Energy-efficient retrofitting with incomplete building information: a data-driven approach

Kailun Feng[1*], *Weizhuo* Lu[1], *Santhan* Reddy Penaka[1], *Erik* Eklund[2], *Staffan* Andersson[1], and *Thomas* Olofsson[1]

[1]Dept of Applied Physics Electronics, Umeå University, 901 87 Sweden
[2]Umeå Municipality, Sweden

Abstract. The high-performance insulations and energy-efficient HVAC have been widely employed as energy-efficient retrofitting for building renovation. Building performance simulation (BPS) based on physical models is a popular method to estimate expected energy savings for building retrofitting. However, many buildings, especially the older building constructed several decades ago, do not have full access to complete information for a BPS method. To address this challenge, this paper proposes a data-driven approach to support the decision-making of building retrofitting under incomplete information. The data-driven approach is constructed by integrating backpropagation neural networks (BRBNN), fuzzy C-means clustering (FCM), principal component analysis (PCA), and trimmed scores regression (TSR). It is motivated by the available big data sources from real-life building performance datasets to directly model the retrofitting performances without generally missing information, and simultaneously impute the case-specific incomplete information. This empirical study is conducted on real-life buildings in Sweden. The result indicates that the approach can model the performance ranges of energy-efficient retrofitting for family houses with more than 90% confidence. The developed approach provides a tool to predict the performance of individual buildings from different retrofitting measures, enabling supportive decision-making for building owners with inaccessible complete building information, to compare alternative retrofitting measures.

1 Introduction

Large shares of buildings such as in European countries are running into the retrofitting phase [1]. Taking Sweden as an example, many buildings were constructed in the so-called Million Homes Program between 1965 and 1975. The residential buildings in Million Homes Program reach a 50-year service life and beyond, and approximately 80% of existing houses in Sweden have low levels of thermal insulation and many have ventilation systems with no heat recovery. It presents a good opportunity to mitigate the energy use of buildings through energy-efficient retrofitting. Various retrofitting measures, such as additional insulation walls, change to multilayer windows, adoption of energy-efficient Heat Ventilation Air Conditioning (HVAC) can be adopted for a retrofitted building. It is significant to develop tools to estimate the performance (e.g. energy savings and investment cost) of retrofitting measures and support the measure selection.

The building performance simulation (BPS) based on physical models is the main tool to estimate energy savings from a set of retrofitting measures. Examples of commercial BPS tools include EnergyPlus, eQuest, and Ecotect. The BPS requires detailed building properties such as building detailed designs, operation schedules, heating, ventilation. and air conditioning design information, the climate, and solar/shading information [2]. However, such complete information may not always be available, especially for older buildings that were constructed several decades ago with incomplete design documentation or with deteriorated components over a long time [3]. The incomplete information can be caused by many reasons, and the examples of them include different design documentation systems, incomplete heating, ventilation, and air-conditioning documentation, facilities, and component deterioration over time.

Previous studies usually set average or other default values for missing building properties [4]. However, deterministic evaluations with default settings obviously cannot provide ranges of uncertainties and risks associated with building retrofitting decisions [5].

To address the challenge, this research proposes a data-driven approach using Bayesian regularization backpropagation neural networks with fuzzy C-means clustering (BRBNN-FCM) to directly model the inner connections among available building properties, retrofitting measures, and corresponding performances. The case-specific missing information is imputed by principal component analysis and trimmed scores regression (PCA-TSR) through considering the relationship among building's characteristics, the missing information, and the entire knowledge described by big data. The proposed data-driven approach overcomes the barrier of incomplete building information in estimating the performance and selecting the retrofitting measures.

* Corresponding author: kailun.feng@umu.se

2 Methods

The developed data-driven approach comprises two sequential modules, (1) performance modelling and (2) data imputation (see Figure 1). In order to properly reflect the influence of incomplete information on buildings' performance, this method will model the performances by prediction intervals, i.e., the value ranges that future performance will locate in defined confidence.

Specifically, the first module, performance modelling consists of data cleaning that improves the reliability of building performance datasets (BPDs) and accuracy of performance modelling. After data cleaning, the second section in the first module is using BRBNN-FCM integrated method to model the building's retrofitting performances based on available building properties. After performance modelling by the first module, the second module aims to impute the case-specific missing data of retrofitting buildings by PCA-TSR method. The retrofitting buildings with imputed data will be input into performance modelling developed in the first module for retrofitting decision-making.

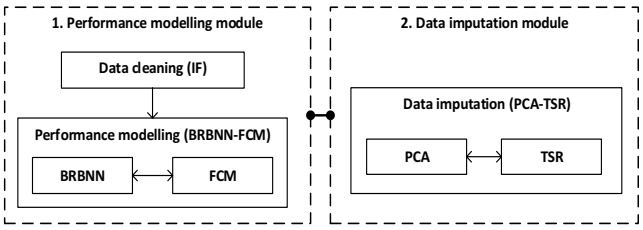

Fig. 1. The overall framework of the proposed data-driven approach.

2.1 Performance modelling module

Data cleaning is the crucial preparation for every data-driven approach, by which the quality of the dataset can be improved and the accuracy of data modelling is ensured. In data cleaning, the invalidity, duplication, anomaly of samples in BPDs are detected and removed. In this study, the Isolation Forest (IF) approach is used to detect the anomaly in building dataset, as it has an advantage in dealing with high-dimensional datasets [6].

Performance modelling is the core function of the first module that can extract the knowledge from BPDs' big data and make modelling of retrofitting's performance. A BRBNN and FCM integrated method is proposed in this study to make performance modelling of building retrofitting. Firstly, FCM is used to cluster the samples based on the features of buildings. And a baseline prediction of building performance is built with BRBNN by establishing the relationship of building properties, retrofitting, and corresponding 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 and performances prediction intervals.

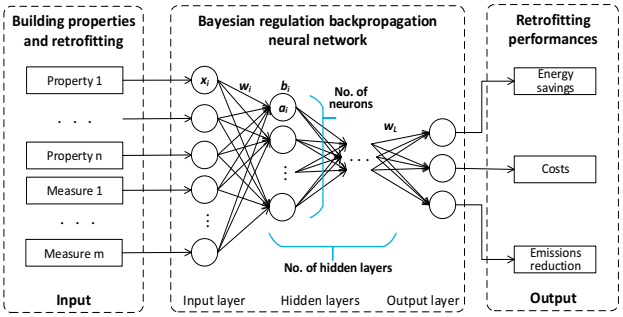

Fig. 2. The overall structure of BRBNN for performance modelling of building retrofitting.

2.2 Data imputation module

It is common that the building is built several decades ago and does not have full access to the complete building information when they need retrofitting. Therefore, it is necessary to supplement the above performance modelling module to address the missing information challenge. The data imputation module in this study will impute the case-specific missing information of buildings.

Therefore, this data imputation method will infer the most possible values for each building in terms of the missing data. The PCA and TSR integrated method that was originally proposed by [7] to address the missing data problem of PCA, is utilised in this study to impute the missing information of buildings that need retrofitting. Folch-Fortuny et al. [8] found that PCA-TSR outperformed other methods by providing the best solution in terms of prediction quality, robustness, and computation time.

Firstly, the missing data will be filled by initial values (usually as zero) in initiation. Then, the data matrix of BPDs will be estimated by PCA, while TSR is used to replace the initial missing data. And the new data matrix will be checked regarding the convergence, and the above missing data replacement iteratively runs until it is converged. The proper missing data of buildings is obtained after the matrix convergence. Therefore, this data imputation method exploits the correlations between variables to impute the case-specific missing data that considers the relationship between building characteristics and the whole knowledge, through considering the coherence of the latent structure of the matrix of BPDs.

3 Empirical study and results

In this study, the empirical shreds of evidence of available building performance datasets from observation records in real life are employed to test the effectiveness and efficiency of the proposed data-driven approach. This research chooses Sweden as the empirical study to validate the proposed method. As the detached and semi-detached house buildings for one/two family are the most popular housing type in Sweden, thus, the empirical study will focus on one/two family houses.

In Sweden, all newly-built buildings, rental, and not rental buildings, such as multi-dwelling stocks and

detached and semi-detached houses must have an energy performance certificate (EPC) according to regulations. In 2016, about 550,000 buildings have been certificated by EPC, which includes detached houses, apartments, and public, commercial or industrial facilities. EPC becomes one of the most informative big data on building performances in Sweden.

EPC comprises four crucial sections of information on the certificated buildings i.e., 1) the basic identification information, such as an address, post number, building types, and cities, 2) the second is the building properties, including building area, used purpose, number of floors, stairs, type of ventilation and more. 3) the building performance, such as electricity, fuel, natural gas consumption, and corresponding contribution from ventilation, heating, cooling, lighting, watering, and so on. 4) the suggested retrofitting measures based on the certification results. An inspiring inner knowledge that in the EPC big data can be expected, which is the inherent connection between building properties, retrofitting, and corresponding retrofitted performances.

After data cleaning, 223,232 one/two family house buildings with EPC records in the Swedish realm are obtained. The data imputation process is performed by PCA-TSR approach based on the big data knowledge from Swedish EPCs. Ten real buildings located from north Sweden (e.g. Norrbottens) to south Sweden (e.g. Skåne) are selected to validate the data imputation. The ten buildings are selected from all of the four different climate regions in Sweden to make the empirical study representative. For privacy, the selected real buildings and their details are anonymous here. The mean squared prediction error (MSPE) is applied to evaluate the accuracy of data imputation and it is 2.78×10^5. And the results of PCA-TSR are compared with the conventional Mean Imputation (MI) and Statistics Imputation (SI) methods. The MSPE of MI and SI methods is 1.66×10^6 and 3.23×10^6, which both are underperformed compared with the proposed PCA-TSR approach (see Table 1).

Table 1. The comparison of data imputation method.

Data imputation method	Mean squared prediction error	Value error
PCA-TSR method	2.78×10^5	-1/+28/+4/+0
Mean imputation method	1.66×10^6	-348/-319/+1
Statistics imputation method	3.23×10^6	-421/-394/-37/+0

After the data imputation, the proposed BRBNN-FCM integrated approach is run on Swedish one/two family house buildings to modelling the energy-savings of retrofitting, the results of modelling have prediction interval coverage probability (PICP) as 90.9%. It means that the model can have 90.9% confidence to accurately predict the energy-savings of different retrofitting measures. The energy-savings modelling results can be seen in Figure 3. The results validated the proposed performance modelling method.

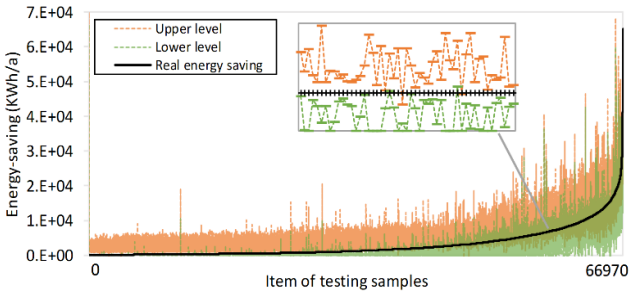

Fig. 3. Modelling of energy-savings for EPC testing samples.

4 Conclusion

The buildings that need envelope, HVAC, or other retrofitting usually already have a long time of operations, causing incomplete building information for performances estimation and retrofit selection. This research proposed a data-driven approach to support retrofitting selection under incomplete information. It provides a tool to predict the individual 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.

This research has limitations that can be addressed in future work. The developed approach assumes that the majority of the building's big data is reliable, except for the anomaly samples. It will be valuable to perform a quantitative investigation of the data accuracy and reliability of building performance datasets. In addition, proposed approach has not been validated by real building cases, which will be handled in future research.

The work is funded by the Formas project Enabling stakeholder engagement with sustainable energy renovation in detached houses, and the EU Horizon 2020 project AURORAL.

References

1. D'Oca S, Ferrante A, Ferrer C, Pernetti R, Gralka A, Sebastian R, Op't Veld P (2018): Technical, financial, and social barriers and challenges in deep building renovation: Integration of lessons learned from the H2020 cluster projects. Buildings **8**, 174
2. Amasyali K, El-Gohary NM (2018): A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews **81**, 1192-1205
3. Mathew PA, Dunn LN, Sohn MD, Mercado A, Custudio C, Walter T (2015): Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. Applied Energy **140**, 85-93
4. Ochoa CE, Capeluto IG (2009): Advice tool for early design stages of intelligent facades based on energy and visual comfort approach. Energy and buildings **41**, 480-488
5. Hiyama K, Kato S, Kubota M, Zhang J (2014): A new method for reusing building information models of past

projects to optimize the default configuration for performance simulations. Energy and Buildings **73**, 83-91

6. Chandola V, Banerjee A, Kumar V (2009): Anomaly detection: A survey. ACM computing surveys (CSUR) **41**, 1-58
7. Arteaga F, Ferrer A (2002): Dealing with missing data in MSPC: several methods, different interpretations, some examples. Journal of Chemometrics: A Journal of the Chemometrics Society **16**, 408-418
8. Folch-Fortuny A, Arteaga F, Ferrer A (2015a): PCA model building with missing data: New proposals and a comparative study. Chemometrics and Intelligent Laboratory Systems **146**, 77-88