Expert Systems With Applications 174 (2021) 114743

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Towards missing electric power data imputation for energy management systems

Ming-Chang Wang [a], Chih-Fong Tsai [b], Wei-Chao Lin [c,d,*]

[a] Department of Business Administration, National Chung Cheng University, Chia-yi, Taiwan
[b] Department of Information Management, National Central University, Zhongli, Taoyuan, Taiwan
[c] Department of Information Management, Chang Gung University, Taoyuan, Taiwan
[d] Department of Thoracic Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan

ARTICLE INFO

Keywords:
Data mining
Electric power data
Energy management system
Missing value imputation
Machine learning

ABSTRACT

Demand for electricity is gradually increasing in many countries. Efforts in related studies have been made for the application of data mining techniques over related electric power data for the development of more effective energy management systems. However, one major challenge is how to compensate for parts of the collected dataset, such as power consumption, voltage, or electric current that may be missing for a specific period of time. In the literature, several methods have been employed for imputation of the missing data, especially single feature value imputation. However, the performance of the different types of imputation methods, i.e. statistical and machine learning methods, for multiple missing features of electric power data has not been fully explored. Moreover, variations in their imputation performance during the summer/non-summer seasons and in the peak/off-peak/semi-peak times have not been investigated. In this paper, the performance of five well-known imputation methods for processing electric power data, two statistical methods, autoregressive integrated moving average (ARIMA) and linear interpolation (LI) models, and three machine learning methods, k-nearest neighbor (K-NN), multilayer perceptron (MLP), and support vector regression (SVR) is compared. The experimental results, based on electric power data for a two-year period in Taiwan, show that the machine learning methods generally perform better than the statistical ones, with K-NN and SVR performing the best. In particular, all of the imputation methods produced higher error rates during the summer season than the non-summer seasons. Moreover, the machine learning methods (especially K-NN) are better choices for the imputation of missing data during peak times, whereas the statistical methods (especially LI) are better for off-peak and semi-peak times.

1 Introduction

Demand for electricity is increasing gradually worldwide. According to the International Energy Agency[1], global electricity demand rose by 4% in 2018, at its fastest pace since 2010. This is because of strong economic growth in many developing countries (Maluf de Lima & Piedade Bacchi, 2019; Krishnamurthy & Kriström, 2015). Moreover, global warming has also had an impact on demand for electricity for heating and cooling to cope with extreme seasonal changes in the weather (Alberini et al., 2019; Dam et al., 2017; Fan et al., 2019).

To this end, energy saving, which is the major objective of developing effective energy management systems, has become more and more important (Huo et al., 2017; Sher et al., 2019; Tekkaya, 2018). An energy management system (EMS) is a computer system for automatically controlling and monitoring related electromechanical facilities that yield significant energy consumption (Doty & Turner, 2012; Weitzel & Glock, 2018). More specifically, related data collected from sensors, such as electricity meters, water meters, thermometers, etc. act as input to the EMS for effective control of electromechanical facilities, including air conditioners, lights, electric power transmission, sprinklers, and related energy saving equipment.

Problems related to control, management, maintenance, and energy consumption can be certainly solved through systematic data analysis of the collected data. Data analytics can be applied to predict electricity

* Corresponding author at: Department of Information Management, Chang Gung University, Taoyuan, Taiwan.
E-mail addresses: mcwang@ccu.edu.tw (M.-C. Wang), cftsai@mgt.ncu.edu.tw (C.-F. Tsai), viclin@gap.cgu.edu.tw (W.-C. Lin).

[1] https://www.iea.org/geco/electricity/

https://doi.org/10.1016/j.eswa.2021.114743
Received 11 August 2020; Received in revised form 24 December 2020; Accepted 13 February 2021
Available online 20 February 2021

demand at different times, short-term (Al-Musaylh et al., 2018) and long-term (Angelopoulos et al., 2019), and for higher air temperatures (Burillo et al., 2019). Moreover, the analytical results can allow us to understand the electric power loadings and energy saving performance, as well as assist in making better related policies (Bigerna et al., 2019; Chesser et al., 2018; Frondel et al., 2019; Leszczyna et al., 2019; Rogers & Polak, 2013; Wang et al., 2018).

However, due to the fact that synchronous transmission between the sensors and the databases in the network is not always stable and for other reasons, such as an interruption in communication transmission, unreliable communications equipment, low quality of network communication, or environmental factors, the data collection process is disrupted or specific electric power levels are not detected. As a result, parts of the collected electric power data stored in the EMS for a specific period of time will be missing information. Fig. 1 shows an example of the industrial electrical current information for July 1st 2018 observed at 3-hour intervals; note that the data from 08:00 to 09:00 are missing.

According to Little and Rubin (2002), there are three missing data mechanisms, which are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Little & Rubin, 2002). In this study, which is based on missing electric power data, the mechanism belongs to the MCAR type. Missing data in MCAR are caused by force majeure, such as the facility problem and poor communication, which relate to environmental factors.

In the literature, one can find many examples of energy related datasets with missing values that need to be dealt with, such for air pollution data (Junninen et al., 2004; Norazian et al., 2008), electric power data (Kim et al., 2017; Turrado et al., 2015), smart meter data (Peppanen et al., 2016), solar data (Demirhan & Renwick, 2018; Žliobaitė et al., 2014), weather data (Kim et al., 2019), and wind speed/turbine data (Martinez-Luengo et al., 2019; Wesonga, 2015). Some related works focus on the problem of missing value imputation for time series datasets (Afrifa-Yamoah et al., 2020; Andiojaya & Demirhan, 2019; Bokde et al., 2018; Cao et al., 2018; John et al., 2019). In such studies, various missing value imputation methods have been used, which can be divided into statistical methods, e.g. autoregressive integrated moving average (ARIMA), linear interpolation and regression, and machine learning methods, e.g. k-nearest neighbor (K-NN), multilayer perceptron (MLP), and support vector regression (SVR) (Lin & Tsai, 2019).

However, the problem of missing value imputation has not been fully explored for the domain of electric power data. For example, Kim et al. (2017) only focused on K-NN and linear interpolation, whereas Turrado et al. (2015) focused on two related regression methods. In this paper, five well-known imputation methods are considered, which include two statistical methods, i.e. ARIMA and linear interpolation and three machine learning methods, i.e. K-NN, multilayer perceptron, and support vector regression.

Moreover, although the amount of training data used to construct the imputation models is an important factor affecting the imputation result, related studies of missing value imputation have rarely investigated this issue (Lin & Tsai, 2019). Here, we also compare imputation performance using 1, 7, and 30 days of electric power training data.

On the other hand, a lot of relevant data can be collected from electric meters, such as real/unit time electric power consumption, voltage, electric current, etc. However, there are two types of missing values in electric power data for a specific time. The first one is single missing values and the second one is multiple missing values (Sridharan & Schultz, 2001; Tram, 2008). In this paper, we focus on multiple missing values for each data, which is a much more challenging problem than the imputation of single missing values.

The rest of the paper is organized as follows. Section 2 overviews the related literature about missing electric power data and the five well-known imputation methods used in this paper. Sections 3 and 4 present the research methodology and experimental results, respectively, while section 5 outlines the conclusion.

2. Research methodology

2.1 Dataset

The dataset was collected from the EMS records of an electric power station in Taoyuan City in Taiwan, recording data for every minute. The dataset covers the period between January 1, 2017 and December 31, 2018. In addition, each data sample is represented by five different features, which are voltage (V), electric current (I), unit time electric power consumption (UPC), electric power factor (PF), and real-time electric power consumption (RPC).

In particular, we focus on multiple missing data within a one day period, rather than shorter or single specific time periods, such as one minute. The imputation of multiple missing data for one day is more difficult than the imputation of a single missing data in one minute. First of all, the data having one or more missing feature values are removed from the original dataset, which results in a complete dataset. Using the complete dataset without missing feature values can be divided into training and testing sets for constructing the imputation models and examining their imputation performances, respectively. More specifically, we consider three types of consecutive periods to generate the training sets in order to construct the imputation models for performance comparison, which are based on one day, 7-day, and 30-day training sets.

Next, each imputation model constructed by these training sets is

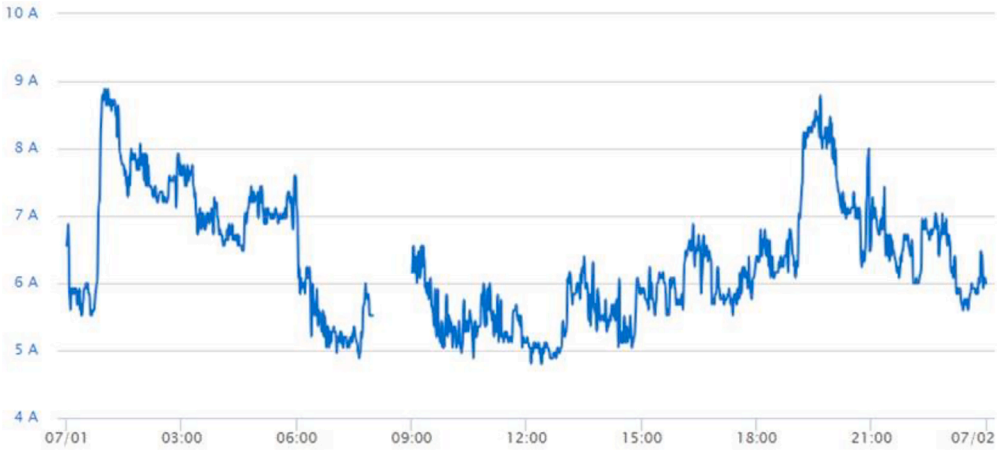


Fig. 1. An example of missing electric power information.

used to impute (or estimate) the feature values of their next days. For example, January 1 of 2017 (one day), January 1 to 7 of 2017 (7-day), and January 30 of 2017 (30-day) are used for the three types of training sets, and their corresponding testing data are January 2 of 2017, January 8 of 2017, and January 31 of 2017, respectively. Consequently, the 1-/7-/30-day experimental datasets contain 625, 596, and 665 groups of training and testing data, respectively.

Moreover, electricity consumption rates in two different time settings are examined. They are peak (week days from 7:30 am to 22:30 pm), off-peak (Monday to Saturday from 00:00 am to 7:30 am and 22:30 pm to 24:00 pm and all of Sunday), and semi-peak times (Saturday from 7:30 am to 22:30 pm) and the summer (June 1st to September 30th) and non-summer seasons.

2.2 Imputation methods

After the training datasets are collected, they are used to construct different imputation models for imputing the missing electric power data for the testing dataset. In this paper, two statistical and three machine learning imputation methods are employed for performance comparison. The statistical imputation methods are linear interpolation (LI) and autoregressive integrated moving average (ARIMA). On the other hand, the machine learning imputation methods include K-NN, SVR, and MLP. These methods are widely used for missing value imputation of energy related datasets. Table 1 lists the imputation methods used in related works.

2.2.1 Linear interpolation (LI)

Linear interpolation (LI) is a popular method for handling missing energy data. It is a curve fitting method, which is based on the use of linear polynomials to construct a curve that best fits a series of data points. Only two data points are required to construct new data points or estimate the missing data (Junninen et al., 2004). However, the estimation quality is reduced when the period of continuous missing data increases.

Suppose the last data point before the missing data A_k is A_n at time K_n and the next data after the missing data A_k is A_{n+1} at time K_{n+1}. In addition, K_k is between K_n and K_{n+1}, and the missing data A_k is associated with the time K_k. As a result, A_k can be estimated by

$$\frac{An - Ak}{An - An + i} = \frac{Kn - Kk}{Kn - Kn + i} \tag{1}$$

In general, linear interpolation (LI) is used to estimate missing data for periods shorter than two hours. Typically, for a period longer than two hours, estimation is based on the historical data for each working day from which to construct the fitting curve for each day. Holidays and other specific conditions are considered individually (Edison Electric Institute, 2000).

2.2.2 Autoregressive integrated moving average (ARIMA)

Table 1
The imputation methods used in related works.

Works	ARIMA	LI	K-NN	MLP	SVR
Afrifa-Yamoah et al. (2020)	V				
Chong et al. (2016)			V		V
Demirhan and Renwick (2018)	V	V			
Junninen et al. (2004)		V	V	V	
Kim et al. (2017)		V			V
Kim et al. (2019)		V	V		
Martinez-Luengo et al. (2019)				V	
Noazian et al. (2008)		V			
Peppanen et al. (2016)		V			
Velasco-Gallego and Lazakis (2020)	V		V	V	V

ARIMA is a typical time series analysis model. An ARIMA model can be used to predict future points in time series data that contain periodical data such as year, season, month, week, and day. ARIMA is composed of AR (i.e. autoregressive), MA (i.e. moving average), and I (i.e. integrated) components, in which AR focuses on the regression of the evolving variable based on its own prior values, MA calculates the regression error as a linear combination of error terms whose values occurred at the same time and various times in the past, and I integrates the data values that are replaced with the difference between their values and the previous values (Mills, 1990).

In particular, AR is based on a set of time series data, denoted as x_1 to x_{t-1} used to predict current data x_t. It is assumed that there is a linear relationship between them, which can be calculated by

$$x = \mathbb{C} + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{2}$$

where c is constant, ε_t means the error terms, and $\varphi_1 \cdots \varphi_p$ indicates the model parameters.

MA assigns different weights to current or past data in order to ensure stability of the model in any circumstance by

$$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

where μ is the average of the time series; $\theta_1 \ldots \theta_q$ mean the model parameters; and $\varepsilon_t \cdots \varepsilon_{t-q}$ stands for the error terms.

Therefore, the ARIMA model is based on

$$\left(1 - \sum_{i=1}^{p} \varnothing_i L^i\right)(1 - L)^d x_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t, \tag{4}$$

where p, q, and d represent the AR, MA, and I parts, respectively; and L is the lag operator.

2.2.3 K-nearest neighbor (K-NN)

K-NN is a supervised learning technique used for classification and regression problems. It is the most widely used imputation method in many domain problem datasets (Lin & Tsai, 2019). The process of K-NN is summarized below (Shakhnarovich et al., 2006). Given a training dataset and testing data to be classified, the distances between the testing data and each of the training data are calculated based on a distance function, usually by the Euclidean distance. For example, the Euclidean distance between x_i and x_j can be obtained by

$$\sqrt{\sum_{l=1}^{n} \left(x_l^i - x_l^j\right)^2} \tag{5}$$

where each data contains n features.

Consequently, the top k nearest neighbors (i.e. the training data) to the testing data can be identified. Therefore, the same class that most of the k training data belong to is assigned for the testing data. In other words, the k value is the major parameter of K-NN.

Generally speaking, the most intuitive nearest neighbor model is 1-NN, which predicts or classifies a data point to the class of its closet neighbor in the feature space. 1-NN is usually used as the baseline classifier since it can provide reasonable performance in many domain problems (Jain et al., 2000). In this study, there are 31 different k values used to identify the best K-NN model, which are 5, 10, 20, 30, …, and 300.

2.2.4 Support vector regression (SVR)

The support vector machines (SVM) algorithm, introduced by Cortes and Vapnik (1995), is based on the statistical learning theory. Generally speaking, the original two-class training data are mapped into a new higher dimensional feature space by a kernel function, which can be a linear, polynomial, or radial basis function (RBF), etc. As a result, a

separating maximum-margin hyperplane in the new feature space is produced to linearly separate the two classes. Moreover, the training data on the margin, which are called the support vectors, are used to distinguish between two different classes.

Support vector regression (SVR) can be used for regression problems. SVR aims at finding the hyperplane for accurate prediction of the real numbers rather than the classes. However, the main idea is the same as for the SVM, which is to minimize the error, and individualize the hyperplane that maximizes the margin.

Given a training dataset denoted as $(x_1, y_1), (x_2, y_2), \cdots (x_l, y_l) \in R^d \times R$, where x means the input features and y the regression values, construct an SVR model as follows:

$$\text{minimize } \frac{1}{2}\|w\|^2; \text{ subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon, \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon, \end{cases} \tag{6}$$

where ε is a free parameter for the threshold and the inner product plus $\langle w, x_i \rangle + b$ is the output prediction for x_i.

In this study, the RBF kernel function is used and the margin of tolerance, i.e. epsilon, is set from 0.1, 0.2, 0.3, …, to 1, and the cost function is set from 1, 10, 20, 30, …., to 90. In total, there are 100 different parameter combinations used to construct the SVR imputation models where the best one will be used to compare with other imputation methods.

2.2.5 Multi-layer perceptron (MLP)

The multi-layer perceptron (MLP) is one of the most widely used artificial neural networks, and is a non-parametric method. It is composed of an input layer, one or more hidden layers, and an output layer, in which each layer contains a number of nodes or neurons. Some pre-defined numbers of nodes residing in different layers are connected to each other with a specific weight. Based on the backpropagation algorithm, the weights between different layers of nodes are adjusted in order to maximize the prediction accuracy over the given training dataset (Bishop, 1995).

Specifically, each node uses a nonlinear activation function, except for the input layer of nodes. The two historically common activation functions are both sigmoids:

$$y(v_i) = \tanh(v_i) \text{and} y(v_i) = (1 + e^{-v_i})^{-1} \tag{7}$$

In addition, the output of the output layer is softmax $(W^{(2)}x_1 + b^{(2)})$, where x_1 means the output of the hidden layer $f(W^{(1)}x + b^{(1)})$. Therefore, the three layers of MLP can be described by

$$f(x) = G\big(b^{(2)} + W^{(2)}\big(s\big(b^{(1)} + W^{(1)}x\big)\big)\big) \tag{8}$$

where x means the input feature vectors, G and s are softmax and sigmoid functions, respectively.

In this study, the hidden layer size is set as 5, 10, 20, 30, …, and 100 and the number of iterations to train the neural network is set from 100, 200, 300, …, to 800. As a result, there are 88 different MLP imputation models constructed for performance comparison.

2.3 Evaluation metric

To assess the performances of the imputation methods, as the regression problem, there are several evaluation metrics that can be considered, where mean absolute percentage error (MAPE) and root-mean-square error (RMSE) are two widely used metrics in related literatures. However, in our study, since each data is represented by five different types of features and the range of these feature values is from about 0.9 to 24,000, MAPE is a more suitable metric than RMSE. This is because RMSE is scale-dependent error meaning that both errors and data are on the same scale (Hyndman & Koehler, 2006). RMSE is not a suitable metric to compare the results of two different forecasts with different units.

MAPE is mostly used to assess the precision of the prediction values for time series related simulations. Not only are the differences between the predicted values and real vales considered, but also the percentage between them. In addition, the average error will not be influenced by extreme values like for our data features.

MAPE can be measured by

$$\frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{h(x^{(i)}) - y^{(i)}}{h(x^{(i)})}\right| \tag{9}$$

where $h\big(x^{(i)}\big)$ means the real value and $y^{(i)}$ is the predicted value.

3 Experimental results

3.1 Machine learning imputation methods

3.1.1 K-Nn

Table 2 shows the lowest MAPE rates for missing electric power data imputation by K-NN trained using the 1-/7-/30-day training datasets in terms of voltage (V), electric current (I), unit time electric power consumption (UPC), electric power factor (PF), and real-time electric power consumption (RPC). In addition, the best performing k values are shown in brackets.

These results indicate that the best imputation performance for missing electric power data is obtained using the 1-day dataset for training, for all five different features. This allows K-NN to perform the best over the other two types of training datasets. The UPC imputation results were the worst. Fig. 2 shows the distribution of the MAPE rates for imputation of the UPC feature values during the two-year period. We can see a rapid increase in the MAPE rates in the second year. This is because the number of missing intervals in the second year is larger than in the first year (c.f. Fig. 3), which limits the training dataset size in the second year. Moreover, there are 45 data in the second year with missing rates in excess of 10% whereas the first year only contains 29 (see Fig. 4).

Since K-NN provides the lowest MAPE rates for the 1-day training dataset, the following analysis is based on the 1-day training dataset. Table 3 shows the MAPE rates for the best K-NN for imputation of the five missing feature values in the summer and non-summer seasons. The imputation results for the voltage feature are better for the summer than the non-summer season. In contrast, for the imputation of the other four features, K-NN performs better during the non-summer than the summer season.

Table 4 shows the MAPE rates for K-NN for the imputation of the five features in the peak, off-peak, and semi-peak times. We can see that the imputation results for the V, I, and RPC features are better during the off-peak time than at other times, whereas the lowest MAPE rates are obtained for UPC and PF feature value imputation in the peak time.

3.1.2 SVr

Table 5 lists the MAPE rates after SVR using the 1-day training dataset for imputation of the five missing feature values as well as the best parameter settings. Note that we only report the imputation results for SVR trained by the 1-day training dataset because the performance is better than for the 7- or 30-day training datasets. Due to the larger

Table 2
MAPE rates for K-NN.

Features	1-day	7-day	30-day
V	0.60% (140)	0.63% (200)	0.68% (180)
I	16.78% (120)	21.29% (150)	32.31% (290)
UPC	693.23% (20)	868.35% (290)	917.53% (290)
PF	0.08% (80)	2.56% (20)	0.013% (10)
RPC	16.71% (80)	21.28% (150)	32.04% (290)

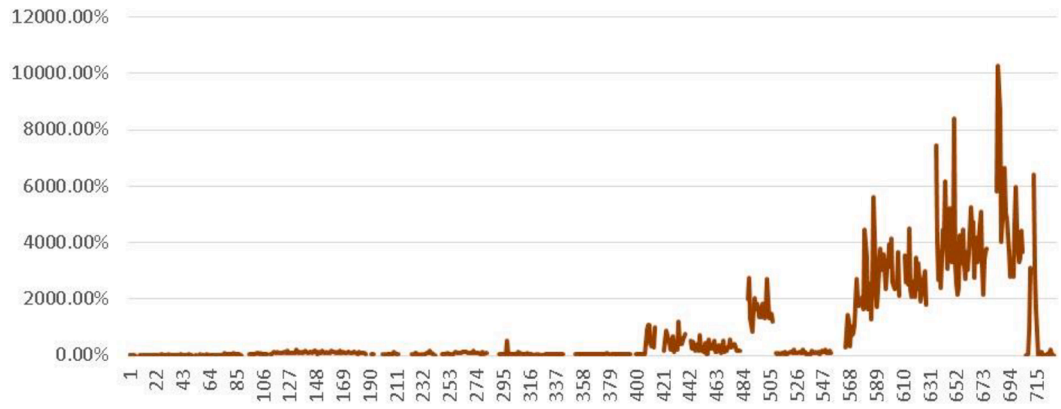


Fig. 2. MAPE rates in the two-year period.

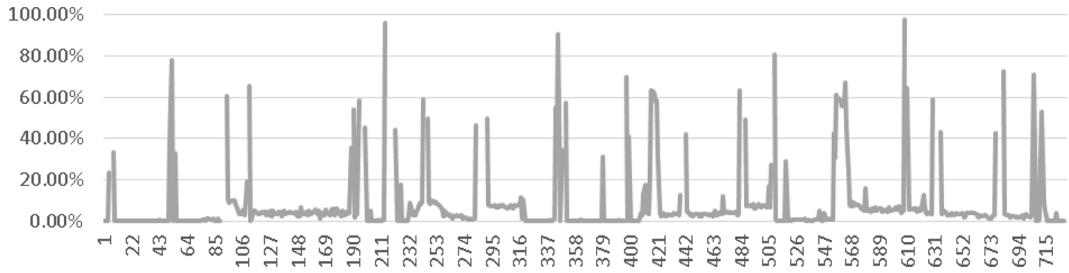


Fig. 3. Missing rates in the two-year period.

number of missing intervals in the second year, the imputation results for the UPC feature are the worst.

Table 6 shows the imputation results after SVR during the summer and non-summer seasons. It can be seen that during the summer season lower MAPE rates can be obtained by SVR for the V and UPC features than for the other features, whereas the imputation results for the I, PF, and RPC features are better during the non-summer seasons in comparison to the V and UPC features.

Table 7 shows the SVR imputation performance for the peak, off-peak, and semi-peak times. These results indicate that SVR produces the lowest MAPE rates for the V, I, and RPC features during the off-peak time, but the lowest rates are produced for the UPC and PF features during the peak time.

3.1.3 MLp

Similar to K-NN and SVR, better MLP performance is obtained using the 1-day training dataset than the 7- or 30-day training datasets. Here, we only report the best MLP results. Table 8 lists the MAPE rates for MLP as well as the best parameter settings for the imputation of the five feature values. Similar to K-NN and SVR, the worst MLP imputation performance is obtained for the UPC features because of the larger missing intervals.

Tables 9 and 10 show the MAPE rates after MLP for the summer/non-summer seasons and the peak/off-peak/semi-peak times, respectively. As we can see, MLP produces better imputation results for the V feature during the summer season than the non-summer seasons, whereas for the other features the results are better during the non-summer seasons. On the other hand, for the imputation of the V, I, UPC, and RPC features, the MLP can produce lower MAPE rates during the peak time, but for the PF feature the results are better in the off-peak time.

3.2 Statistical imputation methods

3.2.1 Li

The LI imputation performance results for the five features are listed in Table 11. It can be seen that during the summer season, LI can produce lower MAPE rates for imputation of the V and PF feature values than for the other three features, whereas the imputation performance for the I, UPC, and RPC feature values is better during the non-summer seasons than the summer season.

On the other hand, LI performs the best for imputation of the UPC feature in the peak time, the V feature in the off-peak time, and the I, PF, and RPC features in the semi-peak time.

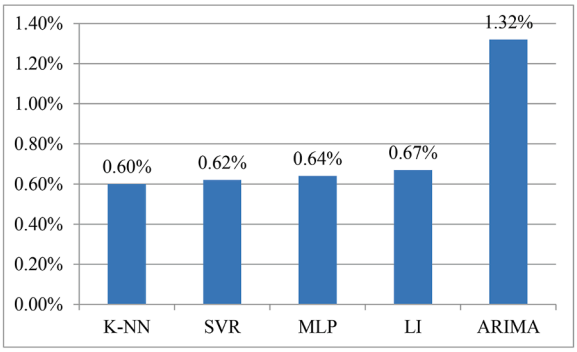
3.2.2 Arima

The ARIMA imputation performance for the five features is summarized in Table 12. As we can see, during the summer season, ARIMA produces lower MAPE rates for imputation of the V and PF feature values, whereas the imputation results for the I, UPC, and RPC features are better during the non-summer season. On the other hand, the lowest MAPE rates for the UPC feature are obtained in the peak time, for the V, PF, and RPC features in the off-peak time, and for the I feature in the semi-peak time.

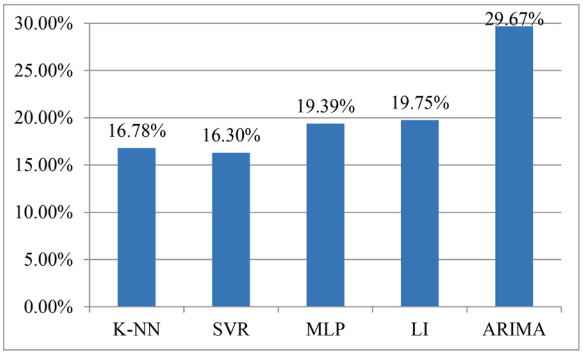
3.3 Discussion

Fig. 3 shows a comparison between the average MAPE rates obtained with five imputation methods. For the V, I, UPC, PF, and RPC features, the best imputation methods are K-NN, SVR, ARIMA, K-NN, and SVR, respectively. On average, except for the UPC feature, K-NN and SVR are the top two imputation methods but there are no significant differences in performance between them. In addition, except for the UPC feature, the machine learning imputation methods usually outperform the statistical ones. This may be because the missing data intervals for the UPC feature are large, meaning that the training data size is small (c.f. Section 3.1.1.). The performance of machine learning imputation methods can be heavily affected by the training data size (c.f. Fig. 3(c)).

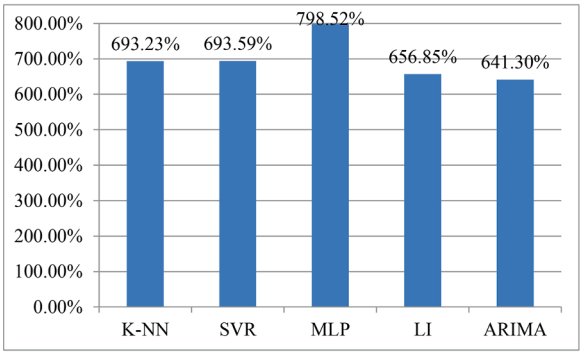
Table 13 shows the results of a comparison of the performance of the



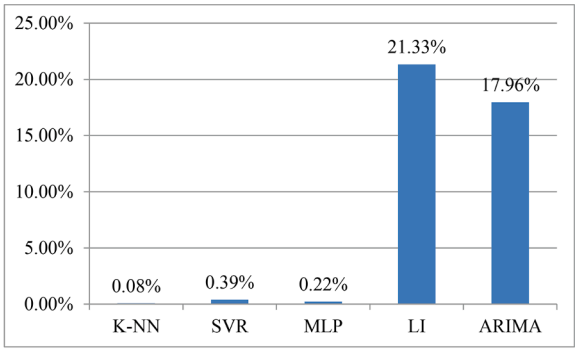
(a) the V feature



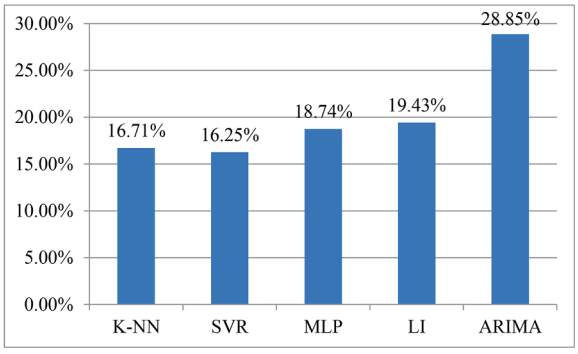
(b) the I feature



(c) the UPC feature



(d) the PF feature



(e) the RPC feature

Fig. 4. Average MAPE rates for the five imputation methods.

Table 3
MAPE rates for the best K-NN for the summer and non-summer seasons.

Features	Summer	Non-summer
V	0.58%	0.62%
I	18.31%	16.42%
UPC	870.42%	793.23%
PF	2.86%	1.21%
RPC	18.26%	16.35%

Table 4
MAPE rates with K-NN for the peak, off-peak, and semi-peak times.

Features	Peak	Off-peak	Semi-peak
V	0.68%	0.56%	0.61%
I	15.1%	12.09%	57.59%
UPC	713.58%	795.02%	1395.52%
PF	6.1%	10.16%	35.35%
RPC	14.86%	12.32%	57.02%

Table 5
MAPE rates and the best parameter settings for SVR.

Features	MAPE	Parameters (cost-epsilon)
V	0.62%	1–0.3
I	16.30%	90–0.2
UPC	693.59%	40–0.1
PF	0.39%	1–0.1
RPC	16.25%	80–0.1

Table 6
MAPE rates for the best SVR for the summer and non-summer seasons.

Features	Summer	Non-summer
V	0.59%	0.63%
I	17.17%	16%
UPC	850.41%	814.69%
PF	5.27%	1.27%
RPC	17.28%	15.9%

Table 7
MAPE rates for SVR for the peak, off-peak, and semi-peak times.

Features	Peak	Off-peak	Semi-peak
V	0.7%	0.56%	0.61%
I	10.51%	10.59%	54.06%
UPC	644.61%	664.85%	1013.93%
PF	6.8%	34.21%	30.41%
RPC	15.23%	10.85%	53.18%

Table 8
MAPE rates and the best parameters of MLP.

Features	MAPE	Parameters (size-iterations)
V	0.64%	60–800
I	19.39%	80–700
UPC	798.52%	70–500
PF	0.22%	30–400
RPC	18.74%	80–600

Table 9
MAPE rates for the best MLP for the summer and non-summer seasons.

Features	Summer	Non-summer
V	0.58%	0.65%
I	21.46%	18.39%
UPC	855.44%	822.2%
PF	0.77%	0.05%
RPC	19.59%	17.85%

Table 10
MAPE rates for MLP for the peak, off-peak, and semi-peak times.

Features	Peak	Off-peak	Semi-peak
V	0.68%	0.58%	0.62%
I	18.31%	18.66%	53.85%
UPC	703.85%	826.05%	1230.76%
PF	7.38%	69.05%	25.36%
RPC	17.43%	15.53%	54.5%

five imputation methods during the summer and non-summer seasons. Note that the best performing methods for the five features are underlined.

As we can see, regardless of the imputation method, performance is only (slightly) better for imputation of the voltage feature value during the summer than the non-summer seasons. This indicates that in most cases, missing electric power data imputation is more of a challenge during the summer.

In short, regardless of the season, the performance of most imputation methods (for the V, I, UPC, PF, and RPC features) is almost the same (K-NN/MLP, SVR, LI, MLP, and SVR, respectively).

Table 14 shows the performance of the five imputation methods during the peak, off-peak, and semi-peak times. With these imputation methods, the lowest MAPE rates are obtained for imputation of the V, I, and RPC feature values during off-peak times, whereas the lowest MAPE rates in the peak times are obtained for imputation of the UPC and PF features. On the other hand, in the semi-peak time, the imputation of most of the features can be very challenging, with most imputation methods producing higher MAPE rates compared to peak and off-peak times.

In short, during peak times, the best imputation methods for the V, I, UP, PF, and RPC features are K-NN/MLP, SVR, LI, K-NN, K-NN, respectively. For the off-peak time, it is interesting that the best imputation methods for the five features are mostly statistical methods, specifically LI, LI, SVR, ARIMA, and LI, respectively. For semi-peak times, the best performing methods for the five features are also mostly statistical methods, namely SVR, LI, ARIMA, LI, and LI, respectively.

These results indicate that the conventional statistical methods are good for imputation of missing electric power data in the off-peak and semi-peak times, whereas the machine learning methods are preferable for peak times.

4 Conclusion

In practice, a certain amount of missing feature values exists in electric power data, which can affect the results of data analysis for the

Table 13
MAPE rates for the five imputation methods during the summer and non-summer seasons.

Features	K-NN	SVR	MLP	LI	ARIMA
Summer					
V	0.58%	0.59%	0.58%	0.61%	0.86%
I	18.31%	17.17%	21.46%	23.5%	31.65%
UPC	870.42%	850.41%	855.44%	674.18%	758.45%
PF	2.86%	5.27%	0.77%	16.56%	16.95%
RPC	18.26%	17.28%	19.59%	23.58%	33.43%
Non-summer					
V	0.62%	0.63%	0.65%	0.72%	1.54%
I	16.42%	16%	18.39%	17.87%	28.72%
UPC	793.23%	814.69%	822.2%	621.35%	701.44%
PF	1.21%	1.27%	0.05%	18.13%	18.44%
RPC	16.35%	15.9%	17.85%	17.42%	26.65%

Table 11
MAPE rates with LI.

Features	Average	Summer	Non-summer	Peak	Off-peak	Semi-peak
V	0.67%	0.61%	0.72%	0.77%	0.49%	0.7%
I	19.75%	23.5%	17.87%	32.21%	9.58%	8.95%
UPC	656.85%	674.18%	621.35%	486.9%	728.92%	743.03%
PF	21.33%	16.56%	18.13%	26.32%	20.4%	0.08%
RPC	19.43%	23.58%	17.42%	31.26%	9.8%	8.96%

Table 12
MAPE rates with ARIMA.

Features	Average	Summer	Non-summer	Peak	Off-peak	Semi-peak
V	1.32%	0.86%	1.54%	1.56%	1.24%	1.41%
I	29.67%	31.65%	28.72%	42.62%	25.77%	25.47%
UPC	641.3%	758.45%	701.44%	490.23%	715.83%	726.4%
PF	17.96%	16.95%	18.44%	40.77%	4.66%	42.97%
RPC	28.85%	33.43%	26.65%	40.25%	25.21%	28.43%

Table 14
MAPE rates for the five imputation methods in the peak, off-peak, and semi-peak times.

Features	K-NN	SVR	MLP	LI	ARIMA
Peak					
V	0.68%	0.7%	0.68%	0.77%	1.56%
I	15.1%	10.51%	18.31%	32.21%	42.62%
UPC	713.58%	644.61%	703.85%	486.9%	490.23%
PF	6.1%	6.8%	7.38%	26.32%	40.77%
RPC	14.86%	15.23%	17.43%	31.26%	40.25%
Off-peak					
V	0.56%	0.56%	0.58%	0.49%	1.24%
I	12.09%	10.59%	18.66%	9.58%	25.77%
UPC	795.02%	664.85%	826.05%	728.92%	715.83%
PF	10.16%	34.21%	69.05%	20.4%	4.66%
RPC	12.32%	10.85%	15.53%	9.8%	25.21%
Semi-peak					
V	0.61%	0.61%	0.62%	0.7%	1.41%
I	57.59%	54.06%	53.85%	8.95%	25.47%
UPC	1395.52%	1013.93%	1230.76%	743.03%	726.4%
PF	35.35%	30.41%	25.36%	0.08%	42.97%
RPC	57.02%	53.18%	54.5%	8.96%	28.43%

energy management system. Most related works employing several imputation methods only consider imputation of single missing feature values in a very short period of time, such as power consumption, electric current, etc. in one minute. In this paper, we focus on assessing two statistical and three machine learning imputation methods for multiple missing feature values in a longer period of one day. Moreover, the differences in imputation performance for the summer and non-summer seasons and the peak, off-peak, and semi-peak times are also examined.

According to our experimental results, the following findings are obtained.

- On average, the machine learning methods, i.e. K-NN, SVR, and MLP, outperform statistical methods, i.e. ARIMA and LI, for four out of five features, which are voltage (V), electric current (I), electric power factor (PF), and real-time electric power consumption (RPC). More specifically, K-NN and SVR perform the best.
- All of the imputation methods produce lower MAPE rates during the non-summer seasons than the summer season. This indicates correct imputation of missing electric power data is more difficult in the summer season than in the non-summer seasons.
- The machine learning methods are more suitable than the statistical methods for imputation of the missing electric power data in the peak times, with K-NN performing better for most of the features.
- The machine learning methods are more suitable than the statistical methods for imputation of the missing electric power data in the peak times, with K-NN performing better for most of the features.
- On the other hand, statistical imputation methods are better choices for the off-peak and semi-peak times, with LI outperforming the other methods for most of the features.

For the practical contributions, this study demonstrates the effectiveness of machine learning methods for imputing missing electric power data, which can be used to replace the existing imputation method, i.e. LI, used in the case company. Moreover, better imputation results by using K-NN and SVR allow the prediction of energy consumption more precisely, which can make the energy management system to effectively control electromechanical facilities.

CRediT authorship contribution statement

Ming-Chang Wang: Conceptualization, Formal analysis, Resources, Data curation, Writing - original draft, Writing - review & editing. **Chih-Fong Tsai:** Methodology, Writing - original draft, Writing - review & editing. **Wei-Chao Lin:** Methodology, Software, Validation, Writing - original draft, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 109-2410-H-182-012, and in part by Chang Gung Memorial Hospital, Linkou, under Grant BMRPH13.

References

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications, 27*(1). https://doi.org/10.1002/met.v27.110.1002/met.1873

Al-Musaylh, M. S., Deo, R. C., Adamowski, J. F., & Li, Y. (2018). Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. *Advanced Engineering Informatics, 35*, 1–16.

Alberini, A., Prettico, G., Shen, C., & Torriti, J. (2019). Hot weather and residential hourly electricity demand in Italy. *Energy, 177*, 44–56.

Andiojaya, A., & Demirhan, H. (2019). A bagging algorithm for the imputation of missing values in time series. *Expert Systems with Applications, 129*, 10–26.

Angelopoulos, D., Siskos, Y., & Psarras, J. (2019). Disaggregating time series on multiple criteria for robust forecasting: The case of long-term electricity demand in Greece. *European Journal of Operational Research, 275*(1), 252–265.

Bigerna, S., Wen, X., Hagspiel, V., & Kort, P. M. (2019). Green electricity investments: Environmental target and the optimal subsidy. *European Journal of Operational Research, 279*(2), 635–644.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bokde, N., Beck, M. W., Martínez Álvarez, F., & Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern Recognition Letters, 116*, 88–96.

Burillo, D., Chester, M. V., Pincetl, S., Fournier, E. D., & Reyna, J. (2019). Forecasting peak electricity demand for Los Angeles considering higher air temperatures due to climate change. *Applied Energy, 236*, 1–9.

Cao, W., Wang, D., Li, J., Zhou, H., Li, Y., & Li, L. (2018). BRITS: Bidirectional recurrent imputation for time series. *International Conference on Neural Information Processing*, 6776–6786.

Chesser, M., Hanly, J., Cassells, D., & Apergis, N. (2018). The positive feedback cycle in the electricity market: Residential solar PV adoption, electricity demand and prices. *Energy Policy, 122*, 36–44.

Chong, A., Lam, K. P., Xu, W., Karaguzel, O. T., & Mo, Y. (2016). Imputation of missing values in building sensor data. *Building Performance Modeling Conference*, 407–414.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*(3), 273–297.

Dam, A., Koberl, J., Prettenthaler, F., Rogler, N., & Toglhofer, C. (2017). Impact of +2 ℃ global warming on electricity demand in Europe. *Climate Services, 7*, 12–30.

Maluf de Lima, L., & Piedade Bacchi, M. R. (2019). Assessing the impact of Brazilian economic growth on demand for electricity. *Energy, 172*, 861–873.

Demirhan, H., & Renwick, Z. (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy, 225*, 998–1012.

Doty, S., & Turner, W. C. (2012). *Energy management handbook* (8th Ed.). Fairmont Press.

Edison Electric Institute (2000). Uniform business practices for unbundled electricity metering, volume two. Available online at: http://www.naesb.org/REQ/req_form.asp.

Fan, J.-L., Hu, J.-W., & Zhang, X. (2019). Impacts of climate change on electricity demand in China: Anempirical estimation based on panel data. *Energy, 170*, 880–888.

Frondel, M., Kussel, G., & Sommer, S. (2019). Heterogeneity in the price response of residential electricity demand: A dynamic approach for Germany. *Resource and Energy Economics, 57*, 119–134.

Huo, H., Shao, J., & Huo, H. (2017). Contributions of energy-saving technologies to building energy saving in different climatic regions of China. *Applied Thermal Engineering, 124*, 1159–1168.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transitions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

John, C., Ekpenyong, E. J., & Nworu, C. C. (2019). Imputation of missing values in economic and financial time series data using five principal component analysis approaches. *CBN Journal of Applied Statistics, 10*(1), 51–73.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*(18), 2895–2907.

Kim, M., Park, S., Lee, J., Joo, Y., & Choi, J. K. (2017). Learning-based adaptive imputation method with kNN algorithm for missing power data. *Energies, 10*(10), 1–20.

Kim, T., Ko, W., & Kim, J. (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Applied Sciences, 9*(1), 204–221.
Krishnamurthy, C. K. B., & Kriström, B. (2015). A cross-country analysis of residential electricity demand in 11 OECD-countries. *Resource and Energy Economics, 39*, 68–88.
Leszczyna, R., Wallis, T. & Wrobel, M. (2019). Developing novel solutions to realise the European Energy – Information Sharing & Analysis Centre. Decision Support Systems, 122, Article No. 113067.
Lin, W.-C., & Tsai, C.-F. (2019). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review, 53*(2), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd Ed.). John Wiley and Sons.
Martinez-Luengo, M., Shafiee, M., & Kolios, A. (2019). Data management for structural integrity assessment of offshore wind turbine support structures: Data cleansing and missing data imputation. *Ocean Engineering, 173*, 867–883.
Mills, T. C. (1990). *Time series techniques for economists*. Cambridge University Press.
Noazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia, 34*, 341–345.
Peppanen, J., Zhang, X., Grijalva, S., & Reno, M. J. (2016). Handling bad or missing smart meter data through advanced data imputation. *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, 1–5.
Rogers, D. F., & Polak, G. G. (2013). Optimal clustering of time periods for electricity demand-side management. *IEEE Transactions on Power Systems, 28*(4), 3842–3851.
Shakhnarovich, G., Darrell, T., & Indyk, P. (2006). *Nearest-neighbor methods in learning and vision: Theory and practice*. The MIT Press.
Sher, F., Kawai, A., Güleç, F., & Sadiq, H. (2019). Sustainable energy saving alternatives in small buildings. *Sustainable Energy Technologies and Assessments, 32*, 92–99.
Sridharan, K., & Schultz, N. N. (2001). Outage management through AMR systems using an intelligent data filter. *IEEE Transactions on Power Delivery, 16*(4), 669–675.
Tekkaya, A. E. (2018). Energy saving by manufacturing technology. *Procedia Manufacturing, 21*, 392–396.
Tram, H. (2008). Technical and operation considerations in using smart metering for outage management. IEEE/PES transmission and distribution conference and exposition.
Turrado, C. C., Lasheras, F. S., Calvo-Rolle, J. L., Pinon-Pazos, A. J., & de Cos Juez, F. J. (2015). A new missing data imputation algorithm applied to electrical data loggers. *Sensors, 15*(2), 31069–31082.
Wang, S., Bi, S., & Zhang, Y.-J. (2018). Demand response management for profit maximizing energy loads in real-time electricity market. *IEEE Transactions on Power Systems, 33*(6), 6387–6396.
Weitzel, T., & Glock, C. H. (2018). Energy management for stationary electric energy storage systems: A systematic literature review. *European Journal of Operational Research, 264*(2), 582–606.
Wesonga, R. (2015). On multivariate imputation and forecasting of decadal wind speed missing data. SpringerPlus, 4, Article No. 12.
Velasco-Gallego, C., & Lazakis, I. (2020). Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. *Ocean Engineering, 218*, 108261. https://doi.org/10.1016/j.oceaneng.2020.108261
Žliobaitė, I., Hollmén, J., & Junninen, H. (2014). Regression models tolerant to massively missing data: A case study in solar-radiation nowcasting. *Atmospheric Measurement Techniques, 7*(12), 4387–4399.