Data-driven prediction models of multi-dimensional energy consumed in public buildings

Chen Yibo¹*, Berardi Umberto²

¹School of Civil Engineering, Zhengzhou University, Zhengzhou 450000, Henan, China
²Department of Architectural Science, Ryerson University, Toronto, Canada
* yb_chen77@163.com

Abstract. Due to the intense pressure from energy shortage and environmental protection, an accurate prediction of building energy consumption is crucial for different energy conservation applications and policies. Besides simulation models and traditional statistical approaches, a data-driven modelling based on energy records provides new opportunities for predicting the building energy demand. This research is conducted based on the whole procedure of data mining with limited datasets, by making use of machine learning techniques and mathematical statistics. Especially, regarding the temporal and the architectural scales, models can be categorized into the short-term prediction, medium-term prediction and long-term prediction of classified energy consumptions, which also represent different modelling characteristics derived from mass data, limited data and poor data respectively. During the modelling process, the fuzzy C-means clustering and the interdisciplinary Lorenz curve were utilized to recognize different energy patterns. Afterwards, models of the nonlinear Support Vector Regression, the Grey model and the traditional polynomial regression were utilized respectively to output the predicted sequence. In summary, with datasets in current energy platforms, this paper presents a study of data-driven models based on energy records considering the nonlinear and uncertain features of different multi-dimensional models.

1. Introduction
The building sector is considered as the major contributor of world climate change [1]. As detailed illustrated in [2], China has become the second largest building energy consumer in the world, with more than 20% being consumed by public buildings. Consequently, the planning and management of building energy consumption have attracted increasing attentions in many countries [3]. Due to the intense pressure from energy shortage and environmental protection, the energy policy in China has been gradually transferred to the double-control of energy intensity and energy quantity [4]. As a result, the accurate and reliable prediction of building energy consumption is believed to be the crucial basis of reasonable energy planning and optimized building operation [5].

Current prediction approaches can be categorized into three types: physical simulations, statistical analysis and hybrid methods [6], which were classified in [7] as white box method, black box method and grey box method respectively. With different modelling mechanics, each approach has its corresponding applicable scope [8]. Firstly, physical models take building design parameters as the input, so that the hourly energy consumption and load sets are exported [9]. Regarding simulation tools, there are already some conclusions in terms of comparison and reliability tests, for example, the ASHRAE Standard 140, or the Annex 21 of International Energy Agency. However, some studies have pointed out that the simulated results of prototypical buildings introduce uncertainties when compared with the results measured in real buildings [10]. At the same time, along with the increasing data and the development of smart metering, the statistical analysis, which pursues the reasonable level of energy consumption based on historical datasets of practical buildings, has attracted increasing attention. Especially, the data-driven models based on practical data records are proven to be an
efficient modelling approach, by making use of different statistical or intelligent techniques, such as the Artificial Neural Network and Support Vector Machine.

However, issues of confidentiality and unwillingness to share data in developing countries have seriously hindered the development of published databases and prediction methods. This is the reason why previous statistical studies in China were conditioned to relatively smaller sample sizes. Actually, reliable energy prediction modelling driven by actual data and information are still absent in China.

2. Methods
This study was conducted based on the whole procedure of data mining, which consists of three processes: data collection and pre-processing, establishment of multi-dimensional data-driven models, and the further application analysis of achieved models (figure 1). As the most important part, the data-driven modelling makes use of machine learning techniques, mathematical statistical methods and the transdisciplinary Lorenz curve in the field of economics. Specifically, the machine learning methods include pattern recognition and nonlinear regression.

**Figure 1.** The normalized modelling procedure based on the whole procedure of data mining.

**Table 1.** A summary of modelling characteristics of multi-scale building energy consumption.

| Characteristics | Temporal scale | Spatial scale | Data details | Modeling basis and influencing factors | Targets of data mining | Techniques of data-mining | Modeling targets |
|-----------------|----------------|---------------|--------------|----------------------------------------|------------------------|--------------------------|------------------|
|                 | Short-term & individual buildings | Medium-term and long-term & individual buildings | Regional prediction |                       |                        |                          |                  |
|                  | Hourly and daily prediction | Monthly and annually prediction | Annually prediction |                       |                        |                          |                  |
|                  | Individual buildings | Individual buildings | Regional buildings |                       |                        |                          |                  |
|                  | Huge data volume and different data types | Limited data volume and limited data types | Poor datasets | Large dependency from historical principles | Limited dependency from historical principles | Overall distribution Principles and the estimation of future planned scenarios | Regional percentiles |
|                  |          |                       |                          | Operation scenarios |                           |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
|                  |       |                       |                          |                          |       |                          |                   |
| Pattern recognition | Pattern recognition by grey correction analysis | Analysis via quota levels | A certain predicted value |                       |                       |                          |                  |
| by Fuzzy C-means clustering (FCM) | Improved grey interval prediction | Lorenz curve modelling | A predicted range and associated seasonal indexes |                       |                       |                          |                  |
| Multi-resolution wavelet decomposition | Curve fitting and parameter extraction | Regional distribution | Regional distribution principles |                       |                       |                          |                  |
From aspects of the temporal scale and the architectural scale, the multi-dimensional models can be summarized as the short-term prediction, the medium-term prediction and long-term prediction of regional or classified energy consumption, which represent different modeling characteristics derived from mass data, limited data and poor data respectively. Secondly, the normalized procedure of data-driven modelling includes data pre-processing, feature analysis, pattern recognition, principles mining and modelling, results validation and applicable analysis etc. Consequently, this paper presents a comprehensive study from the point of view of data-driven modelling, taken the nonlinear and uncertain features of different multi-dimensional models into consideration. Specifically, the establishment and modelling characteristics of three models are summarized in Table 1.

- For the first short-term prediction, the Fuzzy C-means (FCM) clustering was introduced to identify similar days and hours with homogeneous principle as a whole, by taking advantages of membership degrees $\mu_{ij}$ shown in equation (1) and equation (2). After this filtration, the nonlinear relationship hidden behind multi-dimensional parameters and hourly energy consumption could be figured out by Support Vector Regression (SVR) algorithm [11].
- For the second medium-term and long-term energy prediction, the pattern recognition of monthly and annual sequences was managed by grey correction analysis [12], followed by the regression of improved grey interval model.
- For the last regional energy prediction, the nonlinear regional distribution principles were quantified by introducing the interdisciplinary Lorenz curve.

Taking the hotels of a planned district in Beijing as case studies, the energy demand of planned hotels can be predicted following the proposed steps in Figure 2 [13].

$$\text{min } J(U, P) = \sum_{j=1}^{n} \sum_{i=1}^{\epsilon} (\mu_{ij}^n \|x_j - p_i\|)$$

$$s.t. \sum_{i=1}^{\epsilon} \mu_{ij} = 1$$

**Figure 2.** Stepwise modelling of the proposed regional prediction model [13].

3. Results
3.1. Results of short-term prediction for individual buildings
In this study, hourly prediction results are managed for a hotel building, the operation of which is more comprehensive and can be influenced largely by indoor behaviours. Validation of the hourly results was conducted by comparing with measured sequences. The hourly sequences of 21st August and 24th August were taken as the representative changing tendency respectively for stationary
sequence with smaller changing frequencies and non-stationary sequence with larger changing frequencies. The predicted results ($I_p$ series) and the measured electric sequences ($I_{m_{21}}$ and $I_{m_{24}}$) of these two days are given in figure 3.

![Electric power comparison for two days](image)

**Figure 3.** Comparison of the predicted results for the two days to be predicted.

In order to better clarify the role that FCM plays in the prediction accuracy, three kinds of prediction results were carried out, including the results with clustering for both similar days and hours ($I_{p\_d\_h\_21}$ and $I_{p\_d\_h\_24}$), the results with only days-clustering ($I_{p\_d\_21}$ and $I_{p\_d\_24}$), and the results without clustering ($I_{p\_21}$ and $I_{p\_24}$). These three types are in accord with different modelling organization strategies. Firstly, for the former two models with clustering, each predictor is established based on the clustered homogeneous training samples. While the last model takes 30 successive days, which are just in front of the day to be predicted, as the training input. Secondly, for the first model with two-step clustering, a common predictor and the regression equation is shared within a group of similar hours. While for the other two models, a specific prediction of each hour needs to be established for 24 hours respectively. With the error analysis, it can be summarized as follows.

- The results of 21th August (stationary subsequence) achieve quite satisfied accuracy with the smallest mean average percentage error (MAPE) of 3.82%. Meanwhile, the most accurate result of 24th August (non-stationary subsequence) can reach a reasonable MAPE of 9.72%.
- For 21th August, which is distinguished as the stationary Group IV, the hourly prediction with recognition of similar days ($I_{p\_d\_21}$) is validated to have the most accurate results. In addition, even the worst $I_{p\_21}$ is also proven to be reasonable comparing with the measured sequences.
- For the relatively non-stationary sequence on 24th August, the result after clustering for both days and hours ($I_{p\_d\_h\_24}$) is the most accurate, and the worst $I_{p\_24}$ is already unauthentic by giving large deviation from the measured series.

3.2. Results of medium-term prediction for individual buildings

As for the monthly and annual energy prediction, the Grey Model was adopted by considering the non-negligible uncertainty. As shown in table 1, the annual consumption pattern can be recognized by grey correction analysis, thus the predicted range of annual energy consumption can be obtained for each building, as depicted in figure 4. Besides, the monthly prediction of energy consumption can also be achieved with the help of Grey Model and associated seasonal indexes.

3.3. Results for regional buildings

As illustrated in table 1, this model makes use of empirical principles extracted from sample groups to benchmark the regional energy intensity. After the pre-processing, the valid samples in Beijing are taken as the data basis, and there is no obvious relationship between the building areas and the individual EUIs [14]. Consequently, based on the modelling steps in figure 2 and the sample information, the empirical formulae for different building types can be achieved [14].
Figure 4. The predicted range of annual energy consumption for each building sample.

\[
\begin{align*}
EUI_{i\text{-office-}BJ} &= \eta_{\text{office-}BJ} L(a)_{\text{office-}BJ} = 28.54s + 40.03 \\
EUI_{\text{reg-office-}BJ} &= \left[14.27s^2 + 40.03s\right]_{s_l}^{s_u} \\
EUI_{i\text{-mall-}BJ} &= \eta_{\text{mall-}BJ} L(a)_{\text{mall-}BJ} = 4.27s^2 + 39.47s + 74.24 \\
EUI_{\text{reg-mall-}BJ} &= \left[1.42s^3 + 19.74s^2 + 74.24s\right]_{s_l}^{s_u} \\
EUI_{i\text{-hotel-}BJ} &= \eta_{\text{hotel-}BJ} L(a)_{\text{hotel-}BJ} = 13.89s^2 + 8.31s + 73.92 \\
EUI_{\text{reg-hotel-}BJ} &= \left[4.63s^3 + 4.15s^2 + 73.92s\right]_{s_l}^{s_u}
\end{align*}
\]

where \(EUI_i\) is the regional converted EUI (kWh/(m²·a)), \(s_l\) and \(s_u\) are the cumulative areas respectively for the pre-set limitations of \(EUI_i\) and \(EUI_u\) (m²).

4. Discussion and Conclusions
The results show that the data-driven models can provide reliable and accurate prediction of building energy consumption. However, it should be noted that the proposed models are established based on the assumption that the extracted features remain the same for the targeted buildings. This is a shared disadvantage of statistical methods. Thus, in order to guarantee the generalization, refined classifications and respective modelling are essential during the practical applications.

Unlike the previous statistical or physical models derived from multi-dimensional internal and external factors, this study proposed multi-dimensional data-driven models based on available monitoring records at different temporal and spatial scales. The databases of these models are available in current platforms of energy consumption monitoring in China. Besides, the intermediate results are empirical formulae, which are easily to be utilized by engineers for further application.
In summary, this study promotes the completeness of data-driven modelling of building energy consumption, and it also provides technical supports for the application of datasets in current platforms of energy consumption monitoring. Thus, starting from theoretical meanings, practical applicable values emerge.

Acknowledgement
The study was supported by the National Key R&D Program of China (Grant No. 2017YFC0704200). The authors also acknowledge State Grid Corporation of China for the supplied data.

References
[1] Berardi U 2017, A cross country comparison of building energy consumption and their trends, Resource Conservation and Recycling, 123 230-241
[2] Eom J, Clark L E and Kim S H 2012 China’s Building Energy Use: a Long-term Perspective Based on a Detailed Assessment PNNL Rpt. PNNL-21073
[3] Xia J, Hong T, Shen Q, Feng W, Yang L, Lm P, Lu A and Bhandari M 2014 Comparison of building energy use data between the United States and China Energ. Buildings 78 165–75
[4] Zhang Y, He C and Tang B 2015 China’s energy consumption in the building sector: A lifecycle approach Energ. Buildings 94 240–51
[5] Marasco D E and Kontokosta C E 2016 Applications of machine learning methods to identifying and predicting building retrofit opportunities Energ. Buildings 128 431–41
[6] Yang T and Zhang X 2016 Benchmarking the building energy consumption and solar energy trade-offs of residential neighborhoods on Chongming Eco-Island, China Appl. Energ. 180 792–9
[7] Li Z, Han Y and Xu P 2014 Methods for benchmarking building energy consumption against its past or intended performance: an overview Appl. Energ. 124 325–34
[8] Xin Y, Lu S, Zhu N and Wu W 2012 Energy consumption quota of four and five star luxury hotel buildings in Hainan province, China Energ. Buildings 45 250–6
[9] Chalal M L, Benachir M, White M and Shrahily R 2016 Building planning and forecasting approaches for supporting physical improvement strategies in the building sector: A review Renew. Sust. Energ. Rev. 64 761–76
[10] Heiple S and Sailor D J 2008 Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles Energ. Buildings 40 1426–36
[11] Chen Y, Tan H and Berardi U 2017 Day-ahead prediction of hourly electric demand in non-stationary operated commercial buildings: A clustering-based hybrid approach Energ. Buildings 148 228–37
[12] Deng J L 1982 Control problem of grey systems Sys. Cntl. Ltr. 1 25–32
[13] Chen Y, Tan H and Berardi U 2018 A data-driven approach for building energy benchmarking using the Lorenz curve Energ. Buildings 169 319–31
[14] Chen Y, Wu J 2018 Distribution patterns of energy consumed in classified public buildings through the data mining process Appl. Energ. 226 240–51