Building Energy Use Surrogate Model Feature Selection – A Methodology Using Forward Stepwise Selection and LASSO Regression Methods

Erica C. Barnes¹, J.J. McArthur¹
¹Ryerson University, Toronto, Canada

Abstract
Statistical regression models were developed to permit the rapid modelling of large commercial office buildings within a single climate zone. The regression models are developed using a large number of building parameters and their hourly energy model simulated results. In previous building energy regression modelling, there is a research gap in selecting building parameters using statistical approaches. This paper investigates a feature selection method, including forward stepwise selection and LASSO regression, to identify building parameters that, together, have the most significant impact on total building energy load. The regression model, with 25 features selected through this methodology, predicts total energy load at 93.5% accuracy, on average.

Introduction
Building energy simulation software is used in industry to generate reference models and simulate new and existing buildings. The results of such models are often used by building design teams to inform energy conservation measures and verify code and standard compliance. Developing detailed energy models at the early stages of design to simulate and test scenarios is both time and cost intensive and is often not possible due to the lack of detailed information on the design. The ability to quickly evaluate and compare building design scenarios at the earliest design stages can support improved building performance.

Several researchers (Nagpal, et al., 2018) have used supervised machine learning methods, where labelled datasets are used to train predictive mathematical models, to predict building energy loads. These previous studies have generated the sample labels using building energy simulation software. Once the model is trained, the resultant model can predict the building energy load of any combination of the simulated parameters, within the design space defined by the original simulations. When computer-simulated values are used as the output, the supervised machine learning model is often referred to as a surrogate model. A notable gap in the literature is the selection of key input features for such surrogate models, which to date has been based on expert knowledge. This paper discusses the use of wrapper and embedded methods to identify the key input features for the surrogate model.

The feature selection methods used in this research produce a surrogate building energy use prediction model, with an amount of error compared to the simulated result, which can be used to quickly evaluate building energy use based on any combination of features/parameters within the model’s design space. The model algorithm is kept to multivariate regression to allow for comparison of the feature selection methods.

Previous Research
While there is past research related to the development of building energy surrogate models, there is a significant research gap regarding the method for selecting an appropriate subset of building features for use in the models. Past building energy surrogate models of this nature have been trained on an assumed set of variables and have not undertaken a systematic investigation to determine the optimal input set. The previous studies use subsets of 6-40 variables from a total of 120 parameters related to building geometry; building enclosure properties; HVAC system performance characteristics; lighting; equipment; occupants; and schedules. The variables are selected based on previous knowledge of high energy-impacting building parameters and parameters often known in the early stages of design. The final models use the full set of variables. Some researchers have evaluated feature importance in a data pre-processing step by analyzing each variable’s correlation to the simulated result or following model training by analyzing the feature weights (J. S. Hygh 2011, Zhao and Magoulés 2012, Lam and Hui 1996, Tian, Wei 2013, Capozzoli, Mechri and Corrado 2009). These methods are used to interpret the results of the model and not to remove features from the model.

Previous research used various methods for developing the data set to train the learning algorithms and test the models. Different methods to generate the input matrix include: determining building attributes from existing buildings (Tian, Choudhary, et al. 2015); generating sample sets within ranges of building parameters using Monte Carlo sampling (J. S. Hygh 2011, Capozzoli, Mechri and Corrado 2009) and Latin Hypercube Sampling methods (Aijazi 2017, Nagpal, et al. 2018); and randomly generating geometry features using building modelling software plug-ins (Asl, et al. 2016, Aijazi 2017). For each method, the parameters in the sample sets were reflected in computer-simulated building energy models and the simulated results used as the target for training the learning algorithm. The United States Department of Energy (DOE) energy modelling software, EnergyPlus was used by many researchers (Zhao and Magoulés 2012, J. S. Hygh 2011, Aijazi 2017).
Few previous studies have used learned models to select key building features (Zhao and Magoulés 2012). Such feature selection can be used to inform archetype energy model design, code and standard requirements, and assumptions for detailed building energy models. Selecting the most relevant feature set for predicting the output with a certain learning algorithm can have many advantages for the performance of the model including: removing irrelevant features from the model; reducing model overfitting; and reducing model run time.

Methodology
This study starts with a set of 71 building attributes with ranges representative of high to low performing large office buildings in a Toronto, Ontario, Canada climate. A large number of building design samples, filling the building attribute design space, are created. Building energy loads for each sample are determined through building energy simulation software. Wrapper and embedded feature selection techniques, using multiple-variable regression models, are used to determine the optimal set of features representative of Toronto large office building energy critical behaviors. Figure 1 illustrates the work flow for generating the large office building energy data set.

Generating the Building Energy Data Set
The 71 features selected fall into building geometry, building enclosure performance, lighting and plug power densities, heating ventilation and air conditioning system performance, and occupancy categories. They include the varied internal load attributes in the United States Department of Energy (DOE) Large Office Existing Commercial Reference Models (DOE 2012). The ranges of each of the 71 features were selected from a combination of the DOE Existing Commercial Reference Models, ANSI/ASHRAE/IES Standard 90.1 Prototype Building Models (DOE 2016), and industry knowledge of low to high energy performance building attributes.

The DOE Existing Commercial Building Reference Models, developed in collaboration with Lawrence Berkeley National Laboratory (LBNL), Pacific Northwest National Laboratory (PNNL) and the National Renewable Energy Laboratory (NREL) were created as baseline models for 15 building archetypes in 3 construction eras (Torcellini, et al. 2008). The commercial building model inputs are based on the U.S. Energy Information Administration 2012 Commercial Buildings Energy Consumption Survey data which represents 70% of the commercial buildings in the United States. The DOE Large Office EnergyPlus v7.2 Commercial Reference Model (DOE 2012) was used as the base model for this study.

This study uses 4,000 samples, representing random combinations of the values within the ranges set for the 71 features. Latin hypercube sampling (LHS), using MATLAB’s lhsdesign (MathWorks 2018), was used to generate a 4,000 x 71 matrix of space-filling, near-random selections between 0 and 1. Latin hypercube extends the Latin square, a grid with one sample per row and column, to multi-dimensional space, one sample per axis-aligned hyperplane. By multiplying the Latin hypercube sampled values by the feature range standard deviation and adding the feature minimum values, the feature values for each of the 4,000 samples were generated. Features requiring integer values, such as number of storeys, were assigned the rounded value. This sampling method defined a broad design space for the models.

The annual total site energy load was determined using EnergyPlus, a building energy simulation software developed by the United States Department of Energy (DOE). The 71 building features affect 1,756 EnergyPlus Input Design File (IDF) objects in the base model. Minor changes to the base model were made to facilitate surrogate model development. These included: changing the holidays and site data to represent Toronto, Canada; adding 3 additional window and wall types to allow for these parameters to be modified independently for each building elevation; and adding an insulation layer to the slab-on-grade construction.

As per the base model, the mechanical system was assumed to be a central plant with chiller and boiler, and multi-zone variable air volume with reheat distribution system. The DOE selected this mechanical system for large office buildings based on the results of the U.S. Energy Information Administration Commercial Buildings Energy Consumption Survey (CBECS), a survey of approximately 70% of the commercial buildings in the U.S., as reported by Pacific Northwest National Laboratory in their 2006 study (Winiarski, Jiang and Halverson 2006).
The most extensive IDF object transformation required for this study related to the building geometry features. As the geometry is represented as x, y and z coordinates at each vertex, a geometry transformation script was prepared to transform the 25 geometry features to 1,704 IDF objects. The general model construction used in the DOE commercial reference model was maintained with a rectangular footprint, single ground and top floors, repeated basement and middle floors, and plenums on each the ground, middle and top floors. Each floor is made of a single core and four perimeter zones. Windows are represented as strips with the depth modified to suit the sample’s window-to-wall ratio.

Using Eppy (Philip and Tanjuatco 2013), a Python library developed for manipulating EnergyPlus IDFs, the base EnergyPlus IDF was modified by identifying the location of each of the 1,756 objects in the file and replacing them with the values generated by LHS. This was automated to generate the 4,000 IDFs.

The simulations were run on a Windows server, containing the EnergyPlus 8.0.0.008 installation (DOE & NREL 2015). The 4,000 .idf files were divided into 4 batches and transferred to the system using Remote Desktop Protocol where a Powershell script ran the RunEPlus.bat (DOE & NREL 2015) batch script on each individual file while assigning each process to one of the 4 available CPUs. The total site energy load for each simulation was parsed from the output files using a Python script.

Each of the 4,000 data points generated has an input made up of the 71 building features selected through the Latin hypercube sampling and an output of the EnergyPlus simulated total energy load. The full data set was randomly divided into 3 sets; training, validation and testing so that the data set would be consistently randomly split into the same sets using the Python function sklearn.model_selection.train_test_split (Pedregosa, et al. 2011). Of the total 4,000 samples, the training set was 64%, validation set was 16% and testing set was 20%. The training set was used to train the regression model; the validation set was used to compare the accuracy of the final model.

The input matrix for the training set was normalized so that the features have a mean of zero and a variance of one. Rescaling the input matrix causes the features to be on a similar scale and thereby reduces the risk of the model inaccurately assigning importance to the values of the parameters. This also allows for gradient descent to converge in less steps. The validation and test data sets were standardized with the training set mean and standard deviation.

Before applying feature selection methods, a multivariate linear regression model was trained using the training data set and evaluated using the validation set. The multivariate linear regression equation was determined using gradient descent with a mean squared error cost function.

To address non-linear behaviours in the data, several transformations of the input matrix and output vector were attempted to fit the data to a linear model. The transformation that resulted in the highest model performance was combining exponential transformation of the output vector and logarithmic transformation of the input matrix. In exponential and logarithmic transformation, the log of the output vector and the log of the input matrix, respectively, are used to train the model. The prediction equation becomes:

\[
\hat{y} = 10^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}
\]

The remainder of the analysis in this study used exponential and logarithmic transformation of the output and input, respectively.

**Adding Combined Feature Terms Using Forward Stepwise Selection**

In this study additional features were developed to represent linear combinations of several of the original 71 features. This included combining several of the features defining the building geometry into features representative of the building as a whole, such as total conditioned floor area; and combining features that were defined separately for the building elevations and/or floors.

When adding terms that are linear combinations of the existing terms, it is important to remove the original variables to remove the risk of multicollinearity. In regression analysis, correlated inputs can lead to high variance values of the feature weights/coefficients. According to Tian, Choudhary, et al. (2015), in building energy modelling, not many researchers have evaluated feature correlation. It is not clear if previous studies have addressed the issue of multicollinearity when adding terms that are combinations of 2 or more of their original features.

Forward stepwise selection, a common wrapper feature selection method, was used to evaluate and select the combined features for inclusion in the model. The combined features were added to the model one at a time starting with the variables with highest correlation to the total building energy load and a two-tail p-value less than 0.005. Even though the area weighted SHGC p-value was 0.33 and therefore greater than 0.005 threshold, the new term was tested in the model and showed a decrease in average percent error on the validation data set. Based on the model improvement result, the SHGC values per elevation were combined into a single term.

As each combined feature was added, the original features in the combined feature were removed from the model. Using the validation data set, the model accuracy was evaluated at each step. Combined features that improved the accuracy of the model, based on an evaluation of the average percent error, coefficient of determination and root mean squared error, were kept in the feature set.

**Feature Selection Using LASSO and Elastic Net Regression**

After the optimal set of combined features were added, further feature selection was carried out. Variations of an embedded feature selection method were compared and used to select the optimal set of features and regression model. For this study, least absolute shrinkage and selection operator (LASSO) was the embedded feature selection method used. Elastic Net, a linear combination
of LASSO and Ridge regression, was tested to determine if model accuracy improves with the Ridge regulator. Ridge regression, using L2 regularization, adds a penalty term to the multivariate regression mean squared error cost function; the regularization parameter, \( \lambda \), multiplied by the sum of the squared coefficients. LASSO regression, using L1 regularization, applies a similar penalty term, the regularization parameter multiplied by the sum of the absolute value of the coefficients. The regularization parameter controls the trade-off of fitting the training set using the mean squared error term in the cost function and reducing the size of the coefficients. Both methods shrink the coefficients, with a larger \( \lambda \) value shrinking the weights closer to zero. This simplifies the hypothesis and reduces overfitting. The L1 regulator has a unique property that causes some of the coefficients to be exactly zero and therefore removed from the regression model.

The Python module Scikit-learn was used for both LASSO and Elastic Net regression (Pedregosa, et al. 2011). Both Scikit-learn functions apply the regulator term(s) to the mean squared error cost function. The cost function for the LASSO regression is shown in Equation 2 and the Elastic Net regression cost function is shown in Equation 3.

\[
C(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - h_\beta(x_i))^2 + \lambda \sum_{j=1}^{p} \| \beta_j \|_1
\]

\[
C(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - h_\beta(x_i))^2 + \lambda \sum_{j=1}^{p} \| \beta_j \|_1 + (1-\%\text{Lasso}) \lambda \sum_{j=1}^{p} \beta_j^2
\]

A range of \( \lambda \) values for each the LASSO and Elastic Net models were tested, and the model error calculated used the model prediction on the validation data set. The model accuracies and the number of features selected for each were evaluated and final model was chosen.

Results

The parameter ranges selected for the simulations were developed to encompass building attributes applicable to high and low performing large office buildings in Toronto. The simulated building energy use resulted in energy use intensity (EUI) ranges of 0.33 to 2.69 GJ/m² with a range of conditioned floor area between 6,570 to 1,780,000 m². According to the 2014 Natural Resources Canada Survey of Commercial and Institutional Energy Use, the average EUI for non-medical office buildings between 4,646 and 18,580 m² is 1.16 GJ/m² and over 18,580 m² in Canada is 1.09 GJ/m² (NRCan, 2014). This confirms that the parameter combinations for the samples reflect the energy use of existing office buildings in Toronto.

Multivariate Linear Regression Model

When the training data set, with 71 features, was fit to a multivariate linear regression model, the model had poor performance on both the training and validation data. As shown in Figure 2, the residual plot, the linear regression model predicted total site energy load for the validation data set, versus the difference between the EnergyPlus simulated and the model predicted total energy load, had non-random inverted-U shape. The linear model performed well in the mid-range of the total site energy load but reduced in accuracy in the low and high ranges, with negative energy load predictions at the low total energy load range. The negative energy load values are incorrect, arising from error in the model and indicate a need to transform the data. The overall error of this model was high with an average percent error of 58.7% and a coefficient of determination (R²) of 0.812 on the validation set. This type of model behaviour indicates that there are non-linear relationships between the input and output that are not captured in the multivariate linear regression model.
Adding Combined Feature Terms Using Forward Stepwise Selection

As the combined feature terms were added to the model one at a time, in order of highest Pearson correlation coefficient, the original features used to calculate the added combined feature was removed from the model. The R²-score and average percent error were calculated at each step and if the error evaluation metrics indicated an improved model for each the training and validation data sets, the combined feature was kept in the model.

Figure 4: Forward Stepwise Regression - Average Percent Error for Each Subset of Features for Validation Data Set using Multivariate Regression

Figure 4 shows the impact to the accuracy of the model of adding the combined terms to the model.

The optimal model uses 11 additional terms and 22 of the original variables. This method of adding terms to linearly combine the original features reduced the input matrix feature size from 71 to 33 and did not compromise the model accuracy. On the contrary, the combined terms decreased the average percent error on the validation set by 0.6%.

Feature Selection Using LASSO and Elastic Net Regression

LASSO and Elastic Net embedded feature selection methods were used to determine if any of the 33 features could be removed without compromising the accuracy of the model. Both methods were evaluated by testing a range of regularization parameter values which resulted in varying feature subset sizes, feature coefficients, and model accuracies. Figure 5 shows both the LASSO and Elastic Net model average percent error on the validation data as regularization parameters were varied and thereby varied the feature subsets selected by the model. As the regularization parameter decreases, the shrinkage penalty applied to the cost function reduces, the number of features selected increases, and the accuracy of the model improves. At 25 features, both the LASSO and Elastic Net models reach an average percent error plateau with the Elastic Net model performing slightly better than the LASSO model. Therefore, the Elastic Net model that selects 25 of the 33 features is selected as the best model for this dataset based on the models tested. This coincides with a regularization parameter value, λ, of 0.002 and 50% LASSO regularization and 50% Ridge regularization.

Figure 5: LASSO and Elastic Net Models - Validation Data Set Error for Shrinkage Amounts

The 25 features selected are listed in Table 1 in order of importance to the total energy load prediction based on the feature coefficient value in the multivariate regression equation.
Table 1: Features Selected (Listed in Order of Importance Based on the Coefficient Value); * denotes that the term is a combined feature

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>#</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conditioned Floor Area*</td>
<td>14</td>
<td>Window Area Weighted Window Solar Heat Gain Coefficient*</td>
</tr>
<tr>
<td>2</td>
<td>Building Enclosure Surface Area*</td>
<td>15</td>
<td>Temperature Setpoint -Cooling Setback</td>
</tr>
<tr>
<td>3</td>
<td>Area Weighted Lighting Power Density*</td>
<td>16</td>
<td>Opaque Wall Area (East)*</td>
</tr>
<tr>
<td>4</td>
<td>Temperature Setpoint – Heating Occupied</td>
<td>17</td>
<td>Opaque Wall Area (West)*</td>
</tr>
<tr>
<td>5</td>
<td>Supply Air Temperature Cooling</td>
<td>18</td>
<td>Chiller COP</td>
</tr>
<tr>
<td>6</td>
<td>Area Weighted Electrical Plug Load*</td>
<td>19</td>
<td>Elevator Design Level</td>
</tr>
<tr>
<td>7</td>
<td>Infiltration Rate – Ventilation System Off</td>
<td>20</td>
<td>Area Weighted Occupancy Density*</td>
</tr>
<tr>
<td>8</td>
<td>Temperature Setpoint – Cooling Occupied</td>
<td>21</td>
<td>Fan Efficiency</td>
</tr>
<tr>
<td>9</td>
<td>Temperature Setpoint – Heating Setback</td>
<td>22</td>
<td>Building Orientation</td>
</tr>
<tr>
<td>10</td>
<td>Area Weighted Building Enclosure U-Value*</td>
<td>23</td>
<td>Fan Pressure Rise</td>
</tr>
<tr>
<td>11</td>
<td>Conditioned Building Volume*</td>
<td>24</td>
<td>Outside Air Rate</td>
</tr>
<tr>
<td>12</td>
<td>Boiler Efficiency</td>
<td>25</td>
<td>Infiltration Rate – Ventilation System On</td>
</tr>
<tr>
<td>13</td>
<td>Window-to-Wall Ratio*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The test data set, 20% of the 4,000 EnergyPlus simulated samples, was used to determine the accuracy of the final model. As shown in the Box and Whisker plot in Figure 6, the model under and over predicts the total energy load uniformly, with a median percent error on the test data close to zero. Approximately half of the test samples were predicted by the trained regression within 5% of the EnergyPlus simulated total energy load. There are a few outlying data points with prediction percent error in the 20-35% range. These data points fall in the lower range of total energy load values where a small difference between the predicted and actual energy loads can cause a high percent error. The selected model has an average percent error of 6.49% and a R²-score of 0.98 on the test data set.

To confirm that the data set size used was appropriate for the study performed, the test data set average percent error was calculated at decreasing training set sizes as shown in Figure 7. The test data set remained the same and the training data set was taken as random subsets of the original training data set. The graph shows that below a training set size of approximately 500, the Elastic Net multivariate regression model accuracy decreases significantly. Therefore, the training set size selected for this study was appropriate and could have been decreased to reduce computer simulation and model training time.

Discussion and Conclusions

This study illustrates a method for reducing the machine learning algorithm input matrix size used to predict total energy load. The features selected through this process will vary for different building archetypes and climates. In addition, the features with the highest importance to subsets of the total building energy load, such as heating and cooling energy loads, will likely vary. The goal of this study is not to provide a final list of features for building energy surrogate modelling, but instead to introduce a method to reduce the feature subset each time a surrogate model is developed.

Reducing the input matrix size by selecting combined features using a forward stepwise selection method, produced a less complex multivariate regression model with a higher overall total energy load prediction accuracy. This step combined all the building enclosure thermal performance metrics into a single term, and the building geometry-descriptors into six terms.
Previous studies have separated wall, roof, below-grade, and window thermal performance metrics into multiple features, thereby adding model complexity and increasing the risk of data overfitting. The results of this study show that the model accuracy can improve when these building attributes are combined into a single feature in the input matrix.

Using the LASSO regulator as a feature selection method, embedded in the multivariate regression learning algorithm, led to the removal of features with low-importance to the prediction of total energy load and improved the overall model accuracy. When combined with the Ridge regulator in Elastic Net, the model accuracy further improved. The Latin hypercube sampling method used in this study can cause the independent features to be collinear/multicollinear. The Ridge regulator reduces the regression coefficient values, reducing coefficient variance and thereby reducing the impact of collinearity/multicollinearity.

The model developed through this study overall has a high prediction accuracy. However, the model prediction errors can be over 20% for individual data points. There is a risk that by transforming both the input and output to force linearity of the data, that some behaviours of the data are not being captured. Future research will evaluate using non-linear machine learning algorithms on this data set to compare the models’ performance and accuracy.

An interesting outcome of this study is that the building geometry description features included in the final model are not specific to the shape of the building or the interior zoning. More general building geometry terms will allow for flexibility when using this model as an early-stage design tool as the exact geometry is not required. Further research is required to determine if the model can apply to buildings with geometry that differs from the DOE Large Office Commercial Reference Model’s rectangular form.

There are some key considerations that must be taken when carrying out this feature selection methodology. Although the intent is to use statistical methods to select a feature subset representative of the building energy load, the initial feature set (in this study, the original 71 building attributes) must be chosen by the researcher. It is important to use knowledge of the building energy domain when choosing the initial features as well as their respective ranges. Also, as shown in Figure 7, the sample size used to train the regression model has a large impact on the model accuracy. Since the optimal training sample size can vary for each surrogate model, it is recommended that a larger sample size be generated, and the final model checked with differing training sizes to confirm that the selected size was appropriate.

This study tested the feature selection method presented using the total site energy load for large commercial office buildings in Toronto, Canada. Future research will evaluate this method for individual energy end-uses, such as annual heating and cooling energy loads, as well as other building archetypes. Surrogate models where annual heating and cooling energy loads are the target variables will remove the direct electricity loads such as lighting, plug, and fan loads and will isolate those loads dictated by the building.

**Acknowledgement**

This research is funded by the Ontario Centres of Excellence Voucher for Innovation and Productivity I program, Grant #30783 and by RDH Building Science Inc. The student author is also receiving funding from an Ontario Graduate Scholarship.

**References**


