Building stock energy modeling: Feasibility study on selection of important input parameters using stepwise regression

Reza Arababadi\textsuperscript{1} | Hariharan Naganathan\textsuperscript{2} | Mohsen Saffari Pour\textsuperscript{3} | Atefeh Dadvar\textsuperscript{4} | Kristen Parrish\textsuperscript{5} | Oswald Chong\textsuperscript{6}

\textsuperscript{1}Department of Energy, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran
\textsuperscript{2}Department of Construction Management, School of Management, Wentworth Institute of Technology, Boston, MA, USA
\textsuperscript{3}Department of Mechanical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
\textsuperscript{4}Department of Built Environment and Energy Technology, Linnaeus University, Växjö, Sweden
\textsuperscript{5}Del E. Webb School of Construction, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, USA
\textsuperscript{6}Ira A. Fulton Schools of Engineering, School of Sustainable Engineering and the Built Environment, Tempe, AZ, USA

Abstract
Building energy assessment is essential to accomplish the sustainable energy targets of new and present buildings. Retrofitting of the existing buildings by assessing them through energy models is the most prominent method. Studies revealed that there is still blank information about the building stocks, and these affect the valuation of building energy efficiency policies. Literature also recommends that the existing energy models are too complex and unreliable to predict the energy use. Reliability of such energy models would improve through a better alignment of the input parameters and the model assumptions. The authors hypothesized that the reliability of models would be improved through identification of the most relevant energy use parameters for the building stocks in different regions and models. One of the most commonly accepted methods for detecting the most dominant input parameters is sensitivity analysis, though its shortcomings include the need for a large number of data samples and long computing time. In this research, the Energy, Carbon, and Cost Assessment for Buildings Stocks (ECCABS) model is adopted to identify the most important parameters of the presented model. The research team uses the model that has been validated by studies conducted for the UK building stock. Moreover, by assessing the feasibility study with the stepwise regression to identify the significant input parameters have been discussed. Results show that stepwise regression method could produce the same results compared to sensitivity analysis. This paper also indicates that stepwise regression is considerably faster and less computationally intensive compared to common sensitivity analysis methods.

KEYWORDS
building stock, ECCABS model, energy modeling, input parameters, sensitivity analysis, stepwise regression
1 | INTRODUCTION

Existing building stock energy models approximate the baseline energy used for the building stocks to predict the projection of demands for future. There is a fact that the buildings are usually built for the long-term purposes. This will amplify the damaging impacts that the buildings exert on the environment. Most softwares are developed as tools for the preliminary design stages. Therefore, the detailed design phase of their developments is applied as the building performance assessment tools after the final design or complete construction. The ability to rely solely on the natural ventilation for cooling is limited by the climate and thermal loads. The energy models are dependent on the availability of data from buildings and are not verified or tested for their relevance and reliability. Thus, there is a disconnected bridge between the reality and assumptions. Most of the building stock energy models are shaped according to the bottom-up engineering models and require assumptions at several conditions. The accuracy of the building models widely relies on the quality of information and the selection of the input parameters.

Sensitivity analysis is one of the commonly accepted techniques to identify the relevant parameters and components for the models. The reason swirls around its reliability and ease of understanding, which is clearly described in literature. However, the process often forces to make assumptions due to the lack of information and requires high processing capacity to run the models. Also, the high time consumption and the complexity of outputs are few other drawbacks of such technique.

The objective of this research and its contribution to the body of knowledge is assessing the feasibility of applying stepwise regression instead of sensitivity analysis in building stock energy modeling. This promotes rapidness to the models and elevates model utilization. For this reason, the authors have adopted the existing Energy, Carbon, and Cost Assessment for Building Stocks (ECCABS) model, which was previously applied to simulate the UK building stock with 192 samples of residential buildings. The authors selected ECCABS since it is a successful tested model, and it assesses the effects of the energy saving measures (ESMs) for the building stock. The model is a bottom-up engineering model, which means that calculation of the energy demand of a sample of individual buildings is based on the physical properties of the buildings and their energy use (eg, for lighting, appliances, and water heating), and results are scaled-up to represent the building stock of the region studied. The ECCABS model is designed to have limited complexity and limited inputs which facilitates data extraction from available databases and to facilitate short calculation times. The primary outputs of the model are the net energy demand by end uses, the delivered energy to the buildings, CO₂ emissions, and the costs associated with the implementation of ESMs.

The model assumes the buildings as the representative of a region that was evaluated. The paper utilizes the data from the model and mines the feasibility of identifying the important input parameters. It helps in reducing the unwanted or irrelevant input information that increases the complexity of the model and simulation time. Also, reducing the parameters eliminates unavailable information and increases the data quality. Two of the techniques (ie, sensitivity analysis and stepwise regression) are tested with restricted number of parameters, and results are tabulated with justification on the best technique in following parts of the paper.

In summary, the aim of this paper is to compare the results (input parameters), which are derived from the sensitivity analysis (ECCABS Model), with the identified results from the stepwise regression. Also, it discusses the advantages of latter over former. Specific objectives are categorized into:

1. Identify various issues of sensitivity analysis documented in prior studies.
2. Find important parameters through stepwise regression for the available ECCABS model.
3. Identify the relevant input parameters and compare with results from the sensitivity analysis.
4. Document how the stepwise regression mitigates the issues of sensitivity analysis.

2 | RELATED WORKS

This part includes three subsections that contain a more detailed literature review on the existing energy models, utilization of sensitivity analysis on energy models and its drawbacks, utilizing the stepwise regression and its advantages. Finally, it has been tried to present the gap in prior studies and how that drives the research methodology of this paper.

2.1 | Energy models

For energy model calculations, various tactics including building energy simulation, statistical analysis according to the calculations, and experimental measured data can be performed to analyze the energy use of the buildings. In initial energy modeling approach, forward energy simulations are applied, which can be classified as the degree-day method, bin method, or hourly computer simulation tools to predict the annual energy consumption and costs.

Kavgic et al compared different assessments of previous models, which are developed based on the building stock in the UK. In such models, heat balance and empirical relationships were performed to calculate the yearly or monthly energy use of an individual buildings. Table 1 explains various
energy models developed by different organizations in the UK.

The above-mentioned models are different in their separation approach. DECarb is not widely aggregated, which means that a relative data set is arranged to make the 8064 unique combinations for 6 time periods. In UKDCM, buildings are restricted by the climate zones, age, types of construction, number of floors, tenancy, and construction method. In BREHOMES, the housing stock is divided into more than 1000 categories, which are categorized by their form of built, age of construction, tenancy, and the ownership of the central heating. In CDEM, the yearly energy use of only 47 house archetypes is combined, and a unique combinations of form of built type and dwelling age are derived. On the other hand, in the Johnston model, only two “notional” dwelling types (pre- and post-1995) have been developed. All the above-mentioned models need some assumptions according to the available supporting data both in the absence of direct data and in the application of input values.

Building performance simulation (BPS) provides effective resources to support the informed design for the decision-making of energy efficient buildings. Variety of approaches such as the yearly building energy modeling, spreadsheet calculations, and statistical analysis based on empirical data can be employed for performing the whole-building energy analysis.

2.2 Statistics and energy models

Numerical investigation is valuable to extract the respective data regarding the future climate projection, natural disasters, and many other identifiable factors. It is an essential need to establish a framework to interpret the explanations from actual or digital data. A large number of researchers are using the quantitative numerical methods to validate and prove their hypothesis for their presented data. Modelers by running their programs on different software should not expect to extract the same results. These variations in their extracted results are due to different modeling capabilities, various time steps, using disparate calculation algorithms, and many other assumptions during the calculations steps. The statistical models are crucial to forecasting and detecting various real-time issues. Moreover, by including the statistical analysis in reporting the results, the credibility of the findings could be increased. The following sections explain both sensitivity and stepwise techniques and their pros and cons when applied to energy calculations of buildings.

2.3 Sensitivity analysis

Sensitivity analysis methods have been successfully applied in various fields including complex engineering systems, economics, physics, and many others. Sensitivity analysis is a helpful tool to identify the critical control points, prioritizing the data collection or research schemes, and validating a model. Lomas et al verified three necessary sensitivity analysis tools with building energy modeling software. They stated that the Monte Carlo method is realistic although it has a drawback for only defining the total uncertainties. Lam et al showed that it is still unclear for researchers how simulation models are repeated iteratively.

Generally, the sensitivity analyses have been employed in Integrated Method to Assess Greenhouse Effect (IMAGE) by using the Weismann and Morris methods. In this approach, the selected significant input parameters are compared between two methods. Sensitivity analysis approaches need a huge number of iterative runs. Hence, it is essential to computerize the procedure by creating the building energy models with different combinations of inputs. Existing literature classifies sensitivity analysis methods into three categories. The first group of sensitivity analysis methods are based on mathematical approaches and usually calculate outputs for a few values of an input within its possible range. Most methods of this group use one-parameter-at-a-time (OAT) approach. The second group of sensitivity analysis methods are based on statistical analysis. In these group of methods, one or more input parameters may vary at a time. One of the benefits of the statistical methods is ability of quantifying the effect of simultaneous interactions among multiple inputs. Examples of statistical
methods include linear regression analysis (RA), the analysis of variance (ANOVA), the Fourier amplitude sensitivity test (FAST), the mutual information index (MII), standardized regression coefficient (SRC), and standardized rank regression coefficient (SRRC). The third group of sensitivity analysis methods are based on graphical illustrations. These methods use graphs, charts, or surfaces of inputs vs corresponding outputs. This paper investigates a statistical approach for sensitivity analysis and therefore can be classified into the second group of methods described above. The proposed method, stepwise regression, does not need to vary one or more parameters at a time. It actually considers all input parameters and ranks the input parameters based on their degree of importance.

According to Tian, a number of processes in the sensitivity analysis remain the same on most of sensitivity analysis techniques. Examples of these processes include determining input variations, creating and running energy models, collecting results and performing sensitivity analysis, and finally presenting the results. These applications are different according to the variations of input factors for various research purposes, which are often ignored in the field of building analysis.

As mentioned, sensitivity analysis of building energy models requires costly computational procedures that may be the most important problem in adopting a number of sensitivity analysis methods such as variance-based methods. However, existing literature documents the PEAR, PCC, and SRC as sensitivity analysis alternatives, but according to these choices are a compromise between accuracy and computational cost. In addition, states that the rank transformation sensitivity indices may modify the model under study and may produce results that are not reliable. A number of studies (eg, 39, 40) have focused on the Morris method as a computationally cheap method. Their findings indicate that this method was not reliable in ranking input variables based on their level of importance; rather, it can be used to identify unimportant variables of an energy model.

Almost all sensitivity analysis methods include several issues, of which exploring around a base case prevails as the predominant method. The other important issue with sensitivity analysis is the relevance and interaction among the input parameters that is not maintained throughout simulation runs.

### 2.4 Stepwise regression

The regression methods are global sensitivity analysis approaches. In such analysis, the stepwise regression is used as a standard technique. In this technique, the selection of factors is based on their significance level. Therefore, the selection methodology is extracted in respect to the statistical techniques. As stated by Helton et al, the stepwise regression is initially created with the most important variable, which is continued by adding the next most important variable. It means that the sensitivity analysis does not necessitate all of the unclear variables for this regression model.

Dhar et al have used stepwise regression technique to detect the most relevant factors for a temperature-based and generalized Fourier series for energy purposes. Dhar et al describe the Fourier series methods developed for modeling of energy use in commercial buildings. Their study was not focused on finding most important parameters in building stock energy models. Later, Katipamula determined the strength of the variables in three different buildings with a stepwise regression, executed for the dual-duct constant volume (DDCV) systems. Moreover, they compared two models based on their selected variables through such regression analysis. Katipamula was focused on building a multiple linear regression baseline models which could be used in detecting deviations in energy consumption resulting from major operational changes and therefore did not aim to find the most important input parameters for building stock models. It is noteworthy to mention that this regression can be influential on either the design objectives or the constraints when the optimized models are utilized. The forward stepwise method is the most widely used in the analysis according to Helton et al. In this scheme, first, the most important factor enters into the model and then it is followed by the next important one, and this process repeats until no significant variables are left according to the statistical test.

Existing literature failed to examine the possibility of using stepwise technique in a building stock model, which calculates energy consumption of whole buildings in an area. Identifying the vital inputs is crucial when in creating a building stock energy model data on a number of input parameters is missing. Thus, decreasing the number of inputs supports the work to collect and disseminate data for analysis.

### 3 Research Methodology

While sensitivity analysis provides robust and informative results, it requires huge number of iterative runs that are time-consuming and needs high-speed processors. This paper by referring to outputs of the sensitivity analysis and stepwise regression analyzes the benefits of each method over others. Although other data-driven approaches could be used, regression modeling approach is simple to develop and easy to use compared to many other approaches including detailed hourly simulations of energy use in buildings. The methodology to run through the process of identifying the relevant
input parameters through reliable technique is summarized in Figure 1 and explained below.

3.1 | Normality distribution and box-cox transformation

Data used in this study are derived from the baseline model created in ECCABS. For each of the 192 sample buildings, the ECCABS model calculates the energy consumption (response variable) based on the inserted input parameters (predictors). Such set of raw data was formatted as a long list of observations including predictors and responses. The created data set was used to perform the stepwise regression approach.

Normality distribution test is performed to figure out whether the data are normally distributed, which is required to execute the stepwise regression. Prior to performing such analysis, a normality check is done on the response variable (net energy use) to check its distribution. Box-cox transformation is a typical technique used to normalize the data for more robust results. After ensuring that the response variable is normally distributed, the predictors must be arranged for stepwise regression. Some predictors might be strongly correlated, and some may be constant for all sample buildings and therefore do not enter the regression.

As mentioned, a common assumption used for building regression models is usually that the data originate from a normal distribution. In case the variables are not normally distributed, we may need to transform so that this assumption is approximately satisfied. Transformation of independent and dependent variables helps to create the simplest possible regression model. A Box-Cox normality can be often plotted to find an alteration that will roughly normalize the data. The one-parameter Box–Cox transformation from $y$ to $y^\lambda$ for positive variables is defined as follows:

$$ y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases} $$  (1)

3.2 | Performing stepwise regression

Stepwise regression is utilized to examine the importance of the input parameters. It can be performed either in forward method by including a single independent input variable and adding more variables until a certain degree of accuracy is achieved, or backward method by considering all independent input variables and removing the insignificant ones. This paper utilizes the forward selection method to identify the most significant variables.

3.3 | Validation of the regression pattern and analyzing the results

Once the data are examined for normality, normalized using box-cox, and analyzed with stepwise regression, the results are checked for accuracy. Regression model is an approach that determines the obtained results are precise as explanations of the data. The validation for this approach can be done by testing the appropriateness of the regression curve fit, and study of the regression residuals.

3.4 | Comparing results

After validation, results are compared to ones obtained from sensitivity analysis in the ECCABS model and a discussion on the advantages of stepwise regression compared to the sensitivity analysis is presented.

The following sections explain the data collection methodology, considered building types, the results from both sensitivity analysis and stepwise regression. Finally, the conclusions of robustness of different methods and their applicability based on the data type have been presented. At this stage, buildings are divided into two groups and the stepwise regression is done on each group. This division is done to see the more important parameters relative to newer and older buildings.

4 | HOW THE PROPOSED ANALYSIS AND REGRESSION WORK

The aim of the sensitivity analysis is to examine the changes in the output variables of the model according to the minor changes in the inputs. The sensitivity analysis is a helpful tool to determine which variables have the largest effects on the output variables of the model. This step may take a long time if using the meta-model methods with a larger number of input factors. According to Firth et al., sensitivity analysis comprises the following steps:
1. Input parameters should be assigned to set values ($k_i$).
2. Each input parameter faces a small change $\Delta k_i$, while the other input parameters are kept constant, that is, $\pm 1\%$ change in the input parameter.
3. For each change in the input parameters, the model is iteratively run again.
4. New output variables are used to calculate the sensitivity coefficients, and all are normalized with the calculated coefficients.

The sensitivity coefficients characterize the partial derivatives of output variables to input parameters. These coefficients for a model with $n$ output variables and $m$ input parameters are given by the following equation:

$$\frac{\partial y_i}{\partial k_j} \approx \frac{y_j (k_j + \Delta k_j) - y_j (k_j - \Delta k_j)}{2\Delta k_j} \quad (2)$$

where $y_i$ is the $i$th output variable, $k_j$ is the $j$th input parameter ($i$ ranges from 1 to $n$, and $j$ ranges from 1 to $m$), $\frac{\partial y_i}{\partial k_j}$ is the sensitivity coefficient for output variable $y_i$ and input parameter of $k_j$, and $y_j (k_j + \Delta k_j)$ is the value of $y_j$ when the input parameter $k_j$ is increased by $\Delta k_j$ rate.

To be able to compare the sensitivity coefficients of input parameters with different units, the normalized coefficients must be defined. Equation (3) suggested by Firth et al.\textsuperscript{16} is used to calculate this coefficient.

$$S_{ij} = \frac{k_j \frac{\partial y_i}{\partial k_j}}{y_i} \quad (3)$$

Iterative stepwise regression is actually a step-by-step model that automatically selects the independent input variables. As mentioned, in this paper, since the most significant variables are in focus, the forward method is adopted.

5 | RESULTS AND ANALYSIS

Arababadi\textsuperscript{8} used the ECCABS model to calculate the energy demand of the UK building stock. The UK building stock in year 2010 is modeled as represented by archetype residential buildings. The archetypes correspond to:

- Six building types for residential buildings: detached, semidetached, terraced, bungalow, flat, and others.
- Four climate zones represented by the cities of London, Birmingham, Newcastle and Glasgow.
- Six periods of construction namely: Before 1985, from 1986 to 1991, from 1992 to 1995, from 1996 to 2002, from 2003 to 2006, and from 2007 to 2010.
- Two heating system types: central and noncentral.

The model has been validated satisfactorily, for which the results of modeling are in agreement with the statistical data (3% deviation).

The inputs to ECCABS model are listed in Table 2. The table also shows the results for the sensitivity analysis done by Arababadi\textsuperscript{8} based on setting the initial values of the UK building stock. The presented results in Table 1 show that the largest $S_{ij}$ values are all related to the input parameters, which can be influential in the consumption of energy for the space heating purposes. The indoor air temperature results have the highest sensitivity coefficient (1.63 in the residential sector). This can be explained by a 1% increase in the indoor air temperature, and the energy consumption of the buildings increases by 1.63%. Next most important inputs include the external surface and average $U$-value of buildings. The negative sensitivity values mean an increase in those parameters reduces net energy consumption.

5.1 | Stepwise regression

The statistical investigation begins with execution of a normality test on the response variable (total energy use of sample buildings). This paper refers to normality probability plot to interpret concerns and to find the outliers. Figure 2 displays the test results. The deviations from the straight line illustrate the abnormalities, and it is observed that the $P$-value for this data is smaller than the acceptable range (.005). Small $P$-value tells that the hypothesis under consideration may not sufficiently explain the observation. This shows that the null theory is rejected. So, the response variable and the net energy use of sample buildings are not normally distributed.

As mentioned, data transformation and, particularly, the Box-Cox transformation are selected as a technique to make the data normal.\textsuperscript{34} However the Box-Cox transformation does not guarantee the normality of results.\textsuperscript{34} Hence, it is necessary to check whether the transformed data are normally distributed by using a probability plot. Figure 3 shows the normality test done on the transformed data. Comparing Figures 2 and 3 shows a substantial development in the respective test. It is still not a perfectly normal data set but with $P$-value = .015; the null hypothesis is not rejected. Thus, it is assumed that the data are normally distributed.

In this paper, the forward selection technique is used that involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.\textsuperscript{43} Table 3 shows the results of the implemented regression on the 192 sample buildings. Steps in this table represent adding of new variables to the regression model until a reliable model is obtained. Note that some of the
input parameters are not selected in the regression process because they were either strongly correlated or they were constant for all 192 samples. Correlated input parameters in building stock energy models are categorized as structural multicollinearity that means one input parameter is created using other input parameter. For example, one input parameters might be calculated by multiplying another input parameter into a constant.

### TABLE 2 Inputs to ECCABS; the results for sensitivity analysis in residential building stock

| Input parameters | Initial set value \( (K_i) \) | Overall change \( (2^\Delta K_i) \) | Output parameter | Sensitivity coefficient | Normalized sensitivity coefficient \( S_{ij} \) |
|------------------|--------------------------------|--------------------------------|-----------------|------------------------|---------------------------------------------|
| Sw               | 16.475                         | 0.329                          | 0.79            | 2.401                  | 0.070                                       |
| Ts               | 0.700                          | 0.014                          | -0.79           | -56.428                | -0.070                                      |
| I mean           | 1.160                          | 0.023                          | 9.820           | 426.956                | 0.882                                       |
| Wc               | 0.700                          | 0.014                          | -0.790          | -56.428                | -0.070                                      |
| Wf               | 0.677                          | 0.013                          | -0.790          | -60.769                | -0.073                                      |
| Boiler eff.      | 0.760                          | 0.015                          | -6.6            | -440                   | -0.596                                      |
| A                | 90.507                         | 1.810                          | 2.380           | 1.314                  | 0.212                                       |
| Ac               | 4.282                          | 0.085                          | 0.43            | 5.058                  | 0.038                                       |
| S                | 229.218                        | 4.584                          | 9.820           | 2.142                  | 0.875                                       |
| HyP              | 0.048                          | 0.001                          | 0.020           | 20.000                 | 0.001                                       |
| Oc               | 1.148                          | 0.023                          | -0.320          | -13.913                | -0.028                                      |
| Tv               | 24.000                         | 0.48                           | 0.000           | 0.000                  | 0.000                                       |
| Hw               | 5.038                          | 0.100                          | 2.300           | 23.000                 | 0.206                                       |
| Lc               | 0.992                          | 0.019                          | 0.000           | 0.000                  | 0.000                                       |
| Tc               | 22 969 664                     | 459 393.28                     | -0.17           | 0.000                  | 0.000                                       |
| Tmax             | 26                             | 0.52                           | 0.000           | 0.000                  | 0.000                                       |
| Tmin             | 17.188                         | 0.343                          | 18.260          | 53.236                 | 1.63                                        |
| Vc               | 0.036                          | 0.001                          | 0.120           | 120.000                | 0.007                                       |
| Vcn              | 0.309                          | 0.006                          | 0.000           | 0.000                  | 0.000                                       |

**FIGURE 2** Results of normality test on the energy use of sample buildings
FIGURE 3  Normality test on the response variable after the Box-Cox transformation

![Probability Plot of Net Energy Transformed](image)

| Step | 1          | 2          | 3          | 4          | 5          | 6          | 7          | 8          | 9          |
|------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Constant | 0.0001458 | 0.0001707 | 0.0001353 | 0.0001511 | 0.0001527 | 0.0001505 | 0.0006272 | 0.0009421 |
| S    | -0.000001 | -0.000001 | -0.000001 | -0.000001 | -0.000001 | -0.000001 | -0.000001 | -0.000001 |
| U    | -0.00005  | -0.00005  | -0.00005  | -0.00006  | -0.00006  | -0.00006  | -0.00006  | -0.00006  |
| A    | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  |
| Oc   | 0.00001   | 0.00001   | 0.00001   | 0.00001   | 0.00001   | 0.00001   | 0.00001   | 0.00001   |
| Pth  | -0.00002  | -0.00001  | -0.00001  | -0.00001  | -0.00001  | -0.00001  | -0.00001  | -0.00001  |
| Vcn  | -0.00002  | -0.00002  | -0.00002  | -0.00002  | -0.00002  | -0.00002  | -0.00002  | -0.00002  |
| Sw   | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  | 0.000001  |
| Ac   | -0.00009  | -0.00015  |            |            |            |            |            |            |
| Tcc  |            |            |            |            |            |            |            | -0.000001 |
| R-Sq | 52.93      | 90.99      | 92.39      | 94.34      | 95.29      | 95.45      | 95.52      | 96.00      | 96.24      |
| R-Sq(adj) | 52.69      | 90.90      | 92.27      | 94.22      | 95.17      | 95.30      | 95.35      | 95.82      | 96.05      |

FIGURE 4  Residual plots before (left) and after (right) Box-Cox transformation

![Residual Plots for Net Energy use](image)

![Residual Plots for Net Energy Transformed](image)
5.2 | Stepwise regression model validation

As regression results show, the R-Sq (adj) is 96%, and the RMSE indicator is very small. A great $R^2$ cannot guarantee that the presented model is valid for investigation. Therefore, the best choice is to use the adjusted $R^2$ as it can notably improve the level of confidence in the validation process. Figure 4 shows how the Box-Cox transformation helped to improve the fitted regression model by plotting the residual values for both before and after Box-Cox effects. When model is fitted with the data, the residuals are estimated for the random errors that make the relationship between the explanatory variables and the response variable. Thus, if a random appearance happens for the residuals, it can be concluded that the model for the data is fitted correctly. Figure 4 clarifies that the fitted model has improved after the Box-cox transformation.

As Figure 4 shows, the data points are randomly distributed on the fitted plot around zero. Therefore, it can be assumed that the error terms have a mean equal to zero. At this stage, buildings are divided into two groups and the stepwise regression is done on each group. This division is done to see the more important parameters relative to newer and older buildings.

5.3 | Built before 1995 (Old Buildings)

In this section, the results of a stepwise regression on buildings built before 1995 are presented, and the residual plots are analyzed. In the year 1995, the energy policies including heating and ventilation systems to standardize the conservation of fuel and power were updated. The analyzed results are presented in Table 4.

As results of stepwise regression for buildings built before 1995 show, the fitted model still has acceptable R-Sq. (adj) and RMSE (S) but residual plots show (Figure 5) that the previous model that was fitted to the entire data set was more accurate. It is due to the fact that the $R^2$ as a measure of model validity increases by adding more variables into the model.

5.4 | Built after 1995 (Newer Buildings)

The same procedure is done for buildings built after 1995. Table 5 shows the results of stepwise regression for this group of buildings. Note that due to characteristics of data on this group of buildings, more input variables were unable to enter the regression because they are either strongly correlated or constant for all samples.

Residual plots for the model fitted to the buildings built after 1995 are illustrated in Figure 6. The residual behavior of this model is similar to the previous model (built before 1995), and the first model that was fitted to all data set still has better residual behavior.

6 | DISCUSSION

In Table 6, the most important input parameters are compared. These parameters are found by both the stepwise regression and sensitivity analysis. Note that the most important input parameters found by sensitivity analysis are in fact input parameters with highest normalized sensitivity coefficients. It is clear that they have both identified almost the same parameters with relatively the same order of importance.

Stepwise regression in this paper produces relatively accurate results and is considerably faster. Obviously, computational times are highly dependent on the processor capabilities and may vary based on computer system speeds. Although

### Table 4 | Results of Stepwise Regression for buildings before 1995

| Step | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       |
|------|---------|---------|---------|---------|---------|---------|---------|---------|
| Constant | 0.01155 | 0.01430 | 0.01534 | 0.01465 | 0.01310 | 0.01282 | 0.01272 | 0.03718 |
| S    | −0.00001| −0.00001| −0.00001| −0.00002| −0.00002| −0.00002| −0.00002| −0.00007|
| U    | −0.00286| −0.00335| −0.00335| −0.00330| −0.00330| −0.00327| −0.00334| −0.00354|
| Pfh  | −0.00093| −0.00094| −0.00092| −0.00092| −0.00092| −0.00093| −0.00092| −0.00092|
| A    | 0.00002 | 0.00005 | 0.00004 | 0.00004 | 0.00004 | 0.00004 | 0.00007 |        |
| Oc   |         | 0.00035 | 0.00038 | 0.00042 | 0.00042 | 0.00214 |        |        |
| Tcc  | 0.00000 | 0.00000 | −0.00000|         |         |        |        |        |
| Sw   | 0.00002 | 0.00017 |        |         |         |        |        |        |
| Hw   | −0.0038 |        |         |         |         |        |        |        |
| R-Sq | 54.50   | 88.90   | 91.09   | 92.91   | 94.09   | 94.25   | 94.34   | 94.58   |
| R-Sq(adj) | 54.02 | 88.66   | 90.79   | 92.60   | 93.77   | 93.86   | 93.89   | 94.08   |
it would be beneficial to quantitatively report the computational times for stepwise regression and sensitivity analysis, this sort of data was not precisely recorded at the time of this research. However, it was observed that the calculation time is considerably reduced by adopting stepwise regression. To be more specific, in order to perform the sensitivity analysis based on Firth et al., the energy model needs to be run $2 \times N$ times (where $N$ is number of input parameters) and each run takes approximately 30 minutes (using a normal core i3 processor) while performing a sensitivity analysis with the same processing capacity could be done in less than 2 minutes. The other advantage includes the analysis for a different group of samples as it was done in this paper (buildings built before and after 1995). By doing the stepwise regression on different types of sample buildings, one would be able to compare the most important input parameters for each group. Although stepwise regression has some significant benefits over the sensitivity analysis, some drawbacks require more attention. For example, input parameters that are strongly correlated cannot enter the regression process. Notice that in building modeling, some input parameters are calculated by multiplying a constant into another input parameter. For example, lighting energy use and appliance energy use are calculated by multiplying the lighting and appliance energy density by the total floor area; thus, both lighting and appliance energy are correlated with total floor area.

Additionally, input parameters that are identical for all buildings cannot enter the regression process. We know that in building stock modeling, data gathering is time-consuming and labor intensive. In many areas, data on a number of input parameters may be missing. Therefore, it is common to assume the same values of some input parameters for all sample buildings. For example, occupancy, lighting, and plug-load schedules are usually assumed to be identical for different

**FIGURE 5** Residual plots for model fitted to the buildings built after 1995

**TABLE 5** Results of Stepwise regression of buildings built after 1995

| Step | 1     | 2     | 3     | 4     |
|------|-------|-------|-------|-------|
| Constant | 0.01288 | 0.01483 | 0.01223 | 0.01059 |
| S    | -0.00001 | -0.00001 | -0.00001 | -0.00001 |
| U    | -0.00478 | -0.00481 | -0.00480 |
| Ac   | 0.00040 | 0.00058 |
| tcc  |       | 0.00001 |
| R-Sq | 84.57 | 92.71 | 93.53 | 94.88 |
| R-Sq(adj) | 84.40 | 92.56 | 93.32 | 94.66 |
clusters of residential buildings. These sort of input parameters may not be selected in stepwise regression analysis.

7 | CONCLUSIONS

This paper illustrates that stepwise regression is a feasible substitution for sensitivity analysis. It is also provided that the input data parameters are not strongly correlated and not identical for all samples. Since it is usually easy to identify the correlated and nonvarying parameters in building stock modeling, it might still be possible to use a stepwise regression even when the variables are correlated. According to the research content, the following points can be concluded:

- Stepwise regression can help the energy modelers to save the time that is required to run the simulation for sensitivity analysis.
- Results can help the developers to identify the most important inputs for their models. It can help identify those parameters that can be neglected and reduce the number of unwanted input parameters, in turn making simulation easier. It would be easier and less time-consuming to gather the input data when the number of input parameters is reduced. For instance, results of this paper show that parameters such as coefficient of solar transmission of the window or Frame coefficient of the window could be neglected.

As a suggestion for the future work, the same method can be used for the data from other countries to both compare the models to the ECCABS model for the UK and to see whether the important input parameters vary regionally.
Nomenclature

Location_no  Weather region
A          Area of heat floor space
Ac         Average constant consumption of the appliances
HR          Efficiency of the heat recovery system
i          Density of the layer
HR          Height under the roof
Levels      Number of floors of the building
LS          The living space or heated floor area
Form        A parameter which indicates the configuration of
the building
Trmin       Minimum indoor temperature
Ts          Coefficient of solar transmission of the window
Tv          Tint to start opening windows/nat ventilation
U           Mean U value of the building
Vc          Sanitary ventilation rate
Wc          Shading coefficient of the window
Vcn         Natural ventilation rate
Weight      Coefficient to scale up the type to the Building Stock
WF          Frame coefficient of the window
ATT         The attached character of the dwelling
Form        A parameter which indicates the configuration of
the building
LS          The living space or heated floor area
HR          Height under the roof
\( \rho_i \)  Density of the layer
\( C_{pi} \)  Specific heat capacity of the layer
\( S_i \)  Area of the layer
\( d_i \)  Thickness of the layer

ORCID

Reza Arababadi  https://orcid.org/0000-0001-7740-735X

REFERENCES

1. Oree V, Khoodaruth A, Teemul H. A case study for the evaluation of realistic energy retrofits strategies for public office buildings in the Southern Hemisphere. Build Simul. 2016;9(2):113-125. https://doi.org/10.1007/s12273-015-0259-y
2. Xia C, Zhu Y, Lin B. Building simulation as assistance in the conceptual design. Build Simul. 2008;1:46-52. https://doi.org/10.1007/s12273-008-8107-y
3. Malkawi A, Yan B, Chen Y, Tong Z. Predicting thermal and energy performance of mixed-mode ventilation using an integrated simulation approach. Build Simul. 2016;9(3):335-346. https://doi.org/10.1007/s12273-016-0271-x
4. Kavcic M, Mavrogianni A, Mumovic D, Summerfield A, Stevanovic Z, Djurovic-Petrovic M. A review of bottom-up building stock models for energy consumption in the residential sector. Build Environ. 2010;45:1683-1697. https://doi.org/10.1016/j.buildenv.2010.01.021
5. Shorrock L, Dunster J. The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock. Energy Pol. 1997;25(12):1027-1037. https://doi.org/10.1016/S0301-4215(97)00130-4
6. Campolongo F, Braddock R. Sensitivity analysis of the IMAGE greenhouse model. Environ Model Softw. 1999;14:275-282. https://doi.org/10.1016/S1364-8152(98)00079-6
7. Calleja Rodríguez G, Carrillo Andrés A, Domínguez Muñoz F, Cejudo López JM, Zhang Y. Uncertainties and sensitivity analysis in building energy simulation using macroparameters. Energy Build. 2013;67:79-87. https://doi.org/10.1016/j.enbuild.2013.08.009
8. Arababadi R. Energy Use in the EU Building Stock - Case Study: UK. Master’s, Linköping University; 2012.
9. Mata É, Kalagasidis AS, Johnsson F. A modelling strategy for energy, carbon, and cost assessments of building stocks. Energy Build. 2013;56:100-108. https://doi.org/10.1016/j.enbuild.2012.09.037
10. Polly B. Improving the Accuracy of Software-Based Energy Analysis for Residential Buildings, (July); 2011.
11. Zhai ZJ, McNeill JS. Roles of building simulation tools in sustainable building design. Build Simul. 2014;7(2):107-109. https://doi.org/10.1007/s12273-013-0169-9
12. Hong T, Buhl F, Haves P, Selkowitz S, Wetter M. Comparing computer run time of building simulation programs. Build Simul. 2008;1(3):210-213. https://doi.org/10.1007/s12273-008-8123-y
13. Natarajan S, Levermore GJ. Predicting future UK housing stock and carbon emissions. Energy Pol. 2007;35(11):5719-5727. https://doi.org/10.1016/j.enpol.2007.05.034
14. Boardman B, Darby S, Killip G, et al. 40% House. Environmental Change Institute, University of Oxford; 2005.
15. Tian W, Choudhary R. A probabilistic energy model for non-domestic building sectors applied to the analysis of school buildings in greater London. Energy Build. 2012;54:1-11. https://doi.org/10.1016/j.enbuild.2012.06.031
16. Firth SK, Lomas KJ, Wright AJ. Targeting Household Energy-efficiency Measures Using Sensitivity Analysis. London, UK: Routledge; 2009.
17. Johnston D. A Physically-based Energy and Carbon Dioxide Emission Model of the UK Housing Stock, s.l.: s.n.; 2003.
18. Attia S, De Herde A, Gratia E, Hensen JLM. Achieving informed decision-making for net zero energy buildings design using building performance simulation tools. Build Simul. 2013;6(1):3-21. https://doi.org/10.1007/s12273-013-0105-z
19. Zwiers FW, Von Storch H. On the role of statistics in climate research. Int J Climatol. 2004;24:665-680. https://doi.org/10.1002/joc.1027
20. Frey HC, Patil SR. Identification and review of sensitivity analysis methods. Risk Anal. 2002;22:553-578. https://doi.org/10.1111/1042-2277.00039
21. Lomas KJ, Epple H. Sensitivity analysis techniques for building thermal simulation programs. Energy Build. 1992;19(1):21-44. https://doi.org/10.1016/0378-7788(92)90033-D
22. Lam JC, Hui SCM. Sensitivity analysis of energy performance of office buildings. *Build Environ*. 1996;31:27-39. https://doi.org/10.1016/0360-1323(95)00031-3

23. Tian W. A review of sensitivity analysis methods in building energy analysis. *Renew Sustain Energy Rev*. 2013;20:411-419. https://doi.org/10.1016/j.rser.2012.12.014

24. Tian W, De Wilde P. Uncertainty and sensitivity analysis of building performance using probabilistic climate projections: a UK case study. *Autom Constr*. 2011;20:1096-1109. https://doi.org/10.1016/j.autcon.2011.04.011

25. Lam JC, Wan KKW, Yang L. Sensitivity analysis and energy conservation measures implications. *Energy Convers Manage*. 2008;49(11):3170-3177. https://doi.org/10.1016/j.enconman.2008.05.022

26. Rasouli M, Ge G, Simonson CJ, Besant RW. Uncertainties in energy and economic performance of HVAC systems and energy recovery ventilators due to uncertainties in building and HVAC parameters. *Appl Therm Eng*. 2013;50:732-742. https://doi.org/10.1016/j.applthermaleng.2012.08.021

27. Demanuele C, Tweddell T, Davies M. Bridging the gap between predicted and actual energy performance in schools. In: *World Renewable Energy Congress XI*; 2010:1-6.

28. Fabrizio E, Monetti V. Methodologies and advancements in the calibration of building energy models. *Energies*. 2015;8(4):2548-2574. https://doi.org/10.3390/en8042548

29. Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliab Eng Syst Saf*. 2006;91(10–11):1175-1209. https://doi.org/10.1016/j.ress.2005.11.017

30. Dhar A, Reddy T, Claridge D. Modeling hourly energy use in commercial buildings with Fourier series functional forms. *J Solar Energy Eng*. 1998;120(3):217-223. https://doi.org/10.1115/1.2888072

31. Katipamula S. Multivariate regression modeling. *J Solar Energy Eng*. 1998;120(3):177-184. https://doi.org/10.1115/1.2888067

32. Snee R. Validation of regression models: methods and examples. *Technometrics*. 1977;19(4):415-428. https://doi.org/10.1080/00401706.1977.10489581

33. Wang M, Wright J, Buswell R, Brownlee A. A Comparison of Approaches to Stepwise Regression for Global Sensitivity Analysis Used with Evolutionary Optimization; 2013:2551-2558. http://www.ibpsa.org/proceedings/BS2013/p_1047.pdf. Accessed February 28, 2020.

34. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Ser B (Methodological)*. 1964;26(2):211-252. https://doi.org/10.2307/2984418

35. Montgomery DB, Morrison DG. A note on adjusting R^2. *J Finance*. 1973;28:1009-1013.

36. Frey HC, Mokhtari A, Danish T. Evaluation of Selected Sensitivity Analysis Methods Based Upon Applications to Two Food Safety Process Risk Models. Washington, DC: North Carolina State University; 2003.

37. Nguyen A-T, Reiter S. A performance comparison of sensitivity analysis methods for building energy models. *Build Simul*. 2015;8(6):651-664.

38. Giglioli N, Saltelli A. *Simlab 2.2 Reference Manual*. Ispra, Italy: Institute for Systems Informatics and Safety (Joint Research Centre, European Commission); 2008.

39. Confalonieri R, Belloccchi G, Bregaglio S, Donatelli M, Acutis M. Comparison of sensitivity analysis techniques: a case study with the rice model WARM. *Ecol Model*. 2010;221:1897-1906.

40. Wainwright HM, Finsterle S, Zhou Q, Birkholzer JT. Modeling the performance of large-scale CO2 storage systems: a comparison of different sensitivity analysis methods. *Int J Greenhouse Gas Control*. 2013;17:189-205.

41. Tempel M, Kowarik A, Filzmoser P. Iterative stepwise regression imputation using standard and robust methods. *Comput Stat Data Anal*. 2011;55(10):2793-2806.

42. Zhang Y, Zhang X, Huang P, Sun Y. Global sensitivity analysis for key parameters identification of net-zero energy buildings for grid interaction optimization. *Appl Energy*. 2020;279:115820.

43. Brøgger M, Bacher P, Wittchen KB. A hybrid modelling method for improving estimates of the average energy-saving potential of a building stock. *Energy Build*. 2019;199:287-296.

**How to cite this article:** Arababadi R, Naganathan H, Saffari Pour M, Dadvar A, Parrish K, Chong O. Building stock energy modeling: Feasibility study on selection of important input parameters using stepwise regression. *Energy Sci Eng*. 2021;9:284–296. https://doi.org/10.1002/ese3.847