State-of-health estimators coupled to a random forest approach for lithium-ion battery aging factor ranking

Kodjo S.R. Mawonou, Akram Eddahech, Didier Dumur, Dominique Beauvois, Emmanuel Godoy

State-of-health estimators coupled to a random forest approach for lithium-ion battery aging factor ranking

Kodjo S. R. MAWONOU[a,b,*], Akram EDDAHECH[b], Didier DUMUR[a], Dominique BEAUVOIS[a], Emmanuel GODOY[a]

[a] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), 91190, Gif-sur-Yvette, France.*
[b] *Technocentre Renault, 1 Avenue du Golf, 78280 Guyancourt, France.*

Abstract

Electrified vehicles users may expect their vehicle to have a steady autonomy range and available power throughout the lifetime of their cars. The health assessment of Lithium-ion batteries (LIBs), in that regard, represents a critical point for performance evaluation and lifetime prediction. Reliable state-of-health (SoH) assessment is essential to ensure cautious and suitable use of LIBs. To that end, several embedded solutions are proposed in the literature. In this paper, two new aging indicators are developed to enrich the existing diagnosis-based (DB-SoH) solutions. These indicators are based on collected data during charging (CDB-SoH) and driving (DDB-SoH) events overtime. The data are comprised of variables such as distance, speed, temperature, charging power, and more. Both solutions produce reliable state-of-health SoH assessment with significantly good accuracy. Additionally, a data-driven battery aging prediction using the random forest (RF) algorithm is introduced with respect to actual users' behavior and ambient conditions. The proposed solution produced an SoH estimation accuracy of 1.27%. Finally, a method for aging factors ranking is proposed. The obtained ranking is consistent with known aging root causes in the literature and can be used to mitigate fast LIB aging for electrified vehicle applications.

Keywords: Li-ion battery ; SoH estimation ; Aging factors ranking ; Machine Learning; Random Forest.

1. Introduction

General opinion and authorities show increasing concerns over environmental issues. Several policies are conducted worldwide to overcome CO2 emission issues. Accordingly, automotive vehicle manufacturers have to propose more environmentally friendly vehicles [1, 2]. In 2019 the global electric car stock reached 5 million vehicles. Electrified vehicles (EV, PHEV, and HEV) are promising technologies to reduce global transportation CO2 emission. EVs are powered by batteries pack made of numerous electrochemical cells. The control of such a complex system is mandatory and is achieved by the battery management system (BMS) to safely and fully exploit these batteries.

*Corresponding author
Email address: kodjo.mawonou@centralesupelec.fr (Kodjo S. R. MAWONOU)

Among a variety of energy storage solutions, Lithium-ion batteries (LIBs) are broadly accepted as promising candidates for many different applications, mainly due to their high energy, power densities and their longer lifespan[3]. Nevertheless, lately, deep researches are animated on battery technologies to enhance their global cell electrochemical performances compared to today NMC lithium-ion like recent studies on solid-state battery [4, 5, 6].

To make electrified cars more appealing, potential customers need to be comforted in terms of security and benefit such as autonomy and rechargeability. Efficient and cautious use of Li-ion batteries typically requires the monitoring of numerous variables like the state of charge (SoC), the state of health (SoH), the acceptable/available power when charging/discharging, the charge capacity and the internal resistance [7, 8, 9]. Unfortunately, these variables are not directly measurable. The BMS has to estimate each of them, thanks to measured data such as current, voltages, temperatures, and state observers [10]. Besides, these parameters change over time and charge-discharge cycles. These changes have to be taken into account to ensure a maximal level of performance throughout the life of the battery [11]. The lifetime of the battery is maximized by facilitating an ideal operating condition [12, 13]. The autonomy of EVs represents a key element to make CO2 emission mitigation in the transportation sector effective. Unfortunately, LIBs remain complex systems, and the processes of their aging are even more complicated [13]. Several efforts are made in the literature to understand and access Li-ion batteries aging processes. Extensive reviews of the major aging mechanisms are thoroughly conducted in [14, 9, 15, 13]. LIBs aging takes place in the anode, cathode, and electrolyte.

It is widely believed that the major aging phenomena happen at the anode, such as solid electrolyte interface (SEI) formation and growth, lithium plating, loss of lithium, loss of active material, impedance rise, etc. SEI formation is said to be the most relevant phenomenon and is broadly studied, especially in the case of carbon type negative electrodes made [16, 17, 13]. In [18], Safari proposed a SEI model based on electrochemical equations operating a LCO type of LIB. He confirmed that huge SoC variations and high currents increase SEI growth. Prada used this approach in [19] on an LFP type of LIB to identify a capacity loss law. Difficulties arise, however, when an attempt is made to implement this method. In fact, several parameters are to be identified involving sometimes destructive tests for the battery. In addition, the uncertainty from parameter identification makes this model hard to operate.

There are two major aging categories: cycling and calendar aging. The latter consists of all aging processes that lead to degradation when the battery is at rest. This type of aging has been thoroughly studied in the literature [20, 21, 22]. It is asserted to be increased at high SoC, and temperature values. However, it is claimed that its predominant cause is the storage at high SoCs. High SoCs and temperatures induce parasitic side reactions and electrolyte reduction, namely at the anode. The calendar aging is not to be neglected, knowing the EVS spend most of their lifetime (up to 90%) in parking areas [23]. Cycling aging, on another note, is composed of all aging processes that lead to degradation when the battery goes through during charging or discharging [24, 25]. Most manufacturers provide the number of cycles of charge/discharge as the lifespan of their batteries. Lithium plating is a significant aging phenomenon during battery cycling. It can even cause short-circuits, leading to the sudden death of the battery. Lithium platting occurs at high cycling rates. Efforts are produced in the literature to propose health-conscious fast-charging solutions for LIBs [26]. In fact, fast charging implies a high charging current rate. Deep discharge and

overcharging are the next leading cause of aging during battery cycling. Consequently, all these phenomena lead to two substantial macroscopic consequences: capacity loss and power fade. Both outcomes are not desirable, as consumers may expect their vehicle to have a steady autonomy (charge capacity) and power (maximum available) throughout the lifetime of their cars. Several approaches exist to access LIBs aging, such as ampere-hour counting (AhC), remaining useful life (RUL) estimation, differential voltage analysis (DVA) or incremental capacity analysis (ICA), electrochemical impedance spectroscopy (EIS) analysis. These classical approaches are briefly presented in the subsequent section.

Latterly, numerous machine learning techniques have been devised for battery SoH estimation, such as artificial neural network (ANN) [27, 28, 29, 30], support vector machine (SVM) [31, 32], regressive vector machine (RVM) [33, 34], particle filter (PF) [35, 36], Random Forests (RF) and Gaussian process regression (GPR) [37, 9]. Utilizing extracted features from the terminal voltage response of the Li-ion battery under current pulse tests, a novel method is proposed in [31] together with an SVM model to estimate the SoH of an LFP type LIB. The author achieved an SoH prediction error of less than 1%. However, only one cell was considered, and the findings are clearly not applicable to actual EVs.In [33], an RVM model is utilized to predict the remaining capacity of a set of five LIBs. The solution is based on health features extraction and optimization. The solution is based on health features extraction and optimization. Fourteen health features were extracted from charge current, voltage, and temperature profiles to ensure accurate remaining capacity prediction. The findings would have been more interesting if the author had included actual EV data or considered a broader set of LIBs. In [38], a random forest regression for online capacity estimation of lithium-ion batteries was presented. The author developed an RF model to approximate the relationship between characteristic features extracted from the charging voltage-capacity curve and the capacity of an NMC battery. Adopting the resulting model, he reliably predicted the capacity with a mean-square error of 1.3%. The characteristic features were derived from incremental capacity (IC) curves. The distinguishing features were derived from incremental capacity (IC) curves. As an online solution, this approach does not process a large amount of data due to the limited computational capability of the present BMSs. Also, this approach is separately implemented for each user. Therefore, the diverse experiences of EVs users are not exploited. Two chief drawbacks are present in the existing data-driven health prognosis solutions. (1) Most studies use IC curves or an alternative voltage response to current pulse stimuli to extract relevant features. This brings additional computation burden to the BMS and therefore renders these solutions less attractive. (2) Proposed solutions are often designed and tested using very few batteries or even cells. On top of that, actual usage conditions are not considered. In the best-case scenarios, only worldwide light vehicles test cycles (WLTC) are considered instead of actual EV operations.

The present paper aims to propose a complete framework for a high fidelity aging assessment for LIBs in EV application. Because of the limited computational capability of BMSs, an off-line data-driven SoH prediction solution is highly desired. Data from several EVs driving, charging, and parking missions are collected over the years at the pack level. The stored data correspond to variables such as instantaneous power, discharged/charged energy, temperature, estimated SoC by the BMS, etc. We seek a method capable of accurately predicting the health of LIBs,

directly based on real-life EVs usage data collected over many years. Three primary goals are pursued in this paper, as displayed in Fig. 1:

- Design and test aging indicators using recorded data
- Establish a machine learning approach to predict LIBs aging with respect to user's behaviors
- Achieve aging factors ranking.

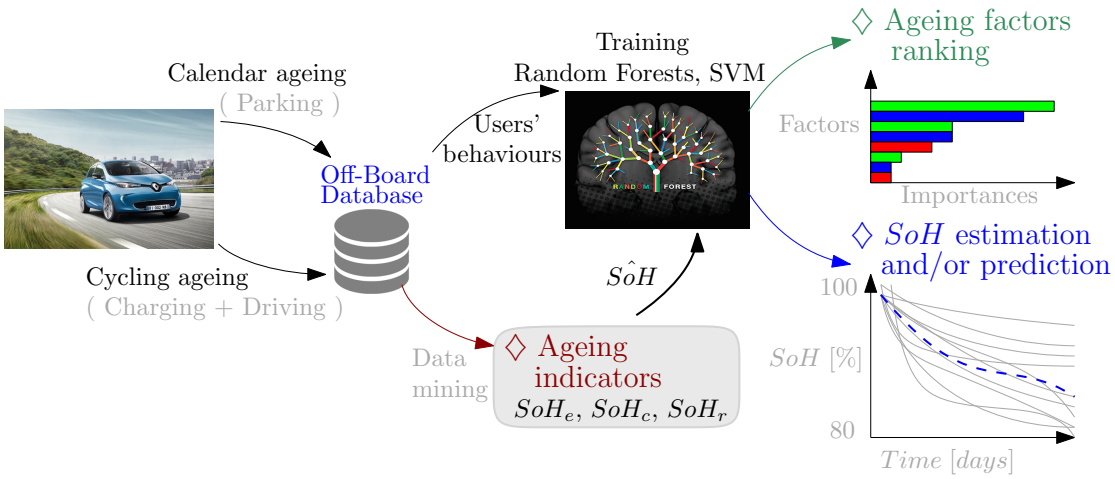


Figure 1: Illustrated steps of the presented study.

This study is accomplished thanks to several collected and stored EVs data in an "off-board database". The intended purpose of this database is to feed tasks such as invoicing, warranty, after-sales, and engineering. The essential characteristics of the collected data are:

- More than 180000 vehicles logged from several countries across Europe and Asia.
- Heterogeneous mix of numerical data such as SoC, temperatures and pack mileage with categorical data such as battery purchase mode (lease, fully paid), Pack Status (new, second life, end of life), and delivery country code.
- No time series. The data is stored event-wise (charging event, driving event, parking event).
- Significant missing data rate.

The outline of this paper is organized as follows. The following section briefly describes existing Li-ion aging estimation solutions. Section 3 displays two new health indicators based on stored data during charging and driving events overtime. A data-driven battery aging prediction using the random forest (RF) algorithm is proposed together with users' behavior and ambient conditions in Section 4. The RF regression is a popular supervised machine learning

algorithm utilized in various fields. RF has the ability to approximate nonlinear relationships accurately, produce a robust performance against outliers, and well handle heterogeneous and missing data. On top of that, an approach for aging factors ranking is proposed in Section 5. Ultimately, conclusions are drawn in Section 6.

2. Health estimation approaches for LIBs

Three global approaches could be used in order to evaluate the battery degradation and estimate its state of health. The first one is the durability model :calibrated using an expensive experimental design plan, this semi-empirical approach uses the history of the battery usage in calendar and cycling to integrate SOC, temperature, and profile impact on battery aging.

The second is the diagnosis-based approach called here DB-SOH, which is based on the evolution with the aging of the time-domain screenshot of some selected battery variables. This latter approach advantage is the adaptability to the users driving and charging habits. However, it could require rich data recording and filtering. That is why, according to the application context, this method could be adapted to match real-time implementation or off-board analysis.

The third one is the black box based perdition approach that uses a tremendous amount of battery history data to approximate its future behavior. Actually, because of the implementation constraints, such a method is not used in real-time BMS applications. However, it is very promoting for off-board analysis, especially with new big-data capabilities and tool developments.

In terms of performance, as stated in the introduction, there are two categories of aging degradation: calendar aging and cycling aging. Energy loss and power fading throughout the lifetime of LIBs are the most relevant issues in the point of view of EV users. Energy loss is correlated to capacity loss, whereas power fading is correlated to internal impedance rising in the LIB. In this section, classical approaches used to assess the health of LIBs are introduced.

In most, if not all, BMSs LIBs aging indicator assessment is done using ampere-hour counting. Through the charging or discharging methods, knowing the initial and the final state of charge (SoC_i and SoC_f) at the initial (t_i) and final time (t_f), one can compute the relative capacity as the state of health of the battery (SoH_c):

$$Q_n = \frac{\int_{t_i}^{t_f} I(t)\,\mathrm{d}t}{SoC(t_f) - SoC(t_i)} \tag{1a}$$

$$SoH_c = 100 \times \frac{Q_n}{Q_{BOL}} \tag{1b}$$

where Q_n is the current nominal capacity, Q_{BOL} is the nominal capacity at the beginning of life (BOL), and I is the current [15]. The precision of this method relies on the quality of the current sensor due to the cumulative error brought by current integration. To reach satisfactory accuracy, the SoC variation $|SoC_i - SoC_f|$ must be adequately estimated or obtained using the measured open circuit voltage (OCV) of the battery in a relaxed state together with the SoC-OCV. For LIBs the SoC-OCV curve is provided by the battery manufacturers. For EV applications, the end of life (EOL) is reached when the relative capacity loss is up to 20%.

A similar approach is utilized to assess the energy loss of LIBs. This approach measures the relative charged or discharged energy of the battery as (SoH_e):

$$\Delta E = \int_{t_i}^{t_f} I(t) \times V_t \, \mathrm{d}t \tag{2a}$$

$$SoH_e = 100 \times \frac{\Delta E}{\Delta E_{BOL}} \tag{2b}$$

where V_t is the terminal voltage of the battery, ΔE is the current energy variation, and ΔE_{BOL} the energy variation at the beginning of the life of the battery [39]. Quite often, this approach is implemented using charged energy to ensure that energy variations are measured in the same conditions. The precision yield by this method highly depends on the accuracy of current and voltage sensors measurements.

Further capacity loss assessment can be conducted using differential voltage analysis (DVA) or incremental current analysis (ICA). These solutions are achieved during charging events using Eq.3:

$$\mathrm{d}Q/\mathrm{d}V = \Delta Q/\Delta V \tag{3a}$$

$$\mathrm{d}V/\mathrm{d}Q = \Delta V/\Delta Q \tag{3b}$$

where ΔV is a voltage variation, and ΔQ is a capacity variation. The DVA is obtained using Eq.3 every time the voltage variation reaches 5 mV, while the battery is being charged with a low current. In [40], it is claimed the DVA curves provide a better understanding of aging. In fact, one can distinguish anode degradation, cathode degradation, and changes in electrode balancing. Different peak values on the DVA or ICA curves can be utilized to assess the state of health of LIBs. Nevertheless, these solutions are barely used in BMSs [41, 42]. ICA and DVA are limited to low current rates, sensitive to measurement noise, influenced by the operating temperature. They are not suitable for chemistry with vast voltage plateaus such as lithium iron phosphate battery (LFP) cells [43].

Like mentioned earlier, LIBs power loss is highly related to internal impedance growth. The commonly associated aging indicator is given as :

$$SoH_r = 100 \times \frac{DCR}{DCR_{BOL}} \tag{4}$$

where DCR is the direct current resistance value of the battery [44]. When it comes to EV applications, the end of life (EOL) is reached when the relative impedance rise is up to 200%. Knowing that the terminal voltage of a lithium-ion cell has an upper bound value, the impedance growth directly hinders the ability to drain high power when the internal impedance growth is large. Alternatively, the internal impedance growth can be accessed using electrochemical impedance spectroscopy data [27, 22, 17]. However, this method of analysis is limited to experimentation in labs due to the heavy pieces of equipment it requires. Therefore EIS based solutions are not suitable for EV applications.

3. Diagnosis Based SOH (DB-SOH) indicators

In this section, data collected during driving and charging events are used to design and test two new aging indicators for LIBs. Ideally, aging performances are measured using well-calibrated protocols and steady ambient

conditions. However, in the case of EV applications, users' behavior, driving, and charging conditions are broadly diverse. For car manufacturers, periodical EV recall for health assessment under a well-designed protocol is too costly and unsuitable. The core idea in this section is to make a statistical use of the stored data by proposing satisfying aging indicators similar to a calibrated protocol.

*3.1. **Driving Diagnosis Based SOH (DDB-SOH)***

The classical capacity based aging monitoring was introduced in section using Eq. 1b. Considering that definition, the following hypothesis are made :

H1: For **identical** driving profiles and conditions, the state of charge variation ΔSoC depends on SoH_c only as for :

$$\Delta SoC = \frac{1}{Q_n}\int_{t_i}^{t_f} I(t)\,\mathrm{d}t \tag{5a}$$

$$SoH_{c,\,dr} \simeq 100 \times \frac{\Delta SoC_{BOL}}{\Delta SoC} \tag{5b}$$

H2: **Identical driving profiles** and **conditions** are identified using covered **distance** d $[km]$, average **speed** $\bar{v}$ $[km.h^{-1}]$ and **temperature** $Temp_{dc}$ $[^o\ C]$.

H3: Accurate state of charge variation ΔSoC is measured using $SoC\ -\ OCV$ curve after a **2 hours minimum rest time** at both beginning and ending of driving events. Fig. 2.3 displays this idea.

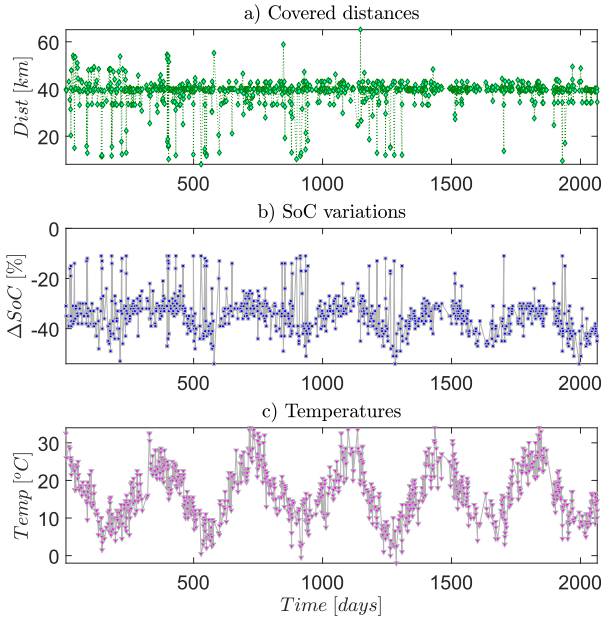
Fig. 2.1 displays driving data example for a typical EV user extracted from our database. The figure shows a) the covered distance for different driving events, b) the resulting SoC variations, and c) recorded battery pack temperatures. One can easily notice that this typical user covers quite often 40 km. The cyclic temperature values matching seasonal temperatures are also noticeable. Numerous reasons may cause a steady shift of the covered distance values in some cases. Most importantly, these shifts have been observed in the database. Therefore, the following equation is used instead of Eq. 5b:

$$SoH_{c,\,dr} \simeq 100 \times \frac{\Delta SoC_{BOL}}{\Delta SoC} \times \frac{d}{d_{BOL}} \tag{6}$$

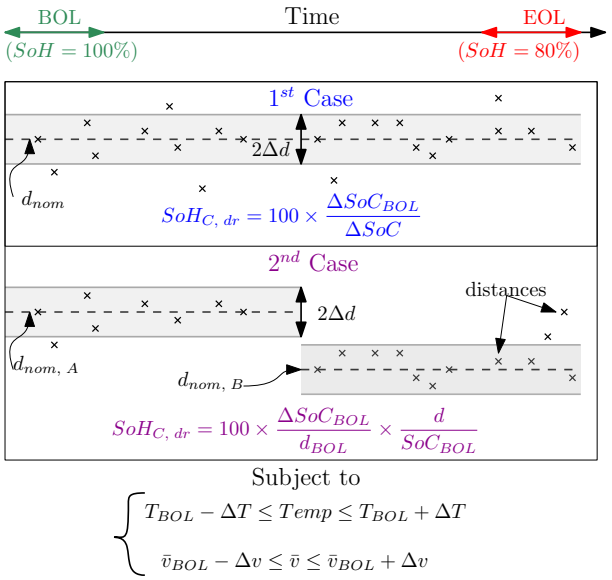
where d is the covered distance, d_{BOL} is the nominal covered distance at the beginning of life (BOL). Both cases (1^{st} with shift, and 2^{nd} without shift) are illustrated in Fig. 2.2.

The measured values of state of charge displayed in Fig. 3.1 a) and Fig. 3.2 a) are filtered using robust weighted local regression algorithm : *"rlowess"* available in Matlab®. LOWESS is a non-parametric regression model introduced in [45]. It combines linear and nonlinear regression by performing separate linear regressions at any given date t in a pre-defined span $l \in [0\ 1]$. Only the data points in l, i.e. the $l \times N$ nearest neighbors of t are used for the local weighted linear regression. The regression weights w_i are calculated for each data point in the span using the "*tricube*" function in Eq. 7:

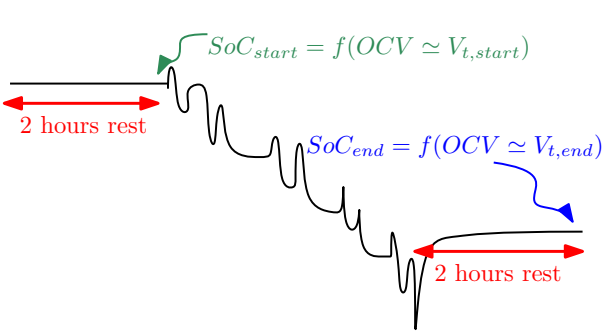
$$w_i = \left(1 - \left|\frac{t - t_i}{d_{max}}\right|^3\right)^3 \tag{7}$$



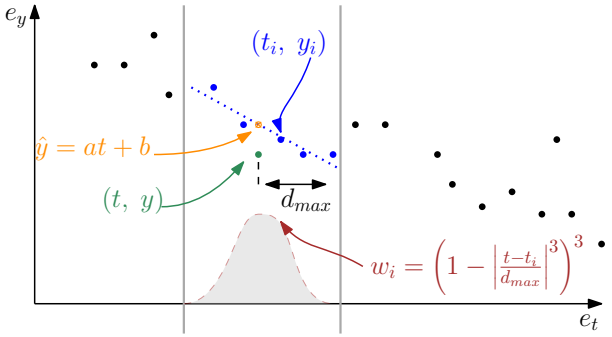
(2.1) Driving data example for typical EV user extracted from our database.



(2.2) $SoH_{c,\,dr}$ indicator diagram.



(2.3) SoC variation using OCV-SoC curve when a minimum rest time is observed.



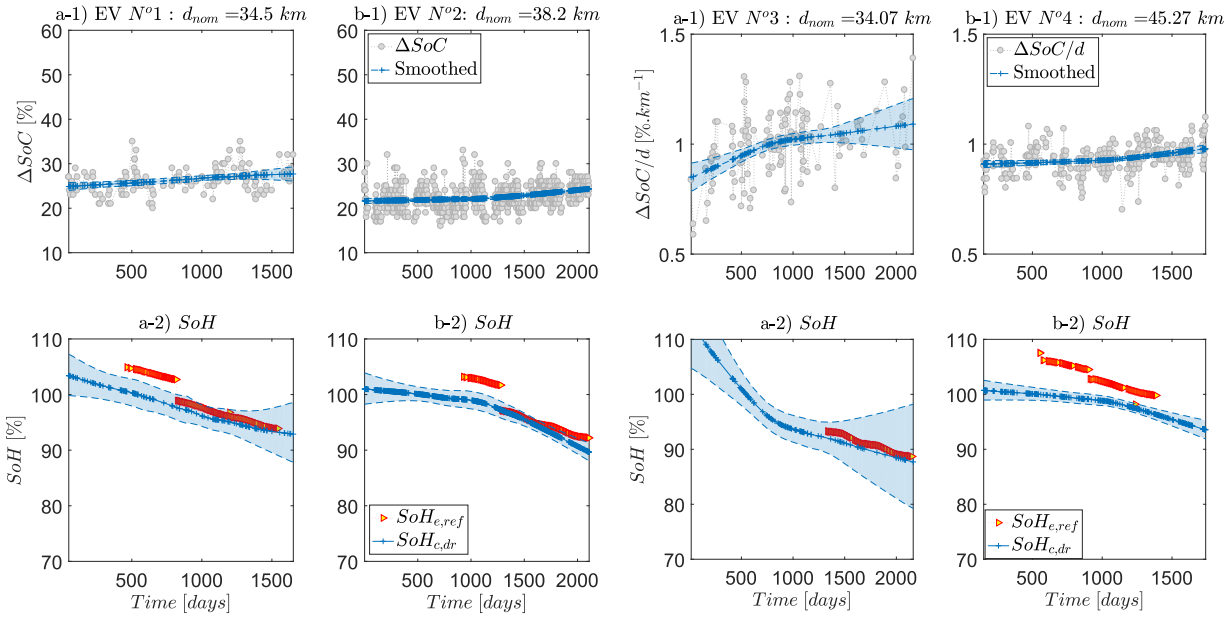
(2.4) Data filtering using robust LOWESS algorithm. Weighted linear mean squares regression is achieved in the sliding window : $(\alpha,\ \beta) = \sum_{i=1}^{n} \omega_i (y_i - t_i \times \alpha + \beta)^2$.

Figure 2: SoH estimation using recorded data during charging events.

where t is the date corresponding to the value to be predicted $\hat{y}$, t_i nearest neighbors of t and d_{max} the distance between y and most remote point among the nearest neighbors.Within the pre-defined span, a weighted linear regression is achieved using Eq. 8a. The smoothed value is then obtained using Eq. 8b.

$$[a,\ b] = \text{argmin} \sum_{i=1}^{n} \omega_i (y_i - t_i \times \alpha + \beta)^2 \tag{8a}$$

$$\hat{y} = at + b \tag{8b}$$



(3.1) 2 EVs in 1^{st} case. $SoH_{c,dr}$ using ΔSoC and driving data. $d_{nom} - 2,5 \leq d \leq d_{nom} + 2,5$ (km), $T_{nom} - 5 \leq T \leq T_{nom} + 5$ (^{0} C) and $v_{nom} - 10 \leq \bar{v} \leq v_{nom} + 10$ (kmh^{-1}).

(3.2) 2 EVs in 2^{nd} case. $SoH_{c,dr}$ using ΔSoC and driving data. $d_{nom} - 12,5 \leq d \leq d_{nom} + 12,5$ (km), $T_{nom} - 5 \leq T \leq T_{nom} + 5$ (^{0} C) and $v_{nom} - 10 \leq \bar{v} \leq v_{nom} + 10$ (kmh^{-1}).

Figure 3: SoH indicator values obtained using recorded driving data.

Using the proposed solution for both cases (with and without distance shift), $SoH_{c,dr}$ values are obtained for 2 EVs in 1^{st} case and 2 EVs more in 2^{nd} case. The results are displayed in Fig. 3. In Fig. 3.1 a-1) and a-2) are displayed the measured SoC variations at several charging events, and their filtered values using LOWESS. In Fig. 3.1 b-1) and b-2) are displayed the resulting $SoH_{c,dr}$ proposed, corresponding to the measurements, in comparison with the reference energy based aging indicator $SoH_{e,ref}$. The proposed $SoH_{c,dr}$ is displayed considering a 3σ confidence interval obtained using a bootstrapping approach together with the LOWESS filtering algorithm. The reference aging indicator $SoH_{e,ref}$ is available 2 years after the vehicle is put into usage.

3.2. *Charge Diagnosis Based SOH (CDB-SOH)*

An energy-based SoH estimation is proposed in [39]. The author therein displayed the linear relationship between the capacity loss and the relative charged energy during constant current - constant voltage (CC-CV) charge for 3 different chemistries of LIBs. The available energy in a battery pack is matched to its capacity. The energy fading can, therefore, become an indicator of aging. An alternative yet efficient approach of battery pack SoH estimation based on exchanged energy is proposed in this paper. The charged energy during a constant power - constant voltage (CP-CV) charging event starting at time t_i and ending at time t_f is given in Eq. 9a. Knowing the initial and final state of charges (SoC_i and SoC_f), one can compute the SoH_e using Eq. 9b. Thus the following hypothesis is made:

H4: For **identical** charging powers P_{ch} in similar temperature conditions $Temp_{ch}$, the health indicator SoH_e is approximated using Eq. 9b.

$$\Delta E_{ch} = \int_{t_i}^{t_f} I(t) \times V_t(t)\, \mathrm{d}t \tag{9a}$$

$$SoH_e \simeq \frac{\Delta E_{ch}}{\Delta SoC} \times \frac{\Delta SoC_{BOL}}{\Delta E_{ch,BOL}} \tag{9b}$$

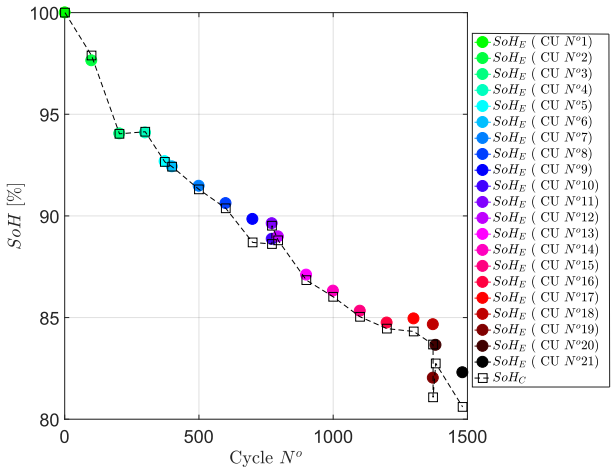
where $\Delta E_{ch,BOL}$ is the charged energy at the beginning of the life (BOL) of the battery, associated with the state of charge variation ΔSoC_{BOL}. Optimally SoC variation values must be close to 100 %. This solution was tested using an EV battery of capacity $Q_n = 128\ Ah$, comprised of 192 cells in configuration $2P96S$. The battery pack was introduced in a climatic chamber and well thermalized. The measured temperature is around $+/-2\ ^oC$ of the target one. Several check-ups (CUs) were conducted after many cycling numbers. Eventually, 21 CUs were done corresponding to 1482 cycles in total. During each CU, the battery pack state of charge, capacity, and charged energy are assessed. Using Eq. 9b and 1b one can compute SoH_e and SoH_c. On Fig. 4.1 are displayed the values of computed SoH_e and SoH_c. The mean absolute difference defined as $|SoH_e - SoH_c|$ is 0.42 %. This minor discrepancy confirms the relevancy of the proposed method.

A test was conducted using data collected from EVs usage for approximately seven years to validate the proposed solution. For these vehicles, during every charge event, instantaneous power measurement data are stored over the years in a database. The sampling time during a charging event is $T_s = 30\ mins$. The available charging powers for the tested vehicles are 43, 22, 11, 7, and 3 kW, as illustrated in Fig. 4.2. The data can be exploited to approximate the corresponding charged energy values. Using Eq. 10 as illustrated in Fig. 4.2, the results of the SoH_e approximation for 2 EVs are displayed in Fig. 4.3. The studied vehicles are effectively purchased and in use.

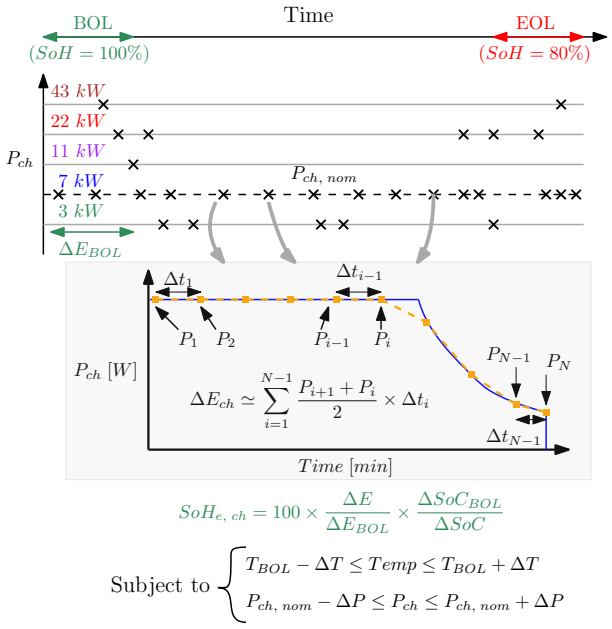
$$\Delta E_{ch} \simeq \sum_{i=1}^{N-1} \frac{P_{i+1} + P_i}{2} \times \Delta t_i \tag{10}$$

In Fig. 4.3 a-1), and b-1) are displayed the measured charged energy at several charging events, and their filtered values. In Fig. 4.3 a-2) , and b-2) are displayed the estimated $SoH_{e,\,ch}$ proposed corresponding to the measurements, in comparison with the reference energy based aging indicator $SoH_{e,\,ref}$. The reference aging indicator $SoH_{e,ref}$ is available 2 years after the vehicle is put into usage. The measured values of charged energies displayed in Fig. 4.3 a-1) and b-1) are filtered using the previously presented LOWESS method.

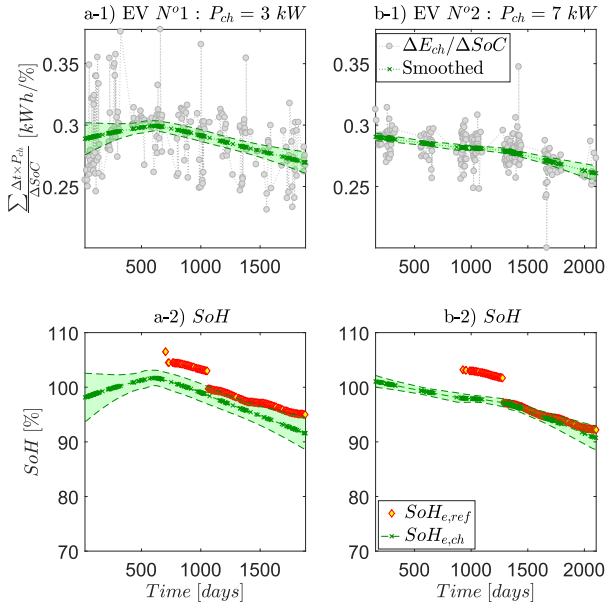
The usage temperature range of the tested EVs is -15^oC to 45^oC, therefore causing significant local variations. A global trend can nevertheless be identified and exploited. In fact, Eq. 9b can only be utilized when the data are recorded in similar conditions: temperature or charging power. Given, for example, that ambient temperature evolves periodically over time, the data displayed in Fig. 4.3 a-1) and b-1) are sparse. The proposed $SoH_{e,\,ch}$ values are tightly close to the reference $SoH_{e,\,ref}$ ones. This suggests that even with approximated values of charged energy, one can derive a satisfactory battery health indicator. The proposed approach can be utilized to support the on-line aging estimation based on capacity fading (SoH_c).



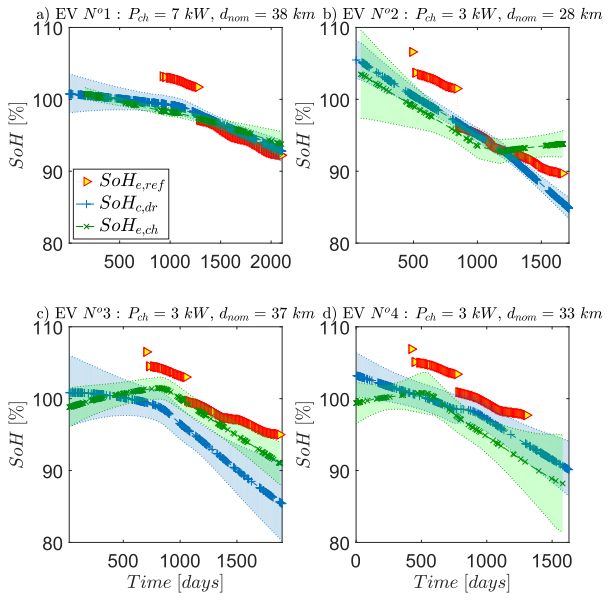
(4.1) Ageing indicator comparison. Absolute mean difference $|SoH_c - SoH_e|$= 0.42%.



(4.2) Charged energy approximation using instantaneous power measurements: T_e = 30 *mins*.



(4.3) Charged energy approximation and $SoH_{e,\,ch}$ values for 2 randomly selected EVs.



(4.4) Proposed health indicators comparison : $SoH_{c,\,dr}$ Vs $SoH_{e,\,ch}$ for 4 randomly selected EVs.

Figure 4: SoH estimation using recorded data during charging events.

3.3. SoH indicators discussion

Two aging indicators were presented using driving data $SoH_{c,\,dr}$ and charging data $SoH_{e,\,ch}$. Both of these approaches should produce similar results, as observed in Fig. 4.1 using test bench data of a $Q_n = 128\ Ah$ battery pack. These indicators were computed for 4 randomly selected EVs and displayed in Fig. 4.4. As expected, both solutions produce similar results and are tightly close to the reference $SoH_{e,\,ref}$ values. There is still room for improvement. The presented solution using driving data can be enhanced using more accurate driving average speed values $\bar{v}$ or even taking into account maximal current and voltage values during driving events to ensure hypothesis H2. The accuracy of the proposed $SoH_{e,\,ch}$ method can be improved by a more accurate approximation of the charged energy ΔE_{ch}. This can be achieved by reducing the sampling time during a charging event to $T_s = 10\ mins$.

4. Machine learning based health prediction

In section 3 two SoH indicators were presented. Both approaches produce satisfying results. Unfortunately, repeated driving or charging conditions are needed to implement said solutions successfully. Those conditions are not always satisfied with all the EVs in the considered "off-board database" at all times. The main goal in this section is to provide a desirable off-line machine-learning approach to predict SoH values based on environmental conditions and user's behaviors. To that extent, the extracted features from the considered "off-board database" are presented. Additionally, an RF approach for SoH prediction is presented. The obtained results are finally compared to a more classical SVR approach.

4.1. Data description and feature extraction

We stated earlier in the introduction that data are stored in the considered "off-board database" event-wise. There are, therefore, no time-series data. Let the proposed $SoH_{e,\,ch}$ using charging data as the output of the model to be designed. Each value of the said indicator will be associated with environmental conditions and users' behavior. As illustrated in Fig. 5.1, at time t_x, user's behavior and environmental conditions are extracted using all the driving, charging and parking data prior to date t_x. The extracted features are then combined with the $SoH_{e,\,ch}$ value at date t_x as $(\mathcal{X}, y)$ where $y = SoH_{e,\,ch}$ is the output and $\mathcal{X} \in \mathbb{R}^p$ is the corresponding input vector of size p, where p is the number of extracted features. In doing so for all the available dates and EVs in the database a learning data-table is obtained : $\mathcal{Z} = \{(\mathcal{X}_1, y_1), \ldots, (\mathcal{X}_n, y_n)\}$ were n is the number of rows. The extracted features are described in Appendix A. Features extracted during charging events have the prefix "*ch_*" or "*charge_*". Similarly, during driving, the prefix "*dr_*" is utilized, whereas "*pk_*" is used for parking events. Significant variables such as mileage, daily covered distance histogram, cumulative discharged energy are considered. The SoC and temperature are taken into account via a "*driving matrix*" and a "*driving matrix*". The cumulative discharged energy (in *kWh*) is stored in the "*driving matrix*" with respect to predefined ranges of SoC and temperature values. Likewise, the total time (in *days*) spent on parking lots is stored in the "*parking matrix*" with respect to predefined ranges of SoC and temperature values.

These two matrices are illustrated in Fig. 5.2 and are present in Table A.2. The total number of charging events is also considered as well as the charging power, temperature, and the SoC variation. The data preparation from $S\hat{o}H$ build to data splitting and RF illustration is displayed in Fig. 5.

4.2. ***Random Forests regression***

Decision trees, also called Classification and Regression Tree (CART), are statistical model firstly introduced in [46]. The random forest regression model is an extension of the CART technique and can offer better prediction performance. The training stage of RF is to construct multiple de-correlated decision trees. Each tree in RF is grown with a randomized subset of predictors and hence the name 'random' forest. Each decision tree consists of decision nodes and leaf nodes. The decision nodes evaluate each fed-in sample by a test function and pass it to different branches based on the features of the sample. Considering our previously obtained dataset $\mathcal{Z} = \{(\mathcal{X}_1, y_1), \ldots, (\mathcal{X}_n, y_n)\}$ were n is the number of rows. The random forest algorithm steps as illustrated on Fig. 5.3 are :

- Sample, with replacement of B sets $\mathcal{Z}_n^i, i = 1, \ldots, B$ (each set has n rows as the original set $\mathcal{Z}$).
- For each set i train a regression tree (CART) that produces the result $\hat{y_i} = G_i(\mathcal{X}, \mathcal{Z}_n^i)$.
- Compute the final output by combining all regression trees as: $G(\mathcal{X}) = \frac{1}{B}\sum_{i=1}^{B} G_i(\mathcal{X}, \mathcal{Z}_n^i)$.

Random forests algorithm fits well our situations as the main advantages are :

- Good results when using large dataset.
- Easily implemented, and very few parameters are needed.
- The produced RF model can be quickly ran for new inputs.
- Missing and heterogeneous data are well handled.

Only two parameters are needed to run RF regression: the number of trees and the number of random features for each split in the forest to build. The main drawbacks of RF algorithms are long tanning process and ill-handled outliers. Fortunately, our study is proposed as an off-line SoH prediction solution. Therefore, a slow training process is not a deal-breaker.

4.3. ***Results***

A classical 75% training and 25% testing split is implemented using the learning data-table, as illustrated in Fig. 6.1. The split is done with respect to users, as shown in Fig. 6.2. This means that the users in the training dataset are not present in the validation dataset. This matches properly our initial goal of predicting the SoH of EVs that are not eligible for our proposed health indicators conditions in Section 3. The experience was conducted using the "*s*cikit-learn" machine learning library available in Python™ programming language. 186 EVs were randomly

selected, producing a learning data-table of more than 150000 rows and 82 columns. The columns represent the selected features as described in Table A.2 of Appendix A. The selected EVs are aged from 4 to 7 years. As a matter of comparison, a support vector regression (SVR) algorithm [47] is also utilized for the SoH_e prediction using the learning data-table. The SVR algorithm is easily implemented using the same "scikit-learn" machine learning library available in Python™ programming language.

The RF results are displayed in Fig. 6.3. In Fig. 6.3 a) the predicted $\hat{SoH_e}$ values (in green) during the training step are superimposed to the reference SoH_e values (in black). In Fig. 6.3 b) the predicted $\hat{SoH_e}$ values (in red) during the validation step are superimposed to the reference SoH_e values (in black). Fig. 6.3 c) represents the testing error values, whereas d) represents the testing error histogram. Note that in Fig. 6.3 a) and b), each curve represents a different user's data and are concatenated together. The prediction errors displayed in Table 1 where MAE is the "mean absolute error", and Max is the "maximum absolute error". One can notice that both SVR and RF algorithms produce similar health prediction accuracy. It is worth noting nevertheless, that better accuracy is achieved using the RF algorithm. A 1.27% SoH prediction accuracy using users' behavior and environmental conditions is satisfactory. Knowing that this solution is proposed as an off-line SoH prediction method, the accuracy will improve over time with more available data and broader environmental conditions.

SoH errors [%]	**SVR**	**RF**
MAE	1,69	1,27
Max	8,80	6,53

Table 1: SoH_e prediction errors. MAE is the "mean absolute error", and Max is the "maximum absolute error"

A simplified CART tree example is displayed in Fig. 7 for the convenience of the reader. In this figure, "value" is the $\hat{SoH_e}$ value associated to the corresponding node. For example, the root node is associated with an SoH_e value of 93.7%. in Fig. 7 "mse" stands for the mean square root error corresponding to a given $\hat{SoH_e}$ value. The root mean square root error is computed as : $\mathbb{E}\left[\left(\hat{SoH_e} - SoH_e\right)^2\right]$. The RF algorithm goal is to minimize that error at leaves level. The displayed example is a simplistic one where the data split depth is 4. Machine learning using decision trees (DT) is more understandable by humans than other classical approaches such ass SVR, neural network or even deep learning solutions. Three paths are displayed : "**path A**", "**path B**" and "**path C**" leading to corresponding SoH_e values : 99%, 87.8% and 86.8%.

- "**Path A**" leads to the highest SoH value and corresponds to youngest EVs with small mileage and very little to no parking at high SoC and temperature.
- "**Path B**" and "**Path C**" in the other hand are related to extreme environmental conditions and produce the lowest SoH values. "**Path C**" being the worst with oldest EVs, stored at very high SoC ($pk\_soc\_90\_100\_temp\_15\_30$) and very low SoC ($pk\_soc\_10\_30\_temp\_0\_15$), driven at high SoC and temperature ($dr\_soc\_70\_90\_temp\_30\_45$).

In regard to this observation, the presented DT in Fig. 7 fits very well the classical LIBs aging root causes described in [13]. Knowing nevertheless that more than 400 trees were used to test the RF approach. Besides, the data split depth considered is 1000. Therefore, going through each tree is not a suitable method to find the most relevant aging factors. In the following section, aging factor importance will be discussed, and a factor raking method will be proposed.

5. Aging factor importance study

5.1. Variable importance

There are three approaches for feature importance computation in the literature: the default mean decrease in impurity importance, permutation feature importance, and drop column feature importance.

The mean decrease in impurity importance of a feature is computed by measuring how useful the element is at reducing variance when creating decision trees within RFs. Built-in feature importance is provided by the library "*s*cikit-learn". However, this approach is biased, as it has the tendency to inflate the importance of continuous features or high-cardinality categorical variables [48]. This approach is often discarded for the drop column or permutation feature importance. Permutation importance is a reasonably efficient and very reliable technique. It measures of variable significance by observing the "model accuracy decrease" (MDA) when randomly shuffling each predictor variable [49]. First, a baseline model is obtained through training. Then for any given feature, the values of the corresponding column are shuffled. The performance deterioration is measured and set as the relative importance of the shuffled feature. Permutation importance does not require the retraining of the underlying model in order to measure the effect of shuffling variables on overall model accuracy. The method is, therefore, inexpensive. The risk is a potential bias towards correlated predictive variables. Drop column feature importance is the most reliable method. It consists of getting a baseline performance score through the first training. Afterward, a column is entirely dropped then the model is retrained. The relative performance is then measured to obtain the actual feature importance. Unfortunately, this method is time-consuming as it requires the training of the model for each dropped column.

5.2. Results and discussion

In the scope of our study, a Python™ library name ELI5 was utilized. ELI5 provides a way to compute feature importances for black-box estimators by measuring how the MDA when a feature is not available. Using the RF model previously obtained in Section 4.2, a permutation feature importance is computed using ELI5. The 20 most important features are displayed in Fig. 8.1. These features are color-coded. Features related to charging events are in blue, whereas those related to driving events are in red, and the one related to parking events are in green. Fig. 8.1 clearly shows that the most relevant features are the battery age ("*days*"), the parking SoC and temperature (*pk_soc*_70_90_*temp*_15_30, *pk_soc*_90_100_*temp*_15_30), the global mileage (*dr_gbl_Dist*) and the global discharged energy (*dr_gbl_EGY*). This observation matches the previous analysis in Section 4.3 using the simplified CART tree in Fig. 7. One can also notice that variables related to parking events (in color green) are globally more important.

This clearly confirms that calendar aging is dominant as most EVs spend up to 90% of their lifetime on parking lots. Unfortunately, parking conditions are highly correlated with seasonal temperatures. It is, therefore, difficult to mitigate LIBs fast aging phenomena regarding the battery temperature. Luckily very high and low SoC storage values can be avoided through charge planning and SoC limitation methods. Charging events can be smartly scheduled to avoid parking batteries at critical SoC values. Eco-charging can be achieved by stop charging events at a defined SoC and temperature threshold.

The presented LIBs aging factor ranking performs well and fits well the available literature. However, users' behaviors are less identified within the 20 most relevant features shown in Fig. 8.1. To that extent the SoH_e sensibility to the battery age (*days*) and discharged energy (dr_gbl_EGY) are utilized. In Fig. 8.2 are displayed for 4 randomly selected EVs, a) the SoH_e values with respect to age, b) the SoH_e values with respect to the discharged energy, and their respective sensibilities (derivatives). Globally different EVs users experience the same climate. Thus the calendar aging of their batteries is similar. However, users may express various behaviors. The core idea in using the SoH_e sensibility (derivative) as the output of the RF model will enable the ranking of the most relevant aging factors that are induced by users' behavior. In doing so, the ranking in Fig. 8 is achieved. This time around, the 20 most important feature ranking is dominated by user induced features. These features are directly related to cycling aging. Observing the most important factors, one can argue that :

- Driving the EVs at high SoC values ($dr\_soc\_90\_100\_temp\_15\_30$, $dr\_soc\_90\_100\_temp\_15\_30$ ie $SoC \in [90\%,\ 100\%]$) is the most aging accelerating factor

- Secondly, the average driving speed has a significant impact on LIBs aging. In Fig. 8 the factor $dr\_speed\_50\_90$ corresponding to average driving speed $\bar{v} \in [50,\ 90] Km.h^{-1}$ is the second-ranked factor. Note that in the context of our study, high values of average driving speed is indicative of high discharge current rate. It is also reported in [13, 16, 50] that high cycling rates cause the loss of active material in LIBs and eventually lead to capacity fade.

- High charging rates are represented in the resulting ranking as ch_power_P22. Like presented in Section 3.2, the available charging powers are 3, 7, 11, 22, and 43 kW. The selected EVs in our study are often charged at 3 and $7kW$. Very few users often charge their car using power higher than 22 kW. However, this feature is ranked as the third most influencing factor nonetheless. It is well known that fast charging has a negative effect on the durability of LIBs. That knowledge is reinforced in this paper.

Overall, these pieces of evidence support the existing knowledge of the root causes of the aging of LIBs. Numerous factors contribute to the decline of the health of LIBs. Interestingly, this study brings a simple and efficient way of ranking the aging factors. It is also possible to rank these factors in terms of calendar aging. Using the SoH sensibility to age, the cycling aging factors can also be ranked.

6. Conclusion

The presented study is dedicated to the health assessment and management of LIBs. A significant dataset has been recorded during real-life EVs usage over close to 7 years of intensive use. This dataset served for the several proposed approaches validation. A full LIBs aging assessment is completed in three parts, including new diagnosis based state-of-health (DB-SOH) estimators, data-driven aging prediction, and relevant aging factors ranking.

The two new health indicators were designed using recorded driving and charging data of real-life EVs, and demonstrate an accurate aging assessment compared to the existing reference. These methods are up-and-coming because they are based on real-life EVs data collected over many years.

Furthermore, a data-driving aging predictor was developed using the collected data and the Random Forest machine learning solution. The proposed estimator displayed an SoH accuracy of 1.27%. We argue that using decision trees is helpful for aging mechanism investigation and is easily interpreted, unlike the existing methods such as neural networks or deep learning approaches. The RF algorithm is used as it produces excellent results when using extensive data, easily implemented, and handles properly heterogeneous and missing data.

Finally, using permutation-based variable importance computation, an approach for aging factors ranking is presented and found to be in total agreement with the classical aging root causes in the literature. What's more, cycling aging factor ranking is achieved based on the extracted user's behaviors and SoH sensibility to age. The proposed factors ranking approach provides an innovative way to explore battery aging comprehension and estimation.

More exciting results can be achieved in the future by storing and taking into account many more features. For example, more information on discharge current rate can be highlighted by recording the maximal EVs acceleration during driving events.

There are many conceivable applications to the presented study, such as warranty adaption regarding users' behavior, scheduled charging to avoid parking batteries at critical SoC values, eco-charging by stop charging events at a defined SoC and temperature thresholds. In late 2019, several works started connecting EVs batteries with the cloud to propose data-based services later in order to improve batteries' performance and extend their lifespan substantially. The presented study in this paper will be of great use in that context.

Nomenclature

Abbreviations

BMS Battery management system

BOL Begining of life

EIS Electrochemical impedance spectroscopy

EOL	End of life
EV	Electric vehicle
HEV	Hybrid electric vehicle
LFP	Lithium iron phosphate
NMC	Nickel manganese colbat
OCV	Open circuit voltage
PHEV	Plug-in hybrid electric vehicle
SoC	State of charge
SoH	State of health

Math Symbols

$\bar{v}$	Average speed of an EV	$km.h^{-1}$
$\Delta E_{ch,BOL}$	Charged energy at the begining of life	Wh
ΔE_{ch}	Charged energy	Wh
d	Covered distance by an EV	km
P_{ch}	Charging power	W
Q_n	Nominal charge capacity	Ah
$Q_{n,BOL}$	Nominal charge capacity at begining of life	Ah
SoC_{BOL}	State of charge at the begining of life	$\%$
$SoH_{c,dr}$	Capacity based SoH indicator using driving data	$\%$
SoH_c	Capacity based SoH indicator	$\%$
$SoH_{e,ch}$	Eergy based SoH indicator using charging data	$\%$
SoH_e	Energy based SoH indicator	$\%$
SoH_r	Current of the battery	A
SoH_r	Impedance based SoH indicator	$\%$
V_t	Terminal voltage of the battery	V

Events	Description	Features
Driving	Total mileage since the first use Total discharged energy since the first use Histogram of the daily covered distances Histogram of average driving speed per driving event Histogram of initial SoC per driving event *driving* matrix corresponding to the discharged energy with respect to temperature {$< 0^o$, $0^o - 15^o$, $15^o - 30^o$, $30^o - 45^o$, $> 45^o$} and SoC {0% – 10%, 10% – 30%, 30% – 70%, 70% – 90%, 90% – 100%}	*dr_gbl_Dist* *dr_gbl_EGY* *dr_dist_0_25*, *dr_dist_25_50*, *dr_dist_50_75*, *dr_dist_75_100*, *dr_dist_100_300* *dr_speed_0_30*, *dr_speed_30_50*, *dr_speed_50_90*, *dr_speed_90_130*, *dr_speed_130_200* *dr_soc_start_0_20*,*dr_soc_start_20_40*,*dr_soc_start_40_60*,*dr_soc_start_60_80*,*dr_soc_start_80_100*, *dr_soc_0_10_temp_0*, *dr_soc_10_30_temp_0*, *dr_soc_30_70_temp_0*, *dr_soc_70_90_temp_0*, *dr_soc_90_100_temp_0*, *dr_soc_0_10_temp_0_15*, *dr_soc_10_30_temp_0_15*, *dr_soc_30_70_temp_0_15*, *dr_soc_70_90_temp_0_15*, *dr_soc_90_100_temp_0_15*, *dr_soc_0_10_temp_15_30*, *dr_soc_10_30_temp_15_30*, *dr_soc_30_70_temp_15_30*, *dr_soc_70_90_temp_15_30*, *dr_soc_90_100_temp_15_30*, *dr_soc_0_10_temp_30_45*, *dr_soc_10_30_temp_30_45*, *dr_soc_30_70_temp_30_45*, *dr_soc_70_90_temp_30_45*, *dr_soc_90_100_temp_30_45*, *dr_soc_0_10_temp_45*, *dr_soc_10_30_temp_45*, *dr_soc_30_70_temp_45*, *dr_soc_70_90_temp_45*, *dr_soc_90_100_temp_45*
Charging	Total number of charging events since the first use Charging power State of charge variation charging temperature	nb_{ch} *ch_power_P3*, *ch_power_P7*, *ch_power_P11*, *ch_power_P22*, *ch_power_P43* *charge_d_soc_0_20*, *charge_d_soc_20_40*, *charge_d_soc_40_60*, *charge_d_soc_60_80*, *charge_d_soc_80_100* *charge_temp_m15*,*charge_temp_m15_0*,*charge_temp_0_15*,*charge_temp_15_30*,*charge_temp_30_45*,
Parking	*parking* matrix corresponding to the time spent on parking lots split with respect to temperature {$< 0^o$, $0^o - 15^o$, $15^o - 30^o$, $30^o - 45^o$, $> 45^o$} and SoC {0% – 10%, 10% – 30%, 30% – 70%, 70% – 90%, 90% – 100%} Total number of days since the first use	*pk_soc_0_10_temp_0*, *pk_soc_10_30_temp_0*, *pk_soc_30_70_temp_0*, *pk_soc_70_90_temp_0*, *pk_soc_90_100_temp_0*, *pk_soc_0_10_temp_0_15*, *pk_soc_10_30_temp_0_15*, *pk_soc_30_70_temp_0_15*, *pk_soc_70_90_temp_0_15*, *pk_soc_90_100_temp_0_15*, *pk_soc_0_10_temp_15_30*, *pk_soc_10_30_temp_15_30*, *pk_soc_30_70_temp_15_30*, *pk_soc_70_90_temp_15_30*, *pk_soc_90_100_temp_15_30*, *pk_soc_0_10_temp_30_45*, *pk_soc_10_30_temp_30_45*, *pk_soc_30_70_temp_30_45*, *pk_soc_70_90_temp_30_45*, *pk_soc_90_100_temp_30_45*, *pk_soc_0_10_temp_45*, *pk_soc_10_30_temp_45*, *pk_soc_30_70_temp_45*, *pk_soc_70_90_temp_45*, *pk_soc_90_100_temp_45* *days*

Table A.2: Selected features description for the data-driven SoH prediction.

Appendix A. Extracted features description

In this annexe the selected features description for the data-driven SoH prediction are presented in Table A.2.

References

[1] IEA. *Global EV Outlook 2018: Towards cross-modal electrification*. International Energy Agency, 2018.

[2] IEA. *Global EV Outlook 2019: Scaling-up the transition to electric mobility*. International Energy Agency, 2019.

[3] Boucar Diouf and Christophe Avis. The potential of Li-ion batteries in ECOWAS solar home systems. *Journal of Energy Storage*, 22:295–301, April 2019.

[4] Shruti Suriyakumar, Sivalingam Gopi, Murugavel Kathiresan, Suriyasarathi Bose, E. Bhoje Gowd, Jijeesh R. Nair, Natarajan Angulakshmi, Giuseppina Meligrana, Federico Bella, Claudio Gerbaldi, and A. Manuel Stephan. Metal organic framework laden poly(ethylene oxide) based composite electrolytes for all-solid-state Li-S and Li-metal polymer batteries. *Electrochimica Acta*, 285:355–364, September 2018.

[5] Kazunori Takada. Progress in solid electrolytes toward realizing solid-state lithium batteries. *Journal of Power Sources*, 394:74–85, August 2018.

[6] Jun Zhang, Chao Zheng, Jiatao Lou, Yang Xia, Chu Liang, Hui Huang, Yongping Gan, Xinyong Tao, and Wenkui Zhang. Poly(ethylene oxide) reinforced Li6ps5cl composite solid electrolyte for all-solid-state lithium battery: Enhanced electrochemical performance, mechanical property and interfacial stability. *Journal of Power Sources*, 412:78–85, February 2019.

[7] Xiaoyu Li, Zhenpo Wang, Lei Zhang, Changfu Zou, and David. D. Dorrell. State-of-health estimation for Li-ion batteries by combing the incremental capacity analysis method with grey relational analysis. *Journal of Power Sources*, 410-411:106–114, January 2019.

[8] Yi Li, Mohamed Abdel-Monem, Rahul Gopalakrishnan, Maitane Berecibar, Elise Nanini-Maury, Noshin Omar, Peter van den Bossche, and Joeri Van Mierlo. A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by Gaussian filter. *Journal of Power Sources*, 373:40–53, January 2018.

[9] M. Lucu, E. Martinez-Laserna, I. Gandiaga, and H. Camblong. A critical review on self-adaptive Li-ion battery ageing models. *Journal of Power Sources*, 401:85–101, October 2018.

[10] Gregory L. Plett. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 1. Background. *Journal of Power Sources*, 134(2):252–261, August 2004.

[11] Yuejiu Zheng, Minggao Ouyang, Xuebing Han, Languang Lu, and Jianqiu Li. Investigating the error sources of the online state of charge estimation methods for lithium-ion batteries in electric vehicles. *Journal of Power Sources*, 377:161–188, February 2018.

[12] Akram Eddahech, Olivier Briat, and Jean-Michel Vinassa. Performance comparison of four lithium–ion battery technologies under calendar aging. *Energy*, 84:542–550, May 2015.

[13] J. Vetter, P. Novák, M. R. Wagner, C. Veit, K. C. Möller, J. O. Besenhard, M. Winter, M. Wohlfahrt-Mehrens, C. Vogler, and A. Hammouche. Ageing mechanisms in lithium-ion batteries. *Journal of Power Sources*, 147(1):269–281, September 2005.

[14] Andrew W Thompson. Economic implications of lithium ion battery degradation for vehicle-to-grid (v2x) services. *Journal of Power Sources*, 396:691–709, 2018.

[15] M Berecibar, I Gandiaga, I Villarreal, N Omar, J Van Mierlo, and P Van den Bossche. Critical review of state of health estimation methods of li-ion batteries for real applications. *Renewable and Sustainable Energy Reviews*, 56:572–587, 2016.

[16] Elie Riviere. *Détermination in-situ de l'état de santé de batteries lithium-ion pour un véhicule électrique*. PhD thesis, Université Grenoble Alpes, l'Université Claude Bernard Lyon 1, 2016. Thèse de doctorat dirigée par Bultel, Yann et Venet, Pascal Mécanique des fluides, procédés, énergétique Grenoble Alpes 2016, 2016GREAI048.

[17] Uwe Tröltzsch, Olfa Kanoun, and Hans-Rolf Tränkler. Characterizing aging effects of lithium ion batteries by impedance spectroscopy. *Electrochimica Acta*, 51(8-9):1664–1672, 2006.

[18] Mohammadhosein Safari. *Vieillissement des batteries à ions lithium : étude expérimentale et modélisation*. PhD thesis, Amiens, January 2011.

[19] Eric Prada. *Aging modeling and lifetime optimization of Li-ion LiFePO4-graphite batteries according to the vehicle use.* Theses, Université Pierre et Marie Curie, November 2012.

[20] Matthieu Dubarry, Nan Qin, and Paul Brooker. Calendar aging of commercial li-ion cells of different chemistries–a review. *Current Opinion in Electrochemistry*, 2018.

[21] Mehdi Jafari, Khalid Khan, and Lucia Gauchia. Deterministic models of li-ion battery aging: It is a matter of scale. *Journal of Energy Storage*, 20:67–77, 2018.

[22] Akram Eddahech, Olivier Briat, and Jean-Michel Vinassa. Performance comparison of four lithium–ion battery technologies under calendar aging. *Energy*, 84:542–550, 2015.

[23] Jean-Michel Vinassa, Akram Eddahech, and Olivier Briat. Strategy for lithium-ion battery performance improvement during power cycling. In *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*, pages 6806–6811. IEEE, 2013.

[24] Martin Petit, Eric Prada, and Valérie Sauvant-Moynot. Development of an empirical aging model for Li-ion batteries and application to assess the impact of Vehicle-to-Grid strategies on battery lifetime. *Applied Energy*, 172:398–407, June 2016.

[25] Akram Eddahech, Olivier Briat, Hervé Henry, J-Y Delétage, Eric Woirgard, and J-M Vinassa. Ageing monitoring of lithium-ion cell during power cycling tests. *Microelectronics Reliability*, 51(9-11):1968–1971, 2011.

[26] Xianke Lin, Xiaoguang Hao, Zhenyu Liu, and Weiqiang Jia. Health conscious fast charging of li-ion batteries via a single particle model with aging mechanisms. *Journal of Power Sources*, 400:305–316, 2018.

[27] Qingxia Yang, Jun Xu, Xiuqing Li, Dan Xu, and Binggang Cao. State-of-health estimation of lithium-ion battery based on fractional impedance model and interval capacity. *International Journal of Electrical Power & Energy Systems*, 119:105883, 2020.

[28] Shuzhi Zhang, Baoyu Zhai, Xu Guo, Kaike Wang, Nian Peng, and Xiongwen Zhang. Synchronous estimation of state of health and remaining useful lifetime for lithium-ion battery using the incremental capacity and artificial neural networks. *Journal of Energy Storage*, 26:100951, 2019.

[29] Guijun Ma, Yong Zhang, Cheng Cheng, Beitong Zhou, Pengchao Hu, and Ye Yuan. Remaining useful life prediction of lithium-ion batteries based on false nearest neighbors and a hybrid neural network. *Applied Energy*, 253:113626, 2019.

[30] Sheng Shen, Mohammadkazem Sadoughi, Meng Li, Zhengdao Wang, and Chao Hu. Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. *Applied Energy*, 260:114296, 2020.

[31] Jinhao Meng, Lei Cai, Guangzhao Luo, Daniel-Ioan Stroe, and Remus Teodorescu. Lithium-ion battery state of health estimation with short-term current pulse test and support vector machine. *Microelectronics Reliability*, 88:1216–1220, 2018.

[32] Dave Andre, Christian Appel, Thomas Soczka-Guth, and Dirk Uwe Sauer. Advanced mathematical methods of soc and soh estimation for lithium-ion batteries. *Journal of power sources*, 224:20–27, 2013.

[33] Peiyao Guo, Ze Cheng, and Lei Yang. A data-driven remaining capacity estimation approach for lithium-ion batteries based on charging health feature extraction. *Journal of Power Sources*, 412:442–450, 2019.

[34] SONG Yuchen, LIU Datong, HOU Yandong, YU Jinxiang, and PENG Yu. Satellite lithium-ion battery remaining useful life estimation with an iterative updated rvm fused with the kf algorithm. *Chinese Journal of Aeronautics*, 31(1):31–40, 2018.

[35] Xianghui Qiu, Weixiong Wu, and Shuangfeng Wang. Remaining useful life prediction of lithium-ion battery based on improved cuckoo search particle filter and a novel state of charge estimation method. *Journal of Power Sources*, 450:227700, 2020.

[36] Jun Bi, Ting Zhang, Haiyang Yu, and Yanqiong Kang. State-of-health estimation of lithium-ion battery packs in electric vehicles based on genetic resampling particle filter. *Applied energy*, 182:558–568, 2016.

[37] S Khaleghi, Y Firouz, J Van Mierlo, and P Van den Bossche. Developing a real-time data-driven battery health diagnosis method, using time and frequency domain condition indicators. *Applied Energy*, 255:113813, 2019.

[38] Yi Li, Changfu Zou, Maitane Berecibar, Elise Nanini-Maury, Jonathan C-W Chan, Peter van den Bossche, Joeri Van Mierlo, and Noshin Omar. Random forest regression for online capacity estimation of lithium-ion batteries. *Applied energy*, 232:197–210, 2018.

[39] Akram Eddahech, Olivier Briat, and Jean-Michel Vinassa. Determination of lithium-ion battery state-of-health based on constant-voltage charge phase. *Journal of Power Sources*, 258:218–227, 2014.

[40] Peter Keil and Andreas Jossen. Calendar Aging of NCA Lithium-Ion Batteries Investigated by Differential Voltage Analysis and Coulomb Tracking. *Journal of The Electrochemical Society*, 164(1):A6066–A6074, 2017.

[41] Limei Wang, Chaofeng Pan, Liang Liu, Yong Cheng, and Xiuliang Zhao. On-board state of health estimation of lifepo4 battery pack through differential voltage analysis. *Applied energy*, 168:465–472, 2016.

[42] Linfeng Zheng, Jianguo Zhu, Dylan Dah-Chuan Lu, Guoxiu Wang, and Tingting He. Incremental capacity analysis and differential voltage analysis based state of charge and capacity estimation for lithium-ion batteries. *Energy*, 150:759–769, 2018.

[43] Yi Li, Kailong Liu, Aoife M Foley, Alana Zülke, Maitane Berecibar, Elise Nanini-Maury, Joeri Van Mierlo, and Harry E Hoster. Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. *Renewable and Sustainable Energy Reviews*, 113:109254, 2019.

[44] Wladislaw Waag, Stefan Käbitz, and Dirk Uwe Sauer. Experimental investigation of the lithium-ion battery impedance characteristic at various conditions and aging states and its influence on the application. *Applied energy*, 102:885–897, 2013.

[45] William S Cleveland and Clive Loader. Smoothing by local regression: Principles and methods. In *Statistical theory and computational aspects of smoothing*, pages 10–49. Springer, 1996.

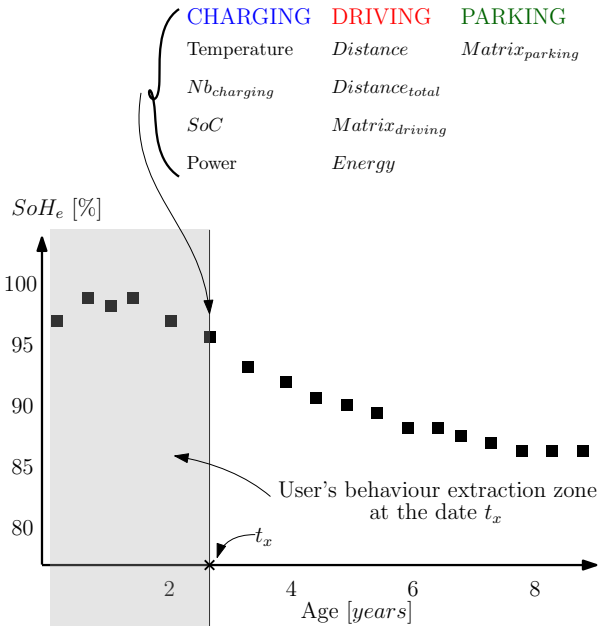
[46] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[47] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

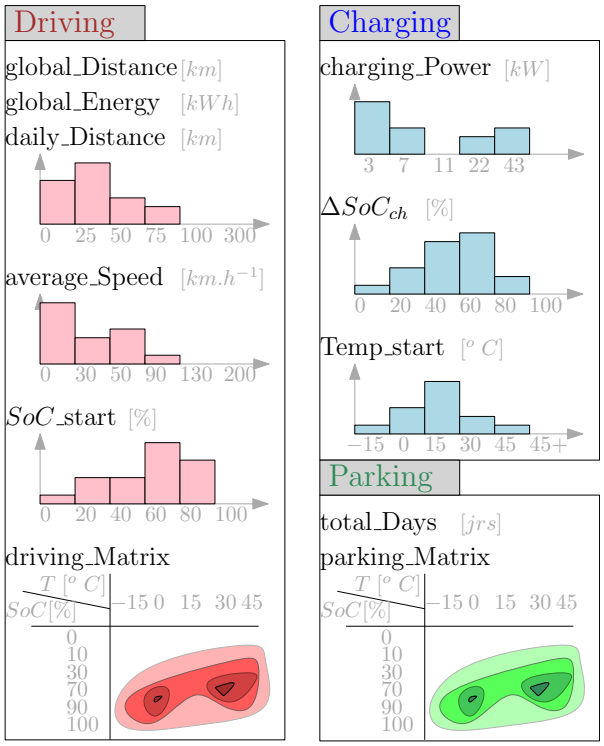
[48] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[49] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

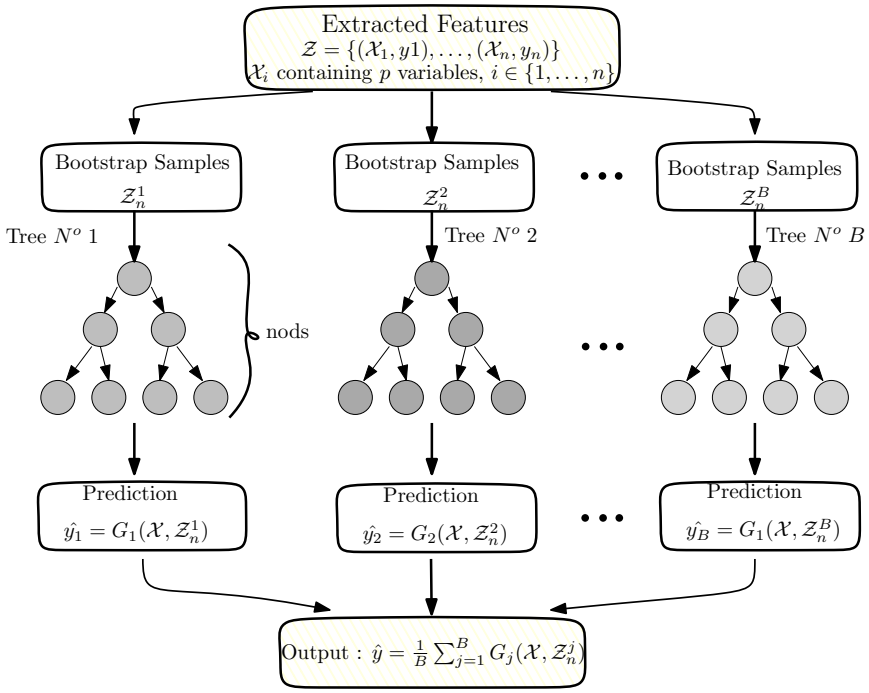
[50] Samuel Pelletier, Ola Jabali, Gilbert Laporte, and Marco Veneroni. Battery degradation and behaviour for electric vehicles: Review and numerical analyses of several models. *Transportation Research Part B: Methodological*, 103:158–187, September 2017.



(5.1) User's behavior monitoring at each $\hat{SoH}$ estimation point

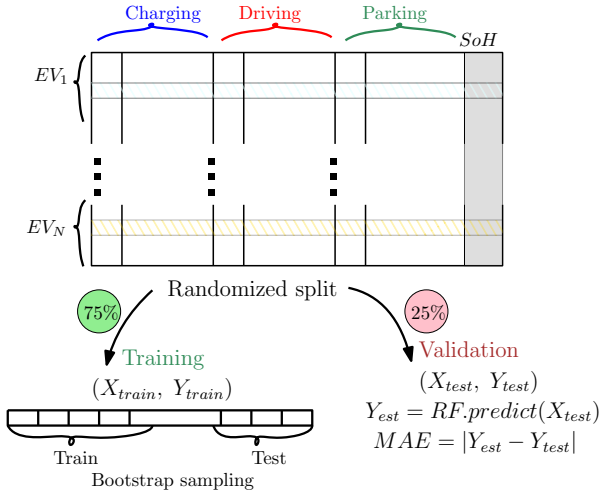


(5.2) Selected features illutration for the data-driven SoH prediction.

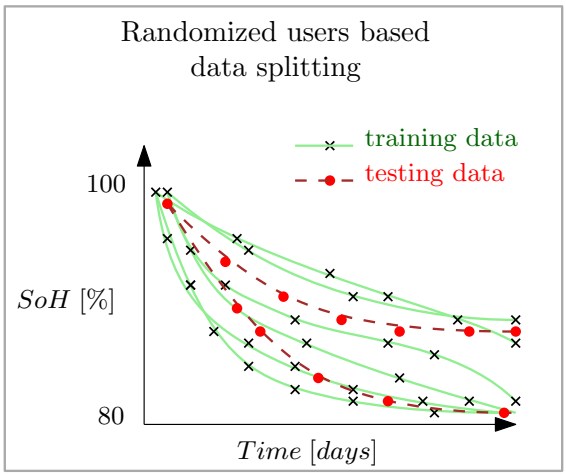


(5.3) Random forest diagram illustration

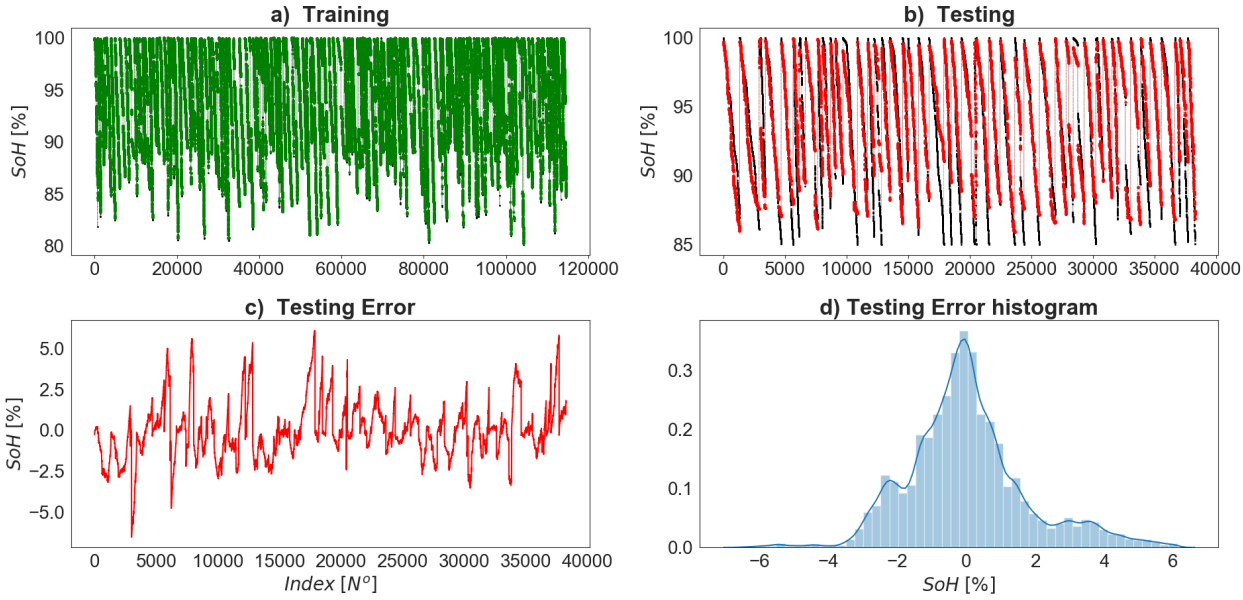
Figure 5: Data preparation and Random Forest illustration .



(6.1) Illustration of the data splitting into traning and validation data sets. The split is done with respect to users.



(6.2) Data splitting into training and testing subsets. Each curve belongs to a random user.



(6.3) Data splitting into traning and validation data sets. The split is done with respect to users.

Figure 6: Data preparation and Random Forest results.

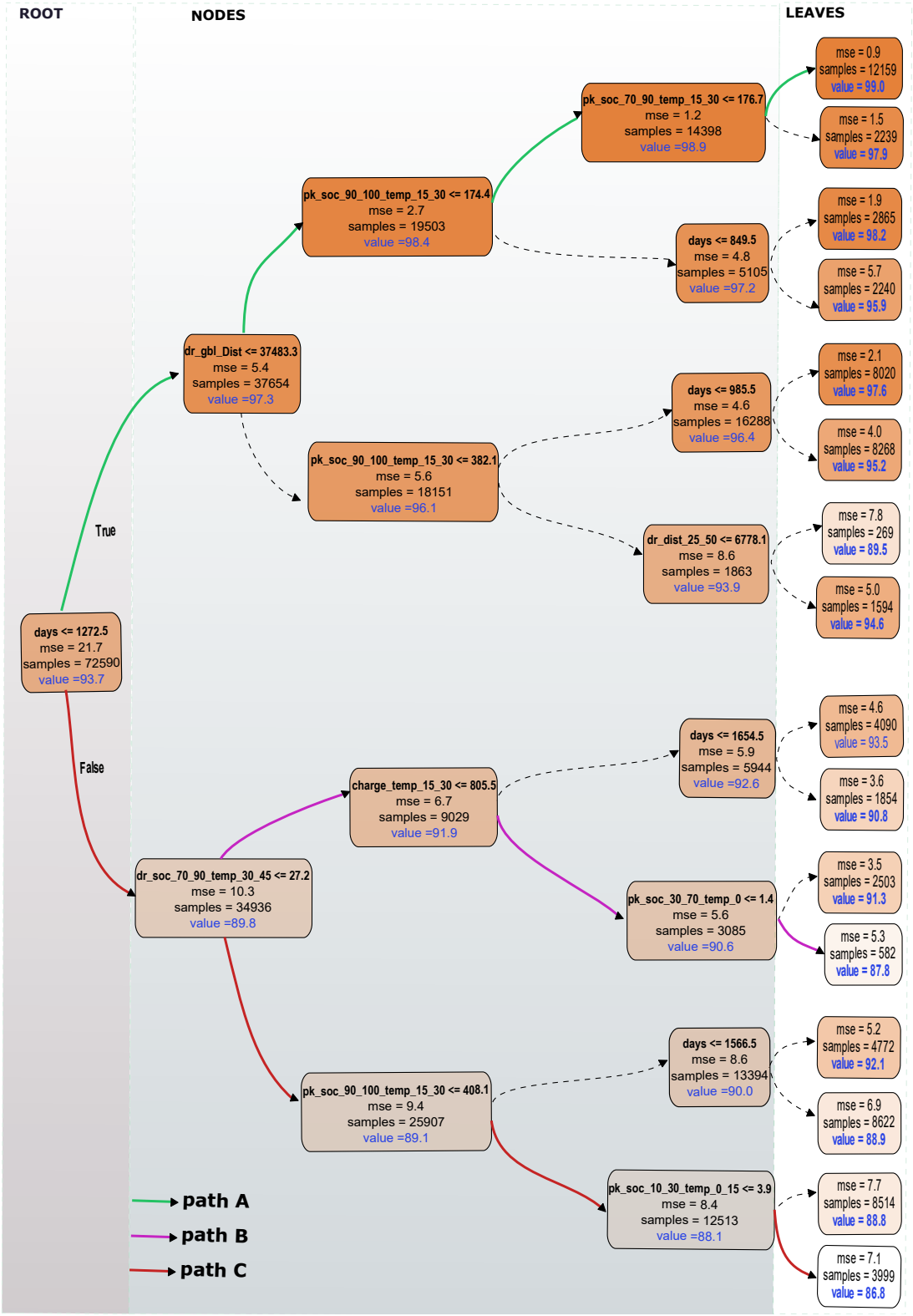
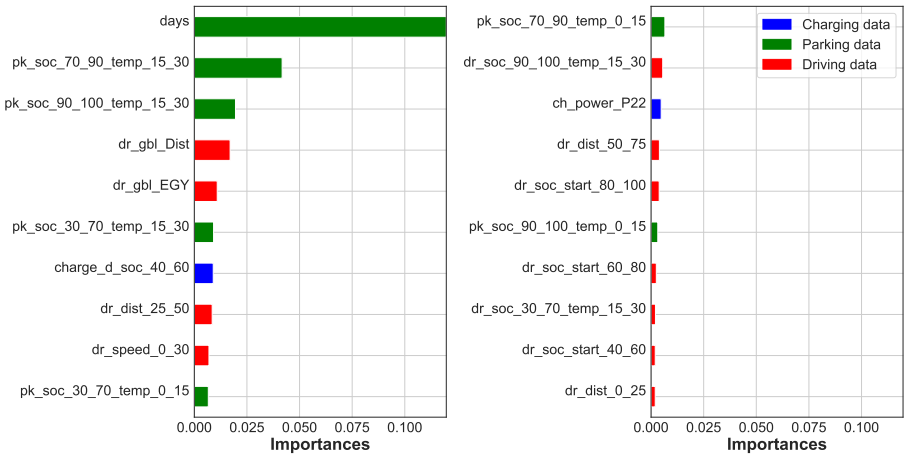
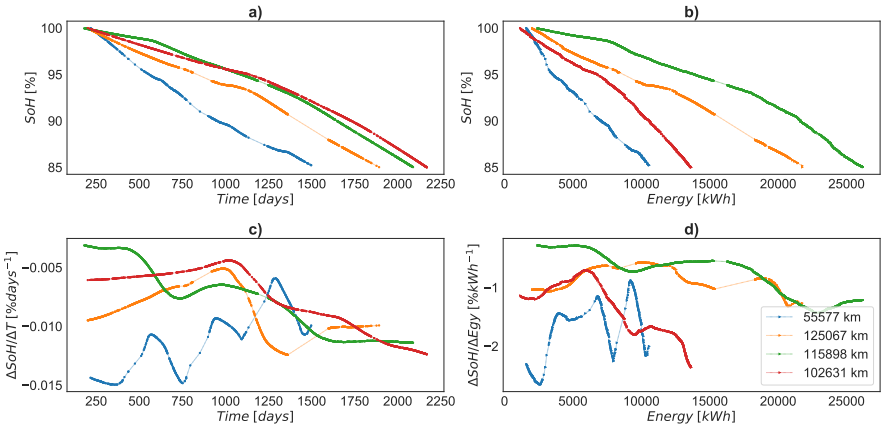


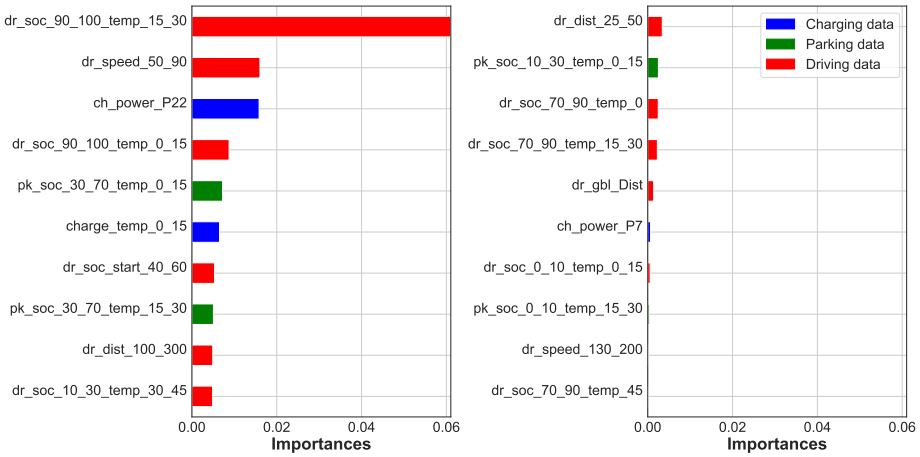
Figure 7: Simplified tree example using RF algorithm. The considered data split depth is 4.



(8.1) Variable importance.



(8.2) RF simple tree example with 3 data split depth.



(8.3) Variable importance.

Figure 8: Illustration and understanding of variables importance.