Investigating the impact of missing data imputation techniques on battery energy management system

Mehdi Pazhoohesh1 | Adib Allahham2 | Ronnie Das3 | Sara Walker4

1Faculty of Computing, Engineering and Media, School of Engineering and Sustainable Development, De Montfort University, Leicester, UK
2Faculty of Engineering, Middle East University (MEU), Amman, Jordan
3Newcastle University Business School, Newcastle University, Newcastle upon Tyne, UK
4Faculty of Science, Agriculture and Engineering, School of Engineering, Newcastle University, Newcastle upon Tyne, UK

Abstract

Effective control of energy storage system (ESS), supplying an ancillary service to a grid, requires effective and critical calculation of state-of-charge (SoC). Charging and discharging values from battery operations are essential in calculating the efficiency and performance of a storage system. This information can also be a key to understand and forecast peak demand performance. Missing data is a real problem in any operations system, and it appears to be more common within powers systems due to sensor and/or network malfunctioning problems. Missing data imputation techniques have evolved in power systems research using smart meter data, but little research has gone into understanding how missing data can be best handled within storage management systems. This paper builds on a year’s worth of charging and discharging data collected from a real 6MW/10MWh lithium-ion storage battery deployed on the distribution network at Leighton Buzzard, UK. Using R Studio version (1.3.959-1) open-source software, eight selected imputation techniques were applied in identifying the best suited technique in replacing various missing data amounts and patterns. Findings from the study open up avenues for discussion and debate in identifying an appropriate imputation technique within the storage management context. The study also provides a pioneering lead in understanding the importance of decomposition in evaluating the right imputation technique.

1 | INTRODUCTION

In recent years, energy storage systems (ESS) have become an important pillar of future smart grids, thanks to their ability to carry out energy arbitrage while providing a range of ancillary services to the network [1–4]. Battery energy storage systems (BESS) can act as a controllable source, and sink of both real and reactive power, making them an important integral part of smart grids.

The energy management system (EMS) of the BESS plays a key role in enabling the BESS to provide designated services or functions required. A typical energy management system also requires supportive data acquisition systems in enabling different actors working to update the control unit about their current operation conditions. In addition, data acquisition systems are also required to control units sending commands to storage units dictating the next operation period according to network requirements [5,6]. Data driven knowledge acquired from an electricity network, including BESS data, can also be used for monitoring and forecasting electricity usage [7–9]. These data are valuable in identifying electricity consumption patterns and role of BESS as a peak demand support mechanism, especially when developing smart city infrastructure [10,11]. Despite the attractive prospects of data driven EMS management, it is important to note that different data-based applications rely heavily on the quantity and quality of data captured. Missing data values resulting from uncontrollable factors, such as, malfunctioning of data acquisition, communication or storage devices can lead to misleading results with greater consequences. Ignoring missing data could lead to critical problem development in controlling an asset providing a service within a vital electricity system. For example, a BESS used for data driven peak reduction within an electrical network can suffer from severe performance loss due to unaccounted missing data. Within peak reduction applications, failure in accurate network demand data capturing framework...
will lead to a failure in required power supply from an allied BESS supporting to reduce power consumption from upstream networks. Furthermore, failure of an EMS due to lack of accurate operational data during a specific operational period can also lead to miscalculated state-of-charge (SoC) instructions being conveyed to associated BESS, resulting in inaccurate scheduling of operational intervals.

The authors aim to address the dangers and consequences of missing data within BESS management systems by investigating and recommending the most appropriate data imputation techniques for handling missing data from a BESS contracted to provide peak reduction service to a local electricity network. Identifying the right data imputation is essential in enabling a BESS to provide its contracted service and allow the EMS to issue correct control actions based on forecasted peak demand management.

1.1 Literature review

A common practice in missing data management, within many applications, dictates complete ignorance of missing entries to reduce computational overload. Application of data deletion methods, in list-wise or pairwise format, reduces the size of usable data leading to biased estimation [12,13]. The methods that were found to be popular in dealing with missing data within electricity systems are interpolation, nearest-neighbour (NN) and non-parametric regression (NPR). In interpolation method, missing data frames are estimated by interpolating adjacent data available. Accuracy of this method is affected by the extent of missing data and consistency between missing data and the adjacent data available. The interpolation method is practical and effective when the period and extent of the missing data is short and the adjacent data have a consistent pattern. Another popular imputation method is the NN that uses available data within a fixed range around missing values. This method is appropriate when electricity system data have a distinct periodic pattern. Unfortunately, this is rarely the case within an electricity system. NPR is logically similar to NN applied when data contains specific patterns. In this method, missing data is estimated using similar patterns.

Authors in [14] presented a correlation based imputation method that supplements missing data using the correlation between demand data from different homes within a microgrid along with proxy parameters (voltage and frequency data of each appliance in these homes). This method has its own limitation, especially within the energy storage context, especially, if proxy parameters are not collected. One study advocated linear interpolation (LI), NN and NPR as best practical methods for data imputation [14]. In their approach, LI method was used when periods of missing data were shorter that one or two hours [14]. If this was not the case, NN or NPR were applied where load profiles were built based on historical data mimicking 'like days' and 'like weekdays' [12,15,16]. Their selection of method appropriateness was essentially based on a trade-off between the calculation complexity and imputation accuracy. In [12], missing data were replaced by data from nearest equivalent day when the period of missing data was registered for more than two hours. For example, if load profile of a Wednesday was missing, then it was replaced by data profile taken from previous Wednesday. If the data from last Wednesday were not also available, then data from last Tuesday or Thursday were used. In special days such as public holidays, data from the nearest Sunday was captured. Missing data imputation in [13] used an equivalent day approach similar to [5]. However, in [13], average of three equivalent days were used to impute the missing data. The work presented in [17] illustrated LI method is only effective if data variation within an interval of one or two hours is small; this performance becomes unsatisfactory if the variation in load pattern becomes unpredictable even for a short period. Imputation method developed in [17] is based on the weighted sum of LI and its own historical average data in a fixed time range. However, this method assumes that obtaining a valid estimation of missing data requires non-missing data within a fixed range. The authors proposed a learning-based adaptive imputation method in [18] that estimates missing data by using feature vectors. These vectors represent past situations and contain load patterns included historical data and their variations. Missing data were then imputed by selecting the \( k \) most similar past situations. To evaluate the effectiveness of their imputation techniques, missing data frames were generated intentionally with different lengths and ratios [18]. Performance of individual technique was evaluated using well known metrics, i.e., mean absolute percent error (MAPE) and the root mean square error (RMSE). In addition to this stream of work on missing data imputation in the field of electricity systems, other notable research literature such as [19,20] have also shown the importance of research progress in the field of data imputation. Performance of six different imputation methods, KNN, fuzzy k-means, singular value decomposition, Bayesian principal component analysis and multiple imputation by chained equations, were compared against different datasets [19]. Evaluation metrics used in this study was RMSE and classification errors. Results from the study shows that Bayesian principal component analysis performed better than other approach. In [20], performance of different imputation methods – random imputation, multiple imputation using expectation-maximization, KNN, and random forests, using error metrics and size effect measures – were compared on a photovoltaic (PV) system dataset, including solar irradiance and temperature. Results showed that KNN and random forests produced the best output.

Our systematic review of literatures in this area demonstrates that little effort has gone into understanding the best fit imputation methods of missing data of a BESS providing a peak reduction service to a local distribution network. This paper evaluates the effectiveness of selected imputation techniques in replacing value and charge characteristics of real life BESS dataset. In addition, our study also focuses on the impact of two different types of data missing patterns that are often side lined in other studies for a standard pattern, in order to reduce imputation complexities.
1.2 | Objectives and contributions

As seen in the previous section, the imputation methods suitable to impute values for a particular type of data and application might not be applicable to another dataset due to different characteristics of various applications. This is demonstrated by the fact that the best imputation methods for replacing missing data for energy demand forecast for an energy consumer are not the same as forecasting the energy generated from a photovoltaic system [17–20].

One of the main applications of BESS in the electrical network is peak demand management where the BESS charges during the low demand period and discharges during the peak period should ensure that the network constraints are within the acceptable limits. Our literature review has flagged up the limited amount of effort that has gone into understanding the best fit imputation methods for an ESS providing a peak reduction service to a local distribution network. Here, the authors focus on handling missing data collected from a real-life battery energy storage system (BESS). Poor treatment of missing BESS data can lead to reduced accuracy and compromised integrity of a storage energy management system [14,15]. In addition, the authors conduct a systematic investigation into a range of imputation methods to identify the best studied technique to impute the missing data of BESS providing peak reduction service to an electrical network. On this note it is important to highlight that our approach considers amount and patterns of missing data as decisive factors to select the best fit imputation method that is paramount for any application/asset. In our endeavour, two different types are introduced of missing data patterns within the standard data frame, continuously missing pattern (CP) and general missing pattern (GP), in order to measure their impact of selection of appropriate imputation method. Eight popular imputation techniques (non-parametric regression (NPR), Markov Chain Monte Carlo (MCMC), k-nearest-neighbours (KNN), regression, mean, zero indicator, and forward filling) were applied to determine how accurately missing data values were predicted within test frameworks of each method, against controlled data frames, using a three stage controlled missing data volume (10%, 20%, 30%). A modified sequential nearest neighbour technique, labelled as KNN in row, was also applied to check its impact on imputation accuracy. Findings from the study not only help to identify the best imputation technique, but the results are also useful in challenging the one size fit model approach to data imputation.

The key objectives of our study can be summarised as:

- Identify the best imputation technique in the energy storage management context that can guide energy storage operators and data analysts to reliably impute missing data obtained from the charging and discharging state (and the value) of an ESS.
- Statistically verify the quality of replaced data, collected from the application domain of ESS providing a peak shaving service to an electrical network, predicted by individual imputation techniques within different types of missing data frames.
- Apply and test the effectiveness of a modified version of sequential near neighbour (KNN) algorithm logic considering the sequential characteristic of an ESS dataset.
- Investigate the importance of decomposition technique in further improving the accuracy of practiced values derived from individual imputation techniques. This is important in the context of ESS as predicting the right SoC is equally paramount as the value of a missing data in order to avoid issuing wrong commands.

1.3 | Organisation of the paper

This paper is organised as follows: The method to select the best performance imputation technique and the evaluation metrics are explained in Section 2. Results are presented and discussed in Section 3. The main findings are summarised in Section 4.

2 | METHOD

2.1 | The electricity system configuration

Figure 1 shows the system of study considered by the authors. The ESS is used to offset the peak demand of a set of loads of a distribution network. The sum of loads' demand is indicated by \( P_D \). The power supplied from the upstream network is indicated by \( P_N \) and it must be within a limit defined by the network operator \( (P_N^{max}) \). In Equations (1) and (2), \( \eta_{ch} \) and \( \eta_{dis} \) represent the efficiency of BESS during the charging and discharging, respectively. \( E_{nom} \) represents the nominal capacity of the BESS, measured in MWh. \( P_{ESS}(t) \) represents the power of BESS exchanged with the network at the current operational period. The positive
value of $P_{\text{ESS}}(t)$ corresponds to the charging state of the ESS however the negative value corresponds to the discharging state. The value of $\text{SoC}(t)$ at the current operational period depends on its value in the previous operational period, i.e. $\text{SoC}(t-1)$, and the energy exchanged through the BESS. The BESS is controlled by a rule-based energy management system, which receives the measured $P_{\text{BESS}}$ and $P_D$ to decide an appropriate set-point of the power to be (dis)charged. Figure 2 shows the flow chart of this energy management system. It shows that one of the most important parameters to consider for issuing an appropriate control command is the SoC. The SoC of the BESS is calculated using the following relationships:

$$
\text{SoC}(t) = \text{SoC}(t-1) + \left( \left( P_{\text{ESS}}(t) \cdot \eta_{ch} \cdot t \right) / E_{\text{nom}} \right) \quad \text{If } P_{\text{ESS}}(t) > 0
$$

$$(1)$$

$$
\text{SoC}(t) = \text{SoC}(t-1) + \left( \left( P_{\text{ESS}}(t) \cdot t \right) / \eta_{dis} \cdot E_{\text{nom}} \right) \quad \text{If } P_{\text{ESS}}(t) < 0
$$

$$(2)$$

In the simple EMS shown in Figure 2, the control horizon is a day (24 h) and the operational period (t), in which the EMS needs to issue an appropriate control action, is 30 min. The EMS must assure a 50% SoC at the end of the control horizon to be sure that the BESS has enough energy for peak reduction in the next day. The EMS assumes that the BESS is optimally sized to provide the peak reduction service in the selected location of the distribution network such that the SoC varies between 20% and 100%. The lower boundaries of SoC is selected to guarantee the smooth operation of the BESS.

It is assumed that there is no missing value in $P_D$ because the main aim is to focus on finding the best imputation method for the missing data of a BESS.

### 2.2 Data framing

The choice of missing data imputation technique and algorithm selection widely varies depending on the characteristics of a dataset [21,22]. The fine balance between deterministic and probabilistic characters of a dataset can often dictate the choice of best imputation method to be applied. In order to instigate our investigation into missing data, it was important to conceptualise all the variables within our raw battery dataset as time dependent with recorded longitudinal information. Variables were sorted individually considering time as the most probable determinant of battery charge and discharge values ($P_{\text{ESS}}$).

The original dataset used in the study was collected from a 6MW/10MWh lithium-ion storage battery deployed on the electricity distribution network at Leighton Buzzard in Bedfordshire, UK. The battery supplied a secure source of energy during peak demands. The raw dataset obtained provided a time series of $P_{\text{ESS}}$ charge and discharge values over a 60 week period. The identified time interval within the raw dataset was 30 min, providing 48 datapoints every 24 h. The positive/negative values within the raw dataset referred to the charging/discharging status of the battery. Further description of the ESS and its characteristics can be found in [23].

The initial raw data was thoroughly inspected for anomalies and cleaned to finalise with no abnormalities or missing value. In total, the cleaned dataset provided 17,883 charge and discharge values over a selected 60 week period and was labelled as the ‘original dataset’ for the experiment. During the next stage, the focus is on developing six test datasets closely resembling real life missing data scenarios. Two important factors were considered in preparing test datasets: (i) amount of missing data and (ii) patterns of missing data.

During the first stage, three data reduced versions of the original dataset were prepared by methodically removing 10%,
20% and 30% of the data following the missing completely at random (MCAR) mechanism [22]. The logic behind the data removal process was inspired by studies [22,24] that classed 10% reduction as small, 20% as medium and 30% reduction as large amounts of missing data reflecting real life scenarios.

Beside data reduction, patterns of missing data are also an important factor that has received little attention in existing research [25]. Therefore, during the second stage, the authors focused on recreating missing patterns within three missing value test datasets prepared earlier. Time data is used as an auxiliary information and introduced the two most commonly observed missing patterns, namely, ‘Continuous missing Pattern’ (CP) and ‘General missing Pattern’ (GP). The rationale behind our data sorting and missing data frame preparation followed two practical scenarios. First, a practical fault situation such as sensor or network failure lasting from several days to weeks [26] leading to a data frame that suffers from a large portion of missing data resembling CP. In a different scenario data can go missing spontaneously following random intervals. In this case the data frame can have scattered random data points resembling a GP. A visual representation of these missing patterns is highlighted in Figure 3. While Figure 3a presents an illustration of GP, Figure 3b presents a typical CP scenario with missing data from days 2, 3, 7 and 8.

Following the test datasets preparation, eight suitable statistical imputation techniques were applied to replace missing data.

2.3 | Imputation techniques

Datasets prepared on battery charge and discharge data were subjected to eight selected sets of statistical imputation techniques: NPR, MCMC, KNN, regression, mean, forward filling, zero Indicator and KNN in row. KNN in row, stemming from sequential-KNN logic, was applied to the test datasets to measure whether sequential imputation can produce better results given the nature of the dataset [27]. Results were compared in a tabular format using statistical validity tests.

All the algorithmic imputation techniques were performed by using R Studio version (1.3.959-1) opensource software. Additional R package libraries such as ‘ggplot2’ [28] was used for the visualisation of data. During the evaluation phase, the imputed test dataset results are compared against the original dataset and applied statistical measures such as standard deviation (SD) and root mean square error (RMSE) to determine the level of accuracy achieved by individual imputation method. In total, 2 validity tests (SD and RMSE) were applied to 6 imputed datasets (3 value reduction, 2 missing patterns), following 8 different imputation methods, giving a total of 96 trial results. In addition, 45 decomposition results were also gathered from the test and the original dataset to verify the characteristics of imputed datasets against the original. Selected imputation techniques are discussed here.

2.3.1 | Indicator adjustment (zero)

This method replaces missing data values with a nominal constant, that is, zero. This method is useful in scenarios where analytical integrity could be compromised with gaps in a series of data [29]. Application of this method can lead to a serious compromise of dataset integrity along with wider statistical insignificance.

2.3.2 | Mean/Mode substitution

Mean or mode-based approach to replace missing data is the simplest and widely used method in generic imputation practice. As part of this technique missing data is generally replaced with a mean or mode value of entire data set within a specific range [30]. Although the method is simple and easily applicable, one of the biggest disadvantages of this technique is that it distorts standard deviation of data distribution, leading to exaggerated number of mean values within the dataset. Cumulative distribution function and variable relationship within the auxiliary datasets are potentially compromised with large error margins between predicted and actual values.

2.3.3 | Near-neighbour imputation (KNN)

The K nearest neighbour or KNN algorithm [31] is a nonparametric method that is widely applied to numeric and
categorical datasets. In this technique, missing data is replaced by calculating average or weighted mean of neighbouring values. Amongst many other possible approaches, the most common practice is to define the position of \( K \)-feature vector within a dataset that accounts for the surrounding number of neighbouring values. Many neighbours assigned to the \( K \) value could have significant influence on the final prediction capability of the algorithm. This method is well known for its relative accuracy in dictating randomly missing data within a dataset, although poor results have been reported when trying to apply this method to replace continuous missing values within a prolong period of time, that is, CP.

### 2.3.4 Forward filling imputation

This method replaces missing data values by propagating the last observed value forward. This method is particularly useful in scenarios where the data has a strong correlation with previous value \([12]\).

### 2.3.5 Sequential near-neighbour imputation (KNN in row)

Sequential KNN resembles random forest algorithm principles as each base classifier is modelled on a forward sequential strategy using majority selection logic \([32]\). However, the technique differs from random forest in its use of nearest neighbour base classifiers instead of a decision tree. The KNN in row is a modified version of sequential near-neighbour principles, in which forward propagation is based on selected sequential values within datasets. \( \hat{P}_{ES} \) values within a specific sequence of time and days in this instance.

For example, if a data frame collects a data value every 30 min, a data point missing on a Monday 11:00 PM will get imputed based on data collected on the previous Monday, as opposed to nearer values on the same day or values from another day (Figure 4). Theoretically, this approach appears to be more appropriate as battery storage and operations data can be largely day and time dependent.

### 2.3.6 Regression method

Regression method is widely used in forecasting practices, but the same principles are applied to impute missing values using one or more auxiliary variables. A linear regression method is more suited to datasets containing non-binary numerical variables, whereas logistic regression is more suited to datasets with binary variables. Despite the simplicity and advantages over more complex imputation techniques, the application of linear regression may lead to distorted distribution of data with poor correlation. Adding a noise term to linear regression often circumvents the problem and reduces the bias while augmenting each predicted score with a residual term. Advantages of such stochastic regression is that each missing value is replaced with a new imputed value as opposed to a repeated one \([33]\).

### 2.3.7 Non-parametric regression (NPR)

Non-parametric regression technique aims to identify associations between the predictor and the response without considering the form of estimated regression function. One popular non-parametric imputation method has been developed using random forest (RF) technique \([34,35]\). In this approach, the missing values are directly predicted from the trained subset of the observed data of RF.

### 2.3.8 Markov Chain Monte Carlo (MCMC)

The Markov Chain Monte Carlo (MCMC) method uses a stochastic model in predicting missing values within a data frame. The concept was coined by mathemetician Andrey Markov to analyse equilibrium distributions of interacting molecules \([36]\). The MCMC technique helps to characterise distribution of value where mathematical properties are unknown; this is particularly useful within a high dimensional and complex pattern of probabilistic distribution. This method had previously been applied to estimate arbitrary missing data within a dataset with promising outputs \([36,37]\).

### 2.4 Performance evaluation

The accuracy of imputed values within test datasets were evaluated using SD, RMSE and data decomposition tests. SD of \( \hat{P}_{ES} \) values within individual imputed datasets was calculated using the square root of variance resembling expression 3. Findings from individual test datasets were compared against the standard deviation of the original dataset. The resulting difference in SD value between individual test datasets and the original dataset was expressed in percentage value (%), labelled as ‘distance’.
The application of Root Mean Square Error (RMSE) offered further insight into the accuracy of imputed data by calculating the difference in standard deviation between the actual and predicted value, as outlined in expression 4. RMSE was applied to each imputed dataset to calculate square root of average square errors in identify error margins.

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_{imputed}^i - X_{obs}^i)^2}{n}} \quad (4)
\]

In both equations, \( n \) denotes the number of \( P_{ESS} \) data points in a dataset. In expression 4, \( X_{obs}^i \) denotes the target of a data point or the original value, while \( X_{imputed}^i \) denotes predicted value of an imputed data point within a test sample.

In addition to these statistical validations, data decomposition technique was applied to investigate the characteristics of selected imputed data. Decomposition method helps to extract seasonal patterns from a time-series dataset and remove excessive noise cycles.

Despite offering an extra means of testing accuracy and validity, the application of decomposition method in data imputation is limited. We used the forecast package within R in applying multiple STL decomposition (MSTL) technique to implement noise smoothing within the decomposed series of original and imputed data. The MSTL logic builds on seasonal trend decomposition (STL) as expressed in Equation (5) [38].

\[
Y_t = T_t + S_t + e_t \quad (5)
\]

\( Y_t \) within the equation represents the time series at time \( t \). \( T_t \) represents the trend-cycle component at time \( t \). \( S_t \) represents the seasonal component at time \( t \). \( e_t \) represents error at time \( t \).

MSTL is a reiteration of STL process that addresses multiple seasonal components within a time-series. The decomposition equation applied in our analysis are as follows.

\[
Y_t = T_t^1 + T_t^2 + \ldots + T_t^n + S_t + e_t \quad (6)
\]

Findings from the decomposition analysis allowed us to compare time-series graph trends against the original dataset in further evaluating the accuracy of individual imputation techniques. This was particularly useful in evaluating +/- replacement accuracy, in addition to value replacement, while trying to understand efficacy of individual techniques.

3 | ANALYSIS AND DISCUSSION

The analysis for the study was carried out in two different stages. During the first stage, a combination of 8 stochastic and deterministic imputation techniques was applied to all test datasets in replacing missing values. Quality and accuracy of imputed datasets were verified using SD and RMSE. During the second stage of analysis, a decomposition method was applied to compare the characteristics of the original and selected test datasets, while a systematic observational pattern recognition process was deployed to determine the accuracy of SoC and value replacement. To reduce subjectivity in our analysis, the time-series graphs were individually inspected by the authors before reaching mutual agreement on findings. To ensure further accuracy, some of the decomposed time-series graphs were standardised using a ratio scale. In total, 141 tests (96 statistical + 45 decomposition) were carried out to ensure the robustness of the process.

3.1 | Results overview

The RMSE (%) and SD (%) results, referred to as the ‘distance’, are presented in Table 1, comparing the original and imputed datasets following GP and CP patterns.

|                | GP       |         |         |         | GP       |         |         |         |
|----------------|----------|---------|---------|---------|----------|---------|---------|---------|
| Methods        | 10%      | 20%     | 30%     |         | 10%      | 20%     | 30%     |         |
| Regression     | 80.99    | 0.13    | 118.56  | 0.92    | 144.91   | 1.21    | 170.10  | 77.67   |
| Mean           | 58.69    | 9.33    | 85.44   | 20.39   | 101.09   | 43.28   | 164.17  | 90.46   |
| Zero indicator | 60.32    | 8.89    | 86.89   | 19.47   | 103.65   | 21.66   | 163.66  | 87.90   |
| MCMC           | 63.49    | 7.23    | 93.90   | 15.92   | 111.51   | 41.63   | 163.66  | 87.90   |
| NPR            | 59.11    | 8.63    | 88.41   | 16.34   | 103.00   | 21.66   | 127.40  | 87.90   |
| Forward        | 30.58    | 0.35    | 50.62   | 0.30    | 58.71    | 0.76    | 304.44  | 113.71  |
| KNN in row     | 33.08    | 0.40    | 47.92   | 3.28    | 59.30    | 1.37    | 297.52  | 113.71  |
| KNN            | 57.77    | 10.04   | 86.77   | 16.37   | 96.54    | 20.19   | 158.87  | 87.76   |

**TABLE 1** RMSE and standard deviation results
The SD distance results observed within the GP indicates that regression and forward filling techniques produced the smallest distance compared to other techniques within this specific missing pattern group. However, it is worthwhile noting that the distance patterns widely vary when the amount of missing values are taken into consideration. For 10% missing dataset, KNN in row and for the 20% and 30% missing dataset, mean and zero indicator achieved the highest distance. This means that these techniques are least reliable when considering small to large missing data with general missing pattern. On the contrary, SD results from the CP group highlight that KNN in row might be the best technique to replace CPs as it produced the closest distance to the original dataset. Here, it is also important to note that KNN applied to GP achieved the third lowest distance position for both 10% and 20% missing data, making it a potential candidate for further experiments and observations.

In addition to Table 1, Figure 5 illustrates RMSE results for general pattern (GP). A notable reduction (i.e., improved results) in error is achieved through the forward filling and KNN methods, while the regression technique achieved the worst result (highest error) in replacing missing values within test datasets. Mean, zero indicator, MCMC, NPR and KNN in row produced comparatively similar performances.

Figure 6 presents RMSE results for CP. Amongst all other imputation techniques, the most favourable RMSE results were achieved by KNN in row. Combining SD and RMSE results it is safe to say that KNN in row stands out as the best imputation technique within CP missing datasets, but the algorithm is not efficient in replacing GP. Here it is also important to note that two unusually high distances produced by forward filling and KNN within the 20% reduced dataset means these two techniques may not be best suited to replace CP missing patterns with moderate amount of data missing from the data frame. Such anomalies can be explained by the logical similarities of these two techniques (i.e., forward filling uses previous value compared to KNN using average of previous neighbours). Existing previous values within the CP missing data frame can be very different depending on the time and day of battery operation, leading to the estimation of misleading values. Such behaviour emphasises the importance of applying appropriate technique to the appropriate context, especially in real life, as over or under estimation could cause serious irreversible damage to a BESS.
3.2 Impact of imputation methods on pattern and characteristics of data

To investigate the impact of various imputation methods on the characteristics of different datasets, a decomposition technique was applied following [38], to both, the original and the CP, 30% imputed datasets containing continuous missing patterns. CP missing pattern with 30% loss was specifically selected for this study as increasing the amount of data loss has shown to pose more challenges to data imputation [22]. Figures 7–11 illustrate the outputs from decomposition analysis addressing total data pattern, trend ratio and daily, weekly and yearly trend analysis.

Figure 7 refers to the graphical time-series patterns arising from the original and imputed datasets along with the battery’s charging and discharging patterns. Findings here demonstrate that only three methods (i.e. regression, MCMC and KNN in row) could considerably match the pattern arising from the original dataset, although regression technique significantly increased the value and frequency of charge and discharge cycle making it less reliable.

Figure 8 illustrates the overall trend (after peak smoothing) arising from each technique. Overall, the original dataset shows an upward curve, while in comparison, relative trends noted within regression, MCMC and KNN in row also show upward curves. In contrast, mean, KNN, zero indicator, NPR and forward filling has presented downward curves, pointing to the wrong estimation of the battery charge state, that is, the charging state falsely represented as discharging state.

Beside the total data patterns and trend analysis, seasonal patterns were captured through more minute details including daily, weekly, and yearly patterns. Figures 9, 10 and 11 present these findings, respectively.

Figure 9 confirms the four techniques including mean, KNN, zero indicator and forward filling have removed the charging and discharging states of the battery by replacing them with neutral states (zero value). Therefore, application of these techniques in real life can lead to serious imputation flaws being posed into battery storage and operations data. In real life, underestimating the charging and discharging states can lead to SoC estimation error issuing inaccurate control commands. In contrast, daily patterns identified through MCMC...
and KNN in row produced peaks resembling the shape of the original pattern.

Figure 10 outlines the weekly power ratio observed as a result of decomposition. Findings here shows additional peaks computed by regression technique, while the range of weekly power ratio had been altered (underestimated) by mean, KNN, zero indicator, NPR and forward filling methods. These findings are significant as real-life application of inaccurate imputation can lead to significant errors in SoC forecast and estimation.

Figure 11 presents important observations regarding yearly patterns. Regression, MCMC and KNN in row presented fairly similar patterns compared to the original data. On the other hand, inner trends observed through mean, zero indicator and NPR were rather non-characteristic in nature. In a real-life scenario this could have a serious implication on analytical capability, SoC forecast and estimation. Although KNN and forward filling were able to identify some trends, their patterns rather appear to be distorted. Despite good results, the range of yearly power ratio was understated by regression in identifying charging and discharging states, while forward filling method strongly understated the ratio for the discharging state. In contrast, mean, KNN, zero indicator and NPR techniques slightly overestimated charging state; while MCMC and KNN in row, once again, produced similar yearly patterns as the original data making them the most reliable.

Table 2 summarises our combined findings from statistical and decomposition analysis. Within CP missing patterns, KNN in row appeared to be the most successful technique in replacing missing values across all missing value ranges. Regression technique came as a close second within this category with few exceptions. GP missing pattern results appeared to be a more mixed bag. Based on the statistical evaluation, KNN and forward filling jointly appeared to be best placed with little margin of error. Regression technique appeared to be second best, particularly when a dataset suffers from a loss of more than 10%.

The following points summarise the important contribution of the study.

1. Importance of patterns of missing data: Within a battery operations data context, imputation techniques in general were found to be more successful in treating general pattern
data loss when compared to continuous missing patterns. For GP, overall 75.94% RMSE average and 15.63% SD average was observed compared to higher overall 158.58% RMSE average and 59.92% SD average for CP. Although RMSE paints a better picture of error margin compared to SD, both findings confirm the characteristics of missing patterns is essential to selecting and recording successful results using the most appropriate imputation technique, that is, KNN in row for CP and KNN/Forward filling for GP.

2. Importance of the amount of missing data: The amount of missing data has often been ignore in imputation research with a one size fits all model predominately being applied, underestimating its impact on the final outcome. This study demonstrates the importance of the amount of missing data within a dataframe and its impact on the accuracy of final outcome. It is observed from the analysis that increasing the proportion of the missing data increased the RMSE and SD value, with few exceptions. Overall 55.5% RMSE average and 5.62% SD average was recorded when imputing 10% missing values, 82.31% (RMSE) and 11.23% (SD) average was recorded with 20% missing data, while 90% (RMSE) and 29.65% (SD) average was recorded with 30% missing values. Therefore, a dataset with higher missing values is prone to less accurate value replacement, with anything above 10% having a higher proportion of miscalculated values. KNN in row appears to be comparatively successful in handling large datasets with CP, while KNN/Forward filling appears to be better at treating large missing datasets with GP.

3. Identification of the best imputation technique(s): From a combined result perspective (SD, RMSE decomposition), KNN and forward filling technique appear to be the most favourable in treating GP missing patterns along with the correct state of charge (+/−). On the other hand, KNN in row and regression appear to be the most favourable techniques, as these two methods produced imputed datasets resembling the original one. This can be confirmed from the total data pattern, the trends, and the daily/weekly/yearly pattern observed.

**FIGURE 9** Daily patterns
4 | CONCLUSION

The existence of missing values in ESS databases is inevitable due to a variety of technical issues ranging from sensor or measurement infrastructure issues to communication network failures. Peak demand management through energy storage and shifting energy load is perceived to be an important area of investigation for ensuring low operational cost and the uptake of carbon efficient storage technologies [39,40]. However, no clear guideline on value error acceptance threshold exist on optimising these operational methods in addition to demand response modelling. Missing data capturing opportunities is a critical issue that can lead to reduced accuracy and compromised integrity of energy storage management [41,42]. Data loss can also significantly reduce the accuracy of peak demand usage forecasts and other analytical conclusions, severely affecting the control and management of an ESS. The main aim of this research was to identify the most suitable data imputation technique that can most accurately replace missing data from BESS. We recreated real life scenarios by replicating quantity and patterns of missing data. Eight different imputation techniques were applied to test their accuracy in replacing missing data totalling 141 imputation trials. The authors presents some unique insights highlighting the importance selecting the right imputation technique in the right scenario. This study provides a methodological roadmap towards approaching data imputation within the energy storage context. The conducted experiment shows that different techniques have their own unique advantages and disadvantages, however, their level of accuracy is highly dependent on the characteristics of missing data frames (namely patterns and sequence). Further, it provided us the opportunity to demonstrate applicability of sequential KNN, in the form of KNN in row, within a storage operations dataset. Although the technique has gained moment in current imputation research, it has shown substantial limitations in replacing GP in our research. Finally, it is emphasised that in addition to the replaced data values, SoC is also an important parameter to be considered within a battery operations dataset. Applying a decomposition technique to our study provided the opportunity open a new avenue for debate and exploration in data imputation research. This is important as some of the techniques applied to our dataset presented distorted results despite replacing statistically significant values within test dataframes.

The work of authors paves the way for future research that could focus on extending the scope of imputation to
larger and more complex datasets. Big datasets obtained from complex BESS projects could be analysed to develop a generalised version of data imputation techniques best suited to custom datasets and individual ESS management scenarios. Battery storage data collected from solar systems can also provide an interesting perspective to future study as their charging and discharging patterns are more versatile depending on the changing climate and weather patterns. More complex machine learning techniques and algorithms along with neural networks (ANN) could be developed to better predict individual missing data values with greater accuracy. Iteration and reinforced learning (RNN) methods could also be developed and applied to complex datasets to increase the accuracy of prediction to larger datasets with missing values accounting for more than 30%. Future research in the area has the potential to complement forecasting techniques by measuring the impact of data imputation accuracy on various forecasting models.

It is appropriate to highlight that the imputation methods suitable to impute values for a particular type of data and application might not be applicable to another dataset with different characteristics of another application.

**FIGURE 11** Yearly patterns

**TABLE 2** Method selection

| Method | RMSE | SD | Trend | Patterns |
|--------|------|----|-------|----------|
| CP     | KNN in row | KNN in row (all) | Regression | Regression |
| KNN    | Forward, KNN (10%) | MCMC | MCMC |
| GP     | KNN | Regression (20, 30%) | KNN in row | KNN in row |
|        | Forward | Forward |
|        | Regression (20, 30%) | KNN |
ACKNOWLEDGMENTS
This research was carried out under the Active Building Centre (ABC), supported by Industrial Strategy Challenge Fund under Programme Grant EP/S016627/1

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID
Mebdi Pazhoohesh https://orcid.org/0000-0001-7268-3355

REFERENCES
1. Rowe, M., et al.: The real-time optimisation of DNO owned storage devices on the LV network for peak reduction. Energies. 7(6), 3537–60 (2014)
2. Nourai, A., Schauer, C.: Changing the electricity game. IEEE Power Energy Mag. 7(4), 42–7 (2009)
3. Manz, D., Pwko, R., Miller, N.: Look before you leap: the role of energy storage in the grid. IEEE Power Energy Mag. 10(4), 75–84 (2012)
4. Soleiman, A., Vahidinavab, V.: Integrating battery condition and aging into the energy management of an active distribution network. In: Smart grid conference (SGC) 2019 Dec 18, pp. 1–6. IEEE (2019)
5. Wong, L.A., et al.: Review on the optimal placement, sizing and control of an energy storage system in the distribution network. J. Energy Storage. 21, 489–504 (2019)
6. Luo, L., et al.: Optimal scheduling of a renewable based microgrid considering photovoltaic system and battery energy storage under uncertainty. J. Energy Storage. 28, 101306 (2020)
7. Chae, Y.T., et al.: Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy Build. 111, 184–94 (2016)
8. Hong, T., et al.: Cost of temperature history data uncertainties in short term electric load forecasting. In: IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems, pp. 212–217. IEEE (2010)
9. Brugnan, PV., et al.: Stacking battery energy storage revenues with enhanced service provision. IET Smart Grid. 3(4), 520–9 (2019)
10. Tureczek, A., Nielsen, PS., Madsen, H.: Electricity consumption clustering using smart meter data. Energies. 11(4), 859 (2018)
11. Meyer, D., Wang, J.: Integrating ultra-fast charging stations within the power grids of smart cities: a review. IET Smart Grid. 1(1), 3–10 (2018)
12. Australian Energy Market Operator (AEMO): Metrology Procedure: Part B: Metering Data Validation, Substitution and Estimation Procedure for Metering Types 1–7. AEMO, Melbourne, Australia (2014)
13. The Electric ESP Handbook, PG & E, San Francisco, CA. (2017)
14. Uniform Business Practices for Unbundled Electricity Metering. Edison Electric Institute (2000)
15. Interim Mindanao Electricity Market (IMEM): Metering Standards and Procedures Issue 1.0. IMEM: Tagaytay, Philippines (2013)
16. Fowler, K.M., et al.: Simplified Processing Method for Meter Data Analysis. Pacific Northwest National Lab.(PNNL), Richland, WA (2015)
17. Peppanan, J., et al.: Handling bad or missing smart meter data through advanced data imputation. In: IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5. IEEE (2016)
18. Kim, M., et al.: Learning-based adaptive imputation method with knn algorithm for missing power data. Energies. 10(10), 1668 (2017)
19. Schmitt, P., Mandel, J., Guedj, M.: A comparison of six methods for missing data imputation. J. Biometr. Biostat. 6(1), 1 (2015)
20. Sundararajan, A., Sarwat, A.I.: Evaluation of missing data imputation methods for an enhanced distributed PV generation prediction. In: Proceedings of the Future Technologies Conference, pp. 590–609. Springer, Cham (2019)
21. Krause, R.W., et al.: Missing network data a comparison of different imputation methods. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 159–163. IEEE (2018)
22. Pazhoohesh, M., Pourmirza, Z., Walker, S.: A comparison of methods for missing data treatment in building sensor data. In: IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), pp. 255–259. IEEE (2019)
23. Greenwood, D., et al.: Scheduling power and energy resources in the smarter network storage project. The 23rd International Conference and Exhibition on Electricity Distribution (2015)
24. Silva-Ramírez, E.L., et al.: Missing value imputation on missing completely at random data using multilayer perceptions. Neural Netw. 24(4), 121–9 (2011)
25. Tremblay, M.C., Dutta, K., Vandermeer, D.: Using data mining techniques to discover bias patterns in missing data. JDIQ. 2(1), 1–9 (2010)
26. Dagnely, P., et al.: Predicting hourly energy consumption. Can regression modeling improve on an autoregressive baseline? In: International workshop on data analytics for renewable energy integration, 2015 Sep 11, pp. 105–122. Springer, Cham.
27. Park, C.H., Kim, S.B.: Sequential random k-nearest neighbor feature selection for high-dimensional data. Expert Syst. Appl. 42(5), 2336–42 (2015)
28. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. Springer (2016)
29. Troyanskaya, O., et al.: Missing value estimation methods for DNA microarrays. Bioinformatics. 17(6), 520–5 (2001)
30. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press (2006)
31. Liu, Z.G., et al.: Evidence combination based on credal belief redistribution for pattern classification. IEEE Trans. Fuzzy Syst. 28(4), 618–31 (2019)
32. Kang, M., et al.: A sequential K-nearest neighbor classification approach for data-driven fault diagnosis using distance-and density-based affinity measures. In: International conference on data mining and big data, pp. 253–261. Springer, Cham (2016)
33. Schunk, D.: A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. ASSTA Adv. Stat. Anal. 92(1), 101–14 (2008)
34. Strehl, D.J., Bühlmann, P.: MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 28(1), 112–8 (2012)
35. Mozharovskyi, P., Josse, J., Husson, F.: Non-parametric imputation by data depth. J. Am. Stat. Assoc. 115(S2), 241–53 (2020)
36. Fearnhead, P.: Markov chain Monte Carlo, sufficient statistics, and particle filters. J. Comput. Graph. Stat. 11(4), 848–62 (2002)
37. Laskey, K.B., Myers, J.W.: Population Markov chain Monte Carlo. Mach. Learn. 50(1-2), 175–96 (2003)
38. Cleveland, R.B., et al.: STL: a seasonal-trend decomposition. J. Official Stat. 6(1), 3–73 (1990)
39. Davis, M.J., Hiralal, P.: Batteries as a service: a new look at electricity peak demand management for houses in the UK. Procedia Eng. 145, 1448–55 (2016)
40. Schmidt, O., et al.: The future cost of electrical energy storage based on experience rates. Nat. Energy. 2(8), 1–8 (2017)
41. Abdulla, K., et al.: Optimal operation of energy storage systems considering photovoltaic and battery energy storage under uncertainty. J. Energy Storage. 28, 101306 (2020)
42. Mozharovskyi, P., Josse, J., Husson, F.: Non-parametric imputation by data depth. J. Am. Stat. Assoc. 115(S2), 241–53 (2020)
43. Fearnhead, P.: Markov chain Monte Carlo, sufficient statistics, and particle filters. J. Comput. Graph. Stat. 11(4), 848–62 (2002)
44. Laskey, K.B., Myers, J.W.: Population Markov chain Monte Carlo. Mach. Learn. 50(1-2), 175–96 (2003)
45. Cleveland, R.B., et al.: STL: a seasonal-trend decomposition. J. Official Stat. 6(1), 3–73 (1990)
46. Davis, M.J., Hiralal, P.: Batteries as a service: a new look at electricity peak demand management for houses in the UK. Procedia Eng. 145, 1448–55 (2016)
47. Schmidt, O., et al.: The future cost of electrical energy storage based on experience rates. Nat. Energy. 2(8), 1–8 (2017)
48. Abdulla, K., et al.: Optimal operation of energy storage systems considering forecasts and battery degradation. IEEE Trans. Smart Grid. 9(3), 2086–96 (2016)
49. Gogotsi, Y., Simon, P.: True performance metrics in electrochemical energy storage. Science. 334(6058), 917–8 (2011)

How to cite this article: Pazhoohesh M, Allahham A, Das R, Walker S. Investigating the impact of missing data imputation techniques on battery energy management system. IET Smart Grid. 2021;4:162–175. https://doi.org/10.1049/stg2.12011