Anomaly detection and missing data imputation in building energy data for automated data pre-processing

K Takahashi¹, R Ooka² and S Ikeda³

¹ Department of Architecture, University of Tokyo
² Institute of Industrial Science, University of Tokyo
³ Department of Innovation Science, Tokyo Institute of Technology

Abstract. A new trend in building automation is the implementation of smart energy management systems to measure and control building systems without a need for decision-making by human operators. Artificial intelligence can optimize these systems by predicting future demand to make informed decisions about how to efficiently operate individual equipment. These machine learning algorithms use historical data to learn demand trends and require high quality datasets in order to make accurate predictions. But because of issues with data transmission or sensor errors, real world datasets often contain outliers or have data missing. In most research settings, these values can be simply omitted, but in practice, anomalies compromise the automation system’s prediction accuracy, rendering it unable to maximize energy savings. This study explores different machine learning algorithms for anomaly detection for automatically pre-processing incoming data using a case study on an actual electrical demand in a hospital building in Japan, namely cluster-based techniques such as k-means clustering and neural network-based approaches such as the autoencoder. Once anomalies were identified, the missing data was filled with prediction values from a deep neural network model. The newly composed data was then evaluated based on detection accuracy, prediction accuracy and training time. The proposed method of processing anomaly values allows the prediction model to process collected data without interruption, and shows similar predictive accuracy as manually processing the data. These predictions allow energy systems to optimize HVAC equipment control, increasing energy savings and reducing peak building loads.

1. Introduction

With the increased interest in effectively reducing building energy consumption and recent improvements to smart building technologies, some buildings have recently started adopting artificial intelligence into their building automation systems (BAS). These systems predict the future demand based on historical trends and proactively control mechanical systems to optimize energy use. Not only can these systems reduce peak demand and annual energy use, but they can also reduce the need for on-site human operators and improve occupant comfort [1].

These smart building systems require high quality datasets which is an issue with working with real datasets. Measured data from sensors or meters have missing data or outlier values due to transmissions issues or disconnections with the BAS. Using contaminated data in carefully tuned machine learning algorithms can cause errors in the prediction results, leading to unexpected building operations. During the research and development stages, these outlier data values are processed manually before being applied to machine learning algorithms, but for a smart BAS to control building systems in real time, the incoming data must be validated, and anomalies must be corrected before being incorporated into the algorithm.
Anomaly Detection is a technique used in many fields to identify outliers outside of the normal expected range. It is applied to detecting fraud, human errors, defects, or novelties and has a range of different methods [2]. Many past research topics have covered using various anomaly detection techniques for equipment fault detection [3], but this study applies anomaly detection techniques to identify and replace the outlier values and clean the dataset before it is used for the prediction model. This paper compares the results of different anomaly detection algorithms to automatically pre-process the dataset which is then used for an existing prediction model.

2. Case Study
A study was done on a hospital building in Ibaraki, Japan which is in the process of implementing a smart BAS to optimize building energy use by managing the control of electrical, cooling, heating, and hot water systems. Currently, development is taking place for a prediction machine learning algorithms model that takes the prior 3 days of each of the end uses and predicts the next 30 hours using an ensemble machine learning algorithm that takes the weighted average of the results from a deep neural network model (DNN) and a random forest regressor model (RF). Further detail of the model structure and prediction accuracy of the existing prediction algorithm will be discussed in Section 3.3.

The hospital has recorded the past 5 years of electrical demand data logged every 30 minutes. Data exploration revealed three types of outlier values in the electrical demand presented in Figure 1. Of the three types of anomalies identified, the delayed reporting is of the greatest concern since a small delay in reporting may only reveal a small error and make it difficult to identify incorrectly logged data. Otherwise, anomalies for electrical demand are often more obvious than the other end uses because the hospital is a 24-hour operation building and has heavy process driven demand. This study focuses on the electrical demand with future studies looking into applying the same algorithm to detect anomalies in the other energy end uses.

![Single Missing](image1)

**Single missing (28 cases)**
Caused by disconnection with the server and no data was recorded

![Delayed Reporting](image2)

**Delayed reporting (2 cases)**
Cause by disconnection with the server that recovered in between recording intervals, causing the value to be added to the next interval resulting in a big value

![Continuous Missing](image3)

**Continuous missing (2 cases)**
Data is not recorded for a long interval, up to 16 days, because of server maintenance.

**Figure 1.** Types of anomalies found in the past data.
2.1. Research Setup
To compare each algorithm’s detection performance, this study implemented a semi-supervised learning method [4] where each model was trained using manually pre-processed data from 2016-2019 with no major outliers or known anomalous values. From the trained dataset, the range of observed errors, or anomaly score, was calculated and a threshold for anomaly classification was determined. Then, to test how well each model can detect anomalies, the models ran 10 different versions of the 2020 data with artificially added anomalous data. The artificial anomalies are random values between zero and double the maximum recorded anomaly, with at least one continuous missing data ranging from three to five days. The final evaluation was the average score of the 10 runs.

The evaluation metrics for anomaly detection used the recall, precision, and F1 score as well as the training time. The recall score represents how accurately anomalies that were correctly identified as anomalous values with a recall score of 1.0 representing all anomalous values were identified. Precision score shows how many of the detected anomalies were real anomalies and not normal data, where a precision score of 1.0 represents all identified values were anomalies. A good algorithm will have a good balance of recall and precision scores; F1 score represents the average of the recall and precision score, where a F1 score of 1.0 represents a perfect score where all anomalies were detected and none of the normal values were misidentified as anomalies.

Once the model detects an anomalous value, it replaces it with the predicted value. The amended data was passed through the existing prediction algorithm to check if the automatically pre-processed data can perform as well as manually processed data. The prediction accuracy was assessed using the average Root Mean Squared Error (RMSE) and the minimum, maximum, and average Expected Error Percentage (EEP). RMSE represents the magnitude of the error given by the square root of the mean squared error which is squared difference between the true electrical demand vector of the time segment \( y \) and the predicted electrical demand vector \( y' \) over the length of the time segment \( n \). EEP represents the deviation of the predicted electrical demand from the real demand and is given by the RMSE of that time segment divided by the observed maximum electrical demand, \( y_{max} \). EEP is the common metric for determining the performance of building thermal load prediction models [5]. The goal for this study was to match or exceed the prediction accuracy of a manually preprocessed dataset.

\[
RMSE = \left( \frac{(y-y')^2}{n} \right)^{1/2}
\]

\[
EEP = \frac{RMSE}{y_{max}} \times 100\%
\]

3. Methodology

3.1. k-Means Clustering
k-means clustering is a type of unsupervised machine learning model that divides the dataset into k clusters. Each data point is assigned to the closest cluster based on the distance to the mean (or cluster centroid). The algorithm randomly assigns each data point a label and calculates the centroid of each set of labels. Then the data is re-assigned a new label based on the closest cluster centroid. The centroid is then recalculated and the process is repeated until the centroid no longer moves.

3.1.1. Model Structure. For this study, the k-means algorithm takes in electrical demand, outdoor air temperature, month, and hour data and divides the data into clusters. The results of clustering can be seen in Figure 2 which shows clear horizontal divides for the clusters on the demand and outdoor air temperature plot. This indicates that the outdoor air temperature has little influence on determining the clusters, which is logical, because this hospital building has high process loads. Using the elbow method, a heuristic approach to determining the number of clusters, dividing the dataset into 5 clusters was determined to be sufficient for this study.
3.1.2. **Anomaly Score.** Euclidean distance from a point to the center of the nearest cluster was used as the anomaly score for the detection criteria. From the training dataset, the distance from each datapoint to the cluster centers were measured. The maximum distance recorded was then set as the maximum allowable threshold. The dataset containing anomalous values was then classified into the nearest clusters and the distances to the centroid were measured. If the distance was greater than the threshold, the datapoint was classified as an anomaly.

3.2. **Autoencoder**

Autoencoders are a special type of unsupervised neural network model that takes the input and returns it back as the output but with reduced noise [3][6]. The input data is first compressed into a “coded” form using an encoder, which is a neural network model that reduces the number of neurons in the hidden layer down to a bottleneck. Then, a decoder neural network takes the coded data and reconstructs it back to the original input data. Instead of directly copying the input data, the autoencoder can learn the most important characteristics of input data and ignore any of the noise.

In its application for anomaly detection, autoencoders learn the normal dataset and is able to capture its main features by removing noise. From the normal data set, a threshold beyond which data is considered anomalous is determined; this threshold is the maximum allowable anomaly score. When data with anomalous points is passed through the trained model, it will fail to reconstruct the input data, leading to higher anomaly score. If the anomaly score is above the threshold value, the input data is classified as an anomaly.

3.2.1. **Model Structure.** For this study, the autoencoder model was built using a one-dimensional convolutional neural network (1D-CNN) model. Convolutional neural network models are commonly used in image recognition with two-dimensional inputs but can be applied one-dimensional to recognize patterns in time series data. Unlike using a simple neural network where each feature is assumed to be independent of one another, 1D-CNN models use the neighboring values of each datapoint, which makes them a good fit for time series data classification problems where the values are affected by previous time steps.

Encoder side takes the prior 72 hours of electrical demand, outdoor air temperature, and month data to compress the data down to a minimum of 18 hidden layer nodes. There are 3 hidden layers of 1D-CNN each with a stride of 2 to compress the data by half each layer. Between each 1D-CNN layer, a dropout layer with 25% dropout rate and a batch normalization layer was added to prevent overfitting. The decoder side then expands the compressed data back to the original data size by using 3 transpose layers to reproduce the electrical demand. Figure 3 presents the simplified structure of the autoencoder model.
3.2.2. Anomaly Score. Mahalanobis distance is a metric that finds the distance away from a point to a distribution. Unlike the Euclidean distance used for k-means, Mahalanobis distance is effective for multivariate data because it considers the covariance between each variable, so highly correlated data do not affect the distance. If the variables have no correlation, the Mahalanobis distance and the Euclidean distance are the same, so Mahalanobis distance is considered a more generalized form of distance.

For input data vector $\mathbf{x}$ and reconstructed data vector $\mathbf{x}'$, anomaly score is given by Equation (3) where $\mathbf{e}$ is the error vector given by $\mathbf{e} = |\mathbf{x} - \mathbf{x}'|$, $\mu$ is the mean error, and $\Sigma^{-1}$ is the inverse covariance matrix. Using Mahalanobis distance as the anomaly score was presented by Malhota et al 2015 [6].

$$D_M = \sqrt{(\mathbf{e} - \mu)^T \Sigma^{-1} (\mathbf{e} - \mu)}$$  

(3)

Another reason the Mahalanobis distance was used for anomaly score over other metrics was because for multivariate data, metrics like mean absolute error examine the average error of the entire segment. The average score gives single large errors a smaller penalty, while an accumulation of small errors is given a higher penalty. Mahalanobis distance, on the other hand, can detect these sudden changes, making it ideal for this study. For example, Figure 4 shows a normal data segment in the summer with higher than average demand compared with a data segment that includes an anomalous value. Because the high demand data has a greater average error, an anomaly score using the mean absolute error can incorrectly identify this data segment, but Mahalanobis distance can correctly identify the large change in error and recognize the segment as an anomaly.

![Figure 3. Architecture of the autoencoder model.](image)

![Figure 4. Comparison of Mean Absolute Error loss and Mahalanobis distance loss as a metric for anomaly score.](image)
3.3. Prediction Model Error based Detection

The prediction model error based detection technique is a proposed approach for identifying anomalies for time series data. Unlike detection techniques based on unsupervised learning algorithms, this approach uses a supervised learning algorithm and takes the predicted value from the trained model and compares its results with the real incoming data. If the new data results in a prediction error above a specified threshold, the data is classified as an anomaly.

A possible risk for this approach is that the model will accidentally learn anomalous data as normal values since the base algorithm is a supervised learning model. In the training stage, extra care must be put in place to eliminate known anomalies. This can be hard when the anomalies are not known or hard to identify.

3.3.1. Model Structure. The existing prediction model is an ensemble model that takes the weighted average results of a DNN and a RF regressor model. The DNN model was found to perform well at predicting the periodic results, while the RF model took the noise of the data into account. A combined weight of 3:1 respectively showed improvements in prediction results compared to each individual model. Since trying to predict the small noise in the data can interfere with anomaly detection, this study only looked at the results from the DNN model to calculate the error.

The DNN model uses the prior three days or 144 timesteps of electrical demand data, the next 30 hours or 60 timesteps of outdoor air temperature data, and the date time information to predict the next 60 timesteps of electrical data. The neurons use a Scaled Exponential Linear Unit (SELU) activation which is a self-normalizing function that converges to a mean of zero and a standard deviation of one and is well adapted for handling many layers of nodes [7]. From previous analysis, it was found that the DNN model can predict the electrical demand at a performance rating of 19.0 RMSE and 1.8% EEP.

3.3.2. Anomaly Score. The relative error between the first timestep of the predicted values and incoming value was used as the anomaly score. Because the DNN model had the risk of learning anomalous data, unlike the other two techniques, which took the maximum training anomaly score, the threshold was determined by fitting the training errors to normal distribution. Any prediction error greater than 6 standard deviations away from the mean was classified as an anomaly.

![Error distribution of the prediction model training data](image)

**Figure 5.** Error distribution of the prediction model training data.

4. Results

4.1. Anomaly Detection Results

| Detection Method          | Recall | Precision | F1    | Time  |
|---------------------------|--------|-----------|-------|-------|
| k-Means Clustering        | 0.94   | 1.00      | 0.97  | 1.0s  |
| Autoencoder               | 0.97   | 0.60      | 0.73  | 131.2s|
| Prediction Model Error    | 0.98   | 1.00      | 0.99  | 581.2s|
Table 1 presents the recall, precision, and F1 scores of the three anomaly detection algorithms along with the training time for each. Figure 6 displays a sample for each algorithm when they did or did not detect the randomly added anomalies.

K-means was able to identify most of the anomalous values in the study but could not identify some of the values when it was within a certain range of values, like dips in the middle of the day. This was likely caused by the determination of clusters being largely dominated by the demand, with other features such as outdoor air temperature and time having a smaller significance.

The autoencoder model had a higher recall than k-means but struggled with misidentifying normal values as anomalous data. The autoencoder model was sensitive to changes in daily pattern, and further inspection revealed that the hospital building operated at a noticeably different pattern in 2020 compared to previous years likely due to a combination of changes in typical operation and the COVID-19 pandemic. This can be seen in Figure 7 where the anomaly score trends upwards in 2020, which may have caused several misidentifications of normal values.

The prediction error model was able to detect almost all the anomalies in this study and had a near perfect score. This method was sensitive to sudden changes in results and was well suited for this study. Viability of this approach applied to other buildings is still unknown since the prediction model was developed for this particular building and will largely depend on the performance quality of the existing prediction model.

![k-Means Clustering](image1)

![Autoencoder](image2)

![Prediction Model Error](image3)

**Figure 6.** Sample detection results for each algorithm.

![Autoencoder – Change in anomaly score over time](image4)

**Figure 7.** Autoencoder – Change in anomaly score over time.
4.2. Prediction Accuracy Results

The three anomaly detection methods showed significant improvements to prediction accuracy compared to the result with no fixes to the test data. Both the autoencoder and the prediction model error methods showed very close performances to the manual pre-processing, which shows that either technique will be a viable option. This also means that while the autoencoder model overpredicted anomalies shown by its low precision score, this overprediction had little impact on the overall prediction accuracy. One concern is that the minimum EEP for all three models was lower than the manual pre-processing, which could indicate that some of the results are biased to favor a lower score for the prediction model, and overfitting has likely occurred.

Table 2. Prediction accuracy results.

| Method                    | RMSE (kWh/30min) | Min. EEP (%) | Max. EEP (%) | Average EEP (%) |
|---------------------------|------------------|--------------|--------------|-----------------|
| No fixes                  | 52.26            | 0.67         | 58.10        | 2.68            |
| Manual Pre-processing     | 21.45            | 1.05         | 7.93         | 1.73            |
| k-Means Clustering        | 23.05            | 0.58         | 12.01        | 1.84            |
| Autoencoder               | 21.99            | 0.55         | 12.29        | 1.71            |
| Prediction Model Error    | 21.68            | 0.58         | 11.88        | 1.72            |

5. Conclusion

The study showed encouraging results for detecting anomalous values and automating the data pre-processing operation using an autoencoder or prediction error model demonstrated similar performances to manual operations. This will allow smart BAS systems to predict and control building energy without interruption and reduced risk of unexpected operation due to corrupted data. The prediction model error method that can utilize an already existing model to detect anomalies is a promising approach, especially for buildings that have already implemented or are in the process of implementing a smart BAS. Such systems can easily apply the workflow without developing a new algorithm.

In future studies, the same methods will be applied to other demands to determine which technique will be appropriate for a robust automated data pre-processing system. Thermal load anomalies are likely harder to detect because they operate at a wider range of values so combination of the autoencoder model that can detect changes of patterns and the prediction error model for sudden changes will also be considered.

References

[1] Zhou K and Yang S 2018 5.11 Smart Energy Management Comprehensive Energy Systems pp 423-56.
[2] Hodge V and Austin J 2004 A survey of outlier detection methodologies Artificial Intelligence Review pp 85-126
[3] Fan C, Xiao F, Zhao Y and Wang J 2018 Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data Applied Energy 211 pp 1123-35.
[4] Chandola V, Banerjee A and Kumar V 2009 Anomaly detection: A survey ACM Computing Surveys 41
[5] Fan C, Liao Y and Ding Y 2019 Development of a cooling load prediction model for air-conditioning system control of office buildings International Journal of Low-Carbon Technologies 14 pp 70-5
[6] Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P and Schroff G 2016 LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection Preprint arXiv:1607.00148
[7] Klambauer G, Unterthiner T, Mayr A, and Hochreiter S 2017 Advances in Neural Information Processing Systems vol 30, ed I Guyon and U V Luxburg et al (Long Beach: Curran Associates) pp 971-80