 CMC, vol.57, no.2, pp.179-194, 2018

A Dual-spline Approach to Load Error Repair in a HEMS Sensor Network

Xiaodong Liu[1] and Qi Liu[1, *]

Abstract: In a home energy management system (HEMS), appliances are becoming diversified and intelligent, so that certain simple maintenance work can be completed by appliances themselves. During the measurement, collection and transmission of electricity load data in a HEMS sensor network, however, problems can be caused on the data due to faulty sensing processes and/or lost links, etc. In order to ensure the quality of retrieved load data, different solutions have been presented, but suffered from low recognition rates and high complexity. In this paper, a validation and repair method is presented to detect potential failures and errors in a domestic energy management system, which can then recover determined load errors and losses. A Kernel Extreme Learning Machine (K-ELM) based model has been employed with a Radial Basis Function (RBF) and optimised parameters for verification and recognition; whilst a Dual-spline method is presented to repair missing load data. According to the experiment results, the method outperforms the traditional B-spline and Cubic-spline methods and can effectively deal with unexpected data losses and errors under variant loss rates in a practical home environment.

Keywords: Electric load data analysis, home energy management, quality assurance and control.

1 Introduction

In recent years, Home Energy Management Systems (HEMS) become increasingly popular due to its features of cross-disciplinary, compatibility and interoperable [Lee, Hsiao, Huang et al. (2016)]. It can be well integrated with Smart Home [Shah, Khalid, Zafar et al. (2017)], Industrial Automation [Lin and Tsai (2014)] and Smart Grid [Lin and Tsai (2014)]. Investigation is also widely made with Cyber Physical Systems [Cintuglu, Mohammed, Akkaya et al. (2017)], Cloud Computing [Sanislav, Zeadally and Mois (2017)] and Big Data Analytics [Sheng, Zhao, Zhang et al. (2016)].

Monitoring and management of electric power consumption is one of the most concerned scenarios for demand side management of Smart Grid [Kumar, Singh, Zeadally et al. (2017); Kumar, Zeadally and Misra (2016)], where the collection, process and analysis of power load data are not only conducive to the industry, but to the end-consumers as well, e.g. in the cases of smart home, home energy management systems, etc. In a domestic

[1] School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh EH10 5DT, UK.

[*] Corresponding Author: Qi Liu. Email: q.liu@napier.ac.uk.

environment, however, present home energy systems suffer from accuracy and stability of collected data, especially in those wireless solutions. According Aloulou et al. [Aloulou, Mokhtari, Tiberghien et al. (2015)], certain factors in the process of data perception and transmission may cause data missing and/or errors, including the failure of meters, battery shortage, communication failure and other reasons.

At present, different methods have been proposed to perform outlier detect and load data repair [Liang, Zhao, Luo et al. (2017)], but few of them focus on load error repair in a domestic environment, where solving missing and/or error data meets following challenges. First, there are no effective solutions when a large portion of data is missing. In addition, it is difficult to know in a domestic area whether relatively large deviation from a gathering point (e.g. a socket) reports an error or an unusual situation (e.g. another appliance being plugged). Finally, the uncertainty of retrieved data caused by changing behaviours due to energy consumption from different occupants, or even from the same one who has different regularity under unexpected circumstances.

In this paper, a method is proposed to detect and fix potential errors of collected load data in a domestic environment. A data verification method to eliminate the hardware failure is firstly presented to ensure no failure data being collected. It then examines the periodicity of data and detects potential errors by focusing on the uncertainty of energy consumption activities with robustness to data loss rates. A Dual-spline method is further designed to fix the missing data in a domestic energy management system.

The remainder of this paper is organized into five sections. Related work on error detection and repair is reviewed in Section II. In Section III, problem definition and preliminaries of core algorithms manipulated in our approach are introduced. Section IV explains the proposed method to achieve detection and repair of missing/error load data in a domestic environment. Results are presented and evaluated in Section V with comparison of corresponding algorithms. Finally, Section VI concludes and identifies future work.

2 Related works

Related work has been performed in the area of anomaly detection, which was extensively investigated using data mining and statistics models. Dean and Dixon proposed a Dixon's Q test, using differences between observation and neighbours divided by a data range as Q values [Dean and Dixon (1951)]. It was then compared with corresponding sampling space with a confidence threshold to detect abnormal values. However, the method can only detect exceptions in a small data set. Grubbs [Grubbs (1969)] examined the differences between attribute means and observation by the standard deviation of attributes as *Z*-values, which was then compared with a 1% to 5% significance level. This method does not require any input parameters but is sensitive to the number of samples. The higher number of samples it has, the more representative. In Chandola et al. [Chandola, Banerjee and Kumar (2009)], a box-plot method was reviewed by adding and subtracting 1.5 times of upper and lower quartiles respectively as confidence intervals. Data beyond the scope were regarded as abnormal ones. In Weron [Weron (2006)], it assumed that a sequence period was given, so mean and median values at the corresponding time points in another period can be used to replace errors or lost

values. The mean values, however, were easily affected by extreme values; whereas the median was not representative for the data. A concept of portrait dataset was proposed, segmenting datasets into a portrait dataset in Tang et al. [Tang, Wu, Lei et al. (2014)], where it assumed that the examined data followed probability distribution, which can be used to estimate and replace outliers. All above statistical methods have tried to find certain distribution in a dataset for outlier detection and further repair.

Time-series relative algorithms, e.g. an Auto-Regressive Moving Average (ARMA) model were also used for error repair. For example, outliers were simulated in a time series, so approximate distribution of the sequence was taken into account with statistical variables to detect the outliers [Ro, Zou, Wang et al. (2015)]. In Ljung [Ljung (1993)], the estimation and detection of outliers in a time series were proposed by Gaussian ARMA processes, which indicated that cumulative outliers were directly related to missing observations and could be detected using likelihood ratios and anomalous scores. However, ARMA-based models assume fixed time series, which is not suitable for domestic load data. Furthermore, historical data is used in an ARMA model for the prediction of future retrieved data along its time series; however, a data repair approach looks forward to discovery and correction of potential errors in a historical dataset. The fact that predictive models assume historical data are correct makes them inapplicable to error detection and fixing in a practical scenario. In addition, traditional statistical methods and ARMA-based models require large-scale reference data for better performance, which makes them not applicable to those scenarios where a large percentage of data is missing.

Data mining algorithms were also investigated for data repair, where the attribute of distance has been mainly used for outlier detection. In Ramaswamy et al. [Ramaswamy, Rastogi and Shim (2000)], a K-Nearest Neighbours (KNN) algorithm was employed to detect outliers, where an outlier could be recognised if n-1 points were closer to a pre-set point m. However, the method suffers high time complexity, resulting in poor performance when the number of neighbours increases. A *K*-means based method was specified in Aggarwal [Aggarwal (2015)] to classify data into a number of groups and iterate them to calculate distances between the centre and data points of each group. All points that are outside pre-defined clusters were then recognised as errors. Compared with the KNN solution, the *K*-means based work reduces the volume by dividing data into several clusters, and therefore effectively downgrades its computational complexity. A weighted KNN-based method was presented in Hilla et al. [Hilla and Minsker (2010)], which calculated weighted distances to identify outliers. An RNN model was proposed to calculate the outliers of data points [Hawkins, He, Williams et al. (2002)]. More usage of such methods can be found in Salvador et al. [Salvador and Chan (2005); Jones, Nikovski, Imamura et al. (2016); Aghabozorgi, Shirkhorshidi and Wah (2015); Gupta, Gao, Aggarwal et al. (2014)]. These methods calculated abnormal scores according to the distances between structured relational data to determine potential outliers, which are not applicable to practical domestic energy management cases.

Recent relevant research work has been undertaken on smoothing techniques. A nonparametric regression method was proposed in Chen et al. [Chen, Li, Lau et al. (2010)] based on kernel smoothing and B-spline smoothing algorithms, where an outlier was

determined depending on confidence intervals of corresponding perception. In Guo et al. [Guo, Li, Lau et al. (2012)], a periodicity function was pre-defined to determine potential errors. It however requires prior knowledge of data periodicity and its period length, which is complicated to implement in a domestic HEMS environment.

All solutions examined above tries to find certain regularity in their retrieved data, whereas in a domestic scenario, randomness is an inevitable feature due to complexity of gathering processes of home appliances and diversity of human activities. A Dual-spline approach is therefore presented in this paper to adapt the uncertainty, which is detailed in the following sections.

3 Preliminaries

3.1 Definition of load errors

In a practical domestic energy management environment, energy consumption is collected at a given frequency, such as once a second or once a minute. Retrieved load data include energy usage of entire home and each appliance, which therefore implies consumers' daily behaviour and habits. The quality of the retrieved data is influenced from multiple factors, such as meter malfunctions, communication failures, and/or even behaviour changes.

Hardware malfunctions in a HEMS system happen, but within a reasonable range. Smart sockets used in this paper, for instance, fulfil the requirements of 50/60 Hz IEC 687/1036 standard, which have their error rate within 0.2%.

In addition, energy data in this paper have been perceived and collected via a wireless sensor network, which suffer from data missing due to communication failures.

In terms of behaviour changes of consumers, an adaptive method is needed to gratify the diversity and uncertainty. For example, a household member always goes home for dinner, but decides to have dinner outside tonight. Such a scenario may cause an outlier using traditional methods, whereas so-called casual patterns happen quite often in a modern family.

Consideration above implies following definition on load data.

Definition 1: Load data can be time-series and denoted as $S = \{(y_i, t_i)\}_{i=1}^{n}$, that is an n-value sequence ordered by time, where t_i represents the i^{th} timestamp, and y_i is the perceived value at t_i.

3.2 A Kernel-based extreme learning machine model

The Extreme Learning Machine (ELM) algorithm was presented as a Single-hidden Layer Feedforward Neural network (SLFN) for faster training speed maintaining with high accuracy [Huang, Zhou, Ding et al. (2012)]. A typical L hidden neurons ELM model can be depicted in Eq. (1).

$$\sum_{j=1}^{L} \beta_j \cdot g\left(\langle \omega_j, x_i \rangle + b_j\right) = y_i \tag{1}$$

where ω_j and β_j are weight values in between input, hidden and output layers, and $g()$

represents the outputs of the j^{th} hidden neuron with b_j as their corresponding bias. In addition, $X = \{x_1, x_2, \ldots, x_N | x_i \in R^D, i = 1,2, \ldots, N\}$ denotes training data with N samples and D dimensions.

Kernel-based Extreme Learning Machine (K-ELM) was presented for the improvement of stability and generalisation performance over the ELM model [Lam and Wunsch (2017)]. A K-ELM algorithm requires no configuration on the number of neurons and the types of activation function but needs to provide a kernel function. In this paper, a Radial Basis Function (RBF) has been employed to conduct the verification of potential load errors, as shown in Eq. (2).

$$K(x_i, y_j) = exp\left(-\sigma \|x_i - x_j\|^2\right), \sigma > 0 \tag{2}$$

4 The proposed dual-spline approach

A Dual-spline solution is proposed in this section to adapt randomness caused by domestic users and to achieve accurate detection and error repair of load data retrieved in a domestic energy management environment.

4.1 System construction

The domestic energy management system established in this paper consists of a collecting gateway for remote communication with servers and local collection of load data, a CT clamp to gather total energy consumption, and multiple smart sockets to monitor the consumption of appliances. The entire system structure deployed in a domestic environment is depicted in Fig. 1. WiFi is employed to enable a star-topology wireless communication between the gateway and appliances.

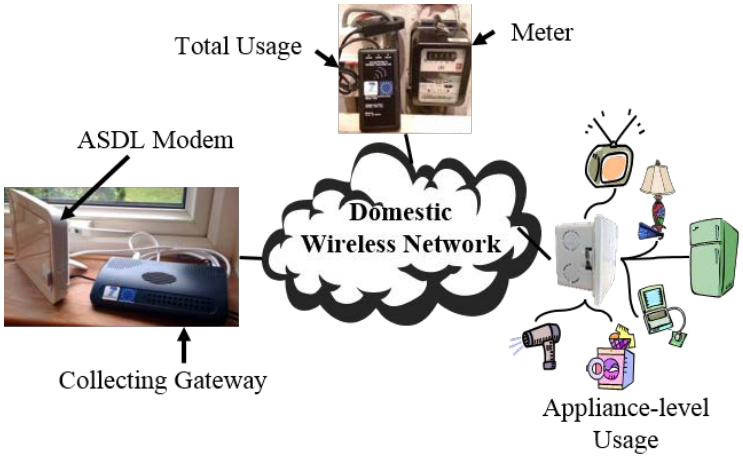


Figure 1: Domestic energy management system

4.2 Analysis of load data characteristics

A general domestic power line is a typical parallel circuit. Eq. (3) can therefore be used to describe the relation between the total power consumption of the examined flat and the power of each sub-terminal appliance consumes.

$$P_{all} = \sum_{i=0}^{n} P_i + \varepsilon \tag{3}$$

where P_{all} represents the total power consumption, P_i is the power of appliance *i*, and ε is the error, which usually refers to line losses and white noises. The line losses in a grid can be a large and complex topic to discuss and model; whereas in the home environment proposed in this paper, it is simply defined as the losses of domestic power circuits. Fig. 2 shows the value of ε during different electrical appliances working within the domestic environment.

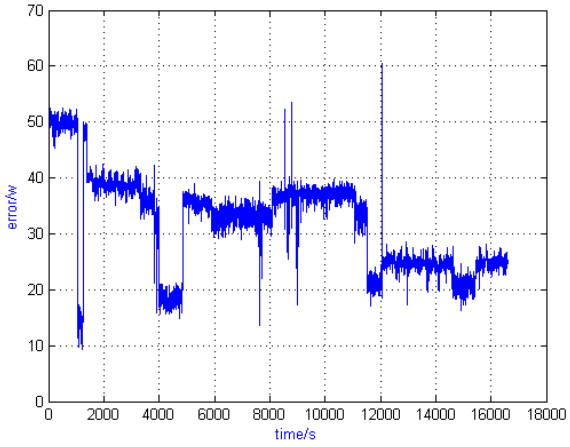


Figure 2: Power line losses in a practical HEMS environment

When all electrical appliances are working properly, the error ε represents the power difference between the total bus power and the consumption of all branches. In this case, major factors affecting ε are run-time loads of working appliances. According to Fig. 2, it can be seen that the error ε keeps stable when working appliances are constant. After sampling, an ELM-based neural network is proposed to learn and predict the trend of ε.

4.3 Verification of load data

After collecting the consumption of the total power and all appliances, according to (1), we can get run-time errors ε_{real}. In addition, predicted errors ε_{pred} can be calculated via the ELM regression. If the difference between a predicted error and the actual one is higher than a predefined threshold θ and occurs 10 times, wrong data could be found. Algorithm 1 gives detailed steps to verify and locate potential outliers.

The verification algorithm above detects passible outliers generated by AC current sensors and/or other integrated hardware at the perception stage. Alternatively, load errors, e.g. energy data losses at the communication stage and other outliers due to behaviour changes of households will be taken into account in next section.

Algorithm 1: Data Validation
Input: Load Data
Output: Outliers
Steps:
1. $count$=0;
2. T'=find(no null data rows in T);
3. While ($t \in T'$)
4. If ($count$>10) return 'ERROR';
5. if ($(\varepsilon^t_{real} - \varepsilon^t_{pred}) > \theta$){
6. $count$++;
7. t++;}
8. else {
9. $count$=0;
10. t++;}
11. end While;

4.4 Repair processes of load errors

After collecting the consumption of the total power and all appliances, according to (1), we can get run-time errors ε_{real}. In addition, predicted errors ε_{pred} can be calculated via the ELM regression. If the difference between a predicted error and the actual one is higher than a predefined threshold θ and occurs 10 times, wrong data could be found. Algorithm 1 gives detailed steps to verify and locate potential outliers.

1) B-spline method

Given a set of load data samples $\{(y_i, t_i)\}_{i=1}^n$, data collection processes can be defined in (4):

$$y_i = m(t_i) + \varepsilon_i \tag{4}$$

where y_i is the data at t_i, ε represents errors. $m(t_i)$ presents the calculated load consumption at t_i. Taking the B-spline smoothing method as an example, it is employed consisting of a set of known basic functions $\{\emptyset_k(t)\}_{k=1}^K$, that are mathematically independent to each other. The main idea is to approximate the function $m(t)$ using a weighted sum or a linear combination of a sufficient number of K basis functions $\emptyset_k(t)$, as shown in (5):

$$m(t) = \sum_{i=1}^K c_k \emptyset_k(t) \tag{5}$$

or in a vector form:

$$m(t) = \vec{c} \cdot \vec{\emptyset}(t) \tag{6}$$

where $\vec{c} = (c_1, \ldots, c_K)$ is a coefficient vector, $\vec{\emptyset}(t) = \overrightarrow{\emptyset_1}(t), \ldots, \overrightarrow{\emptyset_K}(t)$ is the vector of the basis functions. Each function $\emptyset_k(t)$ represents compact support, which ensures that local information is taken into account when estimating the coefficient $\vec{c}$.

In order to estimate $\vec{c}$ from the load samples $\{(y_i, t_i)\}_{i=1}^n$, an $n \cdot K$ matrix is defined in (7):

$$\emptyset = \begin{bmatrix} \emptyset_1(t_1) & \emptyset_2(t_1) \cdots & \emptyset_K(t_1) \\ \vdots & \vdots \quad \ddots & \vdots \\ \emptyset_1(t_n) & \emptyset_2(t_n) \cdots & \emptyset_K(t_n) \end{bmatrix} \tag{7}$$

where $\emptyset[i,j] = \emptyset_j(t_i)$ represents the value of the j^{th} base function at t_i. Then $m(t)$ can be represented as the function $\vec{m} = \vec{c} \cdot \emptyset$ at time $(t_1, \ldots, t_n)$. The coefficients $\vec{c}$ are determined by minimising the SSE to obtain a simple smoothing function:

$$SSE = \sum_{j=1}^{n}\left[y_i - \sum_{k=1}^{K} c_k \emptyset_k(t_j)\right]^2 \quad (8)$$

or in a vector form:

$$SSE = (\vec{y} - \emptyset\vec{c})^T(\vec{y} - \emptyset\vec{c}) \quad (9)$$

where $\vec{y}$ is the vector form of $\{(y_i, t_i)\}_{i=1}^{n}$.

In general, it is necessary to find an estimate function $\hat{m}(t)$ to conduct local averaging procedures, or non-parametric regression. In a domestic case, the curve can be approximated as in (10):

$$\hat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} W_i(t) y_i \quad (10)$$

where $\{W_i(t)y_i\}_{i=1}^{n}$ represents a weight sequence.

2) Cubic-spline method

Given $n+1$ data samples, a Cubic-spline curve $S(x)$ is a sub-defined formula, which satisfies:

a) In every sampling interval $[x_1, x_{i+1}]$, $S(x) = S_i(x)$ is a cubic polynomial;

b) $S(x_i) = y_i$;

c) Derivative $S'(x)$, second derivative $S^{"}(x)$ in the [a, b] interval is continuous; that is, the $S(x)$ curve is smooth.

So n cubic polynomial segments can be defined as in (11):

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (11)$$

3) Dual-spline repair processes

Errors from single appliance

If only one appliance reading is lost, it can be repaired following (12), according to (3):

$$P_k = P_{all} - \left(\sum_{i=1}^{k-1} P_i + \sum_{i=k+1}^{N} P_i + \varepsilon\right) \quad (12)$$

Especially when $P_k < \theta_1$, then $P_k = 0$; that is, detected data loss of an appliance could be caused by line losses, whilst the appliance is possibly switched off.

Errors from multiple appliances

If more than one appliance has data losses detected, a dual-spline repair algorithm is proposed, combining two fitting curve methods to fit load curve data. Two situations need to be considered:

a) Two fitting curves are both located above or below the real value, while one of them is closer to the real one;

b) Two fitting curves are located with the real value in between.

A hybrid curve fitting method called Dual-spline is therefore employed in this paper to repair data losses from multiple appliances. The Dual-spline algorithm pseudo code is described in Algorithm 2.

Algorithm 2: Dual-spline Data Repair

Input: load data $S_j = \{(y_i, t_i)\}_{i=1}^{n}, i \in T, j \in \{1,2,\dots,N+1\}$, missing data by appliance 1: $(D_0, D_1, \dots, D_n)$, missing data by appliance 2: $(E_0, E_1, \dots, E_n)$.

Output: Repaired Data

Steps:

1. P_{a1}=Cubic-spline_Smoothing$(D_0, D_1, \dots, D_n)$;
2. P_{a2}= Cubic-spline_S moothing$(E_0, E_1, \dots, E_n)$;
3. P_{b1} =B-spline_S moothing$(D_0, D_1, \dots, D_n)$;
4. P_{b2}=B-spline_S moothing$(E_0, E_1, \dots, E_n)$;
5. if (($|[P_{a1} + P_{a2} - (P_{all} - P_{known-\varepsilon})]| < \theta$)||($|[P_{b1} + P_{b2} - (P_{all} - P_{kn}$)){
6. then two appliances are ON;
7. return the one in ($\frac{(P_{a1}+P_{b1})}{2} + \frac{(P_{a2}+P_{b2})}{2}$, $P_{a1} + P_{a2}$, $P_{b1} + P_{b2}$) that is closer to $P_{all} - P_{known-\varepsilon}$.}
8. else{
9. there is one or two appliances lost data;
10. if($(P_{all} - P_{known-\varepsilon} - 0) < \theta$){
11. then (two appliances are off);
12. return 'Two appliance are off';}
13. else{
14. then one is ON, the other is OFF;
15. if ($|[P_{b1} - (P_{all} - P_{known-\varepsilon})]| < \theta$) return P_{b1}; else return 0;
16. if ($|[P_{b2} - (P_{all} - P_{known-\varepsilon})]| < \theta$) return P_{b2}; else return 0;}
17. }

In the case of data losses from two appliances, the power consumption of the two appliances and $(P_{all} - P_{known-\varepsilon})$ can be calculated according to (1). Data losses according to the Cubic-spline and B-spline algorithms can be calculated as P_{a1}, P_{a2}, P_{b1}, P_{b2} respectively. If the difference between both curve fitting methods' results and $(P_{all} - P_{known-\varepsilon})$ is within the threshold θ, then the appliances are considered to be ON. According to the two possible situations, data loss can be repaired with the closest value among $\frac{(P_{a1}+P_{b1})}{2} + \frac{(P_{a2}+P_{b2})}{2}$, $P_{a1} + P_{a2}$, and $P_{b1} + P_{b2}$ to $(P_{all} - P_{known-\varepsilon})$. The status with one appliance being switched off can also be examined and repaired.

5 Experiments and performance evaluation

5.1 Construction of experiment environment

A domestic energy management system is established in a real flat with 1 collecting gateway, 1 CT clamp and 10 smart sockets. All load data is gathered once per second. Total load consumption with various combinations of appliances is taken as training data for 1 h.

1) Parameter settings

When training the ELM regression, its kernel function is set to RBF with the kernel parameter being set to 100 and the regression parameter to 1. The fitting performance is shown in Fig. 3.

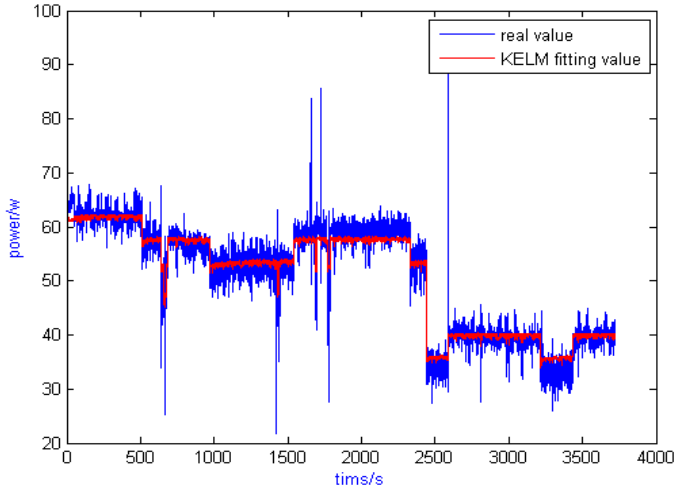


Figure 3: ELM-based fitting values

In the practical system, a desk lamp devotes the minimum power consumption, which varies from 15.7 to 16.7 watts. Line losses, i.e. ε is less than 5 watts according to our test samples. Therefore, taken the line losses in account, data losses from appliances should depict at least 10.7 watts compared to the total energy consumption. θ is therefore set to 10 for both Algorithm 1 and 2.

2) Evaluation criteria

The RMSE and decision coefficient are chosen to evaluate the performance, as shown in (13) and (14).

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(\hat{y}_i - y_i)^2} \tag{13}$$

$$R^2 = \frac{\left(l\sum_{i=1}^{l}\hat{y}_i y_i - \sum_{i=1}^{l}\hat{y}_i \sum_{i=1}^{l} y_i\right)^2}{\left(l\sum_{i=1}^{l}\hat{y}_i{}^2 - \left(\sum_{i=1}^{l}\hat{y}_i\right)^2\right)\left(l\sum_{i=1}^{l}y_i{}^2 - \left(\sum_{i=1}^{l}y_i\right)^2\right)} \tag{14}$$

where l represents the number of testing samples retrieved from the HEMS system; y_i is the true value of the i^{th} sample; $\hat{y}_i$ is the repair value for the i^{th} sample.

5.2 Scenario 1: Data errors on a single appliance

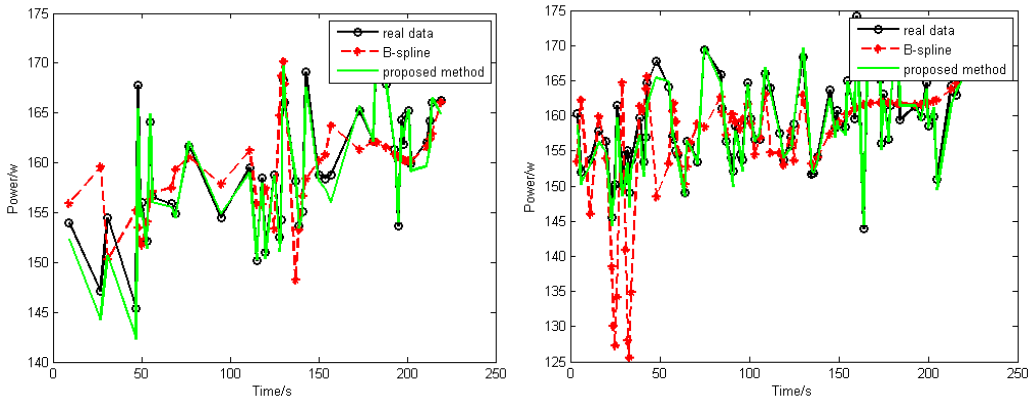
In this scenario, the power of a laptop, i.e. MacBook Pro 13 was collected and randomly removed its load data to repair. The repair accuracy is shown in Tab. 1.

The repair performance at different loss rates is shown in the Fig. 4. According to Tab. 1,

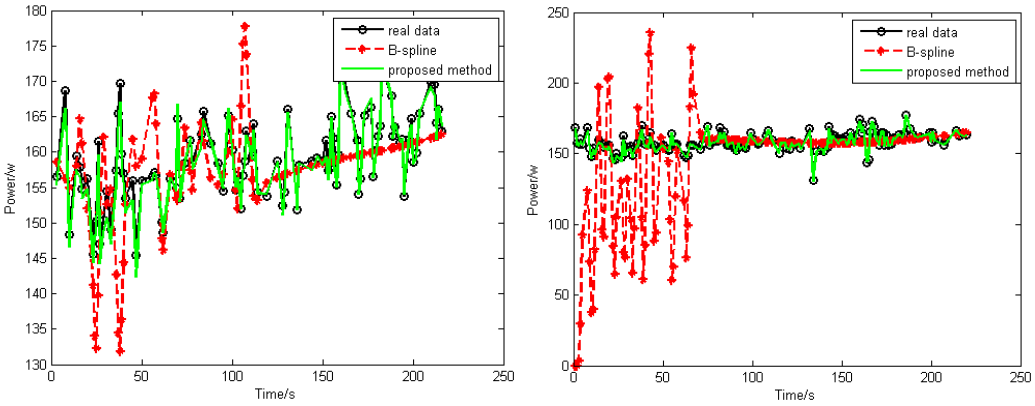
it can be found that the repair accuracy of B-spline drops sharply when the loss rate increases. When the loss rate is at 65%, the B-spline has depicted significant divergences compared to the original load data for the first 70 s. The performance of the Dual-spline method keeps stable and illustrates robust to high data loss rates.

Table 1: Error repair with different loss rates in scenario 1

	B-spline		**Dual-spline Method**	
Loss rate	RMSE	R^2	RMSE	R^2
20%	5.99	0.58	1.33	1.17
35%	8.78	2.46	1.34	1.13
50%	9.07	1.90	1.35	1.12
65%	47.35	43.67	1.08	1.34



(a) 20% (b) 35%



(c) 50% (d) 65%

Figure 4: Performance of dual-spline smoothing repair at different loss rates with single-appliance errors

In addition, when training speed is considered, a K-ELM algorithm with a RBF as its kernel function is employed in this paper, which shows better performance than traditional Back Propagation (BP) and Support Vector Machine (SVM) algorithms, as shown in Tab. 2. B-spline and Cubic-spline are not compared since a smoothing mechanism is employed for the prediction, which is different from the training mechanism used by neural network models.

Table 2: Training time (second) in scenario 1

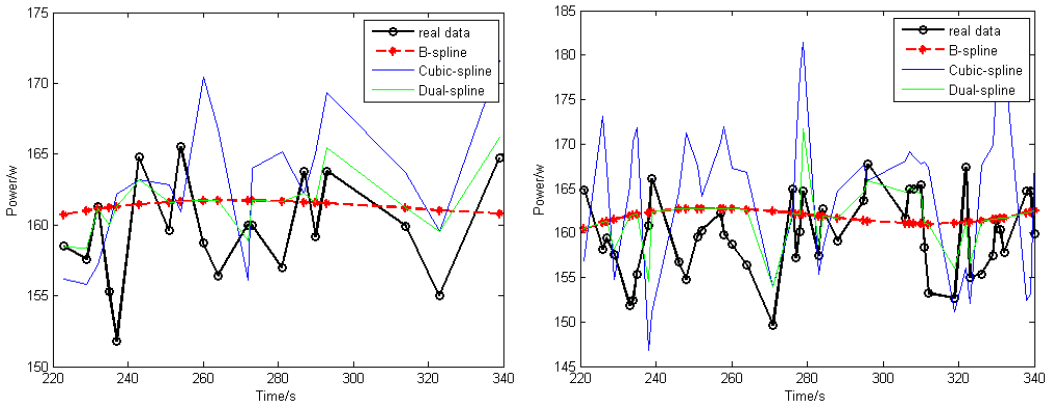
Loss rate	BP	SVM	K-ELM
20%	0.739	0.085	0.065
35%	1.355	0.113	0.093
50%	1.673	0.156	0.103
65%	5.593	0.536	0.114

5.3 Scenario 2: Data errors on multiple appliances

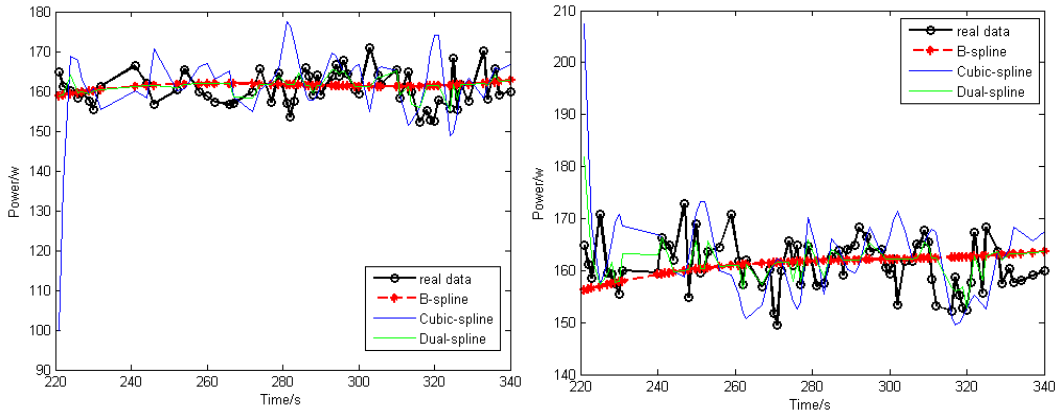
When more than one appliance data is lost, the Dual-spline algorithm facilitates two curve fitting methods to achieve better performance. Tab. 3 shows the performance of the proposed method compared to the Cubic- and B-spline.

Table 3: Error repair with different loss rates in scenario 2

	Cubic-spline		B-spline		Dual-spline	
Loss rate	RMSE	R^2	RMSE	R^2	RMSE	R^2
20%	5.98	2.33	4.02	0.24	3.50	0.54
35%	11.15	5.20	5.03	0.25	4.36	0.65
50%	12.50	6.04	4.45	0.05	0.26	3.91
65%	9.47	3.26	5.40	0.14	4.85	0.61



(a) 20% (b) 35%



(c) 50% (d) 65%

Figure 5: Performance of dual-spline smoothing repair at different loss rates with multiple-appliance errors

As shown in Tab. 3, the Dual-spline is closer to the real value than the B-spline and Cubic-spline. The comparison of repair accuracy is shown in Fig. 5. It can be seen that the Dual-spline is closer to the actual value than the other two smoothing methods.

The training speed is compared with a BP algorithm, as shown in Tab. 4. The SVM algorithm is not implemented in scenario 2 due to complicated configuration of classifiers when multiple appliances are taken into account.

Table 4: Training time (second) in scenario 2

Loss rate	BP	K-ELM
20%	5.733	0.603
35%	6.356	0.819
50%	6.793	0.917
65%	7.39	1.306

6 Conclusion

In this paper, a load data verification and repair method is proposed in a domestic energy management environment. A practical environment has been deployed, where load data of all home appliances and the power bus were collected. Based on the system, a data verification algorithm has firstly been proposed to verify whether an outlier happened or not. A Dual-spline method has then been proposed to repair the missing data. According to the results, potential errors can be verified and found via an improved K-ELM model with shortened training time compared to traditional methods, e.g. the BP and SVM based algorithms. In terms of the accuracy, the Dual-spline method employed in this paper has depicted higher accuracy than B-spline and Cubic-spline techniques. Furthermore, our method shows high robustness to the loss percent of the original data, which helps to be well adapted into a practical domestic energy management environment.

Acknowledgement: This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 701697.

References

Aggarwal, C. C. (2015): *Outlier Analysis*. New York: Springer, pp. 237-263.

Aghabozorgi, S.; Shirkhorshidi, A. S.; Wah, T. Y. (2015): Time-series clustering-A decade review. *Information Systems*, vol. 53, pp. 16-38.

Aloulou, H.; Mokhtari, M.; Tiberghien, T.; Endelin, R.; Biswas, J. (2015): Uncertainty handling in semantic reasoning for accurate context understanding. *Knowledge-Based Systems*, vol. 77, pp. 16-28.

Chandola, V.; Banerjee, A.; Kumar, V. (2009): Anomaly detection: A survey. *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58.

Chen, J.; Li, W.; Lau, A.; Cao, J.; Wang, K. (2010): Automated load curve data cleansing in power systems. *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 213-221.

Cintuglu, M. H.; Mohammed, O. A.; Akkaya, K.; Uluagac, A. S. (2017): A survey on smart grid cyber-physical system testbeds. *IEEE Communications Surveys and Tutorials*, vol. 19, no.1, pp. 446-464.

Dean R. B.; Dixon, W. J. (1951): Simplified statistics for small numbers of observations. *Analytical Chemistry*, vol. 23, no. 4, pp. 636-638.

Grubbs, F. E. (1969): Procedures for detecting outlying observations in samples. *Technometrics*, vol. 11, no. 1, pp. 1-21.

Gupta, M.; Gao, J.; Aggarwal, C. C.; Han, J. (2014): Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250-2267.

Guo, Z.; Li, W.; Lau, A.; Inga-Rojas, T.; 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.

Hawkins, S.; He, H.; Williams, G. J.; Baxter, R. A. (2002): Outlier detection using replicator neural networks. *4th International Conference on Data Warehousing and Knowledge Discovery*, pp. 170-180.

Hilla, D. J.; Minsker, B. S. (2010): Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, vol. 25, no. 9, pp. 1014-1022.

Huang, G. B.; Zhou, H.; Ding, X.; Zhang, R. (2012): Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513-529.

Jones, M.; Nikovski, D.; Imamura, M.; Hirata, T. (2016): Exemplar learning for extremely efficient anomaly detection in real-valued time series. *Data Mining and Knowledge Discovery*, vol. 30, no. 6, pp. 1427-1454.

Kumar, N.; Singh, M.; Zeadally, S.; Rodrigues, J. J. P. C.; Rho, S. (2017): Cloud-assisted context-aware vehicular cyber-physical system for PHEVs in smart grid. *IEEE Systems Journal*, vol. 11, no. 1, pp. 140-151.

Kumar, N.; Zeadally, S.; Misra, S. C. (2016): Mobile cloud networking for efficient energy management in smart grid cyber-physical systems. *IEEE Wireless Communications*, vol. 23, no. 5, pp. 100-108.

Lam, D.; Wunsch, D. (2017): Unsupervised feature learning classification with radial basis function extreme learning machine using graphic processors. *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 224-231.

Lee, Y. T.; Hsiao, W. H.; Huang, C. M.; Chou, S. C. T. (2016): An integrated cloud-based smart home management system with community hierarchy. *IEEE Transactions on Consumer Electronics*, vol. 62, no. 1, pp. 1-9.

Liang, G.; Zhao, J.; Luo, F.; Weller, S. R.; Dong, Z. Y. (2017): A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1630-1638.

Lin, Y. H; Tsai, M. S. (2014): Development of an improved time-frequency analysis-based nonintrusive load monitor for load demand identification. *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 6, pp. 1470-1483.

Lin, Y. H.; Tsai, M. S. (2015): An advanced home energy management system facilitated by nonintrusive load monitoring with automated multi-objective power scheduling. *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1839-1851.

Ljung, G. M. (1993): On outlier detection in time series. *Journal of the Royal Statistical Society*, Series B, vol. 55, no. 2, pp. 559-567.

Ramaswamy, S.; Rastogi, R.; Shim, K. (2000): Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD International Conference on Management of Data*, pp. 427-438.

Ro, K.; Zou, C.; Wang, Z.; Yin, G. (2015): Outlier detection for high-dimensional data. *Biometrika*, vol. 102, no. 3, pp. 589-599.

Salvador, S.; Chan, P. (2005): Learning states and rules for time series anomaly detection. *Applied Intelligence*, vol. 23, no. 3, pp. 241-255.

Sanislav, T.; Zeadally, S.; Mois, G. D. (2017): A cloud-integrated, multilayered, agent-based cyber-physical system architecture. *Computer*, vol. 50, no. 4, pp. 27-37.

Shah, S.; Khalid, R.; Zafar, A.; Hussain, S. M.; Rahim, H. et al. (2017): An optimized priority enabled energy management system for smart homes. *IEEE 31st International Conference on Advanced Information Networking and Applications*, pp. 1035-1041.

Sheng, G.; Zhao, X.; Zhang, H.; Lv, Z.; Song, H. (2016): Mathematical models for simulating coded digital communication: A comprehensive tutorial by big data analytics in cyber-physical systems. *IEEE Access*, vol. 4, pp. 9018-9026.

Tang, G.; Wu, K.; Lei, J.; Bi, Z.; 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.

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