# A method for electric load data verification and repair in home environment

## Qi Liu*

Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing, 210044, China
Email: qi.liu@nuist.edu.cn
*Corresponding author

## Shengjun Li

School of Computer and Software, Nanjing University of Information Science and Technology, 219 Ningliu Road, Nanjing, Jiangsu 210044, China
Email:846579040@qq.com

## Xiaodong Liu

School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh EH10 5DT, UK
Email: x.liu@napier.ac.uk

## Nigel Linge

The University of Salford, Salford, Greater Manchester, M5 4WT, UK
Email: n.linge@salford.ac.uk

**Abstract:** Home energy management (HEM) and smart home have been popular among people; HEM collects and analyses the electric load data to make the power use safe, reliable, economical, efficient and environmentally friendly. Without the correct data, the correct decisions and plans would not be made, so the data quality is of the great importance. This paper focuses on the verification and repair of the electric load data in family environment. Due to the irregularity of modern people's life styles, this paper proposes a system of 'N + 1' framework to handle this properly. The system collects information of every appliance and the power bus to make them verify each other, so it can solve the stochastic uncertainty problem and verify if the data is correct or not to ensure the data quality. In the course of data upload, there are many factors like smart meter malfunctions, communication failures and so on which will cause some wrong data. To repair the wrong data, we proposes a method called LBboosting, which integrates two curve fitting methods. As the results show, the method has a better performance than up-to-date methods.

**Keywords:** data verify; data repair; load data quality; power system; home energy management; HEM.

**Reference** to this paper should be made as follows: Liu, Q., Li, S., Liu, X. and Linge, N. (2018) 'A method for electric load data verification and repair in home environment', *Int. J. Embedded Systems*, Vol. 10, No. 3, pp.248–256.

**Biographical notes:** Qi Liu received his BSc in Computer Science and Technology from the Zhuzhou Institute of Technology, China in 2003, and MSc and PhD in Data Telecommunications and Networks from the University of Salford, UK in 2006 and 2010. His research interests include context awareness, data communication in MANET and WSN, and smart grid. His recent research work focuses on intelligent agriculture and meteorological observation systems based on WSN.

Shengjun Li received his Bachelor's in Software Engineering from the Nanjing University of Information, Science and Technology in 2014, and he is currently pursuing his Master's in Software Engineering at the Nanjing University of Information Science and Technology. His research interests include data mining and activity recognition.

Xiaodong Liu is a reader and the Director of Centre for Information and Software Systems in School of Computing at Edinburgh Napier University. His research interests include context-aware adaptive services, service evolution, mobile clouds, pervasive computing, software reuse, and green software engineering. He is a member of IEEE Computer Society and British Computer Society.

Nigel Linge received his BSc in Electronics from the University of Salford, UK in 1983, and PhD in Computer Networks from the University of Salford, UK in 1987. He was promoted to Professor of Telecommunications at the University of Salford, UK in 1997. His research interests include location based and context aware information systems, protocols, mobile systems and applications of networking technology in areas such as energy and building monitoring.

This paper is a revised and expanded version of a paper entitled 'A method for electric load data verification and repair in home environment', presented at 2nd International Conference on Cloud Computing and Security (ICCCS 2016), Nanjing, China, 29–31 July, 2016.

## 1 Introduction

In recent years, the smart grid has developed rapidly; the electric load data has been essential for electric utilities (Chen et al., 2009, 2010). What is more, more and more domestic appliance manufacturers put an eye on the smart home. With implementation of smart home, the system can collect appliances' status data and analyse it to make domestic appliances intelligent (Han et al., 2010) and make home life more comfortable. Now, many manufacturers have produced their smart home products to meet the customers' demand. Construction of smart grid and the application of smart home, no matter macroscopic or microscopic situation, both need to collect the electric load data to support their decision making. After getting the load data, it can be inferred how appliances consume power and even the electricity consuming patterns. For electricity industry, the analysis of electric load data is important for the day-to-day operation, system analysis, energy saving and energy planning. For energy consumers, they can use the feedback information to know how much power appliances consume and help them modify the power consuming activities to reduce the cost.

In home environment, gathering load data accurately is a challenging task. There is often missing and wrong data in the process of collecting and transferring. It is caused by many factors, such as smart meter malfunctions, communication failures, equipment outages and other factors. These cause a significant deviation in load, and cannot represent the correct energy consuming situation. If we work on these wrong datasets, the results would be useless, even misleading. So it is important to identify and correct the wrong data and fill the missing data.

Currently, many approaches were presented to detect and repair wrong load data, but few researchers focused on the load data repair in home environment. In this paper, we propose a novel method to verify and repair the load data, which means that we not only verify the data is quality or not but also fix the wrong data. Missing data is treated as a special case of wrong data. There are three challenges in solving the problems. First, traditional statistical methods cannot be used if a relatively large portion of data is wrong or missing. Second, it is difficult to judge whether a relatively large deviation represents wrong data or an underlying change in data patterns. Third, due to the home environment, it is hard to deal with the random pattern and uncertainty in electricity consumption.

The contributions of this work are as follows. First, we propose a 'N + 1' system to validate if the load data is correct or not and help repair the wrong data. Second, we focus on the stochastic electricity consumption in home environment; the method can avoid the influence of the load data's aperiodicity in home. Also, the solution can be robust to a relatively large portion of missing data, and it can identify whether the deviation represents wrong data or an underlying change in patterns. Finally, to repair wrong data, we use an idea like Adaboost algorithm (Wen et al., 2015) to fuse two interpolation techniques, which are b-spline and Lagrange interpolation, and propose a new method called LBboosting. The results demonstrate the effectiveness and high performance of the proposed solution.

## 2 Related work

A closely related area to the load data repair is outlier detection, which has been broadly studied in lots of fields, such as data mining and statistics. For outlier detection, researchers usually identify the outliers first, then find a proper way to correct them.

As these methods develop, outlier detection in time series has become a general mathematical concept in statistic field to be studied (Hodge and Austin, 2014). In statistic domain, the mean and median method suggested in Weron (2006) considered a given periodicity for the data,

and replaced missing or corrupted data with the average or medians of the corresponding observations at different periods. This method does not fully take data distribution into consideration. A lot of methods assumed that the data follow an underlying known probability distribution. Tang et al. (2014) split the data into some portrait datasets from a brand new perspective, they assume that the datasets follow some distributions, and find the outliers according to the distribution, replace the outliers with the estimated values finally. Assuming data follow a certain distribution can convert the detection problem to finding the abnormal areas in time series. However, the distribution cannot be precisely known to us all the time, and in different families, the distributions cannot be the same. What is more, traditional statistical methods cannot handle the data which a relatively large portion of it is missing.

Some other methods (Zheng et al., 2015; Ljung, 1993; Gu et al., 2015b; Abraham and Chuang, 1989; Xia et al., 2015; Liu and Xiao, 2016) are regression-based. They assume that the load data have some certain pattern, but to home environment, the randomness is an important feature which need to be considered. People's lives are not set carved in stone, incidents occur every day, resulting in the strong stochastic load data. So these methods inevitably have large error or over-fitting problems. Other methods employ smoothing techniques. Chen et al. (2010) proposed a non-parametric regression method based on b-spline and kernel smoothing to fit the load curve data, then set a confidence interval to judge whether a point is an outlier or not, the point can be seen as an outlier if the point fell out of the confidence interval. Guo et al. (2012) assumed that the load data is periodic and its period is known, judged if a point is an outlier depending on whether it follows the periodicity. But some points does not follow the periodicity may not be outliers in fact. In general, the regression-based methods are based on their parameters. Smoothing techniques are sensitive to their training data, if missing data are out of the training data's range or training data cannot represent the original data, the outcome will be bad. As a result, the consequence is subject to these factors.

In the domain of data mining, a lot of methods based on similarity were proposed to detect outliers. Such as k-nearest neighbour (Ramaswamy et al., 2000; Zhou et al., 2016), k-means clustering (Nairac et al., 1999; Liu et al., 2016a), support vector machine (Gu et al., 2015a) and the neural network method such as RNNs (Hawkins et al., 2002) are used. According to the same features (Yang et al., 2013), they classified the observations, and find the observations which are far from the clusters as outliers. However, most of these methods are designed for structured relational data, and they are also time consuming for training process.

## 3 Problem definition

A load curve data is a time series where the load values are collected at a certain time frequency such as every 1 second or hourly. The load data recorded status of domestic appliances, such as their current, voltage, power. The data quality is influenced by various factors such as meter failures, communication interruptions, and dynamism of customers. So, load curve data consists of not only white noises but also some corrupted data.

Nowadays, the technology for smart meters is mature, the precision of smart meters is very high, and the fault rate is low. In home environment, smart meters are used to collect the data. The meters usually use Zigbee to deliver domestic appliances' status data, due to the short transmission distance of Zigbee and indoor complex structure of family, there will be communication interruptions sometimes, resulting in missing data. So the corrupted data is usually caused by the communication interruptions. As a result, the corrupted data in this paper is referred to as missing data. Under these circumstances, the data repair process is mainly to replace the missing data.

In addition, traditional data cleansing methods check that if the observation follows a certain pattern to identify whether the observation is corrupted or not. But the observation which does not follow the pattern may not be really corrupted. And in home environment, modern people's irregular lifestyle may not follow a so-called pattern. So defining a corrupted data in that way is not precise, and we need to find the corrupted data more precisely.

*Definition 1:* A load curve data is a time series $S = \{(y_i, t_i)\}_{i=1}^{n}$ that is an $n$-values sequence ordered by time where $t_i$ is the timestamp and $y_i$ is the observation at the time $t_i$.

In this paper, we focus on two tasks: First, validate if an observation is corrupted precisely. Second, fill the missing data caused by the communication interruptions.

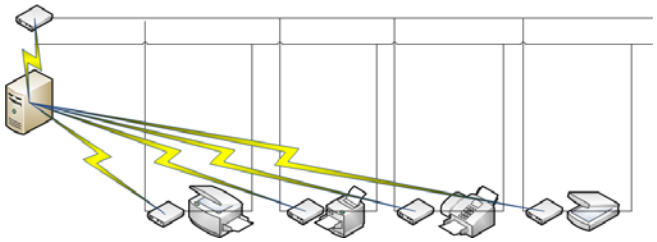
## 4 Proposed methods

As mentioned earlier, modern domestic consumers will not use appliances regularly, that is to say, the habit of using appliances do not follow a certain pattern. So we proposed a new solution to get rid of the influence of domestic consumers' irregular lifestyles.

### *4.1 System construction*

To collect data, we propose a new way to get the data. Assume that the family has N appliances that need to be measured, then we installed N smart meters to get their data respectively. In addition, we installed a smart meter at the power bus for the house. At last, put the server for receiving signals just besides the power bus to ensure the server can receive the signal from the power bus. Figure 1 shows the system layout. In this way, we have constituted our 'N + 1' model to get the data we want. In this paper, the data we collect is power.

**Figure 1** System layout (see online version for colours)



## 4.2 *Validate data quality*

As we know, household electrical circuit is a parallel circuit, so that we can use the following formula to describe the relationship between the power bus and various domestic appliances:

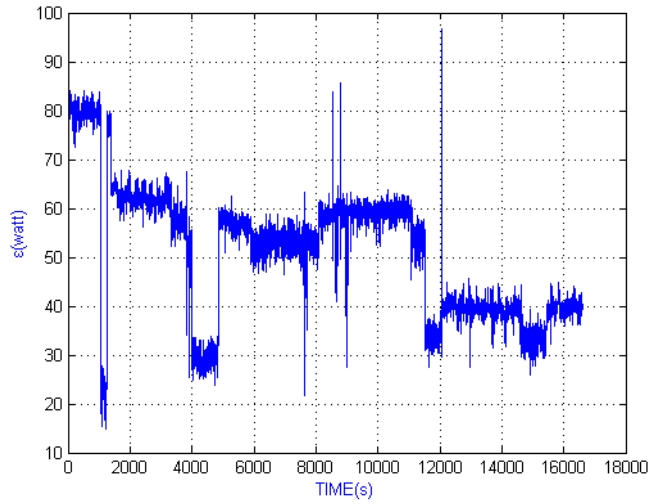
$$P_{All} = \sum_{i=0}^{n} P_i + \varepsilon \tag{1}$$

where the $P_{All}$ is the value of the power bus, $P_i$ is the value of $i^{th}$ domestic appliance and there are n appliances, ε is the error, it mainly refers to the line loss and white noise.

### 4.2.1 *Difference between $P_{All}$ and $\sum_{i=0}^{n} P_i$*

ε is the difference between $P_{All}$ and $\sum_{i=0}^{n} P_i$, in (1) we call it error, but actually it mainly refers to the line loss. The line loss of distribution network is difficult and complex to calculate, but in family environment, the electric circuit is simple, so we collected some training data first. Figure 2 shows some values of ε when various combinations of the appliances work.

**Figure 2** The difference between the values of power bus and sum of appliances when various combination of the appliance work (see online version for colours)




when all domestic appliances and smart meters work well, then the difference between $P_{All}$ and $\sum_{i=0}^{n} P_i$, i.e., ε is subjected to the number of the appliances and their status, from Figure 2, we can easily find there is some relationship between the difference and appliances' status. After getting the training data, here we use extreme learning machine (ELM) to find the relationship.

### 4.2.2 *Extreme learning machine*

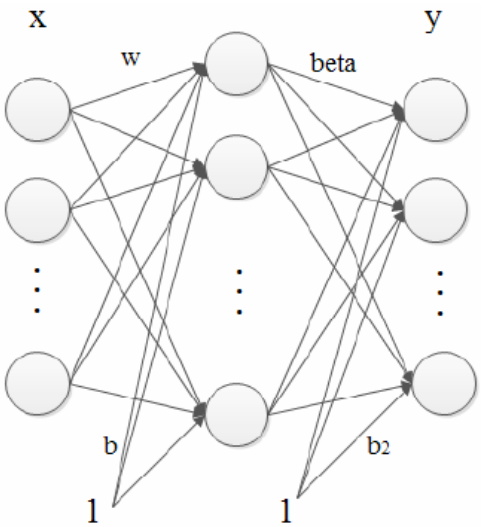
ELM is developed from single-hidden layer feed forward neural networks (SLFNs), it can be used for classification and prediction (Liu et al., 2016b). Here, we use it to study the relationship between $P_{All}$ and $\sum_{i=0}^{n} P_i$.

For N arbitrary samples $(x_i, t_i)$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n$ and $t_i = [t_{i1}, t_{i2}, \ldots t_{in}]^T \in R^m$, SLFNs with N hidden nodes and activation function g(x) can be mathematically expressed as

$$y = g(\mathbf{w} \cdot \mathbf{x} + b) \tag{2}$$

where **x** is the input vector, y is the output of the neuron, $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ is the input weight vector and b is the bias. Activation function g(x) usually can be traingd(·), tansig(·) or sigmoid(·), etc. $\mathbf{w} \cdot \mathbf{x}$ is the inner product of w and x. We can see that a neuron's function is to get the inner product of input vector and input weight vector, then substitute in that activation function the sum of inner product and bias, to get an output response. The schematic diagram of SLFNs is shown as Figure 3. The whole neural network consists of three layer neuron, they are input layer which gets the external information, hidden layer and output layer which feeds back the information to the outside. Among the three layers, the former is the input of the latter.

**Figure 3** Schematic diagram of neuron




In standard SLFNs, the response of the $k^{th}$ neuron of the output layer can be expressed as

$$y[k] = \left[ g(w_1 \cdot x + b_1) g(w_2 \cdot x + b_2) \cdots g(w_L \cdot x + b_L) \right] \cdot \beta_k + b_2[k], k = 1, ..., m$$

where $\beta_k \in R^L$, k = 1, …, m is the weight vector of the $k^{th}$ of neuron output layer. The whole SLFNs can be compact represented as

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \beta + \mathbf{b}_s \tag{3}$$

where **y** is the output vector;

$$h(x) = \left[ g(\mathbf{w}_1 \cdot \mathbf{x} + b_1)\, g(\mathbf{w}_2 \cdot \mathbf{x} + b_2) \cdots g(\mathbf{w}_L \cdot \mathbf{x} + b_L) \right]$$

is the response of all the input vector x, it is called response vector of hidden layer. $\beta = [\beta_1^T\ \beta_2^T \ldots \beta_m^T]^T$ is the combination of output vectors, it is called output weight vector. $b_i$ is bias of the $i^{th}$ neuron of the hidden layer.

What we care is the study ability of the model, that is to say, in function space F, according to the training data $\{x_j, t_j\}_{j=1}^N$, find the $f^* \in F$ that make the loss function C least. The loss function C describe the difference between model f(x) and observations t, the loss function based on the Euclidean distance is expressed as following

$$\{x_j, t_j\}_{j=1}^N, \tag{4}$$

where E is the expectation, D is the sample space, $\|\cdot\|^2$ is the 2 norm operation of the vector.

To make the loss function C least, we used to use back propagation (BP) algorithm, but it is overhead consuming and it will easily sink into local optimum.

ELM proposed and proved two theories.

*Theory 1:* Given a standard SLFNs, and a group of training data $\{\mathbf{x}_j, \mathbf{t}_j\}_{j=1}^N$, $\mathbf{x}_j \in \mathbf{R}^n$, $\mathbf{t}_j \in \mathbf{R}^m$, if the activation function g:R → R is differentiable in any interval, then to any interval in $\mathbf{R}^n$ and R, according to $w_i$ and $b_i$ that generated by any continuous probability distribution, there are:

1 hidden layer response vector H is reversible with probability one

2 $\|T - H\beta\|_F = 0$ is satisfied with probability one.

*Theory 2:* Given any small positive number $\varepsilon > 0$, a standard SLFNs and a group of training data $\{\mathbf{x}_j, \mathbf{t}_j\}_{j=1}^N$, $\mathbf{x}_j \in \mathbf{R}^n$, $\mathbf{t}_j \in \mathbf{R}^m$, if the activation function g: R →R is differentiable in any interval, according to $w_i$ and $b_i$ that generated by any continuous probability distribution, there exists $L \le N$ to make $\|\mathbf{T} - \mathbf{H}\beta\|_F \le \varepsilon$ established with probability one.

With the two theories, what we need to obtain is only the output weight β. To make the loss function C least, then the β is

$$\hat{\beta} = \underset{\beta}{\arg\min} \| \mathbf{T} - \mathbf{H}\beta \|_F \tag{5}$$

Then we can get the whole ELM. This method is fast, its generalisation ability is better than BP algorithm, and it avoids the problem that caused by the local optimum. So that we can train mass data quickly and the consequence will be applicable for other situations properly.

After getting the relationship between power bus and sum of the appliances' power, i.e., the value of ε. If the ε deviates from the calculated value within a threshold, then we declare that in this situation the data is correct and true. Even there are some observations obviously deviate from their so-called regular patterns, the data is real and true. Because we get the values of power bus and various domestic appliances, according to (1), they can verify each other. As a result, this could identify the outliers more precisely.

### *4.3 Fill missing data*

As mentioned earlier, corrupted data refers to missing data and it is caused by the communication interruptions in home environment. And we consider that the data from the power bus is real and correct because of the system layout, it overcomes the problem that the short communication distance of Zigbee brings.

If only one appliance misses its data, we can easily use (1) to fill the data.

$$P_k = P_{All} - \left( \sum_{i=1}^{k-1} P_i + \sum_{i=k+}^{N} P_i + \varepsilon \right) \tag{6}$$

However, there are more than one appliance lost their data sometimes, interpolation techniques are used to help fill it. Interpolation techniques utilise the existing observed points to generate a new interpolation function, and this function will be carried out to get the interpolation to fill the missing data.

#### *4.3.1 Lagrange interpolation*

$x_0, x_1, \ldots, x_n$ are the n + 1 nodes of the time series, utilise them to construct the interpolation polynomial $L_n(x)$. Assume that they meet the following condition:

$$L_n(x_j) = y_j \ (j = 0, 1, ..., n) \tag{7}$$

where $y_j$ is the real value. To construct $L_n(x)$, define the interpolation basis function as following:

*Definition 2:* Given a polynomial of degree n, $L_j(x)$ (j = 0, 1, …, n), it satisfies the following conditions on n + 1 observations $x_0 < x_1 < \cdots < x_n$:

$$L_n(x_j) = \begin{cases} 1, & k = j; \\ 0, & k \ne j. \end{cases} (j, k = 0, 1, ..., n) \tag{8}$$

We call these n + 1 $L_0(x), L_1(x), \ldots, L_n(x)$ polynomials of degree n are the interpolation basis functions of the nodes $x_0, x_1, \ldots, x_n$. The n degree interpolation basis function is:

$$L_k(x) = \frac{(x - x_0)...(x - x_{k-1})(x - x_{k+1})...(x - x_n)}{(x_k - x_0)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_n)} \tag{9}$$
$$(k = 0, 1, ..., n)$$

it obviously satisfies (8). Then the interpolation polynomial $L_n(x)$ which meets the condition (7) could be expressed as following:

$$L_n(x) = \sum_{k=0}^{n} y_k l_k(x) \tag{10}$$

the polynomial like $L_n(x)$ called Lagrange interpolation. We can get the interpolations using the above mentioned method.

### *4.3.2 b-spline interpolation*

Given the control nodes $P_0$, $P_1$, …, $P_n$,so that the b-spline curve of degree k ((k – 1)$^{th}$ power) can be expressed as:

$$P(t) = \sum_{i=0}^{n} P_i N_{i,k}(t) \tag{11}$$

where $N_{i,k}(t)$ are the basic functions of the b-spline curve of degree k, every one of them is called a b-spline.

The b-spline basic function $N_{i,k}(t)$ is k order, it can be expressed as following:

$$N_{i,k}(t) = \frac{1}{k!}\sum_{j=0}^{k-i}(-1)^k \cdot C_{k+1}^{j} \cdot (t+k-i-j)^k \tag{12}$$
$$(0 \le t \le 1, i = 0, 1, 2, ..., k)$$

de Boor Cox's recursive definition:

$$N_{i,0}(t) = \begin{cases} 1, & t_i \le t < t_{i+1}; \\ 0, & t < t_i \text{ or } t \ge t_{i+1}, \end{cases}$$
$$N_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i} N_{i,k-1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} N_{i+1,k-1}(t), k > 0 \tag{13}$$
$$\text{assume that}: \frac{0}{0} = 0$$

The recurrence formula shows that: the i$^{th}$ b-spline $N_{i,k}(t)$ of degree k is obtained depend on $t_i$, $t_{i+1}$, …, $t_{i+k+2}$ nodes.

### *4.3.3 Interpolation fusion method LBboosting*

Here, we propose a new interpolation fusion method called LBboosting to fill the missing data, in the case of more than one appliance losing their information. As we know, modern domestic appliances have not only one mode, some of them have several modes, e.g., a fan, some of them have varied modes, e.g., a computer. So when there is missing data, it is difficult to judge which mode it is, not to mention to fill missing data. The method is based on the system layout, Lagrange interpolation and the b-spline interpolation. As the interpolation technique just simulates the data's trend roughly, it cannot recover the data precisely. The fitting curves produced by the two interpolation techniques have two situations. First, two interpolation methods' outcomes are larger or smaller than the real value, and one of them is closer, as Figure 4 shows. Second, one of them is larger than the real value, and the other is smaller than the real value, as Figure 5 shows.

As mentioned above, we propose a new interpolation fusion method to fill the missing data when there is more than one appliance loses data, the pseudo code of the algorithm is shown in Table 1.

For simplicity, here we consider just two appliances have lost data. The idea is the same when more appliances lose data. $D_0$, $D_1$, …, $D_n$ and $E_0$, $E_1$, …, $E_n$ are the observations of the appliances which have missing data, $P_{l1}s$ and $P_{l2}s$ are the outcomes of the Lagrange interpolation, and $P_{b1}s$ and $P_{b2}s$ are the outcomes of the b-spline interpolation, $P_{known}$ is the sum of the appliances' power which have no missing data.

**Figure 4** First situation (see online version for colours)

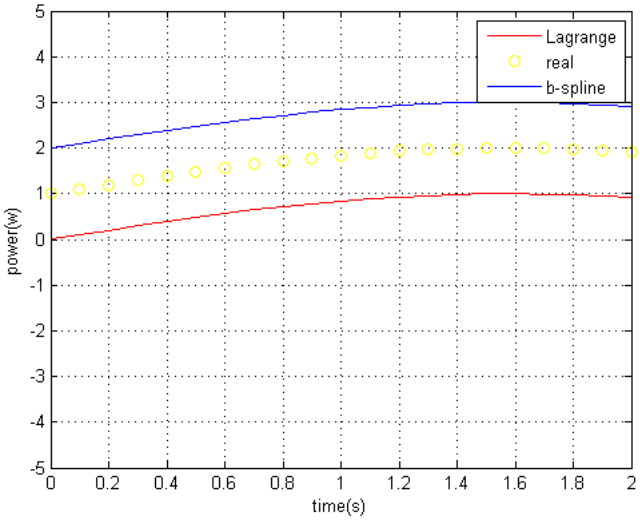



**Figure 5** Second situation (see online version for colours)

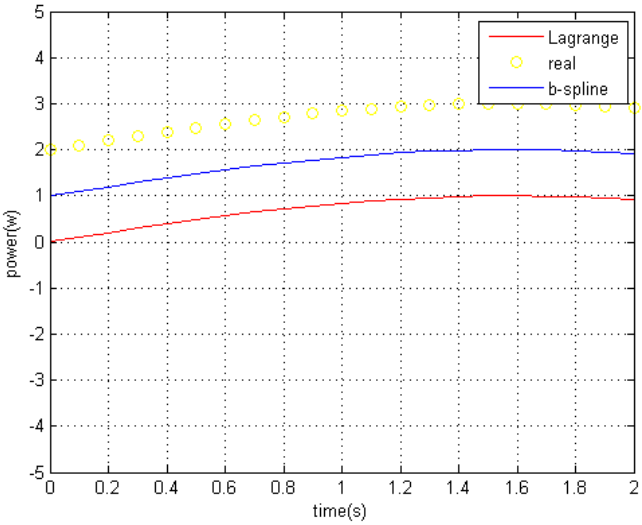



**Table 1** Interpolation fusion method

| | |
|---|---|
| | Algorithm LBboosting ($D_0$, $D_1$, …, $D_n$; $E_0$, $E_1$, …, $E_n$) |
| 1 | $P_{l1s}$ = Lagrange interpolation ($D_0$, $D_1$, …, $D_n$); |
| 2 | $P_{l2s}$ = Lagrange interpolation ($E_0$, $E_1$, …, $E_n$); |
| 3 | $P_{b1s}$ = b-spline interpolation ($D_0$, $D_1$, …, $D_n$); |
| 4 | $P_{b2s}$ = b-spline interpolation ($E_0$, $E_1$, …, $E_n$); |
| 5 | compare $P_{l1}s + P_{l2}s$ with $P_{all} - P_{known} - \varepsilon$; |
| 6 | compare $P_{b1}s + P_{b2}s$ with $P_{all} - P_{known} - \varepsilon$; |
| 7 | compare $\frac{(P_{l1}s + P_{b1}s)}{2} + \frac{(P_{l2}s + P_{b2}s)}{2}$ with $P_{all} - P_{known} - \varepsilon$; |
| 8 | take the value which is more closer to $P_{all} - P_{known} - \varepsilon$; |

## 5 Implementation and performance

In this section, we implement our electric load data verification and repair method in home environment.

### *5.1 Data verification*

In this paper, we used data collected from a lab that models a normal family for one year, and assume that the appliances work as the real situation, i.e., working appliances' number and status are collected just like real situation, so that the ε mentioned before can be figured out using ELM, as Figure 2 shows.

#### *5.1.1 Getting ε by ELM*

After getting the training data, we could find the relationship between $P_{All}$ and $\sum_{i=0}^{n} P_i$, i.e. the value of ε. The precision of the consequence is around 94.4%. In this way, we thought that the ε can be obtained accurately.

As domestic appliances work stably, the error ε is steady, and we can get it in a real-time. Then we get rid of the outliers in training data, and we use the most deviate observation as the bound value, if an observation deviate the average value more than the bound one, this will be thought as an outlier. After training, here we took 100 testing observations randomly. These observations were observed real and correct in advance. Table 2 shows the consequence.

**Table 2** Data verification and outlier detect

| *Mostly deviate from ε* | *Number of test observations* | *Observations that ≥ ε* | *Observations that ≤ ε* | *Accuracy* |
|---|---|---|---|---|
| 9.32% | 100 | 0 | 100 | 100% |

As Table 2 shows, our methods could certificate the data's quality.

### *5.2 Only one appliance missed data*

When only one appliance loses its data, we can use (1) to fix the data and get the following formula:

$$P_{loss} = P_{all} - \sum P_{unloss} - \varepsilon$$

where $P_{all}$ is the value of the power bus, $P_{unloss}$ is the value of every value of the appliance which has no missing data, ε is the error which we got it using ELM.

B-spline interpolation is the up-to-date method in this field, the performance of the b-spline and our proposed method, as Table 3 shows.

First, we made the oven data missed, and the oven usually works steadily. As can be seen in Table 3, when loss rate is low, b-spline which is the up-to-date method in outlier repair is better than proposed method, because the appliance is an oven, its power is relatively steady. So the repair difficulty is low, most of traditional methods can get good results, and proposed method is effected by the precision of *ε*.

**Table 3** Repair accuracy when only one appliance(oven) misses data

| *Number of missing data appliance* | *Loss percent of the data* | *b-spline* | *Proposed method* |
|---|---|---|---|
| 1 | 20% | 95.79% | 91.32% |
| 1 | 50% | 79.37% | 92.47% |
| 1 | 80% | ---- | 89.42% |

To make repair difficulty harder, we intentionally lost some data of refrigerator which has more modes and changes modes continuously. The performance is shown in Table 4.

**Table 4** Repair accuracy when only one appliance (refrigerator) misses data

| *Number of missing data appliance* | *Loss percent of the data* | *b-spline* | *Proposed method* |
|---|---|---|---|
| 1 | 20% | 86.21% | 93.82% |
| 1 | 50% | 54.07% | 91.50% |
| 1 | 80% | ---- | 88.62% |

Compared to the oven missing data repair, obviously the performance of proposed method is better than b-spline interpolation. So our method is robust to the effect of original data quality.

It also can be seen that traditional method is affected by the loss rate of the data, because its principle is depended on the original data, when the loss percent of the data is larger, the repair accuracy is lower. Proposed method's performance is basically better than the b-spline, and it is robust to the loss rate of data. Also, testing data did not take too many actual accidents into the experiment, so the b-spline has a relatively high repair accuracy. When the loss rate is too high, such as 80% in our experiment, b-spline cannot handle the problem under this circumstance, but proposed method worked as well.

Proposed method can even handle the condition that one appliance totally loses its data. As the data's loss rate increases, b-spline and other traditional methods will not work anymore.

### *5.3 More than one appliance missed data*

When more than one appliance loses data, the method for one appliance losing data is not worked. So here we need two interpolation methods and our proposed system to construct the interpolation fusion method to repair the missing data, we called it LBboosting. Table 1 shows how the proposed algorithm works. The LBboosting can handle the situation that more than one appliance missed its data, performance comparison between LBboosting and up-to-date method is shown in Table 5.

Here, we consider only two appliances lose their data, the solution of more appliances losing data is the same. LBboosting fused two interpolation methods, and found the better one, it seems to be just making a relatively little improvement as to b-spline method. But the method would not need training process which traditional methods all

need, and it can work in a real-time with higher accuracy. With the proposed system architecture, we can even take more interpolation methods together to make the performance better.

**Table 5** Repair accuracy when more than one appliance miss data

| *Number of missing data appliance* | *Loss percent of the data* | *b-spline* | *Proposed method* |
|---|---|---|---|
| 2 | 20% | 95.92% | 97.21% |
| 2 | 50% | 73.64% | 78.88% |

## 6 Conclusions and future work

This paper proposed a new data verification and repair method in home environment, it is called a system of 'N + 1' framework. The method considered the real condition that happens in home, and it collected the information of all appliances and the power bus, to make them verify each other. It also can handle the randomness problem which has a big effect on the repair accuracy. Based on system architecture, we proposed a method called LBboosting to fill the missing data. As a result, the system could verify the data quality precisely, and it can repair the missing data better than the up-to-date method. Its advantage is to solve the traditional methods' inherent drawback that they assume the data follows a certain pattern or period. Traditional methods also need a long training time and must learn from representative data, or the performance will be bad. Proposed method does not need a training process and it is robust to the loss percent of the original data.

However, the method is mainly affected by the error, we still need to improve the accuracy of it. The method just considered the missing data caused by the communication failure, and did not take the wrong data caused by the meter malfunction or other factors into consideration. In the future work, we must consider more factors and the repair of wrong data.

## Acknowledgements

This work is supported by the NSFC (61300238, 61300237, 61232016, 1405254, 61373133), Marie Curie Fellowship (701697-CAR-MSCA-IFEF-ST), Basic Research Programs (Natural Science Foundation) of Jiangsu Province (BK20131004) and the PAPD fund.

## References

Abraham, B. and Chuang, A. (1989) 'Outlier detection and time series modeling', *Technometrics*, Vol. 31, No. 2, pp.241–248.

Chen, J., Li, W., Lau, A., Cao, J. and Wang, K. (2010) 'Automated load curve data cleansing in power systems', *IEEE Trans. Smart Grid*, Vol. 1, No. 2, pp.213–221.

Chen, S-Y., Song, S-F., Li, L. and Shen, J. (2009) 'Survey on smart grid technology,' *Power Syst. Technol*., Vol. 33, No. 8, pp.1–7.

Gu, B., Sheng, V.S., Tay, K.Y., Romano, W. and Li, S. (2015a) 'Incremental support vector learning for ordinal regression', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 7, pp.1403–1416.

Gu, B., Sheng, V.S., Wang, Z., Ho, D., Osman, S. and Li, S. (2015b) 'Incremental learning for ν-support vector regression', *Neural Networks*, Vol. 67, No. C, pp.140–150.

Guo, Z., Li, W., Lau, A., Inga-Rojas, T. and Wang, K. (2012) 'Detecting X-outliers in load curve data in power systems', *IEEE Transactions on Power Systems*, Vol. 27, No. 2, pp.875–884.

Han, J., Yun, J., Jang, J. and Park, K-r. (2010) 'User-friendly home automation based on 3D virtual world', *IEEE Transactions on Consumer Electronics*, Vol. 56, No. 3, pp.1843–1847.

Hawkins, S., He, H., Williams, G. and Baxter, R. (2002) 'Outlier detection using replicator neural networks', in *Proceedings of the 4th international conference on Data Warehousing Knowledge Discovery*, pp.170–180.

Hodge, V.J. and Austin, J. (2004) 'A survey of outlier detection methodologies', *Artificial Intelligence Review*, Vol. 22, No. 2, pp.5–126.

Liu, D. and Xiao, P. (2016) 'An energy-efficient adaptive resource provision framework for cloud platforms', *International Journal of Computational Science and Engineering*, Vol. 13, No. 4, pp.346–354.

Liu, Q., Cai, W., Jin, D., Shen, J., Fu, Z., Liu, X. and Linge, N. (2016a) 'Estimation accuracy on execution time of run-time tasks in a heterogeneous distributed environment', *Sensors*, Vol. 16, No. 9, p.1386.

Liu, Q., Cai, W., Shen, J., Fu, Z., Liu, X. and Linge, N. (2016b) 'A speculative approach to spatial-temporal efficiency with multi-objective optimization in a heterogeneous cloud environment', *Security Comm. Networks*, Vol. 9, No. 17, pp.4002–4012.

Ljung, G.M. (1993) 'On outlier detection in time series', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 55, No. 2, pp.559–567.

Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P. and Tarassenko, L. (1999) 'A system for the analysis of jet engine vibration data', *Integrated Computer-aided Engineering*, Vol. 6, No. 1, pp.53–66.

Ramaswamy, S., Rastogi, R. and Shim, K. (2000) 'Efficient algorithms for mining outliers from large data sets', *ACM Sigmod Record*, Vol. 29, No. 2, pp.427–438.

Tang, G., Wu, K., Lei, J., Bi, Z. and Tang, J. (2014) 'From landscape to portrait: a new approach for outlier detection in load curve data', *IEEE Transactions on Smart Grid*, Vol. 5, No. 4, pp.1764–1773.

Wen, X., Shao, L., Xue, Y. and Fang, W. (2015) 'A rapid learning algorithm for vehicle classification', *Information Sciences*, Vol. 295, No. 1, pp.395–406.

Weron, R. (2006) *Modeling and Forecasting Electricity Loads and Prices – A Statistical Approach*, Wiley, New York.

Xia, Z., Wang, X., Sun, X. and Wang, Q. (2015) 'A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No. 2, pp.340–352.

Yang, C., Sheng, M., Li, J., Li, H. and Li, J. (2013) 'Energy-aware joint power and rate control in overlay cognitive radio networks: a Nash bargaining perspective', *International Journal of Embedded Systems*, Vol. 5, No. 3, pp.118–126.

Zheng, Y., Jeon, B., Xu, D., Wu, Q.M.J. and Zhang, H. (2015) 'Image segmentation by generalized hierarchical fuzzy C-means algorithm', *Journal of Intelligent and Fuzzy Systems*, Vol. 28, No. 2, pp.961–973.

Zhou, J., Khawaja, M.A., Li, Z., Sun, J., Wang, Y. and Chen, F. (2016) 'Making machine learning useable by revealing internal states update – a transparent approach', *International Journal of Computational Science and Engineering*, Vol. 13, No. 4, pp.378–389.