


Ensemble machine learning based driving range estimation for real-world electric city buses by considering battery degradation levels

Yongxing Wang¹ | Chaoru Lu²  | Jun Bi³ | Qiuyue Sai³ | Yongzhi Zhang⁴

¹ Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway

² Department of Civil Engineering and Energy Technology, Oslo Metropolitan University, Oslo, Norway

³ School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China

⁴ School of Automotive Engineering, Chongqing University, Chongqing, China

Correspondence

Chaoru Lu, Department of Civil Engineering and Energy Technology, Oslo Metropolitan University, Oslo, 0166, Norway.

Email: chaorulu@oslomet.no

Funding information

National Natural Science Foundation of China, Grant/Award Number: 71961137008; Research Council of Norway, Grant/Award Number: 299078

Abstract

Battery electric buses (BEBs) have been regarded as effective options to address the congestion and pollution problems in the field of urban transportation. However, since the limited driving range of BEBs brings challenges for their promotion, the accurate estimation of the driving range with limited available information has become a critical issue for public transport operators. The real-world data collected from 50 BEBs operated in two different cities is used to develop the driving range estimation method by considering the battery degradation effects. Firstly, the incremental capacity analysis method is introduced to characterize the battery performance, and the battery degradation levels under different charging modes are recognized. Afterward, four types of ensemble machine learning (EML) methods are adopted to model the driving range estimation. The BEB driving data, weather condition data and battery degradation levels are used to train and test the models with consideration of 17 impact factors together with two different charging modes. The results indicate that the ensemble machine learning methods have good performance overall, of which the random forest has the highest accuracy. Furthermore, the importance of influencing factors is analysed, and the relevant insights are discussed.

1 | INTRODUCTION

Traffic congestion and environmental pollution have become the major issues experienced by many cities all over the world with urbanization. In response to the dual pressure of congestion and pollution, battery electric buses (BEBs) have been continuously introduced to public transit systems because of their environmental benefits, such as, zero tailpipe emission. Unlike conventional buses that use fossil fuel, BEBs fully powered the energy stored in the battery, which can be charged by using renewable energy sources. Therefore, system-level emissions are also reduced. However, due to the technical limitation of existing batteries, the driving range of BEBs is significantly shorter than that of conventional buses. The limited driving range brings new challenges to transit system operation and even hinders the further promotion of BEBs. In order to guarantee the service quality and maximize the benefit of BEB, it is

necessary for public transport operators to attain accurate estimations of the driving range and develop proper operational strategies of BEB fleets. Meanwhile, the driving range of BEBs is expectedly affected by the battery performance that tends to deteriorate with aging [1]. As battery performance gradually degrades, public transport operators need to promptly adjust the bus scheduling and even replace the battery to eliminate the adverse impact of battery degradation on driving range variation. Thus, the battery degradation must be considered in the driving range estimation.

Driving range limitation is a critical problem for BEBs, which would further impact the efficiency of public transit systems. Focused on this issue, many efforts have been devoted to accurate estimation of the driving range or related battery state for electric vehicles (EVs) based on various methods [2]. Nonetheless, there are still several issues that should be concerned about the driving range estimation of BEBs, as follows.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *IET Intelligent Transport Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

1. Most of the existing methods concentrate on the driving range estimation for passenger EVs, and neglect the unique features deriving from the vehicle operating environment of BEBs due to lack of data collected from real-world operated BEBs. As a matter of fact, the mobility pattern of passenger EVs has significant differences with the pattern of BEBs, which often operate on the predetermined routes along bus stations. This mobility pattern would exert influence on the driving range of BEBs.
2. Traditional methods usually overlook the impacts of battery degradation on the accuracy of driving range estimation. Battery performance would continuously deteriorate in a long-time window as the vehicle accumulated mileage increases gradually. Therefore, an effective method is needed to assess the long-term effect of the battery performance variation, and the results can be used to improve the driving range estimation.
3. Several previous studies attempt to use data-driven methods to investigate the impact of battery state on driving range, whereas these methods are mainly based on the laboratory test data and unable to reflect the real-world operating environment of EVs, especially for the electric city buses.

In recent years, with the development of sensing technology, a considerable amount of operation data for electric city buses could be collected. Unlike the laboratory test data, the real-world data has ability to embody the effects of actual complex scenarios on BEB operation. To bridge the aforementioned gaps, this study aims to use the BEB operation data collected from two city areas in China to gain deeper insight into the battery degradation and driving range estimation. Besides the operation data of electric city buses, the data of weather conditions would also be employed because the weather and environmental conditions have significant influence on the driving range of EVs [3].

Specifically, the contributions of this study are as follow. Firstly, the real-world data of 50 BEBs operating in two city areas with different charging modes is collected to investigate the battery performance and driving range. The driving pattern and battery status during both charging and driving processes are recorded in the dataset, and the data of weather conditions is also collected to integrate environmental factors into the investigation. The driving pattern often shows as the vehicle parameters during operation, such as driving speed and its variation. The battery status is manifested as the battery parameters for the duration of BEB operation, such as voltage and current. More importantly, in contrast to the data of passenger EVs, the BEB operation data has better capacity to reflect the unique features of electric city buses. Thus, the driving range estimation using BEB operation data has better adaptability to the public transit system with electric city buses. Secondly, in view of the battery degradation effects on driving range, a data-based framework is developed to recognize the battery degradation, which aims to consider the battery degradation levels while estimating driving range. In this framework, the incremental capacity analysis (ICA) method is introduced to characterize the battery performance, and subsequently the battery

degradation levels are recognized. Finally, four types of ensemble machine learning (EML) methods are adopted to model the driving range estimation under different charging modes. Combining the battery degradation levels, the data of BEB driving processes and of weather conditions are used to train and verify the models with consideration of 17 impacting factors. Furthermore, the importance of influencing factors is analysed, and then the relevant insights are discussed.

The remaining portions of this paper are arranged as follows: In Section 2, the literature review is presented. Section 3 presents the collection and processing of the data. In Section 4, the framework for battery degradation recognition is developed to obtain the battery degradation levels. Section 5 introduces the driving range estimation models and analyse the influencing factors. Lastly, the conclusions and future studies are discussed in Section 6.

2 | LITERATURE REVIEW

The limited driving range has been regarded as the main barrier for the promotion of BEBs or other types of EVs. An effective method to alleviate the negative effects of this problem, such as, range anxiety, is the accurate estimation of driving range. Given the fact that the driving range is related to the battery capacity, there are many studies focused on the estimation of state-of-charge (SOC) that represents the ratio of the remaining capacity to the nominal capacity. For example, Xu et al. [4] and Shao et al. [5] used the adaptive filter algorithms to realize SOC estimation. Several studies adopted machine learning methods to improve the SOC estimation [6–8]. However, driving range estimation is a much more complex and challenging issue in comparison with SOC estimation. This is because that the driving range would be affected by many other internal and external factors, besides SOC, such as driving patterns and weather conditions [9]. Therefore, the existing methods regarding SOC estimation are not competent to accurately estimate the driving range.

In view of the complexity of driving range estimation, a considerable amount of research has been conducted to consider more impact factors from different aspects. For example, several studies took into account both of battery capacity and driving speed to improve the estimation accuracy [10–12]. Yuan et al. [13] and Vatanparvar et al. [14] further discussed the impacts of driving patterns on driving range, and then established the models to attain the longest distance that EVs can continually run under various environments. Moreover, with the development of sensing techniques, more related works have put focus on driving range estimation based on data-driven approaches with huge amount of data. Bi et al. [15] utilized the real-world data collected from an operated EV to conduct a study on driving range estimation, where the SOC, speed and temperature conditions would be considered as impacting factors. Lee and Wu [16] proposed a data-based framework to estimate driving range of EVs based on powertrain simulation and driving behaviour analysis. Sun et al. [17] introduced a machine learning algorithm to model the driving range estimation using the EV discharge data coupled with weather information. Furthermore, Bi

et al. [18] considered the driving range estimation as a nonlinear system affected by battery factors and vehicle operating status factors. Subsequently, the radial basis function neural network was employed to model the driving range estimation. In addition, Amirkhani et al. [19] compared the performance of several machine learning methods for the purpose of driving range estimation based on the publicly available dataset. It is worth noting that, as alluded in the aforementioned studies, most of the previous methods aimed to estimate the driving range by considering several factors, whereas they ignored the battery aging effects. In real-world scenarios, the battery degradation would occur as the driving cycle increases and thus exerts influence on the driving range of EVs [20]. For this reason, putting the battery degradation factor in driving range estimation could be regarded as a useful approach to improve accuracy and practicability.

In the previous studies, even though less attention has been focused on considering the battery degradation in driving range estimation, there have existed several studies that attempt to combine the battery aging features with SOC estimation, and still achieve insightful findings [21–23]. In these studies, the state-of-health (SOH) was often applied to represent the battery health condition, which refers to the aging condition of a battery and its ability to deliver the specified performance comparing to the fresh one. Nevertheless, the data used for model training was collected from the laboratory environment, which has limited ability to reflect real-world driving and charging for EV operation. Meanwhile, although the SOH is a commonly-used indicator to define battery degradation, it only applies to the methods that use the data from a special experiment. For the data collected from real-world scenarios, such an indicator may not be suitable for battery aging monitoring, because some specific parameters for SOH derivation could not be detected during EV operating, such as the internal resistance. Fortunately, the ICA has ability to identify the battery degradation by capturing the peak values from IC curves. This feature makes the ICA suitable for application in BEB operating data. In recent years, the ICA has attracted increasing attention and several studies have provided the ICA-based schemes to realize battery aging monitoring [24–26]. Furthermore, He et al. [27] conducted a comparative study with respect to the ICA-based methods and analysed the application boundary. The results from these studies demonstrate that the ICA is an effective method for battery aging monitoring. In addition, the research results for battery degradation could also be applied in other fields besides driving range estimation. For instance, Liu et al. [28] integrated the battery aging effects into charging pattern optimization. Another study further concluded that the battery degradation has significant influence on battery charging management [29].

Overall, even though the previous studies have made fruitful achievements in the estimation of driving range or related battery status for EVs, there are still some limitations, as mentioned above. More importantly, most of the existing studies have put their focus on the passenger EVs, while less attention has been devoted to the BEBs that are applied in the urban transit system. Unlike passenger EVs, BEBs often have their particular operating environment and driving patterns, which further brings significant effects on the driving range estimation. Given such

TABLE 1 Specifications of the studied electric city buses

Attribute name	Value (Units)
Curb weight	7300 kg
Vehicle dimensions	8005 × 2350 × 3105 mm
Maximum vehicle speed	69 km h ⁻¹
Battery nominal capacity	240 Ah
Peak motor power	100 kW
Battery cathode materials	LiFePO ₄

limitations, the driving range estimation models coupled with battery degradation recognition for electric city buses are developed by using the real-world data collected from two city areas in China. The data sources used in this study include the BEB operation data and weather condition data. The methods aim to provide a long-term scheme for accurate estimation of BEB driving range with consideration of the battery aging effects.

3 | DATA COLLECTION AND PREPROCESSING

In this study, the real-time operation data of BEBs is used to investigate the driving range and battery degradation of electric city buses. The 50 BEBs with same specifications and operated in two different cities in China (25 BEBs in each city) were employed to carry out the research. The detailed specifications of the studied electric city buses are listed in Table 1.

The BEBs operate together with conventional buses to satisfy the daily passenger need by following the predetermined schedules of public transport system. During May 2017 to July 2018, the data of BEBs, which contains battery status and driving patterns, was collected and transmitted to the vehicle service centre every 10s interval.

The raw dataset for each BEB includes almost 2 million data points and records the accumulated mileage varies from 6000 to 130,000 km. However, due to the external disturbance during data transmission, missing value and error universally exist in the raw dataset. To deal with such a problem, interpolation methods are adopted to compensate the missing data points and adjust the outliers. Moreover, according to the vehicle status, the operation data is divided into two groups, that is, charging data and driving data. Specifically, the charging data mainly tracks the battery status during the charging processes. Table 2 presents an example of a charging segment, which covers 20 min within a charging process. For the charging data, the driving speed recorded in all the data points equals to zero and accumulated mileage remains constant for each charging process.

Different from the charging data, the driving data mainly tracks the driving patterns and battery status when the BEBs operated in the road network and meanwhile the battery system was experiencing the discharging processes. Table 3 shows an example of a driving segment with 30 km travel distance.

Besides the operation data of BEBs, the weather data source has also been adopted to improve the driving range estimation,

TABLE 2 An example of charging segment in the raw charging data

Timestamp	SOC (%)	Battery voltage (V)	Battery current (A)	Battery highest temperature (°C)	Battery lowest temperature (°C)	Accumulated mileage (km)	Driving speed (km h ⁻¹)
20171019000711	32	561.1	99	36	30	45667.1	0
20171019000721	32	561.3	98	36	30	45667.1	0
...
20171019002701	47	562.3	99	38	32	45667.1	0
20171019002711	47	562.3	99	38	32	45667.1	0

TABLE 3 An example of driving segment in the raw driving data

Timestamp	SOC (%)	Battery voltage (V)	Battery current (A)	Battery highest temperature (°C)	Battery lowest temperature (°C)	Accumulated mileage (km)	Driving speed (km h ⁻¹)
20171018214749	42	546.8	42	38	32	45637.1	50
20171018214759	42	546.9	42	38	32	45637.2	51
...
20171018225609	32	549.3	6	37	31	45667.1	3
20171018225619	32	549.4	4	37	31	45667.1	0

in view of the impacts of external factors on energy consumption. The weather data was acquired from the meteorological science data centre, which records the historical information of daily weather conditions from May 2017 to July 2018. Table 4 provides one sample of the weather records in the dataset.

In the following sections, the charging data would be employed to capture the battery aging characteristics and recognize the battery degradation levels. Afterwards, the BEB operation data were coupled with weather data to develop driving range estimation method, which takes the battery degradation levels into consideration. The BEB operation data is collected from two medium-sized cities from China and the proposed method could be generalized in other similar city areas. Furthermore, the two city areas provided different charge modes to BEB charging, which include fast charging and slow charging modes. Such a feature would be utilized to investigate the impacts of charge modes on battery performance degradation. Note that, the data would be further processed to build the input-output dataset for driving range estimation model, which would be discussed in Section 5.1.

4 | BATTERY DEGRADATION RECOGNITION

4.1 | Battery performance characterization based on incremental capacity analysis

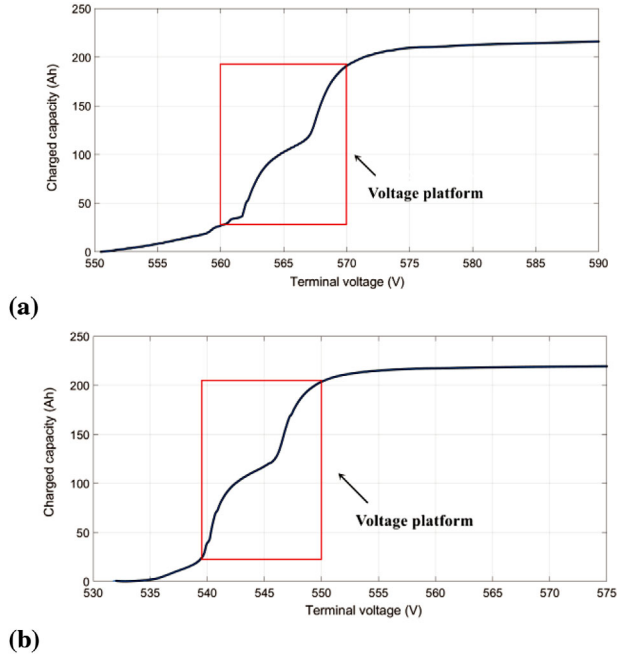
Based on the charging data, we use the ICA method to characterize the battery performance. The ICA method is an in-situ and non-destructive electrochemical technique, which can detect the gradual variation in battery aging behaviour through investigating the evolution of the resulting IC curves with

cycling [30]. Usually, the battery aging assessment has been studied based on the laboratory-collected datasets obtained from accelerated aging tests under certain well-controlled conditions, such as, constant temperatures and low current operations [31]. However, these laboratory-based aging tests cannot sufficiently reflect the real-world operating scenarios of BEBs and thus fail to be used in practical driving environments. Fortunately, the ICA method has been confirmed to be applied in the BEB's battery packs with large current operations and real-world temperature environments [32]. An IC curve is derived by differentiating the voltage relative to the charged capacity under the constant-current-constant-voltage (CC-CV) regime. From the IC curve, the IC peaks can be observed, which are directly indicative of battery health. However, the charging current would be disturbed by the external environment and exhibits the variation in different charging events. In order to appropriately use the ICA technique, the data points with stable charging current need to be extracted. Accordingly, the charging current of 98–99 A and 24–25 A are selected to depict the IC curves for fast charging and slow charging modes, respectively. That is to say, the charging current rates for the fast charging and slow charging modes are respectively close to C/3 and C/10, which are the allowable charging currents for the ICA application, according to the literature [33]. Moreover, to obtain the typical IC curves, the charging processes with SOC being larger than 40% are extracted from the charging data. For example, Figure 1 presents the charging processes for fast charging and slow charging modes, respectively.

It can be observed from Figure 1 that the charged capacity increases as the terminal voltage raises, and there exist the sharp rises in charged capacity between the narrow intervals of terminal voltage. The terminal voltages in such intervals are defined as the voltage platforms, as highlighted in Figure 1, which is

TABLE 4 Record sample of a weather data point

Date	Average air temperature (°C)	Humidity (%)	Wind speed (m s ⁻¹)	Wind direction (i)	Visibility (km)	Cloud amount (%)	Air pressure (hPa)	Rain (mm)
2017-6-1	29.5	76	3.6	183.3	20.9	80	994	0

**FIGURE 1** Example charging processes. (a) An example charging process under fast charging mode. (b) An example charging process under slow charging mode

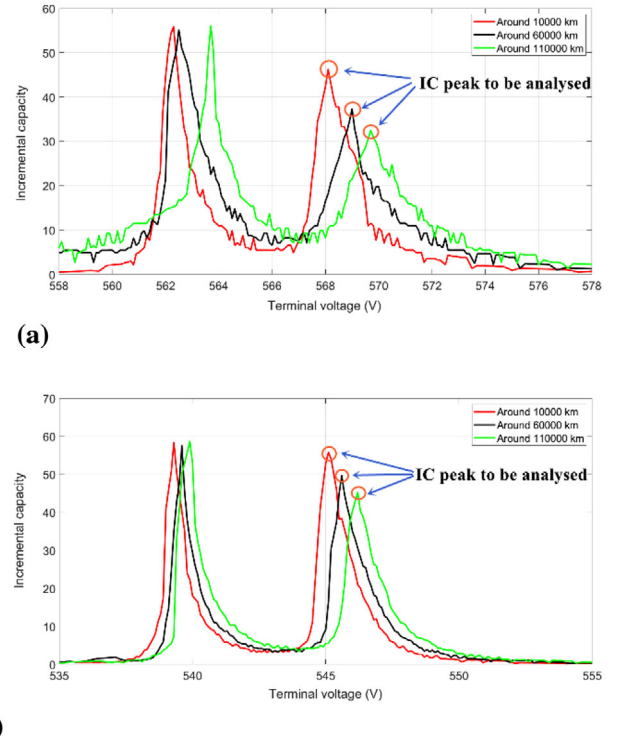
an inherent attribute of Li-ion batteries and able to be used to analyse the battery aging features. By using the ICA method, the curves of charged capacity versus terminal voltage can be transformed into the IC curves that illustrate the relationship between incremental capacity and terminal voltage. Mathematically, the incremental capacity is the ratio of the increment of charged capacity to the fixed increment of terminal voltage, as shown in Equation (1).

$$IC_n = \frac{\Delta Q_n}{\Delta V} = \frac{Q_n - Q_{n-1}}{\Delta V}, \quad (1)$$

where, IC_n denotes the incremental capacity at the voltage step n ; Q_n and Q_{n-1} represent the battery capacity at the voltage steps n and $n-1$, respectively; ΔV denotes the fixed increment of the terminal voltage; and, ΔQ_n represents the charged capacity at the voltage steps n . Moreover, based on the charging current available in the dataset, the battery capacity Q can be computed by Equation (2).

$$Q = \int_0^T I(t) dt, \quad (2)$$

where, Q denotes the charged capacity; and, $I(t)$ denotes the charging current recorded in the data.

**FIGURE 2** Example IC curves. (a) Example IC curves under the fast charging mode. (b) Example IC curves under the slow charging mode

Specifically, to obtain the incremental capacity through the charging data, the data points with consecutive and identical terminal voltages for each charging event are aggregated into one data point, and the corresponding charged capacity is accumulated. By tracking the incremental capacity associated with the voltage steps, the IC curve is generated. To improve the interpretability of the IC curve and better capture the IC peak values, the cubic smoothing spline is employed to smooth the IC curves [34]. Figure 2 illustrates the example IC curves under fast and slow charging modes, respectively. Three IC curves with different aging statuses are presented for each charging mode, and the approximate accumulated mileages are employed to reflect the aging statuses in the figures, according to the fact that the BEB's battery ages as the accumulated mileage increases.

Depending on the terminal voltage profile, the incremental capacity presents similar changing trends for both fast and slow charging modes, and two main peak values could be observed in the IC curves, as shown in Figure 2. In this study, the second IC peaks are analysed, because they have close relation with the loss of lithium inventory that is considered as the primary cause of LiFePO₄ battery degradation [35]. The reason for generating IC peaks is that the voltage steps from voltage platforms

may result in larger capacity increments than the ones resulting from other voltage steps according to the CC-CV regime, as expected from the Equation (1). Meanwhile, from the perspective of battery system, IC peaks are the consequence of the convolution of the electrochemical reactions in the active positive and negative electrode materials. With the battery aging process, the IC peak values would decrease gradually, and thus can be used to assess the battery degradation, as highlighted in Figure 2. Besides IC peak values, voltage platforms also have a certain ability to reflect the battery degradation. This is because the voltage platforms corresponding to the IC peaks tend to increase due to the resistance variation resulting from battery aging. Consequently, the IC peak values and corresponding voltage platforms from 3506 charging processes are obtained based on the dataset of the fast charging mode. And, 2981 IC peak values and corresponding voltage platforms are attained through the dataset of the slow charging mode.

4.2 | Recognition of battery degradation levels

Based on the data obtained through ICA method, the unsupervised machine learning methods are used to recognize the battery degradation levels with consideration of the accumulated mileage. In this way, the data points with similar battery performance are organized into the same cluster, while the data points belonging to different clusters are distinctive, accompanied with the variation of accumulated mileage. As one of the classical unsupervised machine learning methods, k-means clustering algorithm has been widely used in different fields due to its advantages on iterative optimization and computational complexity [36]. This algorithm classifies the data without labels into k different clusters considering the dissimilarity among data points, which is suitable to mine the ICA-based data, and thus employed to deal with the battery degradation recognition. Note that, the battery performance is generally characterized by the capacity, which would decrease as the battery ages. However, the battery capacity cannot be directly measured through BEB operation data. For a BEB, the battery degradation occurs as the accumulated mileage increases to a certain degree. Therefore, this study concentrates on the battery degradation levels based on the IC peak values with various accumulated mileages. The specific capacity with battery aging is not the focus in this study. Moreover, due to the dimension difference among the related variables, the standardized processing is applied in the dataset. Let vector space $X\{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_m\}$ denote the dataset, where the element \mathbf{x}_i represents the i th data point that includes three sub-elements: IC peak value p_i , voltage platform v_i , and accumulated mileage l_i . Equation (3) presents the mathematics for standardization by taking the IC peak value as an example.

$$p_i^o = \frac{p_i - \min(p)}{\max(p) - \min(p)}, \quad (3)$$

where, p_i^o represents the standardized IC peak value recorded in i th data point. $\min(x)$ and $\max(x)$, respectively, denote the max-

imum and minimum values of IC peak. By this way, the original data points are transformed into the nondimensional ones. Afterward, these standardized data points would be used to recognize battery degradation levels using k-means clustering algorithm. Let vector space $C\{\mathbf{c}_1, \dots, \mathbf{c}_j, \dots, \mathbf{c}_k\}$ be the cluster centres, where the element \mathbf{c}_j represents the j th cluster centre that has three corresponding child elements regarding to IC peak value p_j^c , voltage platform v_j^c and accumulated mileage l_j^c . Given the predetermined number of clusters k , the objective function of the k-means clustering algorithm for battery degradation recognition is

$$J(Z, C) = \sum_{j=1}^k \sum_{i=1}^m z_{ij} \left[\left(p_i^o - p_j^c \right)^2 + \left(v_i^o - v_j^c \right)^2 + \left(l_i^o - l_j^c \right)^2 \right], \quad (4)$$

where z_{ij} is a binary variable (i.e. $z_{ij} \in \{0, 1\}$) that equals to 1 if the i th data point belongs to j th cluster; otherwise, it equals to 0. Z indicates the set of z_{ij} . The k-means clustering algorithm aims to minimize the objective function $J(Z, C)$, and the cluster centres are updated through iterative optimization. The detailed principles and procedures of k-means clustering algorithm is referenced in related literature [37].

Note that, for the k-means clustering algorithm, how to determine the number of clusters k is a critical issue. That is, how many significant variations of battery performance occur during the battery aging process? In this section, the gap statistic approach is introduced to obtain the optimal number of clusters k , which is an effective method for estimating the number of clusters and applicable to virtually any kinds of clustering approaches. Mathematically, let D_j denote the sum of pairwise gap for all the data points that belong to the cluster j , which is obtained as shown in Equation (5)

$$D_j = \sum_{x_i, x_{i'} \in c_j} \left[\left(p_i^o - p_{i'}^o \right)^2 + \left(v_i^o - v_{i'}^o \right)^2 + \left(l_i^o - l_{i'}^o \right)^2 \right]. \quad (5)$$

Subsequently, standardize the sum of pairwise gaps under a specific number of cluster k , as given in Equation (6).

$$W_k = \sum_{j=1}^k \frac{1}{2m_j} D_j, \quad (6)$$

where W_k represents the standardized result for D_j under the number of cluster k , and m_j denotes the number of data points belonging to the cluster j . Based on the results, the gap statistic approach aims to compare the $\log(W_k)$ with its expectation under a reference distribution of data, as shown in Equation (7), and search the value of k for which the $\log(W_k)$ has the farthest decline.

$$\text{Gap}(k) = E\{\log(W_k)\} - \log(W_k), \quad (7)$$

where, $E\{\log(W_k)\}$ represents the expectation of $\log(W_k)$, and $\text{Gap}(k)$ is the gap statistic value for k . Furthermore, a specific inequation is used to effectively determine the optimal value of

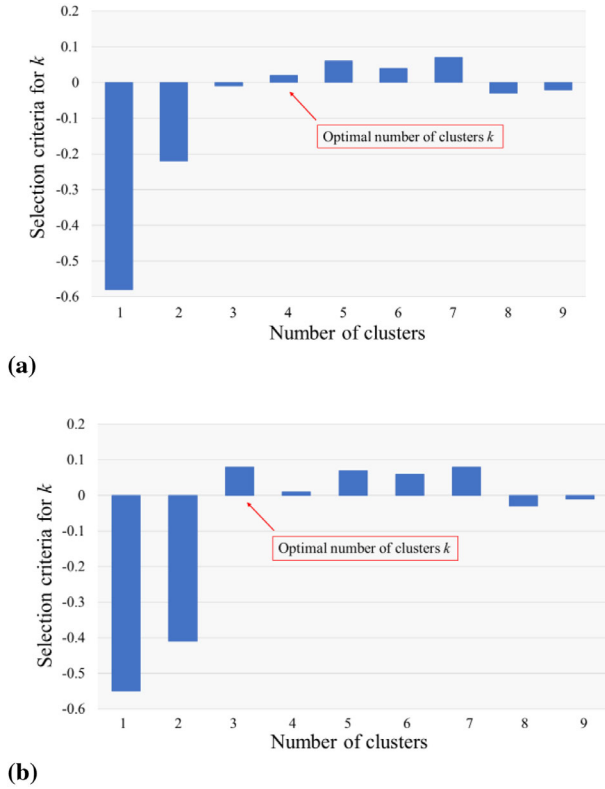


FIGURE 3 Optimal number of clusters k . (a) Optimal number of clusters k under fast charging mode. (b) Optimal number of clusters k under slow charging mode

k , as shown in Equation (8), and the value of k which fits the inequation for the first time is regarded as optimal one.

$$\text{Gap}(k) \geq \text{Gap}(k+1) - \sigma_{k+1}, \quad (8)$$

where σ_{k+1} denotes the unbiased standard deviation for W_{k+1} . More detailed procedures with respect to the gap statistic approach is referenced in related literature [38].

In view of the data characteristics, the framework, which comprises of k -means clustering algorithm and gap statistic approach, is employed to recognize battery degradation levels under different charging modes. Figure 3 illustrates the optimal number of clusters k under two different charging modes, obtained by the gap statistic approach. In the figure, the selection criteria for k is the difference of gap statistic value, and the optimal number of clusters k corresponds to the selection criteria value that is greater than zero for the first time, according to Equation (8). Obviously, the number of clusters k is respectively equal to 4 and 3 under the fast and slow charging modes. Such a result indicates that the fast and slow charging modes bring different effects on the battery performance.

Based on the optimal number of clusters k , the k -means clustering algorithm is applied to recognize the battery degradation, and the recognition results are presented in Figure 4. In the figure, the levels of battery degradation under fast and slow charging modes are respectively denoted as “FDL x ” and “SDL x .” For instance, the FDL 1 represents the level 1 of battery degra-

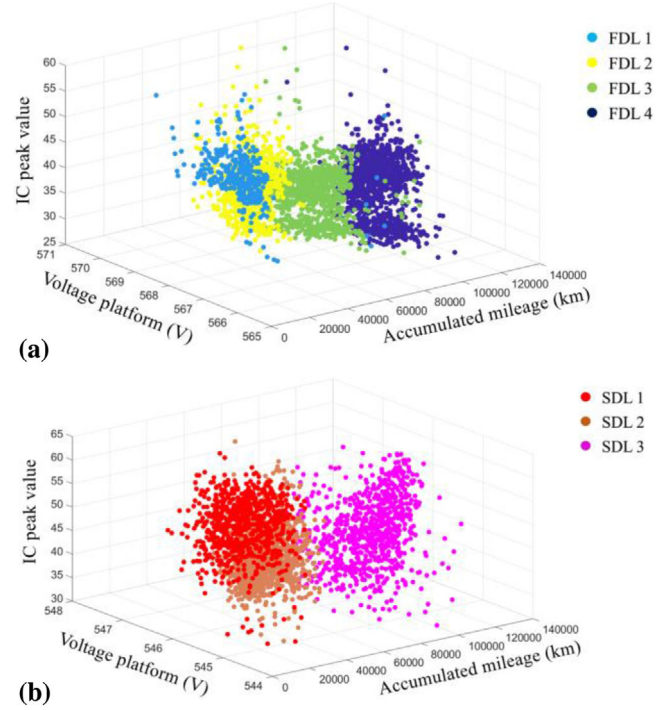


FIGURE 4 Battery degradation levels from recognition results under different charging modes. (a) Battery degradation levels under fast charging mode. (b) Battery degradation levels under slow charging mode

ation under fast charging mode, and the SDL 1 denotes the level 1 of battery degradation under slow charging mode. The higher level indicates the deeper degree of battery degradation for the corresponding charging mode. The recognition results aim to provide support to monitor the battery performance variations and improve the accuracy of driving range estimation. Note that, as the battery degradation levels are considered, the implied temporal evolution of the input-output variables is involved in the driving range estimation. This is because the battery capacity degrades over time, which affects the initial state of related input variables and further influences corresponding output variables.

5 | DRIVING RANGE ESTIMATION MODELS

5.1 | Methodology

The battery degradation levels, driving data, and weather data are combined to model the driving range estimation for BEBs. The experimental data has been processed to realise the research goal. Figure 5 illustrates the flowchart of processes for the data used in driving range estimation.

Firstly, the driving data points are aggregated according to the battery discharge periods. For a single discharge period, it starts as a BEB leaves the charging station and ends as the bus becomes recharged, which covers the itinerary between two charging events in the city transit network. In this way, the

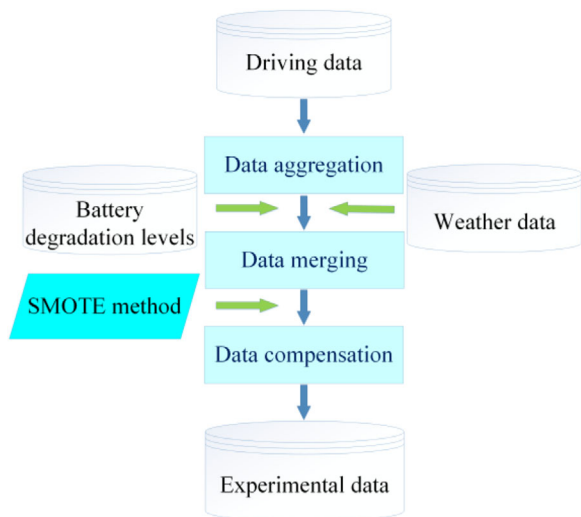


FIGURE 5 Flowchart of the processes for the data used in driving range estimation

incremental depth-of-discharge (Δ DOD) and driving range for each itinerary could be extracted based on the driving data. Note that, in general, the DOD is defined as the SOC discharged from 100%SOC and experiments in the literature measuring the DOD often start at 100%SOC [39]. However, in BEBs, the SOC is rarely 100% and a driving cycle may start at various SOCs, and thus it is difficult to obtain the DOD information from the BEB operating data. Unlike the DOD, the Δ DOD represents the difference in SOC from the initial and final SOCs before a BEB battery is recharged, which has better ability to adapt the battery working state of BEBs. Therefore, the Δ DOD will be used instead of DOD. Besides, the average speed, speed deviation, average voltage, average battery temperature and average current during each discharge period are further obtained. Furthermore, the time-of-day for each itinerary is processed into hours to indirectly reflect the impacts of time-varying traffic status, that is, peak and off-peak hours. In total, 52962 and 47346 itineraries of discharge period are obtained for the BEBs operating in the city areas under fast and slow charging modes, respectively. Secondly, the weather data is merged into the driving data according to the timestamp recorded in the datasets. Before data merging, the day of the week is added in the weather data, which could be used to investigate the week effects, that is, workday and weekend. Thirdly, the battery degradation levels are combined with the driving data based on the accumulated mileage recorded in the datasets. Finally, it is worth noting that the data imbalance regarding Δ DOD is existed in the driving data, and most of the itineraries have the relatively low Δ DOD. This problem would bring negative effects to the generalization ability of driving range estimation models. Concerning the imbalance of driving data, the synthetic minority oversampling technique is introduced to compensate the minority data points. The SOMTE is a k -nearest neighbours (KNN)-based oversampling algorithm, which randomly generates the complementary data points for any original minority instance in the directions of its KNN [40]. By this

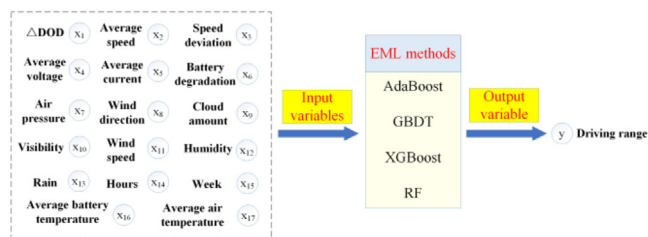


FIGURE 6 Model structure for driving range estimation

way, the experimental data is expanded to 109928 and 90608 itineraries under fast and slow charging modes, respectively.

After data processing, the EML methods are leveraged to establish the models for driving range estimation. The primary idea behind these methods is the combination of basic models to obtain a composite estimation model [41]. In this study, four types of EML methods, including adaptive boosting (AdaBoost), gradient boosting decision tree (GBDT), eXtreme gradient boosting (XGBoost) and random forest (RF), are introduced to model the driving range estimation. The reason for choosing such four EML methods lies in their relatively high computational efficiency and good adaptability for estimation operation with large-scale datasets, as compared to other EML methods, such as, Bayes-based approaches. The characteristics of the four methods are summarised as follows: AdaBoost tandemly trains the basic models and adjusting the weights of data points iteratively; GBDT trains the basic models in a tandem way through reducing the residuals obtained by the previous estimator; XGBoost is a scalable ensemble method based on GBDT, and trains the basic models by further introducing the regular terms and sampling strategy; RF trains multiple basic models in a parallel way by considering both the dimensions of data points and features, and alleviates the over-fitting problem through integrating the estimating results from basic models. During the last decades, the four types of EML methods has gained a significant attention due to their advantages, and more detailed principles and procedures of them are referenced in related studies [42–45]. In addition, the classification and regression tree is chosen as the basic model due to its high computational efficiency and good performance on estimation problems [46].

5.2 | Model training and verification

Based on the processed datasets, the AdaBoost, GBDT, XGBoost and RF are employed to model the driving range estimation with consideration of 17 impacting factors that include the Δ DOD, average speed, speed deviation, average current, average voltage, average battery temperature, battery degradation levels, hours, week, average air temperature, air pressure, wind direction, visibility, wind speed, cloud amount, humidity and rain, as show in Figure 6. The parameters for each method are determined based on the grid search approach with cross-validation test [47]. In this way, the optimal parameters which contribute to the best performance are obtained for each

TABLE 5 Key parameters for the EML models

Methods	Learning_rate	n_estimators	max_depth
AdaBoost	0.35	60	80
GBDT	0.05	300	5
XGBoost	0.05	1060	5
RF	0.35	60	80

TABLE 6 Metric values for the evaluation of the EML models

Charging modes	Methods	RMSE	MAE
Fast charging	AdaBoost	3.6636	2.4368
	GBDT	3.9194	2.8166
	XGBoost	4.0718	2.9319
	RF	3.5704	2.3631
Slow charging	AdaBoost	3.0518	2.0784
	GBDT	3.4742	2.5209
	XGBoost	3.1642	2.2960
	RF	2.9281	1.9662

EML method. The key parameters for the methods are listed in Table 5. In the table, the “learning_rate” represents the learning rate of the model iteration; “n_estimators” denotes the number of the basic models; “max_depth” is the maximum depth for each basic model.

Moreover, the data points are divided into two groups, namely training group and test group, and all the models would be established based on the same datasets. For each charging mode, 75% of the data points are randomly selected as the training group, while the remaining 25 % of them are used as test group to verify the performance of the trained models. Furthermore, we introduce the root mean square error (RMSE) and mean absolute error (MAE) as the metrics to evaluate the performance and accuracy of the models. RMSE is a commonly-used measurement that can exhibit the differences between estimating results and true values observed in actual data, as given in Equation (9).

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}. \quad (9)$$

Moreover, MAE is an arithmetic average of absolute errors, which has better capacity to reflect the true error of estimating results, as shown in Equation (10).

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |\hat{y}_i - y_i|. \quad (10)$$

Table 6 lists the metric values with respect to the RMSE and MAE for the EML models. It is observed that the values of RMSE and MAE exhibit the similar comparison among the four types of EML models, where RF obtains the lowest errors regarding both the metrics. Given the characteris-

TABLE 7 Metric values for the evaluation of the non-EML models and RF

Charging modes	Methods	RMSE	MAE
Fast charging	LASSO	8.3226	5.9898
	KNN	7.3218	5.4178
	SVR	6.0664	4.0072
	RF	3.5704	2.3631
Slow charging	LASSO	8.5139	6.3778
	KNN	6.7832	5.0373
	SVR	5.1717	3.5763
	RF	2.9281	1.9662

tics of the EML methods and datasets used for model training, the underlying reasons for such results mainly lie in the following two points. For one thing, the RF has better ability to deal with high-dimensional data, and thus the RF-based models have better generalization ability than other ones. For another, the RF can parallelly train the basic models, which improves the training efficiency for complex data. This unique advantage makes the RF more suitable to handle the experimental data, which is fused by driving data, weather data and battery degradation levels, than other EML methods. Note that, even though the RF model has the best accuracy as compared to other EML models, the gap of the metric values between RF and other ones is not distinguishable significantly. In a word, the results indicate that all the EML models have good performance on driving range estimation, while the RF has the best accuracy as compared to other ones. In view of this, the RF is selected as the representative of the EML models.

To further compare the performance of EML models with non-EML models, three types of non-EML models, including least absolute shrinkage and selection operator, KNN and support vector regression, are applied to estimate the driving range using the same datasets. These three methods have been widely used, and their principles and procedures are referenced in related literature [48–50]. The optimal parameters for the non-EML models are also determined based on the grid search approach with cross-validation test, which ensures the best performance for each non-EML model. The metric values regarding RMSE and MAE for the non-EML models are presented in Table 7. Moreover, for the sake of comparison, the metric values of RF are also added in the table.

As shown in Table 7, all the values of RMSE and MAE exhibit the significant gap between RF and non-EML models. Such a result indicates that the non-EML models would contribute to larger errors for driving range estimation as compared to the RF model, and thus the EML models has better performance than the non-EML models.

5.3 | Analysis and discussion of influencing factors

In the RF-based models, 17 impacting factors are considered to estimate the driving range of BEBs. Generally, the accuracy of

TABLE 8 Metric values for the evaluation of the RF- Δ DOD and RF-16 factors models

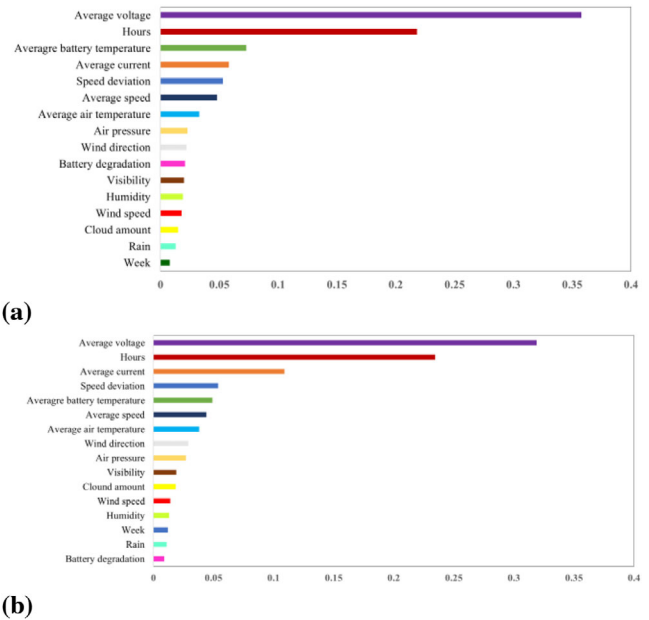
Model types	Charging modes	RMSE	MAE
RF-DOD models	Fast charging mode	9.7718	6.7948
	Slow charging mode	11.4667	7.8968
RF-16 factors models	Fast charging mode	20.1903	13.6469
	Slow charging mode	16.3300	10.8788

an estimation model is increased as the related impacting factors increases [51]. However, more factors considered in the model needs more data sources, and the costs for data collection would be gradually increased to use more data sensors. Some data sources may not even be available in real-world case. As a matter of fact, different influencing factors have different impact degrees to the driving range estimation. In the practical application, how to save the data collection costs or deal with limited data sources also attracts the attention of bus operators, besides the estimating accuracy of driving range. Therefore, the importance of influencing factors is analysed and discussed in this section.

Note that, the existing studies demonstrated that the Δ DOD has dominated impacts on the driving range [52]. Inspired by these research achievements, we first investigate the relationship between driving range and Δ DOD. Firstly, the data that only records the Δ DOD is applied to train and test the RF models under two different charging modes. Moreover, to further explore the impact degrees of other 16 influencing factors, we subsequently adopt the data without Δ DOD to train and test the RF model. The RMSE and MAE for the evaluation of the RF models with only consideration of Δ DOD (RF- Δ DOD) and with consideration of other 16 factors (RF-16 factors) are listed in Table 8. The output variable provided by the models of RF- Δ DOD and RF-16 factors is the driving range estimation.

In comparison to the original RF models as mentioned in Section 5.2, the RMSE and MAE deriving from the RF- Δ DOD models are increased as compared to that of the original RF models, which indicates that the other 16 influencing factors have positive effects to improve the estimating accuracy. Expectedly, the Δ DOD has a linear relationship with the driving range estimation as other inputs are equal, regardless of the initial SOC, because the energy output from the same Δ DOD under the identical battery degradation level is equal [12]. Moreover, by comparing the metric values of RF- Δ DOD models with that of RF-16 factors models, it is observed that, the RMSE and MAE values also indicate that the estimating errors increase as the Δ DOD is removed from impacting factors. These results demonstrate that the model with the other 16 factors can cover most of the information explained by RF- Δ DOD model. Therefore, RF-16 factors model would be able to provide reasonable estimated when the Δ DOD information is not available.

Furthermore, based on the trained RF-16 factor models, the relative importance of influencing factors is investigated to pro-

**FIGURE 7** Relative importance of the impacting factors for driving range estimation. (a) Fast charging mode. (b) Slow charging mode

vide insight into the impact degrees of them. Specifically, the Gini importance is adopted to determine the relative importance of each factor due to its high computational efficiency [53]. Furthermore, the computational results are normalized to obtain the score of each factor. Figure 7 illustrates the relative importance of the influencing factors, where more important factors are endowed with higher scores.

Note that, the 16 influencing factors can be broadly classified into four categories, which include battery status (average voltage, average current, average battery temperature, battery degradation), driving pattern (average speed, speed deviation), weather condition (average air temperature, air pressure, wind direction, visibility, wind speed, cloud amount, humidity and rain) and time factors (hours and week). It can be seen from Figure 7, most of the battery status factors have relatively high impact degrees on driving range, and the driving pattern also has a considerable influence which basically follows the battery status. Thereinto, the average voltage has the highest impact degree among all the factors, because it closely relates to the battery energy output. By comparison, the factors from weather condition have relatively less effective because these factors exert influence on the driving range through indirectly affecting the battery status or driving pattern. For example, the average air temperature has a certain degree of influence on the battery temperature and use of air conditioner; the visibility may make drivers raise concerns about safety and thus affects the driving speed. Note that, in contrast to the visibility, the wind speed, cloud amount, rain and humidity have the relatively limited impacts on the driving speed, and thus get the lower scores. In addition, for the time factors, the factor of hours has a relatively high impact degree. It is because the time-of-day has a close relation to the traffic condition and thus takes significant effects on both battery status and driving pattern. For example,

BEB drivers have to frequently adjust the vehicle speed during peak hours, whereas they can maintain a relatively constant vehicle speed during off-peak hours. By contrast, the factor of week has a relatively low impact degree, which implies that the difference of the passenger load between workday and weekend has limited effect on energy consumption. This is because the BEB often has a limited carriage capacity, while its curb weight is much larger than the containable passenger load. Based on the results, it is recommended that public transport operators should give priority to the battery status and driving pattern factors for driving range estimation. Besides, even though the weather condition factors have a relatively low impact, they also have ability to improve the estimation accuracy. If the higher accuracy is required, the data sources of weather condition should be acquired, which would increase the operating costs. Therefore, public transport operators need to make a trade-off between estimation accuracy and operating cost, and then determine suitable strategies by considering both accuracy requirements and cost budget.

Moreover, when it comes to the influence of battery degradation on driving range estimation, there exists a significant difference between fast and slow charging modes when comparing the case (a) and (b) from Figure 6. It is observed that the impact degree of battery degradation for fast charging mode is higher than that for slow charging mode. Such a comparative result indicates that the fast charging mode contributes to more effects on battery degradation than the slow charging mode. In other words, the battery performance degrades slower under the slow charging mode than that under the fast charging mode. Nevertheless, compared with the fast charging mode, the slow charging mode spends a longer time in charging events, which may affect the efficiency of transit operation. By integrating battery degradation with driving range estimation, accurate estimations of the BEB driving range can be provided considering different charging modes and available operation information. Consequently, public transport operators can develop proper scheduling and dispatching strategies.

6 | CONCLUSION

This study investigates the driving range estimation coupled with battery degradation recognition for electric city buses using real-world data. The ICA method is used to characterize the battery aging performance based on the charging data, and subsequently, the framework comprising the k -means clustering algorithm and gap statistic approach is employed to recognize battery degradation levels. Afterward, four types of EML methods, including AdaBoost, GBDT, XGBOOST and RF, are introduced to model the driving range estimation, in which 17 influencing factors together with two different charging modes are considered. The results show that the EML methods have a good performance on driving range estimation while RF has the highest accuracy among them, which also performs a much better estimation accuracy than that of the non-EML models. This work provides insights for driving range estimation with battery degradation recognition based on real-world data. Furthermore,

the importance of influencing factors is analysed and considered to develop different models. The results indicate that most battery status factors have relatively high impact degrees, followed by the driving pattern factors. The weather condition factors have relatively low impact degrees while also being able to improve estimation accuracy. In addition, the models considering limited available information are developed and discussed to indicate that relatively accurate estimation can be achieved when some important factor is not available. The results show that the battery degradation under different charging modes has different importance on driving range estimation. The developed models could be used by bus operators to accurately estimate driving range of BEBs using on-line operating data and accordingly scheme the vehicle routing. Besides, the models could also be used to determine data sources according to the specific demand of estimation accuracy and save the associate operational cost.

Notably, according to the analysis of influencing factors, we provide the recommendation that public transport operators should consider the accuracy requirements and cost budget to select suitable data sensors, as mentioned in Section 5.3. However, an in-depth investigation regarding the trade-off between estimation accuracy and operational costs is not carried out, because such an issue is outside the scope of this work. Built upon the proposed models, we will explore the correlations between estimation accuracy and operational costs in the future study, and then provide a comprehensive scheme for transit system operation. In addition, this study considers 17 impacting factors, some of which may have correlations. These correlations may have influence on driving range estimation of BEBs. Recently, several machine learning approaches with reliable interpretability have been proposed to quantify the feature importance and correlations in the fields of production, such as the lithium-ion battery manufacturing [54] and electrode production [55]. These advanced methods also have reference significance in driving range estimation and thus will be utilized to improve the estimation performance in our future work.

ACKNOWLEDGEMENTS

This research was supported by the JPI Urban Europe-NSFC project named SMUrTS. The project was funded by the [National Natural Science Foundation of China](#) (Grant 71961137008) and the [Research Council of Norway](#) (Grant 299078). Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.

ORCID

Chaoru Lu  <https://orcid.org/0000-0001-8418-7658>

REFERENCES

1. Bi, J., et al.: State-of-health estimation of lithium-ion battery packs in electric vehicles based on genetic resampling particle filter. *Appl. Energy* 182, 558–568 (2016)
2. Hannan, M., et al.: A review of lithium-ion battery state of charge estimation and management system in electric vehicle applications: Challenges

- and recommendations. *Renewable Sustainable Energy Rev.* 78, 834–854 (2017)
3. Çeven, S., Albayrak, A., Bayır, R.: Real-time range estimation in electric vehicles using fuzzy logic classifier. *Computers & Electrical Engineering* 83, 106577 (2020)
 4. Xu, L., Wang, J., Chen, Q.: Kalman filtering state of charge estimation for battery management system based on a stochastic fuzzy neural network battery model. *Energy Convers. Manage.* 53(1), 33–39 (2012)
 5. Shao, S., et al.: On-line estimation of state-of-charge of Li-ion batteries in electric vehicle using the resampling particle filter. *Transp. Res. Part D* 32, 207–217 (2014)
 6. Xia, B., et al.: State of charge estimation of lithium-ion batteries using optimized Levenberg-Marquardt wavelet neural network. *Energy* 153, 694–705 (2018)
 7. Zahid, T., et al.: State of charge estimation for electric vehicle power battery using advanced machine learning algorithm under diversified drive cycles. *Energy* 162, 871–882 (2018)
 8. Yang, F., et al.: State-of-charge estimation of lithium-ion batteries based on gated recurrent neural network. *Energy* 175, 66–75 (2019)
 9. Varga, B., Sagoian, A., Mariasiu, F.: Prediction of electric vehicle range: A comprehensive review of current issues and challenges. *Energies* 12, 946 (2019)
 10. Vaz, W., et al.: Electric vehicle range prediction for constant speed trip using multi-objective optimization. *J. Power Sources* 275, 435–446 (2015)
 11. Wager, G., Whale, J., Braunl, T.: Driving electric vehicles at highway speeds: The effect of higher driving speeds on energy consumption and driving range for electric vehicles in Australia. *Renewable Sustainable Energy Rev.* 63, 158–165 (2016)
 12. Bi, J., Wang, Y., Zhang, J.: A data-based model for driving distance estimation of battery electric logistics vehicles. *EURASIP Journal on Wireless Communications and Networking* 2018(2018), 251 (2018)
 13. Yuan, X., et al.: Energy and environmental impact of battery electric vehicle range in China. *Appl. Energy* 157, 75–84 (2015)
 14. Vatanparvar, K., et al.: Extended range electric vehicle with driving behavior estimation in energy management. *IEEE Trans. Smart Grid* 10(3), 2959–2968 (2019)
 15. Bi, J., et al.: Estimating remaining driving range of battery electric vehicles based on real-world data: A case study of Beijing, China. *Energy* 169, 833–843 (2019)
 16. Lee, C., Wu, C.: A novel big data modeling method for improving driving range estimation of EVs. *IEEE Access* 3, 1980–1993 (2015)
 17. Sun, S., et al.: A machine learning method for predicting driving range of battery electric vehicles. *J. Adv. Transp.* 2019, 4109148 (2019)
 18. Bi, J., et al.: Residual range estimation for battery electric vehicle based on radial basis function neural network. *Measurement* 128, 197–203 (2018)
 19. Amirkhani, A., Haghanifar, A., Mosavi, M.: Electric Vehicles Driving Range and Energy Consumption Investigation: A Comparative Study of Machine Learning Techniques. In: 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). Shahrood, Iran, December 2019, pp. 1–6 (2019)
 20. Vidal, C., et al.: Machine learning applied to electrified vehicle battery state of charge and state of health estimation: State-of-the-art. *IEEE Access* 8, 52796–52814 (2020)
 21. Shen, P., et al.: The co-estimation of state of charge, state of health, and state of function for lithium-ion batteries in electric vehicles. *IEEE Trans. Veh. Technol.* 67(1), 92–103 (2018)
 22. Chaoui, H., Ibe-Ekeocha, C.: State of charge and state of health estimation for lithium batteries using recurrent neural networks. *IEEE Trans. Veh. Technol.* 66(10), 8773–8783 (2017)
 23. Bonfitto, A.: A method for the combined estimation of battery state of charge and state of health based on artificial neural networks. *Energies* 13, 2548 (2020)
 24. Weng, C., et al.: On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression. *J. Power Sources* 235, 36–44 (2013)
 25. Bian, X., et al.: A novel model-based voltage construction method for robust state-of-health estimation of lithium-ion batteries. *IEEE Trans. Indust. Electron.* <https://doi.org/10.1109/TIE.2020.3044779>
 26. Bian, X., Liu, L., Yan, J.: A model for state-of-health estimation of lithium ion batteries based on charging profiles. *Energy* 177, 57–65 (2019)
 27. He, J., et al.: Comparative study of curve determination methods for incremental capacity analysis and state of health estimation of lithium-ion battery. *J. Energy Storage* 29, 101400 (2020)
 28. Liu, K., et al.: Charging pattern optimization for lithium-ion batteries with an electrothermal-aging model. *IEEE Trans. Ind. Inf.* 14(12), 5463–5474 (2018)
 29. Liu, K., et al.: Lithium-ion battery charging management considering economic costs of electrical energy loss and battery degradation. *Energy Convers. Manage.* 195, 167–179 (2019)
 30. Anseán, D., et al.: Lithium-ion battery degradation indicators via incremental capacity analysis. *IEEE Trans. Ind. Appl.* 55(3), 2992–3002 (2019)
 31. Qu, J., et al.: A neural-network-based method for RUL prediction and SOH monitoring of lithium-ion battery. *IEEE Access* 7, 87178–87191 (2019)
 32. She, C., et al.: Battery aging assessment for real-world electric buses based on incremental capacity analysis and radial basis function neural network. *IEEE Trans. Ind. Inf.* 16(5), 3345–3354 (2020)
 33. Dubarry, M., Liaw, B.Y.: Identify capacity fading mechanism in a commercial LiFePO₄ cell. *J. Power Sources* 194(1), 541–549 (2009)
 34. Sasane, S.: Monotone smoothing splines with bounds. *Acta Applicandae Mathematicae* 169, 613–627 (2020)
 35. Jiang, Y., et al.: Recognition of battery aging variations for LiFePO₄ batteries in 2nd use applications combining incremental capacity analysis and statistical approaches. *J. Power Sources* 360, 180–188 (2017)
 36. Zhu, J., et al.: Efficient registration of multi-view point sets by k-means clustering. *Information Sciences* 488, 205–218 (2019)
 37. Sinaga, K., Yang, M.: Unsupervised K-means clustering algorithm. *IEEE Access* 8, 80716–80727 (2020)
 38. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B* 63(2), 411–423 (2001)
 39. Bhadra, S., et al.: The relationship between coefficient of restitution and state of charge of zinc alkaline primary LR6 batteries. *J. Mater. Chem. A* 3(18), 9395–9400 (2015)
 40. Zhu, T., Lin, Y., Liu, Y.: Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognit.* 72, 327–340 (2017)
 41. Dong, X., et al.: A survey on ensemble learning. *Frontiers of Computer Science* 14, 241–258 (2020)
 42. Hastie, T., et al.: Multi-class adaboost. *Statistics and its Interface* 2(3), 349–360 (2009)
 43. Friedman, J.: Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4), 367–378 (2002)
 44. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, New York, USA, August 2016, pp. 785–794 (2016)
 45. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
 46. Prasad, A., Iverson, L., Liaw, A.: Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9(2), 181–199 (2006)
 47. Huang, Q., Mao, J., Liu, Y.: An improved grid search algorithm of SVR parameters optimization. In: 2012 IEEE 14th International Conference on Communication Technology, Chengdu, China, November 2012, pp. 1022–1026 (2012)
 48. Kukreja, S., Löfberg, J., Brenner, M.: A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes* 39(1), 814–819 (2006)
 49. Altman, N.: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46(3), 175–185 (1992)
 50. Smola, A., Schölkopf, B.: A tutorial on support vector regression. *Statistics and Computing* 14(3), 199–222 (2004)
 51. Smuts, M., Scholtz, B., Wesson, J.: A critical review of factors influencing the remaining driving range of electric vehicles. In: 2017 1st International Conference on Next Generation Computing Applications (NextComp), Mauritius, July 2017, pp. 196–201 (2017)

52. Wang, J., Liu, K., Yamamoto, T.: Improving electricity consumption estimation for electric vehicles based on sparse GPS observations. *Energies* 10(129), 19–22 (2017)
53. Menze, B., et al.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinf.* 10(1), 1–16 (2009)
54. Liu, K., et al.: Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification. *IEEE/ASME Trans. Mechatron.* (2021). <https://doi.org/10.1109/TMECH.2020.3049046>
55. Liu, K., et al.: Mass load prediction for lithium-ion battery electrode clean production: A machine learning approach. *J. Cleaner Prod.* 289, 125159 (2021)

How to cite this article: Wang Y, Lu C, Bi J, Sai Q, Zhang Y. Ensemble machine learning based driving range estimation for real-world electric city buses by considering battery degradation levels. *IET Intell Transp Syst.* 2021;15:824–836.
<https://doi.org/10.1049/itr2.12064>.