

Decision tree model combined with Latin Hypercube Sampling to identify low energy optimal renovation strategies

Ke Li ¹, Tong Zhang ^{*1}

¹ School of Architecture, Southeast University, Nanjing, China

**Corresponding: Tong Zhang, hytong@seu.edu.cn*

Abstract

The problem of low-energy renovation of buildings is receiving considerable attention due to global warming. How to rapidly identify the effective renovating measures in the early design stage has become crucial. We propose a workflow based on decision tree models, incorporating LHS to reduce time. This workflow is designed to identify key design parameters quickly. It identifies their favorable ranges based on the Probability Density Function Plot (PDF) to guide building performance optimization. The proposed method is applied to a case study. The robustness of the results is evaluated by employing different sample sizes for the LHS approach. This approach is shown to be effective and to quickly handle high-dimensional spaces. It can accurately and efficiently identify key parameters and advantageous ranges with smaller sample size and simpler operation, facilitating the exploration of low-energy building renovation.

Highlights

- A workflow based on decision tree model, incorporating LHS was presented to reduce time and cost.
- The workflow can quickly identify the key design parameters and their advantageous range.
- Relationships between renovation strategies for building envelope and heating energy of the case were discovered.
- The robustness of the results is evaluated by employing different sample sizes for the LHS approach.

Introduction

Buildings consume a significant amount of energy and their operation has a significant impact on the environment, especially in this era of global warming. It's necessary to increase the annual global renovation rate for keeping cumulative carbon emissions of the global building stock in check (Ang et al., 2023). And consequently, how to rapidly identify the effective renovating measures in the early design stage has become crucial. These high-cost calculations are caused by a large number of parameters and samples, making the rapid identification of key design parameters a difficult challenge.

In recent years, there have been many studies on the key parameters about building energy demand and several methods were employed, such as traditional statistical analysis methods, optimization algorithms combined with sensitivity analysis methods, etc. Traditional statistical analysis methods are widely used to identify high-impact parameters in real building energy consumption data. (Lang and Lanz, 2022) implemented staggered difference-in-differences regressions to estimate how interventions in energy efficiency affects building-level CO₂ emissions based on data for a portfolio of 548 multi-unit buildings observed over 16 years. (Liang et al., 2018) use fixed effects panel regression models to investigate the effects of retrofits on energy consumption by electricity use and other covariates using pre-post treatment billing data from 2008 to 2013, covering 201 residential buildings and 636 commercial buildings. (van den Brom et al., 2019) used Kruskal–Wallis test instead of a traditional analysis of variance to test the importance about savings per renovation measure including nearly 10 0,000 renovated dwellings. (Filippidou et al., 2016) based on statistical modelling and data analysis to examine the energy efficiency measures and their effects on the energy performance of 757,614 households. The main advantage of regression models is that they are comparatively simple and efficient. However, the method is more applicable in situations where there is a large amount of measured data and samples, and it is not suitable for use in the early stages of building renovation. Much work so far has focused on sensitivity analysis methods combined with optimization algorithms, especially Genetic Algorithm (GA). (Saikia et al., 2020) research the thickness and the location of retrofit by using a new version of GA. (Awada and Srour, 2018) present a GA based framework to model the relationship between potential building retrofit options, the a sensitivity analysis was performed on the effect of each of these factors on the overall percentage improvement. (ZHANG Jingyu, 2020) analyze Spearman correlation coefficient between the parameters and building load based on the data obtained during the optimization process using GA in the grasshopper's plug-in. (Galimshina et al., 2021) identified the most influential parameters using Sobol' sensitivity analysis based on surrogate modelling, which combined NSGA-II and Gaussian process regression. (Li et al., 2018) identify the key design parameters for design optimization using a

multi-stage sensitivity analysis approach and the key design parameters of buildings are optimized to minimize the performance objective, using the GA. This method has a high level of prediction accuracy. However, due to the complex and time-consuming of the process, the models developed using these methods are not understandable and interpretable.

Applying classification algorithms for building energy analysis has the characteristic of simple logic. Due to its ease of use and ability to generate accurate predictive models with understandable and interpretable structures, the decision tree method is widely applied in many scientific and medical fields (Yu et al., 2010). (Namazkhan et al., 2020) developed a decision tree model for revealing the important factors related to household gas consumption based on the data collected from 601 households. (Jeffrey Kuo et al., 2018) build the decision trees (rules) classification rule model for discredited value of energy use to obtain the optimal thresholds factors. (Zhou et al., 2023) developed a Classification and Regression Tree model to quantitatively analysis the relationships between building energy flexibility and external variables of interest based on the data from the Building Management System over a period of three years. Most of the work so far has focused on the relationship between the characteristics of existing buildings and energy consumption, based on a large amount of measured data. However, there is little information on specific renovation measures for individual cases, and it still requires a significant amount of data.

In order to identification the key parameters rapider, (Li et al., 2019) used the Latin Hypercube sampling (LHS) method combined with NSGA-II to identify uncertain design inputs of significant impacts. (Tian et al., 2021) compared one time LHS with replicated LHS to obtain more stable energy predictions for buildings concerning six uncertain parameters. (Yue et al., 2021) used LHS combined with Multilayer Perceptron Artificial Neural Network to optimize a set of 14 variables as well as night ventilation and displacement ventilation strategies. LHS is a type of Monte Carlo simulation (MCS) that uses stratified sampling and requires fewer samples, which is beneficial for reducing the time required for optimization exploration. However, LHS has not yet been combined and applied with the decision tree model to validate its effectiveness in the field of building renovation.

In the early stages of building renovation, identifying key parameters and obtaining their optimal ranges can provide guidance for the design process. We propose a workflow based on decision tree models, incorporating LHS to reduce time consumption. This workflow is to quickly identify the key design parameters and their advantageous range to guide building performance optimization. We apply the method to a case and the robustness of the results is evaluated by employing different sample sizes for the LHS approach. The approach provides simpler operation and require less time about identify key design parameters and their

optimal ranges based on Probability Density Function Plot (PDF), facilitating the exploration of low-energy building renovation during the early design stage.

Methods

Workflow and tool of low energy design optimization

The proposed approach involves the integration of tools from the environmental simulation domain with LHS methods, decision tree model, parametric modelling techniques, and sensitivity analysis. The simulation and analysis methods were tested on a case study. The workflow (Fig. 1), integrates VSCode, GhPython, and EnergyPlus through the Climate Studio package. Initially, input variables and their ranges need to be defined. A LHS technique is employed to obtain a set of sample combinations for the design parameters. To generate the sample sets, (Matala, 2008) recommended using 10 times the number of variables as an acceptable number of training datasets for LHS sampling. In this study, 24 design variables were considered, and 300 sample sets were generated using the LHS method. These sample sets were then used for automatic simulation in EnergyPlus via the Climate Studio package.

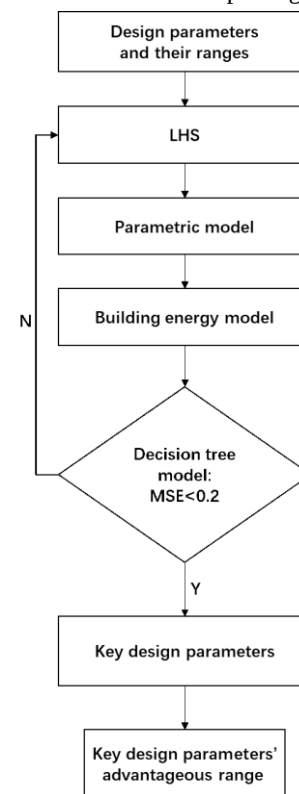


Figure 1: workflow of the methodology

Identification of key design parameters' advantageous range based on PDF

Based on Probability Density Function Plot (PDF), the optimal ranges of key parameters can be determined by considering their range of influence. When a normal distribution is observed, the "three standard deviations rule" can be applied to the optimal ranges of key parameters. According to this rule, for a data set that follows a normal distribution, approximately 68% of data points will fall within the range of the mean plus or

minus one standard deviation. Therefore, the range that includes the peak value plus or minus one standard deviation will contain approximately 68% of the data points. Similarly, the range of the mean plus or minus two standard deviations will contain approximately 95% of the data points, and the range of the mean plus or minus three standard deviations will contain approximately 99.7% of the data points.

However, when the PDF indicates a skewed distribution, the range can be adjusted according to the distribution of the data. In such cases, the following formula can be used to calculate a range that achieves a data proportion of 68%:

$$\text{Range} = [\text{Mean} - \text{Standard deviation} \times k, \text{Mean} + \text{Standard deviation} \times k] \quad (1)$$

Here, k is a constant that can be determined based on the degree of skewness and peak position of the data distribution. Generally, when the skewness coefficient is greater than 0.5, the value of k can be appropriately increased, for example, $k = 1.5$ or 2 . On the other hand, when the skewness coefficient is less than -0.5 , the value of k can be appropriately decreased, for example, $k = 0.5$ or 1 .

To obtain a more accurate range of key parameters, this study used the range corresponding to a data proportion of 68% as the optimal ranges to provide reference for design schemes.

Impacts of different sample sizes of LHS approach

To investigate the efficacy of identifying key parameters and optimizing their ranges using various sample sizes, this study employed sample sizes of 300, 500, 750, and 1000 for analysis. The distribution status of essential parameters was evaluated for each sample size, and the peak axis and standard deviation of these parameters were compared using PDF, the overlap index of advantageous ranges was calculated to determine the effect of different sample sizes on the results.

Case Study

Located in the Sipailou Campus of Southeast University in Nanjing, Jiangsu Province, the building under consideration was initially constructed in the 1980s. The building comprises a large experimental workshop that adopts a reinforced concrete frame structure, prestressed steel reinforced concrete truss, and independent foundation. Presently, the structural condition of the building is good, and the columns in the large experimental workshop can withstand the heavy load of the truss.

During the building renovation stage, renovation methods typically include envelope structure, interior space, and system equipment. Improving the exterior of old buildings is the main potential for energy savings (Lechtenböhrer and Schüring, 2011). Renovation strategies for building envelope structure depend on the specific situation and renovation objectives of the building, and common measures include external wall insulation, window replacement,

roof and floor insulation and waterproofing performance improvement, and structural reinforcement.

During the renovation process, no changes will be made to the primary structure of the building. The renovation measures mainly include the addition of an insulation layer to the external structure, updating window components, and adding shading to the windows. The objective of these measures is to enhance the building's performance and reduce the annual energy consumption demand.

This article discusses a total of 24 relevant parameters of the building, sourced mainly from ASHRAE Handbook 2005 - Fundamentals (Engineers, 2009). The specific variables are shown in Table 1.

Table 1: range of parameters

	Parameters	Range	Index
Wall insulation	Conductivity	[0.001,0.52]	0
	Density	[0.05,140]	1
	Specific heat	[750,1800]	2
	Thermal absorptance	[0.1,0.9]	3
	Solar absorptance	[0.1,0.9]	4
	Visible absorptance	[0.1,0.9]	5
Roof insulation	Conductivity	[0.001,0.52]	6
	Density	[0.05,140]	7
	Specific heat	[750,1800]	8
	Thermal absorptance	[0.1,0.9]	9
	Solar absorptance	[0.1,0.9]	10
	Visible absorptance	[0.1,0.9]	11
Ground screed	Conductivity	[0.2,1.4]	12
	Density	[1050,2350]	13
	Specific heat	[920,1500]	14
	Thermal absorptance	[0.1,0.9]	15
	Solar absorptance	[0.1,0.9]	16
	Visible absorptance	[0.1,0.9]	17
Window	Shgf	[0.09,0.99]	18
	Uval	[0.3,7]	19
	Tvis	[0.05,0.9]	20
Shading	Shading transmittance	[0,1]	21
	Tilt angle/ π	[0,1]	22
	Shading' length/window' height	[0,1]	23

Results and Discussion

key design parameters and their advantageous range

The mean squared error (MSE) of the decision tree regression model is 0.131, indicating that the model has good predictive performance on the data it was trained on. Therefore, the study can proceed with the next step of sensitivity analysis. We quantitatively identified the most influential parameters using Delta Moment-Independent Measure in python. The method is based on the 1st order Minkowski distance between the unconditional and conditional probability density functions (PDFs) of the output (Borgonovo, 2007). It allows for the estimation of both distribution-based sensitivity measures and of sensitivity measures that look at contributions to a specific moment. As shown in the sensitivity analysis of a sample size of 300 (Fig. 2), it is evident that two parameters, namely 22 and 23 of the index, have a significant impact on the building performance simulation, while the influence of the remaining 22 parameters decreases significantly.

As shown in sensitivity analysis of 300 sample size (Fig. 2), which present that two parameters, namely 22 and 23 of index, have a significant impact on the building performance simulation, while the influence of the remaining 22 parameters decreases significantly.

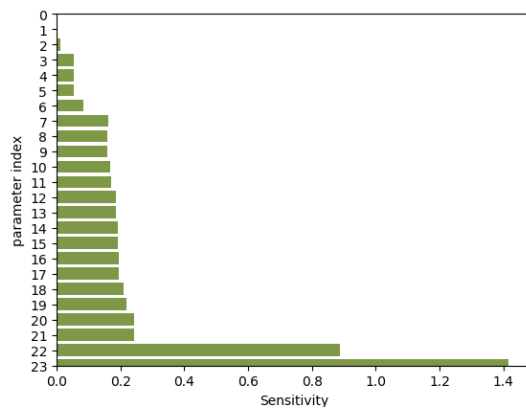


Figure 2: sensitivity analysis of 300 sample size

Based on the box plot (Fig. 3a) of the angle of the shading, it is observed that the mean is smaller than the median, and the skewness coefficient is calculated as -0.86, indicating that the data distribution is negatively skewed. According to the probability density distribution plot (Fig. 4a), the peak of the probability density distribution curve occurs at an angle of 118 degrees for the shading. However, when data is negatively skewed, the range of 68% of data is usually not symmetrical. Therefore, the range of the shading angle was adjusted using a k value of 0.5, resulting in a recommended range of (0.53, 0.79) for the shading angle, meaning that an angle between 95 and 143 degrees relative to the window is more suitable.

Similarly, based on the box plot of the length of the shading (Fig. 2b), it is observed that the mean is larger than the median, and the skewness coefficient is calculated as 1.42, indicating that the data distribution is

positively skewed. According to the probability density distribution plot (Fig. 4b), a shading width that is equal to the window and a length of 0.02 times the window height is more suitable. When the data distribution is positively skewed, the k value is set to 0.15 for adjustment, resulting in a recommended range of (0, 0.26) for the shading length, meaning that it is better to install a shading whose length does not exceed 0.26 times the window height to reduce the heating energy consumption of the building.

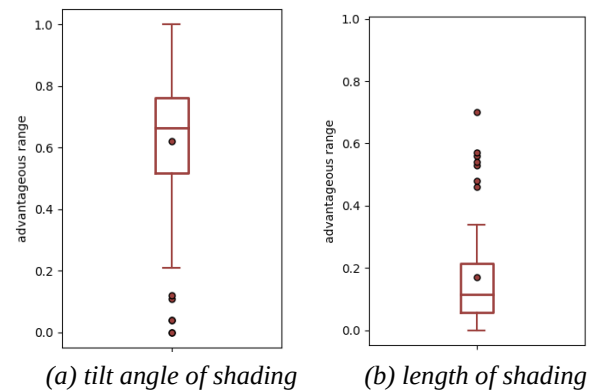
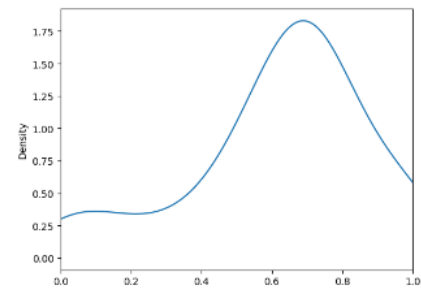
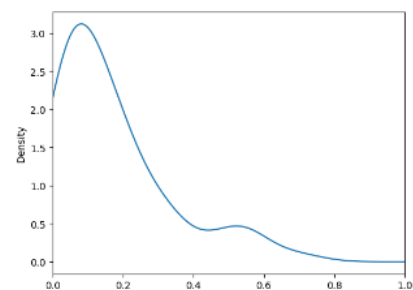


Figure 3: boxplot of key parameters



(a) tilt angle of shading



(b) length of shading

Figure 4: PDF of key parameters

Impacts of different sample sizes of LHS approach

When the sample size was 300, 500, 750, and 1000, the MSE for decision tree induction analysis was less than 0.2, indicating a well-constructed model for each sample size.

The sensitivity analysis results of each parameter under different sample sizes are similar, and the identification results of key parameters are consistent (Fig. 5). As the

sample size increases, the sensitivity of non-key parameters gradually decreases, with the difference between non-key parameters and key parameters becoming increasingly clear.

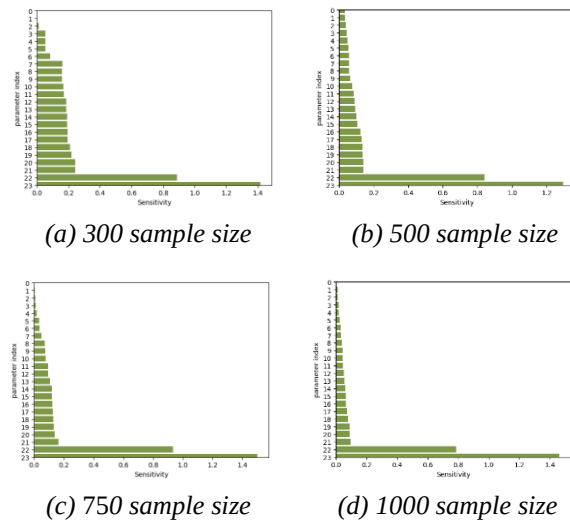
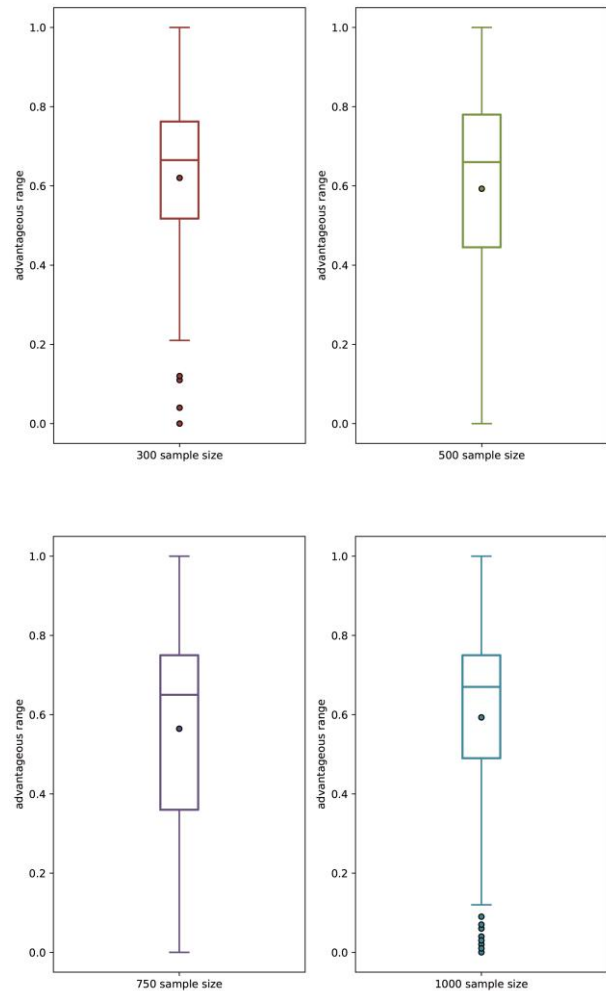


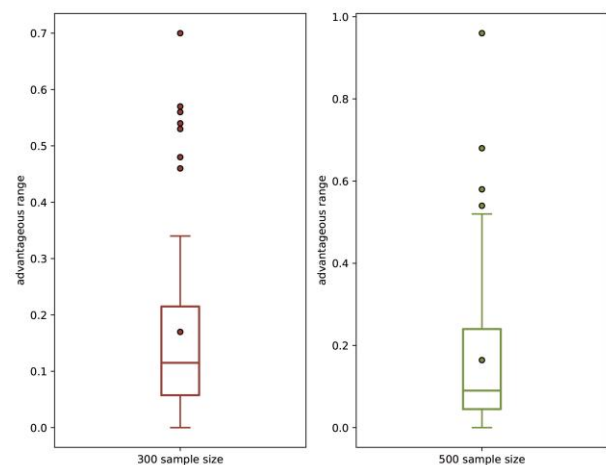
Figure 5: sensitivity analysis of different sample size

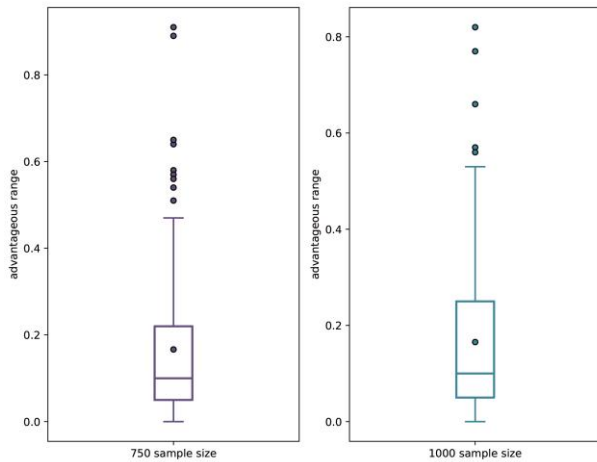
Regarding the recommended range of the tilt angle of the shading, the range becomes more dispersed as the sample size increases to 750. However, as the sample size further increases, the recommended range becomes more concentrated. The median fluctuation amplitude of the samples is relatively small, and the distribution remains consistent, showing a clear skewed distribution.

Comparing the boxplots of the tilt angle of the shading for different sample sizes (Fig. 6a), the mean of the data for all four sample sizes is less than the median, indicating a negative skewness of the data distribution. According to the probability density distribution graph (Fig. 7a), the peak line changes with the increase of the sample size, and it is generally around 0.66, which corresponds to the peak line of the sample size of 300. The corresponding skewness coefficients are -0.86, -0.79, -0.65, and -0.86, respectively. A k value of 0.5 can be chosen to adjust the data range. The corresponding recommended ranges of the tilt angle of shading are (0.53, 0.79), (0.53, 0.79), (0.52, 0.80), and (0.53, 0.79), respectively. It was calculated that the overlap index of each range exceeds 0.74, indicating that they share more than two-thirds of the information and have a high degree of intersection.



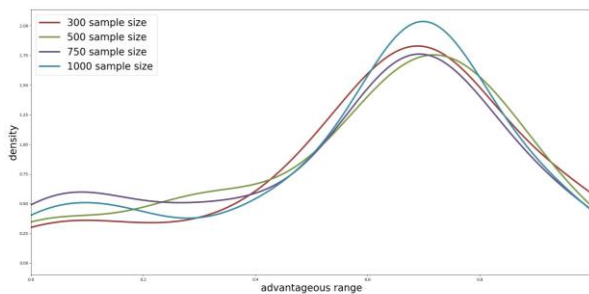
(a) tilt angle of shading



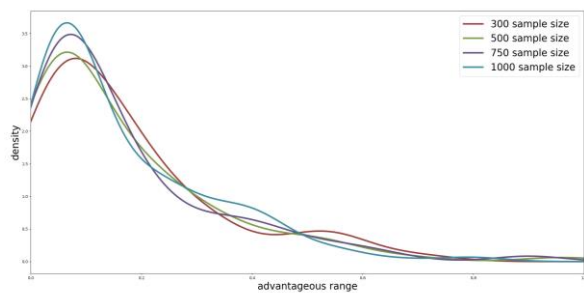


(b) length of shading

Figure 6: boxplot of key parameters in different sample size



(a) tilt angle of shading



(b) length of shading

Figure 7: PDF of key parameters in different sample size

Regarding the recommended range for the length of the shading, as the sample size increases, the median of the samples fluctuates relatively little and the distribution remains consistent, exhibiting a clear skewness.

Comparing the box plots (Fig. 7a) of the tilt angle of shading at different sample sizes, the mean of the data for all four sample sizes is greater than the median, indicating a positive skewness of the distribution. According to the probability density distribution plot (Fig. 7b), the peak value changes very little with increasing sample size, with the abscissa values corresponding to the peak point being 0.02 for the sample sizes of 300, 500, and 750, and 0.01 for a sample

size of 1000. The skewness coefficients are 1.42, 1.89, 1.82, and 1.42, respectively. A value of 1.5 can be chosen as k to adjust the data range. Based on this parameter's range of values, the corresponding range of the ratio of shading' length to window's height are (0, 0.26), (0, 0.28), (0, 0.28), and (0, 0.26), respectively. After calculation, the overlap index between any two of the four datasets is 1.0, indicating a high degree of similarity between them.

Conclusion

This text presents the results and discussion of a study on the sensitivity analysis of key parameters for building performance simulation. The study used a decision tree regression model and LHS approach with different sample sizes to identify the key parameters and their recommended range for reducing the heating energy consumption of a building.

The results showed that two parameters had a significant impact on building performance simulation, while the influence of the remaining 22 parameters decreased significantly. It is recommended that the tilt angle of shading relative to the window should range from 95 to 143 degrees. The recommended range for the length of the shading should not exceed 0.26 times the window height when shading' width that is equal to the window.

The study also found that the sensitivity analysis results of each parameter under different sample sizes were similar, and the identification results of key parameters were consistent. As the sample size increased, the sensitivity of non-key parameters gradually decreased, with the difference between non-key parameters and key parameters becoming increasingly clear. The recommended range for the tilt angle of the shading became more dispersed as the sample size increased to 750, but it became more concentrated as the sample size further increased. The recommended range for the length of the sun visor exhibited a clear skewness, with little median fluctuation and a consistent distribution as the sample size increased.

Overall, the study provided valuable insights into the key parameters and their recommended range for reducing the heating energy consumption of a building. The results could be used to inform the design and construction of energy-efficient buildings.

Acknowledgement

This study is supported by the National Key R&D Program of China (2022YFE0208600) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_0267).

References

- Ang, Y.Q., Berzolla, Z.M., Letellier-Duchesne, S., Reinhart, C.F., 2023. Carbon reduction technology pathways for existing buildings in eight cities. *Nat Commun* 14, 1689. <https://doi.org/10.1038/s41467-023-37131-6>
- Awada, M., Srour, I., 2018. A genetic algorithm based framework to model the relationship between

- building renovation decisions and occupants' satisfaction with indoor environmental quality. *Building and Environment* 146, 247–257. <https://doi.org/10.1016/j.buildenv.2018.10.001>
- Borgonovo, E., 2007. A new uncertainty importance measure. *Reliability Engineering & System Safety* 92, 771–784. <https://doi.org/10.1016/j.ress.2006.04.015>
- Engineers, R.A., 2009. 2009 ASHRAE handbook: fundamentals. *Ashrae Handbook Fundamentals*.
- Filippidou, F., Nieboer, N., Visscher, H., 2016. Energy efficiency measures implemented in the Dutch non-profit housing sector. *Energy and Buildings, Towards an energy efficient European housing stock: monitoring, mapping and modelling retrofitting processes* 132, 107–116. <https://doi.org/10.1016/j.enbuild.2016.05.095>
- Galimshina, A., Moustapha, M., Hollberg, A., Padey, P., Lasvaux, S., Sudret, B., Habert, G., 2021. What is the optimal robust environmental and cost-effective solution for building renovation? Not the usual one. *Energy and Buildings* 251, 111329. <https://doi.org/10.1016/j.enbuild.2021.111329>
- Jeffrey Kuo, C.-F., Lin, C.-H., Lee, M.-H., 2018. Analyze the energy consumption characteristics and affecting factors of Taiwan's convenience stores—using the big data mining approach. *Energy and Buildings* 168, 120–136. <https://doi.org/10.1016/j.enbuild.2018.03.021>
- Lang, G., Lanz, B., 2022. Climate policy without a price signal: Evidence on the implicit carbon price of energy efficiency in buildings. *Journal of Environmental Economics and Management* 111, 102560. <https://doi.org/10.1016/j.jeem.2021.102560>
- Lechtenböhmer, S., Schüring, A., 2011. The potential for large-scale savings from insulating residential buildings in the EU. *Energy Efficiency* 4, 257–270. <https://doi.org/10.1007/s12053-010-9090-6>
- Li, H., Wang, S., Cheung, H., 2018. Sensitivity analysis of design parameters and optimal design for zero/low energy buildings in subtropical regions. *Applied Energy* 228, 1280–1291. <https://doi.org/10.1016/j.apenergy.2018.07.023>
- Li, H., Wang, S., Tang, R., 2019. Robust optimal design of zero/low energy buildings considering uncertainties and the impacts of objective functions. *Applied Energy* 254, 113683. <https://doi.org/10.1016/j.apenergy.2019.113683>
- Liang, J., Qiu, Y., James, T., Ruddell, B.L., Dalrymple, M., Earl, S., Castelazo, A., 2018. Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix. *Journal of Environmental Economics and Management* 92, 726–743. <https://doi.org/10.1016/j.jeem.2017.09.001>
- Matala, A., 2008. Sample Size Requirement for Monte Carlo—simulations using Latin Hypercube Sampling.
- Namazkhan, M., Albers, C., Steg, L., 2020. A decision tree method for explaining household gas consumption: The role of building characteristics, socio-demographic variables, psychological factors and household behaviour. *Renewable and Sustainable Energy Reviews* 119, 109542. <https://doi.org/10.1016/j.rser.2019.109542>
- Saikia, P., Pancholi, M., Sood, D., Rakshit, D., 2020. Dynamic optimization of multi-retrofit building envelope for enhanced energy performance with a case study in hot Indian climate. *Energy* 197, 117263. <https://doi.org/10.1016/j.energy.2020.117263>
- Tian, W., Jin, M., de Wilde, P., Fu, X., Li, G., 2021. Uncertainty analysis of building energy analysis based on replicated Latin Hypercube sampling, in: *Building Simulation. Presented at the Building Simulation 2021, IBPSA*, pp. 1303–1310. <https://doi.org/10.26868/25222708.2021.30246>
- van den Brom, P., Meijer, A., Visscher, H., 2019. Actual energy saving effects of thermal renovations in dwellings—longitudinal data analysis including building and occupant characteristics. *Energy and Buildings* 182, 251–263. <https://doi.org/10.1016/j.enbuild.2018.10.025>
- Yu, Z., Haghighat, F., Fung, B.C.M., Yoshino, H., 2010. A decision tree method for building energy demand modeling. *Energy and Buildings* 42, 1637–1646. <https://doi.org/10.1016/j.enbuild.2010.04.006>
- Yue, N., Li, L., Morandi, A., Zhao, Y., 2021. A metamodel-based multi-objective optimization method to balance thermal comfort and energy efficiency in a campus gymnasium. *Energy and Buildings* 253, 111513. <https://doi.org/10.1016/j.enbuild.2021.111513>
- ZHANG Jingyu, 2020. A parametric approach for performance optimization of residential building design in Beijing. *Building Simulation*.
- Zhou, X., Du, H., Sun, Y., Ren, H., Cui, P., Ma, Z., 2023. A new framework integrating reinforcing learning, a rule-based expert system, and decision tree analysis to improve building energy flexibility. *Journal of Building Engineering* 71, 106536. <https://doi.org/10.1016/j.job.2023.106536>