# Data Cleaning for Prediction and its Evaluation of Building Energy Consumption

**Yun-Yi Zhang[a] , Zhen-Zhong Hu[b] , Jia-Rui Lin[a] and Jian-Ping Zhang[a]**

[a] Department of Civil Engineering, Tsinghua University, China
[b] Shenzhen International Graduate School, Tsinghua University, China
E-mail: hu.zhenzhong@sz.tsinghua.edu.cn

**Abstract –**

**Buildings consume a large amount of energy and a plenty of methods to mine into energy consumption data to aid intelligent management are proposed. However, the data quality issues are inevitable and the influence is lack of discussion. This paper proposed a data cleaning method combing threshold and cluster method. This paper also proposed an index to evaluate the accuracy improvement on big data prediction. A case study is conducted and it is found that the accuracy of data filling is not sure to agree with the improvement of prediction after filling.**

**Keywords –**

**Data Cleaning; Building Energy; Prediction**

## 1 Introduction

Building consume 40% of global primary energy [1] and researchers have found that systematic building energy management can help reduce energy usage by 5% to 30% [2]. Building energy management data contains a lot of useful information and energy consumption prediction is one of the core issues of building energy management, because it can realize pre-reservation and deployment of energy, find abnormal energy usage and assist intelligent decision-making [3]. However, due to the limitations of the hardware conditions of sensors and network transmission environment, the data quality issues is inevitable. According to our investigations of existing building energy monitoring platform, a large amount of data appears quality issues such as missing and anomaly. Although there are plenty of prediction models, the process of data cleaning actually requires considerable efforts [4].

Any prediction algorithm relies on reliable data input, otherwise it will be 'garbage in, garbage out'. Therefore, it is always necessary to clean the data before prediction. However, traditional manual data cleaning methods have problems such as heavy workload and incomplete inspection. Many scholars believe that based on the big data perspective, the quality of big data is different from traditional one. The goal of big data is the analysis or application, so the quality of big data should be to what extend do the data meets the requirements of big data analysis applications [5-7].

However, the current methods for building energy consumption data cleaning mostly are based on the data itself, which fails to make full use of the relationship between energy and the property of the building. The methods for data quality evaluation mostly are based on in-sample evaluation methods, which is not consistent with the application scenario [8].

In response to the problems above, this paper first points out the data quality issues in building energy management platforms. Aiming at data missing and anomalies, this paper then propose the methodology of semi-automatic anomaly recognition and data filling. Based on the application of energy consumption, this paper then proposes an evaluation indicator to show how the data cleaning promotes the accuracy of the prediction. A case study is finally provided to demonstrate the method in this paper and verify its feasibility.

## 2 Methodology

According to investigation into real data collected by sensors and stored in energy monitoring platforms, two major kinds of data faults are missing and sudden change. A sudden change occurs due to an overpass on the survey range of the sensor. Data missing may occur due to loss of internet connection. Based on investigation of existing building energy monitoring platform, we recognized data missing and anomaly as two major data quality issues to be solved in this paper.

The overall methodology for data cleaning and its evaluation for prediction of building energy consumption is shown in Fig 1.

In the data cleaning process, abnormal data are first recognized using methods based on threshold and clustering and the faulty data are eliminated from the data set. Then missing data are filled using methods

based on interpolation, regression and clustering.

In the data prediction process, three commonly used method: multivariable regression, time series analysis and artificial network are chosen and compared. RNN (recursive neural network) model is finally selected to compare the quality of data cleaning.

In the evaluation process, we tested the accuracy elevation using different data cleaning method when data missing rate ranges. And an evaluation indicator is proposed to compare between different data cleaning methods.

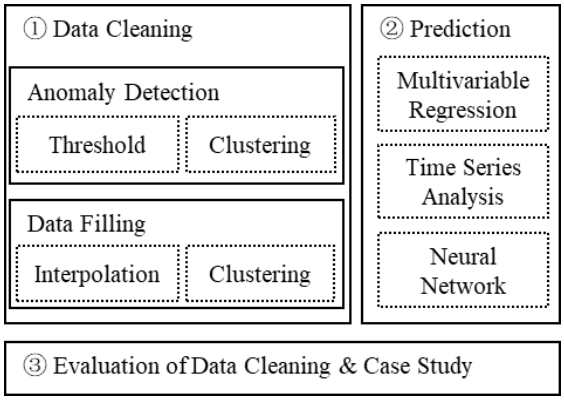


Figure 1. Overall Methodology

## 3 Data Cleaning

### 3.1 Anomaly Detection

According to literature review, anomaly detection for building energy data mainly consists of methods based on threshold and clustering. Our research combines these two methods hoping to get a more precise result.

#### 3.1.1 Threshold Method

A threshold can be selected in advance and used to eliminate all data that exceeds the threshold. This method can be used as preliminary screening. Threshold could come from meaning of the data, for instance, energy consumption cannot be negative, so zero is a lower bound of the data. Threshold can also come from experience or standard, i.e. if the data far surpasses the daily consumption, it is probably an anomaly. According to statistics, we also chose 3 times of standard deviation as a threshold, where 99.7% of the normally distributed data should fall into the range.

#### 3.1.2 Clustering Method

Energy consumption actually may show a mix of different consumption mode. Workdays show a high consumption and weekends show a low consumption. Figure 2 shows a disassemble of energy consumption mode using Gaussian mixture model, and the data shows a composition of two normal distributions. This indicates that we may need to first cluster the data into different groups and detect anomaly in each group, rather than regard the data set as a whole.

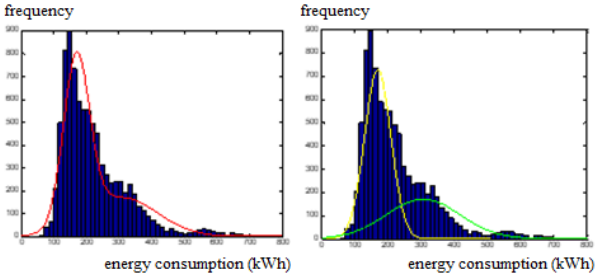


Figure 2 Disassemble of energy consumption mode

In this research, k-means algorithm is applied to perform data clustering, and threshold method is again used to detect anomaly data. Although the principle of k-means algorithm is simple, the hyper parameters are subjective and influences the final result. We implemented a process to automatically perform the detection and the workflow is shown in Figure 3.

Elbow point is used to determine the amount of clusters. For a cluster consisting of n sample data, loss function is defined as the sum of square of distance between data and the center. Different k values are selected and loss function decreases as k increases. We choose the k value which brings the most improvement as the final hyper parameter.

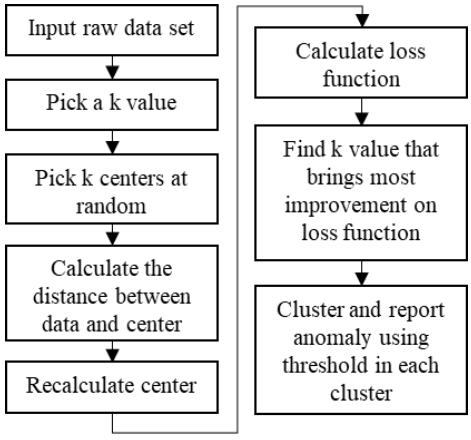


Figure 3. Workflow of data cleaning

### 3.2 Data Filling

After identifying anomaly data, some prediction algorithms can directly remove the missing data, but this method may make it difficult to make full use of some correctly recorded data, and some prediction algorithms require data recording to be continuous. In this regard, after removing anomaly data, we should

also try to fill in these data.

The clustering method uses the knn (k nearest neighbors) algorithm, and the core idea is that in the sample space, the category of a sample is directly determined by the k nearest samples. K nearest neighbors of a missing sample are found and the weighted average of these samples is used as the final filling result. The regression method uses multiple linear regression methods to construct the relationship equations between the monitoring data of different meters, and then calculates and fills in the missing data.

For dealing with missing data, most of the existing processing methods use historical data as a reference data set. For example, interpolation is a commonly used method. This method is simple to calculate, and results can be obtained quickly when there are not many missing values. Another example is the pattern analysis method, which analyzes the pattern of energy consumption data for the same period in the history of each year, and compares it with the energy consumption data before and after the missing value, finds the historical data with the same energy consumption pattern, and uses the weighted average method to fill in. This method also highly requires historical data. Moreover, these methods do not consider the actual meaning of the data, and cannot reflect the true situation when there are too many missing values.

Taking into account the relationship between BIM and monitoring data in the energy management platform, it is possible to identify the data of other functional areas that are similar in function, orientation, area, and energy use mode to the monitored functional area at this point, and different areas at the same time There should be a certain relevance in the energy use rules. If the area, location, function, orientation and other characteristics of the space are similar, the monitoring values of other spaces can be used to calculate the missing value of this point. Using the association relationship shown by the data model in the energy management data platform, if the data at a certain point is missing, you can make full use of other associated data to estimate it.

## 4 Data Prediction

Energy consumption prediction generally refers to the prediction of energy consumption in the future through the monitoring of historical data, in order to estimate the energy consumption that may occur in the future. Currently commonly used methods include multiple linear regression, time series analysis, and artificial intelligence prediction methods. A summary of comparison between these methods is shown in Table 1.

Multiple linear regression is to use of regression analysis to establish the relationship between a number of explanatory variables and the explained variables, so as to achieve the purpose of prediction. The explained variable is the future energy consumption monitoring value of the predicted point, and the explanatory variable usually has four main sources: (1) the historical energy consumption monitoring value; (2) data from other sensors related to the predicted sensor; (3) the physical attributes of the monitored space itself, such as area, function, orientation, etc.; (4) the time attribute, such as seasons, weeks, and holidays. The advantage of multiple linear regression is that it is easy to calculate, saves computing resources, and can quickly obtain an estimate of the predicted value, but the disadvantage is that the predicted result depends on the form of the artificially set regression equation, and the ability to handle the nonlinear relationship between variables is limited. Under the condition of abnormal or missing data, since the historical energy consumption monitoring value will appear as the explained variable, all samples containing abnormal or missing values cannot be used as learning samples to train the model, which may reduce the number of samples.

Time series analysis method is to use autoregressive, moving average and other methods to split the components of the time series into trend items, periodic fluctuation items and residual items, which can be used to analyze the nature of the data itself. The advantage of this method is that it can fully mine the laws of the data itself, decompose the components of the data, and obtain more information that can be understood. The disadvantage is that the time series analysis method has strict requirements for data integrity. Missing data will make the model unable to work. All vacant data must be filled in to make predictions. Therefore, for building energy consumption monitoring data with abnormal or missing data, this method needs to be used with extreme care, and attention must be paid to the sensitivity of the prediction results to the filling data.

Artificial intelligence prediction usually uses the artificial neural network that has developed rapidly in recent years to predict energy consumption. This prediction method has a good generalization ability and can be applied to various prediction problems. The prediction results of these methods are more accurate.

Table 1 Comparison between energy consumption prediction methods

| Method | Advantage | Disadvantage | Influence of missing data |
|---|---|---|---|
| Multi-variable regression | Simple model, easy to train | Only deal with linear relationship | Reduce sample volume |
| Time series analysis | Pattern is comprehensible | Strict requirement on data completion | Missing data have to be filled |
| Artificial intelligence | Most accurate | Complex model, time consuming | Reduce sample volume |

However, the disadvantage is that the complexity of the model is relatively high, so the computing resource requirements are relatively large, and the number of samples is also relatively high. Similar to the multiple regression method, the abnormality and lack of data will also make the samples with these data unable to be used as training samples, which will affect the training effect of the model.

# 5 Evaluation for Data Cleaning

## 5.1 Influence of Data Cleaning on Prediction

Although the prediction of building energy consumption monitoring data is a hot research topic, most studies use relatively complete and high-quality data sets. In fact, it is not reasonable to regard filled data to be the same as the original data. However, these studies did not explore how the filled data will affect the prediction results. Some literature points out that although data filling improves the integrity of the data set, it does not necessarily improve the effect of data prediction. When a small rate of data is missing, the data integrity has little impact on the prediction results, and data filling does not greatly improve the prediction results. When the data missing rate is moderate, data integrity has a great influence on the prediction results. At this time, data filling will greatly improve the accuracy of the prediction results. When the data missing rate is high, both data prediction and data filling lack corresponding training sets, so it is difficult to achieve better results, and it is difficult to improve the data prediction effect after data filling.

Although both the accuracy of data prediction and the accuracy of data filling decrease with the increase of the data missing rate, when the data missing rate is moderate, the improvement of the data prediction effect by data filling is the most significant. In the context of big data applications, the purpose of data filling is not to faithfully restore the original data of the data set, but to improve the data prediction effect through data filling. Therefore, simply comparing the accuracy of the data filling algorithm with original data sets is not the final goal. The data cleaning quality evaluation for data prediction should emphasize on the improvement of the prediction effect. The effect of data cleaning should be evaluated by comparing the prediction effect before and after the data cleaning, the effect of the data cleaning is improved [9].

## 5.2 Evaluation of Prediction

Methods of relative error and absolute error can be used to evaluate the prediction effect. In order to compare various algorithms between different samples, the relative error is often used to evaluate the prediction effect. Commonly used evaluation indicators are Mean Absolute Percentage Error (MAPE), R-squared [10] and Theil Inequality Coefficient (TIC) [11]. Among them, the value range of the MAPE is 0 to infinite, where 0 means a perfect model, and more than 100% means an inferior model. The value range of R-squared is 0 to 1, where 1 indicates a perfect model. The value range of TIC is 0 to 1 and the smaller the value, the smaller the difference between the fitted value and the true value.

Since the value range of the TIC is a limited interval, this research adopts the TIC as the evaluation index of the model's prediction effect. Since the evaluation of data cleaning quality for data prediction should start with the improvement of prediction effect, the effect of data cleaning should be evaluated by comparing the prediction effect before and after data cleaning. Therefore, on a certain data set, the difference in the TIC before and after data cleaning represents the improvement of the data cleaning effect on data prediction.

$$\text{Prediction Accuracy Improvement} = \text{TIC}_{\text{befroe}} - \text{TIC}_{\text{after}} \quad (1)$$

## 5.3 Evaluation Index for Data Cleaning

If we plot the prediction accuracy before and after data filling under different data missing rates, as shown in Figure 4, it can be seen from the figure that when the data missing rate is small, the loss of the TIC is not large. However, when the data missing rate is medium, the TIC increases rapidly, which indicates the rapid attenuation of the prediction effect. When the data missing rate is very large, the prediction accuracy remains at a very low level. The effect of data filling is to postpone the missing rate at the inflection point of the prediction effect attenuation, i.e., when the missing rate is medium, the data filling increases the number of samples available, it can also restore or maintain the data characteristics, so the prediction accuracy can be maintained at a high level at the same time. However, in the case of a high data missing rate, the characteristics of the data itself cannot be reflected, and there are not enough available samples to fill in, and the filling and prediction results are close to random.

To compare the improvement of the prediction effect of different data filling methods, we can compare the difference of the TIC under the condition of different data missing rates, and comprehensively consider the quality improvement under various data quality. Therefore, the area between the two curves can be used to characterize the improvement of the prediction result by the data preprocessing operation, and the calculation formula is as follows. The larger the area in the figure, it means that the data filling method can achieve better

data prediction effect improvement under various data missing rates. Therefore, the data cleaning effect evaluation index for data prediction can be defined as the following formula.

$$\text{Evaluation Index for Data Cleaning} = \int_{0\%}^{100\%} \left(TIC_{after} - TIC_{before}\right)\, d(missing\ rate) \quad (2)$$

## 6 Case Study

This research takes an office building Creative Plaza located in Dalian China as a case study. This building has a building height of 85 meters, with 16floors above ground and 2 floors underground, and has a total building area of 36,500 square meters. Each floor and each zone has a set of sensors to collect electricity consumption every 15 minutes since 2018. We take the electricity consumption of Zone A on the 11th floors from March to July 2020 as an example.

Figure 4 Creative Plaza building

### 6.1 Anomaly Detection

The raw data of 11th floor is shown in Figure 5(a) and we can see an obvious sudden change. The average of the data set is 494.57 and standard deviation is 7368.94. These two anomalies are eliminated using threshold method as in Figure 5(b). After removing the outliers, the data showed a stable fluctuating trend, but there was still a suspicious data anomaly.

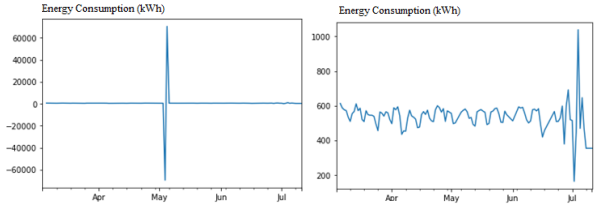


Figure 5. Electricity consumption data (unit: kWh) before (a) and after (b) threshold method

After the elimination of sudden change, we use k-means algorithm to implement the clustering method to detect abnormal data. The number of clusters k is iterated and the loss function under different k values is as shown in Figure 6(a). We can see that when k=3, the decrease in loss function is the maximum using elbow point method.

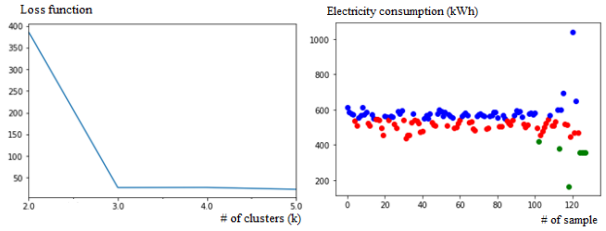


Figure 6 (a) Relationship between loss function and number of clusters (k); (b) Result of clustering

The clustering result divides the original data into 3 categories: high/ medium/ low energy consumption patterns. Anomaly detection is carried out for each type of monitoring data utilization principle. It can be seen that the two highest and lowest data points that deviate from the main trend are abnormal data. It is also caused by the jump of the monitoring meter and should be eliminated. The data after removing all abnormal data is shown in Figure 7, and the whole process is completed automatically.

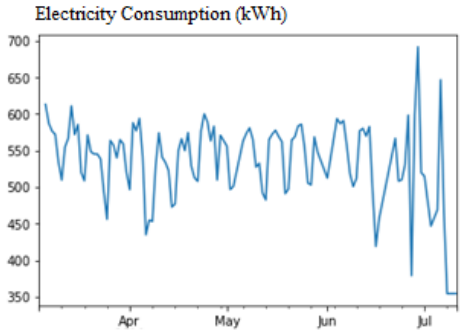


Figure 7. Data after data cleaning

### 6.2 Data Filling

To compare the accuracy of different data filling methods, some of the original data of the electricity consumption of the 11th floor was randomly removed to simulate the lack of data, and the remaining data sets were used for data filling. The error between the filling value and the monitored value is compared in the cases of different missing rates, and the accuracy of different repair methods is compared. There may be two typical cases of missing data. (1) missing at random: this situation often occurs when abnormalities such as the jump of the measurement meter are abnormal, and the missing data is at random. (2) continuous loss: this situation often occurs in abnormal situations such as power failure or stoppage of the measuring meter, or network transmission. Generally speaking, adjacent data

with random missing time can characterize the missing data to a certain extent, and the data law can still be retained. Continuous deletion may make data difficult to recover, and the laws hidden in the data are more difficult to analyze. The accuracy of data filling in these two scenarios are summarized in Table 2 and Table3.

Table 2 Data Filling Accuracy (Missing at Random)

| Missing Rate | Filling Accuracy | | |
|---|---|---|---|
| | Interpolation | Regression | Clustering |
| 10% | 90.29% | 95.09% | 94.03% |
| 20% | 86.25% | 94.64% | 92.84% |
| 30% | 85.88% | 94.24% | 92.67% |
| 40% | 84.25% | 93.77% | 92.55% |
| 50% | 79.06% | 92.99% | 91.59% |

Table 3 Data Filling Accuracy (Continuous Missing)

| Missing Rate | Filling Accuracy | | |
|---|---|---|---|
| | Interpolation | Regression | Clustering |
| 10% | 85.19% | 95.76% | 93.86% |
| 20% | 83.02% | 93.60% | 92.71% |
| 30% | 81.20% | 93.98% | 92.31% |
| 40% | 79.83% | 92.63% | 91.45% |
| 50% | 75.78% | 94.32% | 89.60% |

As can be seen in the tables, the accuracy of regression outperforms interpolation and clustering. In all three methods, the accuracy in continuous missing scenario is lower than missing at random and the influence is more obvious in interpolation method.

## 6.3 Data Prediction

In order to test the accuracy of the building energy consumption prediction algorithm, the 126 data sample are divided into two parts using in-sample test indicators. The first 112 sample points are the training set, and the last 14 sample points are the test set. The prediction result is compared with the last 14 sample points. and the TIC value as an index of prediction accuracy is calculated.

In this case study, considering the data availability, the input to the model consists of three parts. The first part is the daily historical data of power consumption in Area A on the 11th floor. Since the data exhibits obvious weekly fluctuations, the historical data of 7 previous days are used as training data, and the data of the following day is used as output data. The second part is the daily historical data of the power consumption of the 9th and 10th floor of area A. These two areas are related to the predicted power consumption of the 11th floor area A. The functions and energy use scenarios of these areas are similar. The third part is the properties of the space, including the area, floor, orientation and other information that can be obtained from BIM model. In time series analysis, ARMA model is applied. And for artificial network, RNN model with a hidden layer of 10 neurons is applied.

The prediction results of the three models using multiple regression model, time series analysis, and artificial neural network are shown in the figure below. The TIC values are 7.37%, 6.76%, and 6.32%, respectively. The prediction accuracies of the three models are all acceptable. Among them, the artificial neural network can handle the nonlinear relationship between variables, so the prediction The effect is best, as shown in Figure 8.

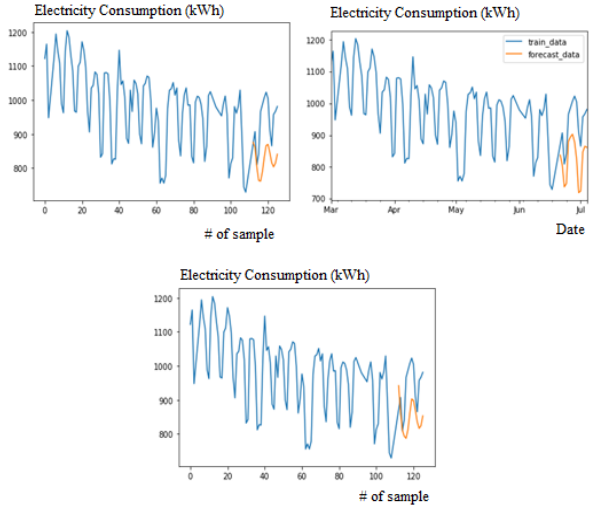


Figure 8. Prediction result of (a) multi-variable regression, (b) time series analysis, (c) artificial network (blue: real data, orange: prediction)

## 6.4 Data Cleaning Evaluation

In order to compare the influence of data filling on prediction, we eliminate different proportions of data and perform neural network on the remaining data after data filling with three different methods: interpolation, regression and clustering as is introduced above. We tested these methods in two data missing scenarios: missing at random and continuous missing. Two evaluation indices are applied: TIC of prediction and accuracy of filled data compared with real data. The results are summarized in Table 4 and Table.5 and is shown in Figure 9 and Figure 10.

Table 4 Evaluation of Data Cleaning (Missing at Random, TIC=6.32% when no data missing)

| Missing Rate | No filling | Interpolation | | Regression | | Clustering | |
|---|---|---|---|---|---|---|---|
| | TIC | Accuracy | TIC | Accuracy | TIC | Accuracy | TIC |
| 10% | 7.61% | 90.29% | 6.79% | 95.09% | 7.47% | 94.03% | 6.92% |
| 20% | 9.91% | 86.25% | 7.02% | 94.64% | 7.56% | 92.84% | 7.22% |
| 30% | 12.23% | 85.88% | 7.24% | 94.24% | 7.90% | 92.67% | 7.56% |
| 40% | 13.09% | 84.25% | 7.30% | 93.77% | 8.43% | 92.55% | 7.62% |
| 50% | 16.58% | 79.06% | 7.76% | 92.99% | 8.72% | 91.59% | 7.97% |

Table 5 Evaluation of Data Cleaning (Continuous Missing, TIC=6.32% when no data missing)

| Missing Rate | No filling | Interpolation | | Regression | | Clustering | |
|---|---|---|---|---|---|---|---|
| | TIC | Accuracy | TIC | Accuracy | TIC | Accuracy | TIC |
| 10% | 9.11% | 85.19% | 7.31% | 95.76% | 7.31% | 93.86% | 7.30% |
| 20% | 11.82% | 83.02% | 7.44% | 93.60% | 7.66% | 92.71% | 7.44% |
| 30% | 13.73% | 81.20% | 9.38% | 93.98% | 8.51% | 92.31% | 7.63% |
| 40% | 16.65% | 79.83% | 10.36% | 92.63% | 9.22% | 91.45% | 8.31% |
| 50% | 19.93% | 75.78% | 12.33% | 94.32% | 9.63% | 89.60% | 8.97% |

In this case, if we compare the accuracy of filled data, the accuracy of regression is the highest. However, using the data filled with clustering method, the prediction is more accurate, as is summarized in Table 6. Therefore in this case, the in-sample evaluation method does not agree with the prediction result, while the latter one is the real target of big data application.

Table 6 Data Filling Evaluation (Missing at Random)

| Filling Method | Average Accuracy of Data Filling | TIC Improvement after Data Filling |
|---|---|---|
| Interpolation | 66.96% | 12.28% |
| Regression | 82.66% | 23.15% |
| Cluster | 79.77% | 23.88% |

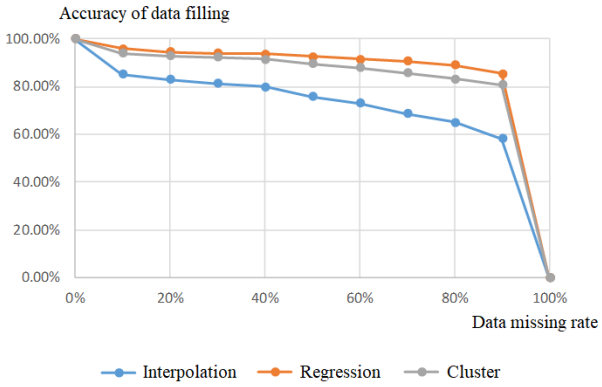


Figure 9 Accuracy of data filling in different data missing rates

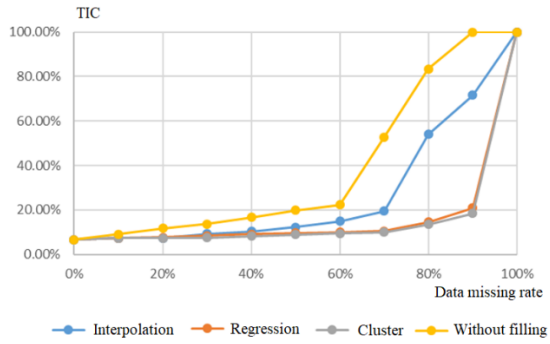


Figure 9 TIC of prediction after data filling in different data missing rates

## 7 Conclusion

Building consume a large amount of energy and a plenty of researchers have proposed methods to mine into energy consumption data to aid intelligent management, however, the data quality issues are inevitable and the influence is lack of discussion. This paper proposed a data cleaning method combing threshold and cluster method. This paper also proposed an index to evaluate the accuracy improvement on big data prediction. A case study is conducted and we found the accuracy of data filling is not sure to agree with the improvement of prediction after filling.

This research is a tentative discussion for the relationship between data cleaning and prediction for building energy consumption and there are several

issues for future work. For example, more data sources from BIM could be taken into consideration so that the relationship between these data could be recognized to improve the data cleaning. More data filling methods and prediction algorithms could be conducted to get a more comprehensive comparison.

## Acknowledgement

This research was funded by the National Key R&D Program of China (grant No. 2017YFC0704200) and the National Natural Science Foundation of China (grant No. 51778336). This research was also supported by Tsinghua University—Glodon Joint Research Center for Building Information Model (RCBIM).

## References

[1] Corry, E., O'Donnell, J., Curry, E., et al., 2014. Using semantic web technologies to access soft AEC data. Adv. Eng. Inform. 28 (4), 370–380.

[2] Costa, A., Keane, M.M., Torrens, J.I., et al., 2013. Building operation and energy performance: Monitoring, analysis and optimization toolkit. Appl. Energy 101, 310–316.

[3] Yun-Yi Zhang, et al. "Linking data model and formula to automate KPI calculation for building performance benchmarking." Energy Reports 7(2021):1326-1337.

[4] A, Imran Khan , et al. "Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques." Energy Procedia 42.1(2013):557-566.

[5] Ismael, et al. "A Data Quality in Use model for Big Data." Future Generations Computer Systems Fgcs (2016).

[6] Li, C. , and Y. Zhu . "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era." Data Science Journal 14.1(2015):21-3.

[7] Rao, D. , V. N. Gudivada , and V. V. Raghavan . "Data quality issues in big data." 2015 IEEE International Conference on Big Data (Big Data) IEEE, 2015.

[8] Y Ma, et al. "Study on Power Energy Consumption Model for Large-Scale Public Building." Intelligent Systems & Applications International Workshop on (2010):1 - 4.

[9] Li, P. , et al. "CleanML: A Benchmark for Joint Data Cleaning and Machine Learning [Experiments and Analysis]." (2019).

[10] Ericsson, N. R. . "Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration." Journal of Policy Modeling 14(1992).

[11] Cowell, F. "Theil, Inequality Indices and Decomposition." Research on Economic Inequality 13.1(2005).