Building Energy Consumption Data Detecting and Recovering Using Bayesian Method

Jun-qi YU\textsuperscript{1,2,*}, Ying TIAN\textsuperscript{1}, An-jun ZHAO\textsuperscript{1,2}, Yun-fei XIE\textsuperscript{1}, Xin-le HUANG\textsuperscript{1} and Lei-lei HUI\textsuperscript{1}

\textsuperscript{1}School of Building Services Science and Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, PR China

\textsuperscript{2}Smart City Research Center, Research Institute of New Urbanization and Human Settlement in Shaanxi Province 710055, PR China

* Corresponding author

Keywords: Building Energy Consumption, Data Detection and Recovery, Maximum Expected Value Algorithm, Naive Bayesian Algorithm, Sparse Bayesian Algorithm.

Abstract. Building energy consumption data plays an important role in building energy analysis and energy saving optimization. However, due to difficulties in collecting, high cost, equipment failure and other reasons, the collected data are prone to be missing, which hinders the mining and analysis of building energy consumption data. In this paper, the Bayesian network is used to check and recover the building energy consumption data. In the case that the amount of time series data missing is less than 50\%, the method G-test is selected to identify abnormal data, and the Naive Bayesian optimizing Expected Maximum Algorithm is used to check the data. When a large number of building energy consumption data missing, the Sparse Bayesian learning algorithm is used to fill in the missing data based on the compressed sensing theory. The results show that the model can effectively deal with the problem of missing data of building energy consumption and can be widely used in practical projects.

Introduction

In recent years, various countries began to vigorously develop building energy consumption monitoring. However, the building energy data is likely to face more variable factors such as sensor failure and the lack of monitoring data. So this situation caused great inconvenience for building energy management, hindered the digging of building energy consumption data analysis \cite{1}.

At present, there are three main processing methods for abnormal data of building energy consumption \cite{2}. First, directly discard the data with missing values. Second, replace all missing data. Thirdly, interpolation and maximum expected value (EM) in statistical theory are used to fill in the missing data. The first and second processing methods do not make full use of the original relevant data, and the effect is not good in actual engineering application. The third abnormal data processing method is closely related to the effective data quality. Although EM algorithm provides an iterative algorithm to calculate the posterior density function and has the advantage of maximum simplicity and stability, it also has the disadvantages of slow convergence speed and low efficiency \cite{3}. Therefore, this paper proposes a data detection and recovery method that combines the traditional statistical method maximization (EM) with machine learning algorithm Bayes to process and fill the missing data in building energy analysis, so as to form a complete data set and then conduct statistical analysis.
Building Data Detection Model of Bayesian Optimization EM Algorithm

Using Bayesian Optimization Method to Correct and Fill Building Energy Consumption Data

For building energy consumption, there are many factors causing the fluctuation of building energy consumption. However, if the influence of weather, personnel and the opening and closing of energy-using equipment on building energy consumption is regarded as a data set, building energy consumption is the specific expression of the output of the data set \[4\]. In order to fill in the missing data, the first step is to divide the data matrix into time series as shown in the table above, so as to identify each quantity and unknown quantity. Then, it is necessary to determine whether the sample miss rate is greater than 50%, and if so, the naive Bayesian optimization EM algorithm is used to check the data. However, this step of data filling does not take into account that the input data itself is missing or error data. When the data missing rate is large, this method is limited. Therefore, the Sparse Bayesian algorithm is introduced to fill in the multi-data deletion when the miss rate is greater than 50%. The specific algorithm is shown in figure 1.

![Figure 1. Framework of data identification model.](image1)

![Figure 2. Model test process.](image2)

For the data of abnormal building energy consumption, this paper selects the common G – test identification method in the medical field, and discriminates the data quality of building energy consumption by checking whether the values of observation variables meet the ratio expected by theory \[5\].

Track the maximum confidence degree of building energy consumption, judging whether it meets our minimum requirements, and using the expected maximum algorithm to correct the abnormal data that does not meet the requirements. In order to solve the expected maximum algorithm easy to fall into local optimum of faults, this paper on the basis of the EM algorithm is introduced into Bayesian network algorithm. The Naive Bayesian classification results using range as the EM algorithm is initialized, and then carried out in accordance with the step E and M iteration. After correction, the integrity of the data set used to analyze the available for the next step.

In the case of large-scale data loss, the above method is prone to local data loss and cannot accurately recover data. Therefore, this paper uses Sparse Bayesian method to fill the data of building energy consumption \[6\].

---

Loss Frequency <50%  Loss Frequency ≥50%

EM-NB: Identify and Correct Building Energy Consumption

Sparse Bayesian Time Series: Fill Building Energy Consumption

Complete Building Energy Data

Data Sample

Random Allocation

Training Dataset Without Missing
Testing Dataset Without Missing

Compares Recovery Models
Interpolation
Average
EM
Bayes Optimization

Set the Energy Consumption Missing Randomly

Testing Dataset with Missing

Identification Recovery Model

Restored Dataset

Calculate RMSE of each Algorithm

Averaged Algorithm Results

Repeat10 times

Interpolation
Average
EM
Bayes Optimization
Determination of Performance Evaluation Indicators

In order to distinguish the ability of building energy consumption data detection and recovery method to restore the original data, the root mean square error evaluation index and average running time are introduced. Based on the known information, the feasibility of the model can be judged by how close the missing value is to the real value. The smaller the deviation between the fill value and the real value, the better the method can restore the original data. Firstly, a certain proportion of samples are selected by random sampling from the real data set as the missing data set and the remaining samples as the training data set. Then, based on the training data set, interpolation filling method and average value method were used to fill the missing data set, and then the filling value generated by each method was compared with the real value to calculate the filling performance. In order to avoid the influence of different samples on the experimental results, the average value of the experiment was filled in for 10 times for the final results. In order to compare the restoring ability of each method to the original data, this paper introduces the evaluation criterion of root mean square error (RMSE) [6]. At the same time, in order to analyze the computational complexity of different building energy consumption data recovery methods, each method was repeated for 10 times to calculate the average running time (ART) [7], and it is the second evaluation index.

Simulation Example

Description of Building Energy Consumption Dataset

In this section, four groups of building energy consumption data are selected, and all of them came from the monitoring platform of Xi'an building energy consumption, which completely records the real-time energy consumption data of several public buildings in Xi'an. The data of building energy consumption includes 10 input variables such as building lighting socket power consumption, air conditioning power consumption, power consumption, special power consumption and time and 1 output of building energy consumption. This paper used the records of building energy consumption published by the platform from January 2017 to December 2017, and the actual number of samples used is 8,240. Four parts (day, week, month and year) were selected to constitute the data set. At the same time, the loss rate of building energy consumption is defined as the ratio of the amount of missing data to the total amount of data. According to different data loss rates, some columns of data were randomly deleted from the complete data set to simulate the missing data. The four sets of building energy consumption data were described in table 1.

| Databases | Size       | Property amount | Categories quantity | Missing items | Missing percentage |
|-----------|------------|-----------------|---------------------|---------------|-------------------|
| 1         | One day    | 4               | 10                  | No            | 0                 |
| 2         | A week     | 4               | 10                  | Yes           | 24.8%             |
| 3         | A month    | 4               | 10                  | Yes           | 52.6%             |
| 4         | A year     | 4               | 10                  | Yes           | 73.4%             |

Building Energy Consumption Data Filling Modeling

Property Information of Building Energy Consumption. For the target building, power monitoring system includes a total of 48 subsystems, including 4 subsystems at level 1 and 44 subsystems at level 2. The main power consumption of the office building includes cooling devices, boilers, lighting sockets in the office area, etc. At the same time, the energy consumption of the
electricity meter and the secondary electricity meter has a good matching degree (less than 2%), which proves that there is no hardware fault in the collection of energy consumption data.\textsuperscript{7}

The accuracy of Naive Bayesian algorithm is mainly determined by the integrity of sample data and the properties of sample attributes. The main attributes and data types of building energy consumption collection are divided into four attributes: energy consumption data, outdoor weather, indoor personnel and date attributes.\textsuperscript{8} Energy consumption data include lighting socket electricity, air conditioning electricity, power electricity and special electricity, while outdoor weather includes temperature, humidity and other factors. Energy consumption attributes of building can be filled according to needs.

**Experimental Process of Building Energy Consumption Identification and Recovery.** At the beginning of the building energy consumption identification and recovery model, based on whether the variables are discrete or continuous, the existing data is checked by the recognition method G-test, and the missing variable data values are initialized by using the expected value of naive Bayesian optimization. Then the no-miss samples were randomly assigned to the training set and the test set, and the building energy consumption data set was restored and reconstructed by manually setting the missing values in different proportions, which were repeated for 10 times and the time and root mean square error were calculated. The experimental process is shown in figure 2.

**Analysis of Experimental Results**

**Example of Building Energy Consumption Filling.** In order to prove the effectiveness of the algorithm model, the total output of building operation energy consumption in the week of November 6, 2017 solstice and November 12 was selected for data monitoring and filling example. The specific energy consumption is shown in figure 4. It can be seen from the figure that energy consumption fluctuates little within a whole week and shows a strong law. The high energy consumption of the building on Saturday and Sunday is due to the fact that the building is in non-working hours on weekends.

Besides, there are many educational institutions in the office building, so there is a large flow of people on weekends and sufficient space utilization. In this special case, the data observation points of 70-100 showed obvious unstable fluctuations. Abnormal data points in building energy consumption were checked according to the Bayes optimization algorithm, and the missing data points were filled in to obtain a complete data set of building energy consumption as shown in the figure 3.

![Figure 3. Example of building energy consumption filling.](image-url)
Analysis of Data Recovery Running Time Results. In order to study the effect of missing data types on algorithm performance, it is assumed that the data are randomly, uniformly or continuously lost. The performance of the proposed algorithm is best when the data is uniformly missing. The reason is that if the data is uniformly missing, then the observed data is also uniformly, so that we can obtain useful information for each period of time. In practice, however, the types of missing data tend to be random. The performance of building energy consumption recovery method is compared with other three algorithms: (1) Interpolation [3]; (2) Average value method [2]; (3) E-M method. We can observe that the performance of Bayes optimization method is still better in the case of random data missing, and even when data is continuously missing, the average running time generated by the proposed algorithm is still small compared with other algorithms.

According to the analysis, when the data set is small, the average running time of the interpolation algorithm is the shortest, because it is only a simple interpolation operation on the data. As the data set gets larger, the superiority of the method presented in this paper becomes prominent, while the running time of the interpolation algorithm becomes significantly longer. This is because the running time of the interpolation algorithm is the superposition of multiple time series, and the algorithm in this paper can simultaneously predict the missing values in multiple time series. Meanwhile, the Interpolation, Average and EM algorithms cannot check the data when the dataset is complete.

Comparative Analysis of Different Building Energy Consumption Filling Algorithms

For comprehensive analysis, it can be concluded that EM-NB data recognition and correction have advantages over interpolation and mean value method in terms of operation time and model accuracy. In other words, when the data is less than 50% missing, EM-NB model in Bayes optimization method is better than EM model and other models. When the data loss rate of building energy consumption is more than 50%, Bayes optimization method sacrifices time but has great advantages in data recovery precision. Therefore, with 50% as the criterion, EM-NB and sparse matrix model are selected in Bayes optimization to correct and fill the data of building energy consumption, which can greatly balance the calculation time and accuracy, and can be widely applied in practical engineering.

Summary

In this paper, the Maximum Expectation Algorithm, Naive Bayesian Algorithm and Sparse Bayesian Algorithm are combined to fill the building energy consumption data missing data. This method can make full use of the time-domain smoothness of time series and use compressed sensing theory to transform the modeling of missing data filling into multi-sparse vector recovery. Run data of office building in 2017 experiment, the simulation results show that even in the case of a large number of missing data, the Bayes optimization method also can effectively restore the whole time series, especially in large data set, data loss serious cases, this method is compared with the traditional interpolation method and average method and expectation method has obvious advantages, to achieve a high precision. The results show that the application of Bayesian optimization in data monitoring and recovery of office buildings is successful.

Acknowledgement

This research was financially supported by Shaanxi Key Research and Development Program No. 2017ZDL-SF-16-5 (entitled Key Technologies and Demonstration of Green Building Energy Management Analysis Platform Based on Big Data), National Key Research and Development Project of China No. 2017YFC0704100 (entitled New Generation Intelligent Building Platform Techniques).
References

[1] Zhao L, Zhang J L, Liang R B. Development of an energy monitoring system for large public buildings[J]. Energy and Buildings, 2013, 66:41-48.

[2] Agrawal S, Agrawal J. Survey on Anomaly Detection using Data Mining Techniques[J]. Procedia Computer Science, 2015, 60:708-713.

[3] Karlis, Dimitris. EM Algorithm for Mixed Poisson and Other Discrete Distributions[J]. ASTIN Bulletin, 2005, 35(01):3-24.

[4] He J, Zhuang F, Liu Y, et al. Bayesian dual neural networks for recommendation[J]. Frontiers of Computer Science (print), 2018(8).

[5] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1977, 39 (1):1-38.

[6] Zhang, Zhilin, Rao, Bhaskar D. Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning[J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5):912-926.

[7] Linda O, Wijayasekara D, Manic M, et al. Computational intelligence based anomaly detection for Building Energy Management Systems[C]//5th International Symposium on Resilient Control Systems (ISRCS), 2012. IEEE, 2012.

[8] Zhao H X, Frédéric Magoulès. A review on the prediction of building energy consumption[J]. Renewable and Sustainable Energy Reviews, 2012, 16(6):3586-3592.