BUILDING ENERGY SURROGATE MODELLING – A FEATURE SELECTION METHODOLOGY

by

Erica Catherine Barnes

Civil Engineering (B.Eng.), McMaster University, 2011

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, 2019

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 – A FEATURE SELECTION METHODOLOGY

Master of Applied Science 2019, Erica Catherine Barnes

Building Science, Faculty of Engineering and Architectural Science, Ryerson University

Abstract

A mathematical regression model, referred to as a surrogate model as it was trained on a set of computer-simulated results, was developed to permit the rapid modelling of large commercial office buildings within a single climate zone. The model was developed using a large number of building features and their EnergyPlus simulated results. In previous building energy surrogate modelling, a research gap in selecting building features using statistical approaches was identified. This thesis investigates a feature selection method, including forward stepwise selection and least absolute shrinkage and selection operator (LASSO), to identify building features that, together, have the most significant impact on annual building energy use. The final model, with 23 features selected through this methodology, predicts annual building energy use at 11.3% error, on average.

Acknowledgements

Thank you to my supervisor Professor Jenn McArthur who provided me with the support and encouragement to take on research that pushed me outside of my previous knowledge and industry experience. Thank you also to my thesis committee member, Dr. Seth Dworkin, who provided valuable insight into the application of this research to industry. And to my thesis committee member, Dr. Russell Richman, who reinforced my love for building science and supported me in spreading my interest in building energy surrogate modelling to his research group.

This research was supported through an Ontario Centre of Excellence VIP1 Grant in partnership with RDH Building Science Inc. Thank you to RDH for supporting me and this research through this grant and, particularly, Steve Kemp for making time to meet regularly and contributing to the development of the building features and ranges as a Toronto building energy modelling industry expert.

Lastly, thank you to Daniel Voshart – your continuous support as a sounding board, keen listener, and person of logic was, as always, appreciated.

Table of Contents

1 INTRODUCTION 1-1

1.1 RESEARCH OBJECTIVE 1-5
1.2 RESEARCH QUESTIONS 1-6
1.3 THESIS STRUCTURE 1-6

2 APPROACHES TO BUILDING ENERGY SURROGATE MODELLING – A SUMMARY OF PREVIOUS RESEARCH 2-8

2.1 SURROGATE MODEL DEVELOPMENT SUMMARY 2-8
2.2 SURROGATE MODEL INTENT 2-9
2.3 BUILDING ARCHETYPE AND SIMULATION SOFTWARE SELECTION 2-12
2.4 LOCATION AND CLIMATE SELECTION 2-15
2.5 BUILDING FEATURES AND RANGE SELECTION 2-17
2.6 SAMPLING PLAN SELECTION 2-19
2.7 TARGET VARIABLE SELECTION 2-21
2.8 SAMPLE SET SIZE, DATA SPLITTING AND CROSS VALIDATION 2-23
2.9 TRANSFORMATION OF INPUT FEATURES AND TARGET VARIABLE(S) 2-25
2.10 COMBINING FEATURES 2-26
2.11 NORMALIZATION 2-27
2.12 FEATURE SELECTION/ELIMINATION 2-28
2.13 TRAINING ALGORITHMS 2-30
2.14 MODEL PERFORMANCE ANALYSIS 2-33

3 BUILDING ENERGY SURROGATE MODELLING – A FEATURE SELECTION METHODOLOGY USING WRAPPER AND EMBEDDED TECHNIQUES 3-36

3.1 METHODOLOGY 3-37
3.1.1 Generating the Building Energy Dataset 3-39
3.1.2 Feature Selection and Trained Model Development 3-47
3.2 RESULTS 3-51
3.2.1 Energy Modeling Data 3-53
3.2.2 Multivariate Linear Regression with Original Feature Set 3-56
3.2.3 Transforming Input Features and Target Variable 3-58
3.2.4 Adding Combined Feature Terms Using Forward Stepwise Selection 3-64
3.2.5 Embedded Feature Selection Using LASSO and Elastic Net Regulators 3-67
3.2.6 Final Model Evaluation 3-77
3.3 MODEL VALIDATION USING DOWNTOWN TORONTO REFERENCE MODEL 3-81
3.4 IMPACT OF LIGHTING AND PLUG LOADS ON MODEL ACCURACY 3-82
3.5 SURROGATE MODEL FINDINGS 3-84

4 DISCUSSION AND CONCLUSIONS 4-85

4.1 KEY FINDINGS 4-86
4.2 FUTURE RESEARCH 4-88

APPENDIX A – SUMMARY OF SURROGATE MODELS PRESENTED IN STUDIES REFERENCED IN CHAPTER 2 91

APPENDIX B – CALCULATION OF COMBINED FEATURES 94

WORKS CITED 105

GLOSSARY 110

Table 1 – Summary of surrogate model development 3-38
Table 2 – Building features and ranges used for dataset development. Integer features are marked with "(I)" 3-42
Table 3 – Combined Features. Calculations are in Appendix B 3-50
Table 4 – Results of multivariate linear regression model with original feature set 3-57
Table 5 – Summary of input and target variable transformations. Transformations with the highest model performance are bolded and highlighted in blue and orange for target and input features, respectively. 3-59
Table 6 – Summary of input and target variable transformation model performance (selected transformations in bold italics) 3-63
Table 7 - Model performance before forward stepwise selection, using the original 71 features, and following forward stepwise selection 3-65
Table 8 – Combined features in order of highest to lowest Pearson correlation coefficient. Last column indicates whether the combined feature was kept in the model. 3-66
Table 9 – Features selected for final LASSO model in order of coefficient values. The order of feature importance for the Elastic Net models where 23 features were selected is presented. * indicates a combined feature 3-72
Table 10 – Features in order of absolute coefficient values following highly correlated combined feature removal. Combined features shown with * 3-77
Table 11 – Final model performance on the log transformed training, validation and test datasets 3-78
Table 12 – Features in order of absolute coefficient values for both the annual total site energy use surrogate model developed in [69] and the annual building energy use surrogate model (excluding lighting and plug energy use) presented in this chapter. Combined features shown with *. 3-83

Figure 1 – Illustrative representation of a surrogate model (orange dashed and dotted curve) fit to training data (blue dashed curve). Adapted from: Mueller, 2014 [9] 1-3
Figure 2 – Graphical representation of the LASSO (left) and Ridge (right) regulators. Adapted from: [12] 1-5
Figure 3 – Surrogate model development decision path. 2-9
Figure 4 – Division of building archetypes used in studies referenced in this chapter. Where the same dataset was used by several reserachers, such as the dataset developed by [35], the archetype used to create the dataset was counted only once. 2-13
Figure 5 – Breakdown of energy simulation software used by researchers referenced in this chapter. Where the same dataset was used by several reserachers, such as the dataset developed by [35], the software used to create the dataset was counted only once. 2-15
Figure 6 – Surrogate model weather file locations for research referenced in this chapter. Where multiple weather files were used in a single study, the points are highlighted in red, with each study represented by

a unique symbol. Locations shown in blue represent studies where a single weather file was used. Map created in Google My Maps [39] 2-16
Figure 7 – Percent of studies summarized in Appendix A using building input features 2-19
Figure 8 – Monte Carlo with uniform distribution of variables (left) – clustering shown in red circles. Latin Hypercube Sampling (right) 2-20
Figure 9 – Summary of surrogate model building energy target variables for studies summarized in this chapter. 2-22
Figure 10 – Breakdown of learning algorithms used for surrogate model development in studies referenced in Appendix A. 2-30
Figure 11 – From studies referenced in Appendix A, number of learning algorithms tested during surrogate model development. Right pie chart shows, for the studies where one algorithm was used, which algorithm was selected. 2-31
Figure 12 – Breakdown of model predictive performance metrics used in studies referenced in Appendix A. Most studies used more than one metric to evaluate model performance. 2-33
Figure 13 – Plot of squared error and absolute error for incremental residual values. Adapted from: [12] 2-34
Figure 14 – Workflow for generating building energy use dataset 3-39
Figure 15 – Wireframe model illustrating modified building geometry 3-40
Figure 16 – Illustration of above grade floor plans 3-41
Figure 17 – Histogram of boiler efficiency input variable for full dataset 3-45
Figure 18 – Input and target variable matrix/vector form 3-46
Figure 19 – Breakdown of annual building energy use target variable 3-47
Figure 20 – Surrogate model training process 3-51
Figure 21 – Histogram plot of annual energy use intensity for full dataset 3-53
Figure 22 – Breakdown of energy use intensity as a percentage of the annual site energy use intensity for the full dataset. Boxplot shows the data within the first to third interquartile ranges with the outliers not shown. 3-54
Figure 23 – Breakdown of annual energy use (GJ) as a percentage of the annual building energy use for the full dataset. Boxplot shows the data within the first to third interquartile ranges with the outliers not shown. 3-56
Figure 24 – Surrogate model training process – Multivariate linear regression 3-56
Figure 25 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) annual building energy use for the validation dataset using the multivariate linear regression model 3-57
Figure 26 – Annual building energy use residual plot for using multivariate linear regression model showing non-linear behaviour 3-58
Figure 27 – Surrogate model training process – input feature and target variable transformations 3-58
Figure 28 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) quadratic transformed annual building energy use for the validation dataset 3-60
Figure 29 – Quadratic transformed annual building energy use residual plot for multivariate regression model 3-60
Figure 30 – Histograms of original, exponential transformed, and Box-Cox transformed target variables 3-61
Figure 31 – Box-Cox transformed annual building energy use residual plot for multivariate regression model 3-62

Figure 32 – Exponential (log) transformed annual building energy use residual plot for multivariate regression model 3-62
Figure 33 – Surrogate model training process – combined feature forward stepwise selection 3-64
Figure 34 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) for the multivariate regression model following forward stepwise selection. Target is log transformed annual building energy use and data presented is from the validation dataset..... 3-67
Figure 35 – Exponential (Log) transformed annual building energy use residual plot for multivariate regression model following forward stepwise selection..... 3-67
Figure 36 – Surrogate model training process – LASSO and Elastic Net feature selection 3-68
Figure 37 – Model coefficient of determination as the number of features in each the LASSO and Elastic Net models change. The selected number of features is shown by the dashed line. 3-69
Figure 38 – Features removed from model (highlighted in red) as penalty weight, λ, increases..... 3-70
Figure 39 – Change in coefficient values for a selection of input features as the shrinkage parameter (λ) was changed for LASSO 3-73
Figure 40 – 'Heatmap' showing the Pearson correlation coefficient of the combined features. The blue lines on the y-axis indicate the features removed in order to eliminate highly collinear input features. 3-74
Figure 41 – Model coefficient of determination as the number of features in each the LASSO with correlated features (green circle) and LASSO with correlated combined features removed (purple triangle) models change. The selected number of features for both methods is shown by the dashed line. 3-75
Figure 42 – Surrogate model training progress – final model evaluation 3-77
Figure 43 - Predicted (surrogate model) vs. actual (EnergyPlus simulated) log transformed annual building energy use for test dataset..... 3-79
Figure 44 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) annual building energy use in GJ for test dataset..... 3-79
Figure 45 – Residual plot for predicted (surrogate model) annual building energy use in GJ for test dataset 3-79
Figure 46 – Predicted annual building energy use in GJ vs. percent error for test dataset..... 3-80
Figure 47 – Distribution of the test dataset percent error. Boxplot shows the data within the first to third interquartile ranges and the outliers..... 3-80
Figure 48 – Impact of training sample size on the mean absolute percent error of the test dataset 3-80

Appendix A – Summary of Surrogate Models Presented in Studies Referenced in Chapter 2.....91
Appendix B – Calculation of Combined Features.....94

1 Introduction

It is widely acknowledged that to achieve the United Nations Paris Agreement on Climate Change [1] global targets, CO_2 emissions related to buildings must be dramatically reduced. Fortunately, this sector has the largest potential to cost-effectively reduce energy use over the long-term [2]. As building owners, industry, municipalities and countries push towards reducing building energy consumption to meet aggressive goals, validated energy-use metrics will be important to inform decisions at each stage of the building design process.

Buildings currently contribute significantly to the total greenhouse gas emissions in Toronto, Ontario, Canada including, but not limited to, on-site energy production, electricity generation, and embodied energy in materials used for construction and retrofits. A 2016 study completed by the City of Toronto found that 45% of Toronto's greenhouse gas emissions were from buildings, with 88% of the emissions from natural gas and 12% from electricity [3]. In Canada, commercial buildings alone account for 14% of end-use energy consumption and 13% of carbon emissions [4]. In Toronto's downtown core, an area designated as a 2030 District, of the 10,030 eGWh/year building energy use, 23% is associated with office buildings [5]. There is a significant opportunity to reduce Toronto's new and existing buildings' energy use, particularly commercial office buildings.

Building energy simulation (BES) software is used in industry to generate reference models and simulate new and existing buildings. The model results are often used by building design teams to inform energy conservation measures and verify code and standard compliance. Using BES software in the earliest stages of building design can be time consuming and costly, and often of limited value due to the unavailability of detailed design information [6]. Simple and fast energy

use prediction tools are needed to evaluate and compare building design scenarios at the earliest design stages [7].

Supervised machine learning methods can be used to predict building energy use based on generated datasets of building attributes and energy use. When computer-simulated values are the variables an algorithm is being trained to predict, the supervised machine learned model is often referred to as a surrogate, meta, response surface or emulator model [8]. The goal of *surrogate modelling*, as defined by Forrester et al. [8], is to fit computer-simulated data to a surface in order to predict results from available data without the use of expensive code, permitting faster computation across the input domain. Using this definition, supervised machine learning algorithms are trained to fit input building features, as close as possible, to a *surface* defined by the building energy simulation target variable. The fit *surface* then becomes the *surrogate* and is used to predict the target variable for a set of input features within the trained *design space* (i.e. within the input feature ranges included in the dataset). The degree of error in the fit *surface* is evaluated on both the dataset used to fit the *curve* (training dataset) and on separate data not used in the model training (validation and test datasets). The surrogate model training and validation/test data behaviour is shown visually in Figure 1.

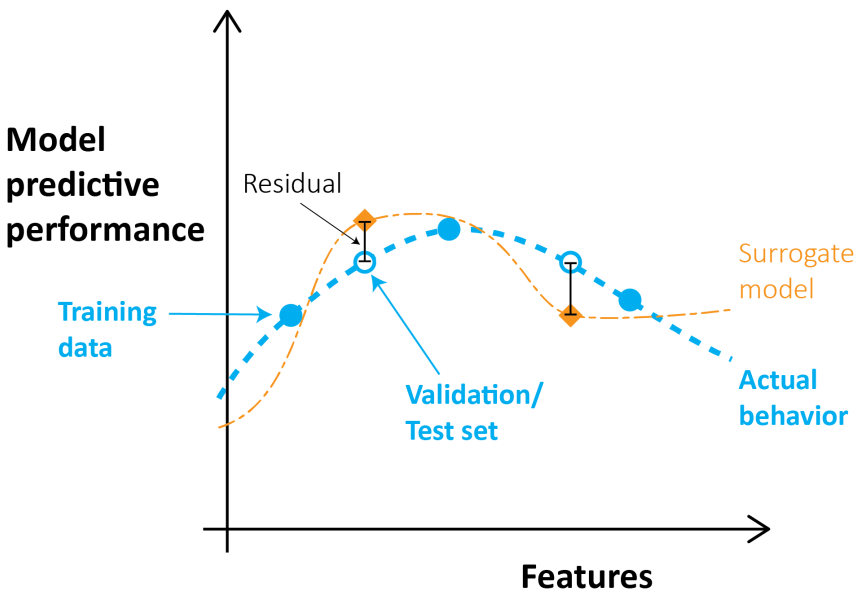

Figure 1 – Illustrative representation of a surrogate model (orange dashed and dotted curve) fit to training data (blue dashed curve). Adapted from: Mueller, 2014 [9]

A notable gap in the literature is the selection of key input features for such surrogate models. Generally, feature selection has been based on expert knowledge. Zhao and Magoulés [10] stated that few previous studies have used learned models to select key building features.

There are three categories of feature selection; filter, wrapper and embedded. Filter feature selection methods are independent of the learning algorithm and often 'score' each feature to the target variable. This score is used to determine which features are kept for model training. Filter methods often do not account for the relationship between features. Wrapper methods use subsets of the features to fit the model and compare model behaviour and performance for each subset. Common wrapper methods include forward stepwise selection where features are iteratively added to the model and backwards stepwise elimination where features are iteratively removed. Embedded methods use properties of specific learning algorithms to select features that best contribute to the model accuracy.

One method of embedded feature selection is the least absolute shrinkage and selection operator (LASSO), also referred to as L1 regularization. L1 regularization, LASSO, and L2 regularization, referred to as Ridge, apply a penalty term to the mean squared error cost function that shrinks the regression coefficients. When the two regularization terms are applied to the cost function, it is referred to as Elastic Net. LASSO uses the L1 norm penalty term, $\lambda \sum_{j=1}^{n} \beta_{(j)}^2$ and the Ridge uses the L2 norm penalty term, $\lambda \sum_{j=1}^{n} \left|\beta_{(j)}\right|$ [11]. The cost function for LASSO is shown in Equation 1 and the Elastic Net cost function is shown in Equation 2.

$$C(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - h_\beta(x^{(i)}))^2 + \lambda \sum_{j=1}^{n} ||\beta_{(j)}|| \quad (1)$$

$$C(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h_\beta\left(x^{(i)}\right)\right)^2 + (\%Lasso)\lambda \sum_{j=1}^{n} \left|\left|\beta_{(j)}\right|\right| +$$

$$(1 - \%Lasso)\lambda \sum_{j=1}^{n} \beta_{(j)}{}^2 \quad (2)$$

The LASSO and Ridge regulators are represented graphically for two-variable regression in Figure 2. The ellipse contours represent the mean squared error term in the cost function, with the black dot representing the minima. The blue circle for Ridge and diamond for LASSO represent the penalty term for each and vary in dimensions based on the λ value. The β_1 and β_2 value with the penalty term become where the ellipse touches the circle or diamond. In LASSO, where the ellipse touches the diamond corner, the coefficient becomes zero (in Figure 2 β_1 is equal to zero) [12].

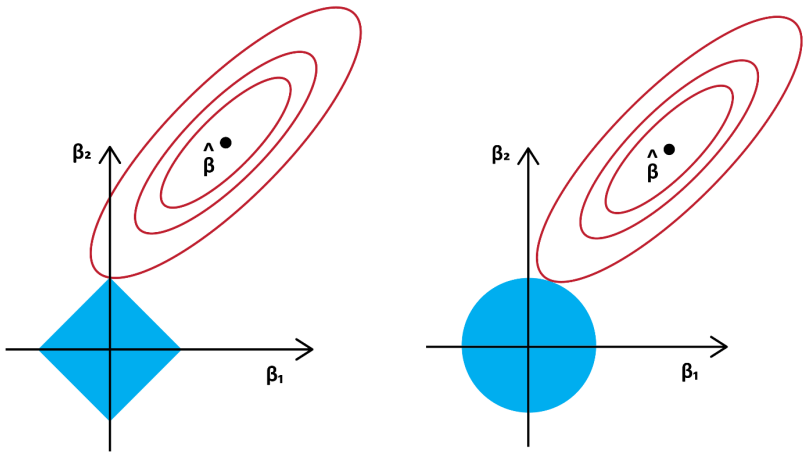

Figure 2 – Graphical representation of the LASSO (left) and Ridge (right) regulators. Adapted from: [12]

The shrinkage parameter (λ) is used to modify the shrinkage penalty applied to the cost function. Larger coefficient shrinkage is associated with a larger shrinkage parameter value which simplifies the hypothesis thereby reducing overfitting. When LASSO and Ridge are combined in Elastic Net regression, the feature selection benefit of LASSO is combined with the reduced over-fitting benefit of Ridge regression.

No previous research was found in which wrapper and embedded feature selection methods were used together to select features for building energy surrogate models. Such feature selection can be used to inform archetype energy model design, code and standard requirements, as well as to develop assumptions for detailed building energy models. Selecting the most relevant feature set for predicting the target variable using a learning algorithm can have advantages for the overall performance of the model. These include removing irrelevant features from the model, reducing model overfitting, and reducing model run time.

1.1 Research Objective

The objective of this research was to evaluate where the research on surrogate modelling for building energy use prediction currently stands and address gaps identified regarding feature selection. Using Toronto, Ontario large office buildings as an example to test a feature selection

methodology, this research used wrapper and embedded feature selection methods and evaluated the impact to the model behaviour and performance.

1.2 Research Questions

This thesis will look to answer the following research questions:

1. Can building energy simulation software (EnergyPlus) results using a set of varying model parameters be used to train machine-learned surrogate models that predict annual building energy use for Toronto, Ontario large office buildings within an accuracy that is acceptable to industry for early-stage design decisions?
2. Can wrapper (forward stepwise selection) and embedded (LASSO regression) feature selection methods be used to simplify the annual energy prediction models and improve model accuracy?
3. Does the surrogate model perform well for building model parameters representative of actual Toronto office buildings?

1.3 Thesis Structure

This thesis has been divided into four chapters. Chapter 2 is a summary of published literature in the building energy surrogate model field. It is a detailed comparison of previous building energy surrogate models, including the stages involved in dataset development, data preprocessing methods, learning algorithms tested and model evaluation metrics used. An analysis and discussion on gaps in current research shaped the approaches used for the surrogate model development described in Chapter 3 of this thesis.

Chapter 3 describes and discusses the development of a building energy surrogate model for large office buildings in Toronto, Ontario, with a focus on input feature selection using wrapper and embedded feature selection methods. The final surrogate model is validated for a combination of

realistic building features extracted from a reference energy model for a downtown Toronto office tower prepared by a local engineering consulting firm.

Chapter 4 provides a conclusion for the research completed and summarizes how future research could fill gaps in the field of building energy surrogate modelling.

2 Approaches to Building Energy Surrogate Modelling – A Summary of Previous Research

As building energy surrogate modelling research increases in popularity, the evaluation and comparison of surrogate model development methods is essential. The studies performed to date followed a similar overall methodology, but the researchers' decisions made at each step had an impact on the learning algorithm that performed best, the input feature selection and importance, and the overall accuracy of the model. This chapter contributes to the current discourse surrounding surrogate model development techniques for building energy prediction by summarizing these decisions and their impacts. Each step in the model development process and the most common methods used by researchers are presented. Next, the impact of methods to surrogate model behaviour are critically discussed and a summary of gaps in the current research are presented.

2.1 Surrogate Model Development Summary

The surrogate model development process for building energy use prediction used in previous studies followed a common path, illustrated in Figure 3. These studies typically began by presenting the surrogate model intent then selecting the building typology, climate and model prediction target variable(s) to guide the surrogate model development. Next, the building features and associated ranges, a simulation software and base model, a sampling plan, and the number of samples were selected. Finally, the simulations were run to develop the labelled dataset. This dataset was trained on one or more supervised machine learning algorithms and a final model selected based on model prediction metrics. Deliberate decisions on process and methodology were required at each step of the surrogate model development process. Researchers following the same decision path could compare the results and behaviour of their surrogate models. However,

as decisions differed between studies, the behaviour of the final surrogate model changed and could not be directly compared to another researcher's model behaviour and results. This was especially true for decisions made at the dataset development stage.

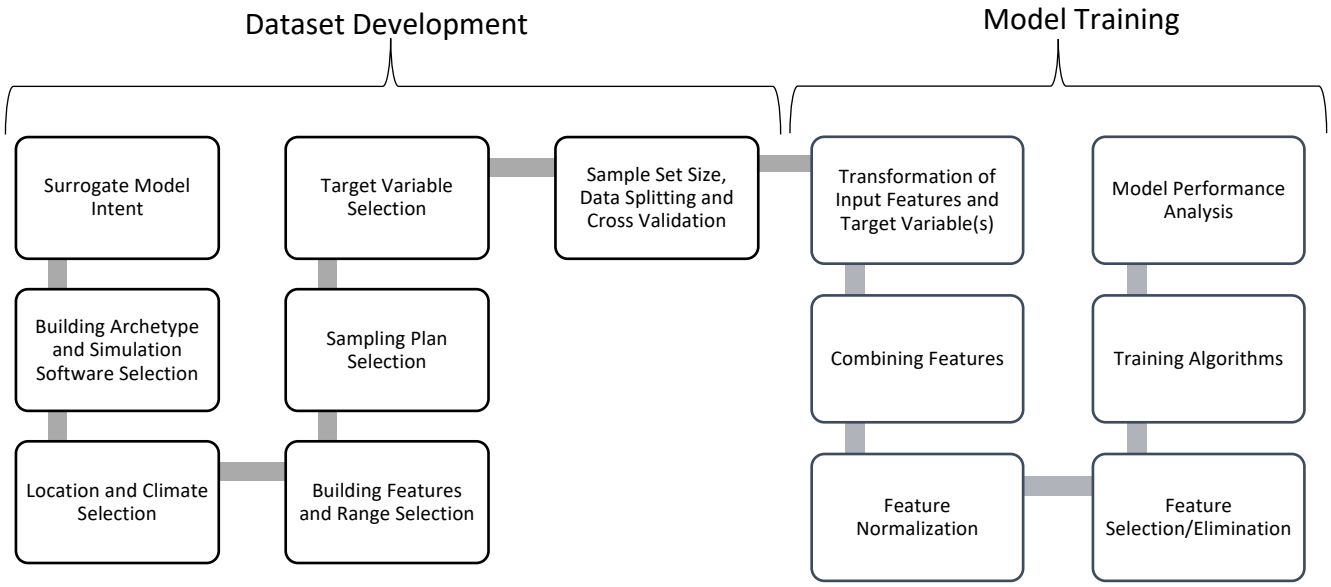

Figure 3 – Surrogate model development decision path.

2.2 Surrogate Model Intent

Many studies have discussed the advantages of using surrogate modelling as an early-stage design tool. Jacobs et al. [13] conducted a survey of design professionals and found that only a quarter of the respondents used computer-based methods, including energy simulation, to inform building form and orientation. Responding to this, Hygh et al. [14] were motivated to develop models that could be used by architects as a practical, early-stage design energy assessment tool for quick feedback on building attributes affecting early-stage design. Catalina et al. [15] presented building energy surrogate models as a middle ground between complex building energy simulation software that require detailed inputs, time and expertise, and simplified models that use one or a few variables to predict energy use, such as the degree day method [16].

Focused on existing buildings, other researchers proposed using surrogate models for retrofit energy conservation measure evaluation. Motivated by the International Energy Agency's Transition to Sustainable Buildings – Strategies and Opportunities to 2050 report [17] which stated that in 2050, 60% of the 2013 building stock will still exist in the United States, European Union and Russia, Tian et al. [18] proposed using surrogate modelling for existing building stock retrofit analysis. Chidiac et al. [19] used surrogate models to evaluate an existing building's potential for energy consumption reduction through energy retrofit measures.

Other researchers used surrogate modelling to focus on a single aspect of building design. Asl et al. [20], Asadi et al. [21] and Catalina et al. [15] focused on the impacts of building form to building energy use. With the goal of informing design decisions using building form's relation to energy use, Asl et al. [20] presented a tool that generated building forms based on site and building constraints and used surrogate models to predict the associated energy use. Asadi et al. [21] trained multiple multivariate linear regression models with differing building forms to evaluate how the feature significance differs between building forms. Catalina et al. [15] used varied building forms in their dataset and described the form with building shape factor (the ratio of the conditioned building volume to the building enclosure surface area).

Some researchers have explored using surrogate modelling for urban-scale building energy models where creating detailed energy simulations for many buildings within an area would not be feasible. Tian and Choudhary [22] used surrogate modelling to develop a model representative of secondary school buildings in the greater London area. They used this model, along with historical data on annual energy use by London's secondary schools, to evaluate the annual energy use impact of energy conservation measures implemented across schools. Mastrucci et al. [23] proposed using surrogate models to address the limitations of current archetype-based urban

building energy modelling in Europe, including building characteristics and occupancy variation adjustability. Nagpal et al. [24] proposed using surrogate models for large university campuses where exploration and prioritization of campus-wide building retrofits for energy conservation are a key priority. Instead of calibrating a detailed energy model simulation for each building on campus, which can take a significant amount of time and expertise, Nagpal et al. [24] proposed using a trained surrogate model to determine unknown building attributes when the measured building energy use is known.

As the field develops, researchers have explored creative and innovative ways to use surrogate building energy models. Nagpal et al. [25] showed how surrogate models specific to individual buildings can be updated as building retrofits are completed and therefore used as a 'living' energy model. Carlo and Lamberts [26] used a multivariate linear regression surrogate model to predict commercial building annual electricity use intensity for Brazil's voluntary commercial, public and service building energy labelling system. Melo et al. [27] built on Carlo and Lambert's [26] research by adding features to the input matrix and highlighted that in a developing country such as Brazil, economic growth comes with increased energy use; therefore a tool developed using surrogate models can help governments minimize building energy consumption as the country develops.

Instead of developing a surrogate model representative only of the building as a whole, Geyer and Singaravel [29] developed multiple building component-based surrogate models and linked them together. They proposed that by developing component-based models, there is more flexibility in applying the models to future scenarios and integrating them into building information modelling (BIM). Korolija et al. [30] trained surrogate models to predict annual energy input requirements for various heating, ventilation and air conditioning (HVAC) distribution systems with varying

simulation inputs related to building orientation, enclosure thermal performance, shading, and daylighting. Papadopouslos and Azar [31] integrated a building energy surrogate model with an agent-based modeling tool that modelled building occupants' dynamic energy use behaviours. Their goal was to provide more flexibility in predicting monthly energy use in relation to occupancy and building operations behaviour. Wong et al. [32] used surrogate modelling to predict daily electricity use in a subtropical climate using daily weather and building enclosure features related specifically to daylighting.

Using surrogate models within optimization algorithms has been of growing interest in this field. A surrogate model can increase the speed of an optimization routine where the energy use results of often thousands of combinations of building features must be determined in the search for the optimal combination(s) [8]. This chapter does not discuss optimization as it is its own field of research and the reader is instead directed to the review conducted by Nguyen et al. [33] summarizing research where building energy surrogate modelling was used with optimization.

2.3 Building Archetype and Simulation Software Selection

Researchers in this field have illustrated the usefulness of surrogate models for a variety of building archetypes. Figure 4 summarizes the surrogate model building archetypes used in the studies referenced in this chapter. A large office was defined as 12 or more storeys, medium offices as 3-11 storeys, and a small office as one to two storeys [34]. A mid/high-rise residential building was defined as four or more storeys. A low-rise residential building was defined as one to three storeys and included single-family homes and low-rise multi-family homes. The 'other' category included buildings such as warehouses, retail, and schools.

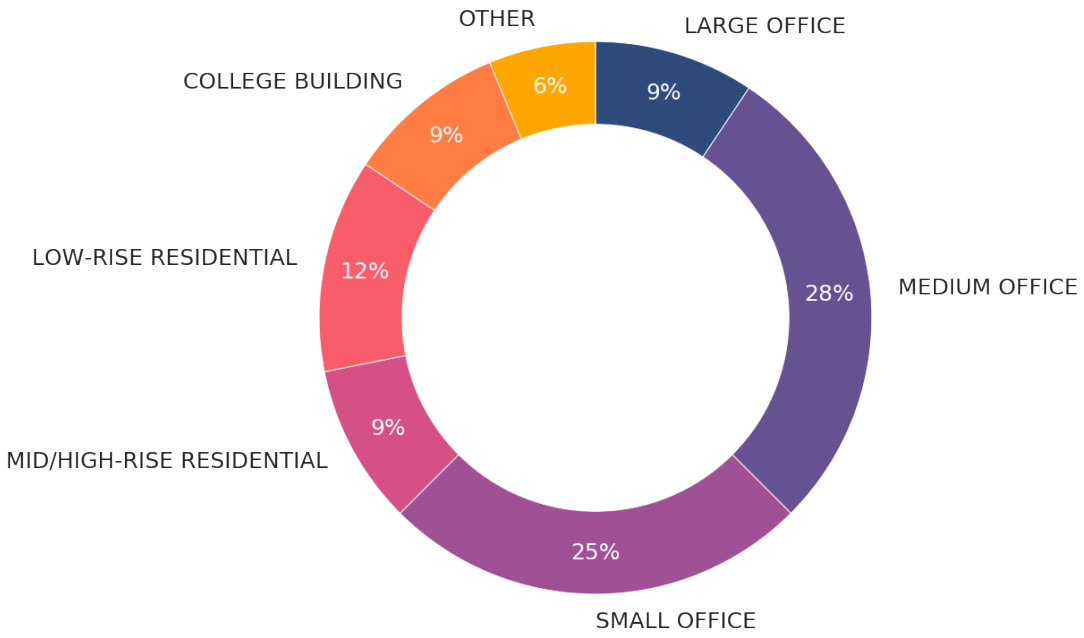

Figure 4 – Division of building archetypes used in studies referenced in this chapter. Where the same dataset was used by several reserachers, such as the dataset developed by [35], the archetype used to create the dataset was counted only once.

Researchers' motivations for selecting one archetype over another varied. Many researchers simply chose a building type as a case study to illustrate their proposed surrogate model development methodology [14, 19, 35, 36]. Other researchers targeted a specific building type to fulfill their surrogate model intent [18, 22, 24, 25]. Tian and Choudhary [22] discussed the need to develop building energy prediction models relevant to the urban scale for non-domestic buildings in the greater London, UK area, stating that data is widely available for the residential sector but not for the non-domestic sector. Tian et al. [18] used university/college campus buildings in the United States, as the real building attribute data for campus buildings was available to them.

Each researcher developed or selected a base model that was modified for the simulated samples. The base model defined the assumptions that were carried through all simulations. Many researchers used previously developed reference models and standards for the base energy simulation model. Aijazi and Glicksman [36] used the American Society of Heating, Refrigeration

and Air Conditioning Engineers (ASHRAE) Standards 90.1-2004 – Energy Standard for Buildings Except Low-Rise Residential Buildings and ASHRAE Standard 62.1-2004b – Ventilation for Acceptable Indoor Air Quality, for mid-rise apartment building enclosure and equipment model attributes. Hygh et al. [14] and Papadopoulos and Azar [31] used the U.S. Department of Energy (DOE) EnergyPlus Commercial Reference Models as the base energy simulation model. These models were developed in collaboration with Lawrence Berkeley National Laboratory (LBNL), Pacific Northwest National Laboratory (PNNL) and the National Renewable Energy Laboratory (NREL) and were created as baseline models for 15 building archetypes in three construction eras for all ASHRAE climate zones [34]. The commercial building model inputs were based on the U.S. Energy Information Administration 2012 Commercial Buildings Energy Consumption Survey (CBECS) data which represented 70% of the commercial buildings in the United States [37].

Researchers with access to operational building information used this data to develop base models for their energy simulations. Lam et al. [38] used a survey of existing commercial buildings in Hong Kong to determine the common building parameters for large office buildings and developed a base model from this. Nagpal et al. [24] developed base EnergyPlus models for operational college campus buildings and validated the base models by comparing the simulation results to the measured building energy use.

Several energy simulation software exist that will calculate the energy use based on user-defined inputs. The simulation software used to develop the dataset for the surrogate model varies by study. Figure 5 shows the relative proportions of the building energy simulation software used to develop datasets in the studies summarized in this chapter. The majority of researchers selected EnergyPlus

for its ability to streamline and automate the modification of the input data file (IDF) text file and because the simulation was based on first principles [14].

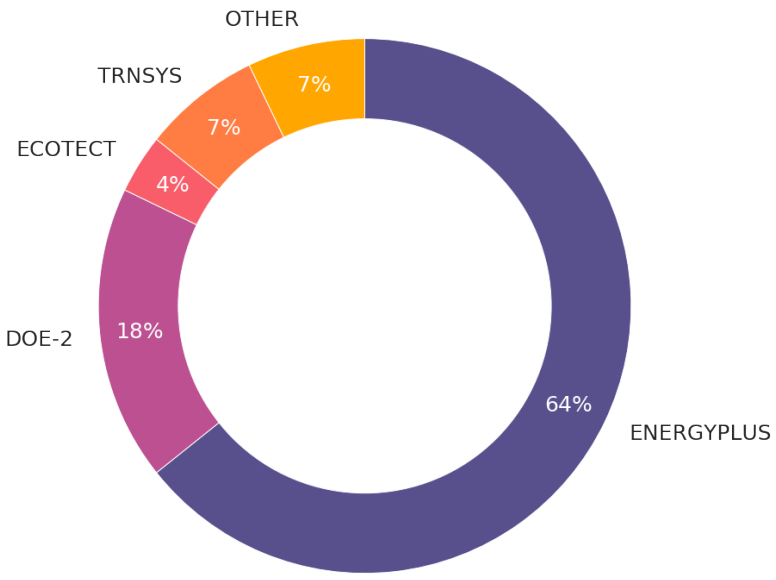

Figure 5 – Breakdown of energy simulation software used by researchers referenced in this chapter. Where the same dataset was used by several reserachers, such as the dataset developed by [35], the software used to create the dataset was counted only once.

2.4 Location and Climate Selection

The selection of location, and associated weather file for the energy simulation, dictates the area for which the developed surrogate model can be used. Figure 6 shows the weather file locations used for the surrogate models referenced in this chapter. Where multiple weather files were used in a single study, the points on the map were highlighted in red, with each study represented by a unique symbol. Locations shown in blue represent studies where a single weather file was used. There is a large cluster of surrogate modelling research focused on buildings in the United States, Europe, Brazil and Hong Kong. As expected, this aligns with where the research is being conducted. The location selected will not only impact the weather file but also the type and range of input features for the model. For example, Lam et al. [38] developed a surrogate model specific

to Hong Kong, where the year-round hot-humid climate results in little need for heating, so features related to heating were excluded from the input feature set.

Figure 6 – Surrogate model weather file locations for research referenced in this chapter. Where multiple weather files were used in a single study, the points are highlighted in red, with each study represented by a unique symbol. Locations shown in blue represent studies where a single weather file was used. Map created in Google My Maps [39]

Some researchers integrated multiple climates into their building energy surrogate model studies. Hygh et al. [14] evaluated the same building features and ranges with four climate files, and therefore trained a medium-sized office building surrogate model for each climate file. They found that the model for annual heating use prediction in the climate with the lowest heating degree day had the lowest accuracy. Aijazi and Glicksman [36] evaluated training a single surrogate model using simulations with multiple climate files by including annual heating degree days as an input feature to represent the climate differences. When they compared surrogate models developed for each climate to the multi-climate model, they found that the multi-climate model had lower accuracy. Catalina et al. [15] developed a model representative of 16 climates in France, representing climate with the input features climate coefficient, the difference between the heating

set-point temperature (a constant value for all simulations); and the monthly average sol-air temperature, a value calculated using monthly average outdoor temperature, horizontal global radiation, and exterior convection coefficient. Further research is required to determine if multiple climates can be included in a single surrogate model without significantly compromising the predictive performance of the model.

2.5 Building Features and Range Selection

There are several considerations that go into selecting the building features and associated ranges that form the input matrix. This step in the surrogate modelling process is of high importance as it impacts the design space for the final surrogate model, and the model behaviour. Tian and Choudhary [22] stated that building features impacting energy use are valuable for analysis and should be focused on. Tsanas and Xifara [35] followed this line of thought and selected building features based on expert knowledge and judgement of which building attributes had the largest impact to the target variable(s). With a goal of developing an early-stage design tool, Hygh et al. [14] selected building features that are typically known during early stage building design and are known to have a significant impact on annual total, heating and cooling energy, such as general building geometry, enclosure thermal performance and shading projection factor. Chidiac et al. [19] chose variables to reflect specific energy retrofit measures for existing office buildings, including lighting loads, daylight sensor integration, enclosure thermal performance and infiltration, and HVAC system efficiencies.

Standards and guidelines were frequently used to guide the ranges selected for each building feature. Amiri et al. [40] used ASHRAE 90.1 to determine building envelope variable values for their discrete features. Papadopoulos and Azar [31] also used ASHRAE 90.1 to determine the ranges for indoor temperature setpoints and lighting and equipment electricity use densities.

Korolija et al. [30] determined values for temperature setpoints, fresh air ventilation, occupant density, and lighting and equipment loads from the ASHRAE Standard and Handbooks, European Standards, and Chartered Institution of Building Services Engineers (CIBSE) Guidebooks for their United Kingdom (UK)-based office building surrogate models. Tian and Choudhary [22] developed their base model for secondary schools in London using the UK Department for Education and Skills Briefing Framework for Secondary School Projects for the geometry, UK National Calculation Method for the scheduling, and CIBSE for the internal heat gains.

Some researchers chose to use or build on previous researchers' input features in order to compare their surrogate models. Roy et al. [41], Papadopoulos et al. [42], Castelli et al. [43], and Chou and Bui [44] used a dataset prepared and made publicly available by Tsanas and Xifara [35]. Al Gharably et al. [45] built on Hygh et al.'s [14] research and added non-rectangular building geometry to the model, keeping all other input variables, targets, sampling plans, and climates the same. Others used features that were unique when compared to other studies. Edwards et al. [46] used a large set of building attribute variables, totaling 156. The majority of these variables were related to building enclosure material properties. Melo et al. [27] combined U-value and thermal capacity values to create 11 unique features representing wall and roof types. Geyer and Singaravel [29] proposed using a construction-level component-based machine learning approach where the features were divided into construction elements to allow for flexibility of building energy surrogate models.

Some researchers had access to real building feature datasets and used these as the feature sets for their building energy software simulations. Tian et al. [18] used building characteristics from University of Pennsylvania and Georgia Institute of Technology campus buildings to build an input dataset, and then used EnergyPlus to simulate the target values.

A summary of the feature categories used by researchers in their surrogate model development is in Figure 7. Further, a detailed summary of the features for the research referenced in Appendix A.

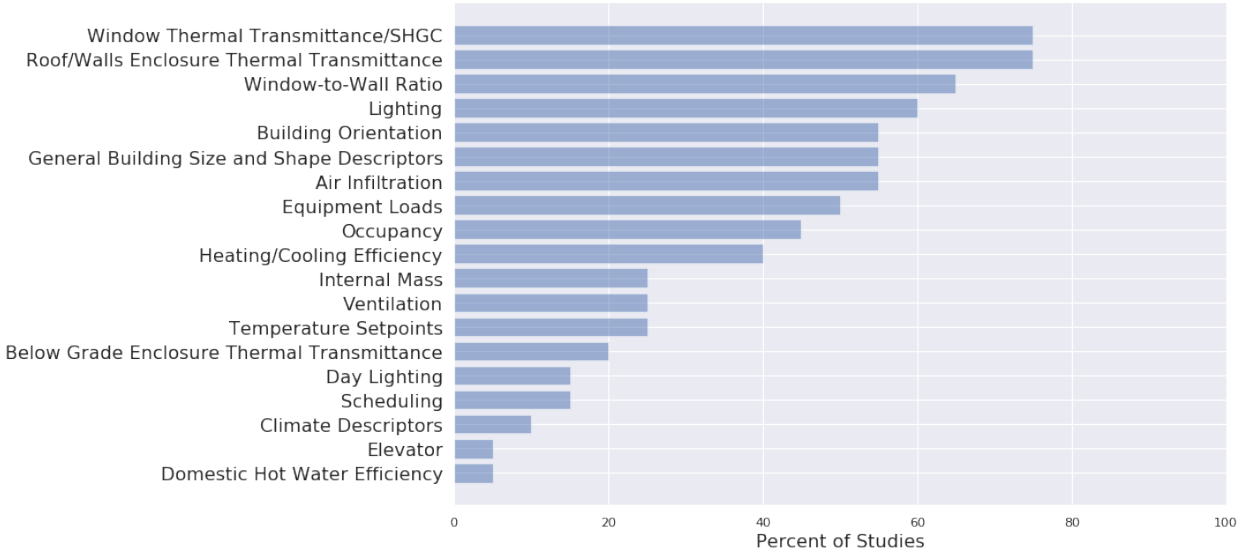

Figure 7 – Percent of studies summarized in Appendix A using building input features

2.6 Sampling Plan Selection

Previous research studies used various methods when developing the datasets used to train learning algorithms and test models. Sangireddy et al. [28] emphasized that the accuracy and validity of the surrogate model is highly dependent on the sampling plan selected. Different methods used to generate the input matrix have included: determining building attribute combinations from existing buildings [18]; generating sample sets within ranges of building parameters using uniform probability distribution Monte Carlo sampling [14, 21, 40] and Latin Hypercube Sampling (LHS) methods [24, 27, 31, 35]; and randomly generating geometry features using building modelling software plug-ins [20, 47].

With Monte Carlo sampling using uniform probability distribution, the variable values are randomly selected from within the ranges for each sample. By contrast, Latin hypercube sampling extends the Latin square, a grid with one sample per row and column, to multi-dimensional space. One sample per axis-aligned hyperplane is generated, creating a matrix of space-filling, near-

random selections where the values within the variable ranges are uniformly selected. Figure 8 illustrates Monte Carlo and Latin hypercube sampling plans for two variables with 100 samples each. The Latin hypercube sampling plan distributes the samples through the entire design space whereas the Monte Carlo leaves gaps in the design space and creates random clusters of data points in some locations. Sacks et al. [48] presented a case for using space-filling sampling plans, such as Latin hypercube, instead of using random sampling plans for computer-generated data, stating that when error is systematic (as is the case in computer-simulated design) the experimental design should fill the design space.

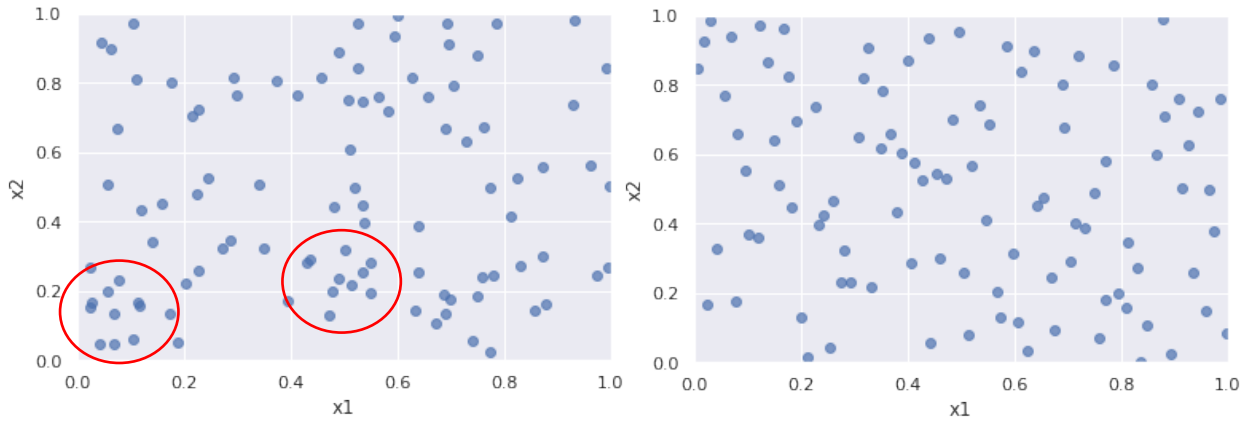

Figure 8 – Monte Carlo with uniform distribution of variables (left) – clustering shown in red circles. Latin Hypercube Sampling (right)

Older studies were found to have created datasets based on a more traditional parametric study methodology. For each simulation, a single variable was modified, and the other parameters were kept in a base case condition. Lam et al. [38] used a combination of parametric and factorial dataset design. A univariate linear regression analysis was performed for each variable to determine the features most sensitive to the targeted annual electricity energy use. The most sensitive features were selected and divided into building load, HVAC system, and HVAC refrigeration plant. All combinations within those categories were simulated.

Wong et al. [32] used several building features affecting daylighting, each with a defined number of discrete perturbations. Each perturbation was simulated with the base case. To test the model trained using the parametric data, Wong et al. [32] randomly generated three samples of variables within the discrete perturbation ranges. With parametric sampling, the interactions between variables are not taken into account. If all interested parameters are changed in each sample, the simulation software behaviour of the interaction between two or more variables can be learned.

One method of setting up the sampling plan to learn the interaction behaviours of the variables is by simulating all combinations of the discrete features. Tsanas and Xifara [35] simulated all combinations of the discrete perturbations for each of the building forms, glazing areas, and building orientations for a total of 768 simulations.

Unique sampling plans were used by some researchers. Sangireddy et al. [28] used the k-means clustering algorithm to cluster all possible combinations of the discrete variables selected, a total of approximately 100,000 combinations. They increased the number of clusters until the cluster model sum of squared errors began to decrease less drastically. The sample from each cluster centre was selected to represent all samples within that cluster and together represented the overall domain space. This method reduced the dataset from approximately 100,000 samples representing all combinations of discrete variables to 200 samples (and therefore 200 clusters). Sangireddy et al.'s [28] sampling plan represented a subset of all combinations of the discrete variables chosen as they were proposing using the surrogate model in lieu of an exhaustive parametric analysis.

2.7 Target Variable Selection

Surrogate models can be trained to predict any of the energy simulation software results. Many researchers used annual and/or monthly total energy, heating and/or cooling energy use or intensity (energy use per floor area) as continuous model target variables. Figure 9 summarizes the building

energy targets researchers trained their surrogate models to predict. Some researchers trained several surrogate models in a single study to predict multiple target variables.

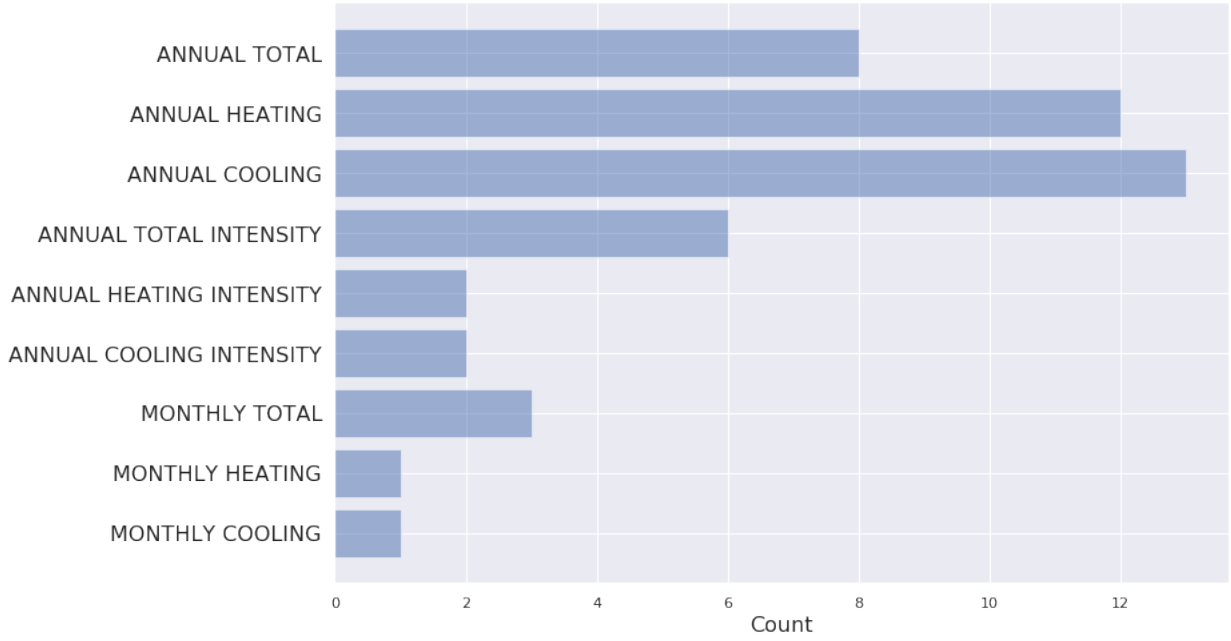

Figure 9 – Summary of surrogate model building energy target variables for studies summarized in this chapter.

Some researchers chose target variables other than building energy for their surrogate models. Ascione et al. [49] and Chen et al. [50] used trained surrogate models to predict targets related to human comfort. Ascione et al. [49] trained a model to predict the percentage of annual discomfort hours, and Chen et al. [50] trained a model to predict ASHRAE55 comfort time and illuminance level. Tian et al. [51] proposed overheating risk and peak heating/cooling use as potential target variables, although they limited their models to annual heating, cooling, and CO_2e emissions.

A few researchers used subsets of building energy target variables. Chidiac et al. [19] separated the targets into end uses and developed surrogate models for annual lighting, equipment, pumps, fans, DHW, chiller, boiler, and electrical. The summed results of all the models provided the annual total energy use. Lam et al. [38] developed separate models for building loads, HVAC systems, and HVAC refrigeration plants. Edwards et al. [46] trained models on 90 of the

EnergyPlus simulation outputs. Instead of training 12 separate models for each month of the year, Papadopoulos and Azar [31] added 12 dummy variables to the input matrix, with the relevant month set to 1 and the remaining 11 variables set to 0. Wong et al.'s [32] model focused on daylighting and predicted daily cooling, heating and lighting electricity energy uses.

Although not as common as using continuous variables for targets, some researchers have converted the building energy use target values into categories such that classification algorithms could be used for the prediction model. Chari and Christodoulou [52] converted the energy modelling results to Irish Dwellings Energy Assessment Procedure Building Energy Rating classifications [53] which combined annual energy use intensity and CO_2 emissions to classify building energy performance into 13 categories. Tsanas and Xifara [35] discretized the continuous annual heating and cooling energy use so that classification algorithms could be used on the dataset.

2.8 Sample Set Size, Data Splitting and Cross Validation

The number of samples and ratio of samples per training, validation and test dataset was not consistent between studies. This is common for machine learning where there are general rules-of-thumb but no defined methodology. This section provides a summary of the sample set sizes and the methods used to split the data into sets.

Hygh et al. [14] started with a training set size of 16,000 samples and found through training a multivariate linear regression (MVR) model, the average percent error of the model reached a minimum between 500-1,000 samples. Aijazi and Glicksman [36] used training set sizes of 50 samples to train more complex algorithms. Both researchers looked at similar sized buildings, in multiple US climate zones, but according to the Hygh et al. [14] analysis, Aijazi and Glicksman [36] may have been using too small a training set size to achieve the most accurate models.

However, since Aijazi and Glicksman [36] used LHS and Hygh et al. [14] used Monte Carlo, the sampling plan may have had an impact on the training size required.

Melo et al. [27] evaluated the model accuracy with different numbers of training samples. They found that the maximum error decreased gradually as the number of training samples increased above approximately 13,000. Chen et al. [50] evaluated the multivariate linear regression coefficient values using different sized training sets, each developed with separate Latin hypercube sampling plans. Sample size increments from 100 to 10,000 were evaluated. Their results showed that with a small training set size, a larger number of input features were important.

A few researchers developed multiple training and validation datasets using either k-fold cross validation [27, 35, 43, 54], and/or bootstrapping [50, 51] in order to evaluate the model variation. K-fold cross validation splits the training dataset into 'K' equal-sized sets. 'K-1' sets are combined and used as the training set and the remaining dataset is used as a validation set. This is completed 'k' times so that each set is used as the validation set once. Bootstrapping, specifically bootstrapping with replacement, is defined as when each sample in the training set is randomly selected from the dataset and when it is selected remains in the original dataset. Therefore, a single sample may be selected multiple times. Using bootstrapping, multiple training and validation datasets can be made from a single dataset. Both strategies are used in surrogate modelling to fit the data subsets to multiple models. The model coefficient variation and predictive performances can be used to assess the stability of the model. From the studies referenced in Appendix A, approximately 20% used cross validation with either k-fold or bootstrapping.

Tian and Choudhary [22] performed both cross-validation and bootstrapping with replacement on their dataset. They found that for multivariate linear regression, the validation set model predictive performance, using coefficient of determination, was lower than for the training dataset.

Conversely, when bootstrapping was used, the model predictive performance was the same for both the training and test dataset. In a later study, Tian et al. [51] tested varying bootstrapping with replacement sample sizes and plotted the standardized regression coefficients for each feature to evaluate at which bootstrapping size the values stabilize. Standardized regression coefficients are the coefficients/weights when the independent and dependent variable variances are standardized to 1.0. Aijazi and Glicksman [36] developed multiple training and validation datasets by creating 10 datasets of 50 samples each from multiple Latin hypercube samples. From the 10 datasets, they randomly selected one as the training set and one as the validation set. They performed this 80 times and calculated model predictive performance for each iteration thereby achieving a model error representative of the mean and standard deviation of the 80 iterations.

2.9 Transformation of Input Features and Target Variable(s)

Target variable transformation can improve the surrogate model behaviour and accuracy. In mathematical modelling, it is common to apply a non-linear transformation to the input features and/or target variable and evaluate how the model behaves. More specifically, in multivariate regression, it is assumed that the input variables are independent of one another and that the target variables are normally distributed. Target variable transformation is commonly completed to force the data into a normal distribution. One method to determine the appropriate target variable transformation is the Box-Cox method [55]. Using this method, a lambda (λ) value is determined for Equation 3 that transforms the target variable values (y) into a normal distribution. Where the optimal λ is 0, the transformation is logarithmic.

$$y' = \frac{(y^{\lambda}-1)}{\lambda} \qquad (3)$$

When comparing the model performance impact of transformed target variables, the model predictive performance can be evaluated in two ways: predictive performance compared to the transformed values, and predictive performance when the values are transformed back to the target variable units. The first evaluates how well the algorithm is fitting the transformed data and the second evaluates how good the transformed data and algorithm are at predicting the target variable compared to the building energy simulations.

Tian et al. [18] quadratically transformed the annual heating and cooling energy use target variables and found an improved model performance for the learning algorithms evaluated. Robinson et al [54] and Melo et al. [27] performed a logarithmic transformation of the target variables. Melo et al. [27] applied the Box-Cox transformation method to determine which transformation of the annual cooling energy use target variable was most appropriate for their data set to reduce the skewed distribution and make distribution more normal.

2.10 Combining Features

Combining original input features into features that represent the whole building or an elevation has been shown to improve model prediction performance in previous research [14, 36, 56]. The impact to the energy use of an uncombined building feature, such as window-to-wall ratio per elevation and floor may individually be minor. However, when combined into the overall window-to-wall ratio for the full above grade building wall area, it may have a larger impact on the annual energy use. Based on knowledge of how the input variables interact in the energy simulation software used, Hygh et al. [14] developed 63 additional variables using linear combinations of 27 variables. Using knowledge of building parameter interactions, Signor et al. [56] improved the linear regression model performance by combining features to create new features more representative of the buildings as a whole. Aijazi and Glicksman [36] used the equation for heat

flux through an exterior enclosure to justify creating additional terms where heating degree days is multiplied by roof, window and wall thermal transmittance and area. Wong et al. [32] combined glazing solar heat gain coefficient and visible transmittance with window-to-wall ratio for a daylighting-focused surrogate model in a cooling-dominated climate where the balance of interior daylighting and solar heat gain was critical for predicting electricity energy use. Asl et al. [20] used building geometry cross terms, such as interior floor area multiplied by interior floor height, in their model. Combining features can improve the usability of the surrogate model in early-stage design when the exact building shape has not been tested. The design team may not have decided on the exact building form and may want to explore, for example, how the wall areas, aspect ratio and window-to-wall ratio impact the annual energy use, knowing what the project specifications are for the conditioned floor area.

It should be noted that when the original variables are linearly combined, they will have a high correlation to the original variables. In regression analysis, correlated inputs can lead to high variance values of the feature weights/coefficients [57]. Therefore, if the original variables are not removed from the model, the coefficients for the multivariate linear regression equation will not be stable and therefore cannot be accurately interpretable. According to Tian et al. [18], not many researchers have evaluated feature correlation in building energy modelling. It is also not clear if previous studies have addressed the issue of multicollinearity when adding terms that are combinations of two or more of their original features.

2.11 Normalization

It is common to normalize/standardize the input features before training a machine learning model [12]. If the input features are not normalized prior to training, this may lead to coefficients biased

by differing input feature magnitudes. Some also normalize the target variable [35] but this is not a necessary step.

Instead of normalizing the input features, Hygh et al. [14] normalized the multivariate linear regression coefficients (standardized regression coefficients) in order to compare the coefficients in terms of sensitivity to the target variables. To do this, they multiplied each coefficient by the input variable's standard deviation and divided by the standard deviation of the target. This means that the input features were not on the same scale when the model was trained and the magnitude of the feature value may have influenced the coefficient value.

2.12 Feature Selection/Elimination

A few researchers used feature selection and/or elimination methods in their studies. Lam et al. [38] evaluated the correlation of 62 input features to annual electricity energy use and selected 28 features with the highest correlation. Ascione et al. [49] used the standardized rank regression coefficients for their annual heating and cooling energy use intensity model determined from a previous study completed by Mauro et al. [58] where the target was thermal energy demand. Variables with coefficient values lower than 0.05 were removed from the input feature matrix.

Tian and Choudhary [22], Lam et al. [38] and Ascione et al. [49] each used the multivariate linear regression standardized coefficients to determine the input features with highest significance to the energy use target. Tian and Choudhary [22] selected four of the original seven features and Lam et al. [38] selected 28 of the original 62 input features with the highest coefficient values to train new MVR models with the feature subsets. Ascione et al. [49] used trained model coefficients determined from a previous study by Mauro et al. [58] to remove input features with MVR coefficient values lower than 0.05 for training an artificial neural network model.

Tian and Choudhary [22] used the standardized coefficients from multivariate linear regression and the total variance contribution for each input variable from multivariate adaptive regression splines (MARS) to determine which input variables had the highest significance to annual heating energy intensity. They selected four of their original seven variables with the highest coefficient values (representing 93% of the model variance according to the MARS model) and trained a new multivariate linear regression model with the four input variables. Roy et al. [41] used Multivariate Adaptive Regression Splines to determine feature importance (degree of participation) of their input features and selected the features with high importance for use in an Extreme Learning Machine (ELM) model. They found that this 'hybrid' approach, with a subset of features produced a more accurate model for predicting heating and cooling energy use than using MARS and ELM, separately.

Hygh et al. [14] used forward stepwise selection to select features that reduced model error. The results showed that a combination of adding expert knowledge-informed combined terms and feature selection through forward stepwise regression can improve model performance. Amiri et al. [40] used forward and backward stepwise regression to remove low significance variables from their input feature set.

Edwards et al. [46] and Sangireddy et al. [28] highlighted the benefits of embedded feature selection when using the L1 regulator, known as least absolute shrinkage and selection operator (LASSO), with multivariate regression. When the LASSO regulator is used, the coefficients for the input features with low or no significance to the target variable are driven to zero and therefore removed from the model.

2.13 Training Algorithms

There is a large focus by researchers on determining the learning algorithms that perform best for building energy surrogate modelling. Aijazi and Glicksman [36] stated that the advantages and disadvantages of different learning algorithms are not well understood in the field. Some researchers test multiple algorithms to determine which performs best for their dataset. Other researchers test a single algorithm and evaluate its accuracy alone. Figure 10 summarizes learning algorithms tested in several of the articles referenced in Appendix A.

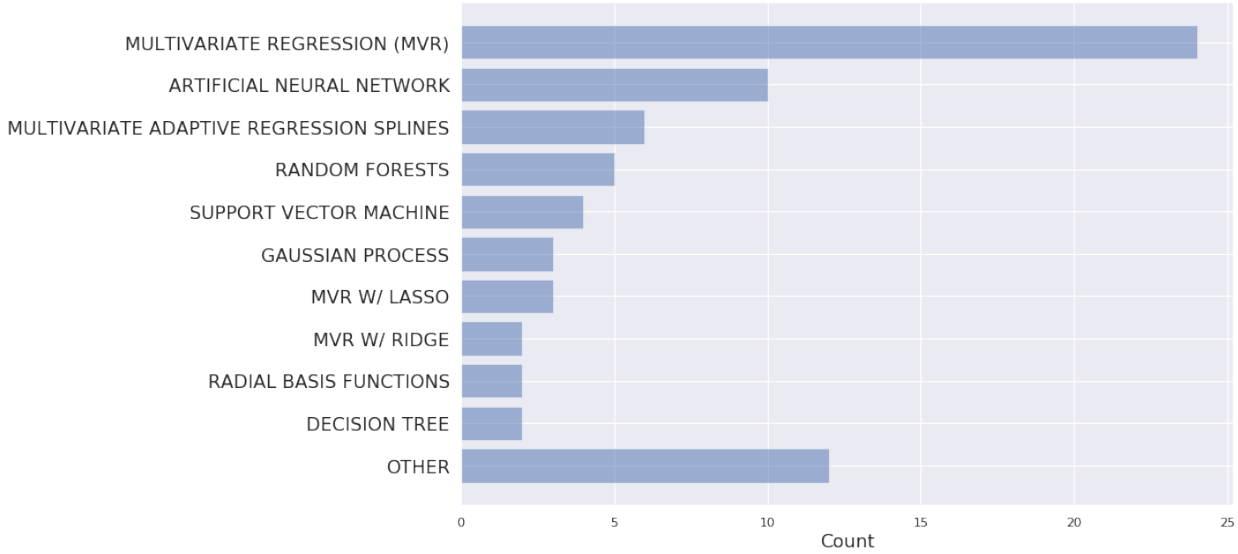

Figure 10 – Breakdown of learning algorithms used for surrogate model development in studies referenced in Appendix A.

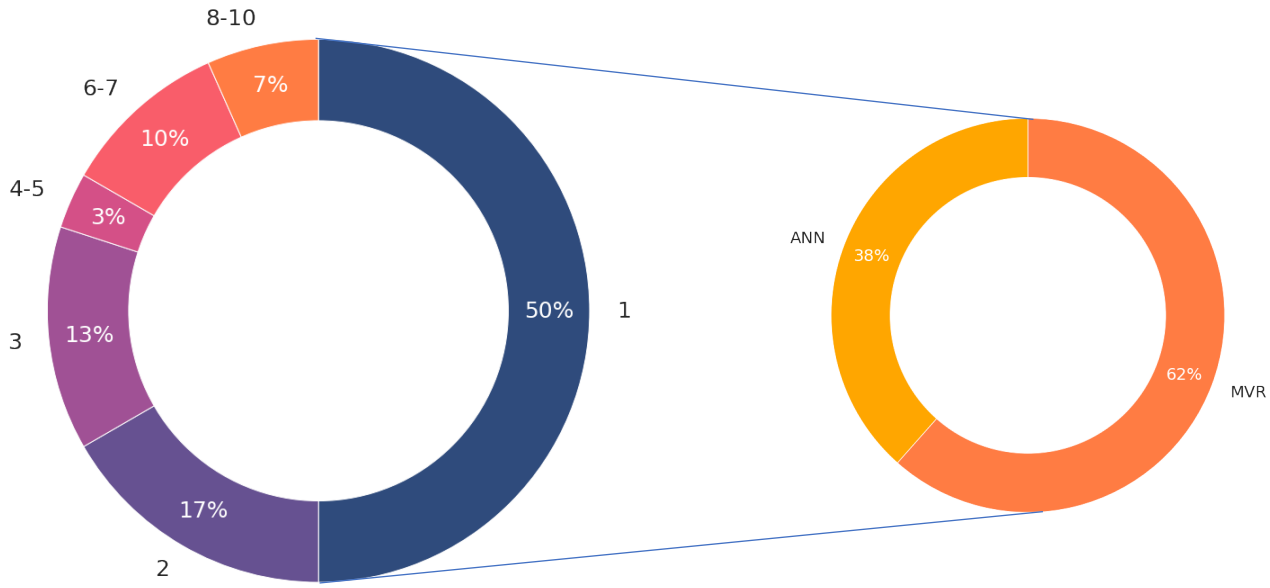

Figure 11 – From studies referenced in Appendix A, number of learning algorithms tested during surrogate model development. Right pie chart shows, for the studies where one algorithm was used, which algorithm was selected.

Figure 11 shows how many learning algorithms researchers used as a percentage of studies referenced in Appendix A. Of the studies where one algorithm was used, all the studies used either artificial neural networks (ANN) or multivariate linear regression (MVR). Many researchers who evaluated 2 or more learning algorithms also used MVR and/or ANN.

Nagpal et al. [24] used a tool developed by Mueller [9] to determine which learning algorithm performed best for their dataset. The tool trained the dataset using random forests and neural networks with 10 combinations of hyperparameters for each and provided the user with the model that produced the minimum error on the validation dataset.

Aijazi and Glicksman [36] used surrogate modelling algorithms discussed in Forrester et al.'s [8] Engineering Design via Surrogate Modelling textbook, Radial Basis Functions and Kriging. Geyer and Singaravel [29] proposed using long short-term memory (LSTM) in the neural network model

to store interactions of the input features that can be used for future surrogate models with different targets.

The hyperparameter tuning techniques including the final model parameters used, were often not disclosed in the studies referenced in Appendix A. Therefore, the researchers may not have evaluated varying hyperparameter ranges and combinations. Instead, they used the parameters built into the algorithm package. Papadopoulos et al. [42] used a dataset prepared by Tsanas and Xifara [35] which was also used by Chou and Bui [44], Castelli et al. [43], and Roy et al. [41] to evaluate tree-based ensemble learning algorithms. Papadopoulos et al. [42] used an exhaustive grid search method to tune the algorithm hyperparameters. Their results showed that when systematic tuning of the hyperparameters was performed for random forests, the model performance significantly improved compared to Tsanas and Xifara's [35] random forest model where hyperparameter tuning was not indicated as being performed and the hyperparameters used were not stated.

Some learning algorithms produce models that allow for interpretability of the building feature sensitivity to the target, using analytical tools such as multivariate regression, decision trees, and in some cases neural nets. Researchers have drawn conclusions about the relative impact of the input features to the target for their specific dataset. This could be used to inform designers on which building attributes will have the largest and smallest impact on energy use.

Many researchers have evaluated feature importance in a data pre-processing step by analyzing each variable's correlation, using metrics such as the Pearson correlation coefficient, to the simulated result or following model training by analyzing the feature weights [14, 18, 35, 41, 42].

Tsanas and Xifara [35] used the weights assigned by the random forests learning algorithm to evaluate the input features' importance to the target. Tsanas and Xifara [35] stressed that

evaluating the sensitivity of the input features to the target(s) using the associated weights in the trained model, does not determine the direct relationship between the input variable and the target, as is the case when evaluating correlation using the Pearson correlation coefficient or performing parametric studies. Instead, the weight values take into account both the relationship of the variable to the target, the relationship between the input variables, and the joint relationship of multiple input variables to the target.

2.14 Model Performance Analysis

There are several common model performance metrics used to determine the ability of the model to predict the target variable. These are often reported on the validation and test datasets, but it is also important to evaluate the model prediction performance for the training dataset and compare to the validation/test dataset to determine if the model is overfitting to the training dataset. Figure 12 shows a breakdown of the model predictive performance metrics used in studies referenced in Appendix A. Most studies used more than one metric to evaluate model performance.

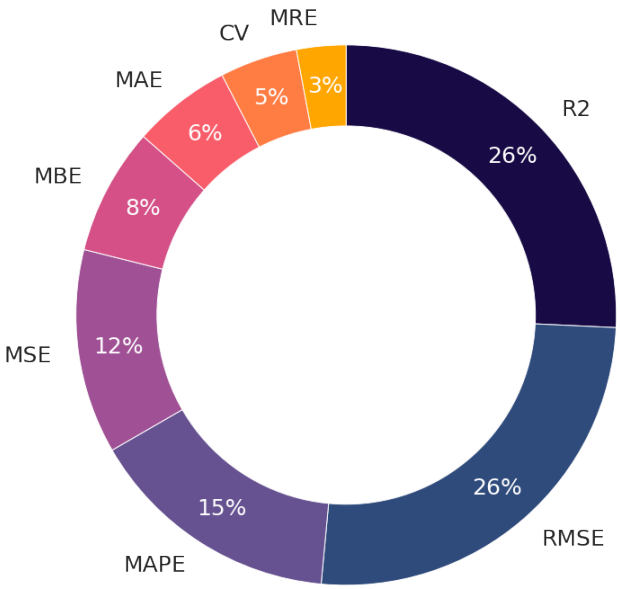

Figure 12 – Breakdown of model predictive performance metrics used in studies referenced in Appendix A. Most studies used more than one metric to evaluate model performance.

Using multiple performance metrics can provide additional insights into the behaviour of the model. For example, root mean squared error (RMSE) and mean absolute error (MAE) are measurements of the average difference, or error, between the target variable and the model predicted value. As illustrated in Figure 13, absolute error is linear across the residual values whereas squared error is exponential. Therefore, RMSE places a higher weight on large errors compared to MAE.

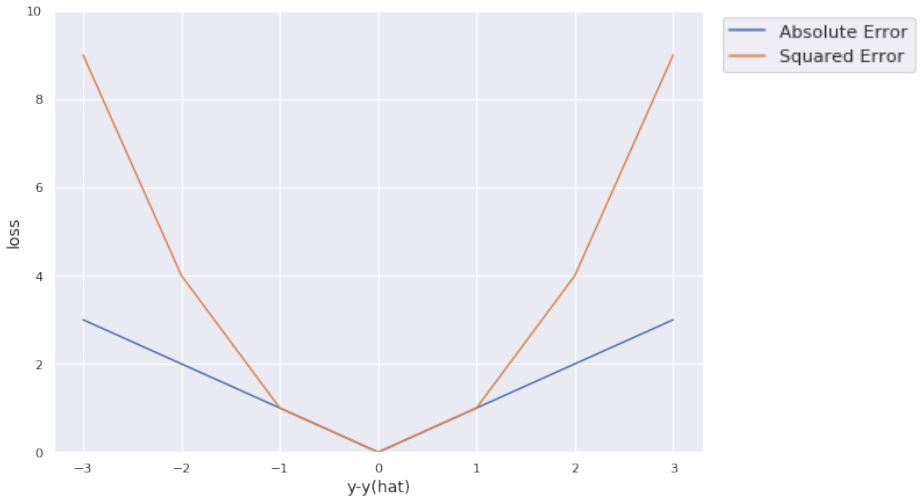

Figure 13 – Plot of squared error and absolute error for incremental residual values. Adapted from: [12]

Chou and Bui [44] proposed using a metric to combine RMSE, MAE, MAPE and R^2 that they referred to as the synthesis index (SI) (Equation 4). It incorporated the number of model performance metrics and the result of the performance measure, P_i. Not only did this average the performance metrics, it also accounted for variation of results from the k-folds cross-validation sets.

$$SI = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{P_i - P_{i,min}}{P_{i,max} - P_{i,min}}\right) \quad (4)$$

Catalina et al. [15] used residual plots to confirm that their multivariate linear regression assumption that the residuals were normally distributed was accurate. Papadopoulos and Azar [31] plotted the model residuals on a histogram to check if the residual distribution was normal.

Sangireddy et al. [28] showed the residual plots for both the training and testing datasets. The training set residual plot showed the residuals evenly distributed above and below the zero-residual line; however, the test set residual plot showed a non-symmetric distribution. They concluded that since the non-symmetric plot did not have a pattern or curve, the surrogate model was appropriate.

A few researchers [27, 36] used model training time, alongside model accuracy metrics, when comparing models. Model training time can be used to compare algorithms, but the absolute time depends on parameters beyond the training set and algorithm, such as computer and algorithm package performance. Unless the goal is to continually train algorithms with incoming data (such as proposed by Geyer and Singaravel [29]), the algorithm training is a very small part of the overall surrogate modelling process.

In summary, there are a variety of methods researchers have used at each stage of the building energy surrogate modelling process. The methodology decisions impacted the dataset and thereby effected the behaviour and predictive performance of the surrogate model. If the surrogate modelling development process differs between studies, the final surrogate model cannot be directly compared.

3 Building Energy Surrogate Modelling – A Feature Selection Methodology Using Wrapper and Embedded Techniques

This study used building energy surrogate model development methodologies similar to previous work summarized in Chapter 2. An identified gap in the literature review was the selection of key input features for such surrogate models, which generally had been based on expert knowledge. Few previous studies used learned models to select key building features [10]. No previous research was found in which wrapper and embedded feature selection methods were used together to select features for building energy surrogate models. Such feature selection can be used to inform archetype energy model design, code and standard requirements, and assumptions for detailed building energy models. Selecting the most relevant feature set for predicting the output with a certain learning algorithm can have many advantages for the performance of the model including removing irrelevant features from the model, reducing model overfitting, and reducing model run time. The feature selection methods used in this research produced a surrogate building energy use model for predicting annual building energy use. The model can be used to quickly evaluate building energy use based on any combination of features within the model's design space. This research therefore contributes to the field of building energy surrogate modelling by combining wrapper and embedded feature selection techniques and using them intentionally to select the features that together, best predict the simulated building energy use.

Multivariate linear regression was used as a starting point for this research, consistent with the majority (63%) of the studies reviewed. Model prediction accuracy was evaluated for both untransformed and transformed input and target variables and the transformations that resulted in the model with the best fit and most uniform residual plot were used. A process of feature combination selection and embedded feature selection using least absolute shrinkage and selection

operator (LASSO) and Elastic Net was evaluated. This multiple step process to surrogate model development is unique to this research and provides a methodology for reducing a feature subset for simplified building energy predictions suitable to building design decisions during the early stages of design when building form, size, enclosure types and mechanical systems are being explored.

3.1 Methodology

Table 1 summarizes the different methods used at each step of the surrogate model development process. By clearly defining the surrogate model development process, the author intends to allow for other researchers in the field to quickly compare their models to the model described in this chapter.

Table 1 – Summary of surrogate model development

DATASET DEVELOPMENT	MODEL INTENT	- Early stage design tool for new buildings and existing building retrofits - Evaluate the sensitivity of the input features to the target variable	
	TARGET VARIABLE	- Annual building energy use (annual heating + cooling + fan + pump energy use)	
	BUILDING ARCHETYPE	Large office (6,570 m^2 to 1,780,000 m^2)	
	LOCATION + CLIMATE	2016 Toronto City Centre, Ontario, Canada – CWEC weather file [59]	
	ENERGY SIMULATION SOFTWARE	EnergyPlus v 8.0.0.008 [60]	
	STATISTICAL ANALYSIS AND MODELLING TOOL	Python v 3.6.5	
	BASE MODEL	U.S. Department of Energy Large Office Commercial Reference Model – EnergyPlus v 7.2 [61]	
	FEATURES + RANGES	- 71 continuous variables - Representative of building geometry, enclosure performance, lighting and electrical power densities, heating ventilation and air conditioning system performance, and occupancy (refer to Table 2)	
	SAMPLING PLAN	- Latin hypercube sampling (LHS) – MATLAB [62] - 4,000 samples	
DATA PROCESSING	TRAIN/VALIDATION/TEST SPLIT	- Training/validation and test random split (70%/15% and 15%) - Training and validation split: random split, repeated 10 times. - Random splitting completed with: *sklearn.model_selection.train_test_split* [63]	
	FEATURE ENGINEERING	- Input feature normalization (mean of zero, variance of 1) - Logarithmic transformation of input and target variables (Box-Cox method used to evaluate normal distribution of data) - Input feature combinations added using forward stepwise selection	
TRAINED MODEL DEVELOPMENT	LEARNING ALGORITHMS + HYPERPARAMETER SELECTION	Multivariate Regression (MVR)	Gradient descent w/ mean squared error cost function
		Lasso (L)	*sklearn.linear_model.lasso* [63]
		Elastic Net (EN)	*sklearn.linear_model.elasticnet* [63]
	ERROR METRICS	- Coefficient of determination (R^2) - Root mean squared error (RMSE) + normalized RMSE - Mean absolute error (MAE) + normalized MAE - Mean absolute percent error (MAPE)	

3.1.1 Generating the Building Energy Dataset

Figure 14 illustrates the workflow used to generate the large office building energy dataset. This study started with a set of 71 building attributes with ranges representative of high to low performing large office buildings in the Toronto, Ontario, Canada climate. 4,000 building design samples, filling the building attribute design space, were created. Building energy use for each sample were determined through building energy simulation software.

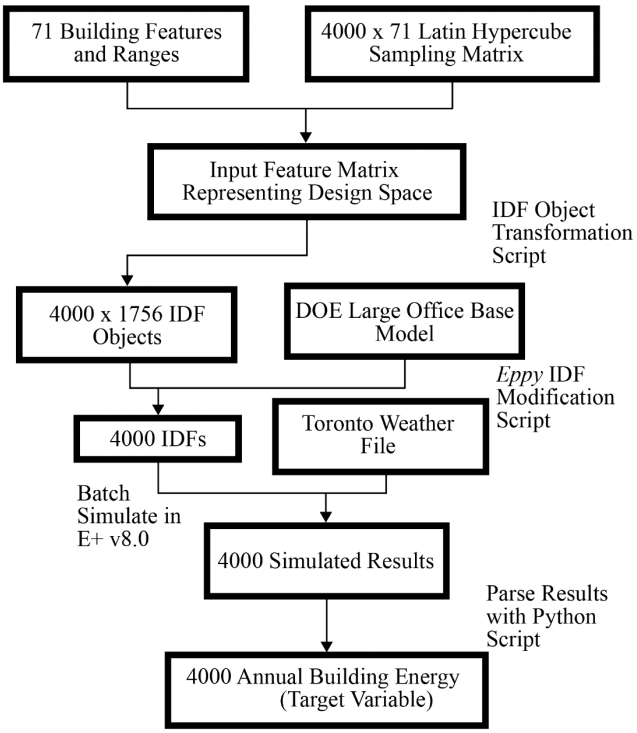

Figure 14 – Workflow for generating building energy use dataset

This study used the U.S. Department of Energy (DOE) Commercial Reference Model for large office buildings [61]. The large office reference model's mechanical system was a central plant with chiller and boiler, and multi-zone variable air volume with reheat distribution system. The DOE selected this mechanical system for large office buildings based on the results of the U.S. CBECS, as reported by Pacific Northwest National Laboratory in their 2006 study [64].

As illustrated in Figure 15, general model construction used in the DOE commercial reference model was maintained with a rectangular footprint, single ground and top floors, repeated basement and middle floors, and plenums on each the ground, middle and top floors. Each floor was made of a single core and four perimeter zones. Windows were represented as strips with the depth modified to suit the sample's window-to-wall ratio.

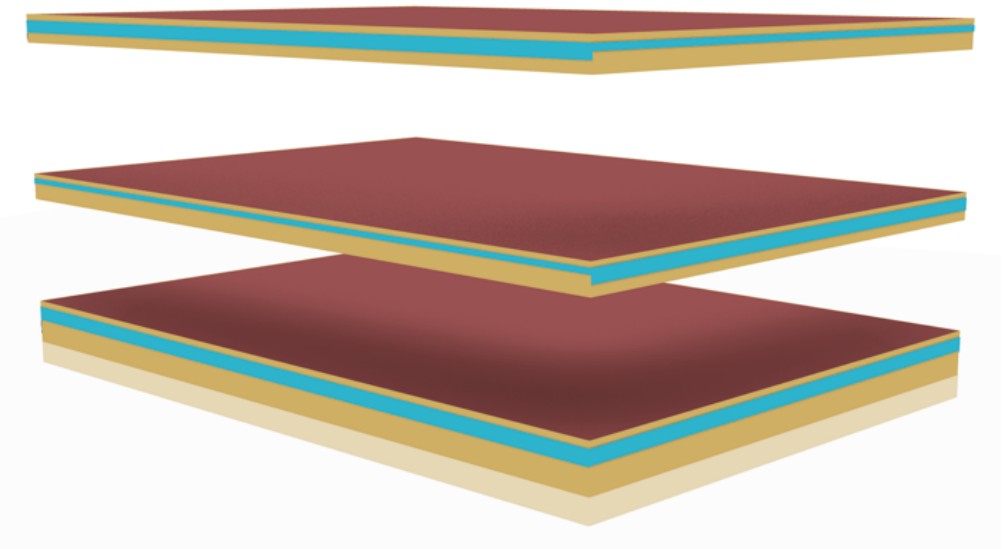

Figure 15 – Wireframe model illustrating modified building geometry

The building floor plan design for the ground, top and repeated floors is illustrated in Figure 16. Each above-grade floor comprised four perimeter zones (one per elevation), and a core zone. The building orientation was represented as degrees clockwise from north and the building elevations were labelled clockwise.

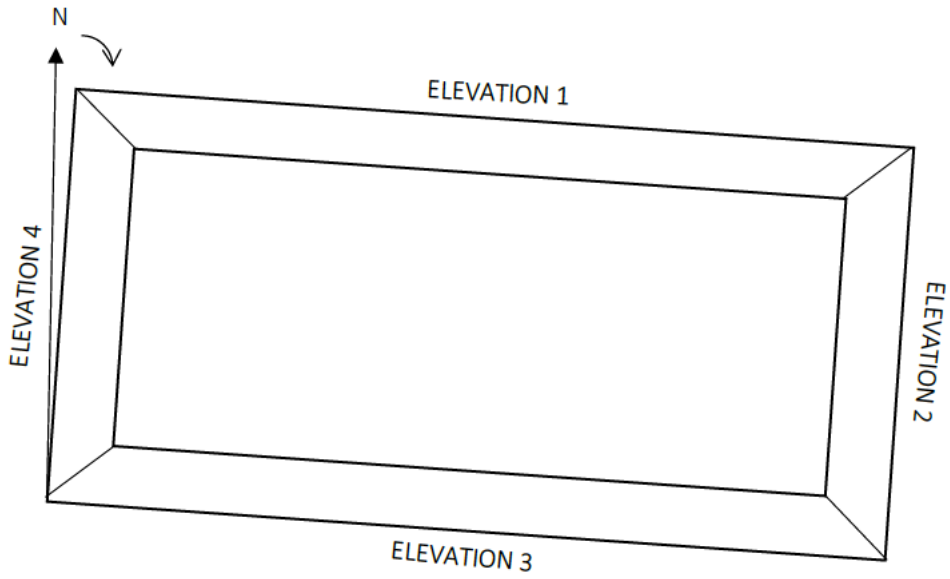

Figure 16 – Illustration of above grade floor plans

The 71 features selected fell into building geometry, building enclosure performance, air infiltration, internal loads, heating ventilation and air conditioning (HVAC) system performance, occupancy, and internal mass categories. The ranges of the 71 features were selected from a combination of the U.S. DOE Existing Commercial Reference Models [61], ANSI/ASHRAE/IES Standard 90.1 Prototype Building Models [65], and industry knowledge of low to high energy performance building attributes [66]. All building attributes that were modified within the DOE Commercial Reference Models [61] for the specific climate zone and construction era reference models were included in the input feature set. The selected 71 features and associated ranges are summarized in Table 2.

Table 2 – Building features and ranges used for dataset development. Integer features are marked with "(I)"

	Feature	Range	Feature	Range
Building Geometry	Number of Repeated Floors	8-80 (I)	Basement/Repeated/Top Floor Height (m)	3-5
	Number of Below Grade Floors	1-2 (I)	Basement/Ground/Repeated/Top Floor Plenum Height (m)	0.1-1
	Building Orientation (° from Elevation 1)	0-90	Perimeter Zone Depth (m)	3-5
	Width/Depth (m)	20-150	Ground/Repeated/Top Floor Elevation 1/2/3/4 Window-to-Wall Ratio (m^2/m^2)	0.1-0.7
	Ground Floor Height (m)	3-10		
Building Enclosure Performance	Whole Window U-Value Elevation 1/2/3/4 (W/m^2K)	0.7-7	Opaque Wall RSI-Value Above Grade Elevation 1/2/3/4 (m^2K/W)	0.35-5.5
	Window Solar Heat Gain Coefficient Elevation 1/2/3/4	0.1-0.8	Below Grade Wall and Slab-on-Grade RSI-Value (m^2K/W)	0.35-3.5
	Window Visible Transmittance Elevation 1/2/3/4	0.2-0.8	Roof RSI-Value (m^2K/W)	0.35-7
Air Infiltration	Air Infiltration Rate – Ventilation System Off ($m^3/s/m^2$)	0.0003-0.002	Air Infiltration Rate – Ventilation System On (% of off)	0.2-0.6
Internal Loads	Basement/Ground/Repeated Perimeter/Repeated Core/Top Floor Light Power Density (W/m^2)	2-30	Basement/Core/Perimeter Equipment Power Density (W/m^2)	8-21
	Elevator Design Level (no. of elevators)	6-30 (I)		
HVAC System Performance	Boiler Efficiency (%)	0.6-0.94	Supply Air Temperature – Heating (°C)	47-55
	Chiller COP (W/W)	3-7	Supply Air Temperature – Cooling (°C)	12.7-18
	Water Heater Efficiency (%)	0.6-0.94	Outside Air Rate (m^2/s-person)	0.00125-0.005
	Temperature Setpoint – Heating – Occupied (°C)	18-22.9	Fan Efficiency (%)	0.5-0.85
	Temperature Setpoint – Cooling – Occupied (°C)	23-26	Fan Pressure Rise (Pa)	1017-1390
	Temperature Setpoint – Heating – Setback (°C less than setpoint)	0-6	Fan Motor Efficiency (%)	0.6-0.95
	Temperature Setpoint – Cooling – Setback (°C more than setpoint)	0-4		
Occupancy	Maximum Occupant Density (m^2/person)	4.5-20	Basement Maximum Occupant Density (m^2/person)	30-40
Internal Mass	Internal Mass (Multiplier of Exterior Enclosure Surface Area)	0.5-5		

Many considerations were taken into account when selecting both the features and their ranges for the dataset generation. The goals were to select features that differed between large office buildings in Toronto, and were included by previous researchers in surrogate model development. The associated feature ranges were selected to include values typical of existing and new buildings in a Toronto climate.

From the studies summarized in Appendix A, there were 39 features used by a minimum of 2 researchers. All features except for roof and wall emissivity, shading, daylighting and scheduling were included, in some form, in the dataset for this study. Building enclosure emissivity was not included as Hygh et al.'s [14] results showed that it had little importance to the annual energy use prediction in cold climates. The impact of shading of adjacent buildings and other objects and shading elements integrated into the building enclosure was excluded from the dataset. Both shading and daylighting are typically not included in early-stage energy models completed in industry and are excluded from the DOE Commercial Reference Model for large office buildings. Lastly, scheduling was kept constant in the datasets as downtown Toronto large office buildings generally operate under the same schedule. It is common in large office building early-stage energy modelling for operation schedules to remain constant as the tenants occupying the building and their scheduling requirements may not yet be known.

There were a few features included in this study that were not included in the studies referenced in Appendix A. Most importantly were heating and cooling temperature off-hour setbacks and supply air temperatures which, as shown in the results section, ended up having high importance to the annual building energy use. In addition, outside air rate was included in the feature set which, out of the studies summarized in Appendix A, only Nagpal et al. [24] had included outside air flowrate.

The building orientation range was 0-90 degrees from elevation 1 (refer to Figure 16) to allow for a building rotation. Since the properties of each elevation were modified independently, this allowed for modification of the elevations to suit which elevation was deemed north.

The perimeter zone depth is typically defined by an energy modeler, typically based on the enclosure properties, and the interior layout and loads. A variation of 2 m in perimeter zone depth was included in the feature set to evaluate the impact of perimeter zone depth on the annual energy use.

The window-to-wall ratio was kept to a maximum of 70% as the windows were only applied to floor height and not the mechanical plenum. This was the maximum window-to-wall ratio that could be applied so that any combination of floor and mechanical plenum height could be simulated. The windows were applied as a strip to the occupied height of the wall as transparent glazing is not typically installed at mechanical plenum locations. The windows spanned the width of the wall and the height varied to achieve the desired window-to-wall ratio.

The range upper limit for lighting and electrical plug power densities was selected to be 30 W/m^2 and 21 W/m^2, respectively, to allow for low efficiency lighting and task lighting, and information technology closets on each floor.

Four sampling plans (parametric, full factorial, Monte Carlo, and Latin hypercube) were initially considered, but only Monte Carlo and Latin hypercube were deemed appropriate. Parametric sampling plans, where one building feature is modified in the energy simulation at a time, do not capture the behaviour of multiple building attributes interacting. Simulating all combinations using full factorial sampling was not feasible given the number of variables, even if the ranges summarized in Table 2 were discretized into bins. Latin hypercube was selected as the sampling strategy for this research due to its space-filling properties.

This study used 4,000 samples, representing random combinations of the values within the ranges set for the 71 input features. The Latin hypercube sample set for the 4,000 x 71 matrix of selections between 0 and 1 was generated using MATLAB's *lhsdesign* [62]. By multiplying the Latin hypercube sampled values by the feature range standard deviation and adding the feature minimum values, the feature values for each of the 4,000 samples were generated. Features requiring integer values, such as number of storeys, were assigned the rounded value.

The use of Latin hypercube sampling to generate the input matrix resulted in uniform distribution of samples for each input feature across the 4,000 samples. As an example, Figure 17 shows a histogram of boiler efficiency for the full data set. The histogram shows that the sample values are uniform across the feature range.

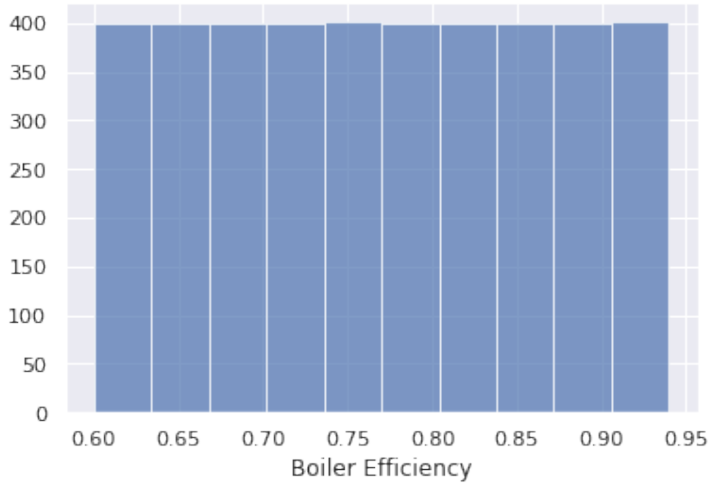

Figure 17 – Histogram of boiler efficiency input variable for full dataset

Minor changes to the DOE large office existing building commercial reference model [61] base model were made to facilitate surrogate model development. These included: changing the holidays and site data to represent Toronto, Canada; adding three additional window and wall types to allow for these features to be modified independently for each building elevation; and adding an insulation layer to the slab-on-grade construction.

As per the DOE Reference model for large office buildings used as the base model, the mechanical system was assumed to be a central plant with chiller and boiler, and multi-zone variable air volume with reheat distribution system.

The 71 building features affected 1,756 EnergyPlus Input Data File (IDF) objects in the base model. To create IDFs for a range of building sizes, extensive IDF object transformation was required. The building geometry was represented in the IDF as x, y and z coordinates at each vertex. A geometry transformation script was prepared to transform the 25 geometry features to 1,704 IDF objects.

Using *Eppy* [67], a Python library developed for manipulating EnergyPlus IDFs, the base EnergyPlus IDF was modified by identifying the location of the 1,756 objects in the text file and replacing them with the values generated by the Latin hypercube sampling plan. This was automated to generate the 4,000 IDFs.

The simulations were run using EnergyPlus v. 8.0.0.008 [60]. The 4,000 .idf files were divided into four batches and transferred to the system using Remote Desktop Protocol where a Powershell script ran the RunEPlus.bat [60] batch script on each individual file while assigning each process to one of the 4 available CPUs. The annual building energy use for each simulation was parsed from the output files using a Python script to create a simulated output matrix, as illustrated in Figure 18.

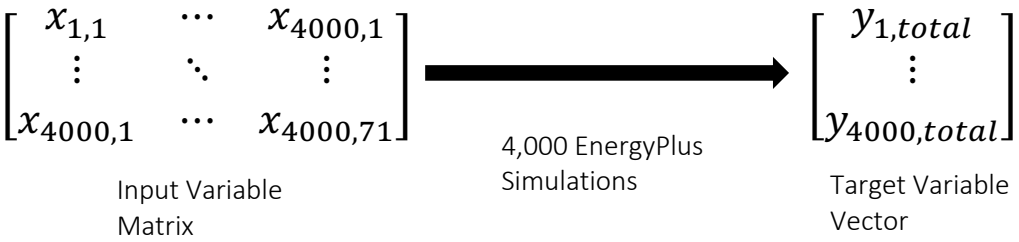

Figure 18 – Input and target variable matrix/vector form

The annual building energy use, used as the target variable in this study, represented the building energy use and did not include the simulated direct energy use governed by the tenant energy loads. For this study, the simulated lighting, interior electricity, and domestic hot water energy use were not included in the annual building energy use target variable. The behaviours of the light and electrical plug power density input features were captured in the surrogate model based on their impact to the building heating and cooling systems. As illustrated in Figure 19, the annual building energy use (i.e. the target variable) is the sum of the annual heating, cooling, fan and pump energy.

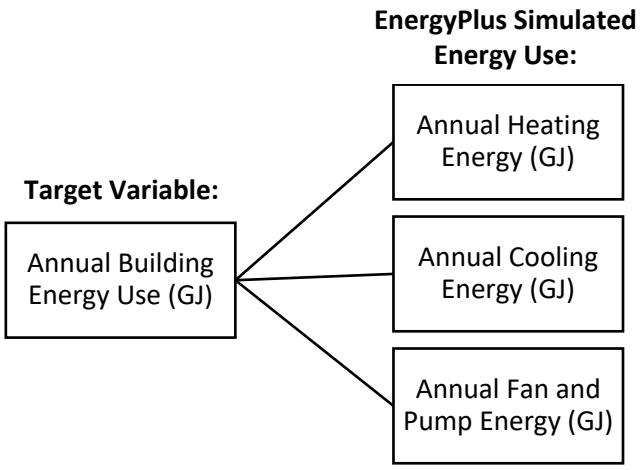

Figure 19 – Breakdown of annual building energy use target variable

3.1.2 Feature Selection and Trained Model Development

The full data set was randomly divided into 2 sets; training/validation and testing. A random seed number was applied to the random split so that the dataset would be consistently randomly split into the same sets using the Python function *sklearn.model_selection.train_test_split* [63]. Of the total 4,000 samples, the training/validation set was 85% (3,400 samples) and testing set was 15% (600 samples).

The training/validation set was further split into a training set of 2,800 samples (70% of the full dataset) and validation set of 600 samples (15% of the full dataset) using random splitting. The

training/validation dataset was sampled 10 times with a random split of data to create 10 training and validation sets. Random seed values 0 through 9 were used for the splitting so that the 10 training and validation datasets would be consistent for the full study. Where the mean and standard deviation model performance metrics are presented, the 10 training and validation sets were used. Where a single model performance metric and model performance graphs are presented in the results section, the training and validation datasets using random seed 0 was used. The Python function, *sklearn.model_selection.train_test_split* [63], was again used for the random split.

For each model training iteration, the training set was used to train the regression model, and the validation set was used to select the appropriate feature and target transformations, select features in the forward stepwise regression, and tune the learning algorithm parameters. Once the highest performing model was selected, the test set was used to evaluate the performance of the final model.

The input matrices for each training set were normalized so that the features were centered with a mean of zero and a variance of 1.0. Rescaling the input matrix causes the features to be on a similar scale and thereby reduces the risk of the model inaccurately assigning importance to the input feature values. This also allows for gradient descent to converge in less steps. The validation and test data sets were normalized with the associated training set mean and standard deviation.

This study evaluated Box-Cox, exponential, quadratic, squared-root and squared transformations of the target variables and compared the model performance for each. The probability distributions for the transformed target variables were analyzed to confirm normal distribution. The model residual plots for the transformed target variables were analyzed to confirm non-skewed distribution. The Box-Cox, exponential (log), and quadratic target variable transformations

produced models with the greatest fit. Therefore, these target variable transformations, in combination with logarithmic transformation of the input features, were evaluated and compared in more detail. Multivariate regression using gradient descent with a mean squared error cost function was used for the data transformation evaluation. Alpha values of 0.001, 0.01 and 0.1 were evaluated. The cost, $C(\beta)$, at each iteration was plotted to confirm that the cost reached zero, and therefore reached the minima, within the number of iterations performed. The remainder of the analysis in this study used the transformed input and target variables selected from this step.

Forward stepwise selection, a wrapper feature selection technique, was used to select combined features that, when added to the input feature set, improved the surrogate model performance. In this study additional combined features were developed to represent linear combinations of several of the original 71 features. A summary of the combined features is presented in Table 3.

The combined features were added to the model one at a time starting with the variables with highest absolute Pearson correlation coefficient to the transformed annual energy use and with a two-tailed p-value less than 0.005. As each combined feature was added, the original features in the combined feature were removed from the model. The model accuracy was evaluated at each step. Combined features that either improved or maintained (within 2 decimal points) the accuracy of the model were kept in the feature set.

Table 3 – Combined Features. Calculations are in Appendix B.

	Combined Features		
Building Geometry	Conditioned Floor Area (m^2)	Above-Grade Building Enclosure Surface Area – including roof (m^2)	Overall Above Grade Vertical Enclosure Window-to-Wall Ratio
	Overall Building Height (m)	Enclosure Surface Area to Volume Ratio (m^2/m^3)	Opaque Wall Area – Elevation 1 (m^2)
	Above Grade Building Height (m)	Aspect Ratio – Depth to Width (m/m)	Opaque Wall Area – Elevation 2 (m^2)
	Weighted Average Floor Height (m)	Above Grade Window-to-Wall Ratio – Elevation 1	Opaque Wall Area – Elevation 3 (m^2)
	Weighted Average Plenum Height (m)	Above Grade Window-to-Wall Ratio – Elevation 2	Opaque Wall Area – Elevation 4 (m^2)
	Conditioned Volume (m^3)	Above Grade Window-to-Wall Ratio – Elevation 3	
	Overall Enclosure Surface Area – including roof and below grade (m^2)	Above Grade Window-to-Wall Ratio – Elevation 4	
Enclosure Performance	Area Weighted Wall and Window U-Value – Elevation 1 (W/m^2K)	Overall Area Weighted Wall and Window U-Value (including below grade walls) (W/m^2K)	Overall Area Weighted Glazing Solar Heat Gain Coefficient
	Area Weighted Wall and Window U-Value – Elevation 2 (W/m^2K)	Overall Enclosure Area Weighted U-Value (W/m^2K)	Overall Area Weighted Glazing Visible Transmittance
	Area Weighted Wall and Window U-Value – Elevation 3 (W/m^2K)		
	Area Weighted Wall and Window U-Value – Elevation 4 (W/m^2K)		
Internal Loads	Area Weighted Lighting Power Density (W/m^2)	Area Weighted Electrical Plug Power Density (W/m^2)	Area Weighted Occupancy Density ($person/m^2$)

Next, a process was completed to evaluate whether additional building features could be removed from the model without significantly compromising the model accuracy. This was done to simplify the model, leading to less inputs required for the surrogate model end-user and the potential for lower model overfitting. This step was completed using an embedded feature selection technique called least absolute shrinkage and selection operator (LASSO). A range of shrinkage parameters (λ) along with combining LASSO with a similar regularization term, called Ridge, into a method

called Elastic Net were evaluated to determine the optimal set of building features and model parameters based on the model performance. This step determined the final surrogate model.

The final surrogate model was evaluated using the remaining 600 samples, referred to as the test dataset. The test dataset was used to evaluate how the model performed on samples not seen at any stage of the model development process. Since the intent of the surrogate model was to predict annual building energy use in GigaJoules (GJ), and not in transformed GJ, the test dataset was transformed back before evaluating the model prediction accuracy.

3.2 Results

The 4,000-sample dataset was developed by generating random samples within the design space using Latin hypercube and then simulating in EnergyPlus to determine the annual building energy use. As a first step, the dataset was evaluated by comparing the simulated annual energy use to published real building energy use data for Canadian office buildings to validate that the sample set captured the range of building energy use expected for real operational performance.

The results of the statistical analysis leading to the final surrogate model development are summarized in this section. Figure 20 illustrates the steps to developing the final surrogate model. This figure is used throughout the results section to highlight the stage of the surrogate development process being evaluated and discussed.

Figure 20 – Surrogate model training process

A training set (2,800 samples) was used at each step of the analysis to fit the data to the model. A validation set (600 samples), comprised of samples not included in the training dataset, was used to evaluate the model. For steps where multiple strategies were compared, ten training and

validation sets were used so that the mean and standard deviation of the model performance could be evaluated and compared. These ten datasets were kept consistent using random seed numbers to split the dataset. Initially, multivariate linear regression was used to evaluate how the data fit to a linear model. Subsequently, the data was transformed using common transformation techniques and the transformed data was fit to a multivariate regression model. The transformation with the best fit was selected and carried through the remainder of the analysis. Since several of the 71 building features used described portions of the building, such as lighting power density for the basement, ground floor, repeated floors core and perimeter, and top floor, these building features were combined to describe each building attribute for the entire building. To evaluate if the combined feature improved the model performance, the combined features were added to the model, one at a time, in order of the highest correlation to the annual building energy use, in GJ. The original building features used to calculate the combined feature were removed from the dataset. If the combined feature improved or maintained the model performance, it was kept in the model and the associated original features removed. The result of this forward stepwise feature selection process was evaluated by comparing the model performance to the model performance with the original 71 variables.

After the optimal set of combined features were added, embedded feature selection using the L1 regulator was carried out. Both LASSO and Elastic Net were evaluated and compared to evaluate how each performs on the data sets and to determine the optimal shrinkage parameter value.

The final selected surrogate model's predictions compared to the EnergyPlus simulations were evaluated using the re-transformed data so that all predictions were in annual building energy use units (GJ).

3.2.1 Energy Modeling Data

The feature ranges selected for the simulations were developed to encompass building attributes applicable to high and low performing large office buildings in Toronto. The simulated building energy use resulted in annual energy use intensity (EUI) ranges of 0.33 to 2.69 GJ/m^2 for a range of conditioned floor area between 6,570 to 1,780,000 m^2. A histogram of the annual EUIs is shown in Figure 21 and a breakdown of the annual EUIs by end-use is shown in Figure 22 for the full dataset. For the majority of the samples simulated, the annual heating energy and annual lighting and electrical plug electricity demands made up a significant portion of the overall annual energy use. The annual heating energy was the most significant portion of the annual building energy use on average but also has the largest variation across the 4,000 data points. Compared to average commercial building data for Canada, this large heating energy use across the sample set was representative of existing buildings in Canada [4]. The cooling energy contributes little to the overall annual energy use, with almost all samples having an annual cooling energy use intensity less than 10% of the total annual energy use intensity.

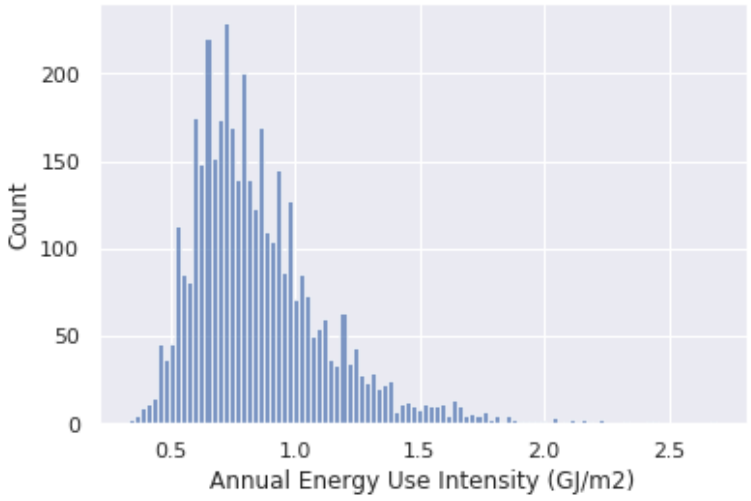

Figure 21 – Histogram plot of annual energy use intensity for full dataset

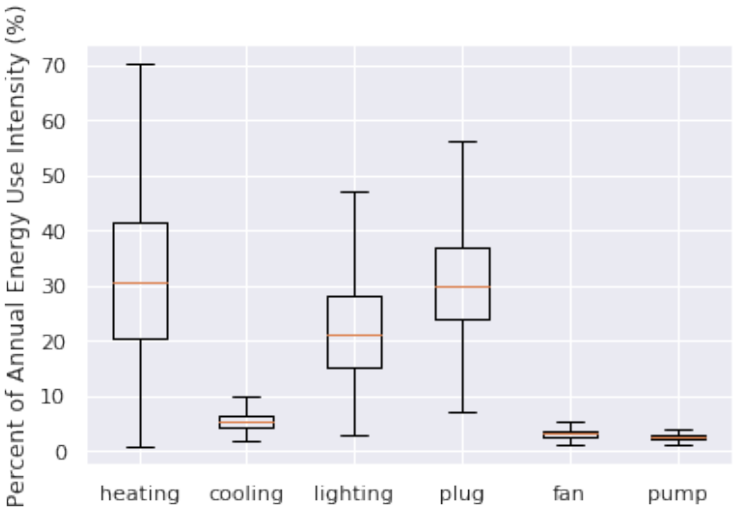

Figure 22 – Breakdown of energy use intensity as a percentage of the annual site energy use intensity for the full dataset. Boxplot shows the data within the first to third interquartile ranges with the outliers not shown.

According to the 2014 Natural Resources Canada Survey of Commercial and Institutional Energy Use [68], the average annual EUI for non-medical office buildings between 4,646 and 18,580 m^2 was 1.16 GJ/m^2 and over 18,580 m^2 in Canada was 1.09 GJ/m^2. The 2009 *Geared for Change Report – Energy Efficiency in Canada's Commercial Building Sector* [4], reported that the average annual EUI for all Canadian commercial buildings constructed pre-1920 to 2004 ranged from 1.3-1.8 GJ/m^2. Based on the dataset of 450 commercial office buildings in Canada used for the 2019 Sidewalk Labs Canadian Commercial Office Buildings Study: Analysis of Energy Use and Performance report [66], the average annual EUI for Canadian commercial office buildings was 1.06 GJ/m^2. The Sidewalk Labs report also presented a normalized EUI value that standardizes for weather, vacancy, occupant density and exceptional loads like data centres and retail. The average annual EUI reported using the normalized values was 0.85 GJ/m^2.

Comparing the EUIs for the simulated dataset used in this study to available EUI data for commercial office buildings indicated that the majority of the samples simulated had lower energy use per conditioned floor area than the national average and were therefore higher performing.

This is explained by the use of Latin hypercube sampling, which resulted in the generation of models using any combination of building features within the specified ranges, which in this case resulted in a sample set that had higher average performance than Canadian existing buildings. The goal of the dataset generation was not to create a set of simulations that were representative of commercial office buildings in Toronto but to generate samples that represent the design space so that actual combinations of building features can be interpolated. The ability of the surrogate model to predict energy use for building feature combinations of an actual Toronto commercial office building is explored in Section 3.3.

For this study, a surrogate prediction model for annual building energy use was developed. The lighting and electrical plug annual energy use, along with domestic hot water, were not included in the annual building energy use target variable to be predicted by the surrogate model. Figure 23 shows a boxplot of the annual energy use included in the target variable as a percent of the annual building energy use. Taking a subset of the total annual energy use was done to remove energy use that was directly related to building occupant behaviours and annual energy use that could be calculated linearly by the building features. For example, the energy use associated with lighting is calculated in EnergyPlus using the light power density for the schedules, multiplied by the building floor area. This relationship can be fit using a linear model. By removing these end-use energy uses from the model, only the more complex relationship between the building features and the annual heating, cooling, fan and pump energy was determined. In Section 3.4, the developed surrogate model is compared to the surrogate model presented in [69] that predicted annual energy use for all end-uses.

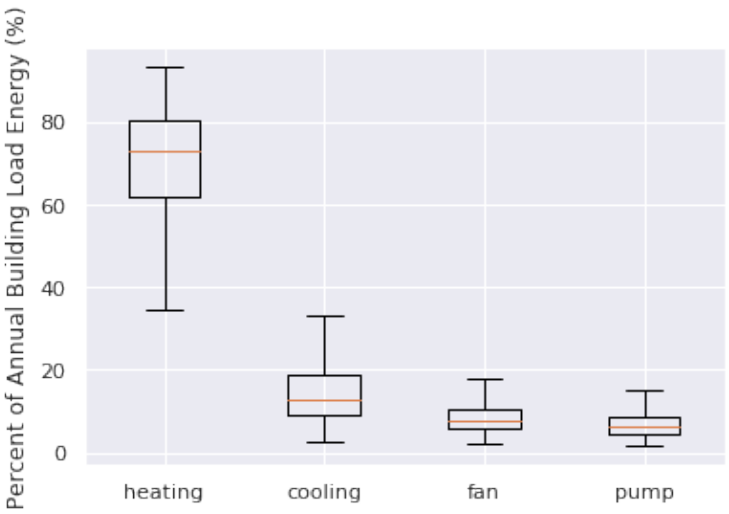

Figure 23 – Breakdown of annual energy use (GJ) as a percentage of the annual building energy use for the full dataset. Boxplot shows the data within the first to third interquartile ranges with the outliers not shown.

3.2.2 Multivariate Linear Regression with Original Feature Set

Multivariate linear regression was used to analyze how the data fit to a linear model. This is a common first step for many surrogate modelling researchers as the resulting model can inform if the data has a linear behaviour and produces a model that is easily interpretable.

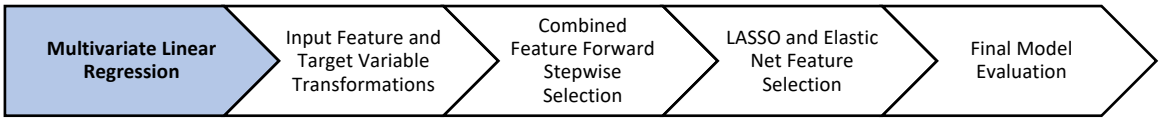

Figure 24 – Surrogate model training process – Multivariate linear regression

When the training dataset with 71 features was fit to a multivariate linear regression model, the model had poor performance on both the training and validation data, as summarized in Table 4. Further, as shown in the predicted (surrogate model) vs. actual (EnergyPlus simulated) (Figure 25) and residual (Figure 26) plots for the validation dataset, there was clear evidence of non-linearity or missing features in the data as evidenced by the non-random inverted-U shape in both plots and particularly the residual plot. The residual plot shows the model predicted target versus the difference between the EnergyPlus simulated and surrogate model predicted target values. Ideally,

the datapoints on a residual plot are evenly distributed above and below the line of zero-residual for the full range of model predicted values. The linear model performed best in the mid-range of the annual building energy use but had significantly reduced accuracy in the low and high ranges, with negative energy use predictions at the low energy use range. The negative energy use predictions are indicative of the model underpredicting the annual building energy use. This type of model behaviour can indicate either that the input features are not adequately describing the behaviour of the data across the full design space and/or that there are non-linear relationships between the input and target that are not captured in the multivariate linear regression model.

Table 4 – Results of multivariate linear regression model with original feature set

		R^2	RMSE	MAE
Annual Building Energy Use (GJ)	Training	0.768	46121.109	30366.171
	Validation	0.745	48631.631	30883.871

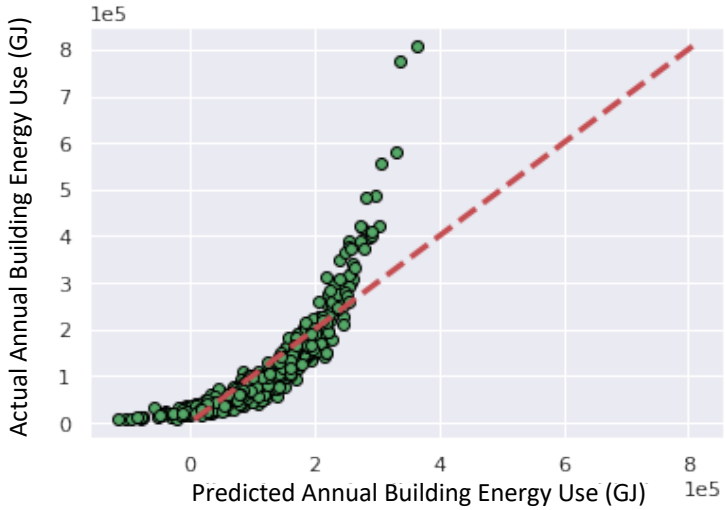

Figure 25 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) annual building energy use for the validation dataset using the multivariate linear regression model

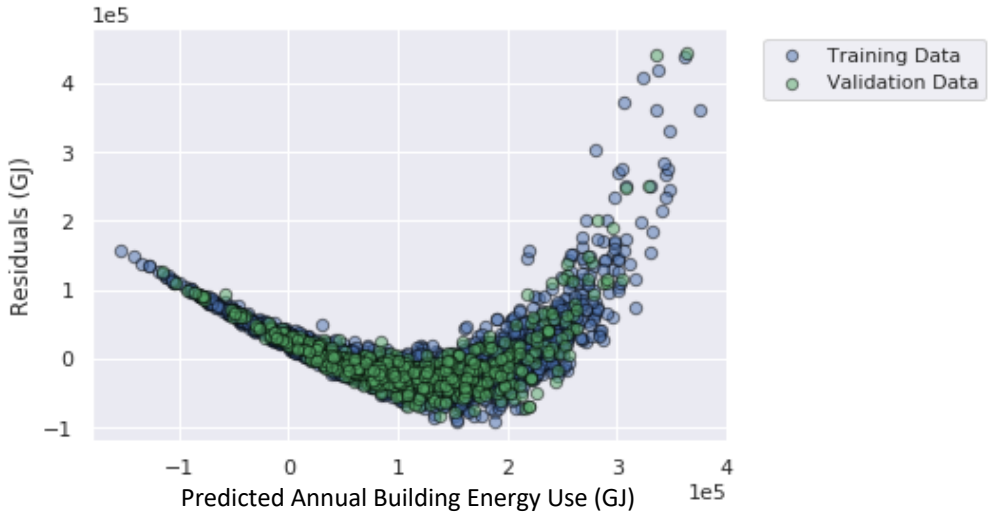

Figure 26 – Annual building energy use residual plot for using multivariate linear regression model showing non-linear behaviour

3.2.3 Transforming Input Features and Target Variable

Since the data exhibited non-linear behaviour when fit using multivariate linear regression, several transformations of the input features and the target variable were analyzed. Transformations of the input features and target variable were completed separately and together and the transformations that resulted in the best model prediction performance, fit using multivariate regression, were selected.

Figure 27 – Surrogate model training process – input feature and target variable transformations

Several common input feature and target variable transformations were evaluated as summarized in Table 5. The coefficient of determination, R^2, for the multivariate regression models for each transformation was calculated and the transformations with the highest R^2 values from each of the target variable and input feature transformations were identified for further analysis. The target variable transformations bolded and highlighted in blue and the input feature transformations

bolded and highlighted in orange in Table 5 led to the highest model prediction accuracy and were therefore further analyzed.

Table 5 – Summary of input and target variable transformations. Transformations with the highest model performance are bolded and highlighted in blue and orange for target and input features, respectively.

Variable	Transformation	Transformation Equation	Training R^2	Validation R^2
Target	Quadratic	$y_i' = \sqrt{y_i}$	**0.9134**	**0.9080**
	Squared	$y_i' = y_i^2$	0.417	0.385
	Reciprocal	$y_i' = \frac{1}{y_i}$	0.648	0.638
	Exponential	$y_i' = \log(y_i)$	**0.953**	**0.954**
	Box-Cox	$y_i' = \frac{(y_i^\lambda - 1)}{\lambda}$	**0.958**	**0.949**
Input	Logarithmic	$x_i' = \log(x_i)$	**0.730**	**0.701**
	Squared	$x_i' = x_i^2$	**0.742**	**0.734**

Of the target variable transformations evaluated, the quadratic, exponential and Box-Cox transformations (highlighted in blue) performed well. The model prediction results for the quadratic transformation showed that although the transformation reduced the skewing of the predicted annual building energy use at the high and low ends (Figure 28) compared to the untransformed model (Figure 25 and Figure 26), the residual plot showed a non-uniform inverted U-shape (Figure 29).

Figure 28 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) quadratic transformed annual building energy use for the validation dataset

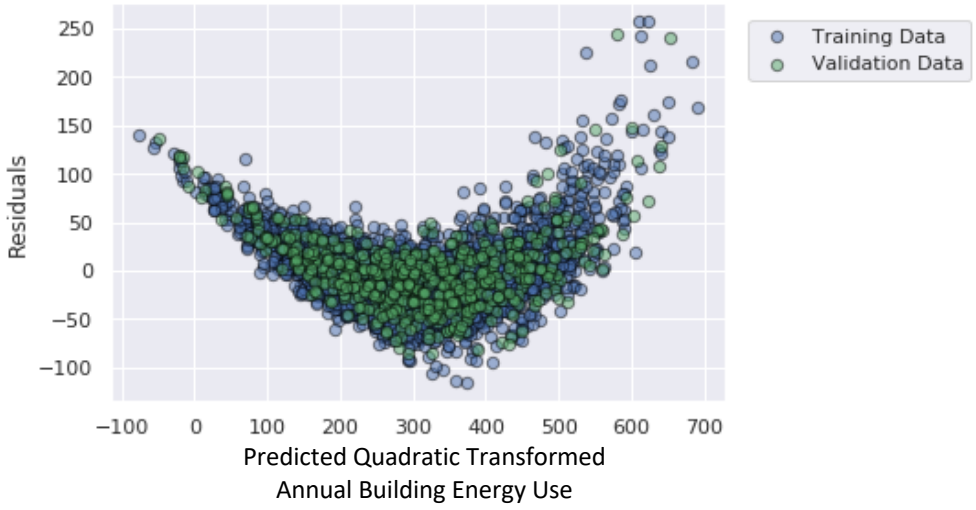

Figure 29 – Quadratic transformed annual building energy use residual plot for multivariate regression model

Figure 30 shows the histograms of the annual building energy use for the untransformed and exponential and Box-Cox transformed target variable. The exponential transformation shifts the data so that it is more normally distributed. The Box-Cox transformation, as it is intended to do, shifts the data so that it is normally distributed. Multivariate regression is based on the assumption that the target variable is normally distributed so transforming the data to achieve normal distribution, is essential to the performance of the model. The lambda, λ, value for the Box-Cox

transformation is 0.1019 +/- 0.0056 using the 10 training datasets. When lambda is zero, the Box-Cox transformation is logarithmic. This explains the target variable distribution similarities between the exponential and Box-Cox transformations.

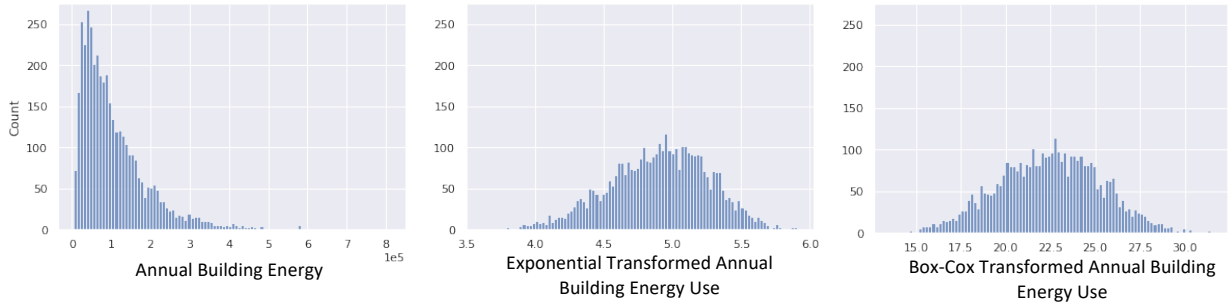

Figure 30 – Histograms of original, exponential transformed, and Box-Cox transformed target variables

The Box-Cox and exponential transformed target variables both showed significantly less skewing on the residual plots (Figure 31 and Figure 32 respectively). The Box-Cox transformed target variable model residual plot, Figure 31, showed that there was still minor skewing at the high and low ends of the target variable. By contrast, the exponential transformed target variable model residual plot, Figure 32, showed the predicted target values fit well with the actual target values across the full design space. It also showed that the residuals are generally uniformly distributed above and below the zero-residual line across the full design space. This is the ideal behaviour of the model prediction.

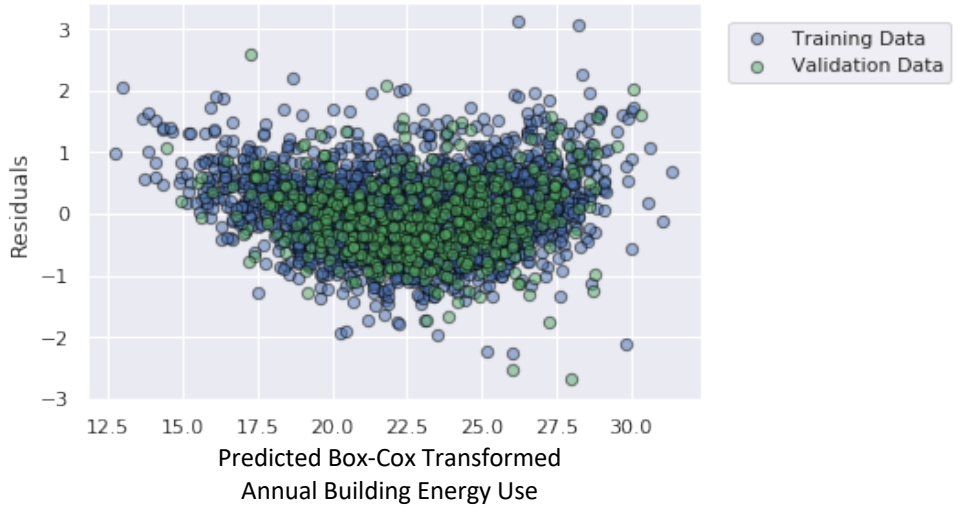

Figure 31 – Box-Cox transformed annual building energy use residual plot for multivariate regression model

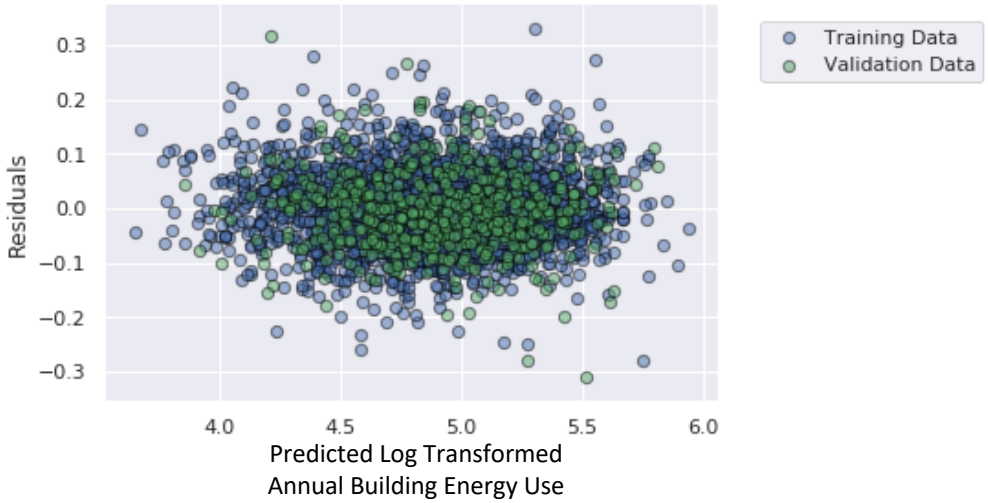

Figure 32 – Exponential (log) transformed annual building energy use residual plot for multivariate regression model

The exponential and Box-Cox target variable transformations and the logarithmic and squared input feature transformations were compared and evaluated in more detail to determine the transformation(s) that produced a multivariate regression model with the best fit. The average coefficient of determination, normalized root mean squared error, normalized mean absolute error, and their respective standard deviations across the ten training and validation datasets for the transformations are shown in Table 6. As the values and ranges of the transformed targets were

different for Box-Cox and exponential transformations, the transformed target variables were normalized prior to training the model to have a mean of zero and a standard deviation of 1.0 so that RMSE and MAE metrics could be used to compare the models.

Table 6 – Summary of input and target variable transformation model performance (selected transformations in bold italics)

ANNUAL BUILDING ENERGY USE			TARGET TRANSFORMATION							
			EXPONENTIAL				BOX-COX			
			Training		Validation		Training		Validation	
			MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
INPUT TRANSFORMATION	ORIGINAL	R^2	0.9539	0.0008	0.9516	0.0038	0.9570	0.0007	0.9542	0.0037
		NRMSE	0.2147	0.0018	0.2250	0.0112	0.2074	0.0018	0.2192	0.0104
		NMAE	0.1703	0.0014	0.1781	0.0088	0.1634	0.0014	0.1727	0.0074
	LOG	R^2	***0.9666***	***0.0005***	***0.9643***	***0.0027***	0.9619	0.0006	0.9585	0.0026
		NRMSE	***0.1827***	***0.0015***	***0.1908***	***0.0106***	0.1954	0.0018	0.2006	0.0061
		NMAE	***0.1411***	***0.0009***	***0.1474***	***0.0073***	0.1512	0.0010	0.1552	0.0032
	SQUARED	R^2	0.8763	0.0020	0.8705	0.0103	0.8846	0.0021	0.8802	0.0093
		NRMSE	0.3517	0.0028	0.3596	0.0143	0.3396	0.0031	0.3436	0.0103
		NMAE	0.2811	0.0024	0.2859	0.0127	0.2711	0.0025	0.2742	0.0072

The results show that the logarithmic transformed input variables along with the exponential transformed target variable resulted in the highest performing model for all model prediction evaluation metrics evaluated. The model with logarithmic transformed input variables along with the Box-Cox transformed target variable performance is close in model performance. However, due to the slight skewing of the Box-Cox transformed residual plot shown in Figure 31, the logarithmic input variable transformation with the exponential target variable transformation was selected and applied to the data for the remainder of the study.

3.2.4 Adding Combined Feature Terms Using Forward Stepwise Selection

Many of the original 71 features were broken out to describe portions of the building that generally have different conditions in real constructed large office buildings. The initial goals of developing the dataset using this method were to allow the user of the surrogate model to modify the building attribute per area and to determine if the feature significance and model behaviour differs when the building attributes are separated by area as multiple features. For example, a designer may decide to improve the enclosure on a single elevation differently from the others such as selecting window glazing coatings with differing solar heat gain coefficients depending on the building elevation. However, having many features describing the same item for different parts of the building may lead to the features being removed from the model as they are not as important as if the item is looked at for the full building. This step evaluated if the model benefited from these features being separate or combined.

Figure 33 – Surrogate model training process – combined feature forward stepwise selection

As the combined feature terms were added to the model, one at a time, in order of the highest Pearson correlation coefficient (and with an associated two-tailed p-value less than 0.005), the original features used to calculate the added combined feature were removed from the model. The equations along with the original input features used to calculate each combined feature are summarized in Appendix B. R^2, RMSE and MAE were calculated at each forward selection step for the training and validation datasets and if the error evaluation metrics indicated an improved or maintained (within 2 decimal points) model prediction for each the training and validation data sets, the combined feature was kept in the model.

Table 8 summarizes the combined features in order of the highest to lowest Pearson correlation coefficient, and therefore the order added to the model using forward stepwise selection, and whether the combined feature was kept in the model following forward stepwise selection.

The optimal model used 22 combined features (calculated from 48 of the original features) and 23 of the original features. This method of adding terms to linearly combine the original features reduced the input matrix feature size from 71 to 45 and did not compromise the model accuracy. On the contrary, the combined features increased the coefficient of determination, and decreased the RMSE and MAE on the training and validation sets as summarized in Table 7.

Table 7 - Model performance before forward stepwise selection, using the original 71 features, and following forward stepwise selection

Model Performance Metric	Dataset	Original 71 Features	Following Forward Stepwise Selection – 22 Combined Features and 23 Original Features
R^2	Training	0.9666 +/- 0.0005	0.9721 +/- 0.0005
	Validation	0.9643 +/- 0.0027	0.9710 +/- 0.0022
RMSE	Training	0.0689 +/- 0.0004	0.0615 +/- 0.0004
	Validation	0.0689 +/- 0.0019	0.0620 +/- 0.0018
MAE	Training	0.0519 +/- 0.0002	0.0474 +/- 0.0003
	Validation	0.0533 +/- 0.0009	0.04780 +/- 0.0012

The predicted versus actual data plot for the validation dataset and the residual plot for both the training and validation datasets are shown in Figure 34 and Figure 35, respectively. The residuals fall within a narrower band following forward stepwise selection indicating that the model is better at predicting the target throughout the design space.

Table 8 – Combined features in order of highest to lowest Pearson correlation coefficient. Last column indicates whether the combined feature was kept in the model.

Combined Feature	Pearson Correlation Coefficient	Two-Tailed P-Value	Kept in Model
Overall Building Enclosure Surface Area – Including Roof and Slab-on-Grade (m^2)	**0.74288498**	**0**	
Above Grade Building Enclosure Surface Area – Including Roof (m^2)	**0.74242409**	**0**	✓
Conditioned Volume (m^3)	**0.72173672**	**0**	✓
Conditioned Floor Area (m^2)	**0.70863331**	**0**	✓
Opaque Wall Area – Elevation 4 (m^2)	**0.56160455**	**1.60E-232**	
Opaque Wall Area – Elevation 2 (m^2)	**0.54850059**	**9.78E-220**	
Overall Building Height (m)	**0.54462322**	**4.63E-216**	✓
Above Grade Building Height (m)	**0.54435992**	**8.19E-216**	
Opaque Wall Area – Elevation 1 (m^2)	**0.53036274**	**5.98E-203**	✓
Opaque Wall Area – Elevation 3 (m^2)	**0.52810868**	**6.20E-201**	✓
Enclosure Surface Area to Volume Ratio (m^2/m^3)	**-0.4598175**	**1.56E-146**	✓
Overall Enclosure Area Weighted U-Value (W/m^2K)	**0.15995443**	**1.66E-17**	✓
Overall Area Weighted Wall and Window U-Value (including below grade walls) (W/m^2K)	**0.12510145**	**3.08E-11**	✓
Overall Above Grade Vertical Enclosure Window-to-Wall Ratio	**0.12190721**	**9.67E-11**	✓
Overall Area Weighted Glazing Solar Heat Gain Coefficient	**0.1206095**	**1.53E-10**	✓
Area Weighted Electrical Plug Power Density (W/m^2)	**-0.1178068**	**4.03E-10**	✓
Weighted Average Floor Height (m)	**0.10524917**	**2.37E-08**	✓
Area Weighted Wall and Window U-Value (includes below grade walls) – Elevation 1 (W/m^2K)	**0.10032451**	**1.04E-07**	✓
Area Weighted Wall and Window U-Value (includes below grade walls) – Elevation 3 (W/m^2K)	**0.08899543**	**2.40E-06**	✓
Overall Area Weighted Glazing Visible Transmittance	**0.08802812**	**3.09E-06**	✓
Area Weighted Wall and Window U-Value (includes below grade walls) – Elevation 4 (W/m^2K)	**0.07592115**	**5.79E-05**	✓
Above Grade Window-to-Wall Ratio – Elevation 1	**0.07458394**	**7.80E-05**	✓
Area Weighted Wall and Window U-Value (includes below grade walls) – Elevation 2 (W/m^2K)	**0.06922872**	**0.00024637**	✓
Area Weighted Occupancy Density ($person/m^2$)	**-0.0673763**	**0.00036016**	✓
Above Grade Window-to-Wall Ratio – Elevation 3	**0.06182926**	**0.00106274**	✓
Above Grade Window-to-Wall Ratio – Elevation 2	**0.06165244**	**0.00109855**	✓
Aspect Ratio – Depth to Width (m/m)	-0.0452357	0.01667462	
Area Weighted Light Power Density (W/m^2)	0.04522417	0.01670241	
Above Grade Window-to-Wall Ratio – Elevation 4	0.03950961	0.0365698	
Weighted Average Plenum Height (m)	0.00326965	0.86270102	

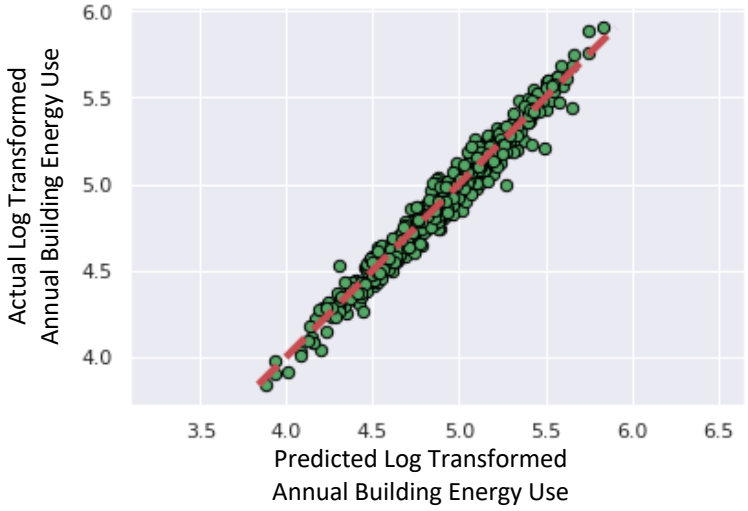

Figure 34 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) for the multivariate regression model following forward stepwise selection. Target is log transformed annual building energy use and data presented is from the validation dataset.

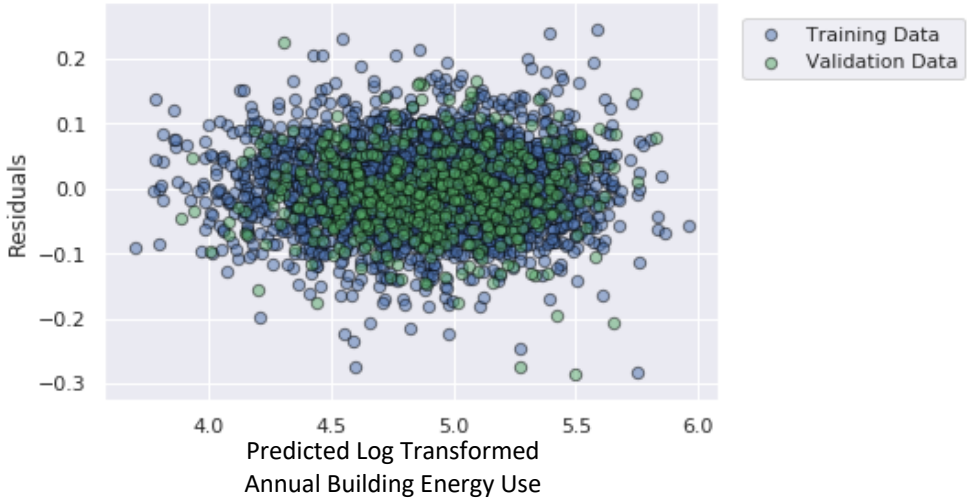

Figure 35 – Exponential (Log) transformed annual building energy use residual plot for multivariate regression model following forward stepwise selection

3.2.5 Embedded Feature Selection Using LASSO and Elastic Net Regulators

The LASSO (L1 regulator) and Elastic Net (L1 and L2 regulators) embedded feature selection methods were used to determine if any of the 45 features in the model following forward stepwise selection could be removed without compromising the accuracy of the model. Both methods were evaluated by testing a range of λ values, which resulted in varying feature subset sizes, feature

coefficients, and model accuracies. 100% LASSO, 75% LASSO and 25% Ridge, 50% LASSO and Ridge, 25% LASSO and 75% Ridge, and 100% Ridge were evaluated for a range of λ values.

Figure 36 – Surrogate model training process – LASSO and Elastic Net feature selection

Figure 37 shows the LASSO, Elastic Nets and Ridge regression models' coefficient of determination, R^2, on the validation data as λ values were varied. The figure shows that when fewer than 20 features were included in the model, the 100% LASSO model performed significantly better than the other models and as more features were included in the model, the model accuracy converged. At 23 features, the model performance reached a plateau when over 50% LASSO was used. The model performance gradually increased as more features were included in the model but the model performance difference between 23 features and 45 features on the validation dataset was 0.005 for each R^2, RMSE and MAE. Such minor increase in model performance was not worth the added complexity of 22 additional features, which may lead to overfitting and will require the model end-user to determine and input more information about the building they intend to model. The LASSO, 100% L1 regulator, model was selected for its higher performance with a lower number of features and its slightly higher performance at 23 features.

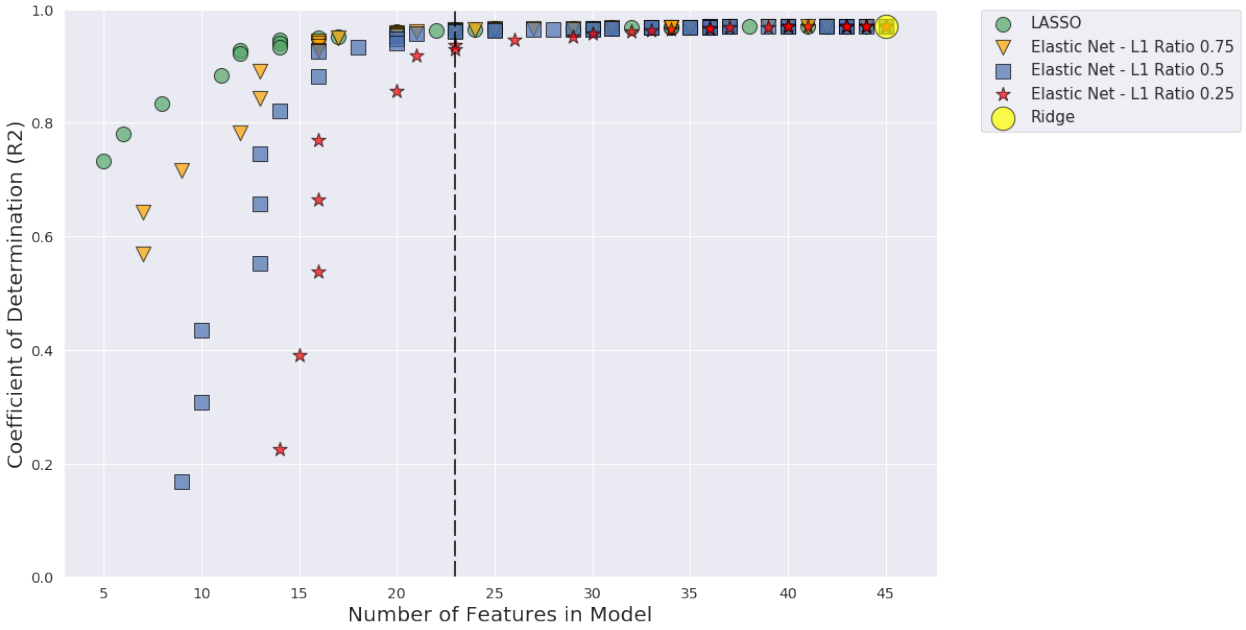

Figure 37 – Model coefficient of determination as the number of features in each the LASSO and Elastic Net models change. The selected number of features is shown by the dashed line.

Insight into the behaviour of the dataset and final model can be gained by evaluating the order in which the features were removed from the model, as shown in Figure 38. The features highlighted in red identify the features removed.

Interestingly, the outside air rate, the amount of ventilation provided per person, was removed from the model early on. Outside air rate is often thought of as contributing significantly to energy use as the outdoor air must be conditioned before it is supplied to the spaces. Nagpal et al. [24] was the only other study which included outdoor air flowrate as an input feature and they did not present the ranking or values of the coefficients (since it was not the focus of their study). Therefore, the comparison of this result to another study could not be completed.

The temperature setpoints are all kept in the model. However, the supply air temperature for heating was removed early and the supply air temperature for cooling remained in the model. This is likely due to the additional latent energy required to cool air versus heat air.

Figure 38 – Features removed from model (highlighted in red) as penalty weight, λ, increases

The feature set selected, and associated coefficients, for the final model are summarized in Table 9. These features are listed in order of importance based on their coefficient weight values for the LASSO model. The same 23 features were selected for each of the LASSO and Elastic Net models tested. The order of feature importance was not the same, however, and is also summarized in Table 9. The LASSO model coefficients presented in Table 9 are the mean and standard deviation (SD) values for the ten training sets. The low standard deviation values are an indication that the coefficients are stable and not significantly changing between different training sets. Some of the standard deviations are significantly higher than others. All features with higher standard deviations are combined features and include: above grade building enclosure surface area, volume, conditioned floor area, overall enclosure area weighted U-value, overall building height, and overall area weighted wall and window U-value (including below grade). A higher coefficient standard deviation can be an indicator of collinearity or multicollinearity between the variables. Collinearity is not an issue for the performance behaviour of the model but can impact the feature importance interpretation for these features. This behaviour is further explored later in this section.

It is interesting to observe that the above-grade building enclosure surface area had almost twice the coefficient value as the next most important feature. Also, both the temperature setpoints and setbacks for heating and cooling had a high significance to the annual building energy use prediction. This is consistent with the findings of Hoyt et al. [70] who found that increasing the cooling set point to 25 from 22.2 reduced the annual cooling energy use by 29% and decreasing the heating set point to 20 from 21.1 reduced the annual heating energy use by 34% in medium office buildings across multiple climates.

*Table 9 – Features selected for final LASSO model in order of coefficient values. The order of feature importance for the Elastic Net models where 23 features were selected is presented. * indicates a combined feature*

Feature	Coefficient		Order of Feature Importance			
	MEAN	SD	Lasso L1 – 100%	EN L1 – 75%	EN L1 – 50%	EN L1 – 25%
Above-Grade Building Enclosure Surface Area – including roof (m^2)*	0.204	0.0063	1	1	1	1
Temperature Setpoint - Heating Occupied (°C)	0.105	0.0005	2	2	2	2
Supply Air Temperature Cooling (°C)	0.069	0.0005	3	3	3	4
Infiltration Rate - Ventilation System Off ($m^3/s/m^2$)	0.059	0.0003	4	4	5	5
Conditioned Volume (m^3)*	0.054	0.0047	5	5	4	6
Temperature Setpoint - Cooling Occupied (°C)	-0.051	0.0007	6	6	6	3
Temperature Setpoint - Heating Setback (°C less than heating setpoint)	-0.049	0.0005	7	7	7	8
Conditioned Floor Area (m^2)*	0.033	0.0026	8	8	8	7
Boiler Efficiency (%)	-0.032	0.0006	9	9	10	10
Overall Enclosure Area Weighted U-Value (W/m^2K)*	0.030	0.0022	10	12	11	14
Above Grade Window-to-Wall Ratio – Elevation 1*	-0.028	0.0008	11	11	12	12
Overall Above Grade Vertical Enclosure Window-to-Wall Ratio*	0.026	0.0007	12	13	13	11
Overall Building Height (m)*	0.026	0.0022	13	10	9	9
Overall Area Weighted Glazing Solar Heat Gain Coefficient*	0.025	0.0005	14	14	14	13
Temperature Setpoint - Cooling Setback (°C more than cooling setpoint)	0.019	0.0004	15	15	15	15
Chiller COP (W/W)	-0.013	0.0005	16	17	17	16
Core Light Power Density (W/m^2)	0.012	0.0004	17	16	16	17
Overall Area Weighted Wall and Window U-Value (including below grade walls) (W/m^2K)*	0.008	0.0024	18	18	18	18
Fan Efficiency (%)	-0.005	0.0005	19	19	19	19
Area Weighted Occupancy Density ($person/m^2$)*	-0.005	0.0005	20	20	20	20
Area Weighted Electrical Plug Power Density (W/m^2)*	-0.001	0.0003	21	22	22	21
Building Orientation (degrees from Elevation 1)	0.001	0.0005	22	23	23	23
Above Grade Window-to-Wall Ratio – Elevation 2*	0.001	0.0006	23	21	21	22

The order of feature importance did not stay consistent throughout the embedded feature selection process, particularly for the combined features. Figure 39 shows how the model coefficient values change for a selection of combined features as λ was increased.

The conditioned floor area had the highest coefficient weight when most of the features were included in the model but as the number of features decreased, the coefficient value for the condition floor area decreased significantly and the above grade building enclosure surface area significantly increased. This indicated that these two features were likely highly correlated. This was confirmed by determining that the Pearson correlation coefficient was 0.91, indicating high correlation between the two features. Where the model performance declined at fewer than 23 features, the coefficient weight of the above grade building enclosure surface area significantly increased meaning that this feature was primarily contributing to the prediction of the target variable.

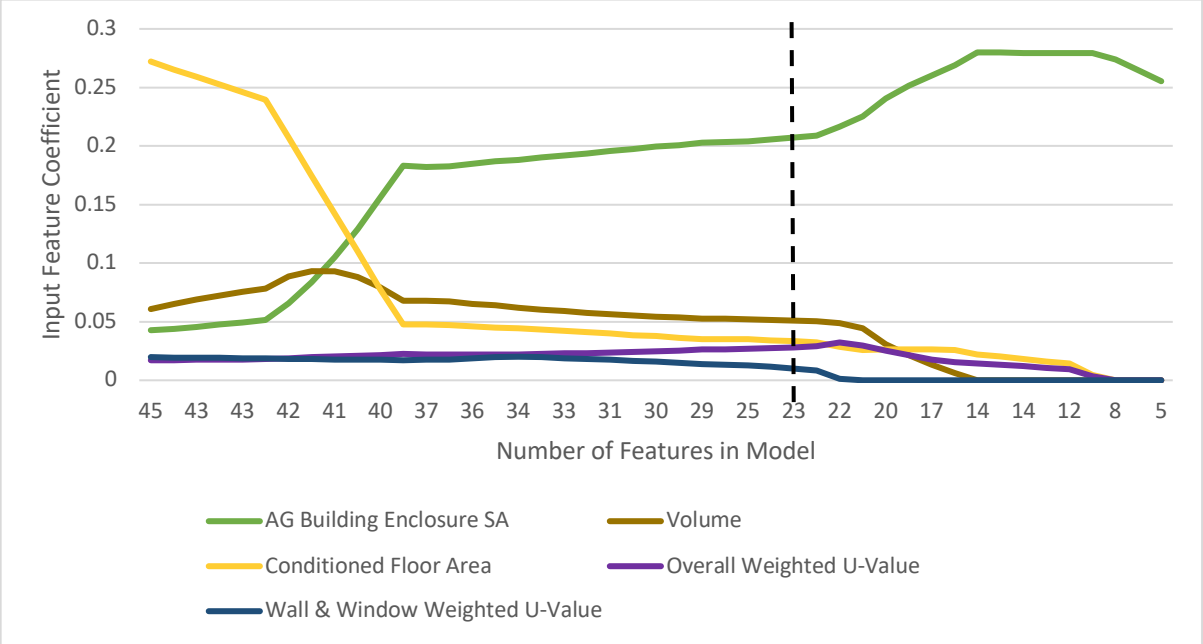

Figure 39 – Change in coefficient values for a selection of input features as the shrinkage parameter (λ) was changed for LASSO

The correlation of the combined features to one another was evaluated using the Pearson correlation coefficient. Figure 40 shows a *heatmap* of the combined feature correlations for the optimal LASSO model features selected. The squares highlighted in dark red and dark blue show two input features that are highly positively or negatively correlated, respectively.

To further explore the impact of the collinear features to the behaviour and prediction accuracy of the model, the highly collinear features were removed as shown by the blue lines in Figure 40.

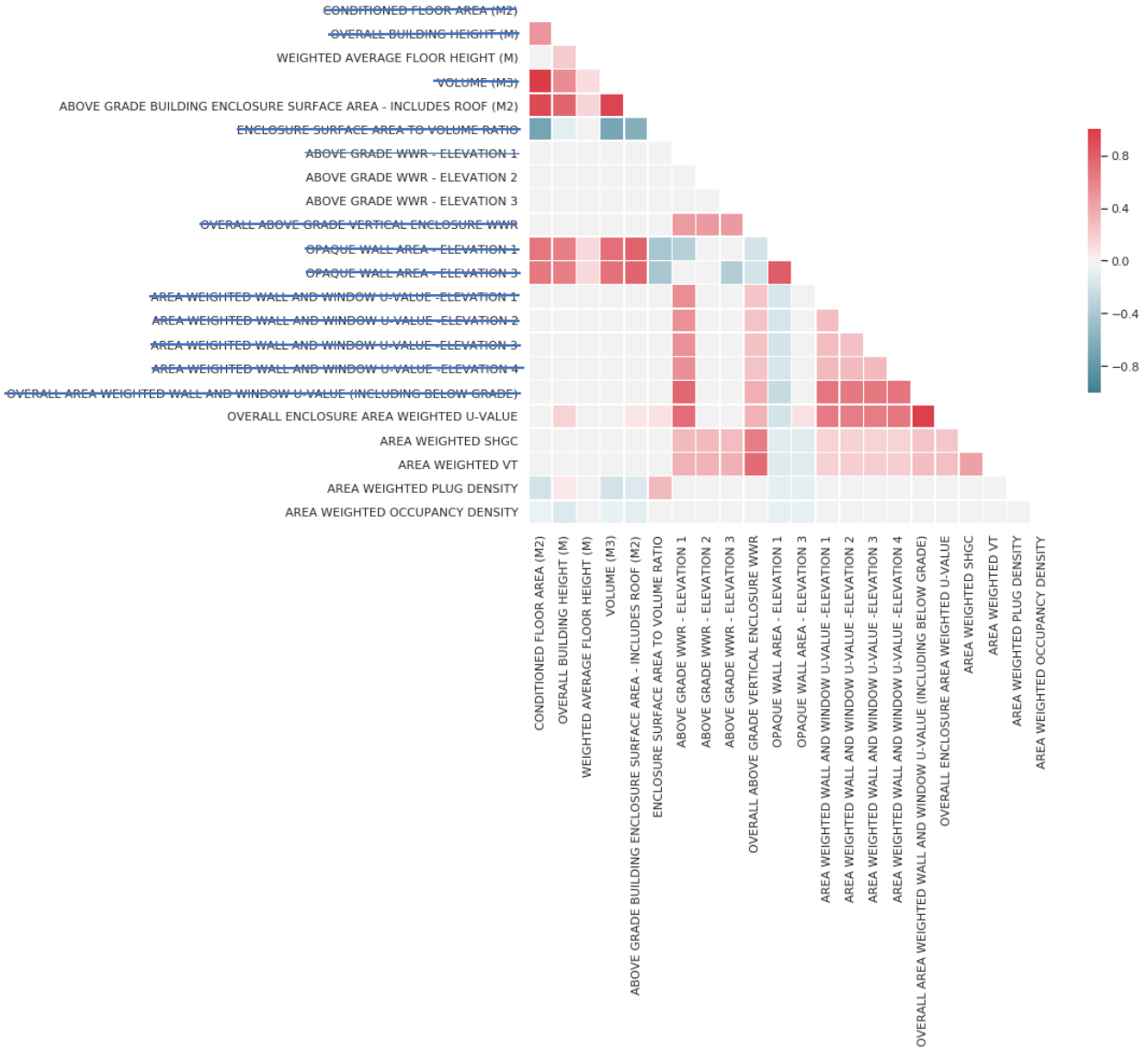

Figure 40 – 'Heatmap' showing the Pearson correlation coefficient of the combined features. The blue lines on the y-axis indicate the features removed in order to eliminate highly collinear input features.

The LASSO model was trained using a reduced input feature set of 32 features, with 13 of the combined features removed due to collinearity. A range of λ values were used to determine the optimal LASSO model for this dataset and illustrate how the model performances compare to the LASSO models with correlated features.

As shown in Figure 41, the models with correlated features removed have a slightly lower coefficient of determination compared to the LASSO models with correlated features. Similarly, the optimal point for model accuracy and number of features is at 23 features. At 23 features the LASSO model with correlated features removed had a coefficient of determination, R^2, of 0.006 less on the training and validation datasets than the optimal LASSO model when the correlated features were included. This indicated that the behaviour of the removed correlated combined features was largely captured by the remaining features.

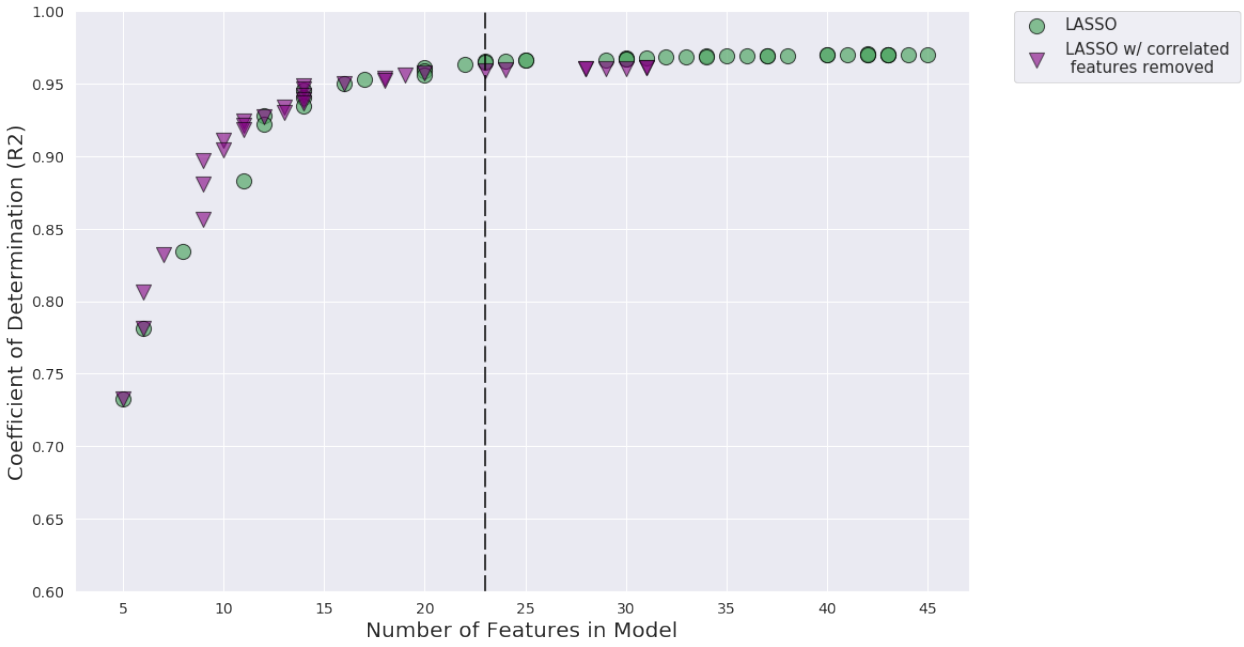

Figure 41 – Model coefficient of determination as the number of features in each the LASSO with correlated features (green circle) and LASSO with correlated combined features removed (purple triangle) models change. The selected number of features for both methods is shown by the dashed line.

In Table 10, the selected features for the LASSO model where the highly correlated combined features were removed is presented in order of coefficient value. The weight distributed between combined features describing the building geometry was concentrated on the enclosure surface area, significantly increasing the significance of this feature. There is risk in doing this as the above-grade building enclosure surface area may describe the building geometry when all samples have a similar rectangular, repeated floor form but if the intent is to use this surrogate model for real buildings where building geometry varies, relying only on the enclosure surface area may lead to surrogate model performance issues. This behaviour and its impact was excluded from this study but is important for further research in this field as simplified geometry building archetypes are often used as the base energy simulation models for dataset development.

After the removal of highly correlated features, the high significance features selected for each feature set were similar; one interesting observation was that the overall area weighted glazing solar heat gain coefficient was of higher importance than the overall enclosure area weighted U-value when the highly correlated features were removed. This behaviour was likely attributed to the high collinearity between these the combined features.

*Table 10 – Features in order of absolute coefficient values following highly correlated combined feature removal. Combined features shown with **

1	Above Grade Building Enclosure Surface Area – including roof (m^2)*	13	Above Grade Window-to-Wall Ratio – Elevation 2*
2	Temperature Setpoint – Heating Occupied (°C)	14	Overall Area Weighted Glazing Visible Transmittance*
3	Supply Air Temperature Cooling (°C)	15	Fan Efficiency (%)
4	Infiltration Rate - Ventilation System Off ($m^3/s/m^2$)	16	Area Weighted Electrical Plug Power Density (W/m^2)*
5	Temperature Setpoint - Cooling Occupied (°C)	17	Above Grade Window-to-Wall Ratio – Elevation 3*
6	Temperature Setpoint - Heating Setback (°C less than heating setpoint)	18	Area Weighted Occupancy Density ($person/m^2$)*
7	Boiler Efficiency (%)	19	Weighted Average Floor Height (m)
8	Overall Area Weighted Glazing Solar Heat Gain Coefficient*	20	Building Orientation (degrees from Elevation 1)
9	Overall Enclosure Area Weighted U-Value (W/m^2K)*	21	Fan Pressure Rise (Pa)
10	Temperature Setpoint - Cooling Setback (°C more than cooling setpoint)	22	Outside Air Rate (m^2/s-person)
11	Core Light Power Density (W/m^2)	23	Ground Floor Perimeter Light Power Density (W/m^2)
12	Chiller COP (W/W)		

For the purposes of drawing conclusions from the feature significance, the LASSO model with highly correlated input features removed was used. However, due to its higher prediction accuracy and ability to consider multiple aspects of building geometry such as building enclosure surface area, conditioned volume and floor area, and overall building height, the LASSO model trained using the 45 input features (22 combined features and 23 original features) was selected as the final model.

3.2.6 Final Model Evaluation

The final model was evaluated for prediction accuracy using the test dataset.

Figure 42 – Surrogate model training progress – final model evaluation

Table 11 summarizes the model prediction results (mean and standard deviation) of the final selected model on the transformed training, validation and test set. The similar model predictive behaviour on the three datasets indicates that the model is not overfitting the training data. The test dataset surrogate model predicted values versus the EnergyPlus simulated log transformed value is shown in Figure 43.

Table 11 – Final model performance on the log transformed training, validation and test datasets

	Training	Validation	Testing
R^2	0.9683 +/- 0.0004	0.9674 +/- 0.0021	0.9695 +/- 0.0001
RMSE	0.0656 +/- 0.0003	0.0658 +/- 0.0014	0.0628 +/- 0.0002
MAE	0.0509 +/- 0.0002	0.0511 +/- 0.0009	0.0495 +/- 0.0002

To evaluate how the surrogate model performs when predicting annual building energy use in GigaJoules, the surrogate model predicted values were re-transformed using Equation 5 and compared to the EnergyPlus simulated target values.

$$\hat{y} = 10^{(\beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)} \qquad (5)$$

Figure 44 shows the test dataset re-transformed predicted values against the EnergyPlus simulated annual building energy use values in GJ. The data generally fits well with some spreading in the higher annual building energy use ranges where there is less data for the model to train on. Figure 45 shows the test dataset re-transformed residual plot and Figure 46 and Figure 47 show the test dataset percent error.

s shown in the Box and Whisker plot in Figure 47, the model under and over predicts the total energy use uniformly, with a median percent error on the test data close to zero. Half of the test samples were predicted by the trained regression within approximately 9% of the EnergyPlus simulated annual building energy use. There are a few outlying data points with prediction percent error in the 38-50% range. These data points fall in the lower range of annual building energy use

values where a small difference between the predicted and actual energy use can cause a high percent error. This is shown in the predicted annual building energy use versus percent error plot (Figure 46) where the higher percent errors are in the low and mid ranges of the predicted annual energy use values. The selected model has an average percent error of 11.38 +/- 0.05% and a R^2-score of 0.9304 +/- 0.0007 on the re-transformed test data set.

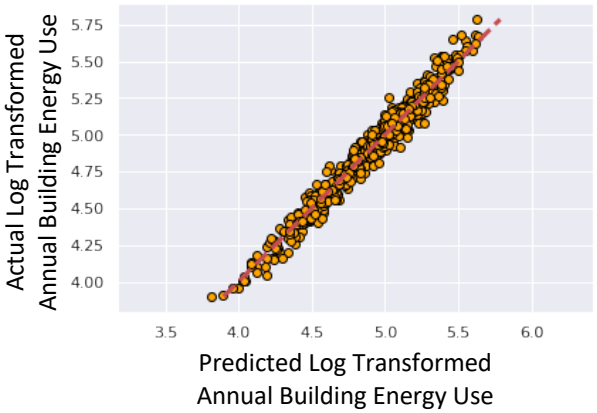

Figure 43 - Predicted (surrogate model) vs. actual (EnergyPlus simulated) log transformed annual building energy use for test dataset

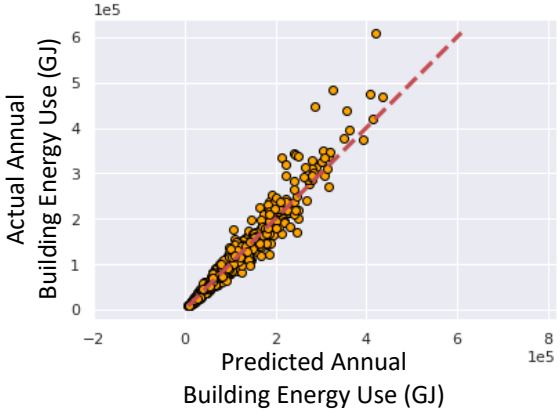

Figure 44 – Predicted (surrogate model) vs. actual (EnergyPlus simulated) annual building energy use in GJ for test dataset

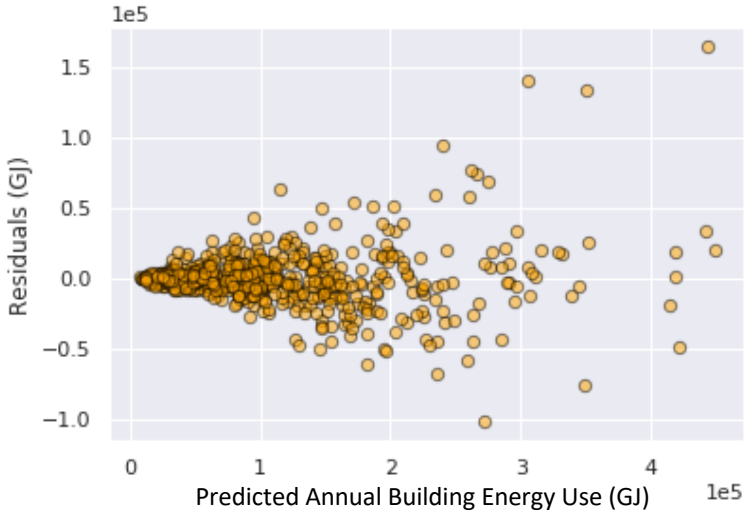

Figure 45 – Residual plot for predicted (surrogate model) annual building energy use in GJ for test dataset

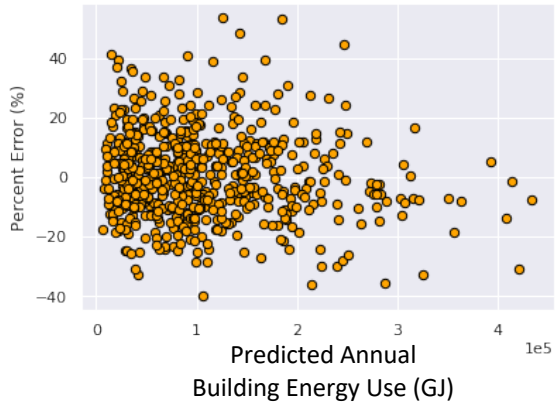

Figure 46 – Predicted annual building energy use in GJ vs. percent error for test dataset

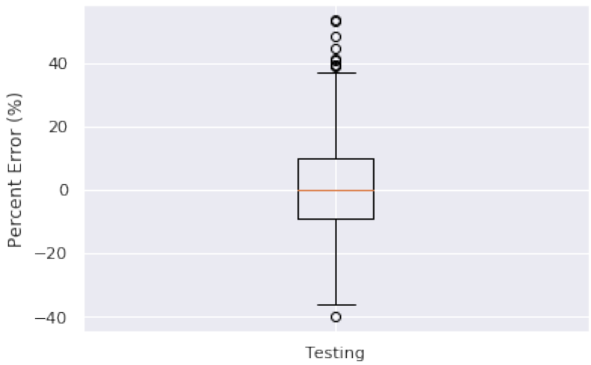

Figure 47 – Distribution of the test dataset percent error. Boxplot shows the data within the first to third interquartile ranges and the outliers

To confirm that the dataset size used was appropriate for the study performed, the test dataset average percent error was calculated on the re-transformed data at decreasing training set sizes as shown in Figure 48. The graph shows that below a training set size of 1,000 training samples, the multivariate regression with L1 LASSO regulator model accuracy decreases significantly. Therefore, the training set size of 2,800 selected for this study was appropriate and could have been decreased to reduce computer simulation and model training time.

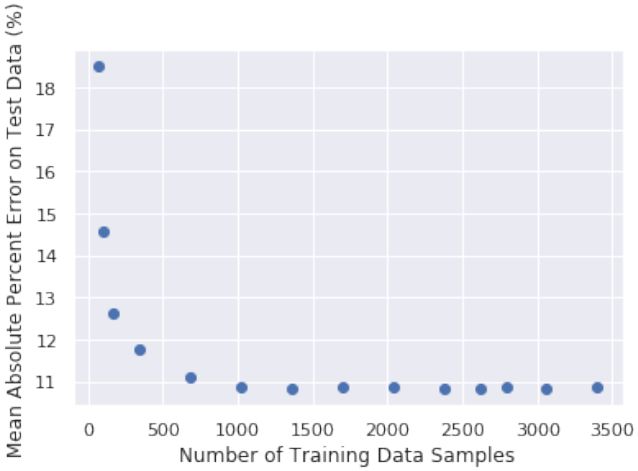

Figure 48 – Impact of training sample size on the mean absolute percent error of the test dataset

3.3 Model Validation Using Downtown Toronto Reference Model

Since the surrogate model was trained on samples that were a random combination of the building features within the defined ranges, the ability of the surrogate model to predict annual building energy use for a realistic combination of feature values had not been evaluated through the test dataset. To validate the surrogate model an eQuest Model National Energy Code for Buildings (MNECB) reference model for a large commercial office tower in downtown Toronto, Ontario was used. The model, referred to herein as Building A, was a 41-storey tower with 5 levels of parking below grade.

eQuest is a building energy modelling simulation tool that uses DOE-2 software to run building energy use simulations. Comparing the surrogate model to annual energy use simulated in the eQuest models was challenging as there were several factors that caused error including, but not limited to, the simulated energy use difference between DOE-2 and EnergyPlus; the error when simplified reference models are used; the error associated with building attributes that remained constant through all samples but vary from the base EnergyPlus model in the eQuest models provided; and the error of the surrogate model itself for a realistic combination of feature values. These can be significant: in a comparison of eQuest and EnergyPlus for a medium size office building archetype, Rallapalli [71] found that the annual pump energy use was 57% higher and the annual natural gas energy use was 94% higher for the eQuest model compared to EnergyPlus.

The eQuest model simulated results could not be readily compared to the surrogate model results to validate the surrogate model. To isolate the error between the surrogate model and the EnergyPlus simulation, the building features from the Building A eQuest model were extracted and those features used in the DOE commercial reference model for large office buildings base model. This involved simplifying the building geometry. Building A had building geometry

similar to the surrogate reference model geometry. The number of storeys were kept the same as were the repeated floors width and depth and the floor heights.

The prediction error for the surrogate model prediction compared to the EnergyPlus simulated value was 14%.

3.4 Impact of Lighting and Plug Loads on Model Accuracy

A previous study [69] used the same original 71 input features and EnergyPlus simulations to predict annual total site energy use, a target variable that included all energy uses simulated by EnergyPlus, including heating, cooling, lighting, electrical, domestic hot water, pump, and fan energy uses and the same input feature and target variable transformation was used as was discussed previously. Combined features were similarly added to the feature set using a forward stepwise selection methodology. LASSO, and Elastic Net with 50% L1 regulator and 50% L2 regulator were evaluated on the training dataset and Elastic Net was found to have the highest model prediction performance on the re-transformed target data. The final model performance on the re-transformed target variable test dataset (20% of the 4,000 samples) was R^2 of 0.98 and the mean absolute percent error was 6.49%, with half the test data predictions falling between +/- 5% MAPE. Therefore, predicting annual total site energy use had a higher accuracy in comparison to the target variable presented in this chapter, annual building energy use. This improved model performance was due to the inclusion of the lighting and electrical plug energy uses that are linearly correlated with the lighting and electrical power densities and building floor area. As shown in Table 12, the lighting and electrical plug power densities were within the top most important features for predicting the annual total site energy use.

*Table 12 – Features in order of absolute coefficient values for both the annual total site energy use surrogate model developed in [69] and the annual building energy use surrogate model (excluding lighting and plug energy use) presented in this chapter. Combined features shown with *.*

#	Annual Total Site Energy Use (GJ) Target [69]	Annual Building Energy Use (GJ) Target
1	Conditioned Floor Area (m^2)*	Above Grade Building Enclosure Surface Area – including roof (m^2)*
2	Above Grade Building Enclosure Surface Area (m^2)*	Temperature Setpoint – Heating Occupied (°C)
3	Area Weighted Light Power Density (W/m^2)*	Supply Air Temperature Cooling (°C)
4	Temperature Setpoint – Heating Occupied (°C)	Infiltration Rate - Ventilation System Off ($m^3/s/m^2$)
5	Supply Air Temperature Cooling (°C)	Conditioned Volume (m^3)*
6	Area Weighted Electrical Plug Power Density (W/m^2)*	Temperature Setpoint - Cooling Occupied (°C)
7	Infiltration Rate – Ventilation System Off ($m^3/s/m^2$)	Temperature Setpoint - Heating Setback (°C less than heating setpoint)
8	Temperature Setpoint – Cooling Occupied (°C)	Conditioned Floor Area (m^2)*
9	Temperature Setpoint – Heating Setback Setback (°C less than cooling setpoint)	Boiler Efficiency (%)
10	Overall Enclosure Area Weighted U-Value (W/m^2K)*	Overall Enclosure Area Weighted U-Value (W/m^2K)*
11	Conditioned Volume (m^3)*	Above Grade Window-to-Wall Ratio – Elevation 1*
12	Boiler Efficiency (%)	Overall Above Grade Vertical Enclosure Window-to-Wall Ratio*
13	Overall Above Grade Vertical Enclosure Window-to-Wall Ratio*	Overall Building Height (m)*
14	Window Area Weighted Window Solar Heat Gain Coefficient*	Overall Area Weighted Glazing Solar Heat Gain Coefficient*
15	Temperature Setpoint – Cooling Setback Setback (°C more than cooling setpoint)	Temperature Setpoint - Cooling Setback (°C more than cooling setpoint)
16	Opaque Wall Area – Elevation 2 (m^2)*	Chiller COP (W/W)
17	Opaque Wall Area – Elevation 4 (m^2)*	Core Light Power Density (W/m^2)
18	Chiller COP (W/W)	Overall Area Weighted Wall and Window U-Value (including below grade walls) (W/m^2K)*
19	Elevator Design Level (no. of elevators)	Fan Efficiency (%)
20	Area Weighted Occupancy Density (person/m^2)*	Area Weighted Occupancy Density (person/m^2)*
21	Fan Efficiency (%)	Area Weighted Electrical Plug Power Density (W/m^2)*
22	Building Orientation (degrees from Elevation 1)	Building Orientation (degrees from Elevation 1)
23	Fan Pressure Rise (Pa)	Above Grade Window-to-Wall Ratio – Elevation 2*
24	Outside Air Rate (m^2/s-person)	
25	Infiltration Rate – Ventilation System On (% of off)	

The features selected and order of significance were compared for the two models with different target variables. Beyond the obvious differences of internal occupant-driven loads, the features selected and their relative order of importance from this analysis were similar. The improved model performance when lighting and electrical energy uses were included in the model was interesting to observe when evaluating the surrogate models developed by other researchers. As summarized in Chapter 2, annual total site energy use was a common target variable used by previous researchers, with 30% of the research included in Chapter 2 using this as a target.

3.5 Surrogate Model Findings

This study developed, tested and evaluated a method for reducing the surrogate model input feature matrix size used to predict annual building energy use. The goals were to introduce a method to reduce the feature subset each time a surrogate model is developed, and evaluate the features selected for the specific dataset. The surrogate model developed not only predicts annual building energy use for any combination of the features (within the initial feature ranges simulated) but the feature significance can be interpreted to better understand the key attributes of a building that impact the energy use. The study focused on the large office building archetype in Toronto, ON but the methodology presented can be applied to different building archetypes and climates.

This study showed that combining features in the original feature set, such as electrical power density for basement, core and perimeter, improved or maintained the model predictive performance. Further, for predicting annual building energy use for large office buildings in Toronto, these features did not need to be separated by area and instead, in the future, could be averaged and represented as a single feature from the onset. This would simplify the sample set EnergyPlus IDF creation and could eliminate the step of combined feature selection. The combined

feature selection step proved to be the most intensive part of the data analysis: it was an iterative process that would be challenging to automate.

Amongst the goals with this study was the evaluation of surrogate model behaviour and performance using a large number of features with wide ranges. Some of the high and low ends of the ranges selected were outside of the normal boundaries for high-rise commercial office buildings in Toronto, both new and existing. This may have impacted the surrogate model behaviour. Future work could include the same features with narrower feature ranges, particularly for building geometry to evaluate the behaviour and predictive performance of the model.

4 Discussion and Conclusions

This thesis provided an overview of surrogate modelling for building energy prediction and methods used by researchers in the field. Through the literature review, a gap was identified in the current building energy surrogate modelling research for selecting a subset of features that have the most important impact on the target variable. To address this gap, a feature selection methodology was presented that reduced the feature subset size without impacting the model predictive performance.

The study presented in Chapter 3 used a large office building archetype in Toronto to explore and analyze the data processing step of feature selection. Initially, 71 features were selected for variation in the dataset. Many of these initial features were combined into *combined features* and added to the surrogate model in a forward stepwise methodology. Where the combined features improved or maintained the model prediction performance, they were included in the model, and the original features were removed. Further, embedded feature selection using the L1, LASSO regulator was used to remove features that had little importance in the prediction of the target variable.

4.1 Key Findings

The final model included 11 of the original features used to develop the sample sets, plus 12 linear combinations of the original features. The final model with 45 features had a coefficient of determination (R^2) on the validation dataset of 0.9674 +/- 0.0021, compared to the surrogate model for the original 71 features, which had a R^2 of 0.9643 +/- 0.0027 on the validation set. Therefore, the surrogate model, following combined feature addition and LASSO feature selection, performed the same as (within error) as the surrogate model with all features included. The final model predicted annual building energy use (the sum of annual heating, cooling, pump and fan energy uses) to an average error of 11.4% on the test dataset, with half the test datapoints within +/-9% error and outlying data points approximately 40% error. This is acceptable for model energy use prediction in an early stage design tool.

To evaluate the performance of a realistic building feature set using the surrogate model, a combination of features from the energy reference model for a newly constructed downtown Toronto, Ontario large office tower was used. The surrogate model prediction was compared to the features simulated using the EnergyPlus base model used to generate the dataset. The predicted annual energy use result from the surrogate model was 14% higher than the EnergyPlus model result. This falls just outside the first interquartile range of error on the test dataset and therefore indicates that this realistic combination of building features is within the design space of the model.

In an interesting outcome of this study, the building geometry description features in the final model were not specific to the shape of the building or the interior zoning. As a result, building geometry terms describing the whole building would allow for flexibility when using this model as an early-stage design tool because exact geometry is not required. The user of the surrogate model would be able to explore design options based on features such as conditioned floor area

and volume rather than heights of individual floors and mechanical plenums. Further research is required to determine if the model can apply to buildings that deviate from the DOE Large Office Commercial Reference Model's rectangular form.

The impact of including a wide range of building geometry would bring value to the field. In this study, the conditioned floor area ranged from 6,570 to 1,780,000 m^2 for the full dataset to reflect the full range of "large" office building sizes. This resulted in the building geometry having a weight almost twice as large as the next most important feature. Narrowing the building size ranges by separating large buildings into further size categories and keeping the building geometry constant for all samples would provide insight into the impact of geometry on the surrogate model behaviour and the feature importance. Future research exploring the impact of including varying building geometry in the feature set is recommended. On one hand, including geometry makes the surrogate model, as a tool, more useful. However, it may impact the ability of the model to differentiate between energy conservation methods on a building where the geometry will be kept constant.

None of the comparable surrogate modelling studies reviewed included temperature setback for both heating and cooling in the model feature set. However, temperature setpoints and setbacks were among the features of highest significance. This shows that the setpoint has an important impact on annual energy use. Using off-hour temperature setbacks is a common easy-to-implement energy saving strategy used by commercial office building owners to reduce their building energy consumption. It is a low or no cost method, and if implemented appropriately, can have little or no impact on tenant comfort. Including this feature in the surrogate model for office buildings is therefore important to exploring energy conservation methods.

The study focused on multivariate regression and did not include complex learning algorithms, such as artificial neural networks and random forest, which have been used by other researchers in this field. The goal of this study was *not* to test several different learning algorithms and select the best one. Instead, the intent was to focus on a process that allows for analysis of selected feature behaviour from the dataset. Future research could evaluate if the feature selection methodology presented improves surrogate model prediction accuracy if implemented before training using a more complex algorithm.

4.2 Future Research

From the review of previous research in the field of building energy surrogate modelling, summarized in Chapter 2 several areas were identified where future research can be advantageous to the further development of this field. The accuracy analysis required for the surrogate modelling intent is often not discussed in the presentation of surrogate model development. As this field of research continues to move forward and surrogate models are integrated into industry-used tools for predicting building energy use and other metrics, development of approved metrics for the surrogate model prediction accuracy will be essential.

Further research into multi-climate surrogate models including the range of climate conditions that can be included in a single model, and the input variables used to describe climate, are research areas that could be thoroughly explored. There is a significant advantage to training multi-climate models, particularly where these models are transformed into industry tools. Rather than developing a surrogate model for each city or region, a surrogate model that is usable for multiple climates can expedite the process.

To develop a methodology for developing surrogate models for several building archetypes in multiple climates, automation of the full process would be advantageous. Mueller [9] developed

an automated surrogate model training program using artificial neural networks and random forests that tested combinations of hyperparameters and selected the model with the highest performance. This was used by Nagpal et al. [24] in their building energy surrogate modelling study. Extending this type of automation to dataset developing, energy model simulation, and data pre-processing would be of value in integrating surrogate modelling into an industry tool.

The impact of the method used to develop the sample set and sampling plan has not been thoroughly evaluated in the context of building energy surrogate modelling. As this step is completed in the early stages of surrogate modelling development and influences the dataset used to train the surrogate models, the sampling plan impact cannot be compared between studies. There is an opportunity to compare and evaluate, in detail, multiple commonly used sampling plans by keeping all other factors of the experiment constant.

The mechanical systems for surrogate models are generally constant, with some studies including varying equipment efficiencies as features. This is limiting when the surrogate model is being used as an early-stage design tool for new or existing buildings, as differing mechanical systems are often explored during this phase. There is an opportunity to explore using multiple mechanical systems in a single surrogate model.

While some datasets are available that used simulation software that is now-obsolete (e.g. [33]), there is benefit for researchers to share their datasets developed using current building energy modelling software commonly used in industry. Benefits of sharing datasets are minimizing the amount of simulations required to analyze the impact of the selected features and associated ranges on surrogate model behaviour and performance, and eliminating variation in datasets used between studies.

There is a large focus in previous research on the highest performing algorithms for building energy surrogate modelling. There is a risk in placing such focus on the last step of a complex process because these algorithms are highly dependent on the input datasets and thus such conclusions will not be valid across the studies. This is evident from the lack of consensus across the literature. There is an opportunity to shift this focus and perform detailed comparisons on earlier stages of the process within dataset development and data processing as well as to look further into feature selection methodologies and use surrogate modelling to evaluate building feature importance to the variable being used as the target, such as monthly or annual energy use.

Another current research gap is the lack of discussion and analysis on how well building energy surrogate models compared to detailed energy models created for real buildings. While some researchers developed a base model using real buildings, for example [24], many studies use existing archetype models as a base model. Comparing surrogate models to detailed energy models may show a different behaviour than comparing reference models to detailed energy models.

As a general comment, it was observed that there is inconsistent reporting of information within the literature critical to interpreting results and enabling study repeatability. This missing information includes, but is not limited to, the versions of energy simulation software and statistical analysis programs and packages used, the hyperparameters used for each learning algorithm, and how the hyperparameters were determined. Thorough documentation is important in order to learn for future work and compare to completed studies.

As this field continues to grow and surrogate models are integrated into other workflows, such as optimization, there are opportunities to focus on each stage of the full process. We as researchers can shape the way the data is formed and test and compare methods to creating datasets that will best meet the needs of the surrogate model.

Appendix A – Summary of Surrogate Models Presented in Studies Referenced in Chapter 2

	Reference	Wong et al.	Tsanas and Xifara	Tian, Choudhary et al.	Tian, Song et al.	Tian and Choudhary	Aijazi and Glicksman	Hygh et al.	Nagpal et al. (2018)	Asl et al.	Chidiac et al.	Catalina et al.
		[32]	[35]	[18]	[51]	[22]	[36]	[14]	[24]	[20]	[19]	[15]
	Other Researchers Using Same Dataset		Roy et al [41] Papadopoulos et al. [42] Castelli et al [43] Chou and Bui [44]					Al Gharably et al. [45] (added wall area per elevation and roof area to feature set)	Nagpal et al. (2019) [25]			
	LOCATION(S)	HONG KONG	ATHENS, GREECE	UNIVERSITY OF PENNSYLVANIA GEORGIA INSTITUTE OF TECHNOLOGY	LONDON, UK	LONDON, UK	RALEIGH, BOSTON, DULUTH	MIAMI, WINSTON-SALEM, MINNEAPOLIS, ALBUQUERQUE	MIT	MIAMI, CHICAGO, SAN FRANCISCO	OTTAWA, EDMONTON, VANCOUVER	FRANCE (16 MAJOR CITIES INCLUDING PARIS, MARSEILLE, CHAMBERY, STRASBOURG)
	BUILDING TYPE	LARGE OFFICE	RESIDENTIAL	CAMPUS BUILDINGS	MEDIUM SIZE OFFICE	SECONDARY SCHOOL	MEDIUM MULTI-UNIT RESIDENTIAL	MEDIUM OFFICE	UNIVERSITY ACADEMIC, LABORATORY, AND RESIDENCE	MEDIUM SIZED BUILDING (DOES NOT SPECIFY USE)	MEDIUM OFFICE	RESIDENTIAL SINGLE-FAMILY HOUSE
	TARGET VARIABLE(S)	DAILY ELECTRICITY LOAD (COOLING, HEATING, LIGHTING, TOTAL)	ANNUAL COOLING LOAD ANNUAL HEATING LOAD	ANNUAL COOLING LOAD ANNUAL HEATING LOAD	ANNUAL COOLING LOAD ANNUAL HEATING LOAD ANNUAL CARBON EMISSION (CO_2e)	ANNUAL NATURAL GAS USE INTENSITY ANNUAL ELECTRICITY USE INTENSITY	ANNUAL TOTAL LOAD ANNUAL COOLING LOAD ANNUAL HEATING LOAD	ANNUAL TOTAL LOAD ANNUAL COOLING LOAD ANNUAL HEATING LOAD	ANNUAL ENERGY USE INTENSITY	ANNUAL ENERGY USE INTENSITY	ANNUAL HEATING LOAD ANNUAL COOLING LOAD ANNUAL FAN, PUMP, OTHER LOADS MONTHLY TOTAL LOAD	ANNUAL TOTAL LOAD MONTHLY TOTAL LOAD
	HYPOTHETICAL/REAL INPUT DATA	H	H	R	H	H	H	H	R	H	H	H
	ENERGY SIMULATION SOFTWARE	ENERGYPLUS	ECOTECT	ENERGY PERFORMANCE STANDARD CALCULATION TOOLKIT (EPSCT)	ENERGYPLUS V 8.1	ENERGYPLUS	ENERGYPLUS	ENERGYPLUS	ENERGYPLUS	DOE 2.2	ENERGYPLUS	TRNSYS
	DISCRETE /CONTINUOUS VARIABLES	D	D	N/A	C	C	C	C	C	?	D	D
Building Geometry	Building Orientation		X				X	X				X
	Building Footprint Area									X		
	Total Conditioned Floor Area			X				X		X		
	Building Volume											
	Total Height		X	X								
	Compactness Ratio (Area/Volume)		X									X
	Aspect Ratio (length/depth)							X				
	Number of Storeys							X				
	Building Depth											
	Building Width											
	Perimeter Zone Depth								X			
	Floor Height								X			
	Total Enclosure Surface Area		X									
	Roof Area		X	X						X		
	Wall Area		X	X						X		
	Window Area		X	X						X		
	Below Grade Area			X								
	Window-to-Floor Ratio											X
	Window-to-Wall Ratio	X	X			X	X	X	X		X	
Building Enclosure Performance	Enclosure Thermal Transmittance											X
	Window Thermal Transmittance			X	X	X	X	X	X		X	
	Window Visible Transmittance	X		X								
	Window SHGC	X			X	X	X	X				
	Glazing/Fenestration Type								X			
	Wall Thermal Transmittance			X	X	X	X	X	X		X	
	Wall Specific Heat											
	Wall Emissivity/Absorptance											
	Roof Thermal Transmittance			X	X	X	X	X	X		X	
	Roof Emissivity/Absorptance							X				
	Slab-on-Grade Thermal Transmittance					X						
	Shading Projection Factor							X				
	Fenestration Shading	X							X			
Air Infiltration	Air Infiltration Rate				X	X	X				X	
	Window Leakage								X			
	Perimeter Outside Air Flowrate								X			
	Core Outside Air Flowrate								X			
Internal Loads	Lighting Power Density			X	X	X	X		X		X	
	Daylighting					X			X		X	
	Equipment Power Density			X	X	X			X		X	
	Elevator Load										X	
HVAC Systems	Heating Efficiency				X	X	X				X	
	Cooling Efficiency				X		X				X	
	Domestic Hot Water Type											
	Fresh Air Ventilation			X		X			X			X
	Heating Temperature Setpoint				X	X			X			
	Cooling Temperature Setpoint				X							
	Unoccupied Temperature Setpoint											
	Service Hot Water Usage								X			
	Economizer Cycle								X			
Occupancy	Occupant Density			X					X		X	
Internal Mass	Internal Thermal Capacity											
	Building Enclosure Thermal Capacity											
	Building Mass			X								
Scheduling	Occupancy Schedule			X		X			X			
	Average Weekday Occupant			X					X			
	Total Lighting Hours			X								
	Average Weekly Lighting			X					X			
	Outside Air Flowrate Schedule								X			
	Total Equipment Hours			X								
	Average Weekly Equipment			X					X			
Climate	HDD						X					
	Monthly Mean Outdoor Drybulb T											X
	Monthly Average Global Radiation											X

	Reference	Sangireddy et al.	Papadopoulos and Azar	Singaravel et al.	Melo et al.	Carlo and Lamberts	Chari and Christodoulou	Amiri et al.	Asasi et al.	Lam et al.	Ascione et al.	Chen et al.
		[28]	[31]	[77]	[27]	[26]	[52]	[40]	[21]	[38]	[49]	[50]
	Other Researchers Using Same Dataset											
	LOCATION(S)	JAIPUR, HYDERABAD INDIA	ABU DHABI	BRUSSELS	FLORIANPOLIS, BRAZIL	FLORIANPOLIS, BRAZIL	IRELAND	SAN JOSE AND BILLINGS	HOUSTON, TEXAS	HONG KONG	NAPLES	HONG KONG
	BUILDING TYPE	SMALL OFFICE	MEDIUM OFFICE	SMALL OFFICE	MEDIUM OFFICE	LARGE OFFICE	RESIDENTIAL HOUSE	SMALL OFFICE	SMALL OFFICE	LARGE OFFICE	SMALL OFFICE	HIGH RISE RESIDENTIAL
	TARGET VARIABLE(S)	ANNUAL ENERGY USE INTENSITY	MONTHLY TOTAL LOAD	MONTHLY COOLING LOAD MONTHLY HEATING LOAD	ANNUAL COOLING LOAD	ANNUAL ELECTRICITY USE INTENSITY	BUILDING ENERGY RATING (BER)	ANNUAL TOTAL LOAD	ANNUAL ENERGY USE INTENSITY	ANNUAL ELECTRICITY LOAD	ANNUAL HEATING LOAD INTENSITY ANNUAL COOLING LOAD INTENSITY ANNUAL DISCOMFORT HOURS	ILLUMINANCE LEVEL AIR CHANGE RATE ASHRAE55 COMFORT TIME
	HYPOTHETICAL/REAL INPUT DATA	H	H	H	H	H	H	H	H	H	H	H
	ENERGY SIMULATION SOFTWARE	ENERGYPLUS	ENERGYPLUS	ENERGYPLUS	ENERGYPLUS	ENERGYPLUS	IRELAND DEAP	DOE 2.2	DOE 2.2	DOE 2.1	ENERGYPLUS	ENERGYPLUS
	DISCRETE /CONTINUOUS VARIABLES	D		C	D	D	D	D	D	D	C	c
Building Geometry	Building Orientation	X		X	X	X		X	X		X	X
	Building Footprint Area											
	Total Conditioned Floor Area						X				X	
	Building Volume						X					
	Total Height											
	Compactness Ratio (Area/Volume)										X	
	Aspect Ratio (length/depth)	X										
	Number of Storeys			X			X				X	
	Building Depth			X								
	Building Width			X								
	Perimeter Zone Depth											
	Floor Height				X		X				X	
	Total Enclosure Surface Area											
	Roof Area										X	
	Wall Area										X	
	Window Area										X	
	Below Grade Area											
	Window-to-Floor Ratio											X
	Window-to-Wall Ratio	X		X	X	X				X	X	
Building Enclosure Performance	Enclosure Thermal Transmittance						X					
	Window Thermal Transmittance			X	X	X	X			X		X
	Window Visible Transmittance											
	Window SHGC			X	X	X	X			X		X
	Glazing/Fenestration Type	X									X	
	Wall Thermal Transmittance			X	X	X	X	X	X	X	X	X
	Wall Specific Heat											X
	Wall Emissivity/Absorptance				X			X	X		X	
	Roof Thermal Transmittance			X	X	X	X	X	X	X		
	Roof Emissivity/Absorptance				X			X	X			
	Slab-on-Grade Thermal Transmittance			X			X	X	X			
	Shading Projection Factor											X
	Fenestration Shading	X			X	X						X
Air Infiltration	Air Infiltration Rate			X	X		X			X	X	X
	Window Leakage											
	Perimeter Outside Air Flowrate											
	Core Outside Air Flowrate											
Internal Loads	Lighting Power Density		X		X	X	X			X	X	
	Daylighting											
	Equipment Power Density		X		X		X			X	X	
	Elevator Load											
HVAC Systems	Heating Efficiency					X	X			X	X	
	Cooling Efficiency					X				X	X	
	Domestic Hot Water Type						X					
	Fresh Air Ventilation						X					
	Heating Temperature Setpoint		X								X	
	Cooling Temperature Setpoint		X								X	
	Unoccupied Temperature Setpoint		X									
	Service Hot Water Usage											
	Economizer Cycle											
Occupancy	Occupant Density				X		X			X	X	
Internal Mass	Internal Thermal Capacity			X	X		X	X	X			
	Building Enclosure Thermal Capacity				X							
	Building Mass											
Scheduling	Occupancy Schedule				X			X	X			
	Average Weekday Occupant											
	Total Lighting Hours											
	Average Weekly Lighting											
	Outside Air Flowrate Schedule											
	Total Equipment Hours											
	Average Weekly Equipment											
Climate	HDD											
	Monthly Mean Outdoor Drybulb T											
	Monthly Average Global Radiation											

Appendix B – Calculation of Combined Features

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	CFA	Conditioned Floor Area (m^2)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S]	$CFA = W * D * (B + S + 2)$
	H	Overall Building Height (m)	- Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$H = (B * BH) + GH + GPH + (S * RH) + (S * RPH) + TH + TPH$
	HAG	Above Grade Building Height (m)	- Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$HAG = GH + GPH + (S * RH) + (S * RPH) + TH + TPH$
	FH	Weighted Average Floor Height (m)	- Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Repeated Floor Height [RH] - Top Floor Height [TH]	$FH = ((B * BH) + (S * RH) + GH + TH)/(S + B + 2)$
	PH	Weighted Average Plenum Height (m)	- Number of Repeated Floors [S] - Ground Floor Plenum Height [GPH] - Repeated Floor Plenum Height [RPH] - Top Floor Plenum Height [TPH]	$PH = ((S * RPH) + GPH + TPH)/(S + 2)$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	V	Conditioned Volume (m^3)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$V = W * D * H$
	OSA	Overall Enclosure Surface Area – including roof and below grade (m^2)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$OSA = (H * W * 2) + (H * D * 2) + (W * D)$
	ESA	Above-Grade Building Enclosure Surface Area – including roof (m^2)	- Width [W] - Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$ESA = (HAG * W * 2) + (HAG * D * 2) + (W * D)$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	ESA:V	Enclosure Surface Area to Volume Ratio (m^2/m^3)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH]	$ESA{:}V = ESA/V$
	AR	Aspect Ratio – Depth to Width (m/m)	- Width [W] - Depth [D]	$AR = D/W$
	WWR1	Above Grade Window-to-Wall Ratio – Elevation 1	- Width [W] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR1] - Repeated Floor WWR – Elevation 1 [RWWR1] - Top Floor WWR – Elevation 1 [TWWR1]	$WWR1 = \frac{W*(((GWWR1*(GH+GPH))+(RWWR1*S*(RH+RPH))+(TWWR1*(TH+TPH))}{W*HAG}$
	WWR2	Above Grade Window-to-Wall Ratio – Elevation 2	- Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR2] - Repeated Floor WWR – Elevation 1 [RWWR2] - Top Floor WWR – Elevation 1 [TWWR2]	$WWR2 = \frac{D*(((GWWR2*(GH+GPH))+(RWWR2*S*(RH+RPH))+(TWWR2*(TH+TPH))}{D*HAG}$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	WWR3	Above Grade Window-to-Wall Ratio – Elevation 3	- Width [W] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR3] - Repeated Floor WWR – Elevation 1 [RWWR3] - Top Floor WWR – Elevation 1 [TWWR3]	$WWR2 = \frac{W * (((GWWR3 * (GH + GPH)) + (RWWR3 * S * (RH + RPH)) + (TWWR3 * (TH + TPH)))}{W * HAG}$
	WWR4	Above Grade Window-to-Wall Ratio – Elevation 4	- Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR4] - Repeated Floor WWR – Elevation 1 [RWWR4] - Top Floor WWR – Elevation 1 [TWWR4]	$WWR2 = \frac{D * (((GWWR4 * (GH + GPH)) + (RWWR4 * S * (RH + RPH)) + (TWWR4 * (TH + TPH)))}{D * HAG}$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	WWR	Overall Above Grade Vertical Enclosure Window-to-Wall Ratio	- Width [W] - Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR1] - Ground Floor WWR – Elevation 2 [GWWR2] - Ground Floor WWR – Elevation 3 [GWWR3] - Ground Floor WWR – Elevation 4 [GWWR4] - Repeated Floor WWR – Elevation 1 [RWWR1] - Repeated Floor WWR – Elevation 2 [RWWR2] - Repeated Floor WWR – Elevation 3 [RWWR3] - Repeated Floor WWR – Elevation 4 [RWWR4] - Top Floor WWR – Elevation 1 [TWWR1] - Top Floor WWR – Elevation 2 [TWWR2] - Top Floor WWR – Elevation 3 [TWWR3] - Top Floor WWR – Elevation 4 [TWWR4]	$WWR = \frac{WWR1*W*HAG+WWR2*D*HAG+WWR3*W*HAG+WWR4*D*HAG}{2*(W+D)*HAG}$
	WA1	Opaque Wall Area – Elevation 1 (m^2)	- Width [W] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR1] - Repeated Floor WWR – Elevation 1 [RWWR1] - Top Floor WWR – Elevation 1 [TWWR1]	$WA1 = (1 - WWR1) * W * HAG$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Building Geometry	WA2	Opaque Wall Area – Elevation 2 (m^2)	- Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 2 [GWWR2] - Repeated Floor WWR – Elevation 2 [RWWR2] - Top Floor WWR – Elevation 2 [TWWR2]	$WA2 = (1 - WWR2) * D * HAG$
	WA3	Opaque Wall Area – Elevation 3 (m^2)	- Width [W] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 3 [GWWR3] - Repeated Floor WWR – Elevation 3 [RWWR3] - Top Floor WWR – Elevation 3 [TWWR3]	$WA3 = (1 - WWR3) * W * HAG$
	WA4	Opaque Wall Area – Elevation 4 (m^2)	- Depth [D] - Number of Repeated Floors [S] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 4 [GWWR4] - Repeated Floor WWR – Elevation 4 [RWWR4] - Top Floor WWR – Elevation 4 [TWWR4]	$WA4 = (1 - WWR4) * D * HAG$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Enclosure Performance	OWU1	Area Weighted Wall and Window U-Value (includes below grade walls) – Elevation 1 (m^2K/W)	- Width [W] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 1 [GWWR1] - Repeated Floor WWR – Elevation 1 [RWWR1] - Top Floor WWR – Elevation 1 [TWWR1] - Window U-Value – Elevation 1 [WU1] - Wall RSI – Elevation 1 [WR1] - Below Grade Wall RSI [BWR]	$OWU1 = ((W * B * BH * 1/BWR) + (W * (HAG) * WWR1 * WU1) + (W * (HAG) * (1 - WWR1) * 1/WR1)) /(W * H)$
	OWU2	Area Weighted Wall and Window U-Value – Elevation 2 (m^2K/W)	- Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 2 [GWWR2] - Repeated Floor WWR – Elevation 2 [RWWR2] - Top Floor WWR – Elevation 2 [TWWR2] - Window U-Value – Elevation 2 [WU2] - Wall RSI – Elevation 2 [WR2] - Below Grade Wall RSI [BWR]	$OWU2 = (\left(D * B * BH * \frac{1}{BWR}\right) + (D * (HAG) * WWR2 * WU2) + \left(D * (HAG) * (1 - WWR2) * \frac{1}{WR2}\right))/(D * H)$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Enclosure Performance	OWU3	Area Weighted Wall and Window U-Value – Elevation 3 (m^2K/W)	- Width [W] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 3 [GWWR3] - Repeated Floor WWR – Elevation 3 [RWWR3] - Top Floor WWR – Elevation 3 [TWWR3] - Window U-Value – Elevation 3 [WU31] - Wall RSI – Elevation 3 [WR3] - Below Grade Wall RSI [BWR]	$OWU3 = ((W * B * BH * 1/BWR) + (W * (HAG) * WWR3 * WU3) + (W * (HAG) * (1 - WWR3) * 1/WR3)) /(W * H)$
	OWU4	Area Weighted Wall and Window U-Value – Elevation 4 (m^2K/W)	- Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Basement Height [BH] - Ground Floor Height [GH] - Ground Floor Plenum Height [GPH] - Repeated Floor Height [RH] - Repeated Floor Plenum Height [RPH] - Top Floor Height [TH] - Top Floor Plenum Height [TPH] - Ground Floor WWR – Elevation 4 [GWWR4] - Repeated Floor WWR – Elevation 4 [RWWR4] - Top Floor WWR – Elevation 4 [TWWR4] - Window U-Value – Elevation 4 [WU4] - Wall RSI – Elevation 1 [WR4] - Below Grade Wall RSI [BWR]	$OWU4 = ((D * B * BH * \frac{1}{BWR}) + (D * (HAG) * WWR4 * WU4) + (D * (HAG) * (1 - WWR4) * \frac{1}{WR4}))/(D * H)$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Enclosure Performance	OWU	Overall Area Weighted Wall and Window U-Value (including below grade walls) (m^2K/W)	All features included in OWR1, OWR2, OWR3 and OWR4	$OWU = ((OWU1 * W * H) + (OWU2 * D * H) + (OWU3 * W * H) + (OWU4 * D * H))/(2 * (W + D) * H)$
	OEU	Overall Enclosure Area Weighted U-Value (m^2K/W)	All features included in OWR1, OWR2, OWR3 and OWR4 - Slab on Grade RSI Value [SR] - Roof RSI Value [RR]	$OEU = ((OWU * 2 * (W + D) * H) + \left(\frac{1}{SR} * W * D\right) + (\frac{1}{RR} * W * D))/(2 * (W + D) * H + 2 * W * D)$
	SHGC	Overall Area Weighted Glazing Solar Heat Gain Coefficient	All features included in WWR1, WWR2, WWR3 and WWR4 - Glazing Solar Heat Gain Coefficient – Elevation 1 [SHGC1] - Glazing Solar Heat Gain Coefficient – Elevation 2 [SHGC2] - Glazing Solar Heat Gain Coefficient – Elevation 3 [SHGC3] - Glazing Solar Heat Gain Coefficient – Elevation 4 [SHGC4]	$SHGC = ((W * HAG * WWR1 * SHGC1) + (D * HAG * WWR2 * SHGC2) + (W * HAG * WWR3 * SHGC3) + (D * HAG * WWR4 * SHGC4))/(2 * (W + D) * HAG)$
	VT	Overall Area Weighted Glazing Visible Transmittance	All features included in WWR1, WWR2, WWR3 and WWR4 - Glazing Visible Transmittance – Elevation 1 [VT1] - Glazing Visible Transmittance – Elevation 2 [VT2] - Glazing Visible Transmittance – Elevation 3 [VT3] - Glazing Visible Transmittance – Elevation 4 [VT4]	$VT = ((W * HAG * WWR1 * VT1) + (D * HAG * WWR2 * VT2) + (W * HAG * WWR3 * VT3) + (D * HAG * WWR4 * VT4))/(2 * (W + D) * HAG)$
Internal Load	LPD	Area Weighted Lighting Power Density (W/m^2)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Perimeter Zone Depth [PD] - Basement LPD [BLPD] - Ground Floor Perimeter LPD [GPLPD] - Repeated Floor Perimeter LPD [RPLPD]	$PERIMETER\ ZONE\ AREA\ PER\ FLOOR\ [PZA] = W * D - ((W - PD * 2) * (D - PD * 2))$ $CORE\ AREA\ PER\ FLOOR\ [CA] = (W - PD * 2) * (D - PD * 2)$

			- Top Floor Perimeter LPD [TPLPD] - Core LPD [CLPD]	$LPD = ((CA * (S + 2) * CLPD) + (W * D * B * BLPD) + (PZA * GPLPD) + (PZA * S * RPLPD) + (PZA * TPLPD))/CFA$

		COMBINED FEATURE	ORIGINAL FEATURES USED TO CALCULATE COMBINED FEATURE	EQUATION
Internal Load	EPD	Area Weighted Electrical Plug Power Density (W/m^2)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Perimeter Zone Depth [PD] - Basement Plug Power Density [BPD] - Perimeter Plug Power Density [PPD] - Core Plug Power Density [CPD]	$EPD = ((CA * (S + 2) * CPD) + (W * D * B * BPD) + (PZA * (S + 2) * PPD))/CFA$
	OD	Area Weighted Occupancy Density ($person/m^2$)	- Width [W] - Depth [D] - Number of Below Grade Floors [B] - Number of Repeated Floors [S] - Perimeter Zone Depth [PD] - Basement Occupancy Density [BOD] - Maximum Occupancy Density [MOD]	$OD = ((W * D * (S + 2) * MOD) + (W * D * B * BOD))/CFA$

Works Cited

[1] UN, "Paris Agreement," United Nations, Paris, 2015.

[2] UNEP, "Buildings and Climate Change - Summary for Decision-Makers," United Nations Environmental Programme, Sustainable Buildings & Climate Initiative, Paris, 2009.

[3] City of Toronto, "TransformTO - Toronto's 2016 Greenhouse Gas Emissions Inventory," 2016. [Online]. Available: https://www.toronto.ca/services-payments/water-environment/environmentally-friendly-city-initiatives/transformto/torontos-greenhouse-gas-inventory/. [Accessed April 2019].

[4] NRTEE and SDTC, "Geared for Change: Energy Efficiency in Canada's Commercial Building Sector," National Round Table on the Environment and the Economy (NRTEE); and Sustainable Development Technology Canada (SDTC),Government of Canada, Ottawa, 2009.

[5] Toronto 2030 District, "Toronto 2030 Platform," 2018. [Online]. Available: https://www.toronto2030platform.ca/. [Accessed April 2019].

[6] S. Attia, J. L. Hensen, L. Beltran and A. De Herde, "Selection criteria for building performance simulation tools: contrasting architects' and engineers' needs," *Journal of Building Performance Simulation,* vol. 5, no. 3, pp. 155-169, May 2012.

[7] T. Ostergard, R. L. Jensen and S. E. Maagaard, "Building simulations supporting decision making in early design – A review," *Renewable and Sustainable Energy Reviews,* vol. 61, pp. 187-201, August 2016.

[8] A. Forrester, A. Sobester and A. Keane, Engineering Design via Surrogate Modelling: A Practical Guide, vol. 1, John Wiley & Sons Ltd., 2008.

[9] C. Mueller, "Computational Exploration of the Structural Design Space," Cambridge, MA, 2014.

[10] H.-x. Zhao and F. Magoulés, "Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method," *Journal of Algorithms & Computational Technology,* vol. 6, no. 1, pp. 59-77, March 2012.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," Department of Statistics, University of Toronto, Toronto, 1994.

[12] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2, Springer, 2009.

[13] P. Jacobs and H. Henderson, "State-Of-The-Art Review: Whole Building, Building Envelope, and HVAC Component and System Simulation and Design Tools," Architectural Energy Corporation; CDH Energy Corp., 2002.

[14] J. S. Hygh, J. F. DeCarolis, D. B. Hill and S. R. Ranjithan, "Multivariate regression as an energy assessment tool in early building design," *Building and Environment,* vol. 57, pp. 165-175, November 2012.

[15] T. Catalina, J. Virgone and E. Blanco, "Development and validation of regression models to predict monthly heating demand for residential buildings," *Energy and Buildings,* vol. 40, no. 10, pp. 1825-1832, 2008.

[16] CIBSE, "Degree-Days: Theory and Application," The Chartered Institution of Building Services Engineers, London, UK, 2006.

[17] IEA, "Transition to Sustainable Buildings - Strategies and Opportunities to 2050," International Energy Agency, 2013.

[18] 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," *Building and Environment,* vol. 92, pp. 61-74, October 2015.

[19] S. Chidiac, E. Catania, E. Morofsky and S. Foo, " A screening methodology for implementing cost effective energy retrofit measures in Canadian office buildings," *Energy and Buildings,* vol. 43, no. 2, pp. 614-620, 2011.

[20] M. R. Asl, W. Xu, J. Shang, B. Tsai and I. Molloy, "Regression-based building energy performance assessment using Building Information Model (BIM)," in *ASHRAE and IBPSA-USA SimBuild 2016*, Salt Lake City, UT, 2016.

[21] 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 and Buildings,* vol. 85, pp. 246-255, December 2014.

[22] W. Tian and R. Choudhary, "A probabilistic energy model for non-domestic building sectors applied to analysis of school buildings in greater London," *Energy and Buildings,* vol. 54, pp. 1-11, November 2012.

[23] A. Mastrucci, P. Perez-Lopez, E. Benetto, U. Leopold and I. Blanc, "Global sensitivity analysis as a support for the generation of simplified building stock energy models," *Energy and Buildings,* vol. 149, pp. 368-383, August 2017.

[24] S. Nagpal, C. Mueller, A. Aijazi and C. F. Reinhart, "A methodology for auto-calibrating urban building energy models using surrogate modeling techniques," *Journal of Building Performance Simulation,* vol. 12, no. 1, pp. 1-16, 2018.

[25] S. Nagpal, J. Hanson and C. Reinhart, "A framework for using calibrated campus-wide building energy models for continuous planning and greenhouse gas emissions reduction tracking," vol. 241, pp. 82-97, 2019.

[26] J. Carlo and R. Lamberts, "Development of envelope efficiency labels for commercial buildings: Effect of different variables on electricity consumption," *Energy and Buildings,* vol. 40, no. 11, pp. 2002-2008, 2008.

[27] A. Melo, R. 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 and Buildings,* vol. 131, pp. 233-247, November 2016.

[28] S. A. R. Sangireddy, A. Bhatia and V. Garg, "Development of a surrogate model by extracting top characteristic feature vectors for building energy prediction," *Journal of Building Engineering,* vol. 23, pp. 38-52, May 2019.

[29] P. Geyer and S. Singaravel, "Component-based machine learning for performance prediction in building design," *Applied Energy,* vol. 228, pp. 1439-1453, October 2018.

[30] I. Korolija, L. Marjanovic-Halburd and V. I. Hanby, "Regression models for predicting UK office building energy consumption from heating and cooling demands," *Energy and Buildings,* vol. 59, pp. 214-227, 2013.

[31] S. Papadopoulos and E. Azar, " Integrating building performance simulation in agent-based modeling using regression surrogate models: A novel human-in-the-loop energy modeling approach," *Energy and Buildings,* vol. 128, pp. 214-223, 15 September 2016.

[32] S. Wong, K. K. Wan and T. N. Lam, "Artificial neural networks for energy analysis of office buildings with daylighting," *Applied Energy,* vol. 87, no. 2, pp. 551-557, February 2010.

[33] A.-T. Nguyen, S. Reiter and P. Rigo, "A review on simulation-based optimization methods applied to building performance analysis," *Applied Energy,* vol. 113, pp. 1043-1058, January 2014.

[34] P. Torcellini, M. Deru, B. Griffith, K. Benne, M. Halverson, D. Winiarski and D. Crawley, "DOE Commercial Building Benchmark Models," in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2008.

[35] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings,* vol. 49, pp. 560-567, June 2012.

[36] A. N. Aijazi and L. R. Glicksman, "Comparison of regression techniques for surrogate models of building energy performance," in *ASHRAE and IBPSA-USA SimBuild*, Salt Lake City, UT, 2016.

[37] U.S. Energy Information Administration, "Commercial buildings energy consumption survey (CBECS)," 2012. [Online]. Available: https://www.eia.gov/consumption/commercial/. [Accessed 06 2019].

[38] J. C. Lam, S. C. Hui and A. L. Chan, "Regression analysis of high-rise fully air-conditioned office buildings," *Energy and Buildings,* vol. 26, no. 2, pp. 189-197, 1997.

[39] Google, "Google My Maps," 2019. [Online]. Available: https://www.google.ca/maps/about/mymaps/. [Accessed July 2019].

[40] S. S. Amiri, M. Mottahedi and S. Asadi, "Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S.," *Energy and Buildings,* vol. 109, pp. 209-216, December 2015.

[41] S. S. 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, no. 3, pp. 4256-4268, 2018.

[42] S. Papadopoulos, E. Azar, W.-L. Woon and C. E. Kontokosta, "Evaluation of tree-based ensemble learning algorithms for building energy performance estimation," *Journal of Building Performance Simulation,* vol. 11, no. 3, pp. 322-332, 2017.

[43] M. Castelli, L. Trujillo, L. Vanneschi and A. Popovic, "Prediction of energy performance of residential buildings: A genetic programming approach," *Energy and Buildings,* vol. 102, pp. 67-74, September 2015.

[44] J.-S. Chou and D.-K. Bui, " Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy and Buildings,* vol. 82, pp. 437-446, October 2014.

[45] M. Al Gharably, J. F. DeCarolis and S. R. Ranjithan, "An enhanced linear regression-based building energy model (LRBEM+) for early design," *Journal of Building Performance Simulation,* vol. 9, no. 2, pp. 115-133, March 2016.

[46] R. E. Edwards, J. New, L. E. Parker, B. Cui and J. Dong, "Constructing large scale surrogate models from big data and artificial intelligence," *Applied Energy,* vol. 202, pp. 685-699, September 2017.

[47] A. N. Aijazi, "Machine Learning Paradigms for Building Energy Performance Simulations," Massachusetts Institute of Technology, 2017.

[48] J. Sacks, W. J. Welch, T. J. Mitchell and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science,* vol. 4, no. 4, pp. 409-435, 1989.

[49] F. Ascione, N. Bianco, C. De Stasio, . G. M. Mauro and G. P. Vanoli, "Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach," *Energy,* vol. 118, pp. 999-1017, January 2017.

[50] X. Chen, H. Yang and K. Sun, "Developing a meta-model for sensitivity analyses and prediction of building performance for passively designed high-rise residential buildings," *Applied Energy,* vol. 194, pp. 422-439, 2017.

[51] W. Tian, J. Song, Z. Li and P. de Wilde, "Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis," *Applied Energy,* vol. 135, pp. 320-328, 15 December 2014.

[52] A. Chari and S. Christodoulou, "Building energy performance prediction using neural networks," *Energy Efficiency,* vol. 10, no. 5, pp. 1315-1327, October 2017.

[53] SEAI, "Introduction to Dwelling Energy Assessment Procedure (DEAP) for Professionals," Sustainable Eeergy Authority of Ireland, Ireland, 2017.

[54] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown and R. M. Pendyala, "Machine learning approaches for estimating commercial building energy consumption," *Applied Energy,* vol. 208, pp. 889-904, December 2017.

[55] G. Box and D. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society,* vol. 26, no. 2, pp. 211-252, 1964.

[56] R. Signor, F. S. Westphal and R. Lamberts, "Regression Analysis of Electric Energy Consumption and Architectural Variables of Conditioned Commercial Buildings in 14 Brazilian Cities," in *Seventh International IBPSA Conference*, Rio de Janeiro, Brazil, 2001.

[57] L. E. Melkumova and S. Y. Shatskikh, "Comparing Ridge and LASSO estimators for data analysis," *Procedia Engineering,* no. 201, pp. 746-755, 2017.

[58] G. M. Mauro, M. Hamdy, G. P. Vanoli, N. Bianco and J. L. Hensen, "A new methodology for investigating the cost-optimality of energy retrofitting a building category," *Energy and Buildings,* vol. 107, pp. 456-478, November 2015.

[59] Government of Canada, "Engineering Climate Datasets: Canadian Weather Year for Energy Calculation (CWEC)," 2016. [Online]. Available: http://climate.weather.gc.ca/prods_servs/engineering_e.html. [Accessed 2018].

[60] US Department of Energy's Building Technologies Office (BTO), and managed by the National Renewable Energy Laboratory, "EnergyPlus," May 2015. [Online]. Available: https://energyplus.net/. [Accessed October 2018].

[61] US Department of Energy (DOE), "Commercial Reference Buildings," November 2012. [Online]. Available: https://www.energy.gov/eere/buildings/commercial-reference-buildings. [Accessed September 2018].

[62] MathWorks, "Statistics and Machine Learning Toolbox," 2018. [Online]. Available: https://www.mathworks.com/help/stats/lhsdesign.html#f3025741. [Accessed November 2018].

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in {P}ython," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[64] D. Winiarski, W. Jiang and M. Halverson, "Review of Pre- and Post-1980 Buildings in CBECS – HVAC Equipment," Pacific Northwest National Laboratory, 2006.

[65] US Department of Energy (DOE), "Commercial Prototype Building Models - ANSI/ASHRAE/IES Standard 90.1 Prototype Building Model Package," 2016. [Online]. Available: https://www.energycodes.gov/development/commercial/prototype_models#90.1.

[66] Sidewalk Labs; Energy Profiles Limited (EPL), "Sidewalk Labs Canadian Commercial Office Buildings Study: Analysis of Energy Use and Performance," 2019.

[67] S. Philip and L. Tanjuatco, "Eppy (Python-based module)," 2013.

[68] N. R. C. NRCan, "Survey of Commercial and Institutional Energy Use (SCIEU)," 2014.

[69] E. C. Barnes and J. J. McArthur, "Building Energy Use Surrogate Model Feature Selection – A Methodology Using Forward Stepwise Selection and LASSO Regression Methods," in *IBPSA Building Simulation* , Rome, 2019.

[70] T. Hoyt, E. Arens and H. Zhang, "Extending air temperature setpoints: Simulated energy savings and design considerations for new and retrofit buildings," *Building and Environment,* no. 88, pp. 89-96, 2015.

[71] H. S. Rallapalli, "A Comparison of EnergyPlus and eQUEST Whole Building Energy Simulation Results for a Medium Sized Office Building," Arizona, 2010.

[72] I. Turiel, R. Boschen, M. Seedall and M. Levine, "Simplified energy analysis methodology for commercial buildings," *Energy and Buildings,* vol. 6, no. 1, pp. 67-83, 1984.

[73] "MapCustomizer," 2019. [Online]. Available: https://www.mapcustomizer.com/. [Accessed April 2019].

[74] C. Buratti, M. Barbanera and D. Palladino, "An original tool for checking energy performance and certification of buildings by means of Artificial Neural Networks," *Applied Energy,* vol. 120, pp. 125-132, May 2014.

[75] M. Rickert, A. Sieverling and O. Brock, "Balancing exploration and exploitation in sampling-based motion planning," *IEEE Transactions on Robotics,* vol. 30, no. 6, pp. 1305-1317, 2014.

[76] A. Prada, A. Gasparella and P. Baggio, "On the performance of meta-models in building design optimization," *Applied Energy,* vol. 225, pp. 814-826, 2018.

[77] S. Singaravel, J. Suykens and P. Geyer, "Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction," *Advanced Engineering Informatics,* vol. 38, pp. 81-90, October 2018.

Glossary

TERM	DEFINITION
Annual Building Energy Use	In this study, annual building energy use is the sum of the annual heating, cooling, pump and fan energy uses generated by the EnergyPlus simulation
Annual Total Site Energy Use	In this study, annual total site energy use is the sum of all annual energy uses affecting the building. It therefore does not include the energy use related to items external to the building, such as exterior lighting.
Base Model	The EnergyPlus input data file (IDF) used as the IDF on which all samples are based. Any inputs that are not modified through the features varied between samples will be consistant for all samples in the dataset.
Box-Cox	Method used to transform data into a normal distribution [58]
Combined Feature	A feature calculated by combining, in a linear equation, two or more of the original 71 features.
Elastic Net	Both the L1 (LASSO) and L2 (Ridge) regulators are applied to the cost function.
Embedded Feature Selection	Embedded methods use properties of specific learning algorithms to select features that best contribute to the model accuracy.
Feature Set	Combination of the input features
Filter Feature Selection	Filter feature selection methods are independent of the learning algorithm and often 'score' each feature to the target variable. This score is used to determine which features are kept for model training. Filter methods often do not account for the relationship between features.
Hyperparameter	In relation to the learning algorithm, the hyperparameter(s) are the terms within the algorithm that can be modified (or tuned).
Input Data File (IDF)	The text file EnergyPlus uses to run the simulation.
Input Feature	The variable(s) describing the target variable. In this study, the input features are the building attributes that were modified in each sample.
Latin Hypercube Sampling (LHS)	A sampling plan that extends the Latin square, a grid with one sample per row and column, to multi-dimensional space. The sampling plan distributes the samples through the entire design space

TERM	DEFINITION
Least Absolute Shrinkage and Selection Operator (LASSO)	Referred to also as the L1 regulator, LASSO applies a penalty term to the cost function that shrinks the regression coefficients, some down to a value of zero, thereby removing it from the model
Normalization	Method of standardizing dataset so that the range of the input features are on the same scale
Pearson Correlation Coefficient	A coefficient describing the linear relationship between two variables. When the coefficient value is 1 or -1, the relationship between the two variables is positively or negatively linear, respectively. A coefficient of 0 indicates no linear relationship between the variables.
Random Seed	A value assigned to the random splitting of data. A random seed of the same value will split the data in the same way each time it is run.
Ridge	Similar to LASSO, Ridge, also referred to as the L2 regulator, applies a penalty term to the cost function that reduces the regression coefficients down to values close to zero
Surrogate, meta, response surface, emulator model	Fitting computer-simulated data to a surface in order to predict results from the available data without the use of the expensive code [8].
Target Variable	The variable the learning algorithm is being trained to predict. In this study, annual building energy use was the target variable.
Test Dataset	A subset of the overall dataset that is held out of the training and validation datasets and is used to evaluate the accuracy of the model to predict data not previously seen.
Training Dataset	A subset of the overall dataset that is used to fit the data to a surface.
Validation Dataset	A subset of the overall dataset that is used to evaluate the trained model on a set of data not used to fit the model. This dataset is used to evaluate and compare the model at the stages to final model development.
Wrapper Feature Selection	Wrapper methods use subsets of the features to fit the model and compare model behaviour and performance for each subset. Common wrapper methods include forward stepwise selection where features are iteratively added to the model and backwards stepwise elimination where features are iteratively removed.