Data-driven modelling of building retrofitting with incomplete physics: A generative design and machine learning approach

To cite this article: Haitao Yu et al 2023 J. Phys.: Conf. Ser. 2654 012053

View the article online for updates and enhancements.
Data-driven modelling of building retrofitting with incomplete physics: A generative design and machine learning approach

Haitao Yu\textsuperscript{1,3}, Kailun Feng\textsuperscript{2*}, Santhan Reddy Penaka\textsuperscript{2}, Qingpeng Man\textsuperscript{1,3}, Weizhuo Lu\textsuperscript{2}, Thomas Olofsson\textsuperscript{2}

\textsuperscript{1} Department of Construction Management, Harbin Institute of Technology, Harbin 150009, China
\textsuperscript{2} Department of Applied Physics and Electronics, Umeå University, 90187 Umeå, Sweden
\textsuperscript{3} Key Lab of Structures Dynamic Behavior and Control of the Ministry of Education, Harbin Institute of Technology, Harbin 150009, China

\*E-mail: kailun.feng@umu.se

Abstract. Building performance simulation (BPS) based on physical models is a popular method for estimating the expected energy savings from energy-efficient building retrofitting. However, for many buildings, especially older buildings, built several decades ago, an operator do not have full access to the complete information for the BPS method. Incomplete information comes from the lack of detailed building physics, such as the thermal transmittance of some building components due to the deterioration of components over time. To address this challenge, this paper proposed a data-driven approach to support the decision-making of building retrofitting selections under incomplete information conditions. The data-driven approach integrates the backpropagation neural networks (BRBNN), fuzzy C-means clustering (FCM), and generative design (GD). It generates the required big database of building performance through generative design, which can overcome the problem of incomplete information during building performance simulation and energy-efficient retrofitting. The case study is based on old residential buildings in severe cold regions of China, using the proposed approach to predict energy-efficient retrofitting performance. The results indicated that the proposed approach can model the performance of residential buildings with more than 90% confidence, and show the variation of results. The core contribution of the proposed approach is to provide a way of performance prediction of individual buildings resulting from different retrofitting measures under the incomplete physics condition.

1. Introduction
The construction industry becomes a major contributor to carbon emissions. Large shares of buildings in many countries are running into the retrofitting phase\cite{1}. Taking China as an example, most of the existing buildings in China were built at the end of the last century, and about 49\% of China's existing buildings were energy-inefficient\cite{2}. It indicates significant potentials to mitigate the energy use of buildings through energy-efficient retrofitting. Various measures, such as change to energy-efficient windows, additional insulation walls, adoption of energy-efficient Heat Ventilation Air Conditioning (HVAC) and so on, can be adopted for building retrofitting. Thus, it is significant to develop tools to estimate the energy performance of retrofitting measures and support decision-making.
At present, the building performance simulation (BPS) based on physical models is the main tool to estimate energy performance from a set of retrofitting measures. Examples of BPS tools include EnergyPlus, Ecotect, and TRNSYS. Using BPS tools requires detailed building attributes such as building detailed designs, HVAC design information, operation schedules, climate, and solar/shading information[3]. However, such complete information may not always be accessible, especially for older buildings that need retrofitting but were built several decades ago[4]. The incomplete information can be caused by many factors for these older buildings, including incomplete HVAC documentation, different design documentation systems, or component deterioration over time.

Several big datasets for building performance have been established, such as the Energy Performance Certification (EPC) developed by European Union, and the Building Performance Dataset (BPD) developed by the United States[5]. These big datasets make it possible to evaluate the building retrofitting performances based on data-driven approaches under incomplete information conditions[6]. Previous studies have indicated that the data-driven approach can learn from big data for building performance predictions[3, 7] and for missing data imputation in wide fields[8].

The scope of this study is to propose a data-driven approach to handle the building incomplete information. It integrates the backpropagation neural networks (BRBNN), fuzzy C-means clustering (FCM), and generative design (GD). The GD technology generates the required big database of buildings. It will overcome the problem of incomplete information during building performance simulation, because based on the big dataset it can directly model the connections of available building attributes, retrofitting measures, and retrofitting performance with BRBNN–FCM. The proposed data-driven approach supports estimating the energy performance of different retrofitting measures under incomplete information conditions.

2. Methods

The proposed data-driven approach contains two sequential modules, (1) building performance database generation module and (2) retrofitting performance prediction module (see Figure 1). Specifically, at the first module the generative design is used to generate building retrofitting solutions with available building attributes and different retrofitting measures. Then, EnergyPlus platform are used to assess corresponding building retrofitting energy performance to construct the required big database, which offers data support for the following modelling sections. After retrofitting samples generation by the first module, the second module aims to use FCM-BRBNN integrated method to model the retrofitting performance by establishing the relationship of retrofitting solutions and their performances, under incomplete building physics.

![Figure 1. The overall framework of the proposed data-driven approach.](image-url)
2.1. Building performance database generation module

Generative design is an emerging technique to derive different design solutions by performing calculations based on certain fixed parameters. It can generate building information models (BIM) automatically through computer aid[9]. In this study, Dynamo visual programming based on Autodesk Revit is introduced to construct the generative design platform. This platform can automatically generate various building retrofitting solutions by inputting available building attributes and identified retrofitting measures. It comprises the following three sections: import section, process section and export section. The import section selects retrofitting measures and the undecided parameters randomly, which is used as import data for the process section. The process section is the core section of this module. It generates different building retrofitting solutions and corresponding building information models in Revit by combining import data randomly. Finally, the export section outputs IDF files that contain detailed building retrofitting solutions and CSV files for recording each building parameters and retrofitting measures.

In this study, EnergyPlus is selected as the building retrofitting energy performance evaluation tool. Building energy simulation with EnergyPlus requires the input of a EPW weather file and a building energy retrofitting solution file in IDF format. The EPW file is EnergyPlus platform-specific weather file that contains hourly weather data for typical weather year, which is used to simulate weather conditions at the building site. The EnergyPlus website provides EPW weather files available for download from numerous sources, such as CSWD, IWEC, TMY, etc., covering many countries and regions around the world. The IDF file provides detailed information of the building retrofitting solutions, including building geometry, space layout, material properties of building components, HVAC system parameters, indoor equipment, building occupant schedules and so on. Energy consumption simulations are performed in the EnergyPlus platform to obtain the energy performance of building retrofitting solutions, resulting in the energy performance database of building retrofitting solutions.

2.2. Retrofitting performance prediction module

Performance modelling is the core function of the second module that can extract the knowledge from the established energy performance database and make modelling of retrofitting’s performance. A Fuzzy C-means clustering (FCM) and Bayesian regularization backpropagation neural networks (BRBNNs) integrated method is proposed in this study to conduct performance modelling of building retrofitting. Machine learning is a set of powerful algorithms in data science that can learn regression and classification from data[10]. The BRBNN is used as a machine-learning algorithm to extract the knowledge and make modelling on building attributes, retrofitting, and corresponding performance such as energy-savings from the established energy performance database. Firstly, FCM is used to cluster the samples based on the features of buildings. Then the BRBNN is used to build a baseline prediction of building performance by establishing the relationship of building attributes, retrofitting measures, and energy performances (see Figure 2). The baseline prediction model will be used as the performance baseline for prediction intervals calculation and modelling. Then the prediction intervals of each cluster and sample are calculated. Based on these results, another BRBNN is used to model the retrofitting performance by establishing the relationship between retrofitting measures and performances prediction intervals (PIs).

In this study, two important metrics, prediction interval coverage probability (PICP) and mean prediction interval (MPI), are introduced to test the reliability and validity of a PI modelling. A reliable PI modelling can predict a target output with a defined confidence level (e.g., 90 or 95%). PICP is defined as the percentage of samples within prediction intervals to represent PI’s reliability. On the other side, an effective PI modelling means the proposed intervals are concentrated to provides meaningful information for decision-making. Based on this, MPI is introduced to represent the validity of PI as the average width of upper and lower intervals. The dataset of building retrofitting energy performance is divided into several sets. 70% of the samples are divided into 70%, 15% and 15% for training, validating,
and testing the BRBNN learning models of baseline, respectively. The remaining 30% of samples are used to test the final PICP and MPI performances of modelling.

![Building attributes and retrofitting measures](image)

Figure 2. The overall structure of BRBNN for performance modelling of building retrofitting.

3. Case study and results

The proposed data-driven approach was applied here to assess building retrofitting of an old residential building located in the north China. The north China is the early developed areas around 1980 to 2000, and there are many buildings need retrofitting after decades of operation. Therefore, a typical old residential building here is selected as study case. This case building is a six-floor and three-unit old residential building constructed in 1982. Based on this, we constructed the BIM model of this case building in Revit (see Figure 3). In addition, the basic information of this case building can be seen in Table 1.

![Energy savings](image)

Figure 3. The model of case building.
Table 1. Basic building information of the case building.

<table>
<thead>
<tr>
<th>Building length (m)</th>
<th>Building width (m)</th>
<th>Building height (m)</th>
<th>Number of units</th>
<th>Number of floors</th>
<th>Floor height (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64.05</td>
<td>11.75</td>
<td>18.80</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

According to the applicable energy-efficient retrofitting measures in selected cities, we finally select several common retrofitting measures considering the economic flexibility. The selected measures are: external wall insulation, roof insulation, replacing energy-efficient windows, replacing energy-efficient lighting system, heating pipeline insulation and thermostat installation. Based on the selected retrofitting measures and the above BIM model, we use established generative design platform to generate various building energy-efficient retrofitting samples. Then, the energy performances of the various retrofitting measures were simulated by EnergyPlus.

After obtaining building energy-efficient retrofitting samples and corresponding energy performances, the proposed FCM-BRBNN integrated approach will be used to establish the relationship of building attributes, retrofitting measures, and energy performances. The model parameters are set as follows. The number of sample clusters affects the width of intervals (MPI) and the validity of the model. The number of clusters for established energy performance dataset can be set as 5, according to the previous studies on residential building retrofitting[11, 12]. In this study, the input neurons are building attributes and retrofitting measures, with a total of 6 input neurons. Meanwhile, the output neuron is only one neuron that represents the energy savings. The number of neurons in the hidden layer is related to the complexity of the neural network, which can be approximately obtained from the number of input neurons and the number of output neurons by the empirical equation (1).

\[ h = a + \sqrt{m+n} \]  

(1)

Where \( h \) is the number of neurons in the hidden layer, \( a \in [1, 2] \), \( m \) is the number of input neurons, \( n \) is the number of output neurons, and the number of neurons in each hidden layer is set to be 5 in this study to ensure the reliability of modelling.

The number of neural network layers is one of the important factors affecting the complexity of neural network. Too many layers of neural network will make the model to be complex. In this study, we determine the number of neural network layers by comparing the root mean square error (RMSE) and the coefficient of determination \((R^2)\) under different neural network layers. Finally, the number of neural network layers is set to be 2 (i.e., 6-5-5-1).

After all the model parameters are set, the proposed FCM-BRBNN integrated approach is run on the case old residential building to model the energy-savings of retrofitting measures. The results of modelling have prediction interval coverage probability (PICP) as 92.2%. It indicates that the model can accurately predict the energy-savings of different retrofitting measures with 92.2% confidence. Meanwhile, the results of modelling have mean prediction interval (MPI) as 3.03e+05. The results validated the reliability and validity of proposed performance modelling method. The energy-savings modelling result is shown in Figure 4.
4. Conclusions and future work

The buildings that need retrofitting usually already have a long time of operations, resulting in the incomplete building information for retrofitting assessment and decision-makings. Without detailed building information, the BPS methods are difficult to accurately predict the energy performances of retrofitting measures. This study proposed an innovative approach from the perspective of data-driven to support retrofitting selection under incomplete information through performance modelling based on the generated retrofitting solutions database. The contribution is to make it not necessary to rely on the complete building information compared with BPS simulation. The study offers a tool for predicting the buildings’ performance from different retrofitting measures, enabling the building owners who normally have limited retrofitting knowledge and inaccessible complete building information to compare alternative retrofitting measures.

In the proposed data-driven approach, the generative design technology is introduced to automatically generate the database of building retrofitting solutions based on Dynamo visual programming and Autodesk Revit platform. EnergyPlus platform is used to obtain the energy performance of building retrofitting solutions, resulting in the energy performance database of building retrofitting solutions. In addition, the machine learning technology using FCM-BRBNN establishes the building performance model that presents the relationship between available building attributes and different retrofitting measures, overcoming the barrier that conventional BPS is not applicable with incomplete building information.

Although the proposed approach is promising, this research has limitations that can be addressed in future work. The developed approach assumes that most of the building big data is reliable. It will be valuable to perform a quantitative investigation of the data accuracy and reliability of building performance datasets. It is also interesting to explore how different algorithms will affect the proposed data-driven approach. Moreover, it will be challenging to apply the data-driven approach to areas without a local dataset. The emerging machine-learning approaches, such as transfer learning, might overcome this limitation. Finally, the proposed approach has not been validated by real building cases, which will be handled in future research.
Acknowledgments
The work is funded by the Formas project Enabling stakeholder engagement with sustainable energy renovation in detached houses (Dnr: 2020-02085), and the EU Horizon 2020 project AURORAL.

References