and

acceleration in $\mathcal{D}_k$. Hence, at each step $t+1$, the most probable acceleration obtained through these probability densities is computed and the corresponding velocity is then derived. An additional state is defined, valued by 0 or 1 depending whether the vehicle is turned off or on, respectively. This state is driven by a probability to turn off the vehicle. This probability is zero if both the acceleration and the velocity are non-null, and higher than 0 (but < 1 to allow restart, following for instance a stop at a traffic light) if both the acceleration and the velocity are null. This probability is computed at each iteration and allow to end the generated cycle. Applying this methodology on the data coming from each cluster allows us to obtain a velocity profile representative to real driving behavior.

C. Real-World Fuel Consumption Estimation

At this stage, for a given complete trip τ, we have at our disposal: i) the road type repartition, ii) the predominant driving behavior in each road type and iii) the associated representative speed profiles. Based on these information, we are thus able to give an estimation of the fuel consumption for the complete trip c_t^*. For a given road type $r \in \mathcal{R}$, let ω_r be the proportion of the given road type, based on the distance, and $c_{r,k}^*$ be the consumption, in L/100km, of the representative driving cycle of the cluster k on the considered road type. Finally, at this road type r is associated a predominant driving behavior b_r equals to the corresponding cluster label obtained from the previous classification e.g. $b_{\text{URBAN}} = 2$ means that the user has the driving behavior associated to the cluster 2 in the URBAN classification. Hence, the estimated fuel consumption for a complete trip c_t^* can be obtained through a weighted average as follows:

$$c_t^* = \sum_{r \in \mathcal{R}} \sum_{k=1}^{K_r} \omega_r c_{r,k}^* \mathbb{1}(b_r = k) \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes the characteristic function that equal to 1 if its argument is verified and 0 otherwise ans K_r in the number of clusters for the road type r.

Based on the complete proposed strategy detailed in this section, we are thus able to i) identify a set of driving behaviors for each road type, and for each of them, ii) simulate a representative speed profile, iii) used to estimate a driving- and vehicle-based representative fuel consumption. Based on the knowledge of driving behavior, we are thus able to estimate a real-world fuel consumption through a simple weighted average of the appropriate representative fuel consumption.

It is important to note that because our data are restricted to smartphone data, we haven't any ground truth data for the consumption. To overcome this lack of information, the choice was to estimate the consumption from the real velocity profiles thanks to the model presented in [5] and use these consumptions as ground truth. The same model is then used to compute the consumption of the generated representative velocity profiles, consumption which is then combined to recover a real-world consumption according to a desired driving behavior. These real-world consumptions are then compared to what we defined as our ground truth to calculate the estimation errors. In the next section, we evaluated this strategy on the data presented in Section III-A.

TABLE II

MEDIAN OF SILHOUETTE SCORES OBTAINED FOR THE GLOBAL CLASSIFICATION PER ROAD TYPE. MEDIAN SILHOUETTE PER CLUSTER IS ALSO REPORTED FOR EACH ROAD TYPE

Median Silhouette	ZONE$_{30}$	URBAN	χ-URBAN	HIGHWAY
all	0.301	0.279	0.260	0.329
C0	0.298	0.232	0.149	0.282
C1	0.148	0.148	0.265	0.347
C2	0.325	0.320	0.100	0.368
C3	0.332	0.304	0.141	0.173
C4	-	-	0.327	0.287
C5	-	-	0.315	0.421

V. RESULTS

In order to validate the complete strategy previously presented, several kind of results and evaluations are required. We thus first present the classification and the proposed interpretation for identifying driving behavior per road type, in terms of normal, slow, dynamic or traffic jam, for instance. We then demonstrates the link with fuel consumption and these identified driving behaviors. We also validated the corresponding representative driving cycles and their use to estimate real-world fuel consumption.

A. Classification and Cluster Interpretation in Terms of Driving Behavior

In this section, we present classification results obtained for each road types (ZONE$_{30}$, URBAN, χ-URBAN, and HIGHWAY). The classification was performed according to the methodology presented in Section III-C. We note that for each road type, we performed a classification for various number of clusters K and the optimal number of clusters K_r is the one that maximize the median silhouette [38]. This approach allows us to select four clusters for ZONE$_{30}$ and URBAN road types, while the optimal number of clusters for χ-URBAN and HIGHWAY is six. Table II shows best median silhouette results obtained per road type and its declination per clusters.

In the same vein as the work of [39], classification results will then be interpreted in terms of driving behavior. As features used for the classification are derived from velocity and acceleration, driving behaviors are characterized as a combination of driving speed and dynamism with respect to traffic conditions. For this purpose, cluster centroids based on five descriptors (v_{mean}, $v_{\max}$, v_{null}, $\bar{v}_{\text{diff}}$ and RPA_{unweight}) are evaluated. For each of these studied descriptors, clusters are ranked and a final interpretation is given according to a combination of these ranking. As displayed in Fig. 5, normalizing the descriptors between 0 and 1, where 1 is the maximal value of the descriptor, allow us to perform a relative comparison between the clusters.

TABLE III

DRIVING BEHAVIOR IDENTIFICATION PER CLUSTER AND ROAD TYPE. THE BEHAVIOR IS DETERMINED THROUGH THE FEATURES USED FOR THE CLASSIFICATION

Road Type	C0	C1	C2	C3	C4	C5
ZONE$_{30}$	Average	Fast	Slow	Dynamic	-	-
URBAN	Dynamic	Fast	Slow	Average	-	-
χ-URBAN	Traffic jam	Crowded traffic flow	Dynamic	Fast	Average 70	Average 80-90
HIGHWAY	Traffic jam	Average 110	Average 130	Fast	Crowded traffic flow	Dynamic

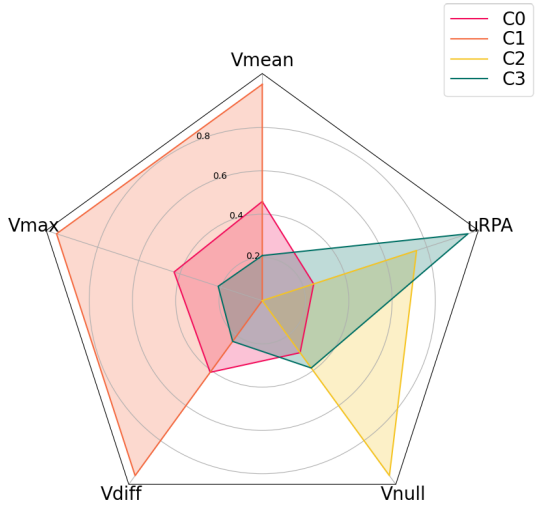


Fig. 5. Radar plot for comparing centroids of four clusters obtained on ZONE$_{30}$. Centroids are restricted to five descriptors (v_{mean}, v_{max}, v_{null}, $\bar{v}_{\text{diff}}$ and RPA_{unweight} (uRPA)) and are normalized between 0 and 1 for a relative comparison between clusters.

Our procedure characterized five main representative driving behaviors: traffic jam, average (representative and limit-followers), fast (speed limit violators), slow (non-aggressive and low speed), dynamic (aggressive). Let us demonstrate our interpretation of classification results obtained for ZONE$_{30}$ road type. In the Fig. 5, Cluster 0 is recovered in the middle for each descriptors with no discriminant direction, with notably a mean speed around 31 km/h (not represented in Fig. 5). Adding the fact that this cluster represent 45% of the road type in terms of distance, this cluster is assigned as *average*. The Cluster 1 has a notable tropism for high speeds as v_{mean}, v_{max} and $\bar{v}_{\text{diff}}$ are maximal (around 43, 68 and 33 km/h respectively). This cluster is thus assign to *fast* behavior. The cluster 2 shows the highest proportion of null velocity with more than 3%; combined to the lowest speed-based descriptors, Cluster 2 is thus assumed as *slow*. Note that in such kind of road type (ZONE$_{30}$), it is difficult to distinguish a real slow behavior of the driver to a slow behavior due to dangerous infrastructure. Finally, we define the last Cluster 3 as *dynamic* as it exhibits speed-based descriptors close to average cluster (limit-follower) but with the highest dynamic criterion RPA_{unweight} (e.g. 0.032m/s^2, against 0.023m/s^2 for the average cluster). This procedure is applied for the three other road types for which results of the interpretation are given in Table III and details about center clusters are given in Appendix B.

To use this methodology on new dataset, the key is to have enough data per behavior to be identified. The number of behavior is obviously not known in advance but can be intuited and we thus recommend at least 5000 trips for a given road type. In addition, the number of trips is not sufficient to take into account but also the distance of the trips. Indeed, a large quantity of very small trips can bias the estimated statistics on them. It is recommended to chose the minimal distance in incoherence with the road type, we thus advice a dataset with trips of at least 1km, 3km, 8km and 30km according to ZONE$_{30}$, URBAN, χ-URBAN and HIGHWAY. Finally, regarding the choice of the road type, we obviously recommend the four road type we choose but the ZONE$_{30}$ and URBAN can be merged according to the purpose of the study and the thinness of the behavior the researchers want to discover.

B. Links Between Driving Behaviors and Fuel Consumption

Based on the classification results, we will first evaluate the statistics regarding the fuel consumption obtained on each cluster and road type, with respect to the interpreted driving behavior. We then present results of the statistical analysis to exhibit the link between driving behavior and fuel consumption.

1) Consumption Analysis With Respect to Driving Behavior: Fuel consumptions per cluster and road type are obtained using the methodology presented in IV-A. As our purpose is to highlight a dependence between driving behavior and fuel consumption, we didn't want to be impacted by the effect of the vehicle. Hence, to avoid this bias at the sub-trip scale, all computed fuel consumptions were estimated with the same vehicle (Volkswagen Golf VII Phase 2, 2018). We notice that the impact of the vehicle will be considered at the trip scale (see Section IV-B). Results are gathered in Table IV and show, for each classes, statistics on fuel consumption through median and mean absolute deviation (MAD).

First of all, the order of magnitude regarding the fuel consumption per road type is coherent with what we can expect. Indeed, based on the effect of the velocity on the fuel consumption, a lower consumption is expected for the intermediate velocity around 70 km/h (χ-URBAN) while a highest one is expected for smaller velocity (ZONE$_{30}$ and URBAN due to low engine efficiency) and higher velocity (HIGHWAY due to predominant aerodynamic impact). Then, for each road type, the link between fuel consumption and driving behavior interpretation also appears coherent. Indeed, for road type related to low and intermediate velocity (i.e. ZONE$_{30}$, URBAN

TABLE IV
MEDIAN AND MEAN ABSOLUTE DEVIATION (MAD) OF REAL CONSUMPTIONS, IN L/100Km, COMPUTED PER ROAD TYPE AND CLUSTERS

Road type		C0	C1	C2	C3	C4	C5
ZONE$_{30}$	Median	6.75	6.63	7.89	7.68	-	-
	MAD	1.14	1.57	1.36	1.30	-	-
URBAN	Median	7.68	6.49	7.31	6.31	-	-
	MAD	1.35	1.19	1.14	0.84	-	-
χ-URBAN	Median	5.80	5.75	5.34	5.87	5.08	5.19
	MAD	0.69	0.88	1.19	0.94	0.52	0.51
HIGHWAY	Median	5.95	5.51	6.15	6.78	4.93	5.45
	MAD	0.47	0.44	0.33	0.51	0.54	0.43

and χ-URBAN), we expect a higher fuel consumption for dynamic and fast behavior compare to the average behavior. For higher velocity (i.e. HIGHWAY), as the velocity factor becomes of first-rate compared to the dynamism, we expect higher consumption for fast behavior while the traffic jam condition (related to highest dynamism) is expected linked to smallest fuel consumption, finally leading to intermediate consumption for average behavior.

2) Statistical Analysis Proves Discriminant Consumptions According to Driving Behaviors: As mentioned in IV-A, for each road type $r \in \mathcal{R}$ and cluster $k = \{1, \ldots, K_r\}$, we are able to obtain $\mathcal{C}_k^{(r)}$, the set of fuel consumptions computed on each sub-trips belonging to the cluster k. Based on these information, the analysis we performed aims at proving that, for a given road type, identified clusters have statistically different consumptions. As clusters are associated to driving behavior, discriminant consumptions across clusters will validate the assumptions that the fuel consumption is dependent on the behavior. Thus, in order to demonstrate this dependence for a given road type, we evaluate whether average fuel consumptions associated to two distinct clusters is equal or not, and repeat this procedure for all pairs of clusters. As the variance of fuel consumption on the different clusters is not equal, multiple comparisons are performed using the pairwise Welch's t-test. To take into account the fact that multiple comparisons are made, an adjustment of the p-value is required and the Benjamini-Hochberg procedure is then applied. We then estimate that clusters have, in average, discriminant consumptions if the adjusted p-value for the given comparison is lower 0.001.

With all p-values lower than 0.001, results regarding the four and six clusters of road type URBAN and HIGHWAY, respectively, indicate that the fuel consumptions are, in average, statistically different between all pairs of clusters, thus suggesting that the associated driving behavior have discriminant fuel consumption. Regarding results on road type ZONE$_{30}$ and χ-URBAN, both show one pair of clusters for which the consumptions can not be assumed discriminant. Indeed, for the two, the statistical analysis indicates that the comparison between cluster 0 and 1 leads to a p-value equals 0.96 and 0.16 for comparison on ZONE$_{30}$ and χ-URBAN, respectively. While the undiscriminating fuel consumption between cluster 0 and 1 in χ-URBAN can be explained by close driving behavior (traffic jam vs crowed traffic flow), the non significant difference between cluster 0 and 1 in ZONE$_{30}$ can be attributed to compensatory effects more than close driving behavior.

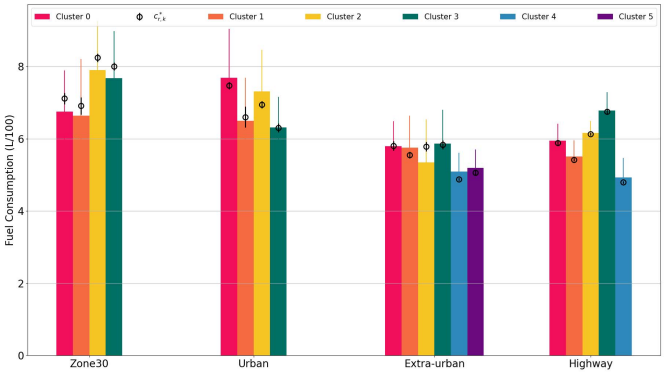


Fig. 6. Median of the real consumption computed on sub-trips per cluster k and road type r. Vertical colored line corresponds to the mean absolute deviation (MAD) for the corresponding set of real consumptions. Black circle markers refers to the median of consumption computed on the generated representative driving cycle $c^*_{r,k}$ and the vertical line the associated MAD over the 25 simulations.

C. Real-World Fuel Consumption Estimation

1) Validation of the Representative Driving Cycle: We first validate the generation of the representative driving cycle. For this purpose, for each road type $r \in \mathcal{R}$ and each cluster $k = \{1, \ldots, K_r\}$, where K_r is the number of cluster associated to the road type r, we first computed, on the real speed sub-trips, the median of the consumptions in $\mathcal{C}_k^{(r)}$ (see Section IV-A). In parallel, thanks to the methodology presented in IV-B, we simulated, for each road type and cluster, 25 realizations of a representative driving cycle and computed the median of the associated consumptions. The figure 6 shows the median of consumption on real sub-trips are then compared to the simulated ones, with respect to the variability obtained in real data and in simulated ones.

We can observe that in presented case, the median of the consumption on simulates representative driving cycle is close to the median of the real consumption and completely included in the variability, in terms of mean absolute deviation (MAD), of the real consumptions. The relative error per road type is in average between 1% and 5%. While the MAD covers a range from 0.33 to 1.57 L/100km for the real consumption, we notice a weak variability of the consumption over the 25 simulated profiles: from 0.04 to 0.29 L/100km for the highest. This results tends to indicate a sufficiently robust process of driving cycle generation. These good results allow us to validate the representativeness of the simulated speed profiles according to driving behavior and road type. We noticed that in addition to a good representativeness in terms of consumptions, we compared two kind of descriptors of the speed profiles to determine whether, beyond consumption, the shape of the speed signal is close. In the same vein that the consumption, we thus evaluate the mean speed and the RPA parameter. We obtained a coefficient of determination of 0.98 and 0.93 for the mean

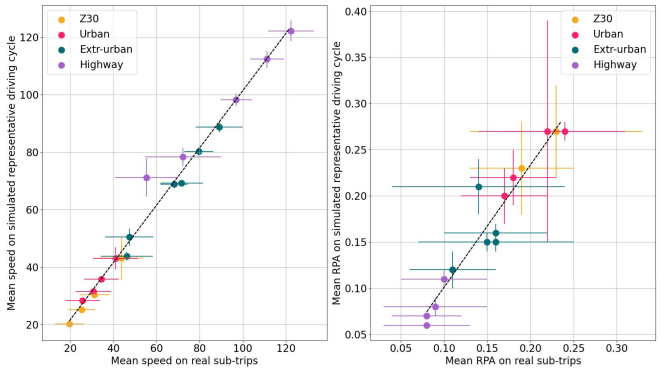


Fig. 7. Parity graph evaluating the correlation of descriptors related to real sub-trips versus simulated representative profiles. Evaluated descriptors are average speed (left) and average RPA (right). Mean and standard deviation are considered for both the real sub-trips vs the 25 simulated representative profiles. Points are colored according to the road type and a linear regression is computed to obtained the coefficient of determination.

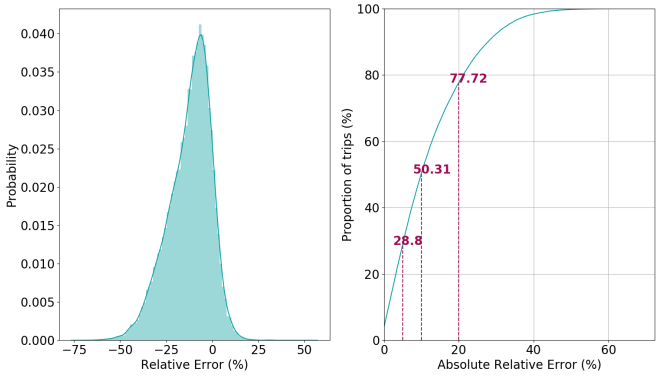


Fig. 8. Histogram of the relative error (left). Cumulative distribution of the absolute relative error (right). Significant absolute relative error thresholds (5, 10 and 20 percent) are highlighted with purple lines and the corresponding proportion of trips satisfying the thresholds are indicated in bold.

speed and the RPA descriptor, respectively (see Fig. 7). This result is based on the comparison of the average values on the real sub-trips on the hand, and the 25 simulated representative profiles on the other hand. Taking now into account the variability of these measures, the standard deviation obtained on the simulated profiles are weaker than the one obtained on real the sub-trips. The standard deviation for the mean speed covers a range from 6.23 to 17.57km/h for the real data while it is significantly lower for the simulated profiles as the range start from 0.64 to 6.68km/h. A similar results is observed on the RPA where the standard deviation associated to the real data varies from 0.04 to 0.08m/s^2 while the one on the simulated data varies from 0 to 0.05m/s^2. We thus observe that these descriptors are relatively stable for the simulated profiles, excepted for the cluster C1 of URBAN road type but for which high variability is also observed on real profiles. These results thus argue for a good correlation between descriptors obtained on real sub-trips and simulated representative driving cycle.

Finally, we can conclude that generated representative speed profiles well simulate both behavioral and consumption characteristics. We thus now present the validation of their use to estimate a real-world fuel consumption.

2) Validation of the Real-World Fuel Consumption Estimation: Let us now place ourselves at a whole trip scale. As previously introduced, this part of our work takes into account an additional variability coming from the vehicle itself. Indeed, at each trip is associated a vehicle, defined by the user. Thus, all computed consumptions were vehicle dependent. As vehicle information are not always sufficiently pertinent for our fuel consumption model, we reduced the number trips to 51004, corresponding to trips for which information regarding the vehicle is useful, and represents 75% of the total number of trips.

Based on the knowledge of its road type repartition and the associated user driving behavior, we estimated the real-world fuel consumption using Equation (1). To validate our approach, we compare this estimated consumption to the real one, computed on the whole velocity profile corresponding to the trip. To perform the comparison, we evaluated the relative error obtained between the real and the estimated ones. In figure 8 (left), the histogram of this error exhibit an asymmetric distribution, more populated for negative errors, thus suggesting an under-estimation of the fuel consumption. It is however important to note that an under-estimation of the fuel consumption estimation given by the vehicle model was observed in a previous study [5]. It is thus difficult to fully explain the resulting relative error by a potential weakness in the quality of the generated velocity profiles. Although the error distribution is skewed, its mode is around 10%, which is close to the median of the relative errors (-10.45 %). The cumulative distribution of this absolute error is given in Fig. 8 (right).

Our proposed methodology allows us to estimate the real-world fuel consumption with an absolute relative error less than 10 percent for half of the studied trips. Each road type is represented by a predominant behavior, and we found across all the trips we studied 818 combinations of driving behavior according to road types. Among these 818 combinations, the 50% of the trips recovered with less than 10% of absolute relative error represents 30% of the possible combinations. In addition, about 30% of the trips are estimated with an error below 5% while 78% of the trips covers an error less than 20%. This 78% of trips covers 70% of our possible combinations of driving behavior over various road type. In other words, we are able to estimate with less than 20% of error, 70% of possible combinations we have in our dataset. The remaining 30% of driving behavior combinations cover about 12 000 trips for which half of the trips are associated to at least one extreme driving behavior. Note that extreme behavior are identified by analyzing clustering results. For each cluster, each datum are compared to the center, and we defined as extreme the datum having a distance to the center higher than the median of the cluster plus or minus the mean absolute deviation of the cluster. As the center of the cluster is used to define the driving behavior, extreme datum are then associated to extreme driving behavior (but not characterized). As representative driving cycles we generated represent the most probable velocity profile for a given driving behavior, we fail to depict

TABLE V

DEFINITION OF STATISTICAL FEATURES USED AS INPUT OF THE UNSUPERVISED CLASSIFICATION. D_{TOT}: TOTAL DISTANCE. T_{TOT}: TOTAL TIME. T_{NULL}: TIME SPENT AT ZERO SPEED. S: NUMBER OF SECTIONS IN THE SUB-TRIP. v_s^{LIM}: SPEED LIMIT ON SECTION s. I: LENGTH OF a. $\lambda = \text{MEDIAN}\ a + \text{MAD}\ a$: THRESHOLD TO DEFINE SIGNIFICANT ACCELERATION

Signal	Feature	Definition
Velocity	average speed	$v_{\text{mean}} = \frac{D_{\text{tot}}}{T_{\text{tot}}}$
	maximal speed	$v_{\text{max}} = \max v$
	percentage of null speed	$v_{\text{null}} = \frac{T_{\text{null}}}{T_{\text{tot}}} * 100$
Velocity sections	average maximal speed	$\overline{v}_{\text{max}} = \frac{1}{S}\sum_{s=1}^{S} \max v_s$
	average (max-limit) speed	$\overline{v}_{\text{diff}} = \frac{1}{S}\sum_{s=1}^{S} \max v_s - v_s^{\text{lim}}$
Acceleration	average acceleration	$a_{\text{mean}} = \frac{1}{T_{\text{tot}}}\sum_i a_i$
	maximal acceleration	$a_{\text{max}} = \max a$
	Hoyer	$H = \left(\sqrt{I} - \frac{\sum_i a_i}{\sqrt{\sum_i a_i^2}}\right)\left(\sqrt{I} - 1\right)^{-1}$
	Gini	$G = 1 - 2\sum_i \frac{a_{(i)}}{\|\|a\|\|_1}\left(\frac{I-i+0.5}{I}\right)$ for ordered data $_{(1)} < a_{(2)} < \ldots < a_{(I)}$
	RPA	$RPA = \frac{1}{D_{\text{tot}}}\sum_i \begin{cases} a_i v_i & \text{if } a_i > 0 \\ 0 & \text{otherwise} \end{cases}$
	unweighted RPA	$RPA_{\text{unweight}} = \frac{1}{D_{\text{tot}}}\sum_i \begin{cases} a_i & \text{if } a_i > 0 \\ 0 & \text{otherwise} \end{cases}$
	number of significant acceleration per km	$A_{\text{km}} = \sum_i \mathbb{1}(a_i > \lambda \text{ and } a_{i-1} \leq \lambda)$

extreme ones, leading to a higher fuel consumption error in the real-world fuel consumption estimation. We also observed that extreme driving behaviors are mainly associated to URBAN area and especially the Cluster 2. This results needs to be weighted as this cluster is also the one for which the median of fuel consumptions for representative driving cycles exhibit the highest difference with the real one obtained in the same cluster. We also remarked that a very few number of trips contains cluster associated to HIGHWAY road type, suggesting a good estimation of the fuel consumption on this specific road type.

Moreover, restricting our analysis to trips containing more than 90% of one road type, we confirmed the previous results as fuel consumption error given by our approach seems to be dependent on the road type, and especially the variability of driving behavior encountered on these road type. Indeed, in HIGHWAY road type, for which the variability of driving behavior is low, the representative driving cycle allows to capture a large part of the driving behaviors as we well estimated 95.56% of the real-world fuel consumptions with an absolute relative error lower than 20%. Inversely, as show in Fig. 9, for the URBAN and ZONE$_{30}$ areas, where the variability of the driving behavior is high, our approach may encounters weakness in correctly capturing deviant behaviors. There again, this weakness is completely

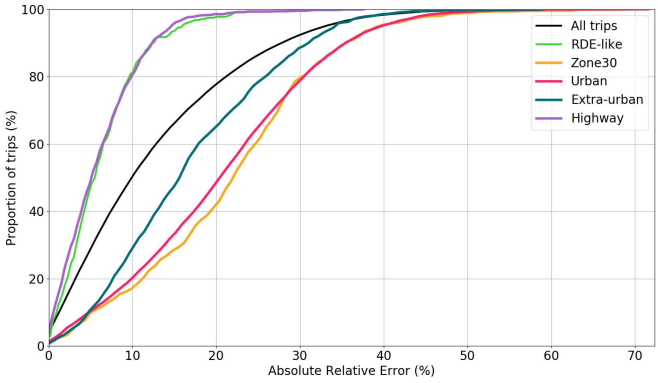


Fig. 9. Cumulative distribution of the absolute relative error for all trips and for each road type, for the subset of trips containing more than 90% of this specific road type. RDE-like trips are also recorded.

inherent to the representative driving cycle generated as the most probable driving cycle which, *de facto*, rules out deviant behaviors.

In the same vein, we performed an analysis on RDE-like trips, by selected trips for each the repartition of (ZONE$_{30}$ + URBAN), χ-URBAN and HIGHWAY are all three between 23% and 40% i.e with a road type repartition corresponding to the Real Drive Emissions standard trip eligibility. Results obtained

TABLE VI

CENTROIDS OF FIVE DESCRIPTORS (v_{MEAN}, v_{MAX}, v_{NULL}, $\bar{v}_{\text{DIFF}}$ AND RPA_{UNWEIGHT}) FOR ALL CLUSTERS AND ROAD TYPE

Road type	Cluster	v_{mean}	v_{max}	v_{null}	$\bar{v}_{\text{diff}}$	RPA_{unweight}
ZONE$_{30}$	C0	31.2	52.1	1.37	18.3	0.023
	C1	43.6	68.0	0.62	32.9	0.020
	C2	19.6	40.4	3.14	8.1	0.029
	C3	25.3	45.9	1.59	13.9	0.032
URBAN	C0	30.7	59.5	1.94	7.2	0.51
	C1	41.0	83.5	1.34	22.7	0.21
	C2	25.7	52.0	2.64	-0.8	0.24
	C3	34.5	66.5	1.69	9.5	0.14
χ-URBAN	C0	47.4	98.3	1.23	4.9	0.07
	C1	46.3	71.5	0.90	-10.3	0.16
	C2	71.6	86.9	0.08	5.1	0.20
	C3	89.0	118.8	0.06	24.4	0.08
	C4	68.1	87.2	0.11	0.6	0.06
	C5	79.6	101.2	0.06	9.1	0.04
HIGHWAY	C0	55.4	120.9	0.78	-5.7	0.03
	C1	96.9	114.1	0.02	-4.0	0.03
	C2	111.2	131.9	0.02	4.8	0.01
	C3	122.3	149.2	0.01	19.5	0.02
	C4	72.2	93.1	0.25	-24.6	0.04
	C5	104.7	112.8	0.00	0.6	0.04

on these trips are promising as about 80% of these trips have a real-world fuel consumption estimated with less than 10% of error (see Fig. 9).

Finally, although our approach encounters some weakness for deviant behaviors, we are able to well estimate the fuel consumption for at least half of our tested population (resp. 78%) with less than 10% (resp. 20%) of error, based on road type repartition and driving behavior only, and thus without recorded any GPS traces. Complementary, we well capture 70% of the driving behaviors present in our population with less than 20% of error.

VI. CONCLUSION AND PERSPECTIVES

This paper describes a original unsupervised methodology for i) identifying, according to road types, trip- and driver-related behaviors from smartphone GPS data, ii) generating a representative velocity profile for all the identified behaviors and road types iii) computing real-world fuel consumptions based on real-world combination of driving behaviors and road types. Such a combinations can be easily accessible through a questionnaire that would be given to the user, allowing us to estimate the real-world fuel consumption only based on basic information and thus without any novel recorded data. An example is the application *"je change ma voiture"* realized for the french Ministry for the Ecological Transition and used to estimate the annual fuel consumption for a user specific vehicle and habits (https://jechangemavoiture.gouv.fr/jcmv/).

Based on a collection of 67 689 recorded trips covered more than 1 million of kilometers, 20 driving behaviors (normal, slow, fast, dynamic, traffic jam, crowded traffic jam, etc) are identified over four road types (30km/h zone, urban, extra-urban and highway). We estimate the real-world fuel consumption associated to this recorded trips with an error

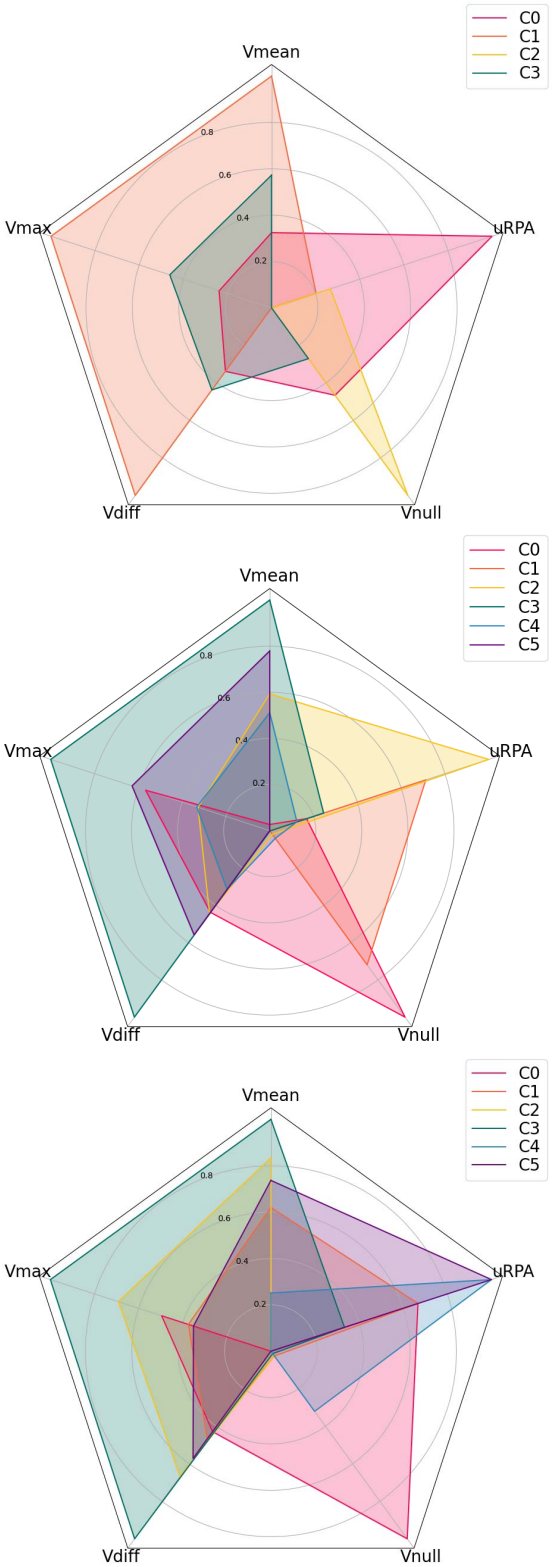


Fig. 10. Radar plot for comparing centroids of clusters obtained on URBAN (top), χ-URBAN (middle) and HIGHWAY (bottom). Centroids are restricted to five descriptors (v_{mean}, v_{max}, v_{null}, $\bar{v}_{\text{diff}}$ and RPA_{unweight}) and are normalized between 0 and 1 for a relative comparison between clusters.

lower than 20% for about 80% of the trips. We better estimate real-word fuel consumptions on road types having less variability and deviant behaviors as in highway than in urban environments.

As a future work, we would refine the classification task, regarding both the chosen features and the algorithm, to be able to better capture deviant behaviors in order to enlarge the span of well estimated real-world fuel consumption. In complement to fuel consumption, impact of driving behavior could also be analyzed from a pollutant emissions aspect.

APPENDIX A
FEATURES USED FOR THE UNSUPERVISED CLASSIFICATION

We present in Table V the definition of features introduced in III-B, ans used as input for the unsupervised classification. We recall that these features are computed for each road type $r \in \mathcal{R}$ and that we simplify the notation by dropping the r exponent, but all signals and calculus are related to sub-trips e.g. the total distance refers to the total distance of the sub-trip related to a road type.

APPENDIX B
CLUSTERS INTERPRETATION

We present in Table VI, center values of five main descriptors (v_{mean}, $v_{\max}$, v_{null}, $\bar{v}_{\text{diff}}$ and RPA_{unweight}) obtained after the classification and used to interpret clusters in terms of driving behaviors. We also provide in Fig. 10 radar plot as in Fig. 5, for road type URBAN, χ-URBAN and HIGHWAY.

REFERENCES

[1] U. Tietge, S. Díaz, P. Mock, A. Bandivadekar, J. Dornoff, and N. Ligterink, "From laboratory to road. A 2018 update of official and 'real-world' fuel consumption and CO_2 values for passenger cars in Europe," Int. Council Clean Transp. (ICCT), Berlin, Germany, White Paper, 2019. [Online]. Available: https://theicct.org/sites/default/files/publications/Lab_to_Road_2018_fv_20190110.pdf

[2] G. Fontaras, N.-G. Zacharof, and B. Ciuffo, "Fuel consumption and CO_2 emissions from passenger cars in Europe—Laboratory versus real-world emissions," *Prog. Energy Combustion Sci.*, vol. 60, pp. 97–131, May 2017.

[3] L. Guzzella and A. Sciarretta, *Vehicle Propulsion Systems*. Berlin, Germany: Springer, 2013.

[4] S. Tsiakmakis, G. Fontaras, B. Ciuffo, and Z. Samaras, "A simulation-based methodology for quantifying European passenger car fleet CO_2 emissions," *Appl. Energy*, vol. 199, pp. 447–465, Aug. 2017.

[5] P. Michel, A. Pirayre, S. S. Rodriguez, and A. Chasse, "From trips database to real-world fuel consumption. Model and large-scale simulation framework," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Rhodes, Greece, Sep. 2020, pp. 1–6.

[6] X. Hu, Y.-C. Chiu, Y.-L. Ma, and L. Zhu, "Studying driving risk factors using multi-source mobile computing data," *Int. J. Transp. Sci. Technol.*, vol. 4, no. 3, pp. 295–312, Sep. 2015.

[7] E. Barmpounakis and N. Geroliminis, "On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 111, pp. 50–71, Feb. 2020.

[8] T. Lajunen and T. Özkan, "Self-report instruments and methods," in *Handbook of Traffic Psychology*, B. E. Porter, Ed. San Diego, CA, USA: Academic Press, 2011, ch. 4, pp. 43–59.

[9] L. Thibault, P. Dégeilh, G. Sabiron, L. Voise, K. Thanabalasingam, and G. Corde, "Sensorless estimation of real-driving emissions from GPS data: An innovative approach allowing large scale measurement campaigns," in *The Future of Road Mobility Research With Impact*. Brussels, Belgium: FORM Forum, 2018, p. 51.

[10] E. Mantouka, E. Barmpounakis, E. Vlahogianni, and J. Golias, "Smartphone sensing for understanding driving behavior: Current practice and challenges," *Int. J. Transp. Sci. Technol.*, vol. 10, no. 3, pp. 266–282, Sep. 2021.

[11] J. F. Júnior, "Driver behavior profiling: An investigation with different smartphone sensors and machine learning," *PLoS ONE*, vol. 12, no. 4, pp. 1–16, 2017.

[12] B. Ciuffo, M. Makridis, T. Toledo, and G. Fontaras, "Capability of current car-following models to reproduce vehicle free-flow acceleration dynamics," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3594–3603, Nov. 2018.

[13] K. Fadhloun and H. Rakha, "A novel vehicle dynamics and human behavior car-following model: Model development and preliminary testing," *Int. J. Transp. Sci. Technol.*, vol. 9, no. 1, pp. 14–28, Mar. 2020.

[14] T. Toledo, H. N. Koutsopoulos, and M. Ben-Akiva, "Integrated driving behavior modeling," *Transp. Res. C, Emerg. Technol.*, vol. 15, no. 2, pp. 96–112, 2007.

[15] T. K. Chan, C. S. Chin, H. Chen, and X. Zhong, "A comprehensive review of driver behavior analysis utilizing smartphones," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4444–4475, Oct. 2020.

[16] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1609–1615.

[17] J. Engelbrecht, M. J. Booysen, G.-J. Van Rooyen, and F. J. Bruwer, "Performance comparison of dynamic time warping (DTW) and a maximum likelihood (ML) classifier in measuring driver behavior with smartphones," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 427–433.

[18] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 234–239.

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

[20] H. R. Eftekhari and M. Ghatee, "A similarity-based neuro-fuzzy modeling for driving behavior recognition applying fusion of smartphone sensors," *J. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 72–83, Jan. 2019.

[21] M. M. Bejani and M. Ghatee, "A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 303–320, Apr. 2018.

[22] J.-H. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2014, pp. 4047–4056.

[23] M. H. Tawfeek and K. El-Basyouny, "A context identification layer to the reasoning subsystem of context-aware driver assistance systems based on proximity to intersections," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102703.

[24] R. Araújo, Â. Igreja, R. de Castro, and R. E. Araujo, "Driving coach: A smartphone application to evaluate driving efficient patterns," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 1005–1010.

[25] J. E. Meseguer, C. K. Toh, C. T. Calafate, J. C. Cano, and P. Manzoni, "Drivingstyles: A mobile platform for driving styles and fuel consumption characterization," *J. Commun. Netw.*, vol. 19, no. 2, pp. 162–168, Apr. 2017.

[26] Y. Yao *et al.*, "Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones," *J. Adv. Transp.*, vol. 2020, pp. 1–11, Mar. 2020.

[27] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (ACM SIGSPATIAL)*, Nov. 2009, pp. 336–343.

[28] T. J. Barlow, "A reference book of driving cycles for use in the measurement of road vehicle emissions: Version 3," TRL Transp. Res. Lab., Wokingham, U.K., 2009.

[29] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[30] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010.

[31] H. Steinhaus, "Sur la division des corps matériels en parties," *Bull. Acad. Polon. Sci.*, vol. 4, no. 12, pp. 801–804, 1956.

[32] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, vol. 1, 1967, pp. 281–297.

[33] S. Kullback, *Information Theory and Statistics*. Gloucester, MA, USA: Peter Smith, 1978.

[34] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[35] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, nos. 1–2, pp. 28–35, 1947.

[36] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. B, Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[37] T. Leroy, J. Malaizé, and G. Corde, "Towards real-time optimal energy management of HEV powertrains using stochastic dynamic programming," in *Proc. IEEE Vehicle Power Propuls. Conf.*, Oct. 2012, pp. 383–388.
[38] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[39] J. Warren, J. Lipkowitz, and V. Sokolov, "Clusters of driving behavior from observational smartphone data," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 3, pp. 171–180, 2019.

Aurélie Pirayre graduated from the Institut Supérieur des BioSciences de Paris in 2013. She received the Ph.D. degree from the Université of Paris-Est in 2017 for her thesis on biological network modeling. She currently holds the position of Data Scientist at IFP Energies nouvelles. Initially interested in graph theory and optimization, her current research interests include oriented toward data science and notably machine learning aspects for multiple applications, such as connected mobility, chemistry, and bio-informatics.

Pierre Michel received the Ph.D. degree from the Université d'Orléans in 2015 for his thesis work on hybrid electric vehicle energy management. After his post-doctoral position at Argonne National Laboratory on connected and automated vehicles, he participated in the HEV powertrains development at PSA Peugeot Citroën. He joined IFP Energies nouvelles in 2019, and his research focuses on innovative fuel and powertrain modeling, eco-driving algorithms development, and more generally mobility analysis by using control theory and data science.

Sol Selene Rodriguez received the M.S. degree (as a General Engineer) from the École Centrale Nantes, France, as part of a double degree program in mechanics with her university UASLP, Mexico, in 2017, and the M.S. degree in energy and powertrains from the IFP School, France, in 2019. She currently makes part of the IFP Energies nouvelles as a Research Engineer working on vehicle dynamics and pollutant emissions modeling and data analysis for connected mobility.

Alexandre Chasse received the Engineering degree from the École Supérieure d'Electricité in 2005 and the École Nationale du Pétrole et des Moteurs in 2007. He is currently a Research and Innovation Project Manager at IFP Energies nouvelles. His research work focuses on the development of services to reduce the energy footprint of mobility. He is also working on the implementation of decision support tools for local authorities to monitor the evolution of mobility using data collected by mobile applications via crowd-sensing.