Determination of the electric vehicles driving modes in real life conditions by classification methods

Mohamed Ben-Marzouk, Guy Clerc, Serge Pelissier, Ali Sari, Pascal Venet

Determination of the electric vehicles driving modes in real life conditions by classification methods

Mohamed Ben-Marzouk[1,2,*], Guy Clerc[1], Serge Pelissier[2], Ali Sari[1], Pascal Venet[1]
[1]Université de Lyon, Ampère (CNRS UMR 5005, Ecole Centrale de Lyon, INSA-Lyon, Université Claude Bernard Lyon 1), F-69100, Villeurbanne, France.
[2]Université de Lyon, IFSTTAR, AME, LTE, 69500 Bron, France
*mohamed.ben-marzouk@univ-lyon1.fr

***Abstract*— In order to study the aging of batteries in automotive applications, it is important to understand how and under what conditions these batteries are operating in electric vehicles (EVs).**

However, because of the specificities of EVs, these uses may be very different from those known from internal combustion engine cars (ICE).

In this paper, we present how, from real-life data, we determine the different driving modes of electric vehicles. The paper presents the different techniques adopted to analyze and classify the data and the different EV running modes, which are obtained.

Keywords— Electric vehicle; life-size experimentation; driving modes; modes of use; classification; batteries

I. INTRODUCTION

Our planet is more and more suffering from the effects of pollution. During the last decades, it has reached records, and takes all forms: pollution of soil, water, air, etc. All of these increase the health and environmental risks.

Air pollution is the most difficult to control. It is caused mainly by agriculture, industrial waste, homes and transportation. In the context of this air pollution, the most difficult source to manage is transportation because it is a source of mobile pollution. In fact, according to CITEPA in France, the transport sector is responsible for the emission of nearly 15% of PM10 (particles with a diameter of less than 10 micrometers) and more than 50% of carbon soot [1]. Thanks to people's awareness of environmental risks and the threats of fossil fuel depletion, the electric vehicle (EV) is now booming with a number of registrations that are constantly increasing (up 23 % between 2015 and 2016) [2].

Despite the policy of countries to reduce pollution in cities by encouraging the purchase of EVs, they remain nonetheless widespread. This is due to the limitations of electric vehicles that are mainly related to batteries problems. It is indeed their weak autonomy, their relatively high prices (400 € / kWh) which can reach 40% of the value of the car and their limited lifetime.

The problem of autonomy seems to be in the process of being resolved since the new models of EVs display a much greater autonomy than that of the models of a few years ago. The high price of batteries can be mitigated by massive production. However, the problem of the lifetime is still relevant. Indeed, mastering the lifespan of storage systems can save on the cost of replacement at the end of life and also on the right energy dimensioning of the system.

There are several research projects around the world dealing with the aging of lithium batteries. However, most of these works only take into account calendar aging or cycling aging.

Yet, in real-life automotive application, lithium batteries are facing an alternation between these types of aging.

Currently, lifetime predictions are based on models and results from standardized cycling tests, with accelerated aging profiles consisting of partial or total charge / discharge, constant currents, or simplified profiles inspired by real uses but which consist of simple impulses and only partially represent the diversity of uses.

To study the problems related to battery lifetime in automotive applications, many authors count on simulations with standardized cycles such as NEDC (New European Driving Cycle) and WLTC (Worldwide Harmonized Light Vehicle Test Cycles) [3], [4]. The results found are often empirical and cannot be very precise. Electric vehicles and their operation depend on many specific factors and parameters. They depend, among other things, on as temperature, SoC, etc.

It is for these reasons that it will be more valuable to seek the different EVs modes of use, starting from real life batteries data recording and their operating conditions, in order to determine their effects on aging later.

In this paper, we will present the methodology followed to search for the different electric vehicles modes of use. We will start with a presentation of the used database. Then we explain the different steps performed to obtain the modes of use.

II. DATABASE DESCRIPTION

This study is based on data extracted from ten EVs. These data are provided by the LTE-IFSTTAR laboratory and come from the CROME project (CROss-border Mobility for EVs). On this project, several French and German partners worked to

design, realize and analyze an electric mobility system between France and Germany [5]. This project has enabled LTE-IFSTTAR to acquire a very large database of EVs in real life application.

In fact, these EVs, having the same architecture and belonging to private and professional volunteers, have been equipped with data loggers that record and transmit via GSM a large number (more than 500) of variables related to the operation (speed, brake use, acceleration rate, etc.), data related to the batteries (current, voltage, cell temperature, etc.) and other related to the engine (torque, engine temperature, etc.).

For each vehicle, the follow-up during a couple of years has allowed to collect thousands of files. Each file represents one use of the EV, in running or in charging conditions. In this file, we find a follow-up of the evolutions of all the variables related to EV over time with a high frequency (0.01s).

Each EV has a battery pack of 16 kWh with a nominal voltage of 325.6 V. These battery packs are made of 88 LEV-50 (LMO) cells mounted in series. Each one has a capacity of 50Ah and a nominal voltage of 3.7V.

As mentioned earlier, we have hundreds of variables to study. To determine modes of use, we are convinced that some variables are more valuable than others. It is for this reason that we will proceed to a selection of parameters that will be described in the following part.

III. SELECTION OF SIGNIFICANT VARIABLES

To select the most significant variables without losing information, we proceeded as follows. At the beginning, we started with the elimination of all the variables that have no relation with the battery usage modes like the temperature of the engine cooling water. So, we kept all the variables related to the batteries (current, voltage, etc.) and the conditions of use of the VE (speed, temperature, etc.).

We have drastically reduced the number of variables to track, going from more than 500 to 140 variables. Nevertheless, the number remains very large and we must still try to reduce it without loss of information.

A. Correlation study

The main purpose of this manipulation is to remove redundancies and keep only one variable on a correlated set. In this part, we have calculated and compared the correlation coefficients between the variables.

According to S. Tufféry [6], even if the variables are continuous, "it is always interesting to compare the two coefficients of Pearson (linear correlation) and Spearman (rank correlation), the most reliable of both is the second", in particular to detect non-linear links.

In the following paragraphs, we will present these two correlation coefficients and the results obtained.

1) Linear correlation coefficient of Pearson "r"

This coefficient is used to study if there is a linear relationship between two variables X and Y. This coefficient is obtained by dividing the covariance between X and Y (σ_{xy}) by the product of their standard deviation, respectively σ_x and σ_y.

$$r_{(X,Y)} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}}} \quad (1)$$

Where n is the observation number, $\overline{X}$ is the mean of X and $\overline{Y}$ is the mean of Y.

2) Coefficient of correlation of Spearman "ρ"

We also used Spearman's correlation coefficient, which allows us to identify a monotonic connection between the variables, whether linear or not. This technique studies the relationship between the ranks of 2 variables. In other words, we do not use the values of the observations in the calculations but we use their ranks. To calculate the Spearman coefficient of X and Y we can use the equation (1) replacing X by $rank_X$ and Y by $rank_Y$. If $X = (10;13;12;22;25;1)$, $rank_X = (5;3;4;2;1;6)$

This coefficient can also be given by the following equation.

$$\rho_{(X,Y)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \quad (2)$$

Where n is the number of observations and d_i is the difference between the $rank_{Xi}$ and $rank_{Yi}$.

3) Correlation conditions

We have found that there are several correlated variables. We subsequently considered that two variables (for example X and Y) are highly correlated if they follow the next condition.

$$(r_{(X,Y)} > 0.75 \ \ AND \ \ \rho_{(X,Y)} > 0.8) \quad (3)$$

For the rest of our study, we choose only one variable among a correlated set. And to properly choose the variables to keep, we computed and compared the amounts of information contained in the variables using Shannon's entropy [7].

B. Features selection

1) Shannon Entropy

It is a mathematical function that can measure the amount of information in a signal. Indeed, the larger the entropy of a signal, the more it contains non-redundant information [7]. The entropy value for each variables is given by the following equation (4).

$$H(X) = -\sum_{i=1}^{n} P_i \log_2(P_i) \quad (4)$$

X being a signal of m individuals, containing n different symbols with ($n < m$). P_i is the probability of the appearance of each symbol i with ($i \in n$).

For example $X = (a;b;a;c;a;b;d)$, so $m=7$, $n=4$ and $P = \left(\frac{3}{7};\frac{2}{7};\frac{1}{7};\frac{1}{7}\right)$,. then $H(X)=1.8424$

Before applying this function, the data have to been normalized. The normalization of each variable vector is calculated by equation (5). We call $Xnorm$ the normalization vector of X.

$$Xnorm = \frac{X - min(X)}{max(X) - min(X)} \quad (5)$$

For each variable, the maximum value takes 1 and the minimum value takes 0. We keep only 4 decimals for *Xnorm*.

When we have a correlated set of variables, we take only the one that has the greatest entropy.

2) Result of the study of the correlation

This followed methodology allowed us to reduce the number of variables to study to 18 non-correlated or weakly correlated input variables. We can find the list of selected variables and a brief description of each one in Table 1.

TABLE I. SELECTED VARIABLES

Variable name	Description	Unit
Soc_init	State of charge at the beginning of driving cycle	%
Max charge current	Maximum current in recovery energy phase	A
Average charge Power	Average recuperated power	W
Average decel	Average deceleration	$m.s^{-2}$
Qperkm	Charge quantity per km	Ah/km
Qchar	Recuperated charge	Ah
Dist	Distance	km
Q disch	Consumed charge	Ah
Average speed	Average speed	km/h
Max speed	Maximum speed	km/h
Max discharge current	Maximum discharge current	A
Effective current	Effective current	A
Average discharge power	Average power when discharging	W
Relative positive acceleration	Relative Positive Acceleration (RPA)	$m.s^{-2}$
Average acceleration	Average acceleration	$m.s^{-2}$
Positive kinetic energy	Positive acceleration Kinetic Energy (PKE)	$m.s^{-2}$
Q auxiliaries	Charge consumed by auxiliaries	Ah
Ambient temperature	Ambient temperature	°C

PKE and RPA are parameters related to the eco-driving and that can also be related to the aggressiveness of driving [8], [9].

All of these variables summarize the electrical parameters of batteries, how and under which conditions they are used. Nevertheless, this number of variables is still very large to make a good classification able to determine the modes of use.

To deal with this problem, we used the Laplacian score for feature selection technique which allows to select the most significant variables.

3) Laplacian score for feature selection

a) Definition and algorithm

Overall feature selection methods, we can distinguish two main categories. The "wrapper" methods and "filter" methods. The wrapper techniques evaluate the features using the learning algorithm that will ultimately be employed. Most of the feature selection methods are wrapper methods. Algorithms based on the filter model examine intrinsic properties of the data to evaluate the features prior to the learning tasks. The filter based approaches almost always rely on the class labels, most commonly assessing correlations between features and the class label [10].

The Laplacian score (LS) is an unsupervised method of selecting significant variables based on the Laplacian Eigenmaps and Locality Preserving Projection. For a set of variables, we measure and compare the LS for each variables and the variables with the highest scores are the most significant.

The application of this method to our variables gave us the results shown in the Fig. 1. This technique is well described in the literature [10], [11].

b) Results

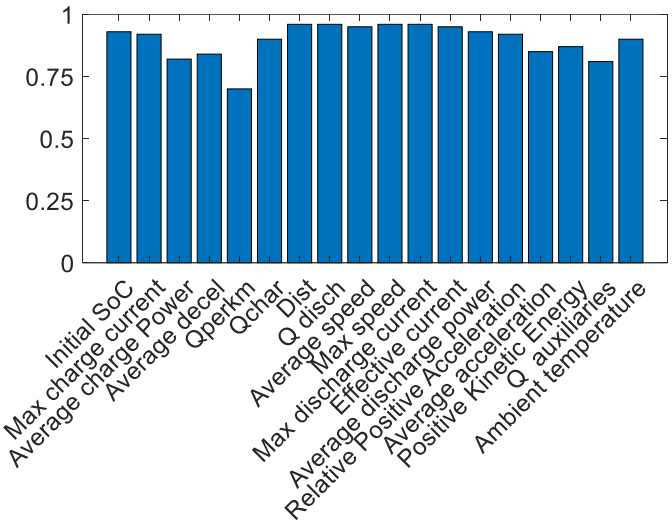

Figure 1 : Laplacian score for each variable

The application of the Laplacian Score method cannot help us reduce the number of variables because as shown in the Fig. 1 we have obtained high and very close scores for all the variables. We have also compared the Laplacian scores results with those obtained by the Infinite feature selection technique [12]. The conclusion is the same; all the scores are very close. When all the variables have high scores, we cannot eliminate anyone of them for the risk of losing information. We then decided to work on the 18 variables together, considering them as the most significant variables.

IV. DIMENSION REDUCTION

Since the number of variables is still high, we applied the principal component analysis (PCA) method in order to reduce the dimensions by projecting the variables on the PCA axes.

A. *Definitions*

Principal component analysis is a fundamental method in multidimensional descriptive statistics. It allows simultaneous processing of any number of variables. The aim of the PCA is to project data on a small space dimension by distorting the reality as little as possible [13].

The PCA projects data on orthogonal axes that means it is a transformation of variables that can be correlated, to new uncorrelated variables. This process not only reduces the number of variables but also makes the information less redundant.

The choice of the number of components is a very important step in PCA. It depends mainly on the quality of the projection of the observations and the variables, on PCA axes.

B. *Choice of components number*

To choose the right number of components to keep, there are several adapted practices. We mention here the elbow criterion (break of slope) and the Kaiser[14] law that allow to keep only the axes that have an inertia higher or equal to the average inertia.[15]

The Fig. 2 shows the inertia for each principal component.

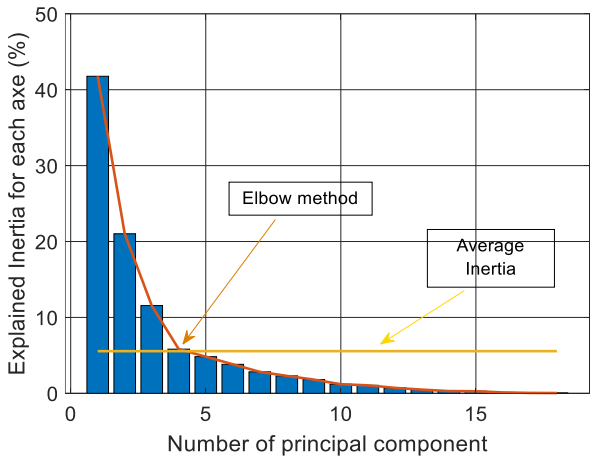

Figure 2: Inertia for each principal component

In our case, following the two criteria mentioned above, we can choose 4 axes of the PCA that explaining 80% of the total inertia (sum of inertia of the 4 axes). However, it is still necessary to check if at only 4 axes, the quality of the projection of the variables is acceptable. The quality of variables projection is calculated as follows [6], [16]:

$$QV_n = \sum_{i=1}^{n} corr^2(V, Ax_i) \qquad (6)$$

Where QV_n is the projection quality of the variable "V" on n PCA axes. $corr^2$ is the square of the Pearson correlation. V is the data vector of the variable called "V". Ax_i is the data vector of the i^{th} PCA axe.

The Table II illustrates the variables projection quality in relation to the number of principal components. The red color is for bad quality and the green color is for a good projection quality.

TABLE II. NUMBER OF COMPONENTS EFFECT ON THE QUALITY OF VARIABLES PROJECTION

	Number of axes						
	1	2	3	4	5	6	7
Initial SoC	1%	1%	1%	95%	95%	99%	99%
Max charge current	43%	51%	53%	53%	68%	75%	77%
Average charge Power	47%	66%	69%	70%	70%	75%	75%
Average decel	0%	84%	84%	84%	87%	94%	94%
Qperkm	1%	3%	74%	74%	78%	80%	89%
Qchar	52%	52%	60%	60%	76%	84%	90%
Distance	61%	86%	86%	88%	88%	94%	94%
Q disch	68%	87%	88%	90%	91%	97%	97%
Average speed	74%	79%	81%	81%	96%	96%	97%
Max speed	85%	86%	87%	88%	88%	89%	89%
Max discharge current	66%	68%	73%	75%	81%	83%	83%
Effective current	81%	92%	94%	94%	95%	96%	97%
Average discharge power	80%	83%	85%	85%	94%	94%	97%
Relative Positive Acceleration	41%	91%	92%	92%	96%	96%	96%
Average acceleration	0%	88%	88%	89%	89%	94%	94%
Positive Kinetic Energy	9%	79%	82%	83%	92%	92%	92%
Q auxiliaries	17%	25%	75%	76%	77%	81%	83%
Ambient temperature	1%	1%	57%	59%	67%	67%	99%

The Kaiser and elbow criterion recommend the choice of 4 principal components for dimension reduction. However, according to Table II, there are some variables that are not well projected. Therefore, to ensure that all the variables are well projected on the axes of the PCA, we chose to keep 7 principal components from 18 (number of variables). The projection quality of all the variables is greater than 75%.

This approach has allowed us to reduce the dimensions and eliminate redundancies which will lead to a better identification of the operating modes.

V. IDENTIFICATION OF THE DRIVING CYCLES MODES

This part focuses on the running modes identification.

In fact, we have several thousands of different driving cycles, we call it also runs. We are interested in identifying the similarities and non-similarities between these runs. We seek to obtain, later, several sets of driving cycles. Each set contains a large number of runs that are very similar and at the same time are very different from another set of runs. So we have tested some classification techniques.

A. *Unsupervised Classification techniques*

The classification methods can be divided into two main families:

- supervised classification where we have already classified elements and where we are interested in adding new elements,

-unsupervised classification where we do not have prior knowledge of classes.

In our case, we need an unsupervised classification method to identify VE usage patterns.

Unsupervised Classification is a data-processing technique that seeks to classify a data set by minimizing intra-class distance and maximizing the inter-class distance as much as possible. There are many distance measurement techniques [17] like Manhattan distance and Euclidian distance. In this part, we used the squared Euclidian distance.

There are several methods of unsupervised classification (called also clustering methods), the most well-known ones are the hierarchical ascending classification (HAC) and the k-means [6].

1) Hierarchical Ascending Classification (HAC)

The HAC regroups iteratively the individuals by aggregating 2 by 2, the closest elements that allows to build progressively a dendrogram (tree diagram) that regroups at the end all the individuals in a single class. In Fig. 3, we present an example that illustrates the algorithm of the HAC. The algorithm starts with regrouping the observations Ob1 and Ob2 because they are the closest elements. Then we regroup Ob3 and Ob4. The end of calculation is reached when all the observations are regrouped in a single group.

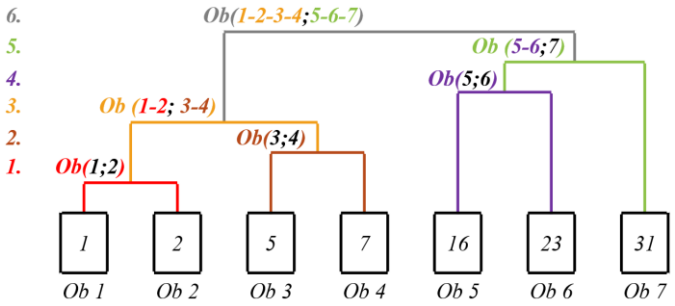

Figure 3: Example of HAC

2) K-means

The k-means method consists in grouping the observations in k groups so that the intra-class distances are minimal and the inter-class distances are maximum. The k-means algorithm operates as follows:

After initializing k points (randomly or not) and considering them as centroids, the algorithm distributes the points (observations) in the k classes thus formed according to their proximity to the centroid. Then the algorithm calculates the classes centroids (centers of gravities) and considers them as new centroids. Then, the algorithm repeats these 2 steps until there is no change. The Fig. 4 illustrates how the k-means algorithm works.

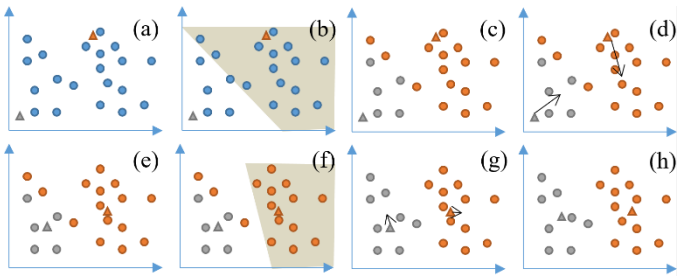

Figure 4: k-means algorithm illustration

The Fig. 4(a) shows the initialization of the centroids, at the beginning the points are in blue to say that they do not belong yet to any class. The Fig. 4 (b and c) show how the point be classed among the 2 classes (grey and orange). Fig. 4 (d and e) present the step of centroids recalculations. Fig. 4 (f and g) represent another iteration of the previous steps. Finally, Fig. 4 (h) shows the final result of this classification.

B. Driving cycles classification

To classify our runs, we chose the k-means method since it is faster and at the opposite of HAC technique, it performs a calculation update of the classes centers after each new assignment to follow the evolution of its content. In HAC, if two individuals are placed in different classes, they are never compared again.

Our results revealed that to properly classify our runs with a minimum of classes, we must keep 5 classes. With only 5 classes, we can explain about 60% of the variance (to explain 100% of the variance, the number of classes must be equal to the number of runs: nearly 8000).

The proportion of the explained variance: R^2 is given by the following equation (7):

$$R^2 = \frac{I_B}{I} = \frac{I_B}{I_B + I_W} \tag{7}$$

Where I is the total inertia, I_B is the inter-class inertia and I_W is the intra-class inertia:

$$I_B = \sum_{i}^{k} distance(C_i, C) \tag{8}$$

$$I_W = \frac{1}{n} \sum_{i}^{k} \sum_{l_{ji}}^{m_i} distance(l_{ji}, C_i) \tag{9}$$

Where k is the number of classes, C_i is the center of class i, C is the center of gravity of all runs, n is the number of runs, m_i is the number of runs belonging to class i and l_{ji} is the driving cycles number j belonging to class i.

C. Obtained results

The driving cycles classification into 5 different operating modes gave the results below.

In the Fig. 5, we present the different classes obtained in a reduced dimension (on only 2 axes of the PCA). The data here are centered and reduced. This figure shows that we have a good classification that separates well the different runs.

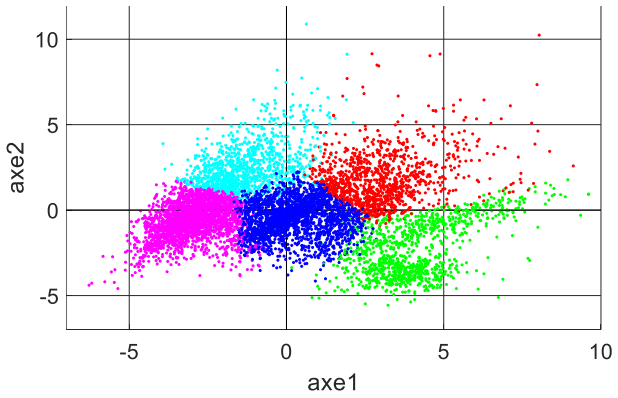

Figure 5: Representation of the classification result on the first two principal components

After the driving cycles classification in 5 parts, we considered that a mode of operation is the average of the runs belonging to the same class.

To understand the differences between the different operation modes, we have represented them in Fig. 6.

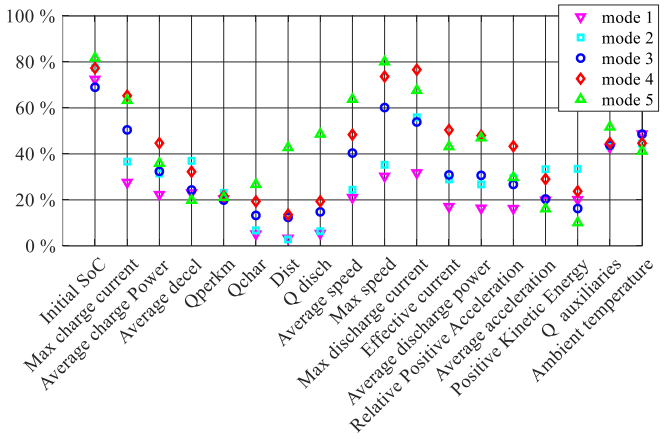

Figure 6: Characteristics of each running mode

Fig. 6 shows the difference between the operation modes according to the 18 variables chosen previously.

To fully understand Fig. 6, we note that we have proceeded to a data normalization so that for each variable the data is between 0% and 100%. 0% being the minimum value and 100% is the maximum value of all the recorded runs.

According to variables distance, average speed and max speed, we can see that modes 1 and 2 represent urban drive cycles and modes 4 and 5 are driving cycles that include highway parts. In addition, according to variables RPA, Average acceleration, average deceleration and PKE, we can see that mode 4 describes a more aggressive behavior compared to mode 5 and similarly for mode 2 compared to mode 1. This may also be related to the traffic density.

We can also notice that, for Qperkm and ambient temperature variables, all the modes have almost the same coordinates. This means that the obtained operating modes do not depend or depend very little on these variables.

This methodology allowed us to obtain several driving cycles classes that depend on different variables. In a next step, we will study the effect of these driving modes on batteries aging.

VI. Conclusion

The present work starts from real-life recording of the batteries solicitation to determine the different operating modes of the batteries in automotive application. The database, which is constructed from about 2 years of data recording of 10 EV, initially contains more than 140 parameters with potential impact on the batteries. By using several selection tools based on correlation coefficients of Pearson and Spearman and Shannon entropy, we decrease the number of parameters to 18. Laplacian Score method cannot manage to reduce anymore this number which attests the difficulty to easily characterize a mode of use of EVs. For a better quality of classification, we used the PCA technique which made it possible to reduce the dimension and to eliminate the redundancies. A K-means analysis enabled us to find 5 different modes of use. Each mode is defined by the value of 18 parameters previously identified. The tools developed in this work will allow to study the potential links between the batteries aging rate and the identified modes of use.

Acknowledgment

We like to thank the region Auvergne-Rhone-Alpes for its financial support for this work.

References

[1] CITEPA, “Poussières en suspension - CITEPA.” [Online]. Available: http://www.citepa.org/fr/air-et-climat/polluants/poussieres-en-suspension.

[2] Avere-France, “Véhicules électriques immatriculés en 2016,” Avere-France. [Online]. Available: http://www.avere-france.org/Site/Article/?article_id=6826.

[3] S. B. Peterson, J. Apt, and J. F. Whitacre, “Lithium-ion battery cell degradation resulting from realistic vehicle and vehicle-to-grid utilization,” J. Power Sources, vol. 195, no. 8, pp. 2385–2392, 2010.

[4] E. Wood, M. Alexander, and T. H. Bradley, “Investigation of battery end-of-life conditions for plug-in hybrid electric vehicles,” J. Power Sources, vol. 196, no. 11, pp. 5147–5154, 2011.

[5] P. Kreczanik, B. Jeanneret, and S. Pelissier, “Construction of Database on Real World Uses of Electric Vehicles - A French Case,” in 2014 IEEE Vehicle Power and Propulsion Conference (VPPC), 2014, pp. 1–5.

[6] S. Tufféry, Data mining et statistique décisionnelle - 4ème édition, 4e édition. Paris: Editions technip, 2012.

[7] R. M. Gray, Entropy and information theory, 2nd ed. New York: Springer, 2011.

[8] K. S. Nesamani and K. P. Subramanian, “Development of a driving cycle for intra-city buses in Chennai, India,” Atmos. Environ., vol. 45, no. 31, pp. 5469–5476, Oct. 2011.

[9] G. Castignani, T. Derrmann, R. Frank, and T. Engel, “Validation study of risky event classification using driving pattern factors,” in 2015 IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT), 2015, pp. 1–6.

[10] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” in Advances in neural information processing systems, 2006, pp. 507–514.

[11] K. Benabdeslem and M. Hindawi, “Constrained laplacian score for semi-supervised feature selection,” Mach. Learn. Knowl. Discov. Databases, pp. 204–218, 2011.

[12] G. Roffo, S. Melzi, and M. Cristani, “Infinite Feature Selection,” in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4202–4210.

[13] G. Saporta, Probabilités, analyse des données et statistique. Paris: Technip, 2011.

[14] D. D. Suhr, “Principal component analysis vs. exploratory factor analysis,” SUGI 30 Proc., vol. 203, p. 230, 2005.

[15] A. S. Beavers, J. W. Lounsbury, J. K. Richards, S. W. Huck, G. J. Skolits, and S. L. Esquivel, “Practical considerations for using exploratory factor analysis in educational research.,” Pract. Assess. Res. Eval., vol. 18, 2013.

[16] R. R. Herrera and D. S.-L. Gac, Initiation à l’analyse factorielle des données : Fondements des mathématiques et interprétations, cours et exercices. Paris: Ellipses Marketing, 2002.

[17] M. Bora, D. Jyoti, D. Gupta, and A. Kumar, “Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab,” ArXiv Prepr. ArXiv14057471, 2014.