BUILDING ENERGY SURROGATE MODELLING METHODOLOGY FOR A DETACHED SINGLE-FAMILY CENTURY HOME ARCHETYPE IN TORONTO, ON

by

Cecilia Skarupa

Mechanical Engineering (B.A.Sc.), Queen's University, 2017

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the program of

Building Science

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

BUILDING ENERGY SURROGATE MODELLING METHODOLOGY FOR A DETACHED SINGLE-FAMILY CENTURY HOME ARCHETYPE IN TORONTO, ON

Master of Applied Science 2020, Cecilia Skarupa
Building Science, Faculty of Engineering and Architectural Science, Ryerson University

ABSTRACT

A surrogate model was developed for a detached archetypal home in Toronto, ON. EnergyPlus was used to perform 1500 simulations within a design space defined by 23 input parameters with ranges based on field study data. Elastic net regression was used to create a surrogate model to predict annual energy use and to perform embedded feature selection. An analysis comparing house size to model performance found that including both small and large homes did not decrease the model accuracy. The final regression model predicted energy use with an average R^2 of 0.946 and MAPE of 6.1% using nested cross-validation. A case study predicted actual annual energy use of two homes in Toronto within 10% error of utility bill data. A preliminary optimization analysis found that several weeks of simulation time could be saved and more optimal solutions could be discovered compared to a brute-force forward stepwise selection optimization.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Russell Richman, for his invaluable guidance and support throughout every stage of this process. Without his continual encouragement, this research would never have been realized to its full potential. Thank you to Dr. Mark Gorgolewski and Professor Jenn McArthur for taking the time to read this thesis and provide invaluable feedback.

The contribution of The Pocket Community Association and the residents who welcomed me into their homes to collect data for my research is truly appreciated. Thank you to my fellow students and work colleagues for your continual support and encouragement. To Erica Barnes – thank you for inspiring my interest in this field. Finally, thank you to my family and friends for always being there as I could not have done this without you. I am truly grateful.

Table of Contents

List of Tables

List of Figures

1 INTRODUCTION

Residential buildings account for 17% of the secondary site energy use in Canada [1]. Single-family homes make up 53.6% of private dwellings in Canada in 2016 [2], and 75% of the homes that will exist in 2030 have already been built [3]. Lowering the energy consumption and greenhouse gas (GHG) emissions in existing homes is essential to achieve the Intergovernmental Panel on Climate Change's (IPCC) requirement for net-zero buildings by 2050 [4]. Understanding our existing building stock and analyzing how to drastically reduce energy use is crucial to this goal. Deep retrofits that involve extensive renovations of the building's systems must be undertaken. These retrofits can have substantial costs, thus determining the combination of retrofit solutions that minimize cost and maximize energy and GHG reductions is essential to incentivize these changes.

To define the best strategy, a national net-zero retrofit plan is needed. To develop a retrofit plan a set of archetypes (used to describe subsets of the housing stock with similar characteristics) that describe the Canadian housing stock must be defined, as there are too many buildings to model individually. Archetypes increase the feasibility of scaling to a national level. Energy conservation measures can be tested and optimized on archetype models, and an optimal set of retrofit solutions can be determined. This research contributes an initial investigation towards describing the Toronto housing stock using a bottom-up archetype development framework that has the capability to be scaled to regional, provincial, and national levels. Figure 1 indicates the overarching goals for this research and a brief description of each step.

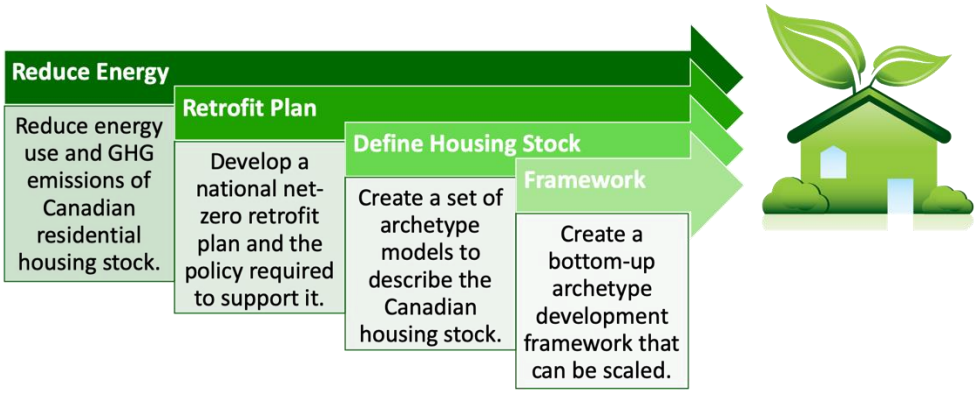

Figure 1 Overarching goals guiding research objectives.

Bottom-up models use detailed housing descriptors to calculate or simulate energy use, which is then scaled to represent trends in larger housing stock subsets. Building energy simulation software is used in physics-based bottom-up models to mathematically model complex systems and predict the energy use of a building. Each simulation can take a few minutes and requires an expert who understands how to use the software. The programs do not always produce accurate results as they rely on assumptions and variability in user expertise. Current bottom-up archetype models are limited by inaccuracies caused by generic assumptions being used to divide the housing stock without sufficient data, and computational requirements to optimize whole building energy models [5]. Fast and accurate tools to predict annual energy use are needed to simplify this process, allowing informative decisions to be made regarding energy conservation measures.

Surrogate models are computationally inexpensive tools that can be used to aid early design decisions for optimization, or perform an uncertainty or sensitivity analysis [6]. They calculate energy use and act as rapid approximations based on an original full building energy simulation model. As Hygh *et al*. describes it, "the resultant multivariate linear regression model is based on a set of detailed simulations that take into account the complex thermal interactions represented within a full scale energy simulation engine, but once developed, can operate independently of

the original, full scale model" [7]. Surrogate energy models are created using supervised machine learning methods. A supervised machine learning algorithm "sees" the input parameters and associated energy output value. It can then determine mathematical relationships that describe the change in the input parameter to the change in the output parameter and predict energy use if new input parameters are given. In a review of using surrogate models to predict energy use, Westermann and Evins found that they could be used to considerably reduce computational time during optimization studies [8]. Hester *et al* explain that not only is the speed beneficial, but having less input parameters required to get an energy use estimate is another advantage [9].

A 3-storey detached century home in Toronto, ON was identified as an archetype that represents 45% of Toronto's detached single family homes by Jermyn [10]. Jermyn developed a baseline energy model for this archetype using average measurements from a field study. The houses that were part of the field study were large versions of this archetype, however a smaller version of this archetype is also very common in Toronto. The detached century home archetype was chosen to explore the use of a surrogate model to describe the energy use of many Toronto homes. Data collected for this thesis from *small* century homes was combined with Jermyn's field study data [10] for *large* century homes. This allowed a larger subset of houses to be captured by one model. It was not known if the surrogate model would be able to predict as accurately when combining both sizes, or if separate models would be more suitable. An investigation was completed to determine how the model would predict on small versus large homes, and if it was appropriate to combine the wide range of sizes into a single model. This begins to address what levels of variation can be modelled by a single surrogate model. This is important because the feasibility of scaling this framework to include larger subsets of the housing stock depends on the level of granularity required to define each archetype. The larger

number of houses that can be represented by a single surrogate model, the more feasible this framework becomes to describe an entire housing stock.

The surrogate model was created by varying a set of parameter ranges and simulating the energy use for each set of inputs. The parameter ranges were developed using Jermyn's collected data [10] and from a field study completed for this research. These ranges along with the new baseline energy model was used to create the surrogate model. This process is shown in Figure 2.

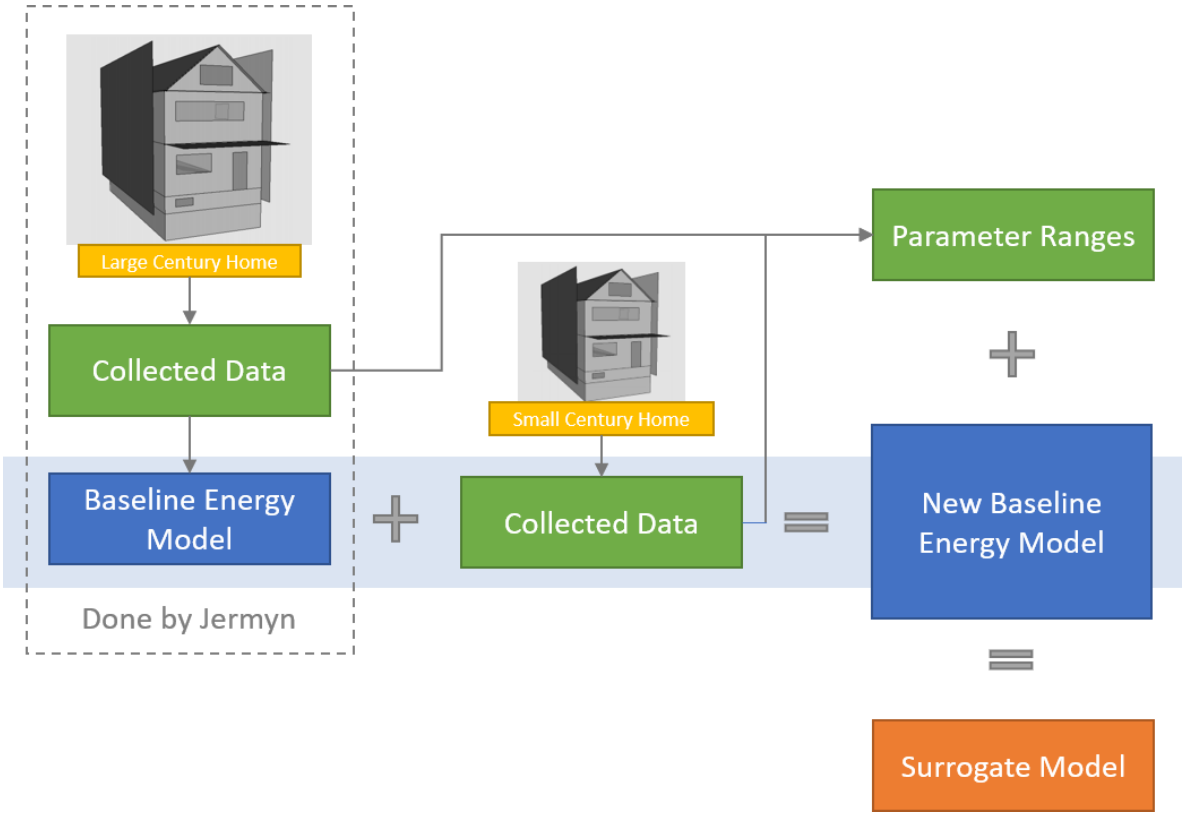

Figure 2 Process to develop surrogate model using existing baseline archetype model.

This research concludes with a case study using utility data from two small century homes and two examples of optimization techniques using the surrogate model.

1.1 Research Objectives

The objective of this research was to create a surrogate model to describe the annual energy use of an archetypal single-family detached century home common in Toronto, ON, and to quantify what accuracy could be achieved. This archetype included both small and large century homes. The size variation in the surrogate model and how it affected the accuracy of the predictions was analyzed. The objective was to determine if wider ranges of house sizes could be incorporated into a single surrogate model. While the accuracy of the model was important, another main objective was to focus on the simplicity of the model. This allows the model to generalize better to new data and it reduces the scope for future research. A simple model is more interpretable, which allows for transparency and a wide range of applications. The final objective was to consider the use of building energy surrogate models to support bottom-up archetype development that can be scaled to represent the Canadian housing stock.

1.2 Research Questions

This thesis aims to answer the following research questions:

1. With what accuracy can a surrogate model developed using multivariate linear regression describe the simulated annual energy use of an archetypal detached single-family home in Toronto, ON?
2. Can a small and large archetype with the same form be described by a single surrogate model, or do separate models provide more accurate results?

To answer these research questions, this thesis followed a phased approach. Phase 1 involved collecting data from archetypal detached century homes in The Pocket neighbourhood in Toronto, ON. Input parameters and ranges were determined to allow a sampling plan to create

data for 1500 “homes” which were run through EnergyPlus to create the dataset required to develop the surrogate model. Phase 2 included creating a surrogate model to predict energy use, an investigation on the impact the size range of the archetype has on model accuracy using two methodologies, reducing the number of input features, testing the model using real utility data, and a preliminary optimization example.

1.3 Thesis Structure

Chapter 2 focuses on the background of the existing archetypal energy model, how the dataset is created, and how the surrogate model is developed and validated. Chapter 3 is a review of existing literature on bottom-up housing stock models and energy surrogate models. Chapter 4 describes the methodology of this research, which is broken into two phases. Phase 1 focuses on developing the dataset used to create the surrogate model. This includes data collection, developing input parameters and associated ranges, devising a sampling plan, creating a baseline energy model, running 1500 simulations, and analyzing the synthetic dataset. Phase 2 describes creating and validating the surrogate model. This includes a house size analysis to determine the impact that size of an archetype has on model performance, preprocessing and transforming the data, comparing different training algorithms, developing the final model, a case study using utility data, an optimization example for the houses used in the case study, and a comparison to Jermyn’s brute-force optimization methodology [10]. Chapter 5 states and discusses the results of each step and follows the same two phases. Chapter 6 explains future work that could be explored and the key findings from this research.

2 BACKGROUND

This Chapter begins with a background of the century home archetype explored in this research. The century home archetype is used to develop a baseline energy model. This baseline energy model is used to create a dataset that is used to develop the surrogate model. The subsequent sections describe the process of creating the dataset and the surrogate model.

2.1 Century Home Archetype

In 2010 Blaszak and Richman [11] developed four archetypes commonly found in Toronto, ON: century detached, wartime, 1970's OBC, and modern. In 2013 Zirnhelt and Richman [12] developed a methodology for modelling and calibrating a single-family home in EnergyPlus. In 2013 Mucciarone [13] designed and tested retrofit wall assemblies and completed a hygrothermal analysis using WUFI. In 2014 Jermyn and Richman [10] developed archetypes for century detached, century semi, and wartime homes, and performed brute-force optimization to determine the most cost-effective retrofit solution to meet a specific energy performance target. In 2016, Niger [14] followed Jermyn's methodology [10] to develop an archetype and optimized retrofit solutions for a 1970's OBC archetype. Blaszak [11] and Jermyn's [10] work led to the Toronto Archetype Project (TAP) initiative created in 2015 by Ryerson University. The Pocket community [15] partnered with TAP in 2016 to allow researchers to use the neighbourhood as an example for archetype classification and net-zero community energy planning. The Pocket – bordered by the CN train tracks, Jones Ave, Danforth Ave, and Greenwood Ave – is on a mission to reduce energy use in their neighbourhood [15]. The relationship between Ryerson University and The Pocket facilitated access to many of the homes to enable a field study on the archetypal homes. The next step in the work done by Blaszak [11] and Jermyn [10] with the

Toronto Archetype Project (TAP) was to investigate a bottom-up modelling approach of the single-family residential stock of Toronto, ON.

Jermyn's methodology for archetype development [10] followed three phases which have been simplified as (1) determine archetypes, (2) collect characteristic housing data to build and calibrate baseline energy models, and (3) develop retrofit strategies and associated costs. Jermyn used the model development procedure proposed by Zirnhelt [12] to model the century home archetype in a Toronto neighbourhood. Jermyn collected data for geometry, envelope constructions, airtightness, internal gains, and HVAC systems. This data was averaged and input into an energy model to create a baseline archetype model. Utility data from the same houses were used to calibrate the model. A brute-force sequential search method was performed using the calibrated baseline model to determine the best retrofit strategy for that archetype. [10]

The Century home archetype described by Jermyn [10] is found throughout the city of Toronto. Jermyn determined that 45% of 33,570 single family homes in Toronto were represented by the century home archetype [10]. There are slight changes in the archetype in different neighbourhoods, such as size. The Pocket neighbourhood has many century homes described by Jermyn [10], however they are much smaller. Figure 3 shows examples of the detached small century archetype in The Pocket.

Figure 3 Examples of century homes in The Pocket neighbourhood in Toronto, ON.

The archetype examined in this research is a detached century home, as defined by Jermyn's archetypal development work [10]. The century homes measured in The Pocket are defined by the characteristics shown in Table 1.

Table 1 Characteristics describing century home archetype in The Pocket.

Characteristic	Century Home
Structure	Wood Frame
House Type	Detached
Number of Storeys	3
Year of Construction	1900-1930
Building Footprint	Rectangular
Roof Type	Peak Roof
Heating	Gas Furnace

2.1.1 Brute-Force Comparison

The baseline archetype model was used by Jermyn to perform a brute-force analysis to determine what retrofits were needed to reduce the energy use intensity to 75 kWh/m2 and 22 kWh/m2. The rule that Jermyn used to decide which retrofit level would be selected at each step was to minimize cost of retrofit per kWh of energy reduction. An energy model was run manually,

upgrading each parameter to Level 1. The retrofit upgrade with the lowest cost per kWh saved would become the new baseline, and the process was repeated until the desired levels were reached. This process was manual, tedious, and time consuming. [10]

The forward selection method selects an optimal solution for that particular baseline; however the true optimal solution is unknown. There could be a combination of upgrades that can get to 75 kWh/m2 with lower costs, but this solution is never presented because of the order that was selected. Surrogate modeling can make the optimization process faster and more versatile and can produce more optimal solutions.

2.2 Simulation Software and Input Modification

There are many simulation software that can calculate the energy use of buildings given enough input information, such as EnergyPlus, DOE-2, Ecotect, and TRNSYS. EnergyPlus is commonly used for surrogate model development. The text file input for EnergyPlus is called an IDF (input data file) and contains a specifically formatted list of the inputs to the simulation software. Figure 4 shows the IDF editing program provided by EnergyPlus, where an object (ex. "*Obj1*") has a list of fields to describe it (ex. "*Name*"), and each object belongs to a class (ex. "*Building*").

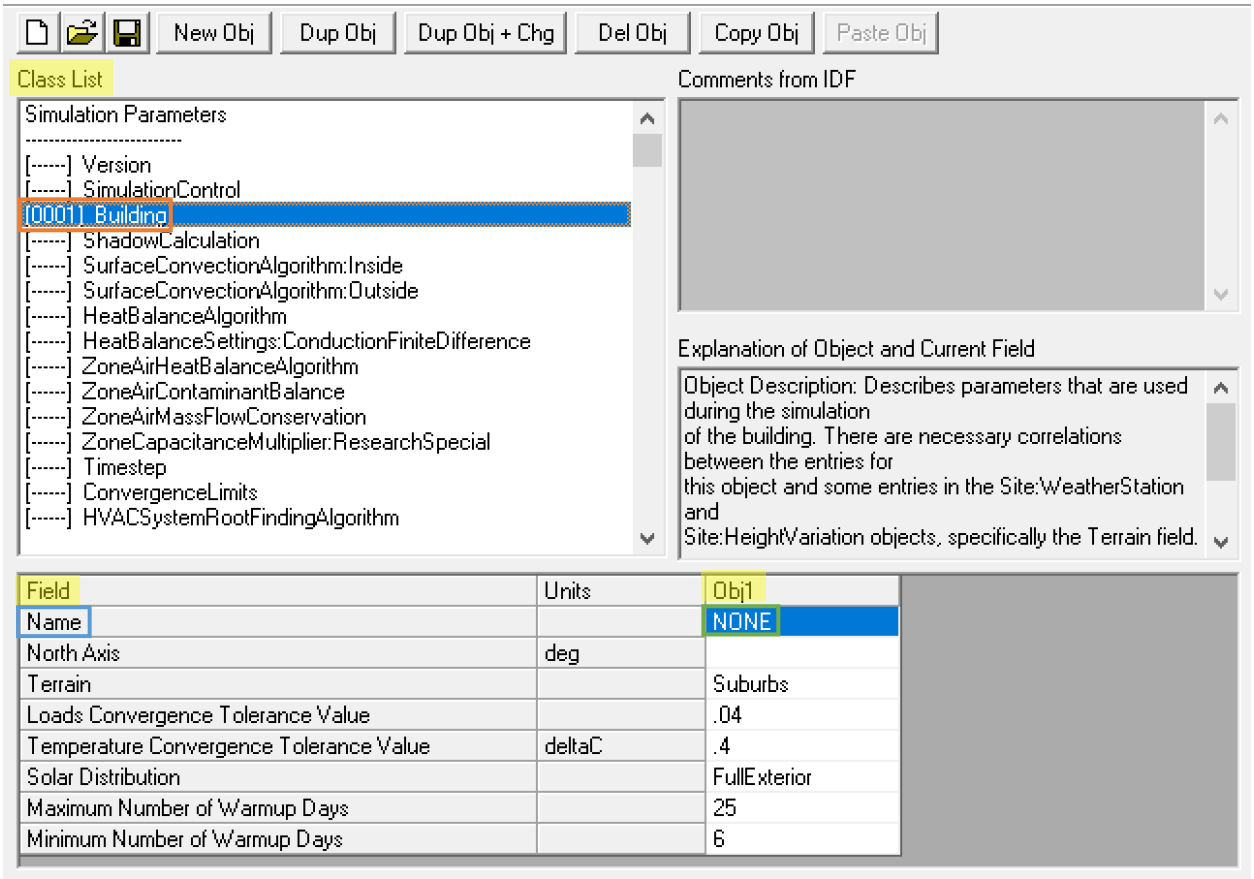

Figure 4 EnergyPlus IDF Editor showing classes, fields and objects.

This information can also be viewed as a text file. This text file can be easily edited using a Python script, allowing many updates to be made quickly. The building object shown in Figure 4 can be edited using Python and the *EPPY* package [16], a Python add-on designed to update IDF files. If the IDF is named "IDF", Figure 5 shows how the *Name* field of *Obj1* in the class *Building* can be updated to "Century Home".

```
1  building=IDF.idfobjects['Building']
2      building.Name='Century Home'
```

Figure 5 Python script to update EnergyPlus IDF.

This process can be repeated for any field in any object and can be used to calculate what the value should be based on other fields. For example, a ventilation rate could be calculated using the volume and airtightness already defined in the model.

2.3 Data Analysis

Tsanas *et al*. underline the importance of statistical analysis before machine learning, and argue that this step is often skipped in similar research [17]. After the data is collected, the next step is data analysis and visualization. This is accomplished using scatter plots, violin plots, Pearson correlation, and variance influence factor (VIF). Scatterplots and violin plots can provide a visual representation of the relationship between input and output variables, as used by Tsanas *et al* [17]. The Pearson's product-moment correlation coefficient acts as a quantitative measure of the strength and direction of an association. It describes the strength of the relationship between two parameters. The sign indicates the proportional relationship, and the magnitude indicates the strength of the relationship. The Pearson's correlation coefficient is always between -1 and 1 [18].

The Pearson's correlation coefficient can only calculate the relationship between two independent variables. The variance inflation factor (VIF) can be used to overcome this limitation, and allows the input data to be examined for multicollinearity [19]. Researchers have determined that VIF values above 5 or 10 indicate multicollinearity [19]–[21]. The VIF takes one predictor and regresses it against every other predictor in the model and calculates the coefficient of determination, R^2. The VIF is then calculated as shown in Eq. 1 [20].

$$VIF = \frac{1}{1-R_i^2} \quad (1)$$

2.4 Sample Size Analysis

A sample size analysis can be used to ensure that the number of samples is large enough to allow the model to produce accurate and reliable results. The analysis involves evaluating the training and validation performance metrics at different sample sizes. The error metric is calcualted for

decreasing sample sizes and plotted on the *y-axis* with number of training samples on the *x-axis*. When validation and training lines converge, the dataset is large enough. If the results show that there is a large difference between the training and validation values, this indicates that there is variance in the model and more samples are needed. If the results show that the error is high, this indicates that the model has high bias. High bias indicates that either a more complex model is required or more parameters. Figure 6 shows an example of a sample size analysis that illustrates high bias and low variance (left), and low bias and high variance (right).

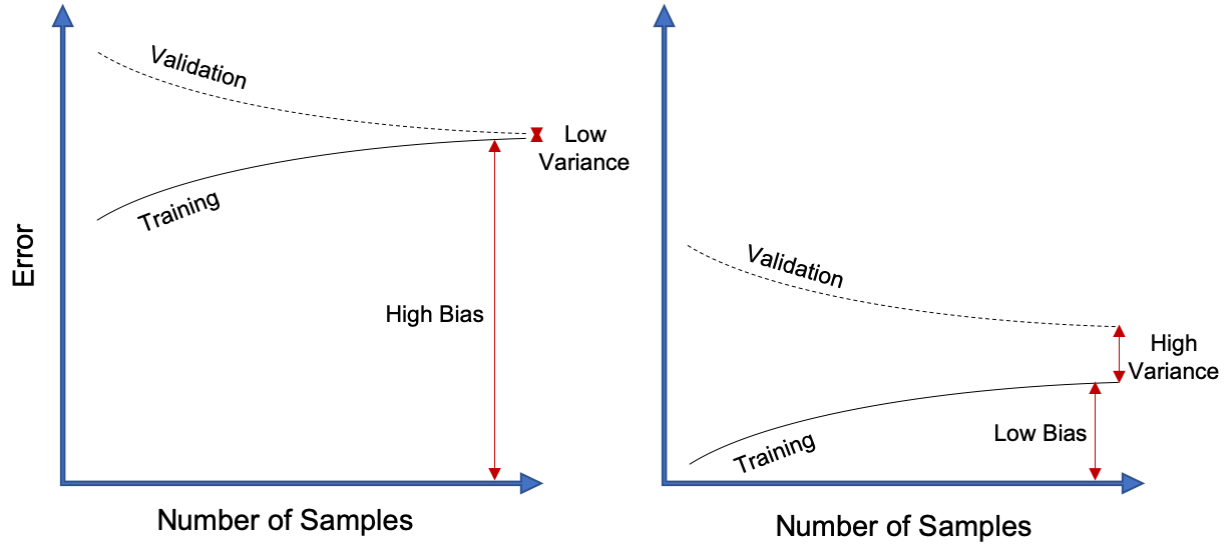

Figure 6 Sample size analysis showing high bias low variance (left) and low bias high variance (right).

2.5 Data Preprocessing

The input variables often have different units making it difficult to interpret regression coefficients. The variables can be standardized to allow them to be compared on equal scales. The mean and standard deviation of a set of input variables is calculated. If regularization is used, standardization is required. The inputs can be standardized by subtracting the mean and dividing by the standard deviation, as shown in Eq. 2 [22].

$$z = \frac{x_i - \mu}{\sigma} \qquad (2)$$

Where x_i is the input parameter values, μ is the mean, σ is the standard deviation, and z is the standardized value.

In linear regression, transformations of the input and output variables can be applied to increase the accuracy of the model. The most common transformations include log, square-root, and inverse. Tian *et al*. describe transformations as the best method to improve a model that has non-linear relationships between input and output parameters while maintaining the underlying linear structure of the regression model [23]. They found that a simple square-root transformation resulted in a greatly improved model. Visually analyzing the residual distribution can determine if linear regression was an appropriate method. If the residuals show evidence of non-linearity, the correct transformation can improve the validity of the assumptions required for linear regression to be appropriate. Standardization and transformations apply for continuous input variables.

Categorical inputs must be processed in some way before they can be handled by the learning algorithm. This is called encoding the categories which is accomplished using *label encoding* or *one-hot encoding*. *Label encoding* replaces a category with an arbitrary numerical value. It is not clear if *label encoding* is able to be processed by the learning algorithm as it imposes an ordinality that is not necessarily true. This could confuse the training algorithm and result in inaccurate coefficient values. To help the algorithm, they can be ordered using knowledge of their effect on the outcome variable. If there is no understanding of the category's effect, *one-hot encoding* is an alternative approach where no assumptions about the data must be made. This works by converting each category into a new parameter and assigning a 1 or 0 to indicate true or false. This allows the model to analyze the effect each category has on the output variable.

2.6 Training Algorithm

There are many algorithms that can be used to predict annual energy use. The most popular methods among surrogate building energy models are linear regression (including lasso, stepwise, and polynomials), artificial neural nets (ANNs), gausian processes (GP), multivariate adaptive regression spline (MARS), support vector machine (SVM), and radial basis function (RBF) [8].

Using a highly complicated algorithm such as ANN could potentially fit a training set perfectly, however it would not necessarily generalize well to new data. The model could be fitting the noise instead of the trends. The "bias-variance trade-off" describes the balance between model complexity and accuracy [22]. The application of the surrogate model will determine which training algorithm best fits for that situation. Although the literature has proven that there are more accurate algorithms, linear regression has the benefit of being highly interpretable. Figure 7 describes the trade-off between interpretability versus flexibility of a model. Lasso is regularized linear regression.

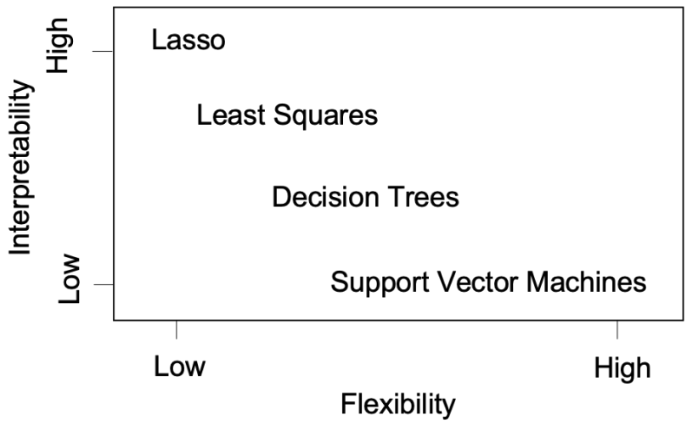

Figure 7 Interpretability versus flexibility of different training algorithms. Figure adapted from [22].

Interpretable models allow conclusions to be made based on the assigned weightings of each input parameter. Flexible models have many hyperparameters that can be used to tune the model to behave in many ways.

2.6.1 Linear Regression

The goal of linear regression is to calculate coefficient values so that a linear model can predict an output value as accurately as possible. This is generally accomplished by minimizing the sum of the squared errors [22]. In the case of multivariate linear regression, the line of best fit becomes a plane. The equation to describe the multivariate linear equation would be in the form shown in Eq. 3.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (3)$$

Where β_0 is the intercept, β_p is the coefficient value for the input x_p, and p is the number of variables.

2.6.2 Elastic Net Regression

Regularization, also known as shrinkage, is a method of reducing variance in a model [22]. Hastie *et al*. explain "the estimated coefficients are shrunken towards zero relative to the least squares estimates" [22]. There are different shrinkage methods, and sometimes the coefficients can be shrunken to equal zero effectively removing them from the equation. The plot on the left of Figure 8 shows an overfitted model. The plot on the right is the same model after applying a

regularization penalty (*i.e*. shrinkage). This results in a reduction of variance.

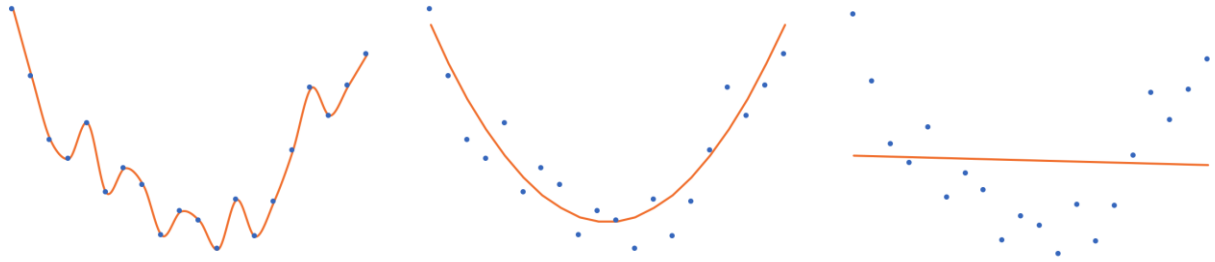

Figure 8 Overfitted model (left), good bias-variance trade-off (middle), underfitted model (right).

Bias refers to model complexity and how well it fits a training set. The plot on the left shows a low bias model that has minimized error in the training set, however it has picked up on the noise instead of the underlying trends. This model would be inaccurate when predicting on unseen data. The plot on the right shows a high bias model that has missed the trends completely and will have a high training error. The model in the middle has a balanced amount of bias so that the error on the training and testing sets will be similar.

Ridge, lasso (least absolute shrinkage and selection operator), and elastic net are all forms of regularized linear regression. The ridge coefficients aim to minimize the function in Eq. 4 [22].

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (4)$$

Where $\lambda \geq 0$ is the tuning parameter and $\lambda \sum_{j=1}^{p} \beta_j^2$ is the shrinkage penalty.

The lasso coefficients aim to minimize the function in Eq. 5 [22].

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (5)$$

Where $\lambda \geq 0$ is the tuning parameter and $\lambda \sum_{j=1}^{p} |\beta_j|$ is the shrinkage penalty.

The tuning parameter controls the impact of the shrinkage penalty. There will be a different set of coefficients every time this value is changed. If the tuning parameter is zero, the least squares fit is being computed (linear regression). As the tuning parameter increases, the coefficient values begin to decrease (this is "shrinkage").

Ridge will shrink the coefficients towards zero as the tuning parameter increases, however no coefficient value will ever reach zero unless $\lambda = \infty$ [22]. With lasso, the tuning parameter could be increased until all coefficients became zero. Ridge will always include all the parameters in the final model. The number of parameters included in the final lasso model will depend on the tuning parameter value. Edwards *et al*. explain that lasso reduces the number of coefficients which results "…in a sparser, more robust model. Note that robustness is defined based on the idea that a simplistic model is most likely to generalize to new scenarios" [24]. Edwards *et al*. based this on model complexity studies [25]–[27]. Tian *et al*. [23] believe that using lasso regression for feature selection could replace the conventional stepwise method that has been so widely used in this field up to this point. This is referred to as embedded feature selection because features are "selected" within the learning algorithm [28].

Elastic net combines the lasso penalty and ridge penalty and minimizes the function in Eq. 6 [29].

$$RSS + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}\left|\beta_j\right|\right) \qquad (6)$$

Where α is the L1-penalty coefficient, and $(1-\alpha)$ is the L2-penalty coefficient. The regularization method is ridge if α = 0, lasso if α = 1, and elastic net if α is between 0 and 1. The tuning parameter and the L1-penalty coefficient are considered hyperparameters and define a specific "model", since a change in either would result in different coefficient values.

Tian *et al*. [23] use the "one-standard-error" rule to determine which hyperparameters to use. This rule indicates that the chosen model should have an error that is no more than one standard error more than the error of the best model [30].

2.7 Validating Algorithm

A machine learning algorithm is trained on a set of data. The objective is to fit a model that most accurately predicts the target variable given the input variables. The model is validated by testing the model developed from the training set on a set of unseen data, referred to as a validation set. An entire dataset is split into training and validation subsets using different splitting methods. To ensure the results are not affected by the location of the split, (for example if the validation set were to contains all the outliers), the split is randomly repeated many times and the average evalutation metrics are used to describe the performance of the model. According to Kuhn and Johnson [31], repeating the resampling may produce different values, but if repeated enough times will estimate the true value.

If this is repeated iteratively, and the performance metrics of the validation set are used to choose the final model, the final model is being selected because it describes the *validation* set well. This is called data leakage. The model is performing well because it was chosen based on the highest validation scores. However, the validation set is just a small subset of data. When the model is passed new data, it may have much lower performance because the validation set was not a representative sample. To overcome this, a "hold-out" set, or testing set, is split off at the very beginning. This is never used by the model to make decisions about which algorithms or hyperparameters to use. The validation set can now be used to choose the model and tune the hyperparameters. When a final model is selected, it can be evaluated on the testing set – data it has never seen before – and the evaluation metrics are a more accurate indication of the model's

ability to predict on unseen data. A proper cross-validation methodology will reduce data leakage and can calculate the model accuracy close to its true value.

2.8 Model Performance Metrics

Root mean squared error (RMSE is) measured in the unit of the output variable and can be interpreted compared to the range of the output values. This makes it difficult to compare across models when the range is not stated, or the units are not the same. If the variables were transformed, for example, the RMSE before and after could not be compared. In existing literature, the values of the output variable(s) are often not stated with the RMSE.

Mean absolute percent error (MAPE) is measured as a percentage of the output variables. It can be compared no matter the unit or mean. MAPE is not always reported alongside R^2 and RMSE, although it is the most useful for comparing models and the most interpretable.

The coefficient of determination, R^2, describes the proportion of variance explained by a line of best fit. It is unitless and always falls between 0 and 1. R^2 can be determined using the total sum of squares and the residual sum of squares as shown in Eq. 7 and 8 [22].

$$R^2 = \frac{TSS - RSS}{TSS} \qquad (7)$$

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (8)$$

Where RSS is the residual sum of squares and TSS is the total sum of squares.

Hastie *et al*. explain that "TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the

regression. Hence, TSS−RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X" [22].

2.9 NSGA-II Optimization

Multi-objective optimization allows more than one objective function to be optimized. This can be accomplished using a non-dominated sorting genetic algorithm (NSGA-II) [32]. NSGA-II optimization can be used to find a Pareto front of solutions using a genetic algorithm. Figure 9 shows an example of a multi-objective optimization problem.

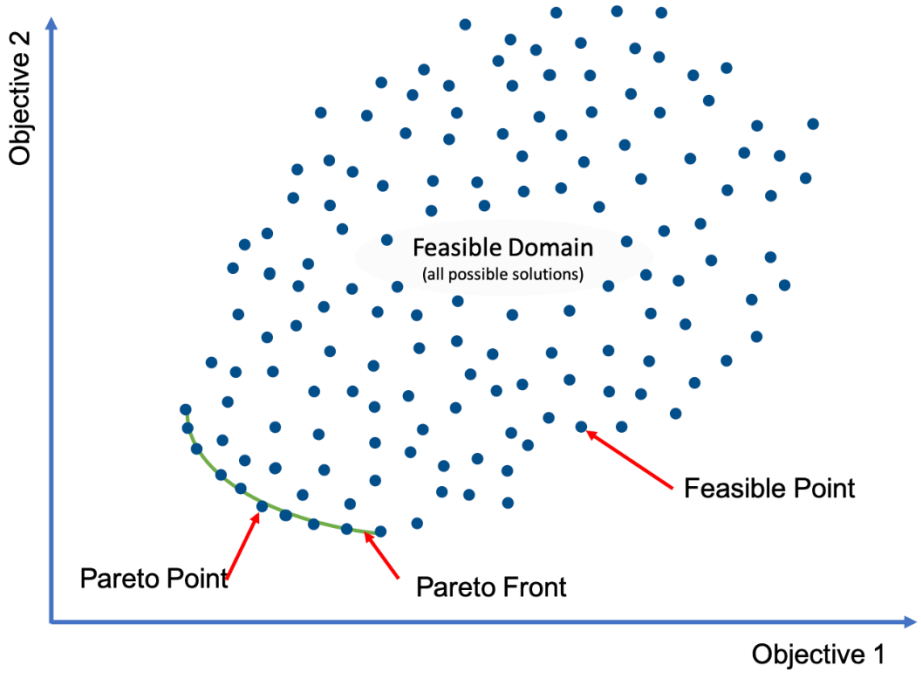

Figure 9 Pareto front diagram for minimizing two objective functions.

The two axes represent each objective. In a trade-off case where both objectives cannot be minimized, for example cost and energy use, there will be no single optimum solution. Retrofit solutions cost money, therefore as cost increases, energy decreases, and vice versa. The Pareto front is a set of solutions that try to find a balance minimizing both functions. The feasible

domain is all possible retrofit solution combinations. The genetic algorithm optimization presents the Pareto front, which is the edge of the feasible domain where both objectives are minimized. It outputs a set of Pareto points that fall along the Pareto front. It is then up to the user to pick a single point as the optimal solution

3 LITERATURE REVIEW

This Chapter begins with a review of existing bottom-up models describing the Canadian residential housing stock to outline the limitations of current methods. The rest of this section outlines current research on energy use surrogate models, including an overview, model intent, development, and validation. The final section summarizes the gaps in the literature that framed the research presented in this thesis.

3.1 Bottom-up Housing Stock Models

There are several different approaches that can be used to model energy consumption of a housing stock. Swan and Ugursal [33] describe the top-down and bottom-up approaches as the two main methods. The *top* and *bottom* refer to the "hierarchal position of data inputs as compared to the housing sector as a whole" [34]. The bottom-up approach considers small sets of houses with similar characteristics (archetypes), and results can be extrapolated to describe a larger subset of the residential housing stock [34]. Swan and Ugursal [33] explain that statistical methods and physical methods are the two groups used in the bottom-up approach. Statistical methods use historical data to identify end-use energy consumptions. This can be accomplished using regression, conditional demand analysis, or neural networks. Physical methods use inputs and known mathematical thermodynamic relationships to calculate end-use energy consumptions. Swan and Ugursal state that for this reason, physical methods are the only option for comparing the impact of retrofit options [33]. Foucquier *et al* [35] suggest hybrid models as a third category of the bottom-up approach. Hybrid models – also called grey box models – combine elements from the physical and statistical approaches.

The BC Housing "Step Code" [36] and The Canadian Hybrid Residential End-Use Energy and GHG Emissions Model (CHREM) [37] are two examples of bottom-up models used to describe

a housing stock. The BC step code is a physics-based model and describes the residential housing stock in British Columbia. The CHREM is a hybrid model and describes the Canadian single-family residential housing stock. These methods are limited by the rigidity of the models, the assumptions that were made, and needing to sample the database due to the computational requirements of simulating the entire design space.

The CHREM is a model based on building performance simulation of 16,952 unique houses that statistically represent the Canadian housing stock [37]. The input data for the models is a sample of real data collected from EnerGuide for Houses Database [38]. The physics-based method was used to simulate the thermal energy transfer and HVAC energy use, while the statistical method was used to estimate occupant-driven loads (appliance and lighting and domestic hot water) [37]. The purpose of the CHREM is to set a baseline energy consumption of the entire Canadian housing stock and allow researchers to model retrofit scenarios and compare the results. It takes approximately 68 seconds to run a simulation for one house, therefore it would take 13 days to run all the files consecutively (this could be shortened with better processors or more computers) and this would have to be repeated for each retrofit scenario investigated [37]. Note that not all houses have to be run for each simulation. There are five interconnected components that are required to run CHREM successfully, and each house is defined using 18-31 input files depending on the housing characteristics [37]. This model is sophisticated but complicated and hard to use. It would be difficult to use for a sensitivity analysis and is not ideal for optimization as the computational requirements would be enormous. Wills [39] adapted the CHREM and used it to perform an optimization algorithm but was limited to a single cost function due to the computational burden. The dataset is not very granular as the intent is to describe the entire Canadian single-family housing stock with 16,952 houses. The geometry inputs vary in size but

not type, as each house's geometry was simplified to be rectangular with a peak roof [37]. This would likely impact the results.

According to the final report "the BC Energy Step Code (the "Step Code") is an amendment to the BC Building Code (BCBC) that provides a performance-based path intended to support a market transformation from current energy efficiency requirements to net zero energy ready buildings by 2032". The retrofit solutions resulted in 54 million combinations of energy conservation measures. They used a dynamic sampling technique to model 60,000-240,000 simulations (which took 12 days). The limitations of this project are that not all the possible solutions were able to be evaluated due to the lack of computing power. Another limitation is that the solutions only applied to homes in British Columbia. [36]

3.2 Surrogate Model Overview

Barnes [40] and Westermann and Evins [8] completed detailed literature reviews of surrogate modelling for building performance simulation in 2019. Barnes [40] reviewed 22 papers and Westermann and Evins [8] reviewed 57. They collected detailed information about the intentions of the models, the model algorithms, the sampling techniques, and the input parameters used.

Westermann and Evins determined there were four stages of the building design process where surrogate models are commonly used. These are: the conceptual design stage, sensitivity analysis, uncertainty analysis, and design optimization. A sensitivity analysis is often an initial step to reduce computational cost before the other three processes. Surrogate models reduce the computational cost of building energy simulation allowing researchers to gain insight into building performance over a wide space of potential design or retrofit options. [8]

Magnier and Haghigat used artificial neural nets (ANN) to create a surrogate model and the NSGA-II optimization algorithm to optimize energy use and comfort. It took three weeks to run the simulations to build the dataset and 7 minutes to run the optimization. If they had simulated the same number of model evaluations that they optimized, the process would have taken 10 years. They were using small time steps and had a computer with low processing power so they noted that the three weeks of simulation time could be greatly reduced. [41]

Westermann and Evins found that in some situations, a trade-off between accuracy and model interpretability was favoured [8]. Ostergard *et al*. compared various machine learning algorithms commonly used to develop energy use surrogate models concluded that the "best" technique should be determined based on time, expertise, and required level of accuracy [42]. Westermann and Evins summarize that Ostergard *et al*. [42] advocate "the use of ANN for extensive analysis, GP for non-experts to get high accuracy, and MLR for quick, automated surrogate modelling" [8].

Figure 10 outlines the process found by Westermann and Evins to describe the surrogate model development process [8].

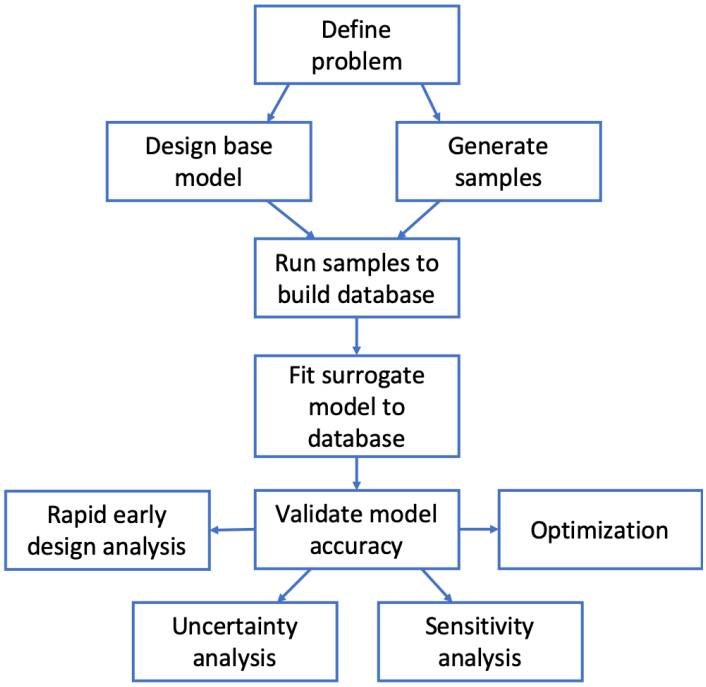

Figure 10 Surrogate development process according to Westermann and Evins. Figure adapted from [8].

There are additional steps to creating a surrogate model and development decisions must be made throughout the process. A few examples include deciding which target variable to predict, which energy simulation software to use, and which error metrics to report. The rest of this section will describe each of these decisions and the options found in previous research.

3.3 Surrogate Model Intent

According to Cerezo *et al*, an archetype is generally *classified* based on properties such as use or construction year and can then be *characterized* using average values for that group. The most commonly used classifiers are type of building, floor area, shape, and age of construction. Once the building has been classified, the archetype must be characterized for necessary parameters to enable energy simulation. This includes non-geometric building and occupant factors (envelope construction details, HVAC system properties, occupancy schedules, internal loads, etc.) that are based on averages of the archetype's parameters. Cerezo *et al*. state that literature data can be

used to characterize archetype parameters, however more granular building data could be collected for the archetype by an audit or field survey. [43]

Tsanas *et al*. [17] explain that modelling whole buildings to determine energy use is a widely used approach even though the result do not necessarily perfectly predict real energy use. They state that the results are an accurate indication of the percentage change and underlying trends and conclude that simulation results "represent actual real data with high probability and as such will be considered as ground truth" [17].

Chidiac *et al*. developed several archetypes for Canadian office buildings, assessed the applicability of energy retrofit measures (ERMs), and selected the most suitable ERMs for each archetype. Their methodology was developed to "simplify the ranking of buildings for retrofit, to select and combine ERMs, and to plan energy and GHG reduction activities" [44]. They used 12 ERMs based on current energy-saving industry standards. Individual equations to describe each end use load (lighting, equipment, pump, fan, domestic hot water, chiller, and boiler) were developed and summed together to calculate overall energy consumption. This allowed the author to gain an understanding of how an ERM affects the individual loads, and the resulting energy reductions. This study simulated individual variable changes but limited multiple interaction simulations to individual variables that effected energy consumption by more than 10%, and only 3 level interactions were considered. The results were used to calculate payback period using installation and material costs from RSMeans. This helped determine the retrofit potential for each archetype. [44]

Hygh *et al*. created a surrogate model for a medium-sized rectangular office building in 4 different climate zones. The model predicted total space heating and cooling and was created using EnergyPlus simulations using a Monte Carlo sampling of the design space. Forward

stepwise regression was used to add combined parameters to the model. They conclude that linear regression models can replace a full energy simulation model when making design decisions during early stages. They found that the regression model for total energy consumption was more accurate than summing the predictions for the heating and cooling models. They used 20,000 EnergyPlus simulation results used to generate the surrogate model, but an analysis of the prediction error versus the number of samples indicated that 1000 samples would be enough for the rectangular office building they modelled. The model was validated using 20% of the 20,000 samples. It is unclear if this result was cross-validated, and the parameters may have been selected on the same set that accuracy was assessed. This can sometimes lead to models with larger bias, *i.e*. not generalizing as well to future data. This emphasizes the importance of researchers in this field clearly reporting all of the steps taken to allow for other researchers to compare and analyze the work that has been done. [7]

Catalina *et al*. created a surrogate model for a multi-unit residential building in 16 cities in France. Their focus was on buildings forms. The model predicted the heating energy use and achieved R^2 values of 0.99 with average error of 2%. They found a quadratic polynomial regression had the best fit. They checked the residuals of the regression to ensure the assumptions of the linear regression were met. The model was validated on simulation results of different building forms that were not used in the training set. [45]

Asadi *et al*. used DOE-2 and Monte Carlo sampling to create a surrogate model for a typical office building in Houston, Texas. The model inputs focused on construction characteristics, shape, and occupancy schedule. A separate model was created for each of the seven building forms. They conducted 10,000 simulations per building and achieved R^2 values of 0.94-0.95 and

5% error. The model was validated using 20% of the 10,000 samples, however it was unclear if the results were cross-validated. [46]

Catalina *et al*. use polynomial regression to predict heating energy consumption in multi-unit residential buildings in Moscow, Bucharest, and Nice. The model had an R^2 of 0.974 and average error of 10% when validating on data from 17 real buildings. They claim that error of more than 30% is acceptable by designers. The only parameters they used were the building global heat loss coefficient, the south equivalent surface, and the difference between the indoor set point temperature and the sol-air temperature. [47]

Hester *et al*. created a surrogate model to predict energy consumption of single-family residential buildings in Chicago, Illinois. The goal was to develop a framework to make informed decisions to improve a building's performance based on early-design decision parameters. They used stepwise forward selection and linear regression and reported R^2 values of 0.968. [9]

Melo *et al*. developed a surrogate model to predict the annual cooling energy of commercial buildings in Florianópolis, Brazil. They tested several algorithms including multivariate linear regression, which performed with a NRMSE of 3.7%. They used LHS and EnergyPlus to perform one million simulations. They pre-processed their data by standardizing the input variables, applied a log-transformation, and *one-hot encoded* their categorical variables. [48]

Sangireddy *et al*. used residual analysis to ensure the appropriateness of their models. They simulated 100,000 input combinations in EnergyPlus for two cities in India. Different machine learning algorithms were trained on the Jaipur dataset and tested on the Hyderabad dataset. Using lasso yielded R^2 scores of 0.877 and a MAPE of 5.8%. [49]

Tian *et al*. modelled a 5-storey office building in London, UK and created a surrogate model to predict heating energy use. They used a distribution-free bootstrap sampling method to compute the sensitivity index variation in building energy analysis. They were able to achieve R^2 scores of 0.983 for multivariate linear regression. [6]

Sekhar *et al*. [50] used Tsanas *et al*.'s dataset [17] and achieves R^2 scores of 0.924 using multivariate linear regression to predict heating load. The input parameters described the building geometry only. They compared many different training algorithms.

3.4 Surrogate Model Development

A sampling plan is used to generate the input matrix from the defined ranges for each parameter. There are many types of sampling plans, such as Monte Carlo, Latin hypercube sampling (LHS), orthogonal array, and full-factorial. It has been suggested by Sacks *et al*. that space-filling sampling plans (such as LHS) should be used when error is systematic [51]. Building energy simulations fall under this category. Westermann and Evins determined that there was a strong preference towards Latin hypercube sampling (28% of the studies reviewed used LHS) [8]. LHS creates a matrix of near-random values that are space-filled which results in values that span the entire design space. This ensures that each building input parameter has an even number samples throughout the whole range. No research was found to indicate the affect of sampling plan choice on the performance of energy use surrogate models.

Barnes [40] determined that 64% of the papers reviewed used EnergyPlus (shown in Figure 11), and Westermann and Evins [8] found a similar value of 56% of the papers. Barnes determined that most of these papers used EnergyPlus for its capability to automate the modification of the IDF text file [40]. It is also the official energy analysis simulation program of the U.S.

Department of Energy, and it is based on first principles instead of simplified algorithms, which can avoid inaccuracies [7].

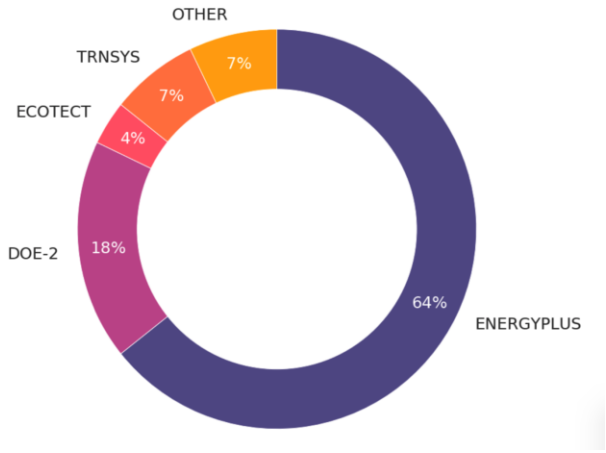

Figure 11 Percentage of energy simulation software used in surrogate modelling. Figure from Barnes [40].

Figure 12 shows the parameters that Barnes [40] determined were the most commonly used in the reviewed papers. The percent of studies that uses each parameters is shown.

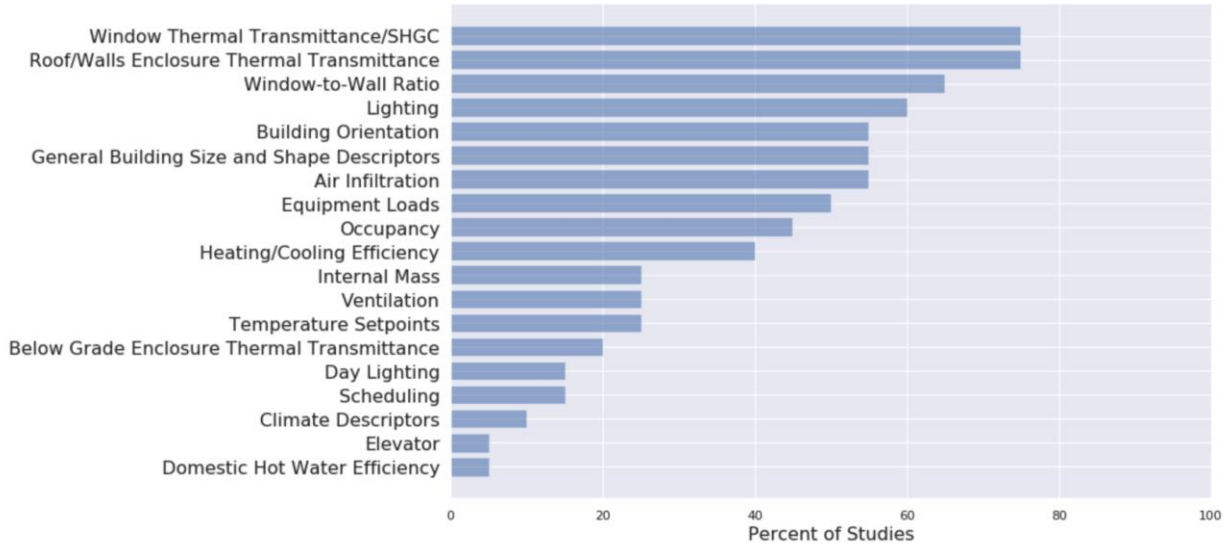

Figure 12 Percent of studies that use each input parameter. Figure from Barnes [40].

Westermann and Evins [8] found that most studies predicted annual energy demand, and the second most common target variable was heating and cooling. Barnes [40] found the opposite.

Several studies that did a sample size analysis with tens of thousands of simulation results concluded that 500-1000 samples would be large enough for a similar situation [7]. Westermann and Evins found that only 26% of the reviewed papers performed a sample size analysis [8]. This should always be done to assess how the sample size impacts the surrogate model performance.

Westermann and Evins found that while most researchers selected the most accurate model, sometimes trading accuracy for interpretability was favoured [22], [52]. Westermann and Evins [8] found that 33% of the papers used linear regression. Figure 13 shows the percentage of the other algorithms.

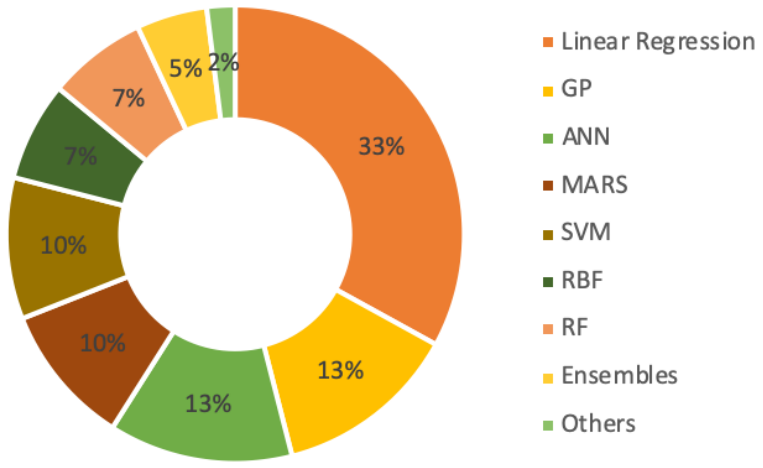

Figure 13 Percent of studies that use each learning algorithm. Figure adapted from [8].

3.5 Model Validation

Castelli *et al*. look at predicting heating and cooling loads in residential buildings from Tsanas *et al*'s dataset [17]. They propose a genetic programming-based framework. They validate their results using 10-fold cross-validation with 100 repetitions. [53]

Chou *et al*. used Tsanas *et al*'s dataset [17] and performed 10-fold cross-validation with 10 repetitions. They compared many different algorithms such as support vector regression,

artificial neural nets, classification and regression tree, etc. in terms of speed and performance when predicting cooling load and heating load. [54]

Hester *et al*. performed cross-validation using a training set of 60%, a validation set of 20%, and a test set of 20%. The validation set was used to tune the initial model developed on the training set. The test set was reserved for evaluating how well the model performed on unseen data. This avoids overfitting and ensures that the reported metrics are accurate. [9]

Melo *et al*. performed 8-fold cross-validation during model development. They also validated their model by testing it against simulation results from a medium office building energy model. [48]

Tsanas *et al*. used 10-fold cross-validation repeated 100 times. They did not have a testing and validation set, only one set was put aside and used to evaluate the algorithm, indicating that the models were validated and tested on the same data. [17]

Tian *et al*. explain that the simplest technique to evaluate prediction error is using a separate subset of the dataset for validation and testing. This method is widely used in building performance analysis. They state that there are several disadvantages to this method and that new methods are required to overcome the shortcomings. [6]

Krstajic *et al*. explain that model *selection* should have a different process (and separately reported performance metrics) than model *assessment*, but many researchers report the cross-validation error that determined which model was the best as the true model performance [55]. Varma and Simon [56] show that this practice gives significantly biased estimates of the true error. They describe the correct practice which requires the parameter tuning to be repeated in each cross-validation loop, and conclude that a nested cross-validation procedure “provides an

almost unbiased estimate of the true error" [56]. None of the studies reviewed included nested cross-validation. It should be noted that this is only required when using a model that requires hyperparameter tuning.

Westermann and Evins, who reviewed 57 papers on surrogate model development, did not mention cross-validation or lasso as embedded feature selection [8]. Barnes determined that only 20% of papers used cross-validation [40].

Barnes [40] determined the percent of evaluation metrics used in the reviewed papers. The most common reported metrics were R^2 and RMSE, followed by MAPE.

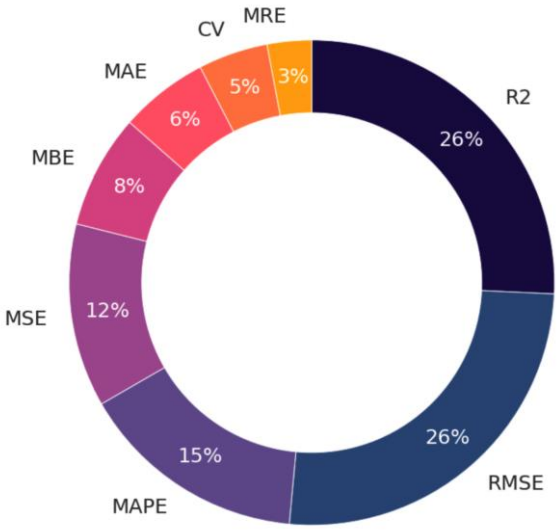

Figure 14 Percent of reviewed studies that use each model performance evaluation metric. Figure from Barnes [40].

Many studies report only R^2 or RMSE, which are not easy to compare. An error metric, such as RMSE, should always accompany a correlation metric, such as R^2. This is because high correlation does not necessarily indicate low error.

Barnes [40] determined that only 12% of surrogate models from the reviewed literature were for low-rise residential buildings. That category includes multi-unit buildings, which made up the

majority of those values. Figure 15 shows the percentage of the building types according to Barnes [40].

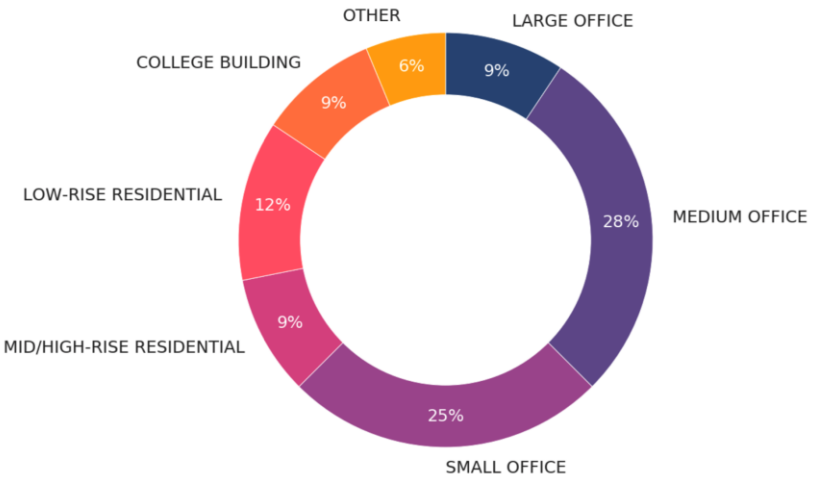

Figure 15 Percent of buildings used to make surrogate model in reviewed studies. Figure from Barnes [40].

3.6 Gaps in Current Literature

The research in this field so far focuses on commercial buildings or multi-unit residential buildings. Not much emphasis has been placed on single-family homes. Although commercial buildings use more energy per building and in total, residential buildings account for 17% of the secondary site energy use in Canada [1]. There is still a great deal of energy savings that can be achieved by retrofitting existing homes. Only one paper by Hester *et al*. [9] was found to look at single-family residential homes, but their objective was to guide sequential design decisions. They used a forward stepwise selection methodology to decide on input parameters. They did not consider ventilation rates or heat recovery efficiencies for input parameters and did not report coefficient values. They conclude that their methodology could be improved with surrogate models that incorporate a wider range of designs, or "by a method to more rapidly generate a metamodel for a particular building type and context" [9].

Many papers do not use cross-validation to choose hyperparameters or obtain true measures of model performance. Nested cross-validation was used in this research to obtain the most accurate model performance metrics. The decisions made in the surrogate model development process are not always made clear in existing literature. This makes it very difficult to learn from and compare to other research in this field. This research aims to be very clear about what decisions were made and what hyperparameters were used.

Many papers explore which training algorithm is the most accurate, however at this point that has been sufficiently covered. Multivariate linear regression is used for its simplicity and interpretability based on the objectives of this paper. No other training algorithms are compared.

None of the reviewed papers explore the bounds of surrogate models and how a variation in building types or size will affect the model performance (however some do explore different shapes). A house size analysis is completed in this research to determine the affect of house size on model performance. Capturing larger variations in houses that can be described by a single surrogate model greatly increases the feasibility of using surrogate models to describe an entire housing stock.

There is currently no bottom-up model of the Canadian housing stock that overcomes the current limitations. Models need to be transparent, flexible, and allow for multi-objective optimization. The use of a surrogate model coupled with the physics-based bottom-up approach will start to fill that gap. More archetypes will need to be developed to incorporate larger subsets of the Canadian housing stock.

This research proposes a methodology for creating a surrogate model for an archetypal single-family home in Toronto, ON. This research is the first step towards creating a bottom-up model that can describe Toronto's residential housing stock and other Canadian municipalities.

4 METHODOLOGY

Phase 1 focused on creating the dataset used in Phase 2 to develop the surrogate model. Phase 1 started with a field study of The Pocket neighbourhood to collect data. This data was used to update an existing archetype model to be the baseline model for the dataset. The data was used to determine a set of parameters and associated ranges that were randomly sampled using Latin hypercube sampling to create 1500 EnergyPlus input files (IDFs) to represent a set of houses within the defined century home archetype. The 1500 models were simulated using EnergyPlus and the annual energy use was appended to each set of input parameters as the output value. Figure 16 shows the outline of the methodology for Phase 1.

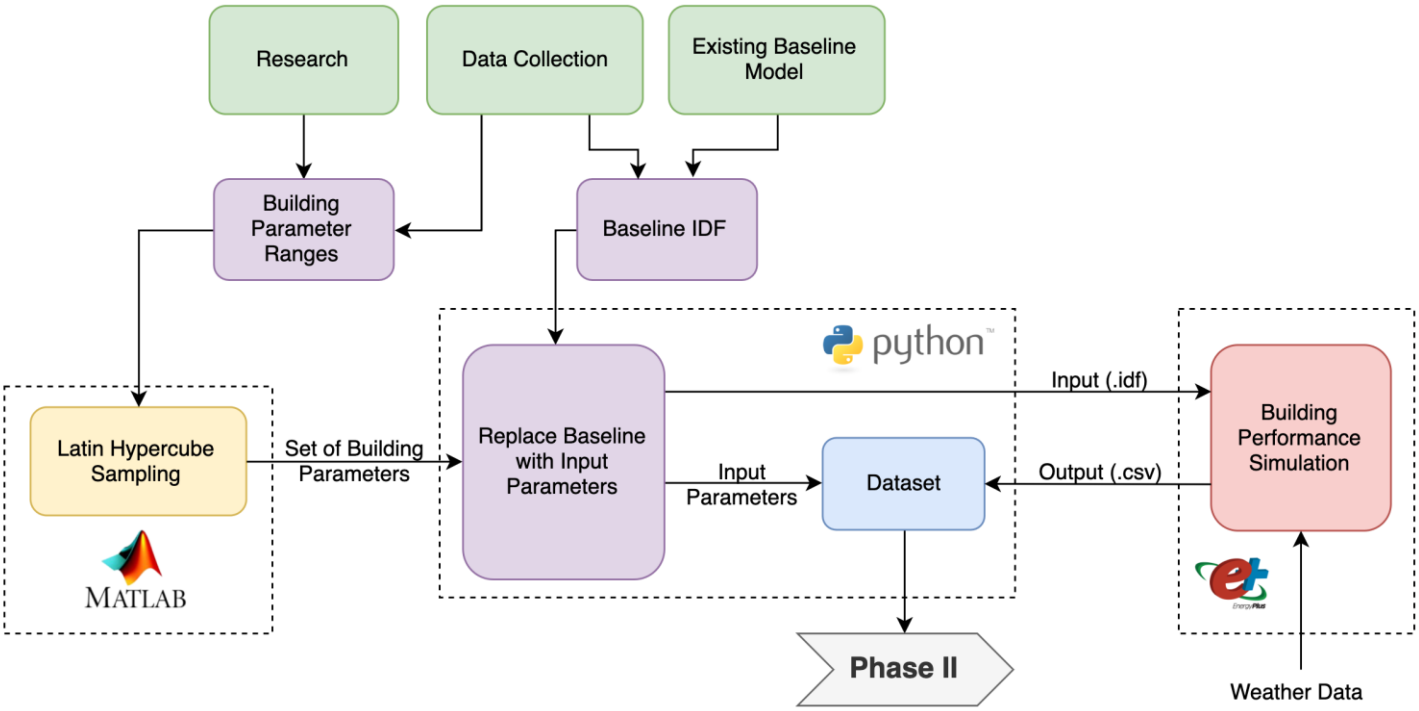

Figure 16 Outline of Phase 1 methodology.

Phase 2 used the dataset created in Phase 1 to develop a surrogate model using multivariate linear regression and regularized regression (elastic net). Phase 2 began with preprocessing the input and output parameters. A house size analysis was performed to determine if a large range of house sizes could be accurately captured by a single surrogate model. Four different models

using multivariate linear regression, elastic net, and different ways of processing the categorical variables (*one-hot encoding* versus *label encoding*) were created and compared. A final model was selected and trained, and the coefficient values were reported. A case study was conducted using the model and utility data from two homes in The Pocket neighbourhood. Finally, a preliminary optimization example was completed using retrofit and costing data from Jermyn's research [10] and compared to their brute-force optimization. Figure 17 illustrates the methodology for Phase 2.

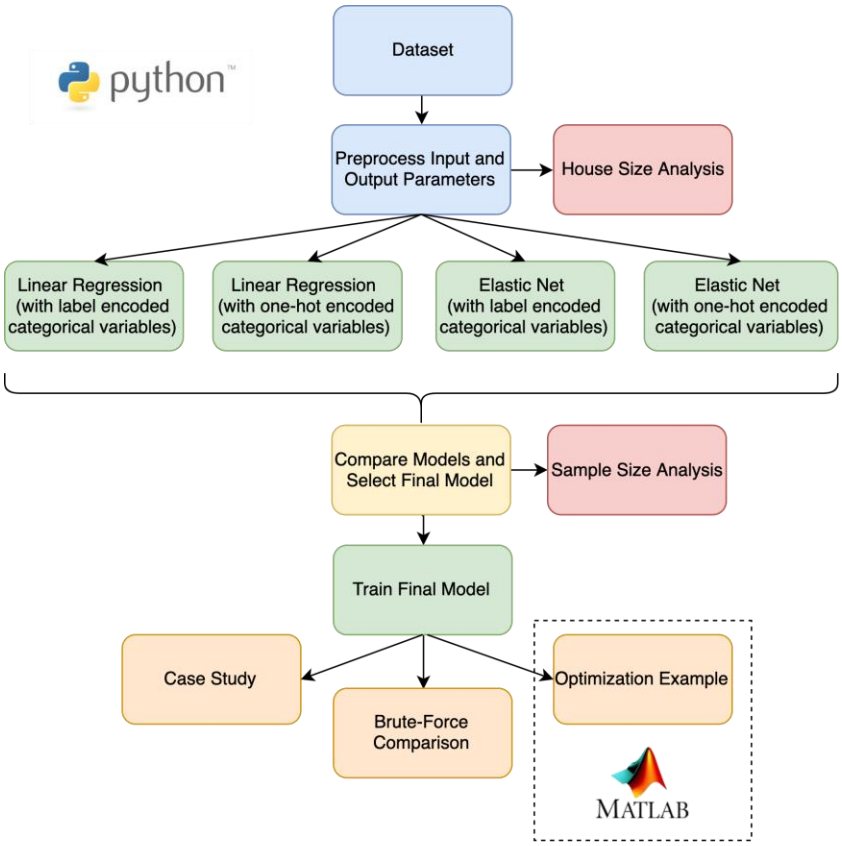

Figure 17 Outline of Phase 2 methodology.

4.1 PHASE I – DATA COLLECTION AND DATASET DEVELOPMENT

4.1.1 Baseline Archetype Model Development

Zirnhelt [12] validated a model development procedure that guided the data that Jermyn [10] collected to create the century archetype model. Characteristic data represents features that are

set in the baseline model and will not be changed. Parameters are features that are modified in the surrogate model and whose values are influenced directly by the ranges decided upon via the data collection. Table 2 shows the collected data and what was considered a parameter versus a characteristic, and what was changed from Jermyn's baseline model [10] and why. The parameters are being updated for each sample; therefore changing the baseline model is not applicable. The rest of this section explains the changes in more detail.

Table 2 Baseline archetype parameters and characteristics, and which were updated in the new model.

Geometry	**Parameter**	**Characteristic**	**Changed?**	**Reason for Change**
Building footprint		x	No	
Storey height		x	No	
Dimensions	x		N/A	
Shading devices and overhangs		x	Yes	Simplification
Glazing and doors	x		N/A	
Floor plan		x	No	
Envelope				
Materials and material properties		x	No	
Window constructions	x		Yes	Simplification
Door constructions		x	No	
Air tightness	x		N/A	
Basement				
Materials and material properties		x	No	
Wall and floor thicknesses		x	No	
Depth of wall below grade	x		N/A	
Internal Gains				
Types of major appliances		x	No	
Occupancy schedules		x	No	
HVAC				
Thermostat location		x	No	
Type of heating and cooling		x	Yes	Auto sized
Total ventilation flow rate		x	Yes	Calculated

The changes were made to prepare the baseline energy model to develop a dataset for the surrogate model. The size of the heating and cooling systems had to be adjusted based on house size, system efficiencies, and envelope thermal effectiveness. The International Energy Conservation Code (IECC) prototype models [57] were used as a reference to update the baseline model. Based on the IECC models, the HVAC systems were auto sized. The windows that Jermyn [10] used were updated to a "simple window" input in EnergyPlus to allow the surrogate model to change the whole window u-value solar heat gain coefficient easily. Simple windows are used in the IECC prototype models [57]. The ventilation rate was updated to the *Outdoor Air Flow per Zone [m3/s]* object in EnergyPlus [58]. Jermyn's model had a constant ventilation rate that cycled on/off with the furnace, resulting in unmet ventilation requirements. The required ventilation for the updated baseline was calculated using Eq. 9 from ASHRAE 62.2 [59].

$$Q_{tot} = (0.15 * A_{floor} + 3.5(N_{br} + 1))/1000 \qquad (9)$$

Where Q_{tot} is the total required ventilation rate in m3/s, A_{floor} is the floor area in m2, and N_{br} is the number of bedrooms.

The infiltration for each sample was calculated using volume and ACH_{50}. The amount of infiltration (up to 2/3 of Q_{tot}) can be subtracted to reduce the total required ventilation rate. Houses with lower air tightness will require larger amounts of mechanical ventilation, therefore the ventilation rate must be calculated for each sample.

Jermyn imported basement heat transfer calculation results from the auxiliary EnergyPlus basement program [10]. This was a very time-consuming process and since then a new Kiva framework has been developed. The Kiva calculation tool replaces the three-dimensional heat

transfer calculations from the basement program to an approximated two-dimensional value that maintains a mean absolute deviation within 3% of the true value [60].

The overhangs were removed from Jermyn's baseline model [10] as it was determined that removing them had a negligible effect on annual energy use. They were very small to begin with based on the century home geometry, and no data was collected about them during the field study to create ranges for them as parameters in the surrogate model.

4.1.2 Data Collection

A survey of The Pocket Community determined that there were many homes with similar characteristics. This was determined by reviewing the houses on Google Earth [61] and from canvassing the neighbourhood. The Pocket Community Association assisted in establishing contact with many residents of The Pocket. When this resource was exhausted, a door-to-door strategy was used to find residents who were willing to participate. Some homes that seemed from the outside to fit the archetype ended up having different characteristics and were not able to be used in the final study. The field study was completed with another student from Ryerson University, Cameron Lawrence, who used the same data for their research on a similar topic. The data that was collected is shown in Table 3.

Table 3 Data collected during The Pocket field study.

Geometry	Windows	Envelopes	HVAC
Overall length and width	Number of panes	Wall insul and construction	Type of heating systems
Floor heights	Frame material	Roof insul and construction	Efficiencies and capacities
Window areas	Window type	Slab insul and construction	

The field study was a non-intrusive investigation, therefore some values such as insulation levels were difficult to determine. Occasionally there was access to the attic or exposed insulation in the basement that allowed some data to be collected. A survey was filled out by the occupants of

each house regarding previous renovation descriptions, appliance use, set point temperatures, and goals for energy reduction renovations, if any.

Jermyn's baseline model included occupancy inputs that were based on averages from surveys filled out by homeowners during the field study. The surrogate model was not developed with any occupancy related input parameters, as it was not within the scope of this research.

4.1.3 Parameter and Range Selection

The selected parameters were broken into the following categories: geometry, enclosure performance, HVAC efficiencies, lighting, and infiltration. Jermyn's detailed work on common retrofit scenarios, feasible values, and associated costs, was used as a reference to choose the retrofit parameters [10]. Jermyn collected data for three "levels" of retrofit upgrades for eight parameters [10]. The parameters and their levels are shown below in Table 4.

Table 4 Retrofit parameters and level upgrades developed by Jermyn [10].

Strategy	Baseline	Level 1	Level 2	Level 3
Walls (RSI)	1.01	4	6	10
Roof (RSI)	2.64	9	10.5	13
Basement Walls (RSI)	0.55	2	3	3.5
Slab (RSI)	0.058	0.75	1	1.75
Windows (U-factor)	2.7	1.9	1.2	1
Air Sealing (ACH at 50 Pa)	10.54	20% Reduction	3	1
Heating and Cooling	80% Eff.	90% Eff.	94% Eff.	97% Eff.
HRV Option	N/A	60% Eff. HRV	85% Eff. HRV	80% Eff. ERV

The data collected from The Pocket included these eight parameters and 14 others that were expected to affect energy usage. Six describe geometry, four describe window-to-wall ratios for each elevation, and the other parameters are orientation, air conditioner efficiency, average lighting, and window solar heat gain coefficient. The parameters were chosen because they

described a variation within the archetype (ex. floor height, window-to-wall ratio) or they described a value that could be changed in a retrofit renovation (ex. wall insulation, furnace efficiency). Ventilation was the only calculated parameter, the rest were sampled from their defined ranges. Table 5 lists all the input parameters.

Table 5 List of input parameters for surrogate model.

Geometry	Windows	Insulation	Other
Orientation	Window U-Value	Wall Insulation RSI	Avg Lighting Density
Depth (side of house)	SHGC	Roof Insulation RSI	ACH @ 50 Pa
Width (front of house)	Front WWR	Slab Insulation RSI	Furnace Efficiency
Basement height above grade	Back WWR	Basement Insulation RSI	AC Efficiency
Basement height below grade	Left WWR		HRV Option
Average floor height (1&2)	Right WWR		Ventilation
Third floor height			

Ventilation was calculated from several of the input parameters during the IDF updating process. The Python script that was updating each IDF calculated the total ventilation rate (Eq. 9) and infiltration credit for each set of input data and set it equal to the ventilation rate in the EnergyPlus IDF.

4.1.4 Sampling Plan

With the design space set, a dataset size of 1500 was determined to be acceptable to accurately model the annual energy use based on previous research [7]. Latin hypercube sampling (LHS) was used to create a 1500 x 23 matrix of values ranging from 0 to 1. This was done using MATLAB's *lhsdesign* [62]. Eq. 10 was applied to convert the LHS generated sample to a value within the specific range.

$$PV_{i,j} = LHS_{i,j} \times \left(PV_{i,max} - PV_{i,min}\right) + PV_{i,min} \qquad (10)$$

Where $PV_{i,j}$ is the parameter value, $LHS_{i,j}$ is the random Latin hypercube sample value, i is the parameter number (1-23), and j is the sample number (1-1500).

This results in a dataset that describes 1500 "houses" with different combinations of parameters within the set design space. The set of 23 parameters that describes each individual "house" will be referred to as a sample.

4.1.5 IDF Modification and EnergyPlus Simulations

A Python [63] script was written to pull the values from each sample and update all the parameters in an IDF and resave it. The 1500 IDF files were run through EnergyPlus [58] and the energy use output was extracted with a Python script and appended to the dataset. It took approximately 36 hours to run all the files. Jermyn's baseline models were created and calibrated in EnergyPlus [10]. Another benefit of using this software was to avoid translating the models into different software.

4.1.6 Dataset Analysis

Since the retrofit parameters affected both heating and cooling, total energy use was selected as the target variable. The EnergyPlus simulation outputs were analyzed. Energy use in GJ, and energy use intensity in kWh/m2 were plotted as histograms to visualize the distribution. Box plots of each were used to analyze the mean, quartiles, and ranges of the energy use and energy use intensity. The end-uses for each home (heating, cooling, DHW, lighting, fans, other) were averaged and compared to the total energy use. Energy use by natural gas versus electricity was compared.

Energy use was plotted against each numerical input in individual scatterplots. This allows for a visual analysis of the relationship between each input variable and the output variables. Positive slopes indicate that as the input variable increases, so does the energy use. Negative slopes

indicate that as the input variable decreases, the energy use decreases. Ventilation was separated as it was not sampled using LHS but calculated from input data.

For the categorical inputs, the energy use of all the samples containing each category were examined individually. Error bar plots were used to show the mean, standard deviation, and maximum and minimum value for the energy use associated with each category.

The Pearson's correlation coefficient was calculated for the entire dataset, comparing each input parameter against each other, and the energy use. This was a univariate comparison, so each input parameter was only compared against the output. The limitation of this method is that it does not capture the relationship of multiple input parameters to each other. The variance inflation factor (VIF) was calculated to examine collinearity between input parameters. This value is similar to the Pearson's correlation except it considers multicollinearity by evaluating the relationship between all inputs and the output instead of just the output. Since all the input parameters except ventilation were randomly generated using LHS, none of those parameters should have a large VIF. However, since ventilation was calculated using combinations of input parameters, the VIF must be calculated to ensure that any collinearity will not affect the interpretability of the model.

4.2 PHASE II – SURROGATE MODEL DEVELOPMENT

4.2.1 Summary of Surrogate Model Development

Table 6 summarizes the decisions made in the surrogate model development process, Phase 2.

Table 6 Summary of surrogate model development decisions, organized as proposed by Barnes [40].

Dataset Development	**Model Intent**	Bottom-up archetype model to predict annual energy use for an archetypal home
	Target Variable	Annual total energy use
	Building Archetype	3-storey detached single-family century home in Toronto, ON
	Location + Climate	2016 Toronto City Centre, Ontario, Canada - CWEC weather file [64]
	Energy Simulation Software	EnergyPlus v8.9 [58]
	Statistical Analysis and Modelling Tool	Python v3.7.1 [63]
	Base Model	Century home model EnergyPlus v8.0 by Jermyn [10]
	Parameters + Ranges	21 continuous variables
		2 categorical variables
		Parameters from field study representative of geometry, envelope constructions, air tightness, internal gains, and HVAC systems
Data Processing	**Sampling Plan**	1500x23 Latin hypercube sampling matrix
	Train/Validation/Test Split	10 times repeated nested 10-fold cross-validation
	Feature Engineering	Input parameter standardization and log transformation of input and output variables
		Elastic net for feature selection
Trained Model Development	**Learning Algorithms + Hyperparameter Selection**	Multivariate regression
		Elastic net multivariate regression
	Error Metrics	Coefficient of determination (R^2)
		Root mean squared error (RMSE)
		Mean absolute percent error (MAPE)

This Table was suggested by Barnes [40] as a way to allow researchers to easily compare and understand the surrogate models that are developed. Many of the reviewed papers did not explain how their surrogate models were created. Some examples include indicating if (or what type of) cross-validation was used, what sampling plan, how the hyperparameters were selected and what

the values were, and what thresholds used to select features were. Some papers only reported RMSE metrics, which can only be compared to output variables of the same unit. This makes it very difficult to compare methodologies and models. This field would benefit from continuity in reporting surrogate model development steps and findings.

4.2.2 Preprocessing Input Variables

The numerical inputs were standardized by calculating the mean and standard deviation of the *training* samples for each input parameter only. The data (training and validation) was standardized by subtracting the mean and dividing by the standard deviation. This was done using the *StandardScalar* package in Python [65]. SKLearn's Pipelines [65] was used to perform the standardizing and model training simultaneously to ensure the K-fold cross-validation technique is not causing data leakage.

Models were developed using both *label* and *one-hot encoded* categorical inputs to determine which is more accurate. The categories were *one-hot encoded* creating four additional parameters (one for orientation and three for HRV options). Table 7 shows how this works for the HRV option.

Table 7 one-hot encoded *HRV option categorical parameter.*

Category	No HRV	HRV 60%	HRV 90%	ERV 85%
No HRV	1	0	0	0
HRV 60%	0	1	0	0
HRV 90%	0	0	1	0
ERV 85%	0	0	0	1

Note that for linear regression one column must be dropped, as the values are inherently coded into *n-1* variables. When performing lasso, ridge, or elastic net, all the categories should be left in. It was unclear if *one-hot encoding* or *label encoding* would produce better results. Both

methods were used and the method that produced the better results was be selected for the final model.

After the initial linear regression was completed, it was clear that a linear relationship was not appropriate. Several of the most common transformations were performed on the data to determine if any improved the model. It was determined that a log transformation of both the input and output parameters was most successful. Therefore, all of the models going forward would use log-transformed data (performed before standardization). Since the energy use is in GJ, the predicted values of the model must be back-transformed to output values in this unit. The model must be evaluated using the back-transformed predicted outputs to ensure that it can accurately predict in the intended output unit.

4.2.3 Model Comparison and Selection

The objective of this research was to create a simple and interpretable model. Multivariate linear regression was used as a starting point to see if an accurate model could be created. The surrogate model was developed using multivariate linear regression and elastic net regression. Elastic net regression was used to reduce the number of parameters required in the model. The linear regression model was compared to see what accuracy must be sacrificed to reduce the number of parameters. For both linear regression and elastic net regression, *label encoded* and *one-hot encoded* categorical variables were used. This resulted in four models that were completed and compared. The models were evaluated on the accuracy, but also on simplicity (number of coefficients). Less parameters are desirable because the model will generalize better to new data. For future work this would also mean less ranges would have to be researched and created, and less optimization would be required.

4.2.4 Evaluation and Validation

The model was evaluated using three metrics that compared the actual output values to the predicted ones. The coefficient of determination, R^2, was used to evaluate the amount of variance that is described by the model. It is independent of unit and can be compared across all models. The mean absolute percent error (MAPE) and root mean squared error (RMSE) metrics were calculated as well. The MAPE is used as it is a more practical and interpretable metric in terms of understanding how much error there is. The RMSE can only be compared to models with the same mean and unit.

K-fold cross-validation splits the data into k sets. Ten folds has been shown to produce test error results that have a good trade-off between bias and variance [22]. Nested 10-fold cross-validation repeated 10 times was used to validate the model. The process for nested 10-fold cross validation is as follows:

1. The whole dataset is split into 10 subsets, or "folds". One fold will later be used as the testing set. The remining nine folds become the new dataset for the inner loop.
2. The inner loop dataset is split again using 10-fold cross-validation. One of these loops is the validation set, and the remaining nine are the training set.
3. A model is trained on the nine folds of the inner loop (training set) and evaluated on one fold of the inner loop (validation set). The R^2 and RMSE for the validation set is reported. This concludes one inner loop.

For elastic net only:

4. Step 3 is repeated for each set of hyperparameters. In this research, 200 tuning parameters and five L1 ratios are investigated, resulting in 1000 different sets of hyperparameters (or

models). The set of hyperparameters that falls within one standard error of the best (lowest) RMSE score is reported.

5. Step 3 is repeated 10 times so that each fold is used as the validation set once. The hyperparameters are averaged and reported. This concludes one outer loop.
6. Steps 1-4 are repeated 10 times so that each set of nine folds of the outer loop is used as the new dataset for the inner loop. The hyperparameters are averaged and reported (which are already the average of 10 inner loops).
7. The result is one set of hyperparameters that are an average of 100 values (10 inner and 10 outer loops). The mean hyperparameters are selected as the optimum hyperparameters for this model and are used to train the elastic net model for the rest of the steps.

For linear regression and elastic net:

8. Step 3 is repeated 10 times so that each fold is used as the validation set once. The 10 R^2 and RMSE are averaged and reported. This concludes one outer loop.
9. Steps 1-3 are repeated 10 times so that each set of nine folds of the outer loop is used as the new dataset for the inner loop. The 10 R^2 and RMSE are averaged and reported (which are already the average of 10 inner loops).
10. The result is one set of R^2 and RMSE values that are an average of 100 values (10 inner and 10 outer loops). The mean R^2 and RMSE value estimates the true performance of the model.
11. Steps 1-10 are repeated for each of the models that are to be tested. A final model is selected based on the mean performance metrics from Step 10.

For the final model:

12. Back to Step 1. The nine folds of the outer loop become the training set. The model is trained on the training set and evaluated on the testing set (which has never been seen before by each inner loop). The R^2 and RMSE values, along with the coefficients are reported.
13. Step 12 is repeated 10 times for each of the outer loops. The mean of the 10 R^2 and RMSE values are the final performance metrics for the selected model on unseen data. The ten sets of coefficients are averaged and the mean and standard deviation are reported. For elastic net, the one set of hyperparameters on each split of data might result in a different number of coefficients. Therefore the reported average number could be a decimal.

This process ensures there is no data leakage by selecting hyperparameters and testing on the same data. Figure 18 shows this process for 3-fold cross-validation. For 10-fold cross-validation the steps would be the same except the outer loop and inner loop would be split into 10 folds.

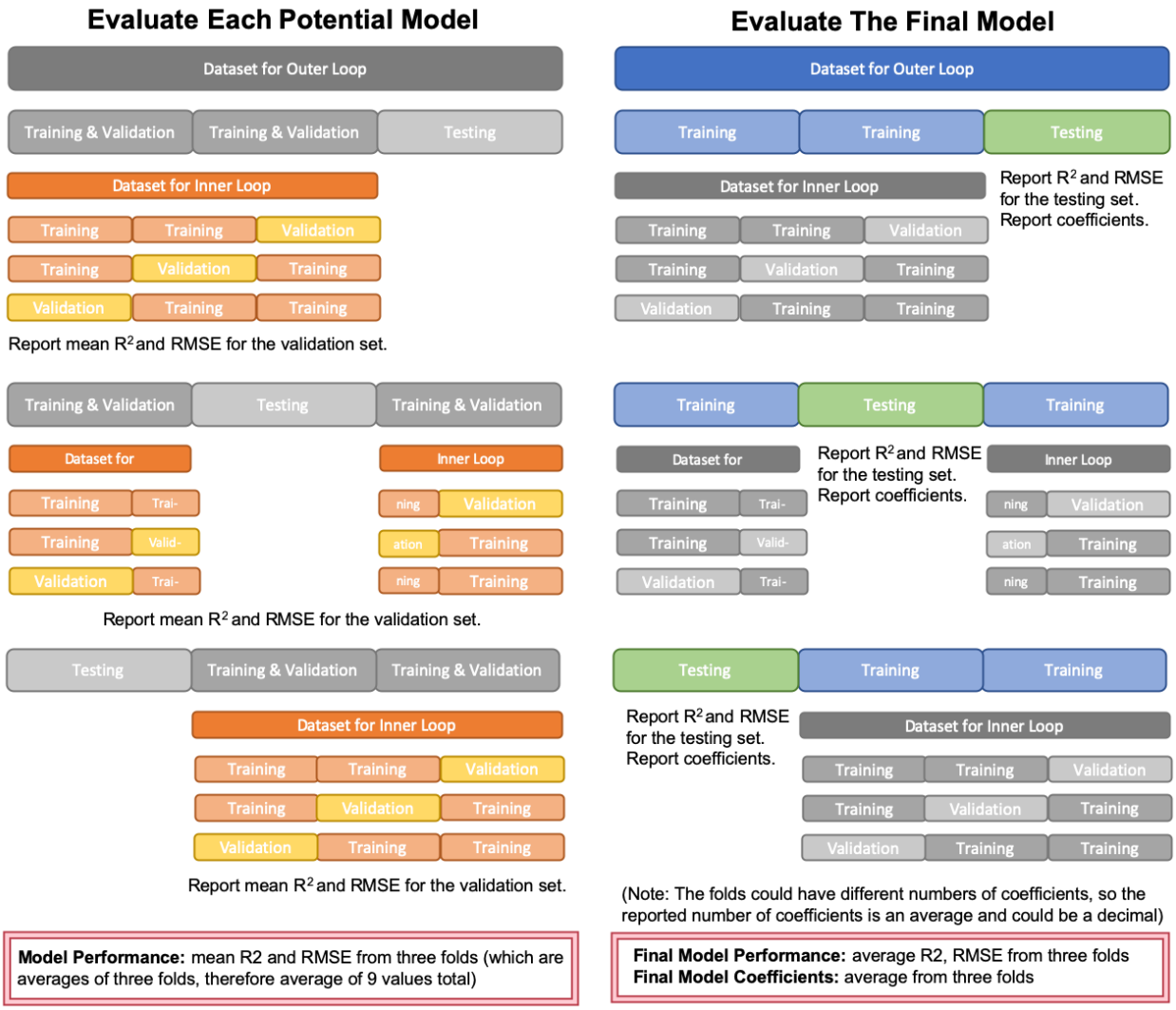

Figure 18 Cross-validation diagram showing 3-fold nested cross-validation.

4.2.5 Sample Size Analysis

A sample size analysis was performed to ensure the number of samples used to train and validate the model was large enough. The sample size analysis shows at what point the evaluation metrics for the training and validation set converges, if at all. This indicates how much variance and bias is present in the model.

4.2.6 Archetype Size Analysis

One of the research questions was to determine if the wide range of size of century homes would require separate archetypes, or if they could be incorporated into one surrogate model. The

flexibility of the surrogate model method allows for the entire design space (large ranges of many parameters) to be modelled. The correlation and error of the model will indicate if it can accurately predict energy use within that design space.

This was tested in two ways to determine how to develop the final surrogate model. The first method involved separating all 1500 samples in three sets of 500 samples based on floor area. Three separate surrogate models for small, medium and large homes were created using the same methodology. A fourth model was created using a random split of 500 samples including all house sizes (so that the sample size across all four models are equal). Performance metrics, sample sizes, residuals, and coefficient values can be compared.

The second method did not separate the data, it trained one model on the whole set. The validation set that this model was tested on was separated into thirds based on floor area and evaluated separately. The performance metrics and residuals can be compared. It was then determined whether three models trained on a specific house size will have similar results to one model trained on all sizes, and will a model trained on all sizes predict similarly on small, medium, and large homes. Since the datasets are smaller, 5-fold cross-validation was performed. This means 20% of the data was used to validate the model. This was repeated 10 times. Only one cross-validation loop with a training and validation set was used to evaluate the model. Since this was for comparative purposes and no model was being selected (or hyperparameters being tuned), no testing set or inner loops were needed. This applies to both methods.

For method 1, 400 samples were used to train and 100 were used to validate. This methodology produced four separate models. For method 2, 1200 samples were used to train and 300 were used to validate. The validation set was split into three subsets based on size, meaning 100 samples were used per size as validation. A fourth subset was created as a random sample of 100

samples from the 300 sample validation set in order to represent all the sizes, and to ensure that the number of samples was equal for each validation subset. Table 8 summarizes each method.

Table 8 The two methods for the house size analysis and the sample sizes.

	Model	Dataset Size	Training (80%)	Validation (20%)		
Method 1	Small	500	400	100		
	Medium	500	400	100		
	Large	500	400	100		
	Combined	500	400	100		
Method 2	All sizes	1500	1200	300	Small	100
					Medium	100
					Large	100
					All sizes	100

Method 1 results in four models that have been trained on data for different sizes of homes. Method 2 results in one model that has been trained on all sizes and evaluated on different sizes of homes. Both methods result in metrics for prediction accuracy on small, medium, and large homes, and a fourth set of metrics representing all sizes.

4.2.7 Case Study

Natural Resources Canada (NRCan) provides an energy performance rating and labelling program that starts with an EnerGuide home energy audit [66]. Two homes in The Pocket had this energy audit performed and were provided with an energy efficiency report which included information that was not always possible to obtain during the field study, such as air tightness, wall insulation values, window U-values, and HVAC efficiencies. Some homeowners provided two years of gas and hydro utility bills. A case study was conducted using this information. The inputs to the surrogate model were taken from the energy audit report and the measured data

collected from these houses. The surrogate model predicted annual energy use. This was compared to the average annual energy use over two years from the utility bills.

4.2.8 NSGA-II Optimization

This research did not include a full optimization scope, however an example of how the surrogate model could be used for optimization was included to demonstrate its capabilities. This was accomplished with the *gamultiobj* function in MATLAB [62]. For this practical application example, the input parameters for the two homes used in the case study was used to create a baseline surrogate model.

Jermyn's thesis included a set of implementation levels for each retrofit parameter and the associated costs, shown in Table 9. The levels were determined by a field study and previous research. The costs were determined from RSMeans and local contractors. The costing shown below has accounted for inflation from 2013. [10]

Table 9 Jermyn's retrofit levels and associated costs [10].

Strategy	Baseline	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Walls (RSI)	1.01	4	6	10	$ 23,284	$ 23,843	$ 25,655
Roof (RSI)	2.64	9	10.5	13	$ 10,911	$ 12,590	$ 18,295
Basement Walls (RSI)	0.55	2	3	3.5	$ 4,794	$ 5,054	$ 5,252
Slab (RSI)	0.058	0.75	1	1.75	$ 5,594	$ 5,922	$ 6,147
Windows (U-factor)	2.7	1.9	1.2	1	$ 20,445	$ 22,751	$ 26,907
ACH at 50 Pa	10.54	20% Reduction	3	1	$ 1,474	$ 1,331	$ 1,305
Heating and Cooling	80% Eff.	90% Eff.	94% Eff.	97% Eff.	$ 3,479	$ 4,048	$ 4,786
HRV Option	N/A	60% Eff. HRV	85% Eff. HRV	80% Eff. ERV	$ 2,347	$ 3,596	$ 4,115

The eight parameters that Jermyn determined retrofit levels and costing for were used to perform an optimization to see which retrofit parameters are the most cost-effective at reducing energy use [10].

In order to calculate the energy use using the coefficients, several preprocessing steps were required. The input variables were log-transformed and standardized with the means and standard deviations of the input parameter data used to train the final model. Then the input parameters were multiplied by their corresponding coefficients and summed together with the intercept. The final value was back-transformed to predict energy use in GJ. The cost for each selected upgrade was summed together to determine the total cost for all the retrofit levels chosen for that specific set of solutions.

The input parameters for the case study homes were used. The eight retrofit parameters described in Table 9 are the only values that were optimized. The other 11 parameters were held constant. The baseline energy usage determined for each house was compared to the energy use after energy-saving retrofits had been applied. A set of cost-effective retrofit solutions were developed very quickly.

To ensure the optimization does not achieve unrealistic results, the ACH_{50} value is only upgraded if wall insulation is upgraded. Jermyn did this for the brute-force optimization [10]. It is assumed that to achieve the ACH_{50} levels outlined by Jermyn, the wall insulation must be upgraded as well [10]. This is a limitation of this method.

The NSGA-II algorithm was given an energy function and a cost function to minimize. The energy function was the final surrogate model – coefficient values multiplied by log-transformed and standardized inputs and a y-intercept – back-transformed to predict energy use in GJ. The cost function was a sum of each of the retrofit costs depending on the chosen level. The set of Pareto solutions returned aimed to minimize both functions.

5 RESULTS & DISCUSSION

5.1 PHASE I – DATA COLLECTION AND DATASET DEVELOPMENT

5.1.1 Data Collection

The field study data collected for this research in conjunction with the field study data collected by Jermyn [10] included 35 homes total for 4-5 different archetypes. The field study included data for geometry, enclosures, and HVAC. This data, Jermyn's data [10], and other sources (described in the next section) were used to determine a range of typical values for each parameter. These ranges define the design space for the model.

5.1.2 Ranges

The maximum and minimum values for each parameter are shown in Table 10. The grey rows are the parameters that Jermyn selected retrofit upgrade levels and costing for [10]. The model is only valid for input parameter values within these ranges. The ranges are designed to incorporate existing baseline conditions as the highest energy use and a heavily retrofitted house as the lowest energy use. The geometry ranges incorporate the smallest to largest values.

Table 10 Input parameters and associated ranges.

Input	Unit	Min	Max	Source
Depth (side of house)	m	9.18	18.59	Field study & [10]
Width (front of house)	m	4.275	7.37	Field study & [10]
Basement height above grade	m	0.54	1.65	Field study & [10]
Basement height below grade	m	0.684	1.87	Field study & [10]
Average floor height (1st/2nd)	m	2.124	3.025	Field study & [10]
Third floor height	m	2.07	2.805	Field study & [10]
Front Window to Wall Ratio	-	0.08	0.35	Field study & [10]
Back Window to Wall Ratio	-	0.08	0.35	Field study & [10]
Left Window to Wall Ratio	-	0.01	0.12	Field study & [10]
Right Window to Wall Ratio	-	0.01	0.12	Field study & [10]
Avg lighting	W/m2	0.46	7.66	[67]
AC Efficiency	COP	2.9	5.0	[68]
Wall Insulation RSI	m2K/W	0	10	Field study & [69]
Roof Insulation RSI	m2K/W	0	14	Field study & [69]
Slab Insulation RSI	m2K/W	0	6	Field study & [69]
Basement Wall Insulation RSI	m2K/W	0	10	Field study & [69]
Air Changes per Hour @ 50 Pa	1/h	1	23	[10] & [70]
Furnace Efficiency	%	0.78	0.98	[71]
Window U-Value	W/m2k	0.71	2.95	[10] & [69]
SHGC	-	0.2	0.7	[10] & [69]
HRV Option	-	Categorical		[10]
Orientation (based on 17˚ tilt)	-	Categorical		Field study

The HRV option and orientation have categorical inputs instead of continuous numeric ranges. Table 11 shows the values that were used for each category.

Table 11 All categories in the categorical input parameters.

	Category			
	1	**2**	**3**	**4**
HRV Option	No HRV	60% Efficient HRV	90% Efficient HRV	85% Efficient ERV
Orientation	North/South Facing (17/197˚)	East/West Facing (107/287˚)	-	-

Most of the houses in The Pocket were constructed between 1900 and 1930 [72], and many of them have no or little insulation. The minimum value for each of the insulation RSI values was 0 (note that the RSI of the constructions was not zero as it included the structural and finish

layers). The maximum values were determined by the levels recommended to reach Passive House Institute US (PHIUS) standard certified house in the Toronto climate zone [69]. PHIUS is an organization dedicated to develop North American passive house practices and certifications for buildings [69].

The windows were simulated in EnergyPlus as a "simple window" object, taking U-value, solar heat gain coefficient (SHGC), and visible transmittance (VT) as inputs. The visible transmittance remained constant at 0.5, and the U-value and SHGC were considered separate continuous variables. The field study, PHIUS guidelines for Toronto's climate zone [69], and Jermyn's values [10] were referenced.

The minimum value for air changes per hour at 50 pascals (ACH_{50}) was determined from Jermyn's research as the lowest possible value for a retrofit of homes constructed within this period [10]. The EcoEnergy Database contains blower door testing results from 500,000 pre-retrofit homes in Ontario [70]. After filtering detached 3 storey homes built between 1900-1930, the results from ~7400 homes were left. The 99^{th} percentile of the values was 22.77. This value was chosen as it was validated by two blower door test results that had been completed on homes in The Pocket.

Most of the furnaces observed in the surveyed homes were high-efficiency furnaces with efficiencies of 0.96 or higher. The range was decided based on the field study and the minimum and maximum efficiency values as outlined by NRCan [71].

The ventilation option is a categorical parameter. The options are no heat recovery ventilator (HRV), 60% efficient HRV, 90% efficient HRV, and 85% efficient energy recovery ventilator (ERV). These values were taken from Jermyn's retrofit values for baseline and levels 1-3 [10]. These values were validated by looking at specifications of Canadian manufacturers of HRV and

ERVs to ensure that these values represented available products. The HRV/ERV was modelled as a part of the existing furnace ducting instead of a separate ventilation system. This was done to match the existing baseline condition of the existing homes. It should be noted that in reality, at low air tightness levels a separate ventilation system would be necessary to ensure the required ventilation levels were met.

Jermyn's baseline model described an average geometry from the measurements of all the surveyed houses [10]. The design space for the surrogate model includes variable geometry. The depth (side of house), width (front of house), basement height above grade (AG) and basement height below grade (BG), average first and second floor height, and third floor height was used to describe the size of the house. The shape will remain the same. The range of values for each of these inputs was chosen as the minimum and maximum value for each dimension collected from The Pocket field study and Jermyn's field study [10]. Both sets of data were included so that a larger variation of the archetype could be represented by one surrogate model, and to be able to analyze the differences between the large and small century home archetype. The geometry minimum and maximum values were expanded by 10% to incorporate additional houses that were not measured but were assumed to exist.

The window wall areas were collected from The Pocket's small century homes. Jermyn reported window-to-wall ratios (WWR) for each elevation of the homes measured [10]. The minimum and maximum value for each elevation (front, back, left, right) was used as the range. The front is the side facing the street, back is opposite the front, left is the left side of the house when facing the front, and right is opposite the left. For example, if the front of the house was facing North, the left side would be facing East. The total WWR for each elevation was divided into

four windows on the front and back sides (one for each floor: basement, main, second and third), and three windows on the left and right sides (basement, main, second).

The orientation is a categorical parameter. The categories are north/south facing or east/west facing. North/south facing would mean the front of the house was facing north or south. Because of the way Toronto's grid is aligned, the cardinal directions were offset 17˚. Figure 19 shows a map of the City of Toronto with The Pocket neighbourhood outlined in red, and Figure 20 shows a map of The Pocket neighbourhood [73].

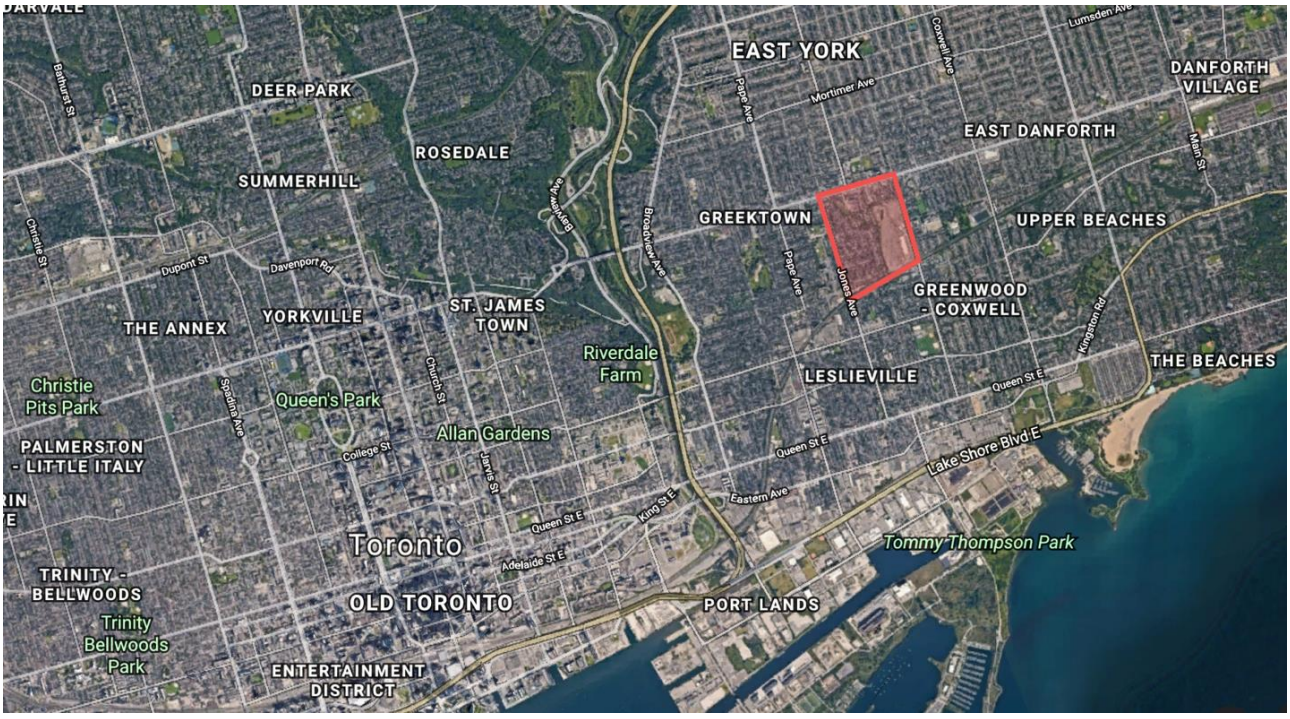

Figure 19 Map of Toronto, ON Canada. The Pocket neighbourhood is outlined in red.

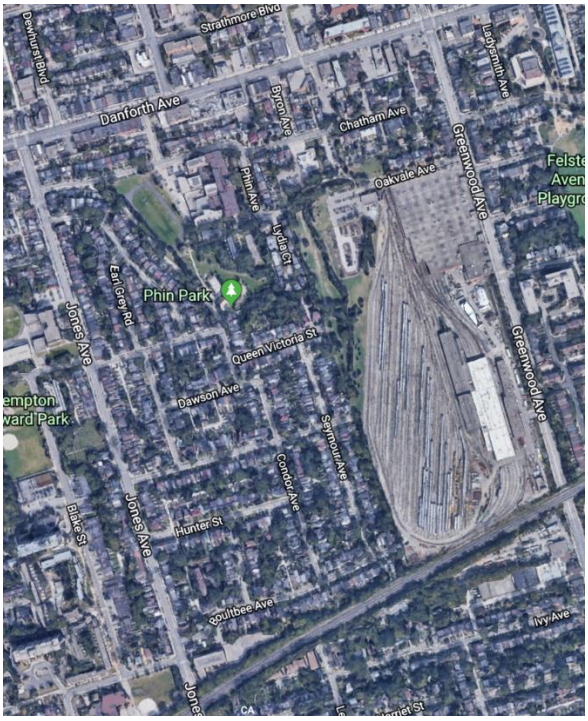

Figure 20 Map of The Pocket neighbourhood.

Air conditioner efficiencies were taken from NRCan's recommendations based on ENERGY STAR products [68].

Statistics Canada reported that there were 27 lightbulbs on average per household in Ontario [67]. Incandescent bulbs have the highest wattage, with the maximum wattage being 100W. LED bulbs have the lowest wattage at around 6W. Using the floor area of the average house size described by the ranges, the average lighting in W/m2 was determined.

The data collection process for future work could be reduced. If minimum and maximum values are already defined, such as insulation values and furnace efficiencies, these parameters do not need to be included in the field study. The most important parameters are the ones that describe the geometry. However, all the data would need to be collected if the house was to be used in a case study.

5.1.3 Baseline Model Development

The baseline model in the OpenStudio plugin [74] for Sketchup [75] is shown in Figure 21. The purple planes mimic the shading that would be experienced by neighbouring houses.

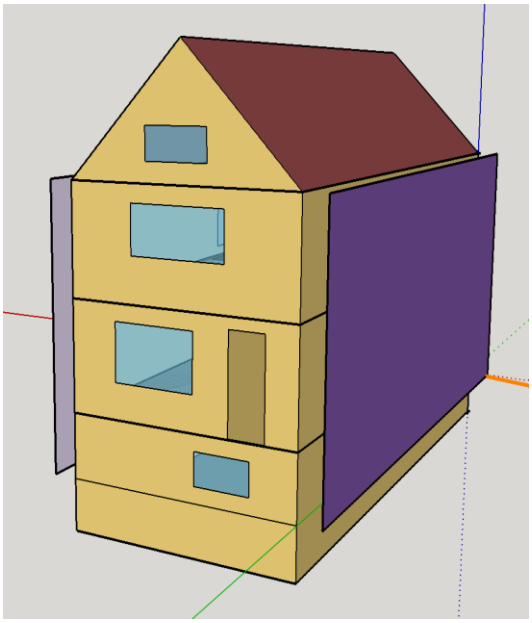

Figure 21 SketchUp model for baseline century home.

The energy use from the updated baseline model is compared to Jermyn's original model in Table 12 [10].

Table 12 Baseline energy values from EnergyPlus.

	Jermyn's Baseline	Updated Baseline	Difference
Energy Use [GJ]	213	168	24%

The difference can be attributed to the auto sizing of the HVAC system. There was a 43% decrease in the size of the heating coil from Jermyn's model [10] compared to the updated model. Figure 22 and Figure 23 show the mean weekly temperatures for each zone in Jermyn's baseline model [10] and the updated baseline model.

Figure 22 Mean weekly temperature for each zone for Jermyn's archetype model [10].

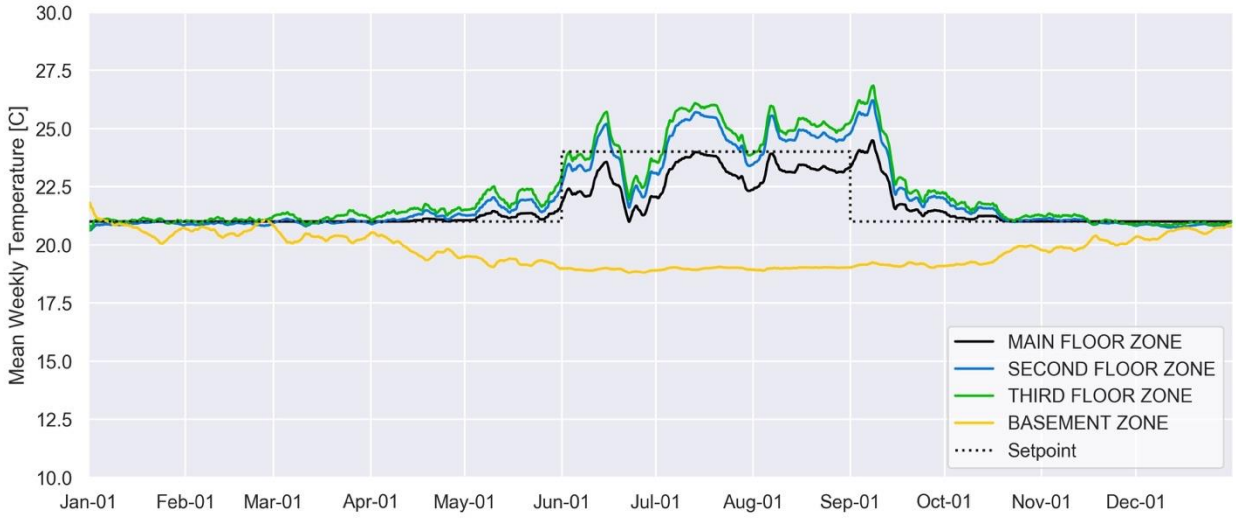

Figure 23 Mean weekly temperature for each zone in the updated baseline model.

The above figures indicate that auto sizing the HVAC system allowed the setpoint temperature to be reached for more of the year. This shows that Jermyn's baseline model [10] is delivering too much heat to the uncontrolled zones. As a result, the third floor is at 26˚C for most of the year. This is likely why the energy use is higher than in the updated baseline model. The house has only one control zone making it very difficult to meet setpoint temperatures in all zones. The system is cooling during most of the time the setpoint temperatures are not being met, which is only attributed to 2% of the total energy use.

5.1.4 Output Data Analysis

The frequency distribution of the 1500 EnergyPlus simulations are shown in Figure 24 versus energy use values in GJ.

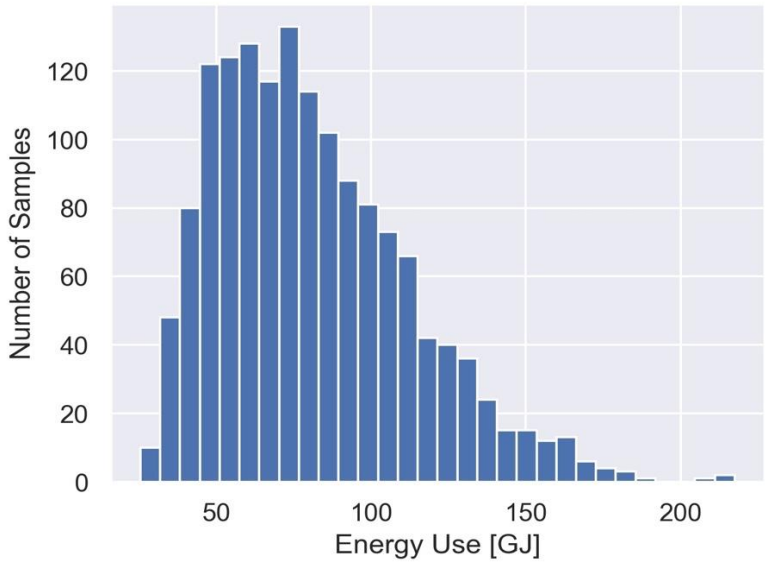

Figure 24 Distribution of energy use output from 1500 EnergyPlus simulations.

Figure 25 shows a boxplot of energy use per household. The red line indicates the mean, the box indicates the 1st and 3rd quartile, and the whiskers indicate the range of the data.

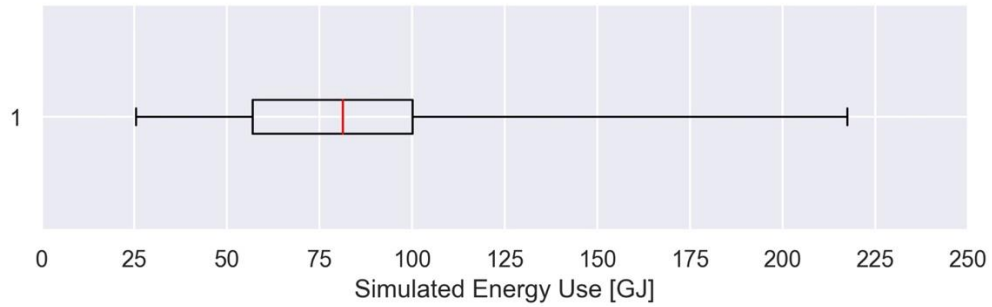

Figure 25 Simulated energy use boxplot.

Energy use intensity for each sample was calculated. Figure 26 shows the results. The red line indicates the mean, the box indicates the 1st and 3rd quartile, and the whiskers indicate the range of the data.

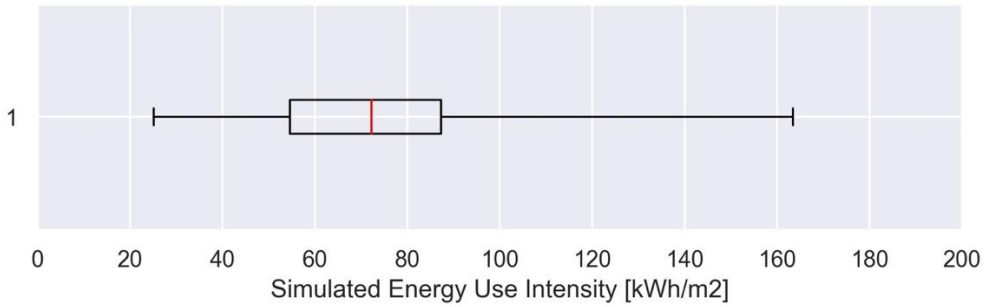

Figure 26 Simulated energy use intensity boxplot.

Distribution of energy consumption by end-use is shown in Figure 27. Natural gas contributes to 80% of the total energy use, and electricity makes up 20%.

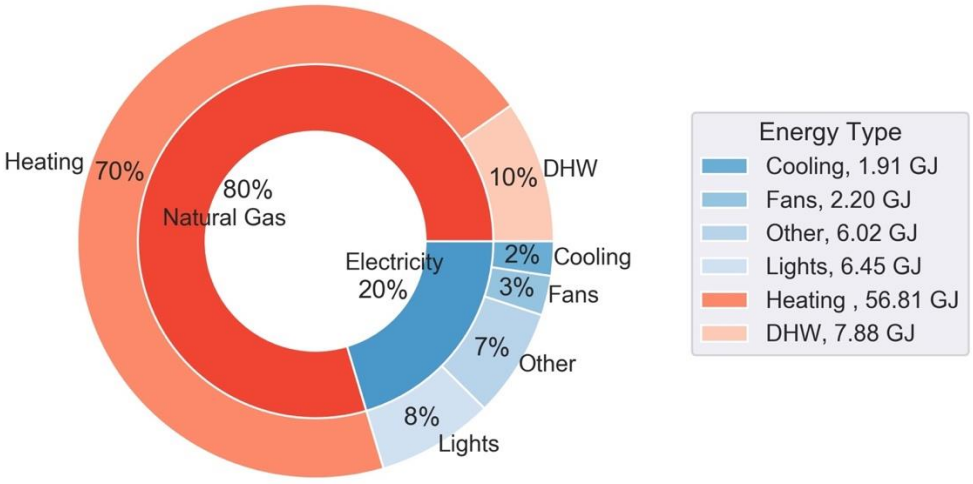

Figure 27 Average end-use distributions for the 1500 EnergyPlus simulations.

5.1.5 Input Data Analysis

Preliminary data visualization was used to analyze the numerical data inputs. Scatter plots for each input parameter were plotted against energy use in GJ. A line of best fit is shown on each scatter plot as a linear relationship between each input variable and the output variable was assumed for the multivariate linear regression. It was included to enable an easier visual comparison of the magnitude of each slope. A positive slope indicates a positive relationship with the output. This is most noticeable in the depth and the ACH_{50}. As depth or ACH_{50} increases, the energy use increases as well. The opposite is true for negative slopes. As the wall

insulation and burner efficiency increases, the energy use decreases. This gives the first indication of what input parameters will be important in the final surrogate model. The scatter plots are shown in Figure 28.

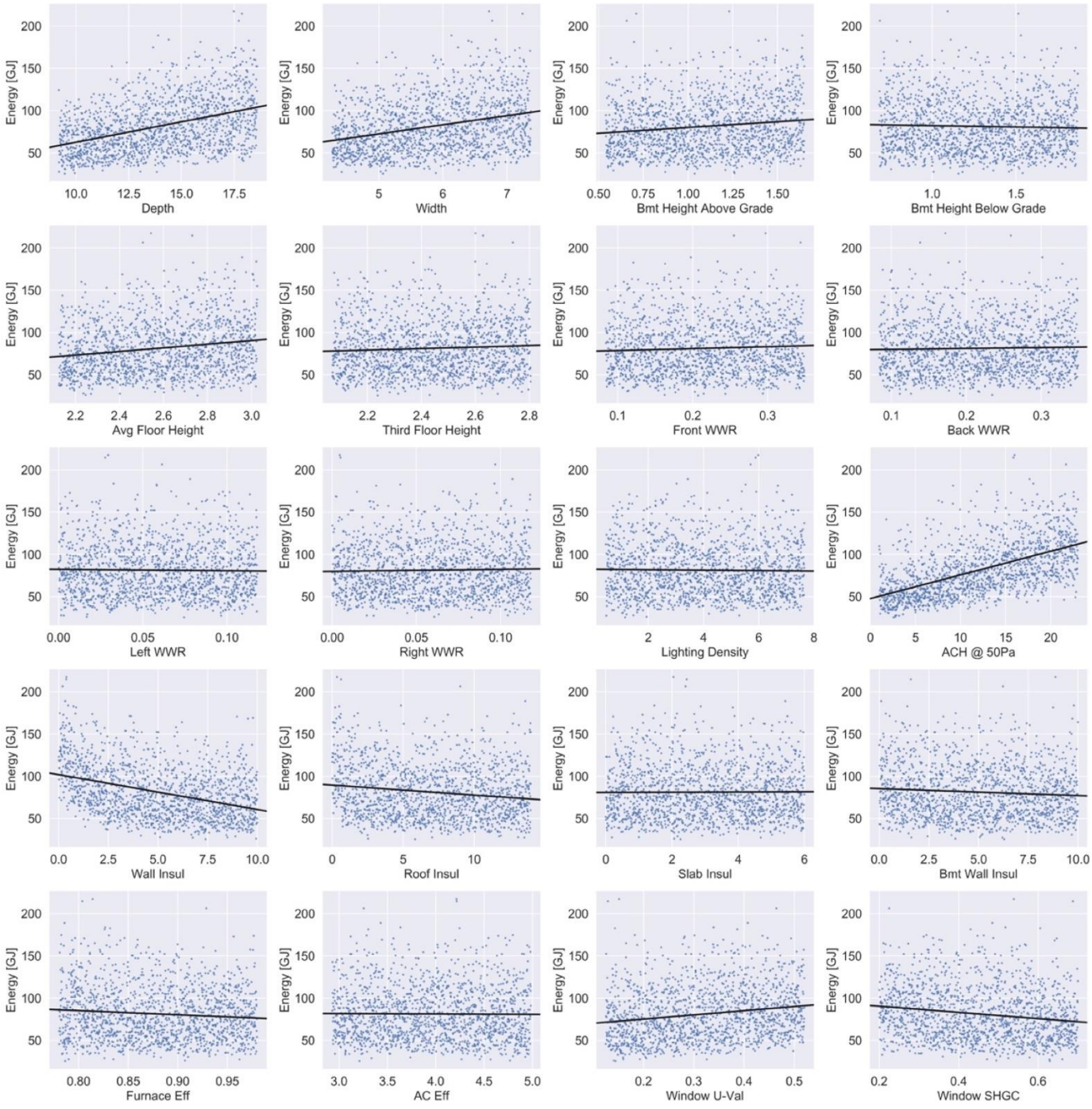

Figure 28 Scatter plots showing energy versus the values for each input parameter.

The distribution plots in Figure 29 show how the Latin hypercube sampling has space filled the range with random values and created a uniform distribution. The Figure shown is for the depth parameter, however the distributions for the other input parameters are the same.

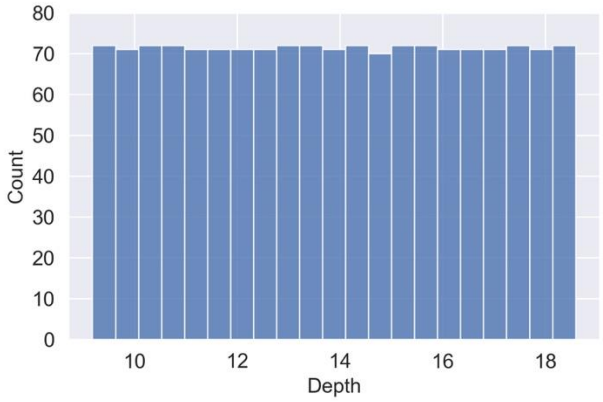

Figure 29 Uniform distribution for the depth input parameter. All other distributions are very similar.

Ventilation was examined separately as it was the only calculated input parameter (not sampled). Figure 30 shows a scatter plot on the left and distribution plot on the right.

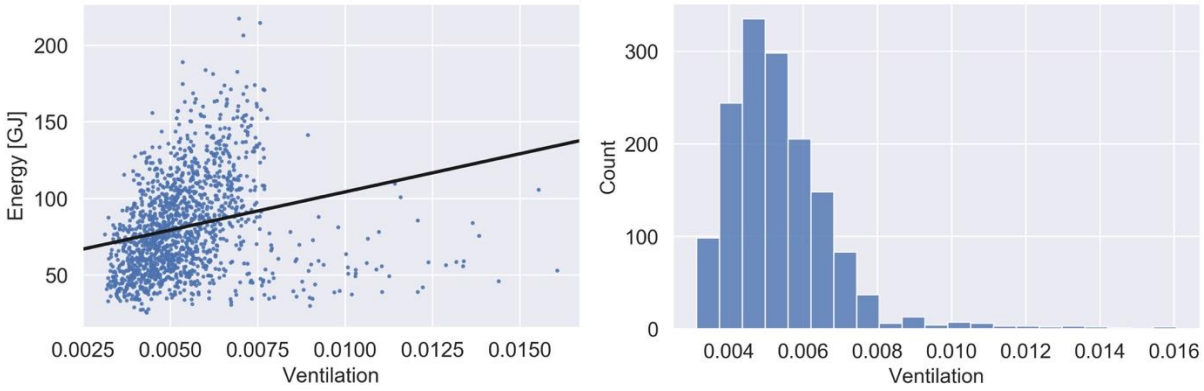

Figure 30 Scatter plot and distribution for ventilation.

It does not have a uniform distribution like the other input parameters because it was not sampled using Latin hypercube sampling.

The categorical inputs are analyzed using violin plots, as shown in Figure 31 below. The left Figure shows the two orientation categories; north-south facing or east-west facing. The right

Figure shows the four HRV categories; no HRV, 60% efficient HRV, 90% efficient HRV, and 85% efficient ERV. These plots compare each category to energy use to visually interpret the relationship between them. The width of each violin describes the distribution of the energy values for that category. The white dot is the median, the thick bar is the 1st and 3rd quartile, and the thin line is 1.5 times each quartile.

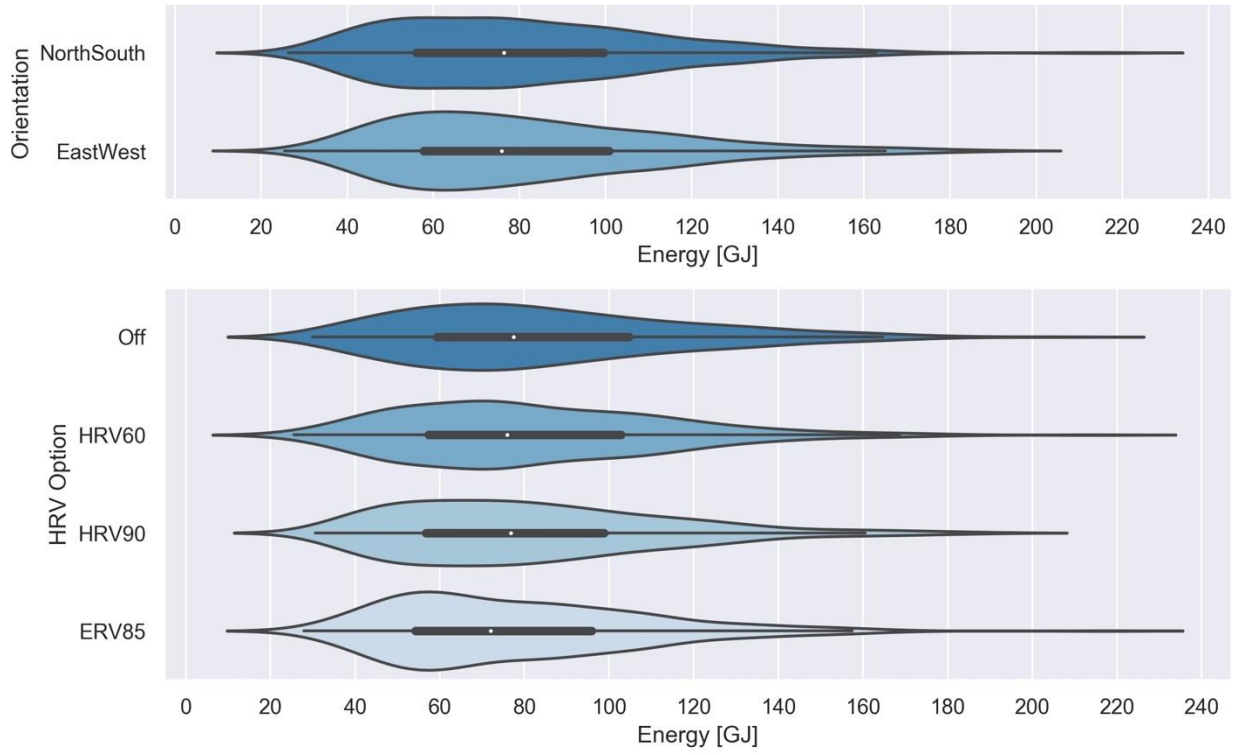

Figure 31 Violin plots for the categorical variables.

The mean, standard deviation, and range of each category is shown in Figure 32.

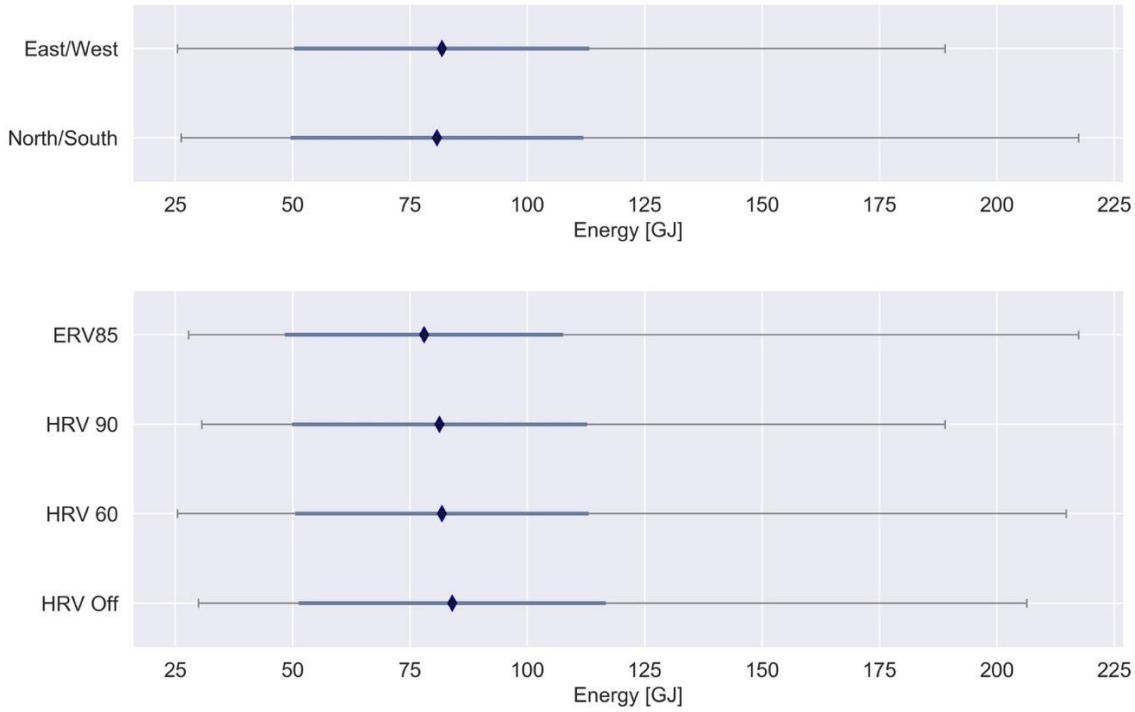

Figure 32 Mean, standard deviation, and minimum and maximum values for each categorical variable.

The two orientations have very similar means, although North/South facing is slightly lower. The 85% efficient ERV is associated with the lowest energy use, then the 90% efficient HRV, 60% efficient HRV, and lastly no HRV with the highest energy use. The HRV is attached to the furnace outdoor air intake and exhaust ducts in the baseline energy model. The furnace is cycled on and off based on the temperature of the control zone, therefore outdoor air is only being brought in while the furnace is on. This explains why there are only small changes in energy use output based on HRV category. The baseline model was left with this system because none of the century home archetypes had a separate ventilation system. As the houses become more airtight, there is not enough ventilation being delivered from the furnace ducts therefore a separate ventilation system would be required. This should be noted as a limitation of this research.

A Pearson's Correlation Matrix, shown in Figure 33, was calculated to analyze the statistical significance between each input variable and the output variable. Most variables were not

correlated to each other because the data was randomly generated with the LHS. The only variable that showed high correlation values is ventilation, because it was calculated using volume and ACH_{50}. The Pearson's correlation between each input and the output is included and shown in the bottom row.

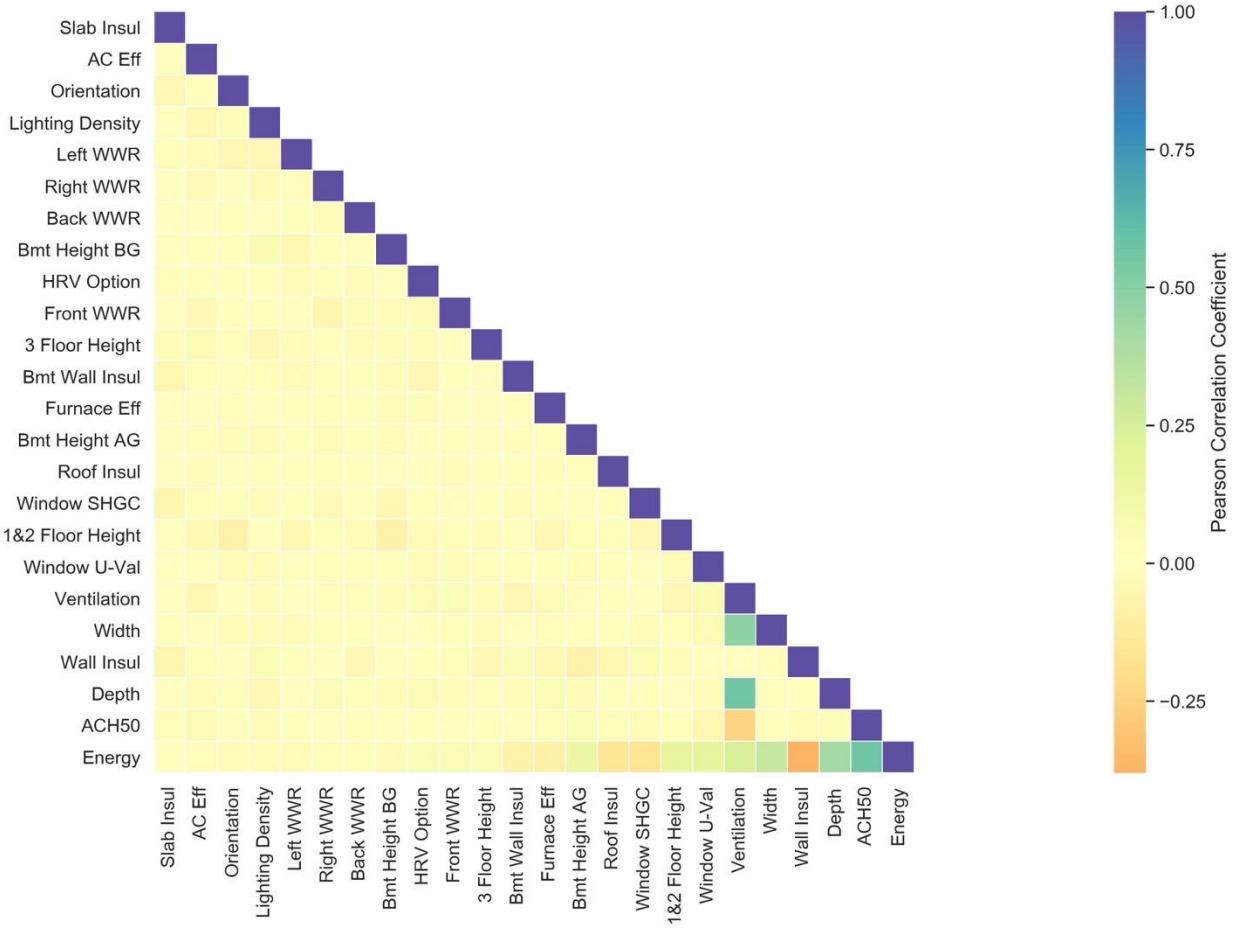

Figure 33 Pearson's correlation value for each input parameter to the output parameter.

The ACH_{50} and depth values are the most negatively correlated, and wall insulation is the most positively correlated. These are the same conclusions that were made after a visual inspection of the scatter plots.

To ensure that the correlation of ventilation with the other parameters did not affect the model, the variance inflation factor (VIF) was calculated. Figure 34 shows the VIF for each parameter.

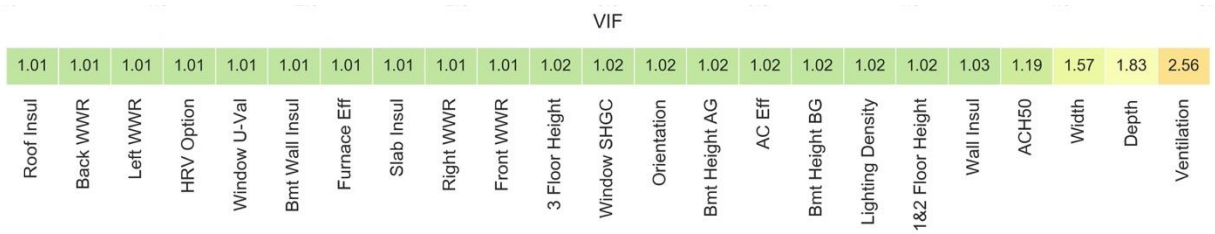

Figure 34 Variance inflation factor for each input parameter.

As expected, ventilation had the highest value followed by depth, width, and ACH_{50}. No values have a VIF larger than 5, therefore it was assumed that collinearity did affect the interpretation of the model's coefficients.

5.2 PHASE II – SURROGATE MODEL DEVELOPMENT

The dataset created in Phase 1 was used to train a model to predict energy use and the surrogate model was developed.

5.2.1 Nested Cross-Validation

Nested 10-fold cross-validation was used to evaluate the final model. The model was split into 10 folds, with each fold being used as the testing set once. This was the outer loop which was only used to *evaluate* the final model (not to choose it). The training set of the outer loop was split again into 10 folds, with each fold being used as a validation set once. This was the inner loop which was used to compare various models and to choose a final model. The final model was evaluated on the testing set of the outer loop. This is summarized in Table 13.

Table 13 Nested cross-validation purpose and split sizes.

	Training		Testing	Purpose
Outer Loop Sample Size	1350		150	Used to evaluate final model
	Training	**Validation**		
Inner Loop Sample Size	1215	135		Used to compare and choose final model

5.2.2 Linear Regression

Linear regression was performed using Python's *SKLearn* package [65]. The model was fit on the training set and evaluated on the validation set, using cross-validation to ensure accurate results. Figure 35 shows the results. The left plot shows predicted energy use versus simulated energy use, where predicted indicates the model's prediction, and simulated indicates the EnergyPlus results. The right plot shows the residuals. The scores shown are for a single fold of the 10-fold cross-validation.

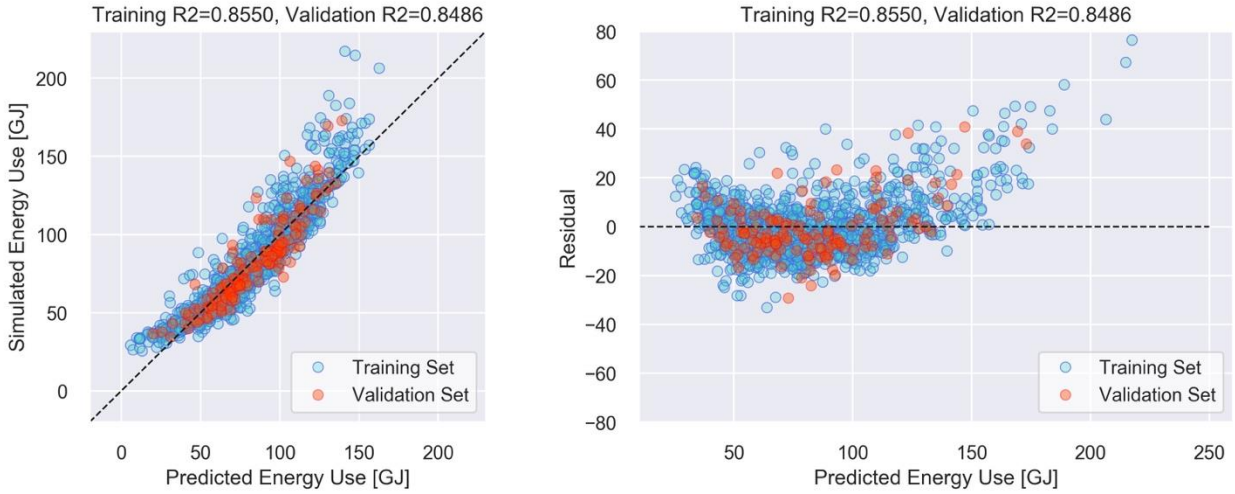

Figure 35 Un-transformed simulated versus predicted energy use and the residual plot.

The residuals are not evenly distributed around zero, indicating that the linear regression model does not fit the data. Linear regression assumes that there is a linear relationship between the expected value (energy use) and each independent variable (building parameters). Certain transformations may improve this linearity.

5.2.3 Regression Variable Transformations

Several transformation methods were evaluated to determine which, if any, were the most effective in improving the model. Table 14 shows the transformations and scores.

Table 14 Transformations performed on the input and output variables.

Method	Regression Equation	R^2 Validation	MAPE
Standard Linear Regression	y = b0 + b1x	0.848	12.0
Reciprocal Model	1/y = b0 + b1x	0.850	12.8
Quadratic Model	sqrt(y) = b0 + b1x	0.883	4.9
Exponential Model	log(y) = b0 + b1x	0.896	2.1
Logarithmic Model	y= b0 + b1log(x)	0.920	9.4
Power Model	log(y)= b0 + b1log(x)	0.958	1.4

The RMSE cannot be compared as it is measured in the unit of the output, so it was omitted. The power model (log-transformation of both the input and output variables) is the most effective transformation and will hereafter be referred to as the "log-transformed model" for simplicity.

The raw energy output distribution, shown on the left of Figure 36, is positively skewed. The log-transformed output, shown on the right, has a more normal distribution.

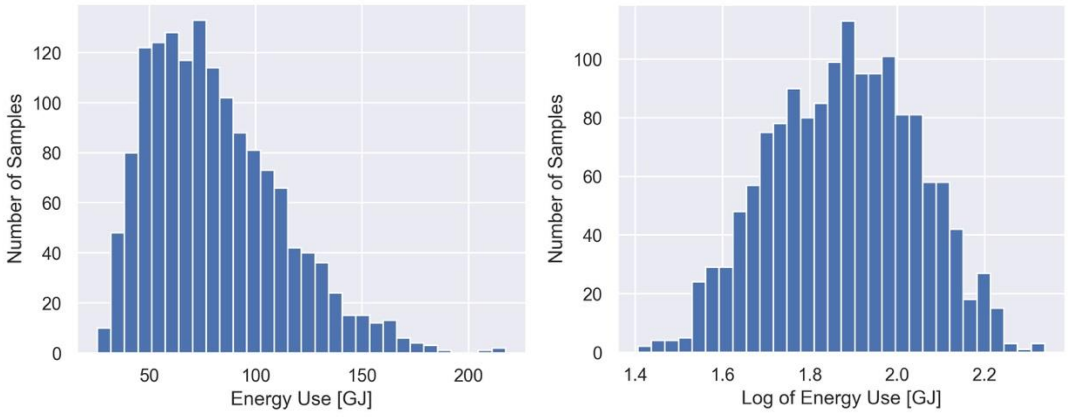

Figure 36 Distribution of energy use outputs un-transformed (left) and log-transformed (right).

Unskewing the distribution increases the linearity between the expected value and the independent variables. Figure 37 shows the new distributions of the log-transformed input variables. While these distributions are not necessarily normal, it emphasizes the linear relationship with the output.

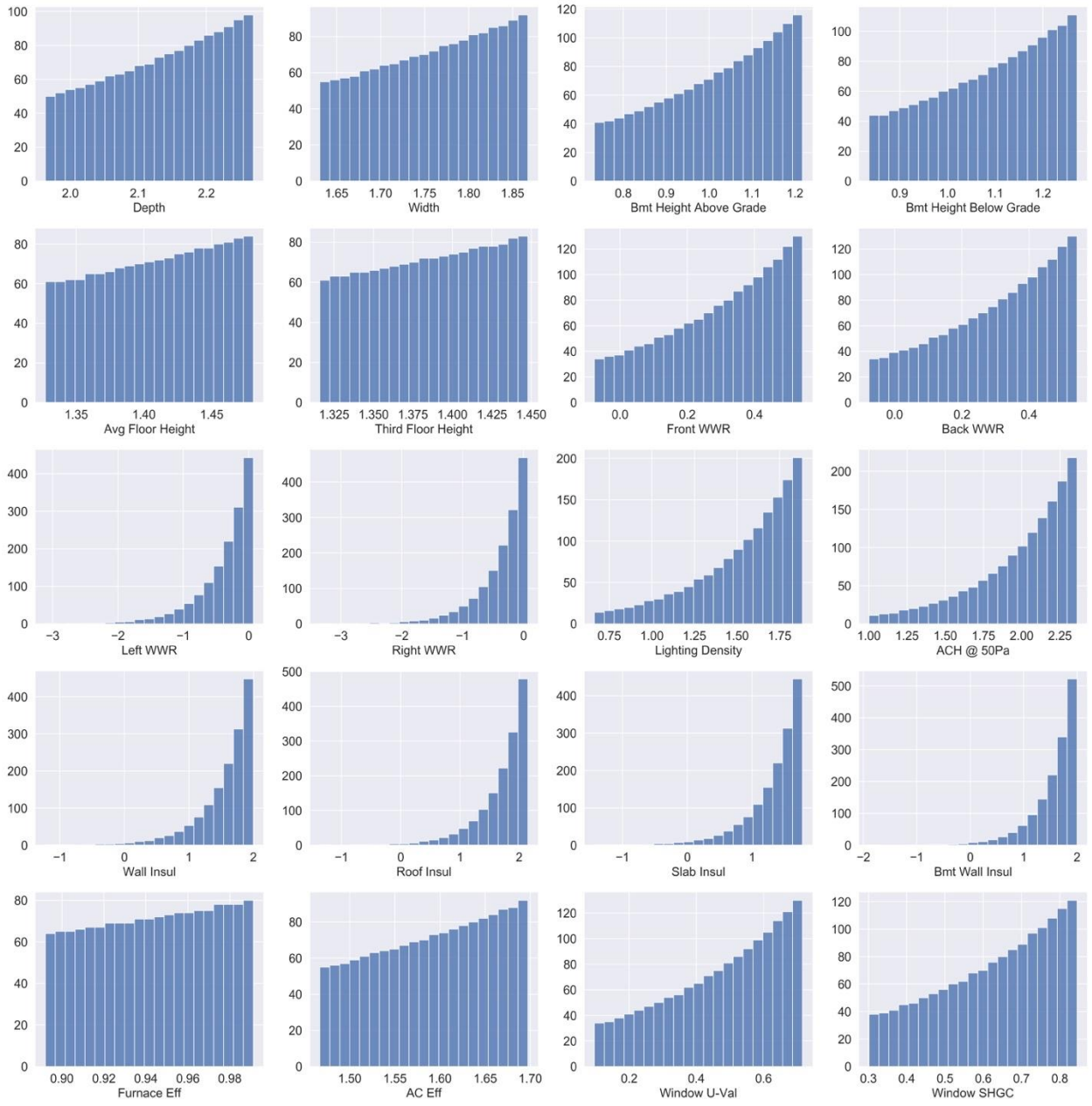

Figure 37 Log-transformed input parameter distributions.

Hair *et al* found that nonnormality in sample sizes of 200 or larger could have negligible effects [76]. The model performance increased which indicates these distributions increase the linear relationship between the input parameters and energy use.

The log-transformation of ventilation has created a more normal distribution, and there is a steep positive slope which indicates that ventilation is highly correlated with energy, shown in Figure 38. Ventilation was examined separately as it is the only parameter calculated from other inputs.

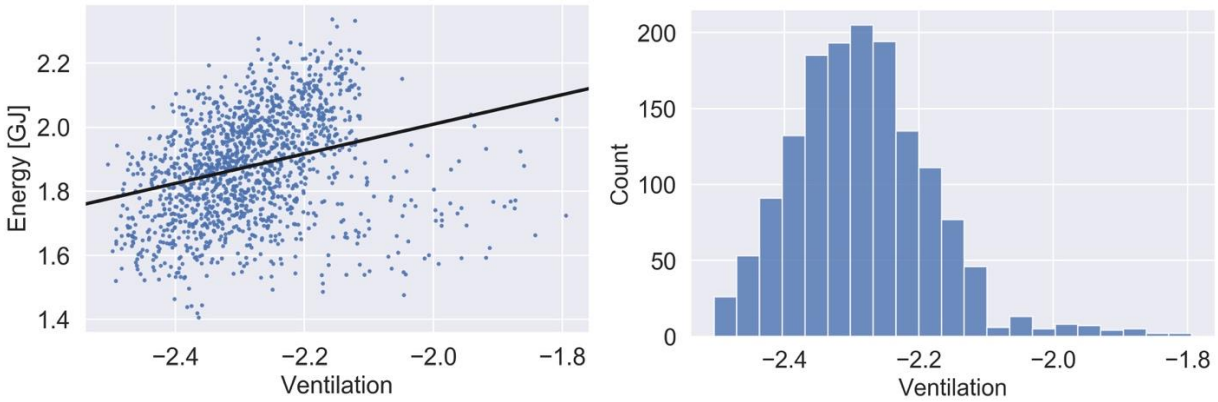

Figure 38 Scatter plot and distribution for log-transformed ventilation input.

5.2.4 Log-Transformation

Figure 39 shows the model with log-transformed input and output variables. The training and validation R^2 scores shown are for a single fold of the 10-fold cross-validation.

Figure 39 Simulated versus predicted energy use and the residual plot for the log-transformed model.

The residuals are evenly distributed around zero indicating the linear regression model of the log-transformed data has a better fit. Barnes reported the same findings [40].

All following models will include a log-transformation of the input and output variables. To predict energy use values in GJ, the output must be back-transformed, shown in Eq. 11 and 12.

$$\log(y) = b0 + b1 \log(x) \qquad (11)$$

$$y = 10^{(b0 + b1 \log(x))} \qquad (12)$$

Only the back-transformed metrics were reported for each of the models being considered.

5.2.5 House Size Analysis

Two methods were used to analyze the effect of house size on model performance. Method 1 included creating four separate models for small, medium, and large homes, and the fourth model would sample all the sizes (*i.e*. the entire dataset). The performance metrics for each model were compared. Method 2 trains a single model on the entire dataset but splits the validation dataset into three subsets for small, medium, and large homes based on floor area. A fourth validation subset is created based on a random sampling of all the sizes. The performance of the single model on each different size subset was compared. The results of the two methods was also compared.

The "medium" homes exist as the middle third so that there is a more definable difference between small and large homes, as the entire range of geometry measurements were sampled evenly. It should be noted that the results for medium sized homes are reported only to offer a more in-depth view of how the model is affected by size.

5.2.5.1 Method 1

The data was split into three evenly sized datasets based on floor area. A random sample of the combined dataset (all sizes) was created to enable comparisons. Each set had 500 samples. Figure 40 shows the energy use versus floor area for small, medium, and large homes.

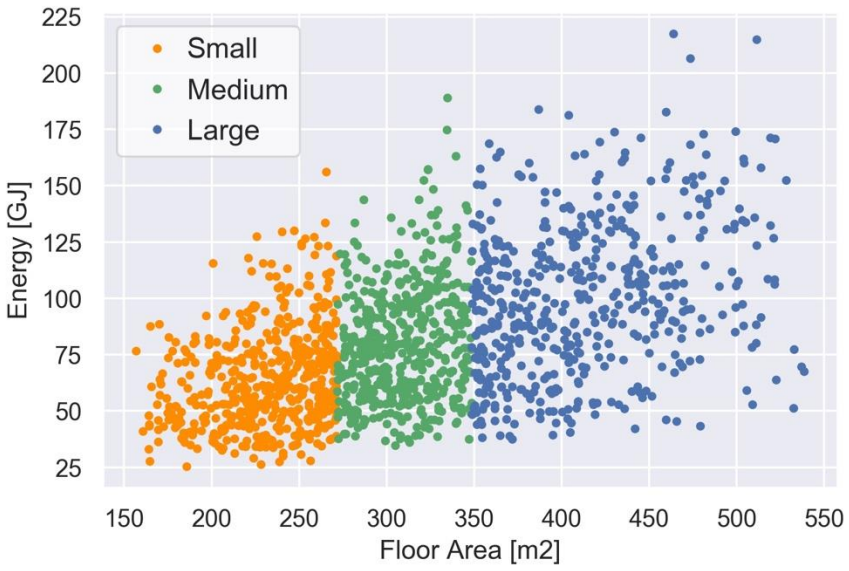

Figure 40 Energy versus floor area separated by house size (separated equally by thirds).

A sample size analysis was done for each dataset. Figure 41 shows the analysis for small, medium, and large homes, and a random sample of the combined dataset (all sizes). Each plot is shown on the same scale to allow for comparison. The shaded areas indicate the standard deviation for each point. This plot is showing log-transformed results instead of back-transformed results because of limitations of the *learning_curve* function [65] used to perform the analysis in Python.

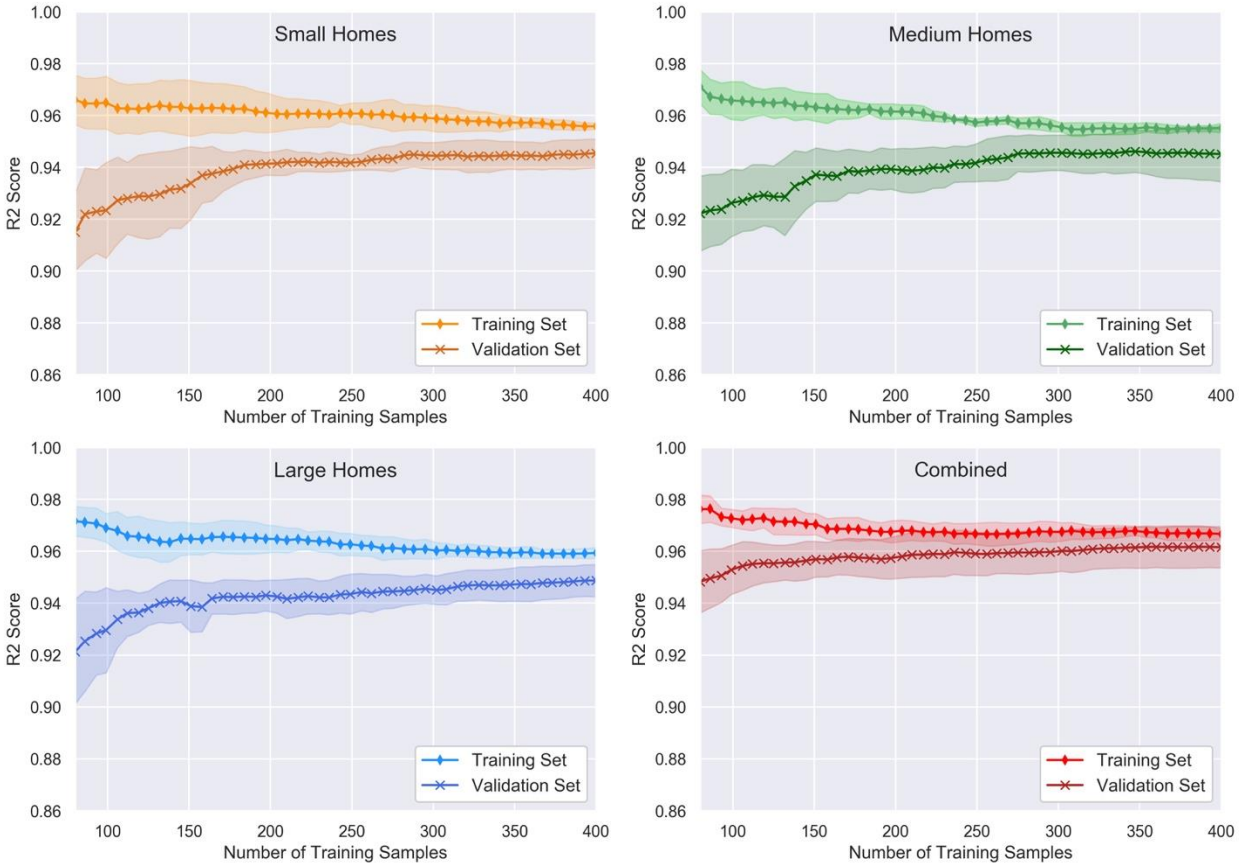

Figure 41 Sample size analysis for each house size using R^2 score.

This analysis indicates that the validation and training scores do not converge at 400 samples for the size separated homes. The combined sample size analysis is the closest to converging and therefore has the smallest variance. Note that the combined sample size is the same as the individual sizes (500 samples). This indicates that the R2 and RMSE would improve slightly with a larger sample size, however not drastically. This is a comparative analysis therefore this small difference is deemed negligible.

A linear regression model with log-transformed input and output variables was chosen for this comparison. The back-transformed model predicts energy use in GJ. Linear regression was chosen to simplify the comparison. The results of the back-transformed linear regression for the four models are shown in Figure 42. The left plot shows the simulated energy use versus the predicted energy use for back-transformed input and output variables. The right plot shows the

residual versus predicted energy use. Note the data points and R^2 scores in the following plots show the values for a single fold in the cross-validation loop. Each type of plot uses the same axes scale to allow for a visual comparison.

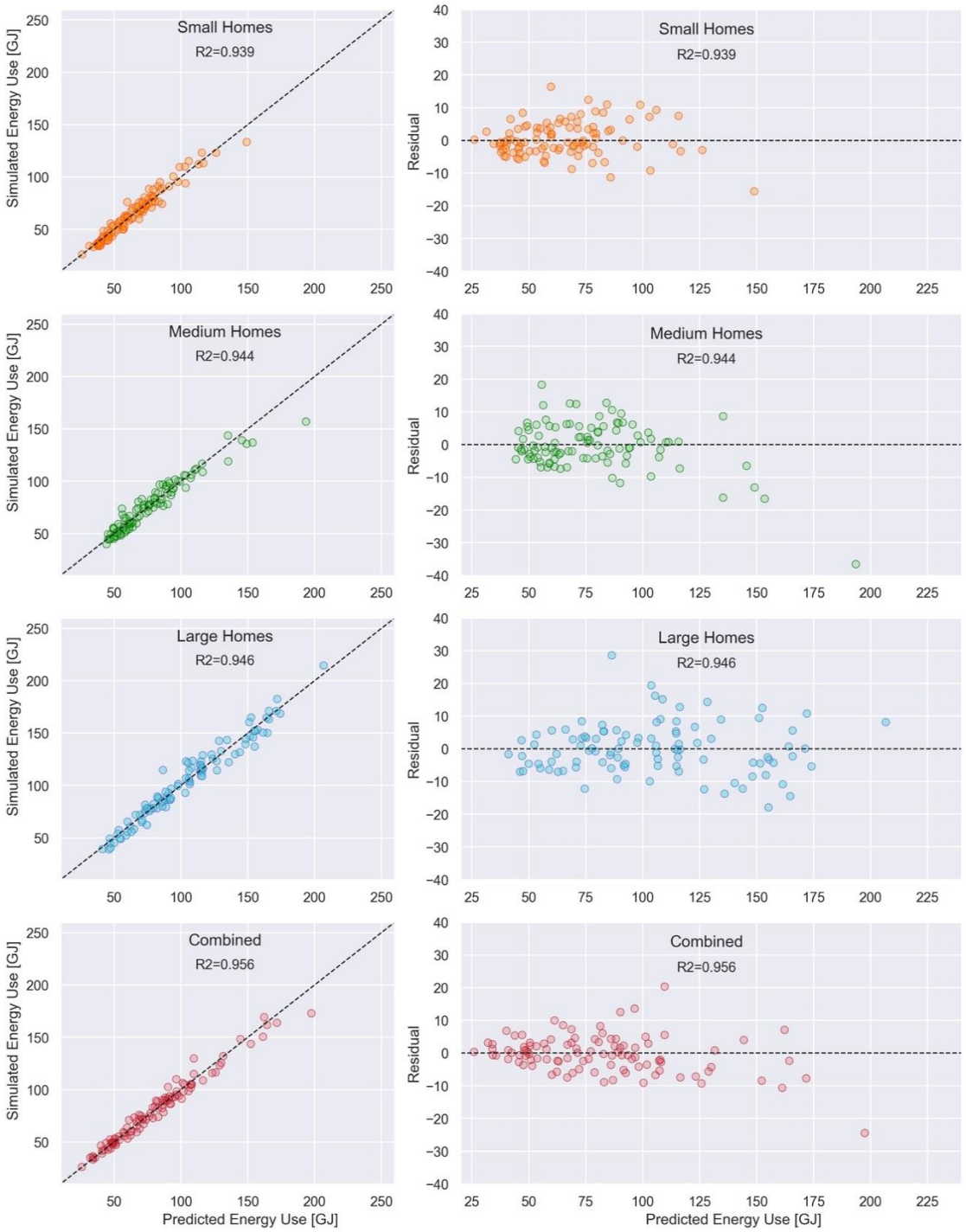

Figure 42 Simulated energy use versus predicted energy use and residuals for each house size.

The residuals are generally evenly distributed around zero and get larger as house size and energy use increases. As the size of the house increases, the range of output energy values increases, and there are few homes with very large energy uses. It also seems that the residuals increase as house size increases.

Ten time repeated 5-fold cross-validation was performed, which means the five folds were randomly generated 10 times resulting in 50 different validation and training combinations that were evaluated. The means of the evaluation metrics are summarized in Table 15.

Table 15 Performance metrics for each house size.

	R^2 Test	RMSE [GJ]	MAPE [%]
Small	0.939	5.43	5.90
Medium	0.944	6.03	5.62
Large	0.946	7.78	6.04
Combined	0.956	6.51	5.82

Figure 43 shows these values along with the standard deviation across the 50 values for each metric. The diamond represents the mean and the error bars indicate the standard deviation.

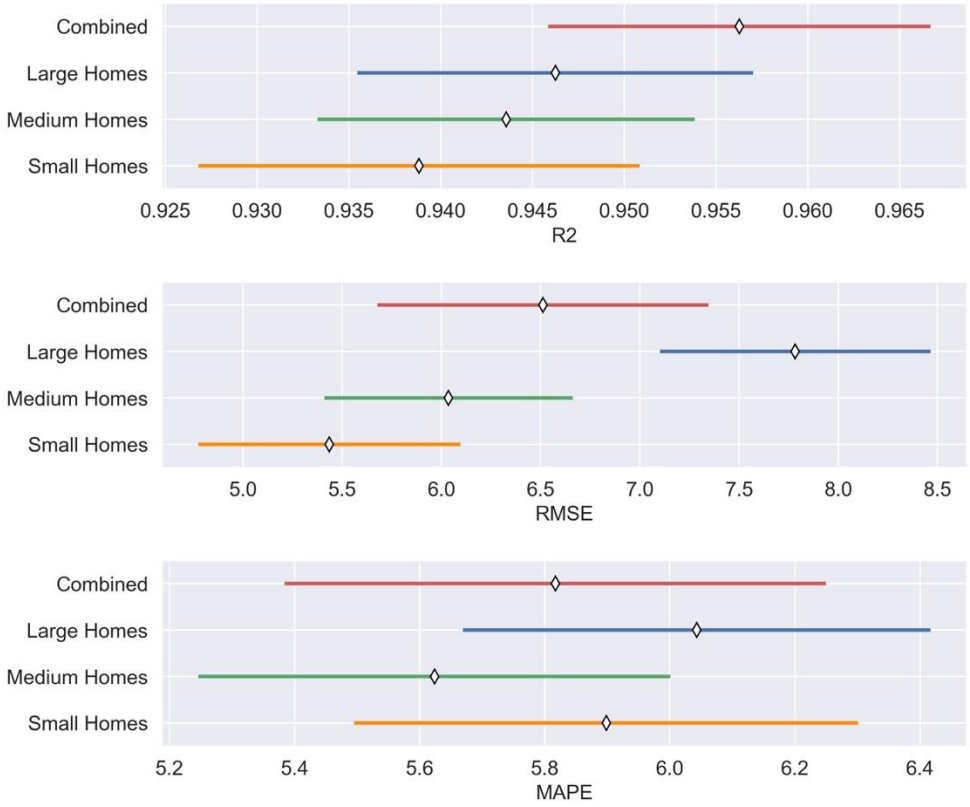

Figure 43 Performance metrics for R^2, RMSE, and MAPE for each house size.

Since there are four models, there are different coefficient values for each. Examining the magnitude and order of each allows meaningful conclusions to be drawn about how the different models are behaving, and if it is feasible to model these differently sizes homes in a single model. The average coefficient values calculated from the 10 times repeated 5-fold cross-validation are shown in Figure 44 for small, medium and large homes. They are organized in descending absolute values for large homes. The error bars represent the standard deviation.

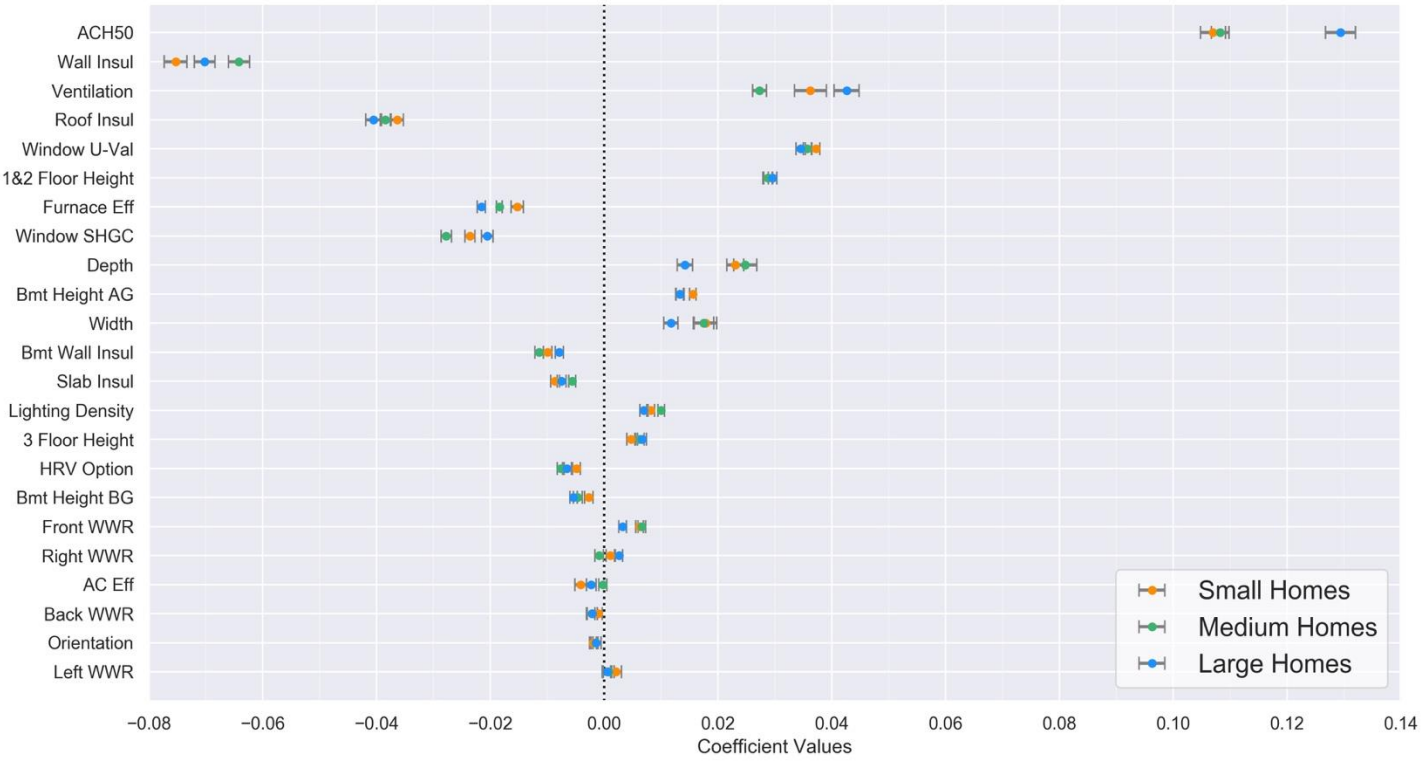

Figure 44 Coefficient values for each parameter for small, medium, and large homes.

The coefficient values are similar, although there are some differences in the order of absolute values. Table 16 shows the coefficient values and their associated standard deviation. The point of reporting the coefficients is to compare the order of the parameters ranked by the absolute magnitude of coefficients, shown by the order number.

Table 16 Coefficient values for large, medium, and small homes.

	Large Homes			Medium Homes			Small Homes		
Parameters	**Coefs**	**SD**	**Order**	**Coefs**	**SD**	**Order**	**Coefs**	**SD**	**Order**
ACH_{50}	0.1294	0.0026	1	0.1083	0.00158	1	0.1070	0.0022	1
Wall Insul	-0.0702	0.0018	2	-0.0642	0.00188	2	-0.0753	0.0020	2
Ventilation	0.0426	0.0022	3	0.0273	0.00121	7	0.0362	0.0028	5
Roof Insul	-0.0405	0.0014	4	-0.0385	0.00088	3	-0.0364	0.0011	4
Window U-Val	0.0345	0.0008	5	0.0357	0.00071	4	0.0372	0.0007	3
1&2 Floor Height	0.0296	0.0007	6	0.0288	0.00073	5	0.0288	0.0009	6
Furnace Eff	-0.0216	0.0007	7	-0.0184	0.00050	9	-0.0153	0.0011	11
Window SHGC	-0.0205	0.0010	8	-0.0278	0.00093	6	-0.0236	0.0009	7
Depth	0.0142	0.0013	9	0.0248	0.00203	8	0.0230	0.0015	8
Bmt Height AG	0.0133	0.0007	10	0.0133	0.00062	11	0.0156	0.0006	10
Width	0.0117	0.0013	11	0.0175	0.00178	10	0.0178	0.0020	9
Bmt Wall Insul	-0.0079	0.0007	12	-0.0114	0.00074	12	-0.0099	0.0007	12
Slab Insul	-0.0075	0.0008	13	-0.0056	0.00062	17	-0.0086	0.0008	13
Lighting Density	0.0070	0.0007	14	0.0100	0.00058	13	0.0082	0.0007	14
HRV Option	0.0067	0.0008	15	-0.0076	0.00061	14	-0.0049	0.0007	16
3 Floor Height	-0.0065	0.0008	16	0.0063	0.00075	16	0.0047	0.0007	17
Bmt Height BG	-0.0054	0.0007	17	-0.0046	0.00076	18	-0.0027	0.0007	19
Front WWR	0.0032	0.0007	18	0.0066	0.00067	15	0.0062	0.0007	15
AC Eff	0.0026	0.0006	19	-0.0002	0.00071	23	-0.0041	0.0010	18
Right WWR	-0.0023	0.0008	20	-0.0009	0.00072	21	0.0011	0.0008	22
Orientation	-0.0021	0.0009	21	-0.0015	0.00046	20	-0.0020	0.0006	21
Back WWR	-0.0014	0.0009	22	-0.0022	0.00094	19	-0.0010	0.0006	23
Left WWR	0.0005	0.0008	23	0.0008	0.00094	22	0.0021	0.0009	20

Some observations can be made by comparing the order numbers. ACH_{50} and wall insulation are the most descriptive parameter in all three models. The furnace efficiency is more important as the home gets larger. The furnace coil is auto sized which means it increases as the house size increases, indicating that the furnace efficiency will have a greater impact on energy use with larger heating coils. The window U-value and SHGC is more important as the home gets smaller. The wall to floor area is increasing, indicating that the thermal resistance of the windows has a larger impact. The width and depth become less important as the homes gets larger. The larger

floor area could indicate that the surface to volume ratio is lower which would result in less heat loss through the envelope. Geometry is important to small homes, then as they get larger and the auto sized HVAC equipment becomes larger, some of the HVAC descriptors can more accurately predict energy use (for example, furnace efficiency and ventilation are more important in large homes). The ventilation and AC efficiency are less important for the medium homes than the small or large, and the HRV option is least important for the small homes. This could be due to the same point made above. Geometry and window descriptors are better able to describe energy use in small/medium homes if the HVAC equipment is smaller.

Note that "importance" is used to mean "the statistical significance within these ranges". This comparison was done to determine if there were significant differences in the order of the parameters with respect to coefficient magnitudes.

5.2.5.2 Method 2

The same linear regression model with log-transformed input and output variables was used as in method 1. The back-transformed model's metrics are reported. Figure 45 shows the actual vs predicted energy use for small, medium, and large homes on the same scale. The residuals are shown on the right.

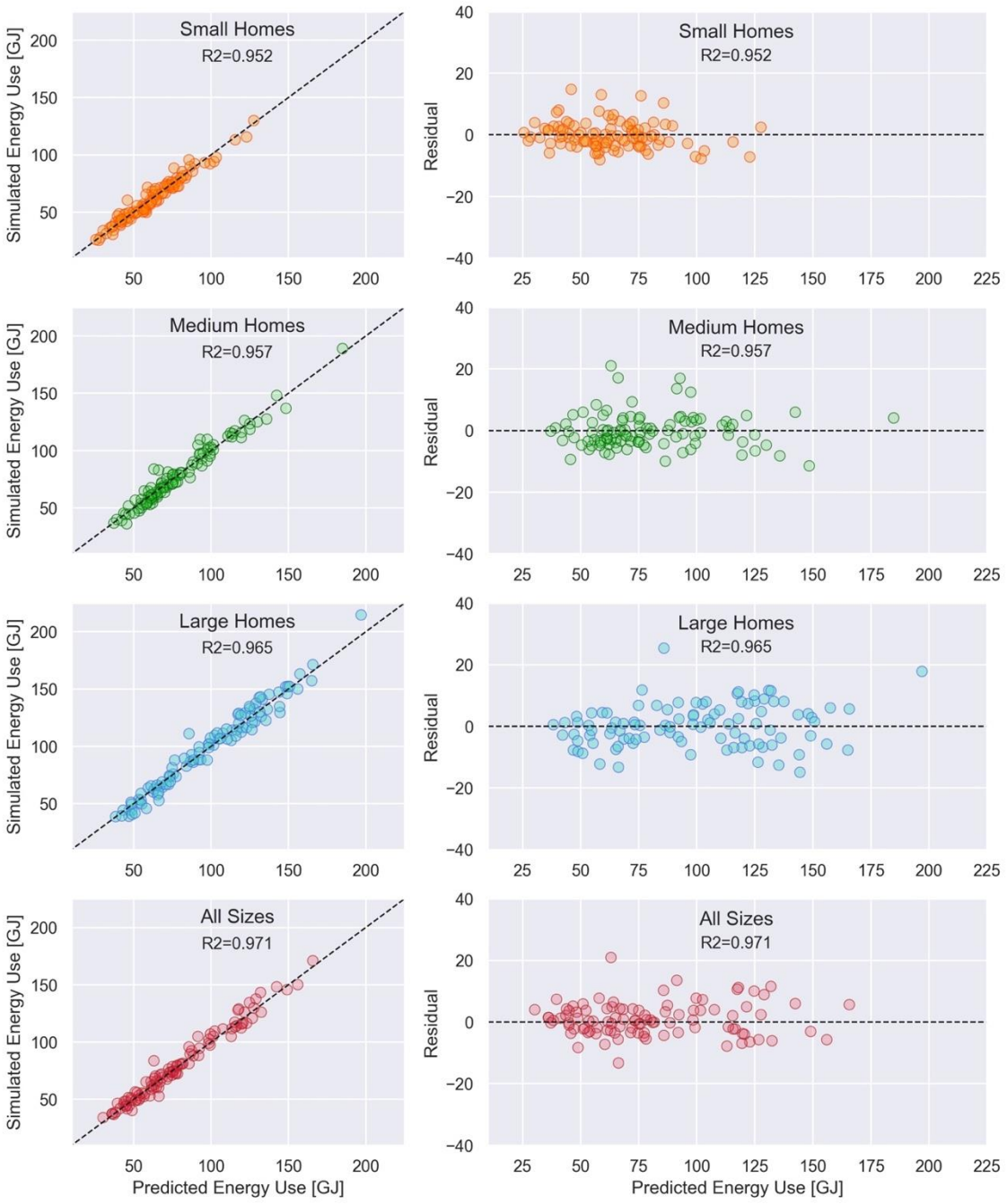

Figure 45 Simulated versus predicted energy use and residuals for each home size.

These residuals look similar to the results of the individual models in Figure 42. As size increases, energy use output increases, and so do the absolute residual values. There seems to be more overlap between energy use values across house sizes compared to method 1.

The evaluation metrics are recorded for each validation subset. The results are shown in Table 17. These values are the means of the cross-validated results.

Table 17 Performance metrics for each house size.

	R^2 Test	RMSE [GJ]	MAPE [%]
Small	0.942	5.22	5.82
Large	0.945	7.83	6.12
Medium	0.949	5.71	5.35
All	0.959	6.23	5.62

The means represented by the diamonds and their associated standard deviations represented by the error bars are shown in Figure 46.

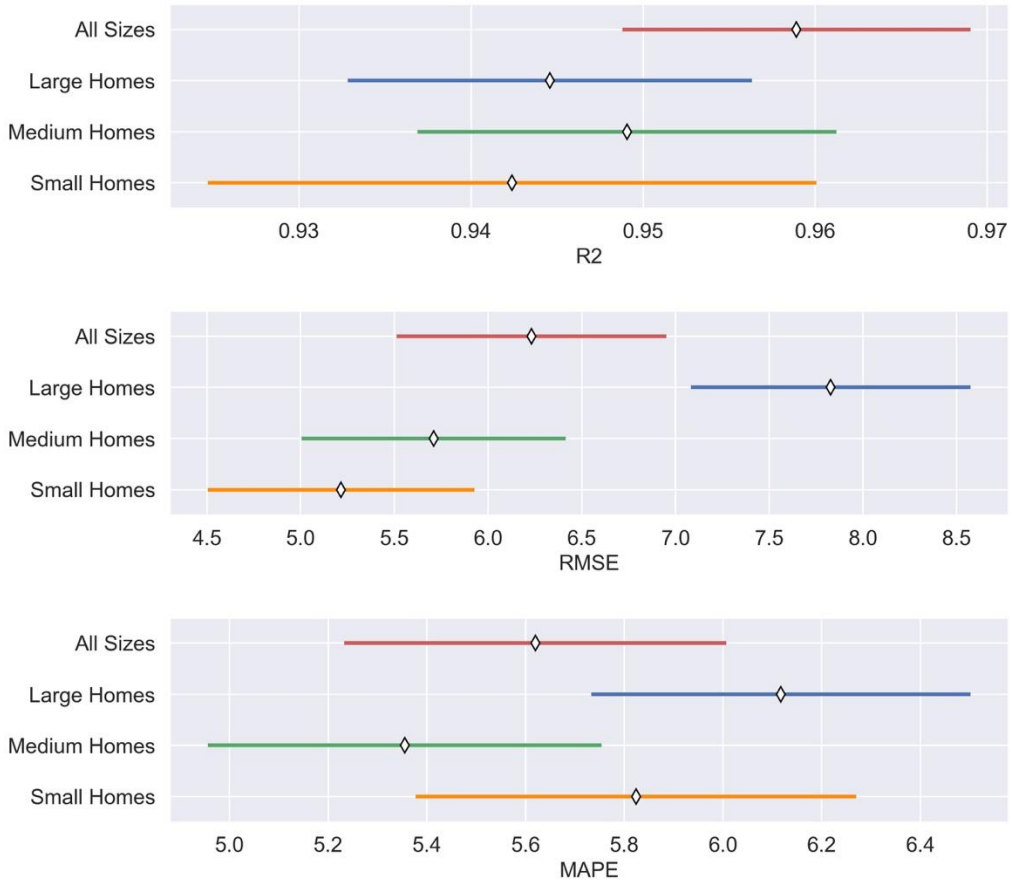

Figure 46 Performance metrics for R^2, RMSE, and MAPE for each house size.

5.2.5.3 Discussion

Some observations can be drawn from the results from methods 1 and 2. It is assumed that the RMSE values are proportional to the increasing magnitude of energy use outputs as the house size increases and will therefore focus on R^2 and MAPE.

- The combined homes have the best R^2 score, and a better MAPE than small and large homes.
- Other than the combined set, the medium homes have the best R^2 and MAPE values.
- The small homes have the worst R^2 score with the largest standard deviation.
- The only difference from methods 1 and 2 is that the medium homes had a worse R^2 score than the small and large homes in method 1, and better in method 2.

The R^2 value is influenced by the span and range of the output values, as it includes the residual sum of squares and total sum of squares in its calculation. Since the total sum of squares value is much larger than the sum of squared residual value, models with larger sum of squared residuals (larger prediction error) that span greater output ranges can have better R^2 scores. That is why an error metric should always accompany R^2, because a good line of best fit does not necessarily mean small prediction error.

The average absolute residual value within energy use bins of 10 GJ are shown in Figure 47, plotted against predicted energy use for each bin. These values are 10 sets of validation data for each size. The RSS is the total sum over the 10 loops, and the numbers above the bars indicate the number of values that are included in that energy bin.

Figure 47 Average absolute residual value for 10 GJ bins of predicted energy use.

The residuals increase as house size increases. Other than the residual being an absolute value that increases as the energy output gets larger, there are several possible reasons for this. At very low energy uses, the more predictable base loads make up a larger percentage of the total energy use, resulting in more accurate predictions. Some input parameters have strong relationships that are only linear for smaller energy uses. As seen in the energy use histogram plot in Figure 36, the distribution of energy uses is normal when log-transformed, however the un-transformed energy uses in GJ show a positively-skewed distribution, resulting in more samples in the low to mid-range of energy use outputs which could explain more accurate predictions for these values.

The RMSE is calculated as the square root of the sum of squared residuals divided by the number of samples. The number of samples in each validation subset is the same. RMSE is an absolute metric and the range is reflected by the magnitude of the energy use outputs, so the

results are expected. The MAPE is a more interesting metric as it describes the percent error which is unaffected by magnitude of energy use outputs.

The MAPE is calculated as the sum of the residual divided by the actual (simulated) value. Lower MAPE values would result from large energy use outputs (a large denominator) and small residuals (a small numerator). The medium homes have low sum of squared residuals and mid-range energy use outputs, resulting in the smallest MAPE. Small houses have low sum of squared residuals but smaller energy use outputs, resulting in a larger MAPE. The large homes have larger energy use outputs, but much larger SSRs, resulting in the largest MAPE. The combined dataset has the full range of energy use outputs, however many of them are focused in the lower energy use range, and mid-range sum of squared residuals. This results in a better MAPE than the small and large homes.

5.2.5.4 Summary

It is important to understand the effect that the size of homes has on model accuracy in order to make decisions on how to create archetypes moving forward. The combined models have a slightly better R^2 and MAPE values than the small *and* large home models for both methods. However, the values are all very similar. Creating a single archetype that combines small and large homes results does not decrease the accuracy of the model. This implies that surrogate models can be created to incorporate more houses that fall under the same characteristics but vary in size. This allows bottom-up surrogate models to describe greater subsets of the Toronto housing stock within a single model. The coefficient value analysis in method 1 determined that the small and large houses had the same top six parameters in terms of absolute coefficient value.

5.2.6 Model Comparison

Four models were compared to determine the approach that was used to create the final surrogate model. Two models used multivariate linear regression, one with *label encoding* the categorical variables and one with *one-hot encoding*. Elastic net was used with both types of categorical variable encoding for the purpose of reducing parameters. The objective was to use the multivariate linear regression model to quantify the accuracy that would be sacrificed by reducing the number of input parameters in the elastic net model.

5.2.7 Linear Regression with Continuous and *Label Encoded* Inputs

The actual (simulated) energy use values were compared to the back-transformed predicted values. Figure 48 shows the back-transformed data. The training and validation R^2 scores shown are for a single training and validation set only.

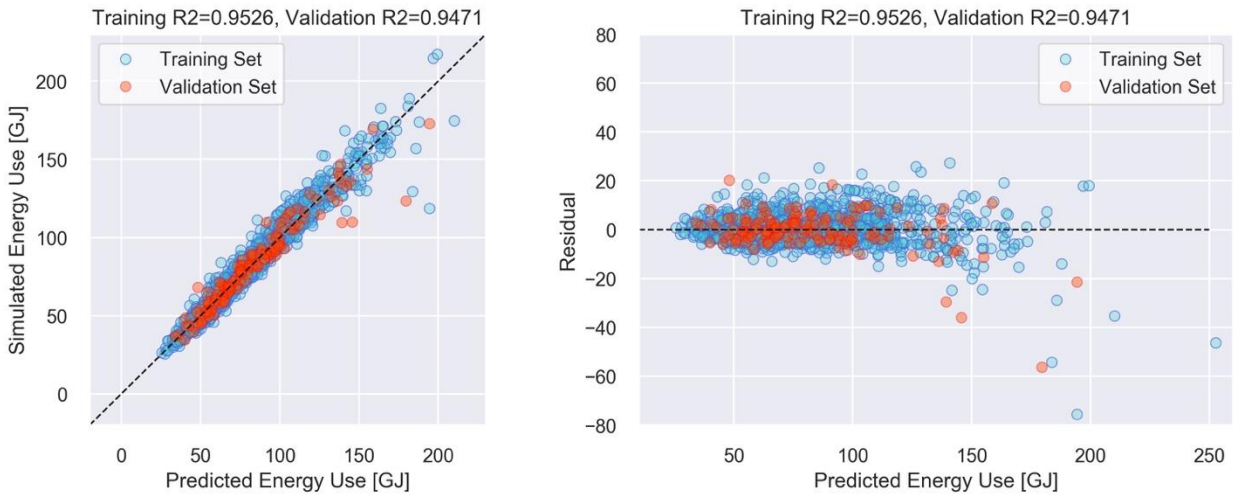

Figure 48 Back-transformed simulated versus predicted energy use and residuals.

The residuals are mostly distributed around zero, however there is slight heteroscedasticity (the residuals become larger as simulated energy use increases) which could be caused by some outliers at the largest energy use outputs. Figure 49 shows the absolute percent error versus simulated energy use on the left, and a box plot of the absolute percent error on the right.

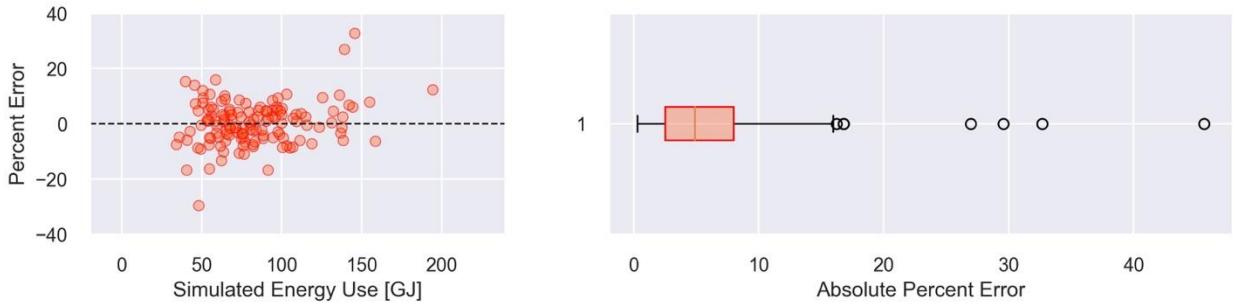

Figure 49 Percent error versus simulated energy use (left) and a box plot (right).

The mean absolute percent error (MAPE) of the data is 5.89%. The box plot indicates that the error ranges from 0-15% (excluding outliers), and the first and third quartile (*i.e.* 50% of the data) range from 3-8%.

The results of the cross-validated linear regression using numerical and *label encoded* inputs are shown in Table 18. This model will be referred to as the "*label encoded* linear regression" model.

Table 18 Back-transformed performance metrics with linear regression with label encoded inputs.

	R^2	RMSE [GJ]	MAPE [%]
Back Transformed	0.9472	7.02	5.89

Only the back-transformed metrics are reported as it predicts values in GJ. The untransformed and log-transformed results are in-between steps and will not be considered.

5.2.8 Elastic Net Regression with Numerical and *One-Hot encoded* Inputs

Elastic net regression was performed for L1 ratios of 0, 0.25, 0.5, 0.75, and 1. For each of these L1 ratios, the R^2 score, RMSE, and MAPE was calculated for 200 tuning parameters ranging from 0.0001 to 0.09. The L1 ratio and tuning parameters used for each calculation are called the model's hyperparameters, and each different set of hyperparameters is considered a new model. For each of set of hyperparameters, 10-fold cross-validation was performed.

Figure 50 shows the effect changing the L1 ratio has on the values of the coefficients as the tuning parameter increases. Lasso regression (left) indicates a L1 ratio of 1, while ridge regression (right) indicates a L1 ratio of 0. Any value in-between 0 and 1 indicates elastic net regression.

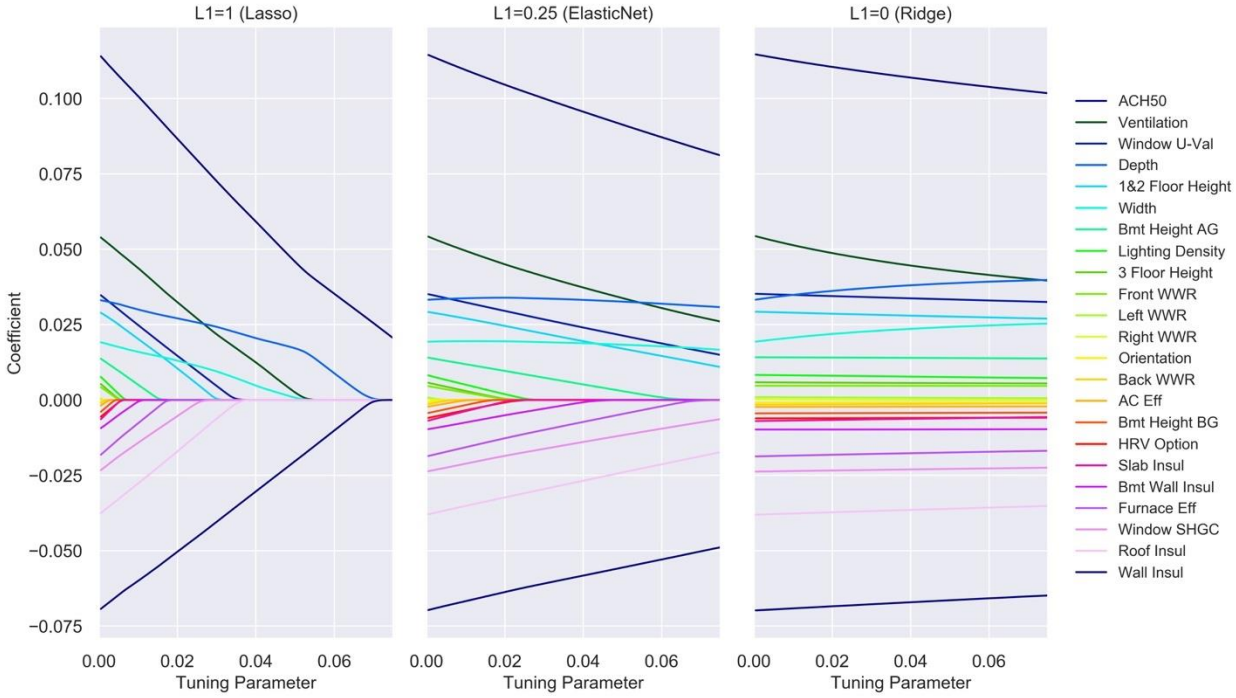

Figure 50 Coefficient values versus tuning parameter for lasso, elastic net, and ridge.

As seen in the right plot of Figure 50, the coefficients are reduced but never removed. In the left plot, almost all the coefficients are removed (if the tuning parameter were to keep increasing, there would eventually be zero coefficients left). The middle plot with an L1 ratio of 0.25 indicates that both ridge and lasso penalties are being applied, resulting in shrinkage of all the coefficients and removal of some.

5.2.8.1 Hyperparameter Selection

The most accurate value for each set of hyperparameters was selected (maximum for R^2, minimum for RMSE), and the standard error for each was calculated. The minimum number of

coefficients that have an accuracy greater than the maximum accuracy with all the coefficients minus the standard error of that accuracy is selected. The equations are shown in Eq. 13-14.

$$R2_{min} \geq R2_{max} - SE \qquad (13)$$

$$RMSE_{max} \leq RMSE_{min} + SE \qquad (14)$$

Figure 51 shows the results (R^2 and RMSE) of the validation set. Only the most accurate point was shown for each L1 ratio and each number of coefficients. The left plots show the entire range of tuning parameters, which shows how the accuracy of the model changes as the tuning parameter increases and the number of coefficients decreases. The right plots show a close-up of the top ten number of coefficients. This is shown for one loop in the cross-validation process.

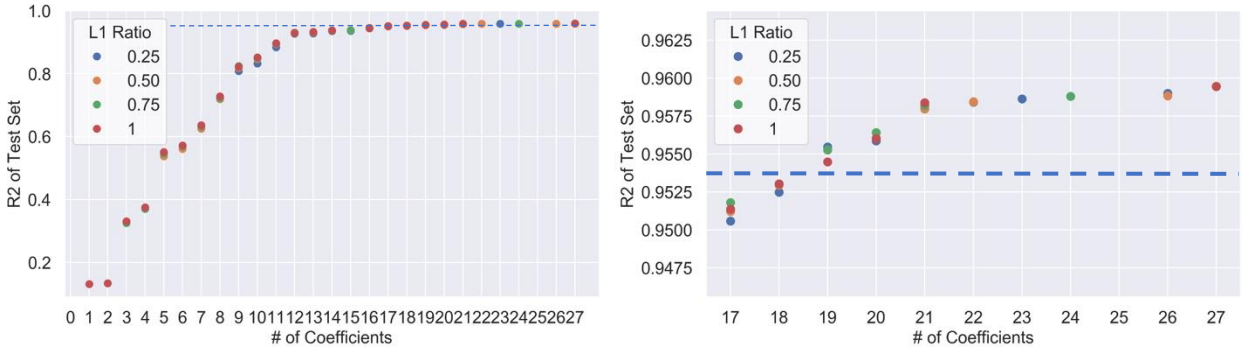

Figure 51 R^2 score versus number of coefficients for various L1 ratios.

The model with the best RMSE within one standard deviation of the RMSE of the best model falls at 19 coefficients. This is shown by the dashed line. Figure 52 shows 19 coefficients only.

Figure 52 R^2 score for 19 coefficients for various L1 ratios.

The tuning parameter value in this case is 0.00236, and the most accurate L1 ratio as seen above is 1. These would be the hyperparameters (*i.e*. the model) selected for this fold. Figure 53 shows the values of each coefficient for the selected hyperparameters. The Figure on the right is a close-up view of the Figure on the left. The dashed line shows where the model was selected as a result of the one-standard error rule.

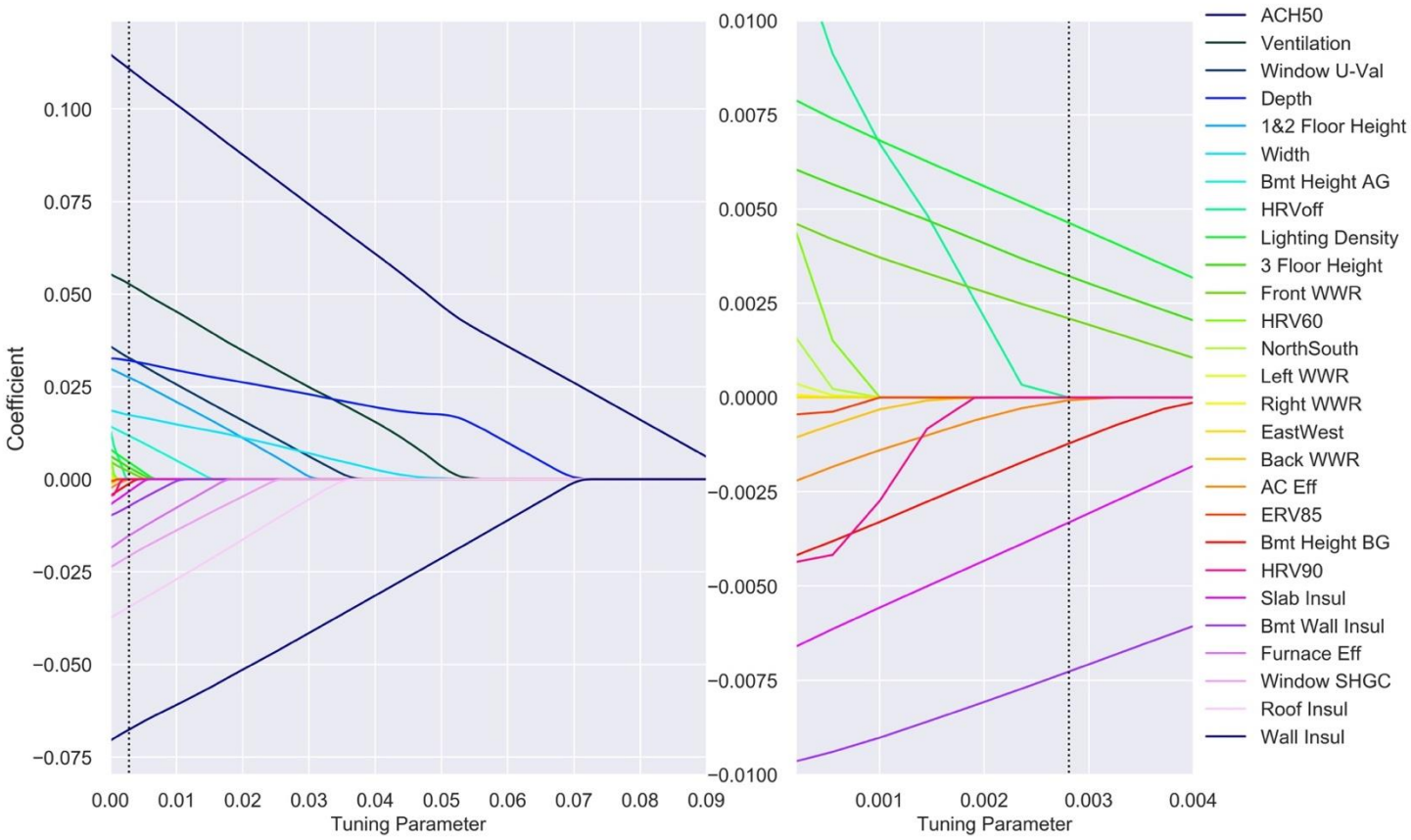

Figure 53 Coefficient values versus tuning parameter for lasso, elastic net, and ridge.

Since 10-fold cross-validation is used, this is performed 10 times. The set of tuning parameters and L1 ratios for each fold are shown in Table 19.

Table 19 Hyperparameters for each fold of the cross-validation.

Fold	Tuning Parameter	L1 Ratio
1	0.003714	0.50
2	0.003714	0.50
3	0.002359	1.00
4	0.002811	0.50
5	0.002359	0.75
6	0.003714	0.50
7	0.002359	0.75
8	0.002359	0.75
9	0.001907	1.00
10	0.002359	0.75
Mean	0.002766	0.700

Elastic net regression using the mean of the hyperparameters was performed using 10-fold cross-validation. Table 20 summarizes the findings for the back-transformed model.

Table 20 Performance metrics for the elastic net model with one-hot encoded *categorical variables.*

	R^2	RMSE [GJ]	MAPE [%]	Tuning Parameter	L1 Ratio	# Coefs
Back Transformed	0.9462	7.11	6.18	0.2766	0.7	19.5

5.2.8.2 Lasso Parameter Elimination

Removing coefficients causes the model accuracy to decrease, however it would be interesting to know in what order the coefficients are being removed (*i.e.* the least important in predicting energy use). Lasso is used (L1 ratio of 1) to show when all the coefficients are forced to zero. Figure 54 shows the 27 parameters and the order in which they were removed from the Lasso regression.

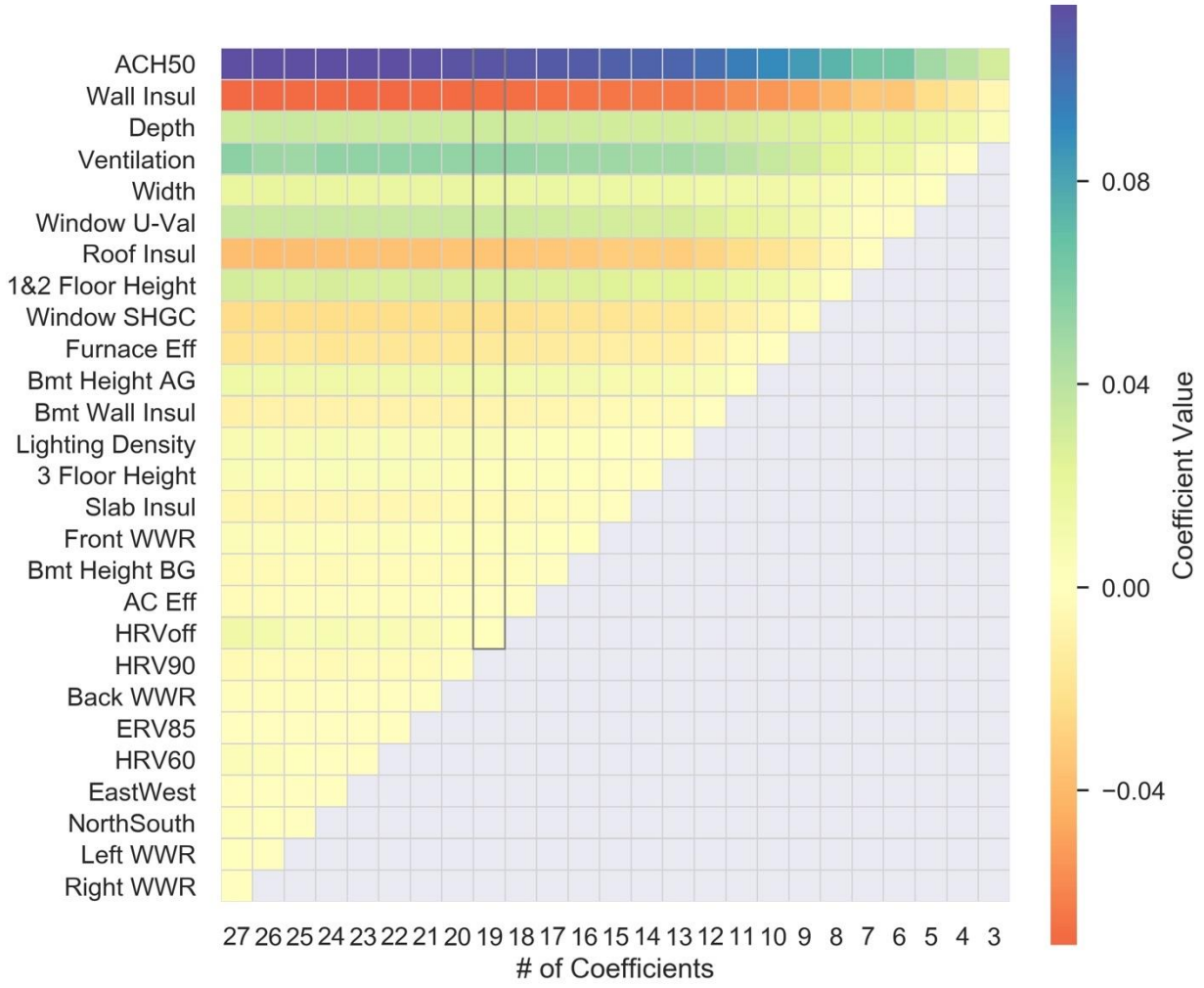

Figure 54 Number of coefficients in order of which they were dropped by the embedded feature selection (elastic net).

The model with 19 coefficients is outlined to show which variables are most likely to be included in the final model.

5.2.9 Final Model Selection

This same process was repeated for linear regression using *one-hot encoding* and elastic net regression with *label encoding*. The graphs and values were not shown as they follow the same trends. Table 21 summarizes the results for all the back-transformed models. Note that R^2 is specifying correlation and should be maximized. RMSE and MAPE are specifying error and should be minimized.

Table 21 Performance metrics of all models developed.

	R^2	RMSE [GJ]	MAPE [%]	# Coefs
LE Linear Regression	0.9472	7.02	5.89	23
OHE Linear Regression	0.9472	7.02	5.88	27
LE Elastic Net	0.9466	7.08	6.10	19.5
OHE Elastic Net	0.9462	7.11	6.18	19.5

The values are the means reported from each value of the 10-fold cross-validation. Figure 55 shows the means (diamond) along with the standard deviation (blue line) and minimum and maximum values (grey lines). The selected model is outlined.

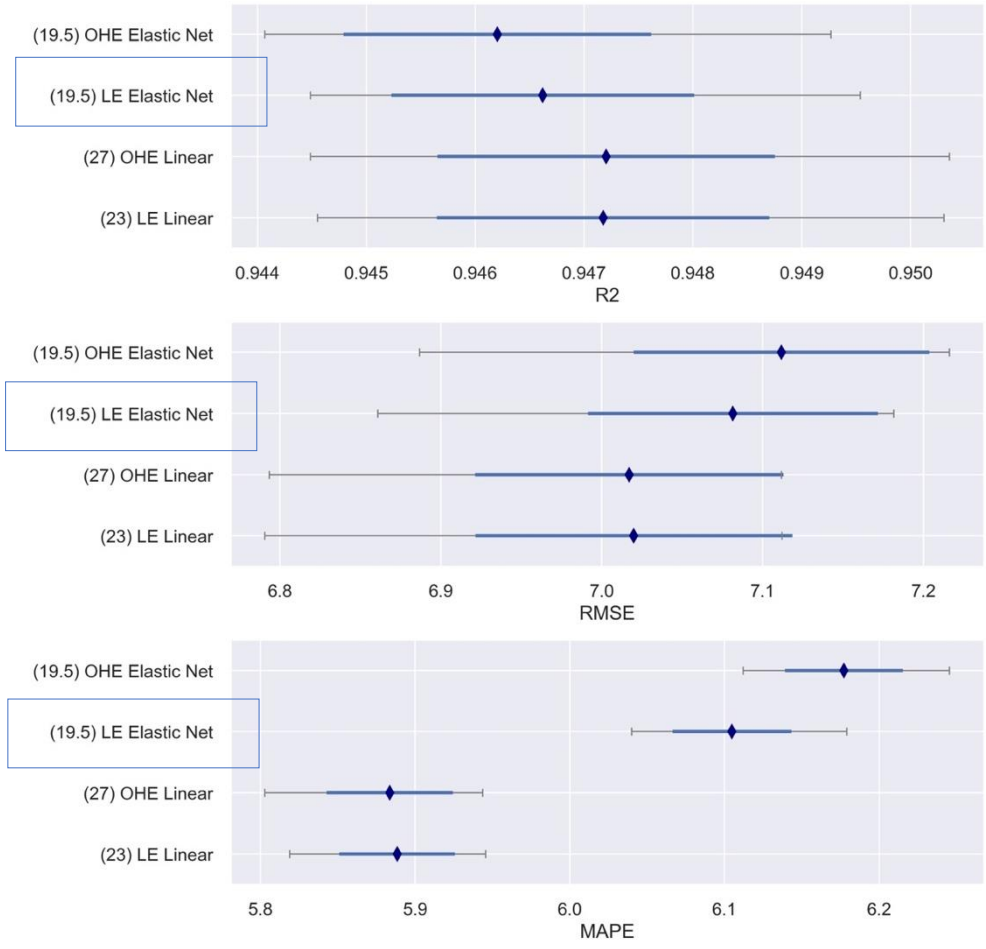

Figure 55 Performance metrics for each model developed.

The elastic net model is chosen because it reduces the number of coefficients and the objective of this research was to select the simplest model without sacrificing a large amount of accuracy. The MAPE is reduced from 5.89 to 6.1% which is negligible when comparing to the energy use of buildings. In a house using 250 GJ of energy annually, which is on the highest end of the simulated results, the decrease in accuracy by selecting the elastic net model would result in a 0.53 GJ difference. Between the *label encoded* and *one-hot encoded* models, the *label encoded* model is chosen as it is simpler than having separate parameters for each categorical option.

The convergence of the training and validation sets during the sample size analysis implies that the model has low variance. This explains why elastic net increases the model error, because generally regularization is used to reduce variance by introducing some bias. In this case elastic net was used as a feature selection methodology at the cost of introducing this small amount of bias.

5.2.9.1 Final Model Results

The training and validation sets are combined to form the new training set of the outer loop. The selected elastic net model is trained on this data and tested on the test set from the outer loop. The model has not seen this data before. The results are shown in Table 22.

Table 22 Performance metrics of final model.

	R^2	RMSE [GJ]	MAPE [%]	# Coefs
LE Elastic Net	0.9465	7.07	6.09	19

This indicates that 94.7% of the variance is accounted for by the model. The mean of the energy use output is 81 GJ. The RMSE is 7.07 GJ, and the MAPE is 6.09%.

5.2.9.2 Coefficients

The coefficient values are calculated for each of the 10 folds in the outer loop. The mean and standard deviation of the coefficient values from the final model are shown in Table 23. They are ordered by largest absolute value.

Table 23 Coefficient values and y-intercept for final model.

Parameter	Coefficient	STD	Parameter	Coefficient	STD
ACH_{50}	0.1117	0.00092	**Lighting Density**	0.0058	0.00034
Wall Insul	-0.0678	0.00066	**Slab Insul**	-0.0045	0.00029
Ventilation	0.0521	0.00115	**HRV Option**	-0.0040	0.00024
Roof Insul	-0.0360	0.00082	**3 Floor Height**	0.0036	0.00034
Window U-Val	0.0332	0.00034	**Front WWR**	0.0029	0.00030
Depth	0.0328	0.00067	**Bmt Height BG**	-0.0023	0.00023
1&2 Floor Height	0.0277	0.00027	**AC Eff**	-0.0005	0.00023
Window SHGC	-0.0218	0.00034	**Right WWR**	0.0000	0.00000
Width	0.0188	0.00065	**Orientation**	0.0000	0.00000
Furnace Eff	-0.0165	0.00029	**Left WWR**	0.0000	0.00000
Bmt Height AG	0.0124	0.00047	**Back WWR**	0.0000	0.00000
Bmt Wall Insul	-0.0080	0.00041	**y-intercept**	1.87836	

The coefficient values are plotted in Figure 56 with the error bars indicating standard deviation. They are ordered by largest to smallest absolute value.

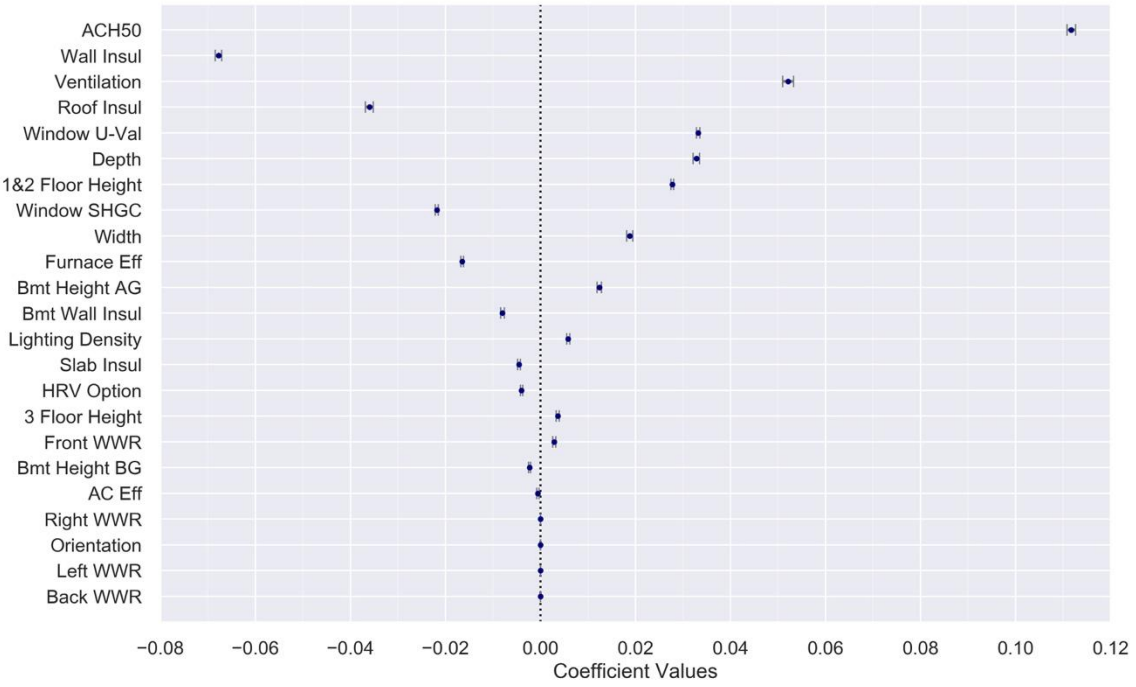

Figure 56 Coefficient values for final model.

The standard deviations are barely visible in most cases, indicating the coefficient values over the 10 folds of the outer loop are very similar. This indicates the model is stable. The elastic net regression removed right, left, and back window-to-wall ratios (WWR) and orientation from the final model by shrinking the coefficient values to zero. The range of modelled WWRs for the sides (left and right) are very small, ranging from 0.01-0.12. The front and back WWRs are slightly larger, ranging from 0.08-0.35. This could indicate why front WWR was still included. Orientation was removed perhaps because the WWRs were not large enough to make the direction they are facing a significant factor for energy use.

5.2.10 Sample Size Analysis

To ensure enough samples were being used for the training and testing of the model, a sample size analysis was conducted. Variously sized training and validation sets were randomly created, with the training set containing 10-90% of the samples from the full dataset, and the validation set containing the remaining samples. A model was fitted to each training set and evaluated

against the validation set using the R^2 score metric. The R^2 score used 10-fold cross-validation to ensure accurate results. The data was split into 10 folds, and each fold took a turn being the validation set, and the remaining nine would be the training set. This is done to ensure one split did not contain specific data that could skew the results, and that every single data point would eventually be in the validation set. Each fold was evaluated, and a mean R^2 was returned. Figure 57 shows the R^2 score for various sizes of training sets. The shaded area represents the standard deviation for the cross-validated scores.

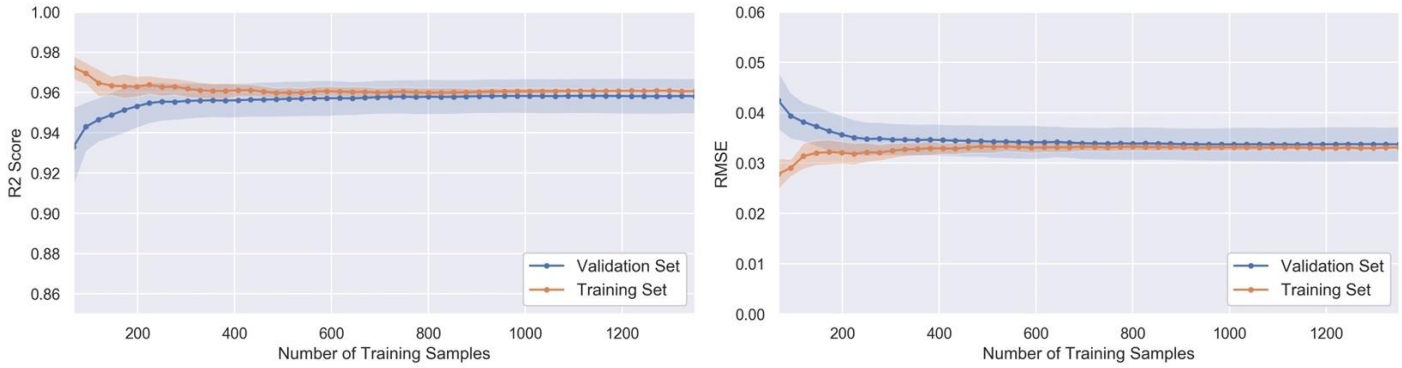

Figure 57 Sample size analysis for the dataset.

The training sets experience smaller standard deviations because it is the set that the model is being trained on. Therefore, when that same training set is used to predict energy use, the model outputs are very reliable. It is also a larger sample size. The R^2 and RMSE scores have almost converged at 1250 samples. Since the marginal increase in accuracy would be negligible (especially at the cost of additional simulation time), this was considered an appropriate sample size.

5.2.11 Case Study

Utility data was obtained for two houses in The Pocket neighbourhood. All inputs required for the model were taken from an EnerGuide energy audit or the field study. This included air tightness testing results, wall insulation values, window U-values, floor areas, and HVAC

systems and efficiencies. There was no information for window SHGC, so a value of 0.7 was assumed based on the window types. A value for lighting density of 2.505 W/m2 was assumed, and a cooling efficiency (for House 1 only) of 3 COP was assumed based on Jermyn's baseline model [10]. The ventilation was calculated using the volume and air tightness. The input values are shown in Table 24.

Table 24 Input values for the case study from the EnerGuide energy audit.

	Unit	House 1	House 2
Infiltration	ACH at 50 Pa	17.05	16.7
Wall RSI	m2K/W	1.3	2.08
Roof RSI	m2K/W	3	2.09
Slab RSI	m2K/W	0.28	0.28
Basement Wall RSI	m2K/W	0.48	1.41
Furnace Efficiency	-	0.8	0.93
Window U-value	W/m2K	0.50	0.41
HRV Option	-	0.24	0.24
Window SHGC	-	0.7	0.7
Ventilation		0.0045	0.0053
Depth	m	13.7	13.8
Width	m	5.1	6.1
Bmt Height Above Grade	m	1.2	1.3
Bmt Height Below Grade	m	1.1	0.8
Average First Floor Height	m	2.6	3.3
Third Floor Height	m	2.5	2.8
Front WWR	-	0.104	0.092
Back WWR	-	0.067	0.102
Left WWR	-	0.017	0.023
Right WWR	-	0.023	0.053
Lighting Density	W/m2	2.505	2.505
Cooling Efficiency	COP	3	3.8
Orientation	-	0.99	0.24

The results indicated that energy use could be predicted to within 10% error compared to utility data using the surrogate model. An EnergyPlus model of each house was simulated to compare to the surrogate model results. Table 25 summarizes the findings.

Table 25 Case study results.

	House 1	House 2
Actual Energy Use (Utilities) [GJ]	129.5	148
Surrogate Model [GJ]	121.53	133.98
Difference compared to Actual [%]	-6.15	-9.47
EnergyPlus Simulation [GJ]	124.4	130.42
Difference compared to Actual [%]	-3.94	-11.88

The House 2 predictions are close to the MAPE value of 6.1%. House 1 is slightly higher. The full EnergyPlus simulation was more accurate in House 1 and less accurate in House 2. This is reasonable as the percent error for the surrogate model ranged from -15 to 15% (excluding outliers). These are promising results; however more case studies would need to be completed to develop a confident estimation of accuracy.

5.2.12 NSGA-II Optimization

Figure 58 and Figure 59 show the Pareto front of solutions for this application. Out of the 65,536 possible combinations of retrofit levels, the optimization algorithm has selected 70 Pareto point solutions that it feels has the best energy-cost trade-offs. Each point corresponds to a set of levels of retrofit upgrades. It should be noted that each point is an individual solution, and they do not build off each other. The Pareto front seems discontinuous. Reviewing the solutions indicate that the discontinuity is caused by retrofit upgrades that drastically decrease energy use. Each "group" of the Pareto front all have the same wall insulation and ACH_{50} level.

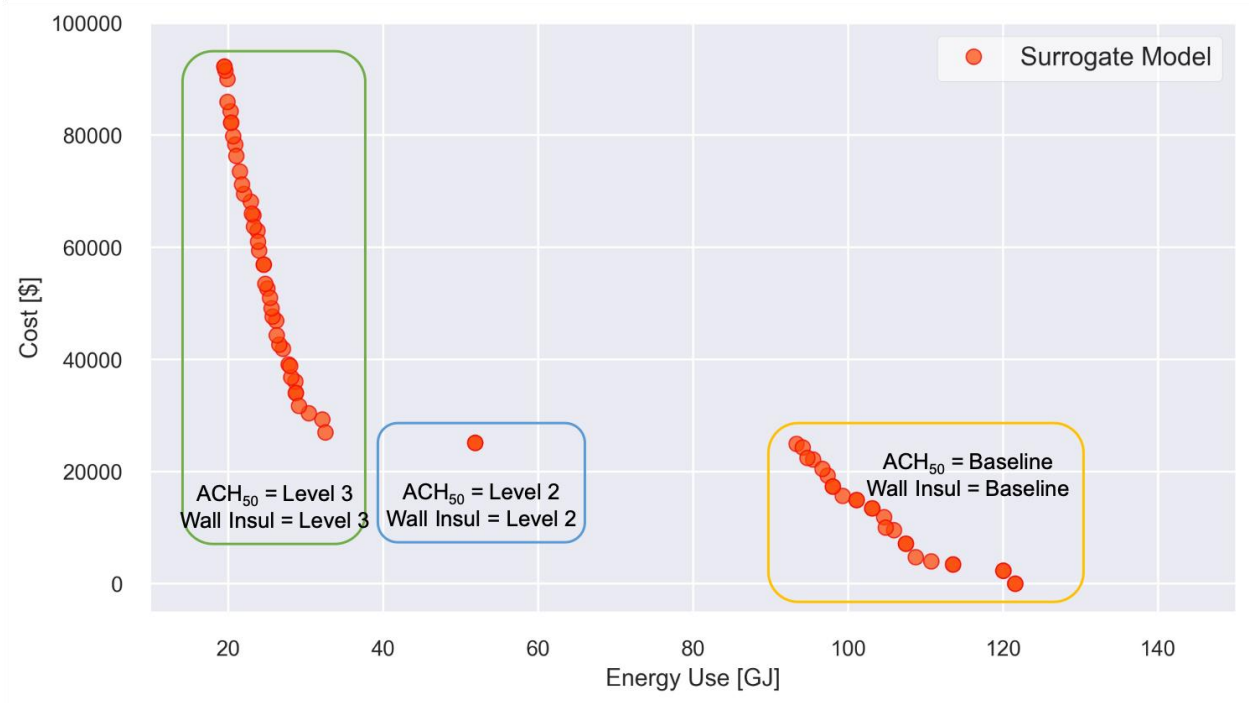

Figure 58 Pareto front solutions for house 1 from the case study.

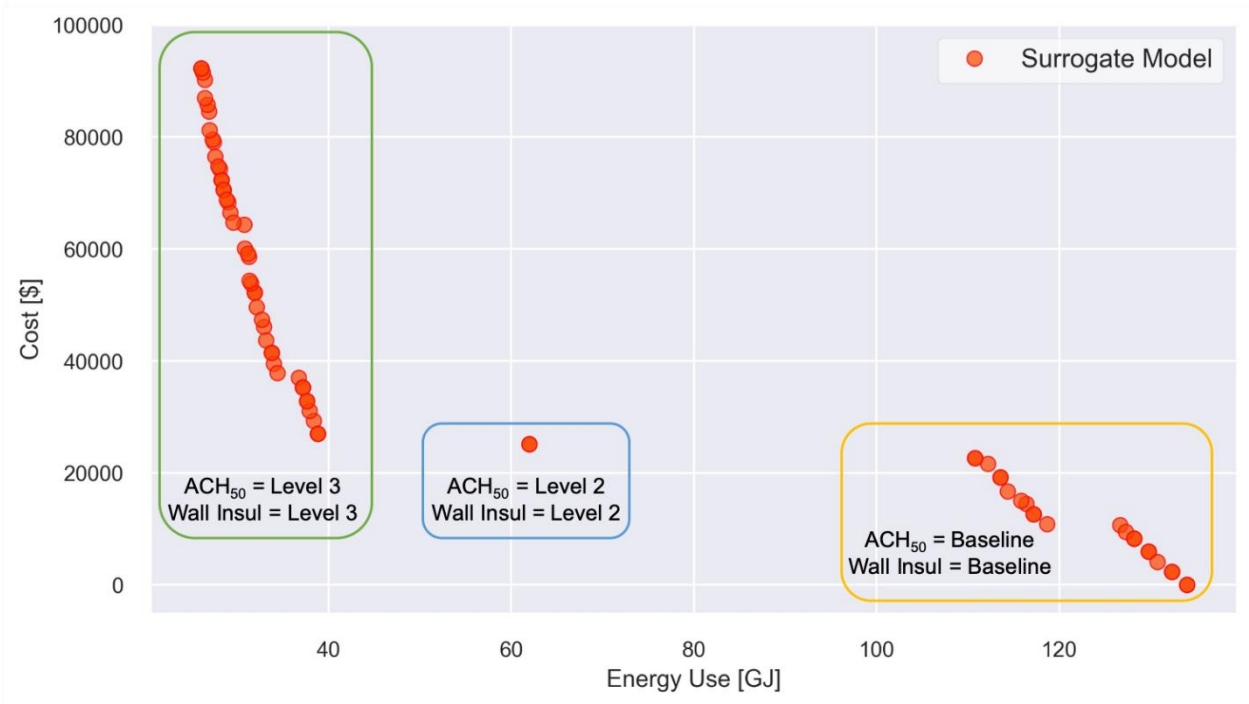

Figure 59 Pareto front solutions for house 2 from the case study.

Since there is a trade-off between cost and energy use, no single optimal solution can be determined. It is now up to the user to decide which point provides the most optimal solution. In

the case of an energy retrofit renovation, there might be certain limitations. If the goal is to maximize energy savings within a certain budget or get to a certain level of energy use, there would be different optimal solutions. The final decision would be made based on the situation, proposed application of the results, and external constraints.

Upgrading the wall insulation and ACH_{50} values reduce the energy more than any other retrofit parameter. The energy reduction seems slightly unrealistic. The ACH_{50} and wall insulation were the two most significant factors in the final surrogate model, and they could be overestimating the energy reduction potential. This example demonstrates the dangers associated with costing data. The level 1 upgrade cost is $23,284, but the level 3 upgrade is only $25,655. This is a minimal increase in cost for large energy savings. And the cost associated with ACH_{50} upgrades is comparatively low. This has to do with the way that Jermyn defined the costing and levels [10]. Likely, the cost could be higher and the energy savings could be lower.

The error on individual surrogate model predictions ranged from -15 to 15% (excluding outliers), although the average was 6.1%. This means these individual points could vary compared to the simulated values. To show how far off the optimization results using the surrogate model were, all 70 optimization solutions were simulated using a full model in EnergyPlus. The results are shown in Figure 60 and Figure 61.

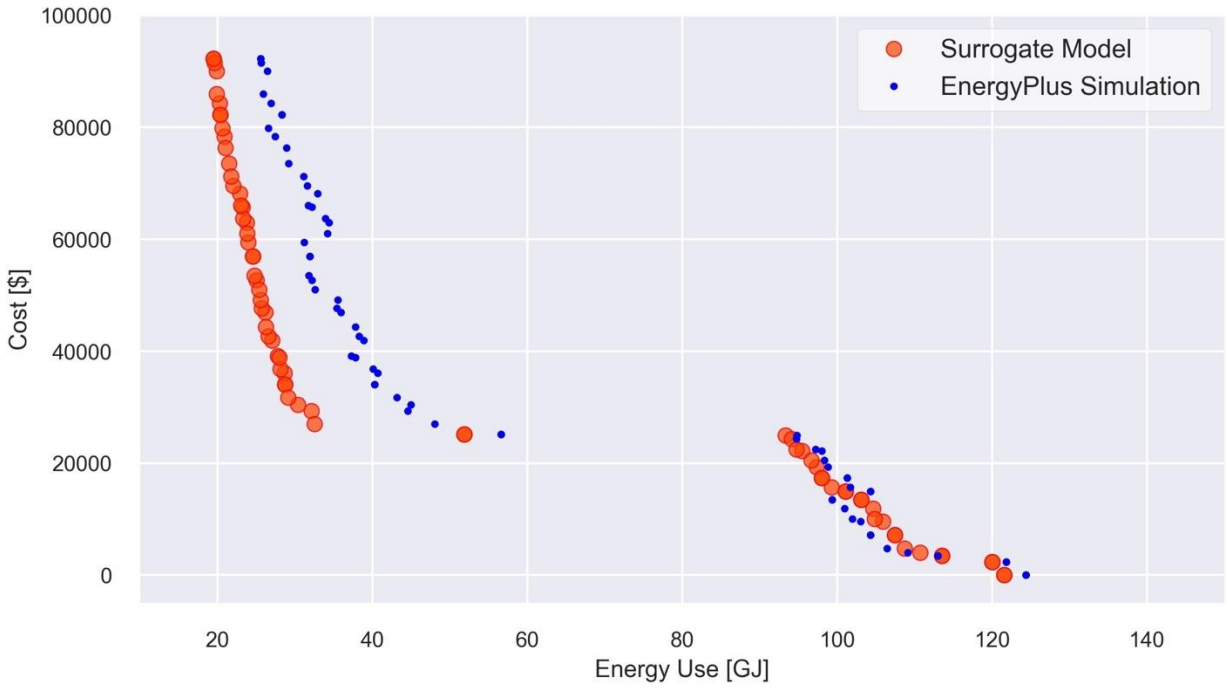

Figure 60 Results of EnergyPlus simulations versus the surrogate model for house 1 from the case study.

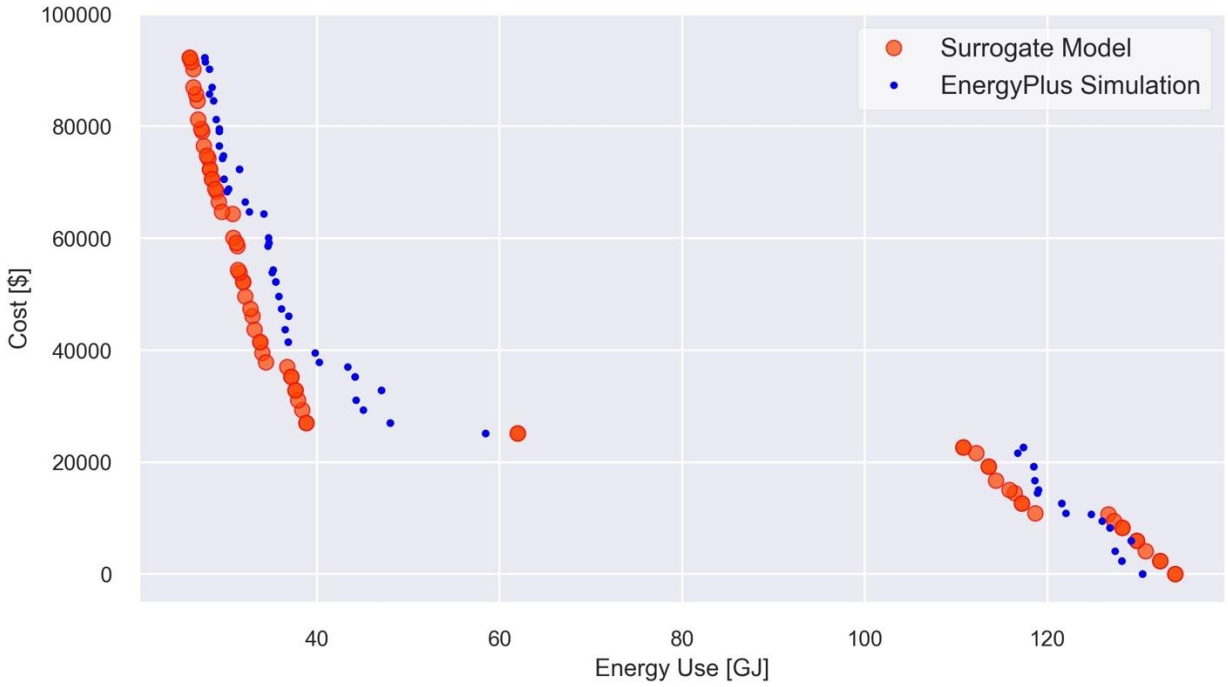

Figure 61 Results of EnergyPlus simulations versus the surrogate model for house 2 from the case study.

The EnergyPlus simulations are close to the predicted surrogate model values at higher energy uses. At lower energy uses, the surrogate model tended to underestimate energy use. If this were

done on multiple houses the error would likely be accurate on average and represent the MAPE of 6.1%. This demonstrates the power and capability of the surrogate model but illustrates the dangers of costing data and using individual houses as it could provide unrealistic retrofit costs and energy savings if the analysis is not done properly.

The purpose of this analysis was to show what an optimization could look like using a surrogate model, to demonstrate the potential issues, and to compare to a brute-force optimization. The purpose was not to conclude which retrofit solution is most cost-effective.

5.2.12.1 Brute-Force Optimization

The results of Jermyn's brute-force optimization [10] are shown in Figure 62.

EUI [kWh/m2]	Cost of Round	Air Sealing	Walls	Roof	Slab	Basement	Furnace	Windows	HRV/ERV
213	-								
188	$ 3,479						Level 1		
181	$ 4,953	Level 1					Level 1		
169	$ 9,747	Level 1				Level 1	Level 1		
144	$15,341	Level 1			Level 1	Level 1	Level 1		
108	$38,625	Level 1	Level 1		Level 1	Level 1	Level 1		
83	$39,184	Level 1	Level 2		Level 1	Level 1	Level 1		
74	$50,095	Level 1	Level 2	Level 1	Level 1	Level 1	Level 1		
57	$51,631	Level 2	Level 2	Level 2	Level 1	Level 1	Level 1		
55	$52,200	Level 2	Level 2	Level 2	Level 1	Level 1	Level 2		
54	$54,547	Level 2	Level 2	Level 2	Level 1	Level 1	Level 2		Level 1
46	$56,359	Level 2	Level 3	Level 2	Level 1	Level 1	Level 2		Level 1
38	$62,038	Level 3	Level 3	Level 3	Level 1	Level 1	Level 2		Level 1
31	$82,483	Level 3	Level 3	Level 3	Level 1	Level 1	Level 2	Level 1	Level 1
22	$84,789	Level 3	Level 3	Level 3	Level 1	Level 1	Level 2	Level 2	Level 1

Figure 62 Jermyn's brute-force optimization results. Figure adapted from [10].

The brute-force optimization was completed with a different baseline energy model and slightly different inputs than these case studies, however comparing the results show that the NSGA-II optimization algorithm is showing results with lower costs and lower energy uses. This shows that adding the most cost-effective options in order is overestimating the total costs required to

achieve lower energy uses. The NSGA-II optimization shows that upgrading the wall and ACH_{50} results in lower cost per energy use saved overall, although it was not necessarily the lowest cost per energy use saved compared to each possible retrofit upgrade as in the brute-force optimization. This shows that the surrogate model can find more optimal solutions than the brute-force optimization.

5.2.12.2 Optimization Comparison

For this optimization example, there were 65,536 possible solutions. Table 26 compares the time it would take to complete this optimization using mathematical optimization (simulating all possible solutions), brute-force optimization, and NSGA-II optimization.

Table 26 Time comparison with and without surrogate model for mathematical, brute-force, and NSGA-II optimization.

Without Surrogate Model		**With Surrogate Model**
Mathematical Optimization	**Brute-Force Optimization**	**NSGA-II Optimization**
45 days to run 65,536 simulations	3 hours to run 184 simulations	2 days to run 1500 simulations and develop surrogate model
Several hours of post-processing	Several days to perform manual brute-force optimization	Minutes to perform optimization
Set of optimal solutions determined	*Sub-optimal solution determined*	*Set of optimal solutions determined*

Another limitation of optimizing without a surrogate model is that any changes in the optimization problem would require all the simulations to be rerun. The surrogate model can be used in any application or repeated without requiring additional simulations. One of the limitations of using the surrogate model is the slight loss of accuracy. The final model had a MAPE of 6.1% with an error range of -15 to 15% (excluding outliers).

6 CONCLUSIONS

6.1 Future Work

Future work must be done to determine how many models are needed to describe the municipal, provincial, and national housing stock. Is there a point where the surrogate model framework does not apply? In different climates? Urban versus suburban areas? More varied surrogate models need to be developed in different parts of the city, with different house types, in different climates, to determine what accuracy is possible to describe a large subset of the housing stock, and at what point a single surrogate model does not apply.

Energy use was predicted within ~10% error of actual utility bill data for the case studies investigated. The case study only included two homes due to the availability of the EnerGuide energy audit results and the utility data. If more homes were included in the case study, a confident representation of the accuracy of the model compared to utility data could be understood.

The preliminary NSGA-II optimization demonstrated the capabilities and power of the surrogate model. However, it also indicated potential dangers associated with costing data and only using a single home. A more in-depth optimization study should be completed to expand upon the findings from this research. Future work should be done to optimize multiple houses to provide an accurate understanding of the trends for cost and energy reduction in terms of retrofit solutions for this archetypal home. Once enough surrogate models have been developed to describe Toronto's residential housing stock, the best retrofit strategies could be optimized. Adding life-cycle cost or occupant comfort as another objective to the optimization problem could be investigated.

The practicality and applicability of the model could be improved by beginning the research with the end goal in mind regarding how it would be used. The input from stakeholders (industry professionals or policy makers) could make the model more valuable in terms of practicality of use. Additionally, future work could include a user interface to allow the model to be easily used.

The performance and applicability of the surrogate model is heavily dependent on the dataset. The Latin hypercube sampling plan was chosen because it was most common in the reviewed literature. Further research should be completed on the effect of sampling plans on the accuracy of the model. The samples with high predicted error could be examined to see if there are any repeated patterns that could be remedied to reduce the error of the model.

One of the largest limitations of this surrogate model was fixing the occupant determined loads in the baseline model. This is not a limitation of the methodology itself, as these inputs can – and should – be added and the impact on the accuracy of the model and the practicality of using it to describe larger subsets of housing stocks should be investigated. It was not within the scope of this research.

A natural gas furnace was fixed in the baseline energy model. The surrogate models need to be able to include a retrofit option to upgrade existing furnaces to gas-free options such as air source heat pumps. This could potentially be included as a categorical variable; however further research must be completed to analyze if the model could achieve the same levels of accuracy.

Another limitation of the current model is that there is no separate ventilation system. It was modelled as such because that is the state of the existing baseline condition in the surveyed homes and adding a separate ventilation system as a retrofit requires a very large renovation with high costs. Since heavily retrofitted homes with low airtightness are being investigated, at some point a separate ventilation system would be necessary to ensure the required ventilation levels

are being met. Future work could be completed to determine if it is possible to have a model that can switch the HVAC system based on input parameter values.

6.2 Conclusions

An existing archetype for large detached century homes in Toronto, ON was updated using data from a field study of small century homes in The Pocket neighbourhood. The baseline model used to develop the dataset was updated and improved. Ranges for 23 parameters were determined using data from 35 homes which were measured in Jermyn's field study [10] and The Pocket field study. EnergyPlus was used to run 1500 simulations within the defined design space to create the synthetic dataset.

The final model used elastic net regression with *label encoded* categorical variables. It was the simplest in terms of number of coefficients (19.5 on average) and only sacrificed 0.2% MAPE compared to the full model with all the coefficients. Testing on a held-out dataset, the final model was able to achieve an R^2 value of 0.947 and a MAPE of 6.1%. The final model predicted energy use within 10% of annual utility bills for two houses in the field study. The first research question can be answered: **a surrogate model developed using multivariate linear regression can describe the annual energy use of an archetypal single-family home in Toronto, ON within 6% error on average.**

To answer the second research question, a house size analysis was completed to determine whether including a much smaller version of the century home archetype in the surrogate model would affect the performance. The results showed that including the wide range of house sizes sustained the model accuracy compared to the small or large houses on their own. The absolute value of the coefficients for each parameter were similar for all the home sizes, with air tightness and wall insulation RSI being the two most important parameters. The second research question

can be answered: **small and large archetypes with the same form *can* be described by a single surrogate model without losing accuracy**.

A sample size analysis of the small and large homes showed that more samples may have increased model performance slightly, as the validation and training errors had not yet converged at 400 samples. The increase would be minimal (R^2 of ~0.01) and would be at the cost of additional simulation time. The purpose of this analysis was to compare models and each model had the same amount of variation therefore a larger sample size would likely produce similar results. The sample size analysis of the entire dataset showed the model was much closer to converging (more samples would only change R2 by <0.001) therefore this was considered an appropriate sample size.

A preliminary optimization investigation was conducted using the NSGA-II optimization algorithm and Jermyn's collected data for retrofit upgrade values and associated costs [10]. This proved to be more effective than brute-force forward stepwise optimization performed by Jermyn [10]. The results of the optimization analysis indicated that increasing airtightness and the wall insulation RSI values would result in a large energy reduction. This option was not selected by the brute-force optimization because this upgrade has a large cost, therefore it was not selected as the most "cost-effective" option until later in the forward stepwise selection process. At that point the total cost became very expensive and many of the other retrofits that had been selected before did not necessarily provide a cost-effective solution at that point.

Using NSGA-II optimization, a design space with 65,536 potential combinations was narrowed down to 70 optimal solutions in less than a minute. To simulate all these solutions would have taken more than 45 days (assuming one minute per file, a conservative estimate). The 1500 EnergyPlus simulations for this research were run in approximately 1.5 days. Once the final

model was selected, the surrogate model took seconds to create. The optimization algorithm took a few hours to set up and one minute to run. This results in substantial time savings (potentially weeks), more flexibility, and the dataset that was created only needs to be simulated once. The developed model can be used repeatedly without additional simulations.

This thesis describes a bottom-up surrogate modelling approach to describe energy use in an archetypal house in Toronto, ON. The results indicate that further investigations must be conducted to determine the applicability of this framework in more varied situations. This research found that multivariate linear regression can be used to develop an accurate surrogate model which can incorporate a wide range of house sizes. There is great potential for this framework to contribute to eventually developing bottom-up surrogate models to describe the entire residential Canadian housing stock.

REFERENCES

[1] NRCan, “Energy Efficiency Trends in Canada 1990-2013,” 2013.

[2] Statistics Canada, “Census in Brief: Dwellings in Canada, Census year 2016,” 2017. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016005/98-200-x2016005-eng.cfm. [Accessed: 01-Dec-2019].

[3] NRCan, “Build Smart – Canada’s Buildings Strategy,” 2017.

[4] V. Masson-Delmotte *et al.*, “Global Warming of 1.5°C,” 2018.

[5] C. F. Reinhart and C. Cerezo Davila, “Urban building energy modeling - A review of a nascent field,” *Building and Environment*, vol. 97. pp. 196–202, 2016, doi: 10.1016/j.buildenv.2015.12.001.

[6] W. Tian, J. Song, Z. Li, and P. de Wilde, “Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis,” *Appl. Energy*, vol. 135, pp. 320–328, 2014, doi: 10.1016/j.apenergy.2014.08.110.

[7] J. S. Hygh, J. F. DeCarolis, D. B. Hill, and S. Ranji Ranjithan, “Multivariate regression as an energy assessment tool in early building design,” *Build. Environ.*, vol. 57, pp. 165–175, 2012, doi: 10.1016/j.buildenv.2012.04.021.

[8] P. Westermann and R. Evins, “Surrogate modelling for sustainable building design-A review,” in *Energy & Buildings*, 2019, vol. 198, pp. 170–186, doi: 10.1016/j.enbuild.2019.05.057.

[9] J. Hester, J. Gregory, and R. Kirchain, “Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption,” *Energy Build.*, vol. 134, pp. 202–211, 2017, doi: 10.1016/j.enbuild.2016.10.047.

[10] D. Jermyn and R. Richman, “A process for developing deep energy retrofit strategies for single-family housing typologies: Three Toronto case studies,” *Energy Build.*, vol. 116, pp. 522–534, 2016, doi: 10.1016/j.enbuild.2016.01.022.

[11] K. M. Blaszak, R. Richman, and P. Eng, “Prioritizing Method for Retrofitting Toronto’s Single-Family Housing Stock to Reduce Heating and Cooling Loads,” 2013, doi: 10.1061/(ASCE)AE.1943-5568.0000102.

[12] H. E. Zirnhelt and R. C. Richman, “The potential energy savings from residential passive solar design in Canada,” *Energy Build.*, vol. 103, pp. 224–237, 2015, doi: 10.1016/j.enbuild.2015.06.051.

[13] A. Mucciarone, “Towards a Proposed Framework for Analyzing Sustainable Renovation Building Envelope Assemblies,” Ryerson University, 2011.

[14] S. Niger, “HIGH PERFORMANCE RETROFIT OPPORTUNITIES OF TORONTO ’ S 1970S RESIDENTIAL DETACHED AND SEMI-DETACHED HOUSES,” Ryerson University, 2016.

[15] “About the Pocket - thepocket.ca.” [Online]. Available: https://www.thepocket.ca/who-we-are/about-the-pocket/. [Accessed: 18-Dec-2019].

[16] P. Santosh, “eppy: Scripting language for E+ idf files, and E+ output files.” .

[17] A. Tsanas and A. Xifara, “Accurate quantitative estimation of energy performance of

residential buildings using statistical machine learning tools," *Energy Build.*, vol. 49, pp. 560–567, 2012, doi: 10.1016/j.enbuild.2012.03.003.

[18] J. A. Wass, *Statistics in a Nutshell*, vol. 26, no. 1. O'Reilly Media, 2009.

[19] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," *Appl. Energy*, vol. 140, pp. 385–394, 2015, doi: 10.1016/j.apenergy.2014.12.011.

[20] A. Bager, M. Roman, M. Algedih, and B. Mohammed, "Addressing Multicollinearity in Regression Models: a Ridge Regression Application," *J. Soc. Econ. Stat.*, vol. 6, no. 1, pp. 30–45, 2017.

[21] N. Kock and G. S. Lynn, "Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations," *J. Assoc. Inf. Syst.*, vol. 13, no. 7, pp. 546–580, 2012, doi: 10.17705/1jais.00302.

[22] T. Hastie, R. Tibshirani, G. James, and D. Witten, *An Introduction to Statistical Learning with Applications in R*, vol. 102. 2006.

[23] W. Tian, R. Choudhary, G. Augenbroe, and S. H. Lee, "Importance analysis and meta-model construction with correlated variables in evaluation of thermal performance of campus buildings," *Build. Environ.*, vol. 92, pp. 61–74, 2015, doi: 10.1016/j.buildenv.2015.04.021.

[24] R. E. Edwards, J. New, L. E. Parker, B. Cui, and J. Dong, "Constructing large scale surrogate models from big data and artificial intelligence," *Appl. Energy*, vol. 202, pp.

685–699, 2017, doi: 10.1016/j.apenergy.2017.05.155.

[25] G. Schwarz, “Estimating the Dimension of a Model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.

[26] H. Akaike, “Information Theory And An Extension Of The Maximum Likelihood Principle,” *Springer Ser. Stat.*, 1998.

[27] H. Bozdogan and D. M. A. Haughton, “Informational complexity criteria for regression models,” *Comput. Stat. Data Anal.*, vol. 28, no. 1, pp. 51–76, 1998, doi: 10.1016/S0167-9473(98)00025-5.

[28] H. X. Zhao and F. Magoulès, “Feature selection for predicting building energy consumption based on statistical learning method,” *J. Algorithms Comput. Technol.*, vol. 6, no. 1, pp. 59–77, 2012, doi: 10.1260/1748-3018.6.1.59.

[29] Q. Li and N. Lin, “The Bayesian Elastic Net,” *Bayesian Anal.*, vol. 5, no. 1, pp. 151–170, 2010, doi: 10.1214/10-BA506.

[30] R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[31] M. Kuhn and K. Johnson, *Applied predictive modeling*. 2013.

[32] K. Deb, “Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction,” 2011.

[33] L. G. Swan and V. I. Ugursal, “Modeling of end-use energy consumption in the residential sector: A review of modeling techniques,” *Renew. Sustain. Energy Rev.*, vol. 13, no. 8, pp.

1819–1835, 2009, doi: 10.1016/j.rser.2008.09.033.

[34] N. Fumo, “A review on the basics of building energy estimation,” 2013, doi: 10.1016/j.rser.2013.11.040.

[35] A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay, “State of the art in building modelling and energy performances prediction: A review,” *Renew. Sustain. Energy Rev.*, vol. 23, pp. 272–288, 2013, doi: 10.1016/j.rser.2013.03.004.

[36] “Energy Step Code 2017 Metrics Research - Full Report.” [Online]. Available: https://www.bchousing.org/research-centre/library/residential-design-construction/energy-step-code-2017-full-report&sortType=sortByDate. [Accessed: 09-Dec-2019].

[37] L. G. Swan, V. I. Ugursal, and I. Beausoleil-Morrison, “Hybrid residential end-use energy and greenhouse gas emissions model-development and verification for Canada,” 2011, doi: 10.1080/19401493.2011.594906.

[38] S. Blais, M. Sc, A. Parekh, P. Eng, M. A. Sc, and L. Roux, “ENERGUIDE FOR HOUSES DATABASE-AN INNOVATIVE APPROACH TO TRACK RESIDENTIAL ENERGY EVALUATIONS AND MEASURE BENEFITS.”

[39] A. D. Wills, “On the Modelling and Analysis of Converting Existing Canadian Residential Communities to Net-Zero Energy,” p. 447, 2018.

[40] E. Barnes, “BUILDING ENERGY SURROGATE MODELLING – A FEATURE SELECTION METHODOLOGY,” Ryerson Univeristy, 2019.

[41] L. Magnier and F. Haghighat, “Multiobjective optimization of building design using TRNSYS simulations, genetic algorithm, and Artificial Neural Network,” *Build. Environ.*,

vol. 45, pp. 739–746, doi: 10.1016/j.buildenv.2009.08.016.

[42] T. Østergård, R. L. Jensen, and S. E. Maagaard, “A comparison of six metamodeling techniques applied to building performance simulations,” *Appl. Energy*, vol. 211, pp. 89–103, Feb. 2018, doi: 10.1016/j.apenergy.2017.10.102.

[43] C. Cerezo, J. Sokol, S. AlKhaled, C. Reinhart, A. Al-Mumin, and A. Hajiah, “Comparison of four building archetype characterization methods in urban building energy modeling (UBEM): A residential case study in Kuwait City,” *Energy Build.*, vol. 154, pp. 321–334, 2017, doi: 10.1016/j.enbuild.2017.08.029.

[44] S. E. Chidiac, E. J. C. Catania, E. Morofsky, and S. Foo, “A screening methodology for implementing cost effective energy retrofit measures in Canadian office buildings,” *Energy Build.*, vol. 43, no. 2–3, pp. 614–620, 2011, doi: 10.1016/j.enbuild.2010.11.002.

[45] T. Catalina, J. Virgone, and E. Blanco, “Development and validation of regression models to predict monthly heating demand for residential buildings,” *Energy Build.*, vol. 40, no. 10, pp. 1825–1832, 2008, doi: 10.1016/j.enbuild.2008.04.001.

[46] S. Asadi, S. S. Amiri, and M. Mottahedi, “On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design,” *Energy Build.*, vol. 85, pp. 246–255, 2014, doi: 10.1016/j.enbuild.2014.07.096.

[47] T. Catalina, V. Iordache, and B. Caracaleanu, “Multiple regression model for fast prediction of the heating energy demand,” *Energy Build.*, vol. 57, pp. 302–312, 2013, doi: 10.1016/j.enbuild.2012.11.010.

[48] A. P. Melo, R. S. Versage, G. Sawaya, and R. Lamberts, “A novel surrogate model to

support building energy labelling system: A new approach to assess cooling energy demand in commercial buildings," *Energy Build.*, vol. 131, pp. 233–247, 2016, doi: 10.1016/j.enbuild.2016.09.033.

[49] S. A. R. Sangireddy, A. Bhatia, and V. Garg, "Development of a surrogate model by extracting top characteristic feature vectors for building energy prediction," *J. Build. Eng.*, vol. 23, pp. 38–52, 2019, doi: 10.1016/j.jobe.2018.12.018.

[50] S. Sekhar Roy, R. Roy, and V. E. Balas, "Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM," *Renewable and Sustainable Energy Reviews*, vol. 82. Elsevier Ltd, pp. 4256–4268, 01-Feb-2018, doi: 10.1016/j.rser.2017.05.249.

[51] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments," *Stat. Sci.*, vol. 4, no. 4, pp. 409–423, 1989, doi: 10.1214/ss/1177012413.

[52] S. Shan and G. G. Wang, "Metamodeling for high dimensional simulation-based design problems," *J. Mech. Des. Trans. ASME*, vol. 132, no. 5, pp. 0510091–05100911, May 2010, doi: 10.1115/1.4001597.

[53] M. Castelli, L. Trujillo, L. Vanneschi, and A. Popovič, "Prediction of energy performance of residential buildings: A genetic programming approach," *Energy Build.*, vol. 102, pp. 67–74, 2015, doi: 10.1016/j.enbuild.2015.05.013.

[54] J. S. Chou and D. K. Bui, "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy Build.*, vol. 82, pp. 437–446, 2014, doi: 10.1016/j.enbuild.2014.07.036.

[55] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, “Cross-validation pitfalls when selecting and assessing regression and classification models,” *J. Cheminform.*, vol. 6, no. 1, 2014, doi: 10.1186/1758-2946-6-10.

[56] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, Feb. 2006, doi: 10.1186/1471-2105-7-91.

[57] U.S. Department of Energy, “Residential Prototype Building Models,” *U.S. Department of Energy*, 2017. [Online]. Available: https://www.energycodes.gov/development/residential/iecc_models. [Accessed: 03-Dec-2019].

[58] “EnergyPlus | EnergyPlus.” [Online]. Available: https://energyplus.net/. [Accessed: 17-Dec-2019].

[59] ASHRAE, *Standard 62.2-2019 -- Ventilation and Acceptable Indoor Air Quality in Residential Buildings*. 2019.

[60] N. Kruis, “Development and Application of a Numerical Framework for Improving Building Foundation Heat Transfer Calculations,” 2015.

[61] “Google Earth.” [Online]. Available: https://www.google.com/earth/. [Accessed: 28-Dec-2019].

[62] “MATLAB - MathWorks - MATLAB & Simulink.” [Online]. Available: https://www.mathworks.com/products/matlab.html. [Accessed: 16-Dec-2019].

[63] “Welcome to Python.org.” [Online]. Available: https://www.python.org/. [Accessed: 16-Dec-2019].

[64] “Weather Data | EnergyPlus.” [Online]. Available: https://energyplus.net/weather. [Accessed: 20-Dec-2019].

[65] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” in *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2825–2830.

[66] “EnerGuide energy efficiency home evaluations | Natural Resources Canada.” [Online]. Available: https://www.nrcan.gc.ca/energy-efficiency/energuide-canada/energuide-energy-efficiency-home-evaluations/20552. [Accessed: 26-Dec-2019].

[67] Statistics Canada, “Households and the Environment: Energy Use: Table 5 — Number and type of light bulb use, by province, 2011,” 2011.

[68] Natural Resources Canada, “Air Conditioning Your Home.” [Online]. Available: https://www.nrcan.gc.ca/energy/publications/efficiency/residential/air-conditioning/6051. [Accessed: 14-Dec-2019].

[69] Passive House Institute US, “Certification Guidebook,” 2019.

[70] “ARCHIVED - ecoENERGY Retrofit – Homes Program | Natural Resources Canada.” [Online]. Available: https://www.nrcan.gc.ca/energy-efficiency/energy-efficiency-homes/what-energy-efficient-home/ecoenergy-retrofit-homes-program/5003. [Accessed: 26-Dec-2019].

[71] Natural Resources Canada, “Heating with Gas,” 2012.

[72] “Atlas of the City of Toronto and suburbs : founded on registered plans and special

surveys showing plan numbers, lots & buildings (volume I) : Digital Archive : Toronto Public Library." [Online]. Available: https://www.torontopubliclibrary.ca/detail.jsp?Entt=RDMDC-912_71354GOA_V1&R=DC-912_71354GOA_V1. [Accessed: 18-Dec-2019].

[73] "Google Maps." [Online]. Available: https://www.google.com/maps/@43.6647973,-79.346555,15z. [Accessed: 18-Dec-2019].

[74] "OpenStudio | OpenStudio." [Online]. Available: https://www.openstudio.net/. [Accessed: 18-Dec-2019].

[75] "3D Design Software | 3D Modeling on the Web | SketchUp." [Online]. Available: https://www.sketchup.com/. [Accessed: 18-Dec-2019].

[76] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. 2000.