energies

MDPI

Article

Probabilistic Forecasting Based Joint Detection and Imputation of Clustered Bad Data in Residential Electricity Loads

Soyeong Park [1], Seungwook Yoon [2], Byungtak Lee [3], Seokkap Ko [3] and Euiseok Hwang [1,*]

[1] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Korea; soyeongp@gist.ac.kr

[2] School of Mechatronics, GIST, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Korea; ysw1207@gist.ac.kr

[3] Honam Research Center, Electronics and Telecommunications Research Institute, Gwangju 61012, Korea; bytelee@etri.re.kr (B.L.); softgear@etri.re.kr (S.K.)

* Correspondence: euiseokh@gist.ac.kr; Tel.: +82-62-715-3223

Abstract: Residential electricity load data can include numerous types of bad data, even clustered bad data, as they that are typically captured by simple measurement instruments. For example, in the case of a time-series of Not-a-Number (NaN) errors, the values before or next to a NaN may appear as the sum of actual values during the times of the NaN series. To utilize load data that includes such erroneous data for prediction or data mining analysis, customized detection and imputation should be conducted. This study proposes a new joint detection and imputation method for handling clustered bad data in residential electricity loads. Examples of these data are known invalid data points, such as consecutive NaN or zero values followed by or being ahead of an outlier. The proposed joint detection and imputation scheme first investigates the neighbors of the invalid data points, using probabilistic forecasting techniques. These techniques are implemented by the next valid neighbors to determine whether there is an anomaly or not. Then, adaptive imputations are applied on the basis of the detection , the candidate point should be imputed simultaneously or not. To assess the potential of the newly proposed scheme to characterize the clustered bad data, we analyzed the electricity loads of 354 households. Moreover, joint detection and imputations are conducted to test with the randomly injected synthesized clustered bad data (containing NaNs of various lengths) that is followed by the summation of the actual NaN values. The proposed scheme succeeded in detecting clustered bad data with an accuracy of 95.5% and a false alarm rate of 3.6% for all households in the dataset. Outlier detection-assisted imputation schemes are evaluated for NaNs with optional outliers. Results demonstrate that these schemes improve the overall accuracy significantly compared to schemes without outlier detection.

Keywords: bad data detection; probabilistic forecasting; residential electricity load

check for updates

Citation: Park, S.; Yoon, S.; Lee, B.; Ko, S.; Hwang, E. Probabilistic forecasting based joint detection and imputation of clustered bad data in residential electricity loads. *Energies* **2021**, *14*, 165. https://dx.doi.org/10.3390/en14010165

Received: 1 November 2020
Accepted: 25 December 2020
Published: 30 December 2020

1. Introduction

With the growing concerns on energy and environmental sustainability, a huge research effort has been made to achieve a smart and efficient energy management for decreasing the carbon footprint [1,2]. For this reason, an energy management system (EMS) has been introduced. This system efficiently controls the energy flexibility during generation, distribution, and consumption, considering the distributed energy resources (DERs) [3]. It is also applied in various areas such as factories, buildings, and houses, to conducts energy consumption forecasting and facility operation optimization on the basis of energy data analysis [4]. For prediction and optimization, various studies, which particularly focus on commercial or residential buildings containing several individual entities, are being conducted [5,6]. There are also numerous studies on independent residential homes, including load pattern analysis and clustering [7], machine learning-based prediction [8], and optimization considering renewable energy and the energy storage system [9].

Residential load raises a major energy management concern as it accounts for 56.9% of the final energy consumption of buildings [10]. Data-driven methods require data preprocessing, especially in the case of smart meter data. These methods are mainly used for processing residential electricity load data that contains numerous outliers caused for example by electricity theft [11]. Therefore, anomaly detection and imputation methods play a crucial role in obtaining reliable energy data.

With regard to anomaly detection, previous studies have employed traditional algorithms, such as the k-nearest neighbors (k-NN), support-vector machine (SVM), decision-tree (DT), as well as deep-learning methods, such as convolutional neural network (CNN), recurrent neural network (RNN), and generative adversarial network (GAN) to obtain energy consumption data [12]. Especially regarding residential load data, Xu et al. [13] suggested a detection method that combines RNNs and quantile regression. Various imputation techniques have been used to enhance imputation. These include clustering-based imputation with data located geographically close [14], bidirectional imputation combining long short-term memory model (LSTM) and transfer learning [15], and learning-based imputation based on the load pattern [16].

However, real residential energy consumption data has the tendency to include bad data, such as Not-a-Number (NaN) or zero points that are found scattered in several cases, even in the shape of clusters; and anomalies whose value is the sum of the actual values during the clustered bad data points. These outliers can significantly affect the performance of data-driven methods when handling clustered bad data. Anomaly detection and customized imputation are highly valuable in this case.

In this paper, a new joint detection and imputation method is proposed to complement the bad data in residential electricity load that contains clusters of NaNs and invalid data points before the clustered bad data. First, to determine whether the value is an outlier or not, probabilistic forecasting and probability distribution is performed, and the z-score of each invalid data point is obtained. These z-scores are utilized to detect the outlier with selecting the threshold obtained obtained using the loss function, the mean absolute error (MAE) after the imputation. Then, on the basis of the detection result, joint imputation based on the forward-backward joint auto-regressive (AR) model [17] is applied to the clusters of invalid data. If the value before the clustered bad data is detected as a normal one, imputation is applied only to the range of clustered bad data. However, if the value before the clustered bad data is judged it is an anomaly, it have to be handled with the clustered bad data during imputation. That is, the imputation range should be changed to include the previous point with clustered bad data when the detection result suspects that the previous point is the outlier.

2. Methodologies

The overall schematic of the proposed method is presented in Figure 1. As shown in the figure the data format is a vector, in particular univariate time-series data. Therefore, the dataset is usually treated as a shape of vector throughout the paper. We consider the following specific types of bad data: clustered bad data and outliers with accumulated values. Clustered bad data (CBD) is a set of consecutive bad data such as NaN and zero points. An outlier with accumulated values indicates the type of outlier that has a certain value that is estimated by the sum of actual values in data.

First, to figure out whether the value is an anomaly or not, we use probabilistic forecasting on the candidate points located before the CBD. The candidate point indicates the point located before the CBD because this location has the probability of having an outlier. The z-score of these points, which can be calculated using the mean and standard deviation, is classified by the threshold to minimize the MAE following imputation. Second, we implement the adaptive imputation schemes according to the anomaly detection schemes, based on the joint bidirectional models.

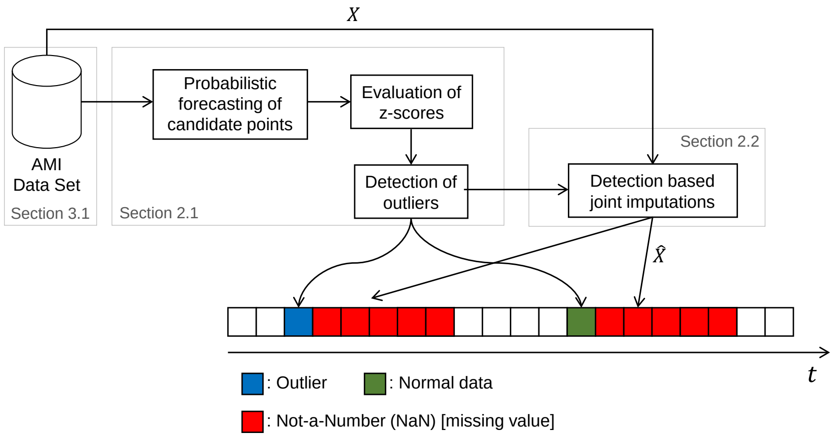


Figure 1. The overall schematic of the proposed methodology.

2.1. *Probabilistic Forecasting Based Anomaly Detection*

In this section, we propose a detection algorithm based on probabilistic forecasting. The residential dataset includes several types of bad data, and the outlier with accumulated values is usually generated before or after missing points appear in our dataset. As Algorithm 1 shows, the proposed detection algorithm that is based on the statistics and metrics is used to detect the outlier with accumulated values.

Algorithm 1: Detection algorithm for accumulated outlier

```
Input: observed values in candidate points $\boldsymbol{X}_{candidate} = \{x_{d_1,t_1}, ..., x_{d_n,t_n}\}$,
    corresponding workday types $\boldsymbol{v}_{candidate} = \{v_{d_1}, ..., v_{d_n}\}$, selected length of NaNs $T_{nan}$,
    searching range of threshold $thld\_max$, searching step of threshold $step$;
Output: Outlier detection result (0 or 1) for candidate points;
for $i = 1$ to $n$ do
    GET $\boldsymbol{N}_{d_i,t_i}$ in Equation (1);
    $z_{d_i,t_i} \leftarrow \frac{x_{d_i,t_i} - \hat{\mu}_{d_i,t_i}}{\hat{\sigma}_{d_i,t_i}}$;
end
thld = 0;
for $j = 1$ to $\lceil thld\_max/step \rceil$ do
    for $i = 1$ to $n$ do
        if $z_{d_i,t_i} < thld$ then
            imputation for $\{x_{d_i,t_i+1}, ..., x_{d_i,t_i+T_{nan}}\}$;
        else
            imputation for $\{x_{d_i,t_i}, ..., x_{d_i,t_i+T_{nan}}\}$;
        end
        $accuracy_j \leftarrow accuracy_j + \frac{MAE(imputation)}{n}$;
    end
    $thld$ += $step$;
end
$thld^* \leftarrow (\mathrm{argmin}_j(accuracy_j) - 1) * step$;
if $x_{d_i,t_i} < thld^*$ then
    $result_i \leftarrow 0$;
else
    $result_i \leftarrow 1$;
end
return: $Result$ ;
```

First, we implement the probabilistic forecasting to the candidate points. Here, we select the k-nearest neighbors (k-NN) algorithm as a probabilistic forecasting algorithm. The k-NN algorithm is not typically used in probabilistic forecasting, but we did so in this study as it reveals some statistical information, such as μ, σ and the distribution. With the basic k-NN algorithm, the physical location is only considered when collecting values; however, in this case, the hour time index and workday type are also utilized due to the daily and weekly pattern of the load profile. Let $x_{d,t}$ be the measured data at the time t of the d and v_d the indicator of the workday. Then, the nearest neighbors values are collected from the data points that have the same time t in the day d of the same workday type v_d with the target point among the previous days data, as follows:

$$\boldsymbol{N}_{d_i,t_i} = \left\{ x_{d_i+\alpha,t_i} \,\middle|\, v_{d_i+\alpha} = v_{d_i},\ \alpha = -k, -(k-1), \ldots, -1, 1, \ldots, k-1, k \right\}. \tag{1}$$

$\boldsymbol{N}_{d_i,t_i}$ denotes the nearest neighbors samples for the i-th target point x_{d_i,t_i}. In this method, the z-score is used to detect the outlier with accumulated values. The z-score is calculated by subtracting the mean value from the specific value and dividing the difference by the standard deviation, as follows:

$$z_{d_i,t_i} = \frac{x_{d_i,t_i} - \hat{\mu}_{d_i,t_i}}{\hat{\sigma}_{d_i,t_i}}, \tag{2}$$

where $\hat{\mu}_{d_i,t_i}$ and $\hat{\sigma}_{d_i,t_i}$ denote the mean and standard deviation of the probabilistic forecasting from the nearest neighbors values $\boldsymbol{N}_{d_i,t_i}$. Because the z-score measures the difference between the observed value and the sample mean in unit of the sample standard deviation, the large z-score value can be regarded as an indicator of the outlier. Thus, the threshold that determines whether the value is outlier or not should be selected in advance. In this study, the proper thresholds of z-scores are determined by minimizing the MAE imputation accuracy.

2.2. Forward-Backward Joint Imputation

This section presents an accumulated outlier aware joint imputation method. Figure 2 is the framework of the joint imputation method based on outlier detection. The proposed method considers only power data, not environmental data. It also considers time-series specification and error information. The load data is a time-series data that is affected by past and future data. Thus, unlike the prediction model, the imputation model can be constructed using the past and future data.

Proper application of the regression model is critical to estimating the CBD. Several models have been used for imputation, with the linear interpolation (LI) being the simplest model, expressed as follows.

$$\hat{x}^{LI}_{d,t_{n_0}} = x_{d,t_{n_0}} + \frac{x_{d,t_{n_0}+T_{nan}+1} - x_{d,t_{n_0}}}{T_{nan}+1} \times (t - t_{n_0}), \tag{3}$$

when the NaNs occur across the day, day and time indexes may need to be modified to make the sequence to be consecutive, for example $x_{d,n_d+1} = x_{d+1,1}$, where n_d is the number of data points in a day. For simple description, however, $x_{d,k}$ is used throughout the paper for $k \leq 0$ or $k > n_d$.

In Equation (3), x_t denotes the observed power data at time t; $t_{n_0} + T_{nan} + 1$, the time index right after CBD; t_{n_0}, the time index right before CBD; $t \in \{t_{n_0}+1, \ldots, t_{n_0}+T_{nan}\}$, the time indexes for CBD; and T_{nan}, the length of CBD. If the CBD length is shortened and data fluctuations are reduced, the LI model can demonstrate robust performance. However, the LI model is not suitable for data rebuilding because of the high variation of the household's power data. Therefore, the AR method, can be applied to the proposed model. AR is the statistical method that describes the value of a certain time point with respect to the past data. If the data has high autocorrelation, AR generates more accurate results. Because the electricity load data has autocorrelation and numerous researchers have investigated the whether the AR method is suitable for energy prediction, the data can be properly described by the forward-backward joint AR model. The AR method using the past data was designed as follows:

$$\hat{x}^{fwd}_{d,t_{n_0}} = \sum_{k=t_{n_0}-l^{fwd}}^{t_{n_0}} (w^{fwd}_{d,k}\, x^{fwd}_{d,k}), \tag{4}$$

where l^{fwd} denotes the length of the past data for the imputation of forward directions, and $w^{fwd}_{d,k}$ denotes the weight parameter of the AR method for forward directions. In addition, the future data can be obtained in the same way as follows:

$$\hat{x}^{bwd}_{d,t_{n_0}} = \sum_{k=t_{n_0}+T_{nan}+1}^{t_{n_0}+T_{nan}+1+l^{bwd}} (w^{bwd}_{d,k}\, x^{bwd}_{d,k}), \tag{5}$$

where l^{bwd} denotes the length of future data for the imputation of backward directions and $w^{bwd}_{d,k}$ denotes the weight parameter of the AR method for forward directions.

Finally, the forward-backward joint imputation is designed by combining Equations (4) and (5) because both past and future data can be considered for imputation, unlike forecasting. The forward-backward joint imputation is expressed as follows:

$$\hat{x}^{joint}_{d,t_{n_0}} = \sum_{k=t_{n_0}-l^{fwd}}^{t_{n_0}} (w^{fwd}_{d,k}\, x^{fwd}_{d,k}) + \sum_{k=t_{n_0}+T_{nan}+1}^{t_{n_0}+T_{nan}+1+l^{bwd}} (w^{bwd}_{d,k}\, x^{bwd}_{d,k}). \tag{6}$$

The proposed model performs the imputation process in the forward and backward directions. The first term considers the past time index starting with $t_{n_0} - l^{fwd}$ and ending with t_{n_0}, and the second term considers the future time index from $t_{n_0} + T_{nan} + 1$ to $t_{n_0} + T_{nan} + 1 + l^{bwd}$. That is, with past and future sequences simultaneously for the omitted range, $\{t_{n_0} + 1, ..., t_{n_0} + T_{nan}\}$. Meanwhile, the other case that includes the outlier with accumulated values in front of CBD, the imputation range should contain the outlier point, t_{n_0}. In other words, the imputation range should be replaced with the range $\{t_{n_0}, t_{n_0} + 1, ..., t_{n_0} + T_{nan}\}$, including the outlier point with accumulated values t_{n_0} as follows:

$$\hat{x}^{joint}_{d,t_{n_0}} = \sum_{k=t_{n_0}-l^{fwd}}^{t_{n_0}-1} (w^{fwd}_{d,k}\, x^{fwd}_{d,k}) + \sum_{k=t_{n_0}+T_{nan}+1}^{t_{n_0}+T_{nan}+1+l^{bwd}} (w^{bwd}_{d,k}\, x^{bwd}_{d,k}). \tag{7}$$

Unlike (6), the time point t_{n_0} is excluded in the upper bound of the first term, it becomes $t_{n_0} - 1$. With this change, the omitted range is replaced with $\{t_{n_0}, ..., t_{n_0+T_{nan}}\}$. Generally, in (6) and (7), both sequences are considered containing both $x^{fwd}_{d,t_{n_0}}$ and $x^{bwd}_{d,t_{n_0}}$ in the vector $\mathbf{x}^{fb}_{d,t_{n_0}}$, while calculating the weight matrix $\hat{\mathbf{w}}_{d,t_{n_0}+k}$ as follows:

$$\begin{aligned}
\mathbf{x}^{fb}_{d,t_{n_0}} &= \left[\mathbf{x}^{fwd\ T}_{d,t_{n_0}}\ \mathbf{x}^{bwd\ T}_{d,t_{n_0}} \right]^T \\
\hat{\mathbf{w}}_{d,t_{n_0}+k} &= (\boldsymbol{X}^{fb\ T}_{d,t_{n_0}} \boldsymbol{X}^{fb}_{d,t_{n_0}})^{-1} \boldsymbol{X}^{fb\ T}_{d,t_{n_0}} \mathbf{x}^{tg}_{d,t_{n_0}+k}\ (k = 0, ..., T_{nan}) \\
\hat{\mathbf{x}}_{d,t_{n_0}} &= \left[\hat{x}_{d,t_{n_0}}\, \hat{x}_{d,t_{n_0}+1} ... \hat{x}_{d,t_{n_0}+T_{nan}} \right]^T = \begin{bmatrix} \hat{\mathbf{w}}^T_{d,t_{n_0}} \\ \hat{\mathbf{w}}^T_{d,t_{n_0}+1} \\ \vdots \\ \hat{\mathbf{w}}^T_{d,t_{n_0}+T_{nan}} \end{bmatrix} \mathbf{x}^{fb}_{d,t_{n_0}}.
\end{aligned} \tag{8}$$

where $\mathbf{X}^{fb}_{d,t_{n_0}} = [\dots \mathbf{x}^{fb}_{d-b_i,t_{n_0}}{}^{T} \dots]$ and $\mathbf{x}^{tg}_{d,t_{n_0}+k} = [\dots x_{d-b_i,t_{n_0}+k} \dots]^T$ for b_is where $v_{d-b_i} = v_d$, $b_i \in \{-b, ..., -1, 1, ..., b\}$.

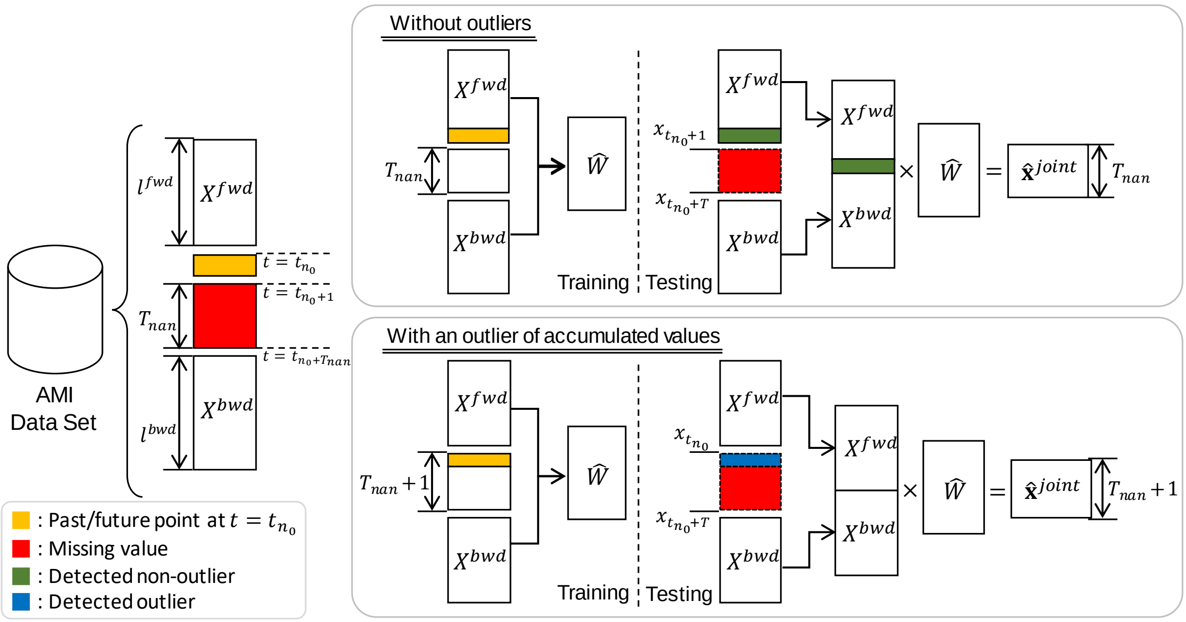


Figure 2. The framework of the forward-backward joint imputation.

3. Numerical Evaluation

To verify the feasibility of the newly proposed method, we applied it to 354 residential electricity load data [18] to verify the feasibility of the proposed method. We obtained the energy consumption data from households in Incheon, South Korea, with a resolution of 15 min. But the dataset was preprocessed to change the resolution from 15 min to 1 h by the cooperating organization, because of invasion of privacy issues. This data includes CBDs obtained from metering system faults or communication errors.

The simulation involves artificially injecting CBD to the valid points to calculate the accuracy and compare performance of the proposed method. Two types of anomaly group are injected: one is the plain CBD that includes only NaNs, and the other one is the CBD that includes an outlier with accumulated values in front of it. Although it is possible to vary and treat the length of injected CBDs, their length is fixed to five to prevent disturbances during simulation. The valid points, wherein the outlier will be placed, are the same as in Section 2.1.

3.1. Data Analysis

In this section, the used data was analysed about outlier with accumulated values. Dataset have to contain CBDs and outliers in front of CBDs to apply the proposed method. To confirm whether these pattern is included or not in the dataset rapidly, 5 households was randomly selected and analyzed. It was verified whether there are the values that increase in proportion to the length of NaNs, and the ratio of potential outliers was analyzed with a conventional method.

As stated previously, in residential electricity load data, NaN and zero padding appear as points or clusters, and the outliers are estimated as the sum of existing actual data. In particular, we analyzed whether the outlier with accumulated values exists or not in the actual data, because the proposed method is concerned with adaptive imputation. The analysis was conducted on five random households. As the dataset was preprocessed to change the resolution from 15 min to 1 h, it was estimated that an outlier with accumulated values will appear at 1/4 ratio due to NaN processing method. The point in front of CBD is termed 'candidate point', and the z-score of candidate points according to the length of CBD is shown in Figure 3. The z-score was calculated in the same manner as the k-NN method in Section 2.1. The blue ellipse in the figure indicates potential outliers according to increasing the length of CBD, which means that the sum of NaNs exist in used data.

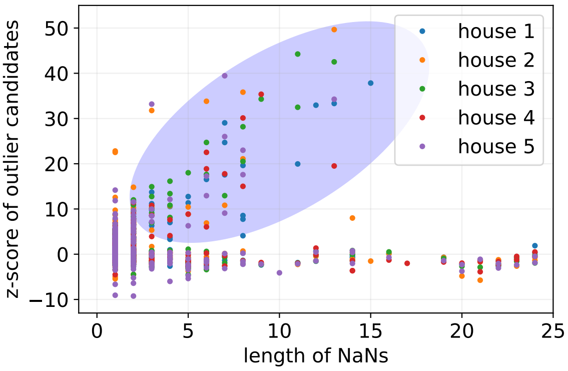


Figure 3. Analysis of the potential outliers and the length of not-a-number (NaN) sequences.

The result of calculating the ratio of values classified as outliers based on the 3-sigma rule is shown in Figure 4. The case that has a length of nine or more was excluded, as there are only 1 or 2 sequences for each household. The total ratio of potential outlier was 0.32, which is close to the estimated value (1/4). Considering that the outlier can appear for other reasons, we did not reject the assumption that the outlier with accumulated values occurs at a quarter ratio. Therefore, outliers with accumulated values have been observed in the dataset and can be imputed with customized methods.

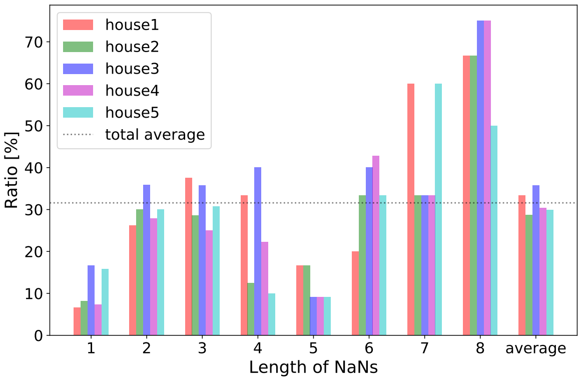


Figure 4. Analysis of the potential outliers and the length of NaN sequences with a preliminary 3-sigma threshold.

Figure 5 presents normalized residue analysis results, where we can see whether the z-score follows the standard normal distribution [19]. Due to the uncertainty of the probabilistic forecasting, the residue is not identically distributed as standard norma, and the threshold for decision may need to be set larger than the preliminary three sigma, which will be discussed in the following subsection.

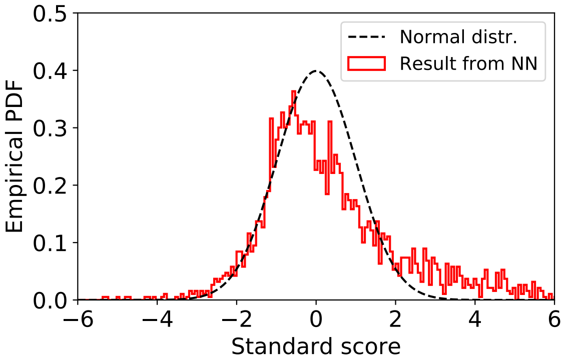


Figure 5. Normalized residue analysis: comparison between normal distribution and result from nearest neighbors.

3.2. Detection Results for Residential Data

In this section, we evaluate the performance of the anomaly detection based on probabilistic forecasting. The proposed method is compared with the 3-sigma rule in statistics, which is the case when the decision criteria is 3.0 and prove the performance of proper z-score threshold. First, while searching the threshold, MAE after imputation for varying thresholds is shown as Figure 6. The test result in this particular household indicates that when the threshold is changed to 7.8, the total imputed MAE is the lowest, which is indicated by the red dashed line in Figure 6. The proper thresholds are selected depending on each household.

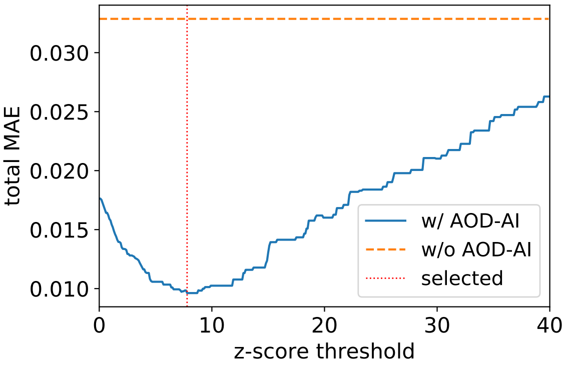


Figure 6. Total mean absolute error (MAE) with varying z-score thresholds.

Second, the result of the anomaly detection for the proper threshold is presented as a confusion matrix and compared with the result for the threshold of 3, as presented in Table 1. The number of the overall injected bad data series is 72,882, 24,113 of which include anomaly in front of NaNs. The total accuracy is 95.5% and 91.6% with the results from selected threshold and preliminary threshold, respectively. The proposed method demonstrated a false alarm rate of 3.6%, while the preliminary case demonstrated a false alarm rate of 9.7%. The scanning process enhanced the performance in terms of true negative (TN) and false positive (FN). Comparing by cases, TN and false positive (FP) were improved in the proposed method; TN was increased by 6.8%, and FP was decreased by 63.0%. Conversely, true positive (TP) and false negative (FN) slightly deteriorated due to the trade-off in binary classification. Although the accuracy decreased in the case of true anomaly, this result can be significant in terms of the FP has been notably improved. Because the normal value point is corrupted during the imputation, the FP must be minimized as much as possible. The proposed method demonstrated the lowest total

MAE at the proper threshold. It also demonstrated better performance compared with conventional methods.

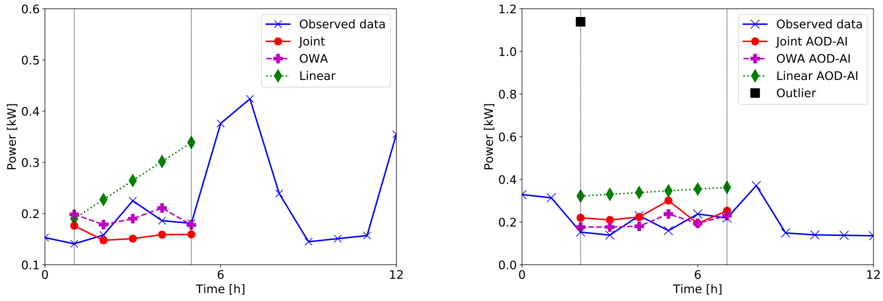
Table 1. Detection results of the residential dataset (**a**) with the selected threshold from scanning and (**b**) with a preliminary threshold of 3-sigma.

		(a)	
		Prediction	
		Normal	**Anomaly**
True	**Normal**	70,272	2610
	Anomaly	1765	22,348
		(**b**)	
		Prediction	
		Normal	**Anomaly**
True	**Normal**	65,821	7061
	Anomaly	1065	23,048

3.3. Imputation Results of the Residential Data

In this section, we implemented two cases to compare the performance of the accumulated outlier detection aware imputation (AOD-AI)—one without AOD-AI and the other one with AOD-AI. To compared the accuracy of the joint imputation, we used two methods: LI as a baseline, and optimally weighted average (OWA) [20] as one of the recent promising imputation scheme. OWA is the imputation method for smart meter data that combines LI and historical average (HA) imputation by the weighted sum with optimized weights.

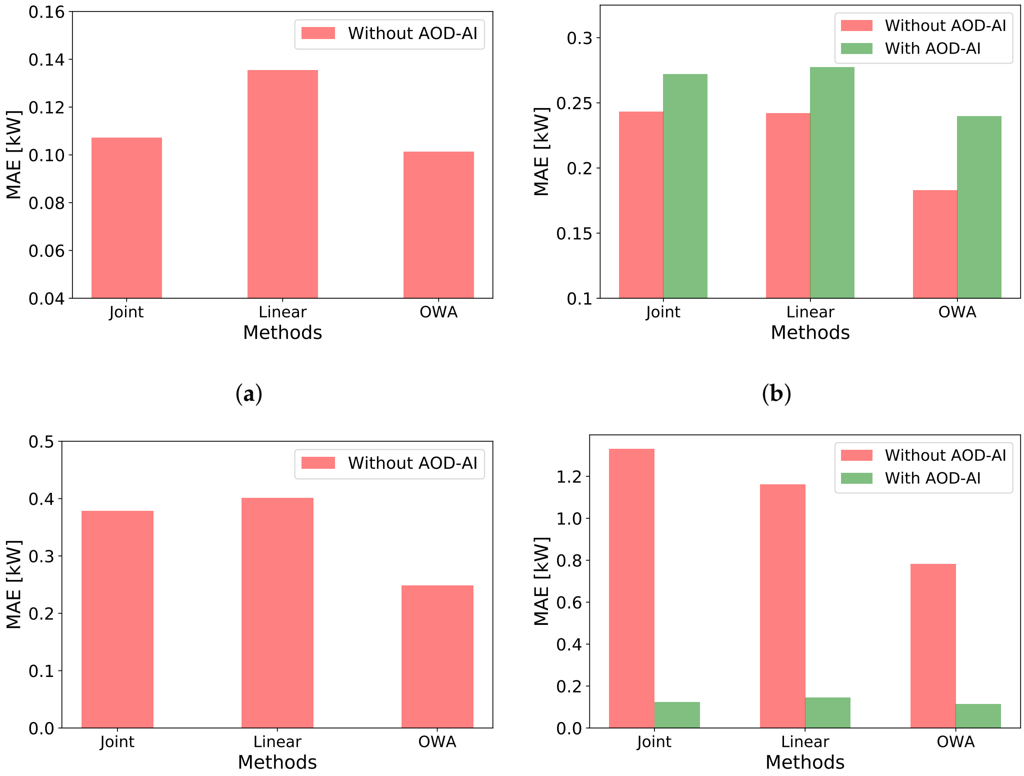
In Figure 7a, the imputation results without outliers are presented in three methods. LI method exhibits low accuracy as the power data is affected by human patterns and has a specific shape by household. Conversely, the performances of the joint imputation and OWA methods were found to be better compared with those of the LI model, considering the historical data. In Figure 7b, the imputation results with outliers are presented only AOD-AI applied. If the outlier with accumulated values are not removed, all methods will derive terrible results. When the outlier was removed and imputed with CBD, the imputation performance was improved in all cases.



(**a**) (**b**)

Figure 7. The imputation results of the residential dataset (**a**) without outliers when detected as non-outliers (TN) and (**b**) with outliers of accumulated values when detected as outliers (TP).

Figure 8 presents the evaluation of the residual errors by the result of outlier detection. In Figure 8a, the results indicate how the imputation model handles the historical or surrounding data well. OWA accuracy is the best, and it seems that the combination of LI reflected the level of imputation range to the imputation result. But with joint method, it is possible to derive better performance with introduce additional criterion in searching training set. In Figure 8b, the application of AOD-AI is worse than the case without AOD-AI. The corruption of normal values in candidate points produced poor imputation results. In Figure 8c, the best method is OWA and the worst is LI. While the joint method and LI are influenced by remaining outliers in front of CBD, it seems OWA is robust due to the historical average method. In Figure 8d, outliers are detected precisely and imputed with CBD. So, the result of AOD-AI shows similar accuracy as in the case (a).



(**a**) (**b**)

(**c**) (**d**)

Figure 8. The imputation error evaluation with the residential dataset in case of (**a**) true negative (**b**) false positive (**c**) false negative (**d**) true positive.

The overall performance of the imputation is summarized in Table 2. From the table, we observe that the overall performance of the joint method with AOD-AI is 0.108, OWA method with AOD-AI is 0.100, and LI with AOD-AI is 0.131 in MAE, which correspond to the reductions in MAE by 71.9%, 60.3%, and 63.9% respectively, relative to the schemes without AOD-AI. OWA shows the best performance considering the detection result. Except for OWA, other methods were also improved with AOD-AI. This result shows the strength of AOD-AI and the necessity of outlier detection and adaptive imputation.

Table 2. Total imputation results of residential dataset.

Method	**AOD**	**MAE [kW]**
Joint method	without AOD-AI with AOD-AI	0.385 0.108
Linear interp.	without AOD-AI with AOD-AI	0.363 0.131
OWA	without AOD-AI with AOD-AI	0.252 0.100

4. Conclusions

In this paper, we introduce the joint detection and imputation scheme for residential electrical load data. To impute the missing values, the proposed method considers two types of bad data—CBD and outliers with accumulated values. Since outliers with accumulated values deteriorate imputation performance, the proposed method estimates the missing values while detecting the accumulated outlier. It also provides numerous benefits:

- The outlier can be detected without any labeled data about bad data. Since the annotated data is rare in smart meter data, it is appropriate for the residential load dataset.
- The data was analyzed that some outliers are located in front of CBD. The possibility of outlier is considered and excludes not only bad data but the point before and after CBD during probabilistic forecasting, enabling more accurate detection.

In numerical evaluations, the imputation models are simulated using a residential dataset. The imputation accuracy can be increased by 60.3% to 71.9% by the proposed AOD-AI approach with the imputation schemes of LI, OWA, and joint feed-forward. Since the proposed imputation method combined outlier detection and joint imputation, only the method without AOD-AI can be compared as a previous research. Therefore, the joint imputation with AOD-AI shows better performance than any other method. As shown as the total imputation result, the AOD-AI can be applied to any imputation model and enhance the accuracy of imputation enormously. The proposed imputation model is influenced by the detection errors. Thus, it is important to properly control the threshold for the detection to reduce errors. Also, deep learning can be applied to the imputation and detection models to increase the performance of the proposed scheme.

Author Contributions: S.P. devised the idea and designed the methodogies in this manuscript as the first author. S.Y. assisted with the imputation method and analyzed the simulation results. B.L. analyzed the electricity load data and conducted preprocessing the data and S.K. characterized and analyzed residential electricity load abnormality. E.H. instructed and supervised the research as the corresponding author. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Trade, Industry & Energy(MOTIE) of the Republic of Korea (No. 20181210301570).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/etri/Power_Usage_Dataset.

Acknowledgments: This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Trade, Industry & Energy(MOTIE) of the Republic of Korea (No. 20181210301570) and by the Energy AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea(NIPA) funded by the Ministry of Science and ICT (No. S1601-20-1011).

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

The following nomenclatures are listed including the abbreviations to referring terms and variables used in equations:

AOD-AI	Accumulated outlier detection aware imputation
AR	Autoregression
CBD	Clustered bad data
k-NN	k-nearest neighbors
MAE	Mean absolute error
NaN	Not-a-Number
FN	False negative
FP	False positive
TN	True negative
TP	True positive
LI	Linear interpolation
OWA	Optimally weighted average
Candidate point	The point located in front of clustered bad data
$x_{d,t}$	Observed values at day d, hour t
d	Day index
t	Time index [hour]
v_d	Workday type (Binary) in day d
t_{n_0}	The time point that occurred accumulated outlier (AO)
T_{nan}	The length of clustered bad data (CBD)
$z_{d,t}$	Z-score of $x_{d,t}$
$\hat{\mu}_{d,t}$	Mean from the result of probabilistic forecasting
$\hat{\sigma}_{d,t}$	Standard deviation from the result of probabilistic forecasting
$\hat{\mathbf{w}}_{d,t}$	Weight matrix for autoregression (AR)

References

1. Wood, N.; Roelich, K. Tensions, capabilities, and justice in climate change mitigation of fossil fuels. *Energy Res. Soc. Sci.* **2019**, *52*, 114–122. [CrossRef]
2. Destek, M.A.; Aslan, A. Disaggregated renewable energy consumption and environmental pollution nexus in G-7 countries. *Renew. Energy* **2020**, *151*, 1298–1306. [CrossRef]
3. Pfeifer, A.; Dobravec, V.; Pavlinek, L.; Krajačić, G.; Duić, N. Integration of renewable energy and demand response technologies in interconnected energy systems. *Energy* **2018**, *161*, 447–455. [CrossRef]
4. Lee, D.; Cheng, C.C. Energy savings by energy management systems: A review. *Renew. Sustain. Energy Rev.* **2016**, *56*, 760–777. [CrossRef]
5. Park, K.; Yoon, S.; Hwang, E. Hybrid load forecasting for mixed-use complex based on the characteristic load decomposition by pilot signals. *IEEE Access* **2019**, *7*, 12297–12306. [CrossRef]
6. Yoon, S.; Hwang, E. Load guided signal-based two-stage charging coordination of plug-in electric vehicles for smart buildings. *IEEE Access* **2019**, *7*, 144548–144560. [CrossRef]
7. Diao, L.; Sun, Y.; Chen, Z.; Chen, J. Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy Build.* **2017**, *147*, 47–66. [CrossRef]
8. Lee, J.; Kim, J.; Ko, W. Day-Ahead Electric Load Forecasting for the Residential Building with a Small-Size Dataset Based on a Self-Organizing Map and a Stacking Ensemble Learning Method. *Appl. Sci.* **2019**, *9*, 1231. [CrossRef]
9. Ahmad, A.; Khan, A.; Javaid, N.; Hussain, H. M.; Abdul, W.; Almogren, A.; Alamri, A.; Azim Niaz, I. An optimized home energy management system with integrated renewable energy and storage resources. *Energies* **2017**, *10*, 549. [CrossRef]
10. Jeong, Y.S. Assessment of alternative scenarios for CO_2 reduction potential in the residential building sector. *Sustainability* **2017**, *9*, 394. [CrossRef]
11. Gu, Y.; Liu, T.; Wang, D.; Guan, X.; Xu, Z. Bad data detection method for smart grids based on distributed state estimation. In Proceedings of the 2013 IEEE International Conference on Communications (ICC), Budapest, Hungary, 9–13 June 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 4483–4487.
12. Himeur, Y.; Alsalemi, A.; Bensaali, F.; Amira, A. Anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *arXiv* **2020**, arXiv:2010.04560.
13. Xu, C.; Chen, H. A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data. *Energy Build.* **2020**, *215*, 109864. [CrossRef]
14. Park, S.M.; Park, S.Y.; Kim, M.; Hwang, E. Clustering-Based Self-Imputation of Unlabeled Fault Data in a Fleet of Photovoltaic Generation Systems. *Energies* **2020**, *13*, 737. [CrossRef]
15. Ma, J.; Cheng, J.C.; Jiang, F.; Chen, W.; Wang, M.; Zhai, C. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy Build.* **2020**, *216*, 109941. [CrossRef]

16. Kim, M.; Park, S.; Lee, J.; Joo, Y.; Choi, J.K. Learning-based adaptive imputation methodwith kNN algorithm for missing power data. *Energies* **2017**, *10*, 1668. [CrossRef]
17. Tufts, D.W.; Kumaresan, R. Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood. *Proc. IEEE* **1982**, *70.9*, 975–989. [CrossRef]
18. ETRI Power Usage Dataset. Available online: Https://github.com/etri/Power_Usage_Dataset (accessed on 29 November 2020).
19. Park, S.Y.; Park, S.M.; Hwang, E. Normalized Residue Analysis for Deep Learning Based Probabilistic Forecasting of Photovoltaic Generations. In Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Korea, 19–22 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 483–486.
20. Peppanen, J.; Zhang, X.; Grijalva, S.; Reno M.J.. Handling bad or missing smart meter data through advanced data imputation. In Proceedings of the 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Minneapolis, MN, USA, 6–9 September 2016; IEEE: Piscataway, NJ, USA, 2020; pp. 483–486.