Advanced Cluster and Predictive Analysis Tool Development for Commercial Office Real Estate Energy Usage

by

Carleen Lawson

BAS, Carleton University, 2016

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Building 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.

ACKNOWLEDGEMENTS

I would first like to thank my academic supervisor Jennifer McArthur of the Department of Architectural Science at Ryerson University. Professor McArthur was an endless source of in-depth knowledge and passionate support. She was available at all hours to help me push past the struggles of my research journey and deserves a standing ovation.

I would also like to thank my second reader, Dr Michael Brooks, whose company REALPAC chose to collaborate with Ryerson, providing their 20 by ‘15 Dataset for this Major Research Project. I am gratefully indebted to him for providing a summer internship at his company and for his very valuable comments on this thesis.

In addition, I owe Jennifer McArthur, Michael Brooks and Mitacs gratitude and recognition for securing me a Mitacs accelerate grant to support my research.

Finally, I must express my very profound gratitude to my Mother, Careen Lawson, for providing me with unfailing support and continuous encouragement throughout the evolution of my education, character and passions. This accomplishment would not have been possible without her. Thank you.

Carleen Lawson

Advanced Cluster and Predictive Analysis Tool Development

for Commercial Office Real Estate Energy Usage

Master of Building Science 2019, Carleen Lawson

Building Science Program, Department of Architectural Science, Ryerson University

ABSTRACT

From 2009-2015, REALPAC collected monthly energy usage and building characteristics for over 500 buildings in the 20 by ‘15 Energy Benchmarking Survey (REALPAC, 2009). While preliminary analysis had been completed on this dataset, this research undertook an in-depth statistical analysis of the data to identify trends and important variables. Eight machine learning algorithms were employed to predict energy usage as a function of previous energy use and select physical features. The dataset did not possess the appropriate variables to predict such usage accurately. Characteristics such as building system efficiency, construction assemblies, condition, compactness, and window to wall ratio are thus recommended for inclusion in future data-gathering initiatives.

Contents

1 Introduction

Climate change is an urgent global concern. Human driven green house gas emissions have launched the planet into the Anthropocene era where we are experiencing a drastic surge in sea levels, extreme weather events and a loss of biodiversity. The results have led to severe health and habitat complications among vulnerable populations (Allen, et al., 2018). Currently, renewable energy sources have yet to retire the burning of fossil fuels for electricity. In 2017, over 33 million tonnes of coal and 8 million cubic meters of natural gas were burned for electricity production in Canada (Statistics Canada, 2017). This practice, along with a high dependence on fossil fuel (e.g. natural gas) heating significantly increases atmospheric pollution levels, which trap heat and fuel the global warming crisis (Warren & Lemmen, 2014).

In Canada, the commercial building sector accounts for 15% of the national energy consumption (Statistics Canada, 2012). As energy conservation is the most environmentally and financially sustainable energy resource (Ontario Ministry of Energy, 2017), there is presently a strong push to understand energy usage and trends within buildings. Voluntary energy benchmarking has been established, through mediums such as Energy Star Portfolio Manager and the Real Property Association of Canada (REALPAC) 20 by '15 Energy Benchmarking Survey (20 by '15), to increase public awareness of energy consumption. Mandatory programs, such as Ontario's *Energy & Water Reporting and Benchmarking* (EWRB) for Large Buildings, have also been introduced by governments to improve the collective understanding of energy consumption and inform energy improvements in the built environment.

1.1 Research Motivation

The motivation of this Major Research Project (MRP) is to analyze the results of the large Canadian office energy benchmarking survey, 20 by '15 (REALPAC, 2018). To date, this dataset has only been the subject of preliminary analysis, and REALPAC felt that an in-depth statistical analysis would provide the office sector with increased insight into both their consumption and efficiency strategies.

1.2 Research Objective

The objective of this research is to perform an advanced statistical analysis of annual office energy consumption and identify a predictive model to inform sustainability initiatives. The outcomes of this will hopefully empower sustainability campaigns, such those operated by

REALPAC and the Canada Green Building Council, with the ability to target office real estate predicted to have poor energy performance. Because such real estate possesses a higher improvement potential than top performers, directing resources towards these buildings has the potential to significantly reduce Canada's overall greenhouse gas emissions.

1.3 Research Questions

Two key research questions have framed this research:

1) What are the most significant predictors of Energy Use Intensity (EUI)?

2) How accurately can one predict a building's EUI based with past energy consumption and limited physical building data?

2 Literature Review

The following section reviews the current literature surrounding commercial office real estate energy mandatory reporting, the history and structure of REALPAC's 20 by '15 survey, energy benchmarking studies, machine learning algorithms, and their use for building energy consumption prediction.

2.1 Commercial office Real Estate Energy Mandatory Reporting

Initially, benchmarking was a word exclusive to topography and indicated a geological reference point (Pérez-Lombard, et al., 2009). In the 1970's, companies began using benchmarking tools as means to compare productivity of processes with similar parameters (Pérez-Lombard, et al., 2009). The concept of benchmarking buildings first arose in the beginning of the 1990's to compare energy consumption of buildings with similar characteristics (Pérez-Lombard, et al., 2009).

This development was spurred by government concerns over political instability in regions supplying energy (Pérez-Lombard, et al., 2009). Two such examples are the Iranian revolution and the first Gulf war which threatened secure access to oil. Many nations responded by focusing on energy efficiency, specifically in the building sector, which was targeted as its energy consumption surpassed both industry and transportation sectors (Pérez-Lombard, et al., 2007). Pérez-Lombard et al (2009) define energy efficiency as "consuming less energy while providing equal or improved building services".

In 2013, the Ontario Ministry of Energy adopted a *Conservation First* policy in their *Long-Term Energy Plan*. Energy use reduction is the most environmentally and financially sustainable energy resource (Ontario Ministry of Energy, 2017). The province pushed energy conservation as an alternative to the mass expansion of energy infrastructure and as a solution to consumers struggling with rising energy costs (Ontario Ministry of Energy, 2017).

To further support the implementation of the *Conservation First policy*, on February 6th, 2017 the province of Ontario filed a regulation for the *Reporting of Energy Consumption and Water Use* under the Green Energy Act (O. Reg 20/17). The regulation required all commercial and multi-unit residential properties with a gross floor area of 50,000 sf or greater to report their annual energy usage through the *Energy Star Property Manager* by July 1st of the following calendar year (O. Reg. 20/17). The regulation had implemented phased deadlines which required

250,000 sf buildings to file in 2018, 100,000 sf buildings to file in 2019 and 50,000 sf buildings to file in 2020 (O. Reg 20/17).

2.2 20 by '15 Initiative

Before the mandatory energy and water benchmarking was legislated, the Real Property Association of Canada initiated its own benchmarking survey to guide commercial office buildings further down the path of sustainability. The survey challenged participants to achieve a Total Building Energy Use Intensity of 20 ekWh/ft²/yr by 2015 (REALPAC, 2009). The project was dubbed '20 by '15' for short (REALPAC, 2009). REALPAC predicted that reaching this target will save $1.85 billion and 7.5 megatonnes of greenhouse gas emissions every year (REALPAC, 2009). To this end, REALPAC surveyed its commercial office members to collect monthly energy usage and building characteristics for over 500 Canadian commercial buildings from 2009 to 2015 (REALPAC, 2009). Participating buildings were required to be primarily used as a commercial office facility, possess a minimum exterior area of 20,000 sf, and have a maximum vacancy rate of 30%.

Data was entered by the commercial office members into the REALPAC Energy Normalization Database (the Database) using an online portal according to the process described in detail in Appendix I. This portal was fashioned after Energy Star *Property Manager*, an online tool created by the US Environmental Protection Agency to benchmark building energy, water and greenhouse gas emissions (ENERGY STAR, n.d.). During each year of the survey, REALPAC prepared a report summarizing key trends and the cumulative findings to date, summarized in Table 2-1.

Table 2-1 Summary Statistics of Annual Normalized Energy Use Intensity of Canada-wide Data Set

Year	No. of Buildings	Data Set Range Min. (ekWh/ft2/yr)	Data Set Range Max. (ekWh/ft2/yr)	Mean Normalized EUI (ekWh/ft2/yr)	Median Normalized EUI (ekWh/ft2/yr)	No. of Buildings at the 25th Percentile or lower	No. of Buildings at or Below 20.0 ekWh/ft2/yr	Proportion of Data Set at or Below 20.0 ekWh/ft2/yr
2010	357	5.5	77.9	29.4	28.1	89	40	0.11
2011	367	8.2	70.6	27.7	26.7	92	53	0.14
2012	370	10	85.7	26.6	24.8	93	83	0.22
2013	487	9.7	119.4	28.4	25.6	122	102	0.21
2014	470	12.1	140	29	25.9	118	81	0.17
2015	437	11.6	137.8	29	25.7	109	94	0.22

Reprinted (adapted) from (REALPAC, 2017)

In order to compare the data, REALPAC normalized the Energy Use Intensity for building characteristics, weather, location, and against the base year to improve comparability between each individual building despite variations in building operation, climate and site (REALpac, 2015).

2.2.1 Normalized Building Energy Use (in the current year)

First, the Database calculated the total building energy use intensity per square foot. Then the value was used to normalize the EUI for other building characteristics such as High Intensity or Exceptional Tenant Energy Use, Annual Vacancy, Occupant Density and Operating Hours (REALpac, 2015).[1] This metric permitted the comparison of buildings with various characteristics; however, they still must pertain to the same location and year.

2.2.2 Weather Normalized Building Energy Use to Base Year 2009

Next, the Database used the "Normalized Building Energy Use" metric and removed the influence of weather and climate change (REALpac, 2015). The impact of a particular location's climate was calculated via Heating Degree Days (HDD) and Cooling Degree Days (CDD). The Database used the annual HDD and CDD data from Environment Canadas weather stations located in airports by the closest major city to the building. The data was accessed via the National Climate Data and Information Archive website (www.climate.weatheroffice.gc.ca).

As HDD and CDD trends are shifting due to climate change, the energy use was normalized to the 2009 base Year, so it did not appear as if a building's performance was gradually worsening over time.

[1] This document will visually distinguish dataset variable names through Verdana font.

Equation 1 Weather Normalized Building Energy Use to Base Year 2009

$$Weather\ Normalized\ Building\ Energy\ Use\ to\ Base\ Year\ 2009$$
$$= \frac{\text{annual}\ \frac{\text{HDD}}{\text{CDD}}\ \text{in the current year}\ *}{\text{annual}\ \frac{\text{HDD}}{\text{CDD}}\ \text{in 2009}\ *} \times 100\%$$
$$\times\ Normalized\ Building\ Energy\ Use\ (in\ the\ current\ year)$$

**for the Closest Major City*

The metric "Weather Normalized Building Energy Use to Base Year 2009" permitted comparability between buildings experiencing different weather patterns across different survey years.

2.2.3 Location & Weather Normalized Building Energy Use to Base Year 2009 in Toronto, ON

Lastly, the Weather and Use Normalized Building Energy consumption was normalized according to weather (HDD and CDD) differences between the Closest Major City to the building and Toronto's Lester B. Pearson International Airport weather station (REALpac, 2015). The metric "Location & Weather Normalized Building Energy Use to Base Year 2009 in Toronto, ON" permitted comparability between buildings despite variations in location. This is the final metric used to compare the energy use of all of the buildings across Canada.

2.3 Energy Benchmarking Studies

Building energy surveying and statistical analysis has been used by many researchers to benchmark buildings using various methodologies.

Mills (2016) investigated the potential for Action-Oriented Benchmarking in the non-residential built environment, criticizing conventional energy benchmarking for inspiring sustainable action without any practical guidance. Mills stated that this shortcoming of conventional energy benchmarking can be addressed through disaggregated approaches as opposed to conventional whole buildings methods. A deeper investigation and documentation of the building's energy systems, metrics, details, and end uses can lead to the identification of cost effective, individually customized energy saving measures. Mills predicted that benchmarking will become more popular as it is pushed forward by public policies**.** He suggested enhanced utility bills and incentivized benchmarking to add future research initiative needed to improve the actionability of benchmarking results.

Capozzoli *et al.* (2016) analyzed 100 Healthcare centers to evaluate energy use reference values. The researchers applied a Linear Mixed Effect Model to heterogeneous data sets and found a best fit of 0.01% error, average fit of 15% error and worst fit of 38% error. The testing R^2 error was found to be 0.96 and residuals were randomly distributed with a mean of zero. These results indicate a robust model with a high estimation ability.

Lee and Lee (2009) employed data envelopment analysis to benchmark a sample of 47 Taiwanese government buildings. This method, developed by (Charnes, et al., 1978) has been commonly used in other fields to assess production efficiencies and involves separating variables into management factors and scale factors. Using this technique, the data set was normalized to remove the effect of scale factors and focus on energy management. In this study, final efficiency assessments were given as a percentage and the authors found that poor management, not scale, was the key indicator of poor energy performance.

2.4 Machine Learning

Machine learning is a valuable technique for data analysis and can be categorized as either supervised or unsupervised. Unsupervised learning creates a model without cross referencing the correct data labels and instead relies on patterns and grouping between the independent variables. In supervised learning, the algorithm is provided with a labelled ("training") dataset to develop a model. One unsupervised method – Principal Component Analysis (PCA) – and six supervised learning methods – Linear Discriminant Analysis (LDA), k-Nearest Neighbours (KNN), Multiple Linear Regression (MLR), Support Vector Machines (SVM), Decision Trees, and Artificial Neural Network (ANN), have been considered in this research and are described in the following sections. The specified algorithms were selected to test both supervised and unsupervised methods and based on the outcomes of previous studies which are discussed in Section 2.3.

2.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised learning method that identifies the combinations of variables accounting for the most significant variation within the dataset. PCA is one of the oldest and most prevalent multivariate statistic techniques and is used in almost every scientific discipline (Abdi & Williams, 2010). The origins of PCA can be found in the works of many mathematicians such as Cachy, Pearson, Jorden or Cayley, Selversler and Hamilton. The

instantiation of PCA is accredited to Harold Hotelling after his 1933 paper, “Analysis of a Complex of Statistical Variables into Principal Components” (Hotelling, 1933). Hotelling’s research mainly pertains to the psychological discipline, which referred to the fundamental variables of a data set as the ‘mental factors’. Hotelling chose to instead refer to these as components so as to not be confused with the mathematical definition of ‘factors’.

Hotelling’s Method of Principal Components uses the Eigenvalues and the Eigenvectors of a data set’s correlation matrix to determine the combinations of independent variables accounting for the most variation in the dependant variables (Hotelling, 1933). The Eigenvalues and Eigenvectors are then used to transform each data point and project them on to the principal component axis. In other words, the Method of Principal components finds linear combinations of the existing variables and uses them to attain new components. Principal components (PCs) are those which account for the most variability. To prevent variations, PCA first involves scaling the data to ensure variables with differing units can be compared to one another. Hotelling writes that the data must be normalized with a mean of zero so the variance between the standard deviations of each component may be compared. The number of PCs cannot be higher than the original number of variables. The first principal component (PC1) is defined as the component with the most variance. The second principal component (PC2) is orthogonal to the first principal component and accounts for the second most variance. This pattern continues for each consecutive PC, with all PCs mutually orthogonal to avoid confounding. There are no more PC once the entirety of the variance is accounted for. It is important to note that the relative position of each data point to another remains the same; the axes are simply rotated.

Figure 2-1 illustrates the two PCs representing the dataset both in the initial vector space (top) and in a simplified (2D) vector space (bottom).

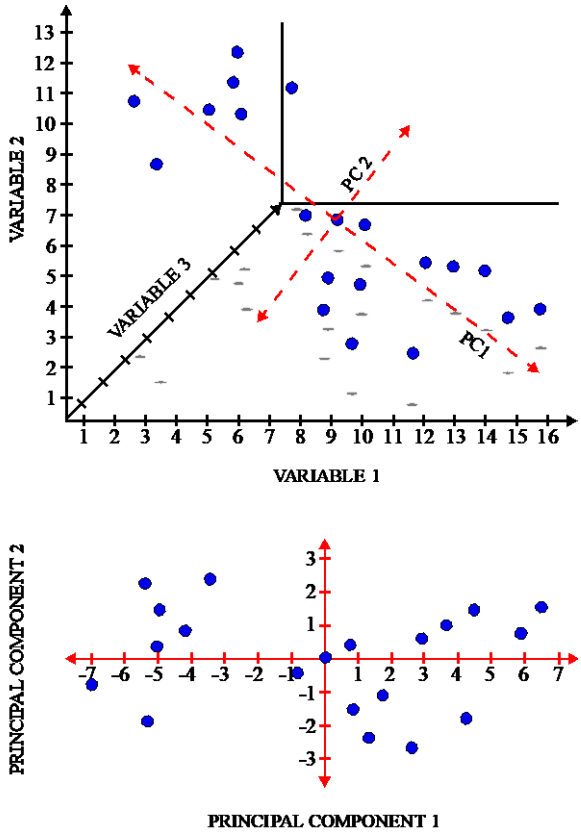


Figure 2-1 Illustration of PCA Concept

Abdi and Williams stated the goal of PCA is to compress the data set so only important information is retained and the unnecessary variables are discarded (2010). This concept is referred to as dimensionality reduction. In addition, PCA is used to analyze the structure of the observations and the variables.

After a PCA is performed, the Scores and loadings are returned (Bro & Smilde, 2014). The Scores are the new lengths for each observation. They are typically represented on a scatterplot graph with a principal component on each axis. If clusters are observed, it indicates that it is possible to classify a sample given the variables represented by the principle components on the graph (Bro & Smilde, 2014). Score plots are also useful for quickly detecting outliers, which are fall outside of data clusters.

PCA will return a loadings plot which is a matrix containing the Eigenvectors in each column, arranged from largest to smallest (Abdi & Williams, 2010). Each column represents a principal component, with Principal Component 1 being the first column, Principal Component 2 being the second and so on. The loadings rate the contribution of each original variable to each new

principal component. In his “Principal Component Analysis” monograph, Jolliffe recommends that a loading threshold above or 0.7 or below -0.7 should be considered significant (1986). Loadings are typically visualized using biplot one may quickly analyze the relationship between the loadings of two principal components (Abdi & Williams, 2010).

In addition, the Eigenvalue for each principal component will be returned. The Eigenvalues are used for calculating the variance captured by each principal component. To do so, a particular Eigenvalue is divided by the sum of all the Eigenvalues. The result is the variance retained by the principal component represented as a percentage.

Both the Eigenvalues and the loadings are used to perform dimensionality reduction on the data set. First, the boundary is determined with the aid of a scree plot (Abdi & Williams, 2010). The Eigenvalues are plotted according to size and the resulting graph is visually analyzed for the presence of a noticeable change in slope or an ‘elbow’ as seen in Figure 2-2. If an elbow is identified, the principal components to the right of that point are disregarded.

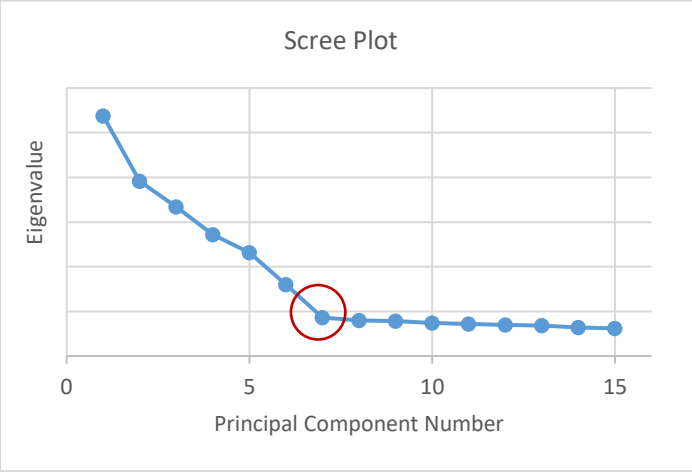


Figure 2-2 Ideal Scree Plot with Elbow

If an elbow is not identified, then it is considered good practice to retain enough principal components to explain 90% of the variability (Ringnér, 2008). Refer to Figure 2-3 and Figure 2-4.

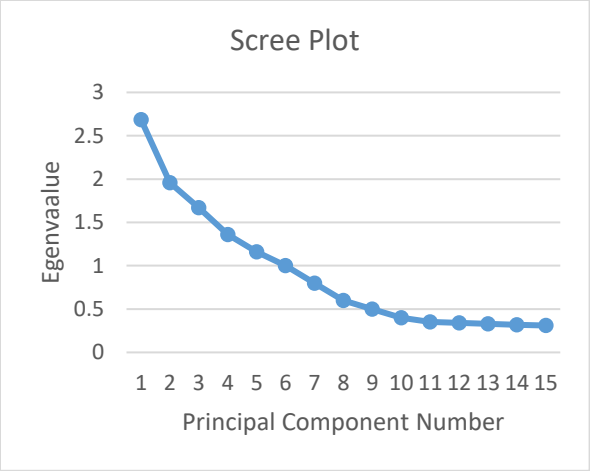


Figure 2-3 Non-Ideal Scree Plot with no Elbow

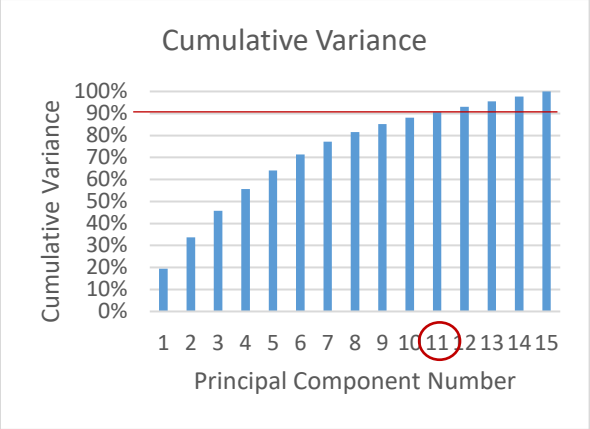


Figure 2-4 Sample PC Cumulative Variance Graph

Once the important principal components are determined, the loadings plot is reviewed to determine which variables are significant. Joliffe outlines two different approaches for this.

First, the largest principal component is analyzed. Any loading above 0.7 or below -0.7 marks a variable which contributes significantly to the corresponding principal component. Of all the significant variables, the one with the highest absolute loading is selected to represent that component. The approach proceeds by analyzing the next largest component and continues until the smallest retained principal component. If a variable has been selected for a principal component with a higher variance, it cannot be selected again; the variable with the next highest absolute loading value is selected.

The second approach is performed in reverse. It uses principal components which were not retained to determine which variables should be discarded. The most significant variable for the smallest principal component would be removed from the data set. This method continues for each progressively larger discarded principal components. The justification for this procedure is that the smaller principal components highlight redundant variables. In other words, insignificant principal components are represented by insignificant variables.

PCA can also be used to predict new observations given the same variables (Abdi & Williams, 2010). This is performed by training the PCA model on only a sample of the dataset and later testing by predicting the remaining observations according to a random effects model. The PCA model performance is typically evaluated using computer executed resampling techniques. Two examples are the bootstrap and cross validation methods. The predicted results are compared against the actual observations and the accuracy of the model is assessed.

2.4.2 Linear Discriminant Analysis

LDA is a supervised learning algorithm where the dataset is transformed such that it can be projected onto axes that show the maximum differentiation between different classes and minimum differentiation within each class. The resulting visualization is a graph where the individual classes are clustered as best as possible with the most distance between each the centroids of each class. Figure 2-5 illustrates this concept graphically.

Tharwat, Gaber, Ibrahim and Hassanien (2017) describe the main goal of LDA as identifying redundant or dependant variables to be removed by transforming features in a higher dimensionality space to a lower dimensionality space. LDA is not useful when the number of dimensions is greater than the number of samples (Tharwat, et al., 2017).

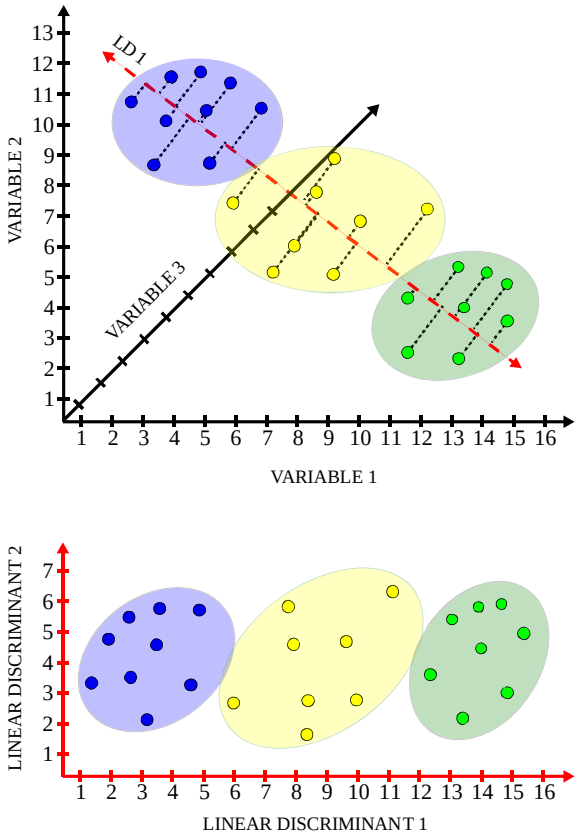


Figure 2-5 Illustration of LDA Concept

2.4.3 k-Nearest Neighbours

k – Nearest Neighbours (KNN) is a supervised learning algorithm used to classify new observations (Stamp, 2017). The method was first introduced as the k-nearest neighbour rule by Fix and Hodges in their technical report "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties" (1951).

Using a training set of classified data and a distance metric the KNN algorithm classifies a new observation, X, by evaluating the known classes of the training data points neighbouring nearest to X. The number of observations evaluated is represented by *k*. The class of X is determined by the most frequent class appearing within the *k* nearest neighbours. For example, in Figure 2-6, when k=3, the majority (2) of the three closest points are red, and thus the unknown point would be tagged red. Conversely, if *k*=5, the algorithm would predict the unknown point would be green, as majority (3/5) of the five closest points are green. A challenge of KNN models is

correctly predicting the class of datapoints that fall near the threshold of regions as they may be mislabelled due to their proximity.

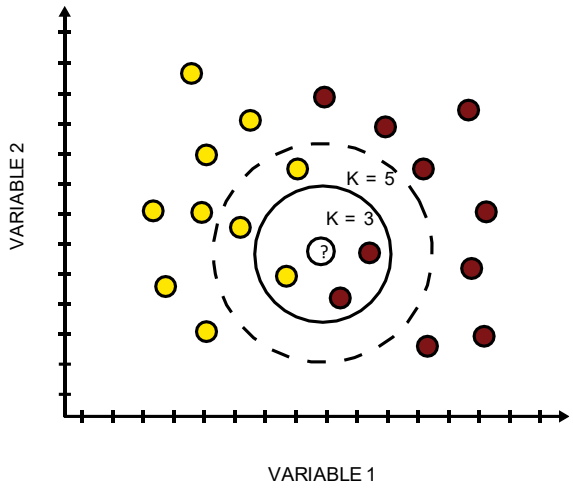


Figure 2-6 KNN Concept Diagram

The KNN algorithm does not assume relationships between the variables and the classes (Shumeli, 2017). The class is simply determined by *k* and the distance metric selected – options include Euclidian, Mahonobis, and Chebychev —and therefore, the chosen distance metric is very important to the success of the algorithm: "the most desirable distance function is the one for which a smaller distance among samples implies a greater likelihood for samples to belong to the same class" (Mucherino, et al., 2009).

The size of *k* also affects the quality of the KNN outputs (Knox, 2018). Low *k* values are more sensitive to individual data points, resulting in noise and more erratic class boundaries. Large *k* values are less sensitive to local variations in class clusters, producing smoother boundaries yet sacrificing accuracy. Moderate k values are ideal as they achieve a balance between sensitivity and noise reduction. Duda *et al.* (2012) suggest that the ideal k is the square root of the number of instances in the data set. However, the cross validation of K-fold training and validation sets is well established statistical technique for testing various values of *k* and determining the values with the lowest misclassification rate (StatSoft Inc, 2013).

When graphed, the KNN plotted probabilities show a boundary between 'winning' and 'losing' classes. This allows one to quickly determine which class a new observation would be categorised as.

2.4.4 Support Vector Machines

Support vector machines (SVM) is a supervised learning technique that, unlike LDA and PCA, directly generates a classification without outputting a score (Stamp, 2017). SVM is considered to be a powerful function and is successfully employed by a wide range of disciplines. The function provides the following advantages (Steinwart & Christmann, 2014):

1. High ability to learn with few free parameters
2. Robust against outliers and disruptions in the model
3. High computational efficiency

SVM can be applied to either raw data or to a set of Scores output by a different algorithm (i.e. PCA) (Stamp, 2017). SVMs are typically used for binary classification and therefor the classification loss function is used with two response values, +1 or -1, which correspond to class labels (Steinwart & Christmann, 2014).

The governing concepts of SVMs (Stamp, 2017) are listed below and illustrated graphically in Figure 2-7:

1. Hyperplane separation – The labelled data is separated into multiple classes based on a hyperplane. A hyperplane is a subspace with one less dimension than the dataset.
2. Margin Maximization - during the construction of the hyperplane, the margin separating the classes are maximized. The margin is the smallest distance between any data point and the hyperplane.
3. Use a high dimensional space - This concept may be counter intuitive to previously discussed models, but SVMs perform best on higher dimensional datasets as they can better inform the discovery of a parting hyperplane.
4. Kernel Trick - Uses a kernel function to spatially transform the non-linearly separable data with the hope of improving separability.

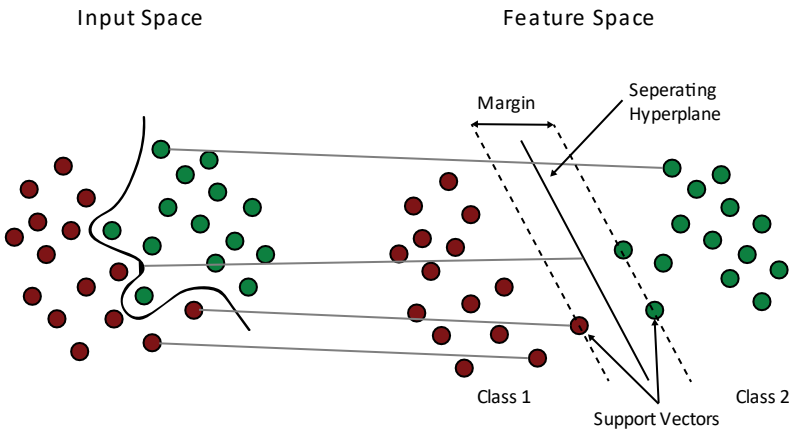


Figure 2-7 Support Vector Machine Conceptual Illustration

When the data is not separable in *input space* (given by the actual dataset), a kernel can be used to map the data points into *feature space* which is defined by a different dimensionality (Campbell & Ying, 2011). This may improve the separability.

Though SVMs were originally used for binary classification, they can be adapted to allow for multi-class classification (Campbell & Ying, 2011). This is typically performed using a series of one-against-all classifiers: *C* separate SVMs are constructed, with *C* representing the number of classes in the data. For each SVM, a different class is labelled as positive and the remaining class samples are labelled as negative. It should be noted that this one-against-all approach is weak when the number of classes is close to the number of samples as it results in a large imbalance within each separate SVM. SVMs can also be used for regression analysis.

As with most machine learning algorithms, the success of SVMs are also deteriorated by noise and outliers (Campbell & Ying, 2011). To improve the generalization of noisy SVMs, the algorithms can adopt 'soft margins' which allow for some data points to fall inside the margin. A validation study can be performed by training the SVM on training data using various margin width to discover the best value for the parameter.

The success of an SVM can be evaluated via the classification accuracy or the validation error.

2.4.5 Decision Trees

The decision tree technique is a supervised learning algorithm which is popular, intuitive, and comprehensible as it is closely modeled on human reasoning (Kotsiantis, 2013). Decision trees combine a sequence of logical tests where each test evaluates a numeric feature against a

threshold or a nominal feature against a range of possible values. The model decides an outcome based on if the threshold was achieved or nominal value observed and then branches towards a subset of different logical questions which split the data. After an observation is filtered through all of the model's decisions, it is classified according to the most frequent class within that same region. Depending on the type of decision tree (regression or classification) the final return is either a numeric value or class prediction for each now observation. The success of the model is either evaluated by the error rate which is the percentage of misclassified observations, or the accuracy rate which is the percentage of correctly classified observations.

The key elements of a decision trees are: root, node, branch, and leaf, as illustrated in Figure 2-8. A node represents a test that the observation is put through. An example could be "Is the Closest Major City Vancouver, BC?". The result determines the subsequent test. The root is the first node and is the test with the highest predictive capacity. A leaf represents the model's final prediction (i.e. Good or EUI =19 eWh/sf/year). Branches represent the junctions between roots, nodes and leaves.

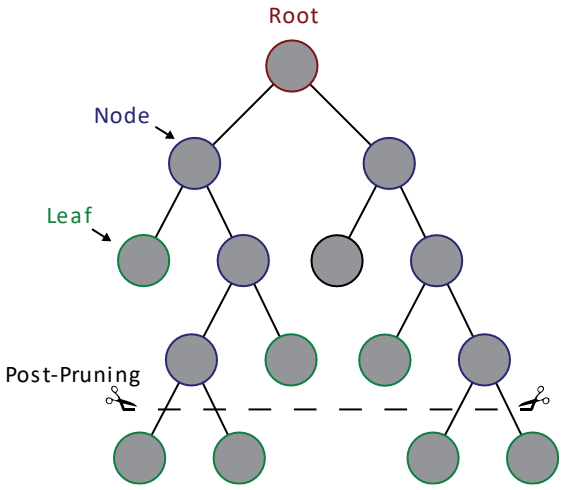


Figure 2-8 Decision Tree Concept Diagram

Decision tree algorithms develop models through automatic induction, which contains two main phases: growth and pruning (Kotsiantis, 2013). In the growth phase, a dataset with known classes (or numeric outputs for regression trees) is provided for training. The algorithm looks for patterns or generalizations in the data by evaluating all possible decisions to determine which

best split the data; this is evaluated based on a minimization of entropy (the degree of mixing) in the resultant split nodes. This process is repeated until all observations within a decision (node) fall within the same class or a stopping criterion is reached.

The more complex the decision tress, the lower the generalization and predictive accuracy rate, therefor pruning is performed resulting in near-optimal models which employ efficient heuristics (Kotsiantis, 2013). The following procedures can be preformed to reduce overfitting (Kotsiantis, 2013):

- Pre-pruning: terminates a branch prematurely during training when a stopping condition is reached.
- Post-pruning: after a model is generated, certain branches are retroactively removed
- Data pre-processing: reduces the features of the dataset until an optimal number of characteristics are reached to build a simpler tree.

Many different decision tree algorithms have been established to address common disadvantages such as poor generalization or time-consuming model training. According to the Law of Large Numbers (Bernoulli, 1773), as the quantity of random events increases, the variance between the probable value and the average actual value minimizes. To reduce variance without increasing bias through overfitting, a "Random Forest" algorithm (Breiman, 2001) was developed.

Typically, decision trees consider all features to split a node, however it has been found advantageous to only evaluate a random subset of variables when performing a node split (Breiman, 2001). Radom forests makes use of the principle of *bagging* and randomly subsets the variables of a training dataset with replacement. The quantity of random variables in each subset is referred to as *F* and is kept consistent for each node split (Breiman, 2001). Breimen suggests that F be "the first integer less than log2M+1, where M is the number of inputs" (2001).

Multiple decision trees, or a 'forest', are created for each data subset and are generated until a stopping condition, such as a predetermined number of nodes, is reached (Breiman, 1994). A majority vote is performed to predict an object class, the class receiving the most votes is selected. To predict a numerical outcome as with regression trees, the predicted values are averaged. This technique has been demonstrated to increase classification accuracy rates over individual decision trees with a reduced likelihood for overfitting. Note that the risk of

overfitting is slightly higher for regression analysis as the results are selected by an average rather than by a majority vote.

2.4.6 Artificial Neural Networks

An Artificial Neural Network is a complex supervised learning algorithm modelled after the nonlinear, parallel processing of the human brain (Haykin, 1998). The network of neurons in the human body are capable of experiential learning and knowledge-storing. Neurons communicate with each other as follows: one neuron will receive stimulus, an electrical message, from a receptor. If the stimulus is strong enough to reach the neuron's activation threshold (determined via prior learning), the neuron will pass the message forward to an effector which produces a discernable response. If the stimulus is too weak, the information is passed back from the neuron to the receptor, exhibiting feedback.

ANNs engage a large web of simple, interconnecting processing units or nodes, which mimic neurons (Haykin, 1998). Refer to Figure 2-9 for a Conceptual illustration of an ANN diagram. There are three types of nodes: input nodes, output nodes and hidden nodes (Wu & Feng, 2017). In a trained model, each input (i.e. data variable) is connected to a hidden node which is also connected to each output via synaptic weights (Haykin, 1998). These weights have a numeric value which stores the model's knowledge (gained through training). The larger the absolute value of a synaptic weight between two nodes, the stronger their connection.

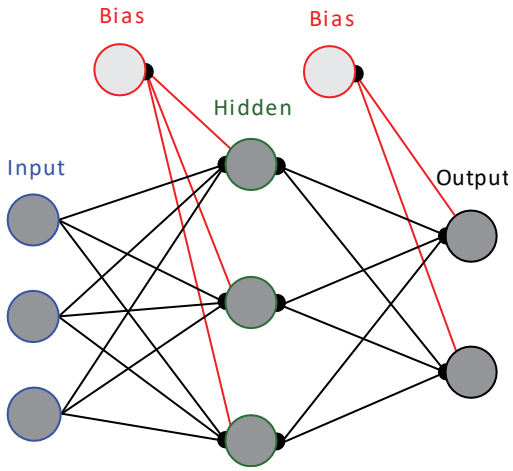


Figure 2-9 ANN Concept Diagram

Model learning is improved by bias units. Each hidden node and output nodes are influenced by a bias unit which is consistent for all nodes in a layer. A Bias unit stands alone and does not proceed another unit or layers.

To reduce the error rate and create a more robust ANN model back propagation (Werbos, 1974) or resilient backpropagation (Riedmiller & Braun, 1993) are often employed.

Back Propagation calculates the partial derivative of a single weight, while keeping the remaining weights the same (Werbos, 1974). The goal is to find the weight that returns the lowest error rate. The sign (+/-) and magnitude (-1 to +1) of the weight are combined with a small learning rate (the same used for all weights) to calculate the weight change. The error rate is recalculated at each interval of weight change. Once the error rate increases it indicates that the minimum was identified at the previous weight which is then selected for the final model. The process is repeated for the remaining weights. The disadvantage of back propagation is that it is slow as it involves many calculations.

Resilient Backpropagation (Rprop) was developed to increase the speed of back propagation (Riedmiller & Braun, 1993). Here, only the sign of the partial derivative is calculated and the learning rate is different for each weight (adapted during training). If the signs are the same for both previous and current partial derivatives, this indicates the weight value has not moved past the point of minimum error and to continue in the same direction. Larger learning rates may be used to increase the speed of the process. Once the signs are not the same, indicating an increase in the error rate, the previous smaller learning rate is used with the previous iteration, to reduce the speed of weight change. This process continues until the error minimum has been found.

ANNs provide many advantages such as adaptability, self organization and self learning. It is not without its disadvantages though. ANNs are black box models and the hidden layers of the model are generally too complex to be comprehended and analyzed. Therefore, it is difficult to interpret the data structure from the model. ANN can only be used to predict an outcome. It is also sensitive to overfitting, especially with a small or low variance data set. A solution to this shortcoming is K-fold cross validation.

2.5 Use of Machine Learning for Energy Prediction

The previous section touched on machine learning as a form of analysis for benchmarking energy consumption. A range of machine learning approaches have been applied to energy prediction. Several papers provide a detailed summary of the body of literature on this topic.

Yu et al reviewed benchmarking analysis via the regression and ANN methods on a dataset with 55 residential buildings in Japan (Wu & Feng, 2017). They investigated the accuracy and benefits of developing decision trees. They found the results were easier for users to extract information. When applied to the training accuracy in this study was 93% while test data accuracy was similar at 92%.

Artificial Neural Networks have been used to predict energy consumption in a wide range of studies Aydinalp et al applied the neural network method to the 1993 Survey of Household Energy Use dataset to develop a model which accurately predicts the residential energy consumption of appliances, lighting and space cooling (Aydinalp, et al., 2002). The resulting ANN displayed a strong prediction performance, $R^2 = 0.90$, which performed well even with residences with abnormally high or low energy usage. The study highlighted the robustness of the ANNs.

A 2004 study by the same researchers used NN models on the same dataset to predict energy consumption due to space heating and domestic hot water (Aydinalp, et al., 2004). The ANNs resulted in R^2 values of 0.91 and 0.87 indicating the high confidence.

3 Methodology

The data was collected on a web-based platform, cleaned and normalized using Excel (Microsoft, 2018) and analyzed with the open source software (The R Foundation, 2018) and R Studio (RStudio, 2018). The following methodology in Figure 3-1 was performed:

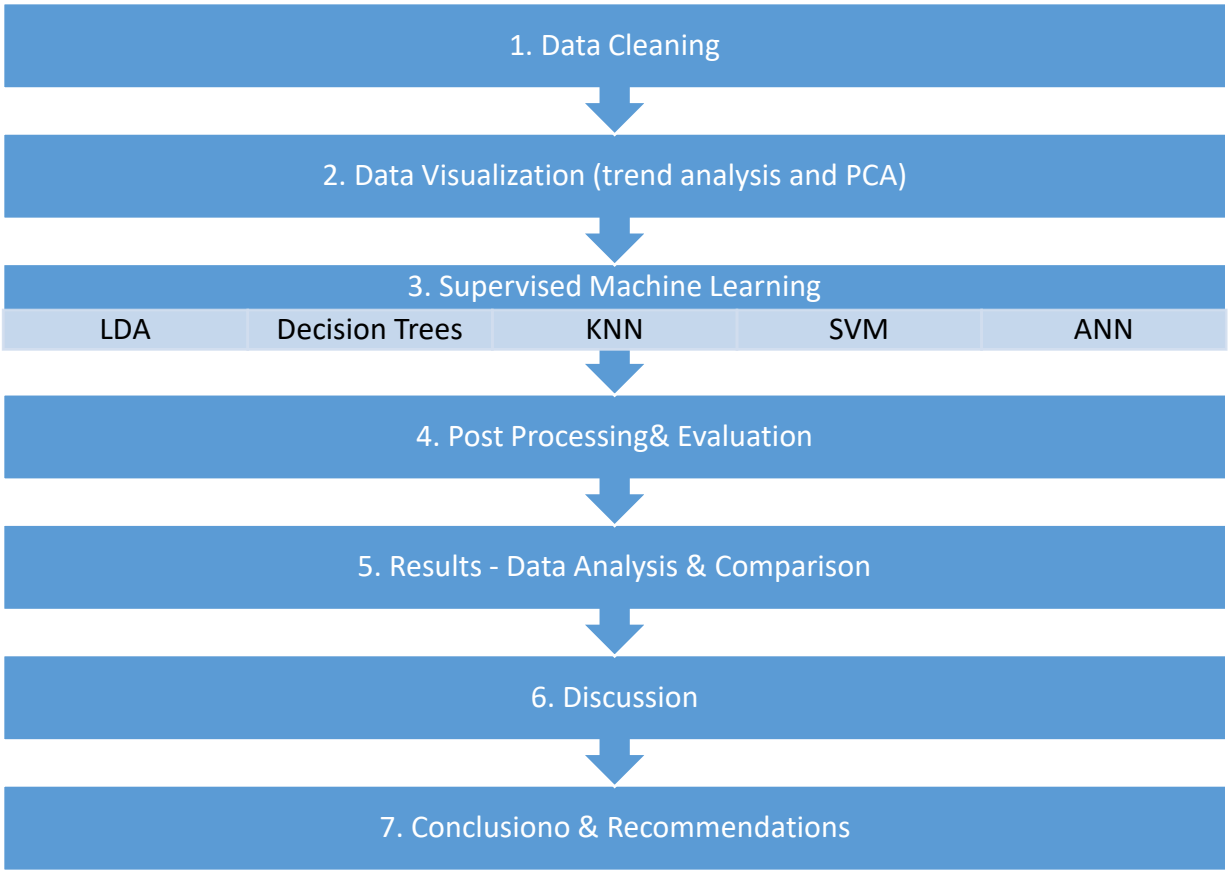


Figure 3-1 Report Methodology

3.1 Data Cleaning

The raw data was extracted from the REALPAC 20 by '15 site and was investigated for inconsistencies and outliers. These typically include user input errors, missing energy meters and building area updates, which lead to inconsistency across the years. Property Managers were contacted to provide confirmations and revisions. Building years (building data entered for a particular year) were excluded from the survey if errors remained unresolved.

The datasets used in this paper are not identical in size to the datasets in the published REALPAC 20 by '15 reports. The REALPAC reports removed more individuals from the database during cleaning. This research used a slightly less strict criteria in the effort to retain as many individuals as possible and to help train more robust models.

Table 3-1 lists the amount of eligible buildings which were retained after cleaning by province and year.

Table 3-1 Amount of Eligible participating buildings by year and province

Province	2010	2011	2012	2013	2014	2015
AB	65	72	77	94	98	97
BC	50	50	58	72	69	76
MB	7	6	7	7	7	5
NS	1	3	6	7	6	3
ON	236	225	217	288	287	259
QC	11	9	13	18	16	14
SK	2	2	1	11	16	16
Total	**372**	**367**	**379**	**497**	**499**	**470**

3.2 Data Visualization

The analysis of the data first began with the exploration of overall trends through the use of data visualization. The aim was to identify key variable relationships to integrate into predictive models.

Each variable in the 2010|2015 dataset was plotted against another using linear regression. This was carried out using the **plot**[2] (The R Foundation, 2018) function in [r]'s base package. The data points were coloured red, black and green according to the class of their 2010 EUI to highlight possible '*Poor'*, '*Fair'* and '*Good'* clusters.

The class and provinces of each building across each survey year were tabulated, along with their proportions. Excel was used to compare the effect of Exterior Area, Asset Manager and Closest Major City on the EUI or Qualitative EUI.

As part of data visualization, the unsupervised learning method, PCA, was explored first to identify any meaningful clusters or associations within the dataset without the use of user-defined (and potentially arbitrary) classes.

3.2.1 Principal Component Analysis

A principal component analysis (PCA) was employed to analyze the potential for feature reduction on the multivariate and multiscale dataset. PCA leads to weak results when applied to

[2] This document will visually distinguish function names with **bold** font.

data with low variance, however; its application on wide varying datasets leads to results that are sensitive to outliers (Akinduko & Gorban, 2013). Therefore, it may be difficult to reveal accurate and valuable underlying structures (Akinduko & Gorban, 2013). For this reason, two outliers were discarded prior to performing the algorithm. No other pre-processing was performed

The *FactoShiny*[3] package (Vaissie, et al., 2016) was used in Rstudio to open the **PCAshiny** application (Vaissie, et al., 2018), which allowed for the iterated exclusion of variables and the export of results and graphs. To ensure all variables were independent, a new iteration was performed for each combination of the following variables, illustrated in Figure 3-2

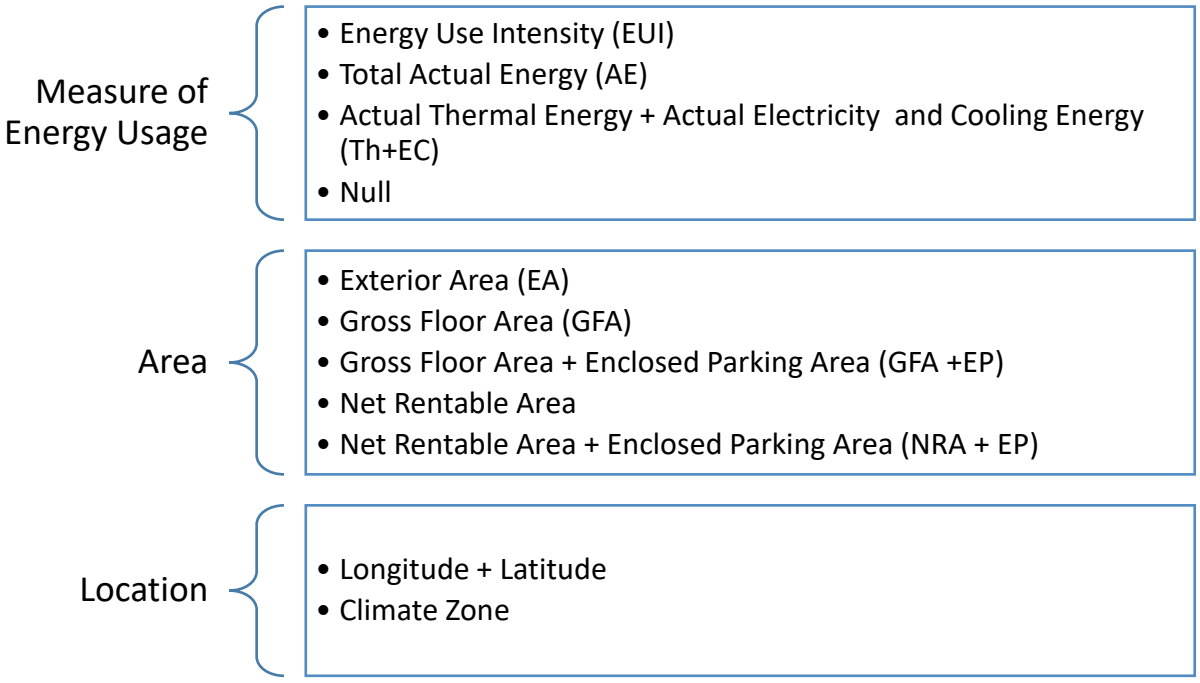


Figure 3-2 Interdependent variables for Energy Usage, Area and Location

The energy data for 2015 was considered the most robust as it possessed the most observations compared to previous years. First, the PCA was performed on 44 combinations of variables for the 2015 energy data and the eigenvalues, cumulative variance, the first ten Scores and first ten loadings for the 5 largest PCAs were recorded for each iteration. After analysis of the results indicated the alternating combinations resulted in limited variability, subsequent iterations for other years only interchanged the outcome variables. Only Exterior Area and Longitude +

[3] This document will visually distinguish package names with *italics* font.

Latitude were used for the Area and Location variables; this resulted in 44 iterations for each year. Due to the analysis of the results, only the 2015 Dataset was used.

One drawback of PCA shiny is the inability to easily access and record the full set of loadings and Scores. It is best used to quickly perform a series of iterations, review the variance of each principal component and produce graphics. To address this shortcoming, the **prcomp** function was used in R Studio to perform a principal component analysis on datasets where more in-depth results were desired. This function is also compatible with **predict** which estimates the principal components of new data given a trained model.

The **prcomp** function was applied on two versions of the 2010-2015 data set. The principal components PC1 and PC2 were plotted against each other and the loadings were classified according to *'Fair', 'Poor', 'Good'*.

*Table 3-2 2010-2015 dataset variables used in iteration 1 and 2 of PCA using the **prcomp** function*

Iteration	*2010 Actual Electricity and Cooling (E&C)*	*2010 Actual Thermal*	*2010 Actual Total*	*Exterior Area*	*Gross Floor Area*	*Enclosed Parking*	*Latitude*	*Longitude*	*Construction Year*	*Number of Structures*	*Building Class*	*Climate Zone*	*Electrical Heat*	*Cooling Tower*	*Soft landscaping*	*Occupant Density*	*Vacancy Rate*	*Weekly Operating Hours*
1	X	X			X	X	X	X	X	X	X		X	X	X	X	X	X
2			X	X					X	X	X	X	X	X	X	X	X	X

A threshold of 90% was set for the desired cumulative variability provided by the PCA dimensions. This means the sums of the variation explained by each PCA dimension were added together until a minimum of 90% was achieved. Only dimensions prior to the threshold were considered significant and any variable that without a loading more than 0.7 or less than 0.7

corresponding to any retained PCs any of those dimensions would be flagged for removal in subsequent algorithms. If the loadings threshold was not reached, the PCA results were considered unsuccessful for the dimensionality reduction.

The Scores plots were analyzed for clusters, either discrete or overlapping, within the 2010 Qualitative EUI, 2015 Qualitative EUI or Qualitative EUI Change graphs. The plots were also analyzed to interpret which independent variables lead to better results. These interpretations were used to reduce the amount of iterations needed in future algorithms.

3.3 Supervised Machine Learning

After identifying key features using PCA, supervised learning methods were used to identify patterns and predict an outcome.

3.3.1 Linear Discriminant Analysis

The Linear Discriminant Analysis was performed on the 2010|2015 Dataset. During preprocessing, the outliers identified during the PCA (i.e. 189, 256, 289) were removed from the dataset. The data was not normalized before hand as this processing is later performed by the applied [r] package. Due to difficulties extracting Scores with other packages, the **LDA** function from the *momocs* package (Bonhomme & Claude, 2018) was chosen as it allowed a more in-depth look at the results (i.e. score extraction) and customization of graphs.

The LDA iterations varied the sets of input variables as well as the outcome variables as seen in Table 3-3 and Table 3-4 to improve the likelihood of discovering a successful combination. Sets A and B were used in the first six iterations and used the outputs of the two PCA **prcomp** iterations. Set C was developed from the best-performing variable combination of these initial iterations and consists of the same input and outcome variables as in iteration 1 but removed the area variables, which are not entirely independent of other input variables, to determine whether this would significantly improve accuracy of the results. Set D is comprised of all features in set C which accounted for more than 3% of the variance to linear discriminants.

Table 3-3 LDA Iteration Description

Iteration	Input Variables (Refer to Table 3-4)	Outcome Variable
1	Set A	2010 Qualitative EUI
2	Set A	2015 Qualitative EUI
3	Set A	Qualitative EUI Change
4	Set B	2010 Qualitative EUI
5	Set B	2015 Qualitative EUI
6	Set B	Qualitative EUI Change
7	Set C	2010 Qualitative EUI
8	Set D	2010 Qualitative EUI

Table 3-4 2010-2015 dataset variables used in iterations of LDA

Input Variable Set	2010 Actual E&C	2010 Actual Thermal	2010 Actual Total	Exterior Area	Gross Floor Area	Enclosed Parking	Latitude	Longitude	Asset Manager	Construction Year	Number of Structures	Building Class	Climate Zone	Electrical Heat	Cooling Tower	Soft landscaping	Occupant Density	Vacancy Rate	Weekly Operating Hours
A	X	X			X	X	X	X	X	X	X	X		X	X	X	X	X	X
B			X	X					X	X	X	X	X	X	X	X	X	X	X
C	X	X					X	X	X	X	X	X		X	X	X	X	X	X
D	X	X					X	X	X	X							X		X

The following information was recorded for each iteration:

1. Linear Discriminant loadings:

   Similar to PCA, each loading or eigenvalue was divided by the sum of all of the loadings for a particular linear discriminant. This calculated the strength of the loading and the significance of the corresponding variable to the linear discriminants. The strongest variables were identified as important features and the variables with weak loadings were considered unimportant.

2. Score plots:

The score plots were analyzed for class clusters. A score plot with a visibly tight group of clusters implies that classification is possible given the structure of the data.

3. Overall classification accuracy rates:

   The LDA model performed a *leave one out* cross-validation to train the model. The 'left' out data was used by the algorithm to test the classification accuracy. The **LDA** function determined the final classification accuracy as the average correct classification rate for all of the cross validated iterations.

4. Within-class accuracy rates and;
5. Misclassification rates

3.3.2 k-Nearest Neighbours

The k-Nearest Neighbours (KNN) algorithm was employed to analyze the potential for classification of new samples and to illustrate class separability between various combinations of independent variables. The **knn** and **knn.cv** functions (Ripley & Veneables, 2002; Ripley, 1996) from the *class* package (Ripley & Venables, 2015) were used. The **knn** function performs the KNN algorithm, using one test and training set to develop the model while the **knn.cv** function employs leave-one-out cross validation to create the model (Ripley & Venables, 2015).

The analysis was performed on the 2010-2015 dataset. Four iterations were performed (Table 3-5). K was determined by taking the square root of the number of observations in the training set.

Table 3-5 KNN iterations

	Independent Variables									Dependent Variables (Classifiers)			knn function used
Iteration	Actual E & C Energy	Actual Thermal Energy	Building Manager	Latitude	Longitude	Construction Year	Occupant Density	Vacancy Rate	Weekly Operating Hours	2010 Qualitative EUI	2015 Qualitative EUI	Qualitative EUI Change	
1 (Test/Train)	X	X	X	X	X	X	X	X	X	X			**knn**
2 (Test/Train)	X	X	X	X	X	X	X	X	X		X		**knn**
3 (Test/Train)	X	X	X	X	X	X	X	X	X			X	**knn**
4 (Cross Validation)	X	X	X	X	X	X	X	X	X	X			**knn.cv**

All iterations used the same independent variables which were selected based on the results of the LDA feature reduction results (See Section 4.2).

For iterations 1 to 3 the KNN model was trained on 70% data and tested using the remaining 30% of individuals. The only difference between the iterations was the outcome variable. The results were analyzed and the outcome variable used for the highest classification accuracy was chosen for iteration 4. The fourth iteration was performed using cross validation, in an attempt to improve the accuracy of the classification results.

For each iteration, the Confusion Matrix, Classification Accuracy, Confidence Interval, and P-value were recorded.

The KNN algorithm was then applied to create graphs for iterations 1 to 3 with the purpose of identifying variables combinations which illustrate class clusters. Within each iteration the following variables were mapped against each other.

Table 3-6 Variables in each KNN model iteration

Sub-iteration for iteration X	Actual E & C Energy	Actual Thermal Energy	Property Manager	Latitude	Longitude	Construction Year	Occupant Density	Vacancy Rate	Weekly Operating Hours
X.1	X	X							
X.2				X	X				
X.3			X			X			
X.4							X	X	
X.5	X								X
X.6	X		X						
X.7	X			X					
X.8		X		X					

The graphs use colour to indicate which class the KNN model would assume each point belongs to and point size to illustrate the KNN model's assumed probability of each point belonging to that particular class. Data points with a black border indicate an actual individual from the original dataset, not a predicted value from the model. The graph success was assessed based on a visual assessment of the class separability.

3.3.3 Multiple Linear Regression

The Multiple Linear Regression models were created in [r] with the **lm** function for linear models on the 2010-2015 dataset. To ensure comparability across variables, each variable was scaled between 0 and 1, using their existing maximum and minimum values. To save time, the model focused only on the independent variables identified as most significant by the PCA and LDA feature reduction results, with the exception of energy variables. The scope of this paper will only include the assessment of overall EUI and EUI Change across years; therefore, the model did not evaluate the models against the more specific Electricity and Thermal data. Please refer to Table 3-7 for a list of the variables included in each MLR iteration.

Table 3-7 Variables in each MLR model iteration

	Input Variables							Outcome Variables		
Iteration	Building Manager	Latitude	Longitude	Construction Year	Occupant Density	Vacancy Rate	Weekly Operating Hours	2010 EUI	2015|2010 EUI	Qualitative EUI Change
1 (Test/Train)	X	X	X	X	X	X	X	X		
2 (Test/Train)	X	X	X	X	X	X	X		X	
3 (Test/Train)	X	X	X	X	X	X	X			X

For all iterations The MLR model was trained on 70% data and tested using the remaining 30%. Each training and test set included the same proportion of classes as the original dataset in an attempt to ensure that smaller classes, such as *'Good'* and *'Poor'*, were significantly represented in both sets. The aim was to improve upon the low classification accuracy rates observed in previous methods.

The model used linear regression to evaluate the influence of each independent variable from the 2010 data on the 2010 EUI (current year), 2015 EUI (future year) or the change in EUI between 2010 and 2015. After the models were trained, they were applied to the test data to evaluate its performance when confronted with new observations. The predicted values from both the training and test set were denormalized. See Equation 2 and Equation 3 for details on how the data was denormalized using the max and min values from the pre-normalized dataset.

Equation 2 Normalization Formula

$$X_{normalized}^{actual} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$where\ X_{min} = minimum\ value\ of\ X\ within\ the\ vector\ containing\ X$$

$$where\ X_{max} = maximum\ value\ of\ X\ within\ the\ vector\ containing\ X$$

Equation 3 Denormalization Formula

$$Y_{denormalized}^{predicted} = Y_{normalized}^{predicted} * \ [X_{max} - X_{min}] + \ X_{min}$$

For each iteration the overall classification accuracy rate, within-class accuracy rate, predicted training set values, predicted test set values and P-values for each variable were recorded.

After the initial results were unsatisfactory, the algorithm was revised to predict the percent change in EUI between 2010 and 2015 in proportion to the starting EUI. This value was calculated (2010 EUI – 2015 EUI) / 2010 EUI as expressed as a percent change. The relevant classification metrics are listed in Table 3-10.

Also, for each test and train set of each iteration, the predicted and actual values were plotted against each other. The graphs were visually assessed based on the perceived correlation between the two variables.

3.3.4 Support Vector Machines

The Support Vector Machines (SVM) models were created in R Studio (R version 3.3), using the **tune** and **svm** functions from the *e1071* (Meyer, et al., 2018) package on the 2010-2015 dataset. To avoid bias in the inputs, each variable was scaled. To reduce time costs, the models focused only on the independent variables identified as most significant by the PCA and LDA feature reduction results. The energy variables 2010 EUI, 2015|2010 EUI and 2010 Qualitative EUI were used as outputs; the model did not evaluate the more specific Electricity and Thermal data. Table 3-8 summarizes the variables included in each SVM iteration.

Table 3-8 SVM iterations

	Independent Variables							Output Variables			
Iteration	Building Manager	Latitude	Longitude	Construction Year	Occupant Density	Vacancy Rate	Weekly Operating Hours	2010 EUI	2015|2010 EUI	Qualitative 2010 EUI	Qualitative 2015 EUI
1 (Test/Train)	X	X	X	X	X	X	X	X			
2 (Test/Train)	X	X	X	X	X	X	X		X		
3 (Test/Train)	X	X	X	X	X	X	X			X	
4 (Test/Train)	X	X	X	X	X	X	X				X

For all iterations, the SVM model was trained on 70% data and tested using the remaining 30% of observations. As with SVM iterations, each training and test set included the same proportion of classes as the original dataset in an attempt to ensure that classes with a smaller population, such as *'Good'* and *'Poor'*, were significantly represented in both sets. Iterations 1-3 were run with each of the following kernels: linear, polynomial, sigmoid and radial. Iteration 4 was trained using the best performing kernel in iteration 3 to save time. The range of kernel selection allowed the testing of different data transformations to improve the class separability and subsequent prediction accuracy. Two different types of models were also tested: the first two iterations were run using a regression SVM model and the second two iteration were trained using a classification model to investigate which model would return the best predictive accuracy.

After the models were trained, they were tested to evaluate performance when confronted with new observations. The predicted and actual values for both the train and test datasets were compared. The Root Mean Square Error (RMSE), overall classification accuracy rate, within-class accuracy rate, predicted training set values, predicted test set values were also recorded (See Appendix II).

3.3.5 Decision Trees

The decision tree models were created in R Studio (R version 3.3). The random forest decision tree algorithm was specifically used as Breimann's research indicates that it offers improved

classification accuracy rates when compared to individual decision trees (1994). The models were created using the **ranger** function from the *ranger* package (Wright, 2017) on each dataset. **Ranger** was selected as it quickly conducts recursive partitioning, is appropriate for high dimensional datasets, and supports classification and regression forests. The training and test datasets were then created through random sampling with one third of the observations placed in the test set. The independent variables were the same for each iteration for the random forest decision tree models: Latitude, Longitude, Construction Year, No. of Structures, Building Class, Climate Zone, Electrically Heated, Cooling Tower, Soft Landscaping, Occupant Density, Vacancy Rate and Weekly Operating Hours.

For regression trees, the minimum node size was set to 1, the number of trees was left at the default value of 500 and the variable importance mode was set to impurity. Classification was set to FALSE. The random forest algorithm used the training data to grow numerous random forests; the majority class vote or the average numeric vote was selected as the model output. After each model was created, the random forest model and the test data were applied as arguments in the predict function. A confusion matrix was created comparing the predicted outcome variable and the actual outcome variable from the test dataset. To compare the results of the four different outcome variables (EUI, Actual E&C Energy, Actual Thermal Energy or Actual Total Energy) each output was normalized by dividing the residual by the actual amount. In total, twenty-four iterations of the regression trees were run. The accuracy of each model was then calculated and graphed.

For classification prediction, separate vectors with class labels were not required. Two vectors containing only the Qualitative EUI labels were created for the test and training set. The formula and training dataset were put into the **ranger** function. As with the regression iterations, the minimum node size was set to 1. The number of trees was left at the default value of 500. Classification was set to TRUE as it is required when the data set is in the form of a data matrix. The variable importance mode was set to impurity; this selects the Gini index for classification trees and measures the variance of the responses for regression trees (Wright, 2017). In total, six iterations of the classification trees were run. As with the regression trees, the accuracy of each classification model was calculated and graphed.

3.3.6 Artificial Neural Network

The Artificial Neural Network algorithms were run with the *neuralnet* package (Fritsch, et al., 2016), which trains neural networks through resilient backpropagation with weight backtracking. In order to ensure comparability, the independent variables were scaled between 0 and 1, using the existing maximum and minimum values for each respective variable. For predicting EUI Change in 5 years, the dependant variable was also scaled between 0 and 1; however, a maximum of 100 and a minimum of -100 were used. For predicting current EUI, a maximum of 50 and a minimum of -50 were used.

The neural net was tuned using 1-12 nodes in each hidden layer of the model. The number of hidden layers varied from one to two. For each model, convergence depends on a set threshold of 0.01. The seeds were set to encourage reproducibility in the results and the stepmax was set to 10,000 to prevent over-training. Five (5) repetitions of the **ANN** function were run for each set of parameters to explore the various starting weights, improving the likelihood of finding a model that converges and reducing the chance of an overfit model. The resulting accuracy was recorded as an average of all successful repetitions. If a particular dataset did not converge in all five (5) repetitions, the accuracy was recorded as 'N/A'.

The tuning model was run using 10-fold cross-validation. The folds were originally split automatically by [r] (The R Foundation, 2018). This led to folds that did not contain all three (3) classes. To address this concern, the data set was manually split into 10 different sets, with each set containing a relatively evenly distributed amount of Poor, Fair and Good classes. For each set of parameters, individual fold test classification accuracy as well as the corresponding the maximum, minimum and average were summarized. The calculation of classification accuracy was originally attempted setting *linear.output = false* (classification). As the results showed very little divergence, the classification methodology was revised and a regression approach (implemented by setting *linear.output = True*) was used. A function was then written to de-normalize the output, which was then classified according to a metric appropriate to the dependant variable being predicted. If the outcome variable and the actual variable class matched, the result was considered correct, no matter the numeric distance between the two values. A function was created to record the actual and predicted dependant variable values as well as their residuals.

The algorithm was first employed to predict 2015 EUI using 2010 data. After the results were unsatisfactory, the algorithm was revised to predict the percent change in EUI between 2010 and 2015 in proportion to the starting EUI. This value was calculated as (2010 EUI – 2015 EUI) / 2010 EUI and is expressed as a percent change. To further improve results, the algorithm was also trained to predict the actual change in EUI between 2010 and 2015, expressed in the same units as EUI (eKWh/sf/yr). Actual change was calculated as 2010 EUI minus 2015 EUI.

This approach was employed for seven sets of input variables, to explore if the presence or absence of certain variables affected the classification accuracy of the test data. The input variables combinations for each set are seen in Table 3-9. Iterations combined different variable sets with one, two or three hidden layers to predict either actual EUI change or percent EUI change between 2010 and 2015. Seven iterations of each one- and two-layer ANN models were completed for both outcome variables. Due to non-converging models, only five iterations were completed for the three-layer ANN model. They the actual EUI change outcome variable and input variable sets 3 to 7.

Table 3-9 Variables engaged with each ANN iteration

Input Variable Set	2010 EUI	Exterior Area	Property Manager	Latitude	Longitude	Construction Year	Num. of Structures	Class	Climate Zone	Electrically Heated	Cooling Tower	Soft Landscaping	Occupant Density	Vacancy Rate	Weekly Operating Hours
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X			X	X	X
3	X	X	X		X	X	X	X	X	X			X	X	X
4	X	X	X		X	X		X	X	X			X	X	X
5	X	X	X		X	X		X	X				X	X	X
6	X	X	X		X	X		X	X					X	X
7	X	X	X		X										

The results were analyzed and the best combination of hidden layer nodes were recorded for each iteration. The classification accuracy of each neural net iteration was plotted for simplified

comparison. A visualization of the neural net structure with the highest overall accuracy was output along with the corresponding weights.

3.4 Evaluation

To evaluate the accuracy of predictive models, the datasets were split into training and test sets. The models were trained on the training dataset, then applied to the test dataset and the predicted responses were compared to the actual values. The success of a model was evaluated according to overall classification and within-class accuracy rate and RMSE. Overall classification accuracy rate is the proportion of correctly identified classes (*'Good'*, *'Fair'* or *'Poor'*)[4] out of all the responses.

Equation 4 Overall classification accuracy rate

$$Overall\ classification\ accuracy\ rate = \frac{Amount\ of\ Correctly\ Classified\ Observations}{Total\ Amount\ of\ Observations} \times 100\%$$

The within-class accuracy rate is the portion of correctly identified instances of each class out of instances of that class.

Equation 5 Within-class accuracy rate

$$Within\ class\ accuracy\ rate = \frac{Amount\ of\ Correctly\ Classified\ Observations\ in\ Particular\ Class}{Total\ Amount\ of\ Observations\ \ in\ Particular\ Class} \times 100\%$$

The overall classification and within-class accuracy rate are primarily used for classification algorithms, but were also used to evaluate the results of regression algorithms in an effort to explore possible improvements to the usefulness of the results. The responses of a regression model were classified during post-processing based on the following conditions in Table 3-10 and evaluated as if they were classification model results using Equation 4 and Equation 5.

Table 3-10 Classification Boundaries for various Dependent Variables

	Outcome Variable		
Class	EUI	EUI Actual Change	EUI Percent Change

[4] This document will visually class names with a combination of 'quotation marks' and *italic* font

	(ekWh/sf/yr)	(ekWh/sf/yr)	
Poor	X => 40	X <= 0	X <= 0%
Fair	20 < X < 40	0 < X < 10	0% < X < 25%
Good	X <= 20	X => 10	X => 25%

The RMSE was used to solely evaluate results from regression models. RMSE was chosen against other regression metrics as it punishes larger residuals more severely. As the RMSE were calculated for different outcome variables, different RMSE thresholds exist for different outcome variables (Table 3-11).

Table 3-11 RMSE Threshold for a successful and ideal RMSE evaluation

	Outcome Variable	**Threshold for Successful RMSE**	**Threshold for Ideal RMSE**
De-Normalized Outcome	EUI	5 (ekWh/sf/yr)	2 (ekWh/sf/yr)
Normalized Outcomes	EUI	.09	.04
	Actual Total	.08	.03
	Actual E&C	.05	.02
	Actual Thermal	.03	.01

The threshold for successful RMSE was determined to be as 5 ekWh/sf/yr for denormalized EUI as this value was believed to be useful for building managers, and will still successfully identify very well performing buildings and flag very poor performing buildings. The threshold for an ideal RMSE was set as 2 ekWh/sf/yr to expand the potential actionability of the results. The thresholds for the normalized EUI and Actual Total was determined by normalizing the previous thresholds using maximum and minimum values from the 2010 dataset. To determine the thresholds for the normalized Actual Total, Actual E&C and Actual Thermal Energy Use, the 20 by 15 model building was reviewed. Approximately 63.5% of energy within this building is consumed through electricity use; the remaining 36.5% is attributed to thermal energy. Therefore, the Actual E&C and Actual Thermal threshold were determined by multiplying the Actual Total energy use by their respective proportions.

4 Results

The following section displays a graphical visualization of the key datasets and explains the results gained from each algorithm. The expected outcomes for each algorithm are explained and tables and graphs are provided to display and compare the results. An analysis of the findings is provided to evaluate the model success and interpret observed data trends.

4.1 Data Summary

It was anticipated that data visualization was that linear relationships between input and outcome variables would identify important features. This was tested by plotting each variable against one another. No strong relationships were observed between variables that were not definitively related to one another (e.g. GFA and EA). The graph below, Figure 4-1, illustrates a sample of the resulting plots with the '*Poor'*, *'Fair'* and *'Good'* classes respectively represented by red, black and green datapoints. The plots are coloured to better highlight class clusters when two variables were compared with one another. Clusters are only observed in plots with EUI as that was the metric by which the classes were defined.

Figure 4-2 highlights the quantity of buildings in the data set for each year by province. The scale is logarithmic so that less represented buildings are visible. Most buildings were located in Ontario and only 1 - 17 buildings were located in Manitoba, Nova Scotia and Saskatchewan. The amount of buildings participating in each province did fluctuate year-to-year, however the quantity stayed within a comparable range.

As can be observed in Figure 4-3 to 4-5, the buildings were primarily located near major cities. The graphs only display locations of buildings from the 2015 dataset however, the datasets from other years display similar distributions. The classes of building EUI, *'Poor'*, *'Fair'* and *'Good'*, are relatively well distributed throughout all of the location clusters. The darker colours represent a denser population of building of that class in each region.

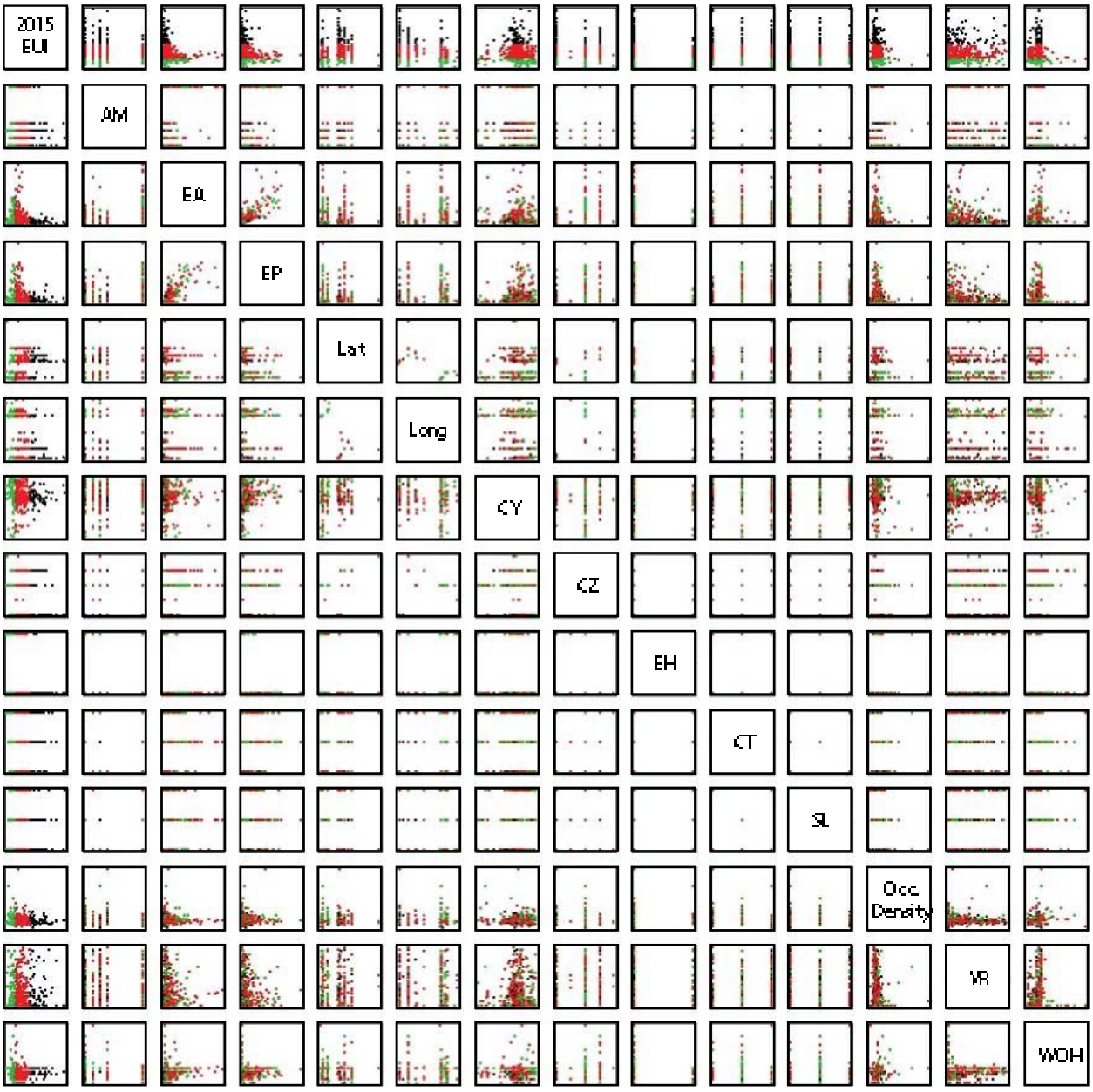


Figure 4-1 2015 dataset plot with 2015 Energy Use Intensity (EUI), Asset Manager (AM), Exterior Area (EA), Enclosed Parking (EP), Latitude (Lat), Longitude (Long), Construction Year (CY), Climate Zone (CZ(, Electrical Heat (EH), Cooling Tower (CT), Soft landscaping (SL), Occupant Density (Occ. Density), Vacancy Rate (VR) and Weekly Operating Hours (WHO) plotted against one another. All datapoints are classified according to 2015 Qualitative EUI.

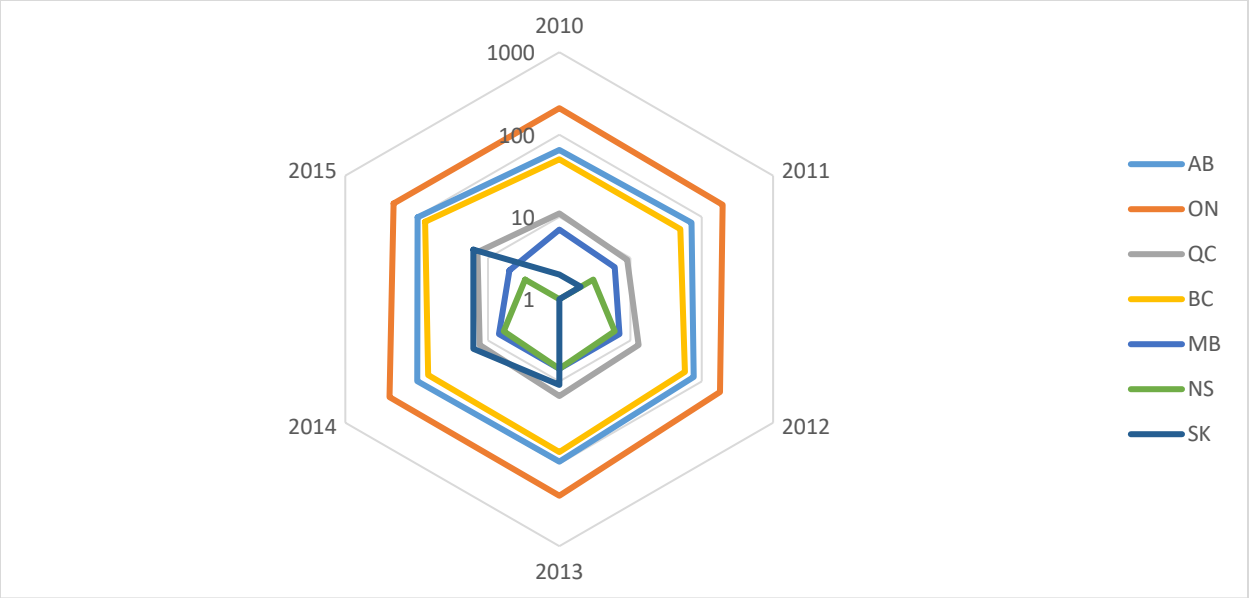


Figure 4-2 Logarithmic representation of the quantity of buildings in the 20 by '15 dataset by year and province

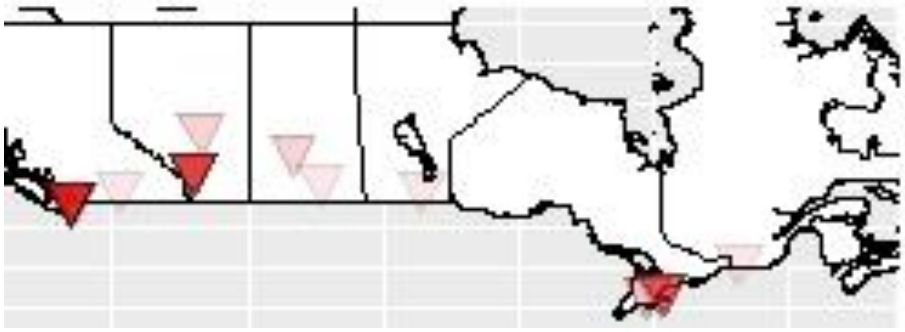


Figure 4-3 Geographic distribution of 'Poor' buildings across Canada using the 2015 dataset

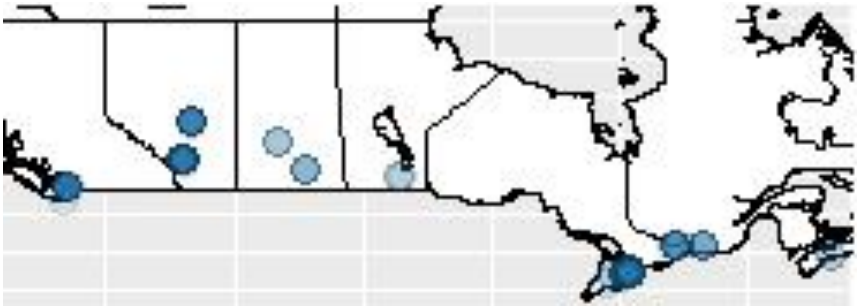


Figure 4-4 Geographic distribution of 'Fair' buildings across Canada using the 2015 dataset

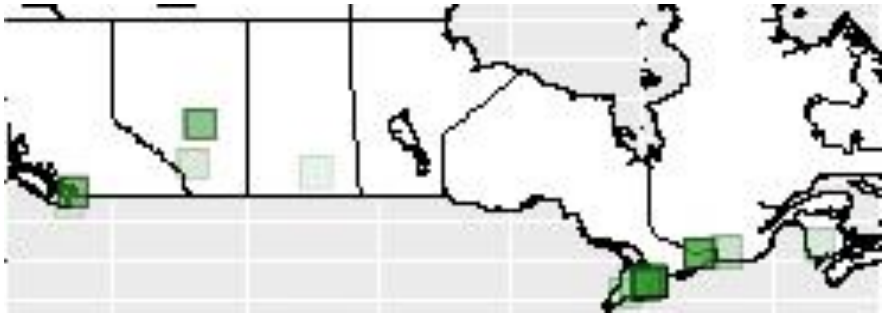


Figure 4-5 Geographic distribution of 'Good' buildings across Canada using the 2015 dataset

To test the second expectation, that geographical trends would be observed in the data, the average EUI for each Closest Major City (>3 buildings each year) was graphed for the years from 2010 to 2015. The buildings in Vancouver, BC performed the worst. This is assumed to be because Vancouver would rarely view its energy consumption normalized against the weather and climate of another city. Its actual energy usage may appear comparatively low due to its warmer climate leading to a perception of better performance and less focus on energy efficiency. From years 2010 to 2012, Ottawa was the best performing city, however between 2013 and 2015 the Toronto buildings improved and became the best performers.

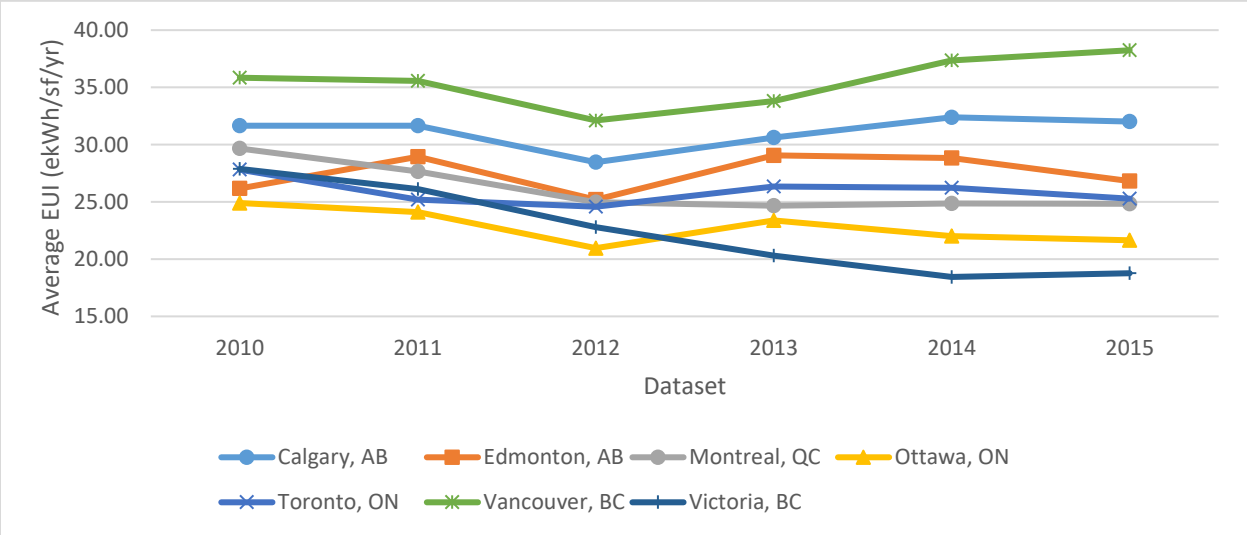


Figure 4-6 Average EUI for each Closest Major City for 2010 to 2015 Datasets

Table 4-1 describes the number of buildings in each province by year and class. It can be seen that most buildings are located in Alberta, British Columbia and Ontario.

Table 4-1 Number of individual buildings in each province by dataset and class with the class proportions for a given year and province contained in brackets.

Prov.	Class	Dataset					
		2010	2011	2012	2013	2014	2015
AB	Poor	7 (11%)	8 (11%)	7 (9%)	15 (16%)	20 (20%)	19 (20%)
	Fair	53 (82%)	62 (86%)	61 (79%)	71 (76%)	73 (74%)	68 (70%)
	Good	5 (8%)	2 (3%)	9 (12%)	8 (9%)	5 (5%)	10 (10%)
ON	Poor	17 (7%)	6 (3%)	6 (3%)	22 (8%)	22 (8%)	16 (6%)
	Fair	187 (79%)	178 (79%)	152 (70%)	191 (66%)	196 (68%)	164 (63%)
	Good	32 (14%)	41 (18%)	59 (27%)	75 (26%)	69 (24%)	79 (31%)
QC	Poor	1 (9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (7%)
	Fair	10 (91%)	9 (100%)	11 (85%)	15 (83%)	14 (88%)	9 (64%)
	Good	0 (0%)	0 (0%)	2 (15%)	3 (17%)	2 (13%)	4 (29%)
BC	Poor	19 (38%)	16 (32%)	13 (22%)	17 (24%)	17 (25%)	30 (39%)
	Fair	30 (60%)	34 (68%)	39 (67%)	43 (60%)	45 (65%)	36 (47%)
	Good	1 (2%)	0 (0%)	6 (10%)	12 (17%)	7 (10%)	10 (13%)
MB	Poor	1 (14%)	0 (0%)	2 (29%)	1 (14%)	1 (14%)	1 (20%)
	Fair	5 (71%)	3 (50%)	5 (71%)	5 (71%)	6 (86%)	4 (80%)
	Good	1 (14%)	3 (50%)	0 (0%)	1 (14%)	0 (0%)	0 (0%)
NS	Poor	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Fair	1 (100%)	1 (33%)	3 (50%)	4 (57%)	5 (83%)	3 (100%)
	Good	0 (0%)	2 (67%)	3 (50%)	3 (43%)	1 (17%)	0 (0%)
NT	Poor	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Fair	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)
	Good	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)
SK	Poor	2 (100%)	0 (0%)	0 (0%)	1 (11%)	6 (38%)	3 (19%)
	Fair	0 (0%)	2 (100%)	0 (0%)	7 (78%)	10 (62%)	12 (75%)
	Good	0 (0%)	0 (0%)	1 (100%)	1 (11%)	0 (0%)	1 (6%)
Annual Average	Poor	13%	8%	7%	11%	13%	15%
	Fair	77%	79%	72%	68%	70%	63%
	Good	10%	13%	21%	21%	17%	22%

The Average Exterior Area for each Current Year Qualitative EUI Class was graphed for the years between and including 2010 and 2015. Figure 4-7 Average Exterior Area for each Current Year Qualitative EUI Class for 2010 to 2015 datasets

The Average EUI for each Asset Manager was graphed for the years between and including 2010 and 2015. The largest Asset Managers were anonymized using numbers 1 through 4. The number 9 represents all the other buildings whose property management companies did not represent a substantial portion of the dataset. Figure 4-8 shows that those buildings with smaller asset

managers were likely to perform better than the others. Asset Managers 3 and 4 performed the worst.

Figure 4-7 shows that '*Poor*' buildings are more likely to be smaller in size; in particular less than 300,000 sf. This may indicate that the area normalization skews the results for smaller buildings as it is more difficult to reduce at this size.

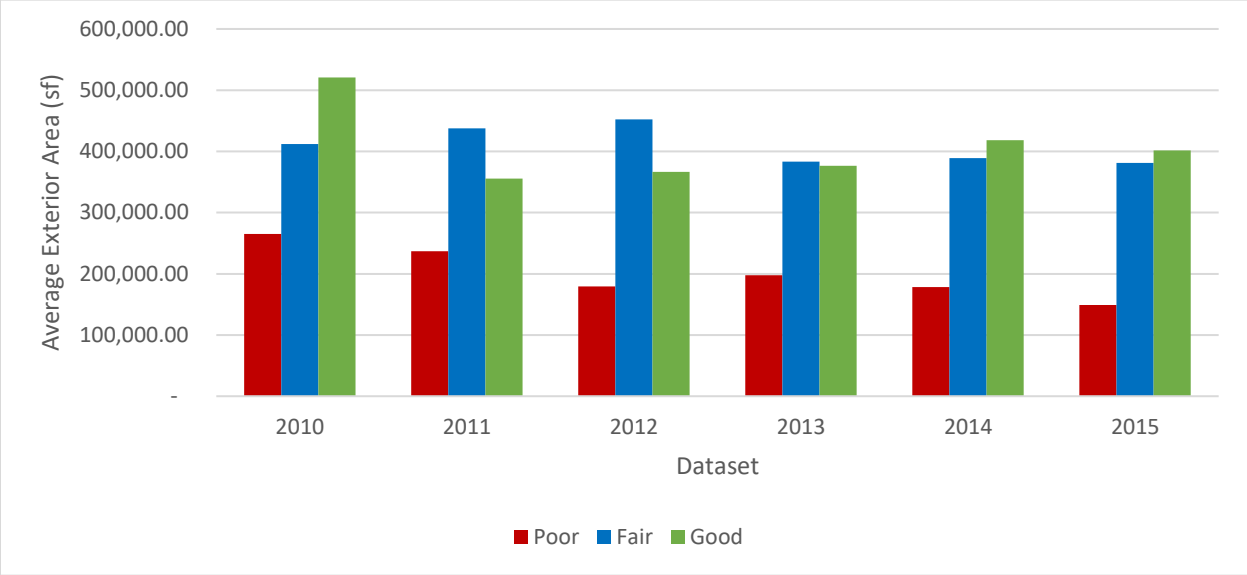


Figure 4-7 Average Exterior Area for each Current Year Qualitative EUI Class for 2010 to 2015 datasets

The Average EUI for each Asset Manager was graphed for the years between and including 2010 and 2015. The largest Asset Managers were anonymized using numbers 1 through 4. The number 9 represents all the other buildings whose property management companies did not represent a substantial portion of the dataset. Figure 4-8 shows that those buildings with smaller asset managers were likely to perform better than the others. Asset Managers 3 and 4 performed the worst.

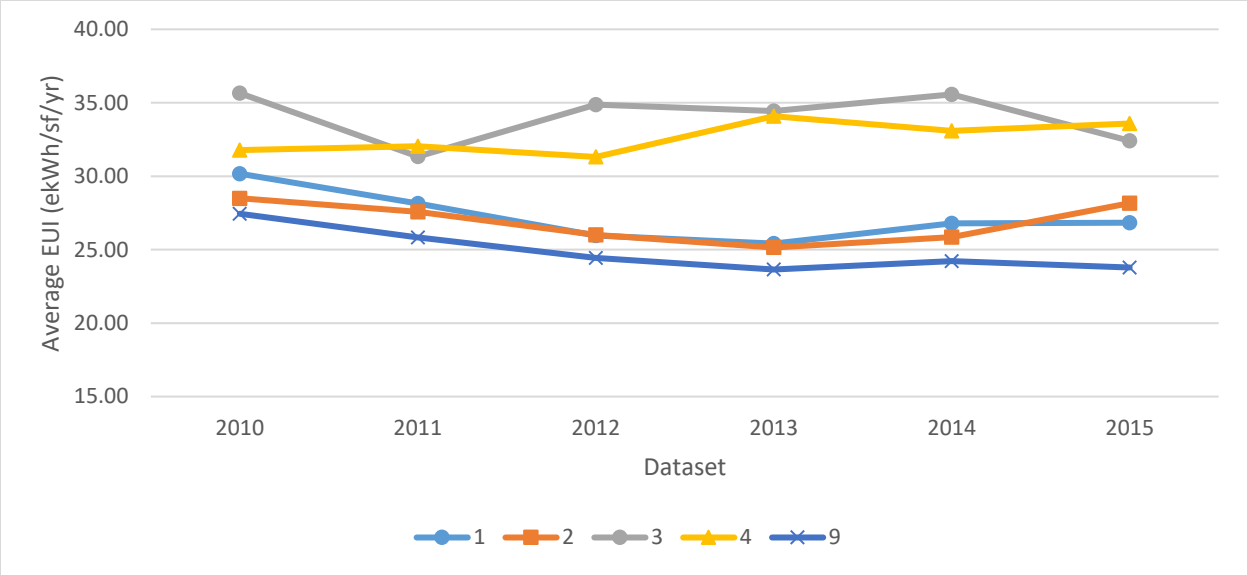


Figure 4-8 Average EUI for each Asset Manager for 2010 to 2015 Datasets

The data visualization was less successful than expected. No linear relationships were highlighted by plotting all of the variables against one another. It was determined that Nova Scotia, the Northwest Territories and Manitoba are under represented in the dataset and cannot provide useful insights on the commercial office energy usage in these provinces.

It was observed that smaller buildings are more likely to be classified as '*Poor'*, as are buildings located in British Columbia. Toronto buildings on average have the least amount of energy consumption and show the most amount of improvement over time than the other Closest Major Cities.

4.2 Principal Component Analysis

Principal Component Analysis was performed for dimensionality reduction and, following the methodology presented by Joliffe (1986), feature extraction. For dimensionality reduction, principal components (PCs) were determined with the intent to reduce the number of dimensions in the dataset, compressing it by minimizes the number of PCs required to account for at least 90% of the variation. Several subsets of features (44 iterations) were considered in each dataset, and the results are presented in Figure 4-9 and Figure 4-10, which summarize the iterations performed with **PCAshiny** on the 2015 Dataset. This was not particularly successful; a minimum of 10 principal components were required for all combinations of features considered. For all iterations involving EUI or Total Actual Energy, eleven principal components were consistently

required to reach the 90% threshold. For iterations that disregarded energy usage variables, ten dimensions were required. This is understandable as removing one entire feature reduced the number of principal components. The iterations which used a combination of Actual Thermal Energy and Actual E&C Energy always required 12 principal components, as they had one more dimension than the iterations using Actual Energy or EUI.

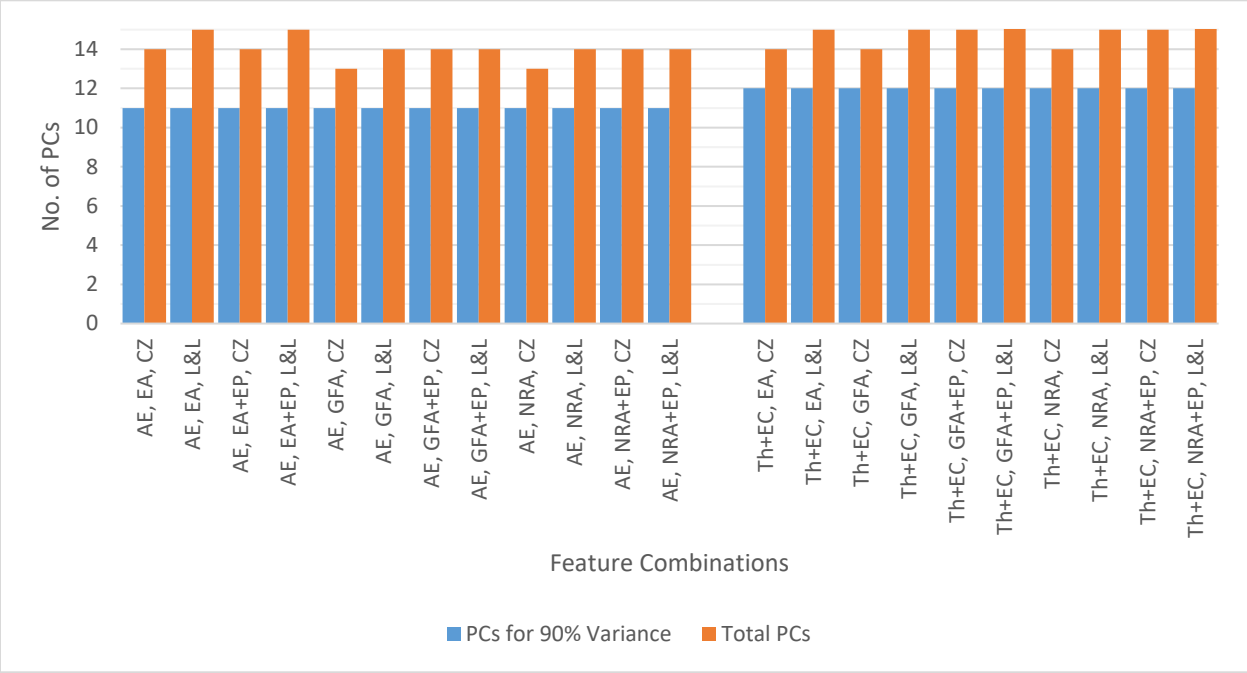


Figure 4-9 Number of PC required for 90% Cumulative Variance for the 2015 Dataset iterations using Total Actual Energy and a combination of Actual Thermal Energy + Actual Electricity and Cooling Energy

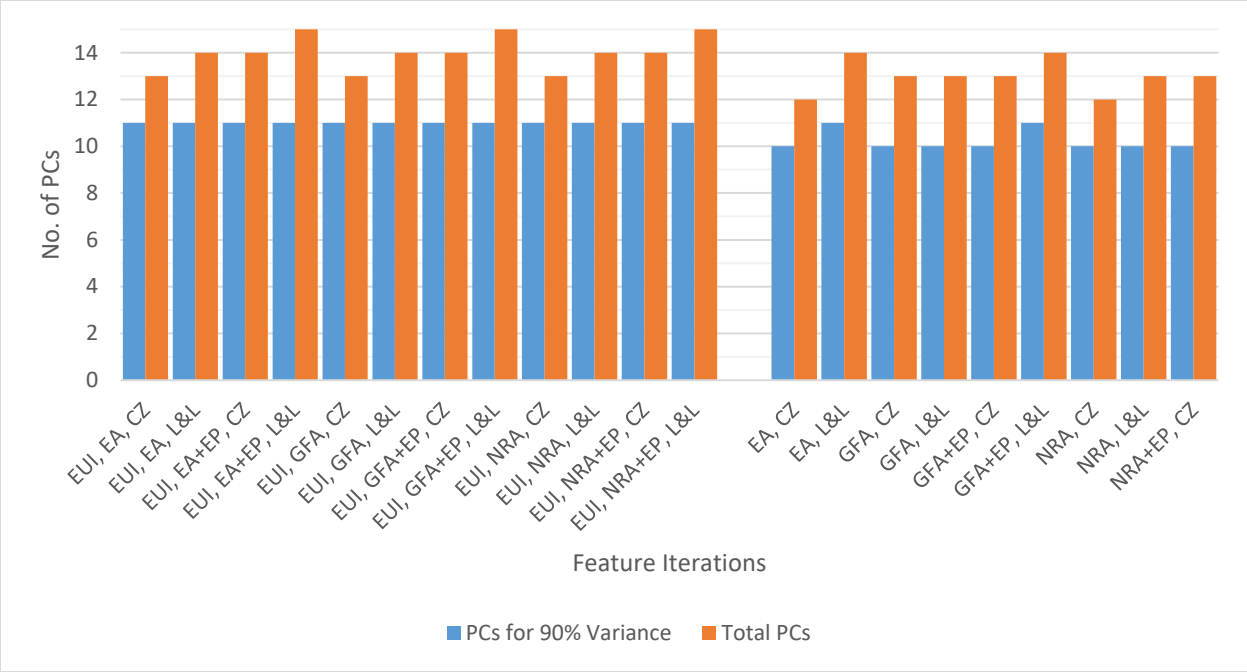


Figure 4-10 Number of Principal Components required for 90% Cumulative Variance for the 2015 Dataset iterations using either EUI only or no energy variables.

It was anticipated that alternating the energy use, area and location variables would affect the PCA results. However, it was observed that the cumulative variability and feature contribution changed little with each iteration. In fact, no change was observed between iterations with the same measure of energy. Only iterations without energy use features exhibited a slight fluctuation; between 10 and 11 principal components are required to reach the threshold. It can be assumed the change was due to the lower number of dimensions in those iterations.

In addition, it was anticipated that clusters could be found within the data; it was hoped the clusters would correspond to different Qualitative EUIs such as *'Good'*, *'Fair'* or *'Poor'*.

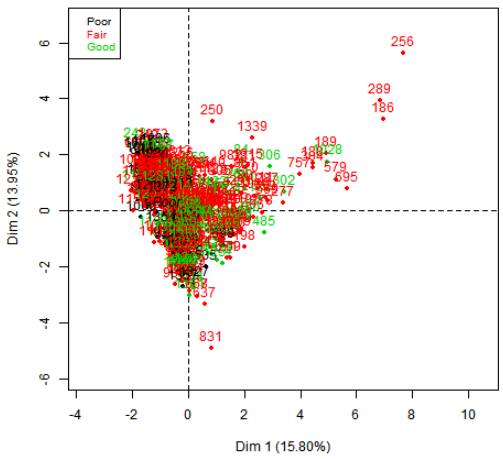


Figure 4-11 Scores Plot, 2015 dataset, EA, CZ

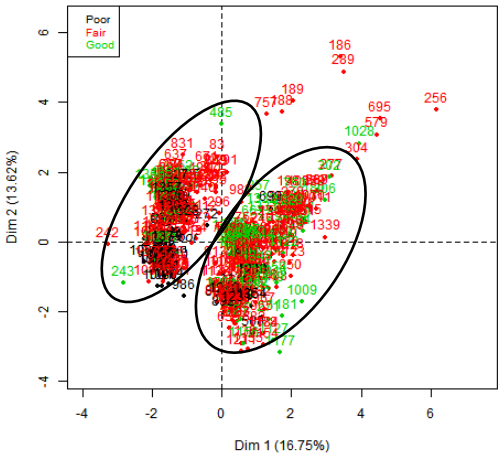


Figure 4-12 Scores Plot, 2015 dataset, EA, L&L

None of the score plots using the Climate Zone feature, as seen with Figure 4-11, illustrated clusters. However, iterations involving Longitude and Latitude, as seen in Figure 4-12, consistently showed two clusters. Unfortunately, these clusters were not separated according to Qualitative EUI. This indicates the clusters cannot reveal anything useful regarding a building's energy performance. It does however inform us that using iterations with Latitude and Longitude is more useful than iterations with Climate Zone. This is understandable as Latitude and Longitude provide more specificity than Climate Zone.

As the cumulative variability graphs indicated the PCA results could provide little opportunity for dimensionality reduction, PCA was not performed on the 2010-2014 Datasets, which are considered less robust (refer to methodology). PCA shiny was used to explore the relationships between feature and principle components by producing exploratory loading maps. The magnitude of a feature's contribution is represented by the length of the line and its proximity to the PC axis. If a variable is closer to PC 1 than PC 2, it indicates that it contributes to both PCs, just more to PC 2. Refer to Figure 4-13 and Figure 4-14.

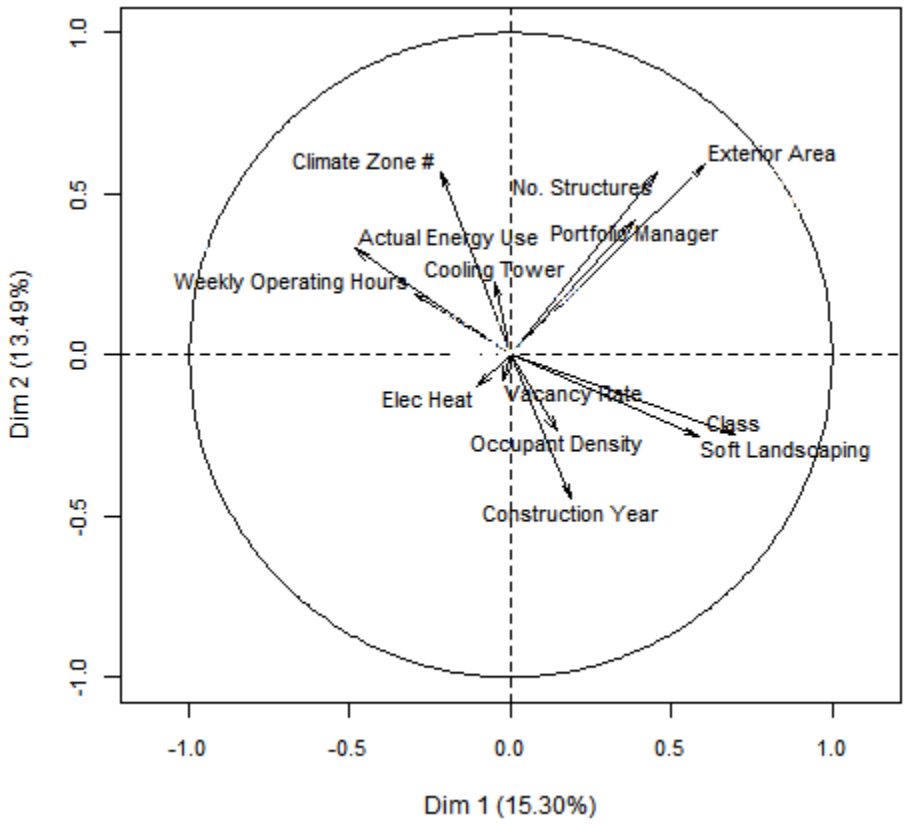


Figure 4-13 Loadings Map, 2015 dataset, AE, EA, CZ, PC 1 (Dim 1) vs PC 2 (Dim 2)

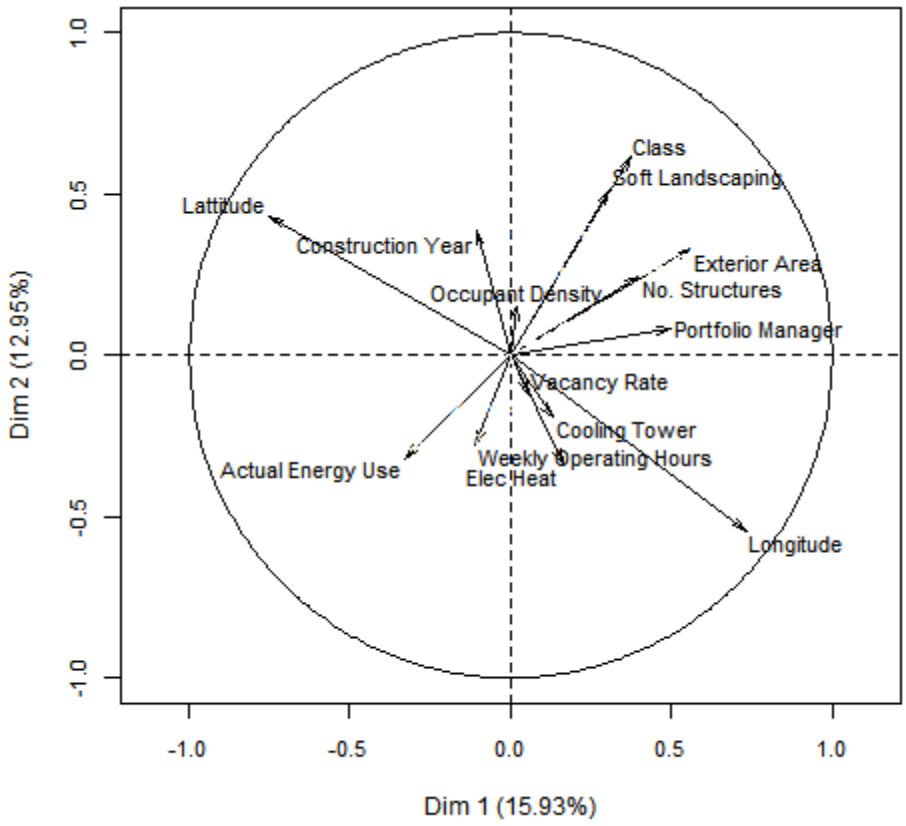


Figure 4-14 Loadings Map, 2015 dataset, AE, EA, L&L, PC 1 (Dim 1) vs PC 2 (Dim 2)

Neither of the loading maps illustrated feature loadings that contributed a significant amount of variance to one single axis. However, Figure 4-13 illustrated the Exterior area was the variable of most significance, contributing towards both the PC1 and PC2 relatively evenly. Figure 4-14 showed Longitude and Latitude being the primary variables, but neither contributed strongly to one axis. Both maps show Vacancy Rate, Occupant Density, Cooling Tower, Weekly Operating Hours and Electrically Heated as being among the least significance to PC 1 and 2.

As shiny does not output the (numerical) loadings for all of the features and principal components, it was solely employed to quickly visualize many iterations and determine which ones to analyze further. Since no particular iteration showed unique promise for dimensionality reduction, a deeper investigation into the loadings was conducted by using the **prcomp** function on two randomly selected iterations from the 2010-2015 dataset. It was anticipated that using this dataset would highlight clusters in the 2010 data according to the future 2015 Qualitative EUI results or the Qualitative EUI Change.

The plot in Figure 4-15 shows the scree plot and the cumulative variance for Iteration 1. The iteration, which originally had 16 features, required 13 out of 16 principal components to reach the 90% threshold. The scree plot was also analyzed for elbows, or changes in slope. It was assumed a sloped change would occur within a close range of the 90% cumulative variance threshold. One elbow was identified at PC 4, which only captured 65% of the variance. As the elbow is still 25% away from the threshold, it was not chosen as a boundary for dimensionality reduction.

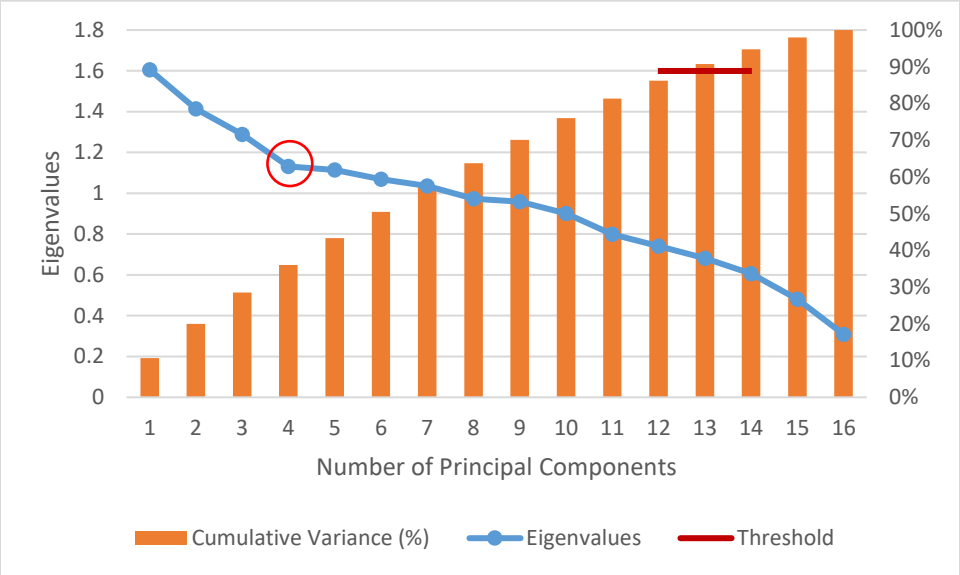


Figure 4-15 Iteration 1 Scree Plot and Cumulative Variance for 2010-2015 dataset

The plot in Figure 4-16 shows the scree plot and the cumulative variance for iteration two. The iteration, which originally had 13 features, required 11 out of 13 principal components to reach the 90% threshold. The scree plot was also analyzed for elbows, or changes in slope. Three elbows were identified at PC 5, PC 8 and PC 12 which captured 48%, 71% and 96% of the variance respectively. The combination chart indicates that PC 11 is an appropriate boundary for dimensionality reduction. For feature extraction, the intention was to identify the most significant features and remove the insignificant ones. A threshold criterion of Eigenvalue ≥ 0.7 or ≤ -0.7 within a principal component was used to determine if a feature was considered significant.

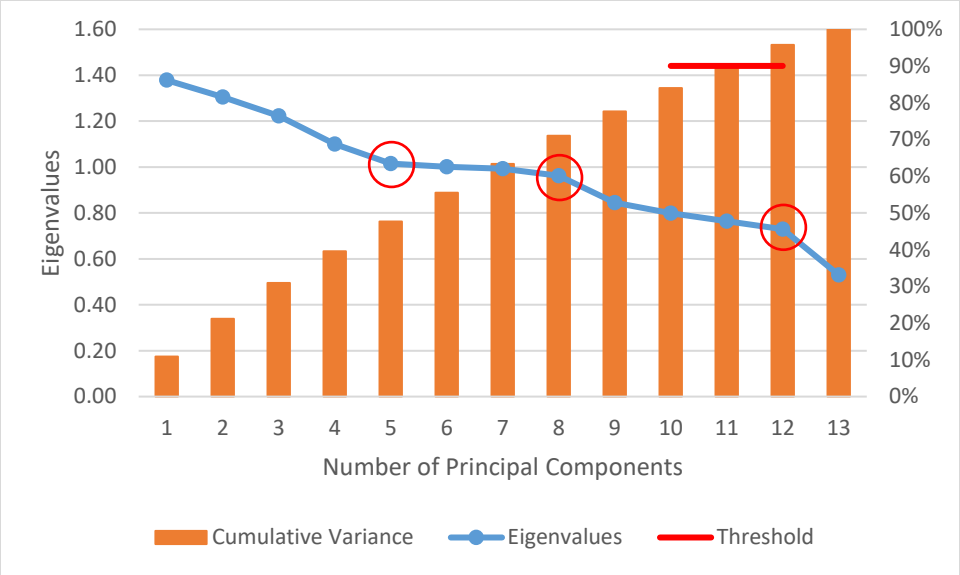


Figure 4-16 Iteration 2 Scree Plot and Cumulative Variance for the 2010-2015 dataset

The PC loadings, were further analyzed for Iteration 2. It was expected that each feature would significantly contribute to at least one principal component. The factors plot for each iteration were reviewed. Plots

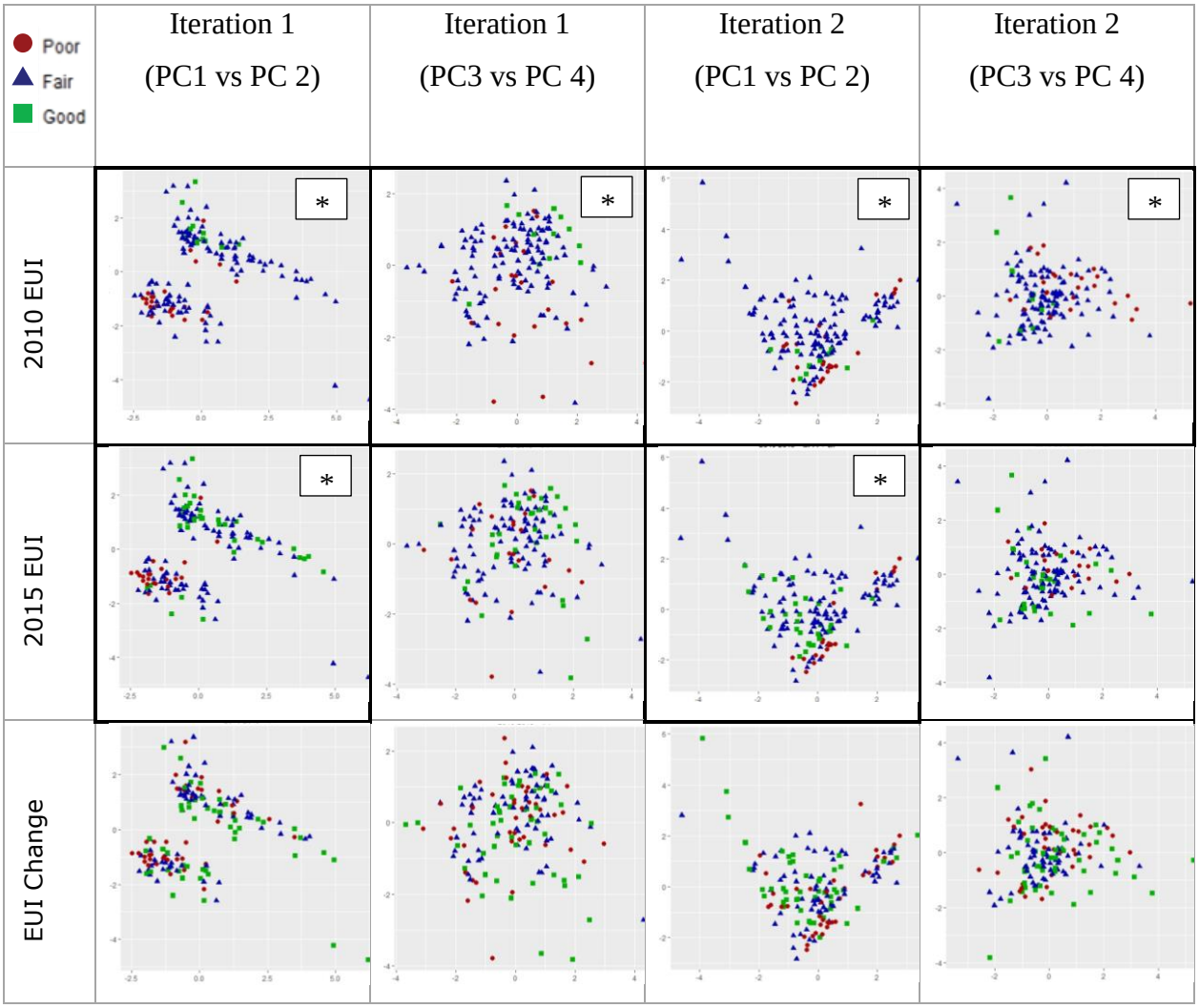
To the contrary, only one feature, Electrically Heated, reached the significance threshold for *any* principal component. It was assumed that perhaps the threshold was set too high. The results were again analyzed by lowering the significance threshold to >0.5 or <-0.5. Table 4-2 summarizes the findings, with loadings above original threshold highlighted dark grey and those above the revised threshold light grey. Only seven features reached the threshold, and none of PC 2, PC 3, PC 10, PC 11, or PC 12 had any significant loadings, thus PCA was not successful for feature extraction for this data set.

Table 4-2 Iteration 2 loadings for the 2010-2015 dataset

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13
2010 Actual Energy	0.13	0.20	0.44	-0.08	-0.07	0.16	-0.25	**-0.60**	0.37	0.02	0.24	-0.26	-0.16
Occupant Density	-0.22	-0.31	0.42	-0.08	-0.02	-0.23	-0.19	0.23	0.39	0.44	-0.40	0.01	0.16
No. of Structures	-0.26	0.40	0.07	0.35	-0.04	0.15	-0.07	0.41	0.05	0.43	0.47	0.05	-0.19
Exterior Area	-0.43	0.46	0.04	-0.04	-0.09	-0.03	-0.04	0.08	-0.05	-0.29	-0.17	-0.47	0.49
Operating Hours	0.24	0.21	-0.02	-0.30	**-0.56**	-0.15	-0.46	0.17	-0.35	0.09	-0.18	-0.04	-0.27
Asset Manager	-0.39	0.29	-0.03	0.04	0.29	-0.21	0.04	-0.46	-0.32	0.27	-0.35	0.21	-0.28
Building Class	**-0.54**	-0.26	-0.07	-0.02	-0.28	0.06	0.08	0.06	0.21	-0.39	-0.04	-0.09	**-0.58**
Climate Zone	0.11	0.48	-0.10	-0.27	-0.12	-0.27	0.26	0.07	**0.52**	-0.14	-0.05	0.47	0.00
Electrically Heated	0.13	-0.02	-0.20	0.36	0.20	**-0.73**	-0.28	0.01	0.16	-0.15	0.15	-0.26	-0.12
Vacancy Rate	0.13	0.11	-0.27	**0.57**	-0.10	0.38	-0.31	-0.06	0.25	-0.07	-0.48	0.15	0.03
Soft Landscaping	-0.35	-0.16	-0.29	-0.25	0.05	0.04	**-0.58**	-0.15	0.04	-0.05	0.32	0.38	0.29
Cooling Tower	0.08	0.09	-0.42	-0.43	0.46	0.24	-0.11	0.18	0.24	0.20	-0.13	-0.38	-0.23
Construction Year	-0.07	-0.12	-0.49	0.04	-0.48	-0.10	0.29	-0.32	0.13	0.46	0.08	-0.23	0.17

Score plots, which project the transformed data onto a pair of principal component vectors, were analyzed to help identify any patterns or clusters. The score plots are shown for iterations 1 and 2 in Table 4-3 and results were classified according to 2010 Qualitative EUI, 2015 Qualitative EUI, and Qualitative EUI Change. The six graphs that display clusters by class are indicated with an asterisk ("*").

Table 4-3 Score Plots for iterations 1 and 2 using the 2010 - 2015 Dataset and classified according to 2010 EUI, 2015 EUI or five-year EUI Change, respectively.



As seen above, *'Poor'* and *'Good'* building clusters are visible in all graphs classified according to 2010 EUI and three of the four graphs classified by 2015 EUI. This shows that the data points do cluster according to current and future energy performance. No clusters were observed in graphs classified according to EUI Change. Looking at the data as a whole, it was noted that the PC 1 vs PC 2 comparisons for iteration segmented the data into two discrete clusters; this appears to have been caused by the influence of Latitude and Longitude, which have a non-linear relationship with heating degree days, as included features rather than Climate Zone, which has a linear relationship.

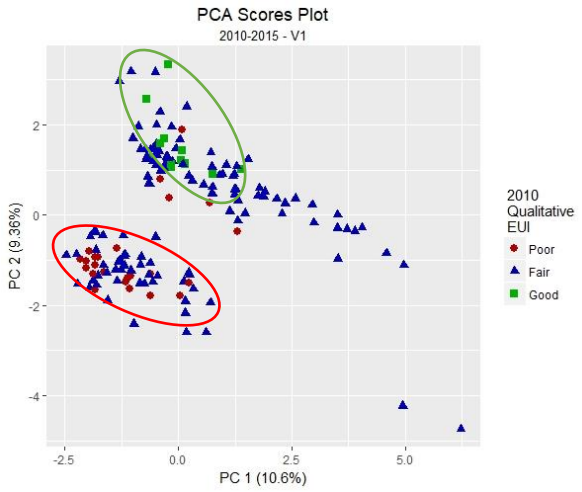


Figure 4-17 Iteration 1 PC1 vs PC 2 PCA Score Plot classified by 2010 Qualitative EUI

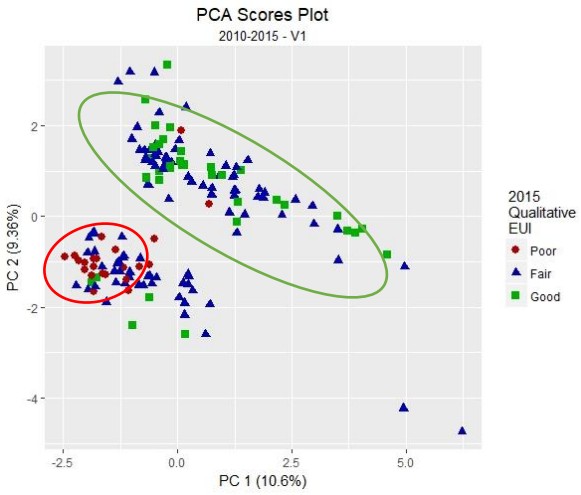


Figure 4-18 Iteration 1 PC1 vs PC 2 PCA Score Plot classified by 2015 Qualitative EUI

Figure 4-17 and Figure 4-18 highlight the class clusters in Iteration 1, PC 1 vs. PC 2. In Figure 4-17, the *'Good'* scores were best clustered with no data points falling outside the group. The majority of these points were also segregated from the *'Poor'* scores into separate clusters. The *'Poor'* scores were less tightly clustered than *'Good'*; five data points fell outside the main group, compared with zero for the latter. The *'Fair'* Scores were distributed between both clusters, likely caused by its location between the other classes, divided by an arbitrary boundary.

In Figure 4-18, the *'Good'* Scores were well clustered with four data points falling outside the group. The Poor Scores were better clusters with only 2 data point outside the main group. The *'Fair'* Scores were again distributed between the 2 large clusters.

Overall the PCA was unsuccessful for dimensionality reduction; however, it was useful for describing the structure of the dataset. The results demonstrate a measure of natural separation according to energy performance. It also illustrates that any energy usage or area variable can be used without negatively impacting the results. There is therefore no need to explore every variable combination in future algorithms. Lastly the results favour iterations with Latitude and Longitude vs Climate Zone, as illustrated by the ability of the PCA to demark the *'Poor'* and *'Good'* classes through clustering when iterations with Latitude and Longitude were used. This result indicates that only iterations with Latitude and Longitude needed to be explored for future

algorithms; separate iterations for Climate Zone are not expected to lead to better results. These observations were useful for reducing the number of iterations performed on future algorithms.

4.3 Linear Discriminant Analysis

It was expected that the application of LDA would identify clusters allowing classification of new samples, and allow important features to be distinguished from unimportant features.

The overall and within-class correct classification rates for each iteration were compared to assess whether LDA was able to find class clusters within the data (Table 4-4). For the initial six runs, iteration 1 and iteration 4 resulted in the highest overall classification accuracy. This shows the model is good at separating data according to the current year (2010) Qualitative EUI*;* however, it is 20-40% less successful at achieving separation with future (2015) Qualitative EUI and Qualitative EUI Change classes. When the within-class accuracy rates were compared, it was observed that all of the models, including iterations 1 and 4, were very poor at classifying the *'Good'* individuals.

Table 4-4 LDA Classification Accuracy rates by iteration

Iteration	Input Variables	Outcome Variable	Overall Classification Accuracy	Class *'Poor'* Accuracy	Class *'Fair'* Accuracy	Class *'Good'* Accuracy
1	Set A	2010 Q.EUI	84.3%	61.3%	94.1%	28.6%
2	Set A	2015 Q.EUI	66.2%	59.3%	80.8%	30.4%
3	Set A	Q.EUI Change	44.9%	32.0%	67.0%	23.3%
4	Set B	2010 Q.EUI	84.3%	**64.5%**	94.1%	21.4%
5	Set B	2015 Q.EUI	67.7%	48.1%	88.8%	21.7%
6	Set B	Q.EUI Change	42.4%	34.0%	63.6%	18.3%
7	Set C	2010 Q.EUI	84.3%	58.1%	94.8%	28.6%
8	Set D	2010 Q.EUI	**85.4%**	51.6%	**96.1%**	**42.9%**

Results in the first six iterations indicated pointedly larger loadings for any area related variables. For iteration 1, GFA had a loading that was on average four orders of magnitude greater than the other loadings. For iteration 4, Exterior Area had a loading that was on average seven orders of magnitude greater than the other loadings. Instead of being an indicator of significance, the large loading likely indicates over-representation within the dataset and a dependent variable. The energy variables, occupant density, and vacancy rate were all calculated as a function of area and

are likely confounding the model. This was explored by performing another iteration with all area variables removed. Gross Floor Area and Enclosed Parking Area were removed from iteration 1 to create iteration 7. Iteration 1 was chosen as it was 7.2% better at classifying *'Good'* buildings than iteration 4.

The loadings of iteration 7, as seen in Table 4-5, displayed a tight range with no abnormally high values. This permitted the observation of which features contribute more to the separability of the data. The overall classification accuracy remained the same, implying that the model did not lose integrity when the area variable was removed.

Table 4-5 LDA Iteration 7 Loadings and Variance

Variable	Loading	Variance		
	LD1	LD2	LD1	LD2
2010 Actual E&C	-3.33	0.68	**21%**	**5%**
2010 Actual Thermal	-1.89	-1.31	**12%**	**10%**
Building Manager	0.12	-1.11	1%	**9%**
Latitude	-0.85	-0.87	**5%**	**7%**
Longitude	**6.75**	2.18	**43%**	**17%**
Construction Year	2.35	**4.45**	**15%**	**35%**
Number of Structures	0.02	-0.11	0%	1%
Building Class	0.06	0.15	0%	1%
Electrically Heated	0.02	0.15	0%	1%
Cooling Tower	0.13	-0.15	1%	1%
Soft Landscaping	0.03	0.04	0%	0%
Occupant Density	-0.02	0.50	0%	**4%**
Vacancy Rate	-0.14	0.20	1%	2%
Weekly Operating Hours	0.05	-0.72	0%	**6%**

After reviewing the loadings plot, any feature with a loading below 3% was considered insignificant and removed to create iteration 8. This last iteration showed a slightly improved overall classification accuracy which was 1.2% better than iterations 1 and 7. This indicates that the integrity of the model is not compromised if Number of Structures, Building Class, Electrical Heat, Cooling Tower, and Soft Landscaping are not included. In fact, the ability to classify *'Good'* building improved by 14.3%. Refer to Table 4-6 for the iteration 8 loadings and variances. It can be seen that the separability of LD1 corresponds greatest to Longitude and is inversely related to Actual E&C Energy. In addition, the separability of LD2 is mostly explained by Construction Year and Longitude.

Table 4-6 LDA Iteration 8 Loadings and Variance

Variable	Loading		Variance	
	LD1	LD2	LD1	LD2
2010 Actual E&C Energy	-3.43	1.17	**22%**	**8%**
2010 Actual Thermal Energy	-1.95	-1.50	**12%**	**11%**
Building Manager	0.12	-1.43	1%	**10%**
Latitude	-0.88	-1.03	**6%**	**7%**
Longitude	6.97	2.17	**44%**	**15%**
Construction Year	2.46	5.46	**15%**	**38%**
Occupant Density	-0.02	0.64	0%	**4%**
Weekly Operating Hours	0.05	-0.92	0%	**6%**

The score plots were all compared. Linearly separated class clusters were observed in LDA score plots that were classified according to current year EUI. The best separation was observed in iteration 8 as seen in Figure 4-19. The plots classified according to 2015 Qualitative EUI and EUI Change were not observed to have linearly separated classes as seen in Figure 4-20.

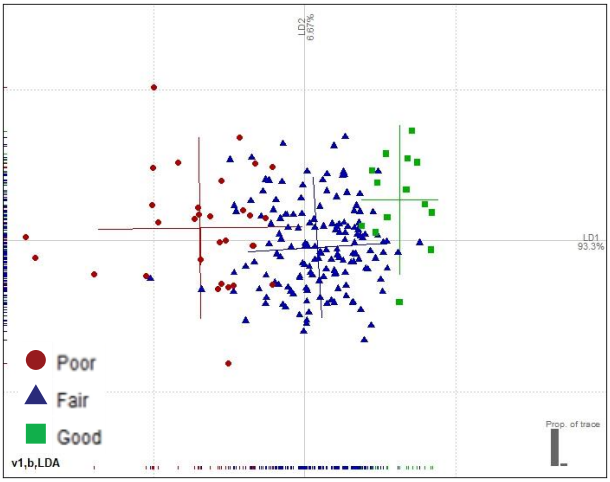


Figure 4-19 Iteration 8 LDA Score plot classified by 2010 Qualitative EUI. Optimal separability is observed.

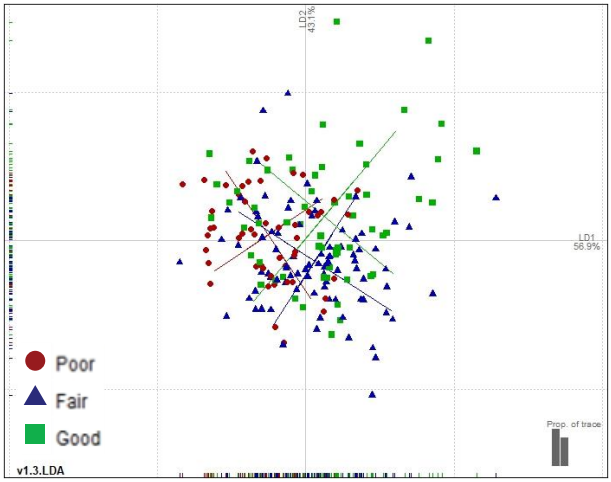


Figure 4-20 Iteration 3 LDA Score plot classified by EUI Change. Poor separability is observed.

The results indicate that data structure contains natural separation according to the current year Qualitative EUI. Also, the results suggest that the most important variables for class separability are Actual Electricity and Cooling Energy, Actual Thermal Energy, Property Manager, Latitude, Longitude, Construction Year, Occupant Density and Weekly Operating Hours. Moreover, the LDA algorithm can be used for predicting new samples with 85.4% accuracy, however the likelihoods of misclassification for *'Poor'* and *'Good'* buildings are 49% and 58%, respectively.

Overall LDA successfully analyzed the data structure and significant features and identified clusters according to current year Qualitative EUI. However, LDA does not offer confident prediction rates.

4.4 k-Nearest Neighbours

For k-nearest neighbours, two outcomes were anticipated. First, that the KNN algorithm would be able to correctly classify the new samples. Second, that KNN graphs would illustrate separability between classes when different variable combinations were explored. As illustrated in Table 4-7, the first expectation proved to be reasonable for the overall classification and true for *'Fair'* class accuracy, but untrue for *'Poor'* and *'Good'* classes.

Table 4-7 KNN Classification Accuracy rates by iteration

Iteration	Outcome Variable	Overall Classification Accuracy	*'Poor'* Classification Accuracy	*'Fair'* Classification Accuracy	*'Good'* Classification Accuracy
1	2010 Qualitative EUI	**83.3%**	12.5%	83.1%	0.0%
2	2015 Qualitative EUI	71.7%	20.0%	92.9%	**23.1%**
3	2010-20215 Qualitative EUI Change	48.3%	**38.5%**	76.9%	19.0%
4*	2010 Qualitative EUI	78.4%	6.5%	**100.0%**	0.0%
*Cross validated KNN Model					

The overall classification accuracy rate was the highest with iteration 1. The corresponding KNN model cannot be considered a success as it performed poorly at classifying buildings with *'Poor'* and *'Good'* 2010 Qualitative EUI values. The remaining iterations were also heavily biased to classify a building as *'Fair'* despite either a different outcome variable or a different KNN function being used. This same bias was seen with the PCA and LDA prediction results and is likely because as with each outcome variable, the quantity of *'Fair'* buildings is greater than *'Poor'* or *'Good'*.

Iteration 2 had the highest accuracy for classifying *'Good'* 2015 Qualitative EUI in buildings, indicating that it is easier for a KNN model to correctly categorize a *'Good'* 2015 Qualitative EUI given 2010 data. It is thereby considered the most successful model.

The poorest performing overall accuracy was from iteration 3, though it performed the best in classifying *'Fair'* buildings and was the second-best iteration for classifying *'Good'* buildings. The amount of buildings falling into each class for Qualitative EUI Change were much more evenly distributed than with other outcome variables. The result indicates that KNN was better able to define divisible boundaries between 2015 Qualitative EUI classes. This likely explains why the within class accuracy rates were slightly better. Overall the results indicate that 2010 data cannot be used to successfully predict the Qualitative EUI Change.

The second expectation was disproved as the KNN graphs were generally poor at clustering by class. For each iteration the classification probabilities output by the KNN model were plotted behind a scatter plot displaying the actual data points. The probabilities are communicated via size and transparency; the larger and darker the point, the higher the probability a point belongs to a particular class. The solid large points represent the actual individuals.

A successful graph would display individual data points grouped by class, residing in a space with matching class probabilities predicted by the KNN model. Refer to Figure 4-21 for a concept diagram of a successful KNN graph.

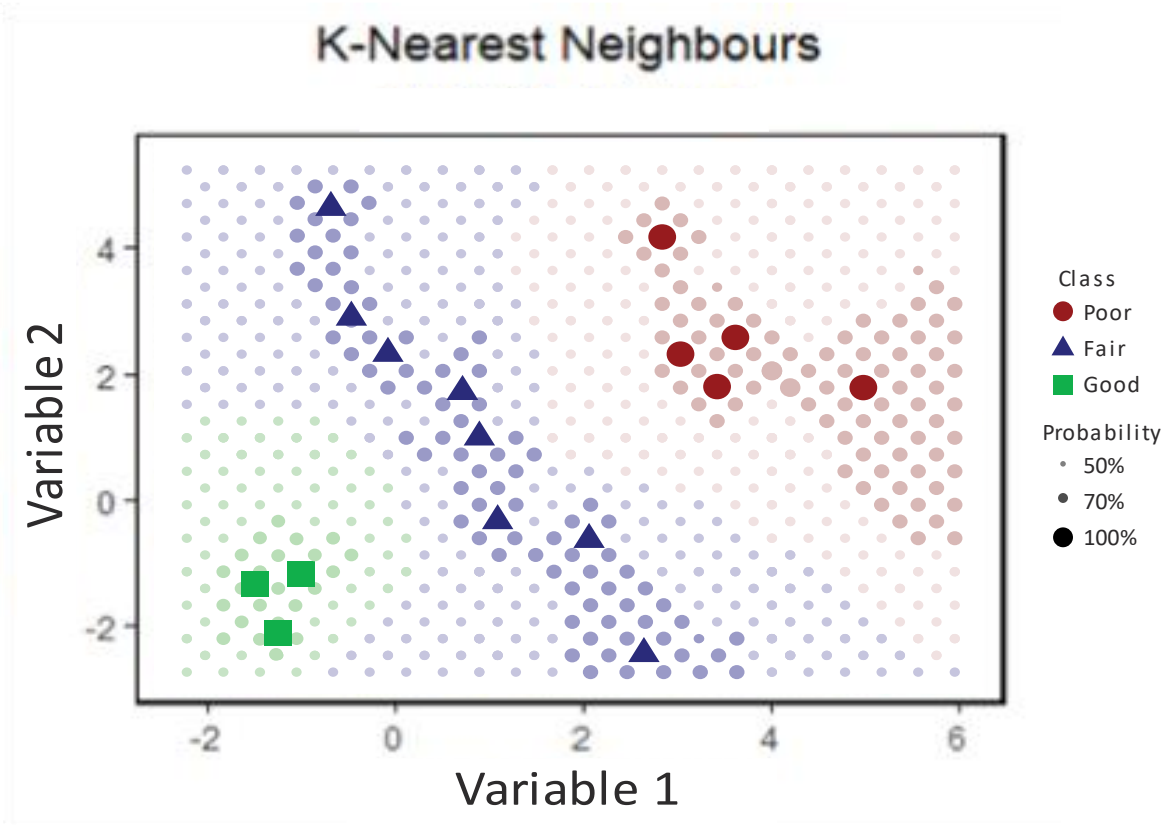


Figure 4-21 Example of the overlaid scatter plot showing a successful KNN model

Generally, the graphs did not cluster well. Iteration 1 – 2010 E&C Energy vs 2010 Thermal Energy (see Figure 4-22), resulted in a graph with moderately successful separability of the 2010 Qualitative EUI classes. The *'Good'*, *'Fair'* and *'Poor'* buildings can be seen to group amongst themselves naturally, with some overlap. However, the KNN modelled probabilities do not line up neatly with the clusters of actual individuals.

The graph likely demonstrates class separability because 2010 Qualitative EUI is dependent on 2010 Thermal Energy and 2010 Actual E & C. This graph was produced as an exploratory measure and is not useful in producing insight about energy data trends.

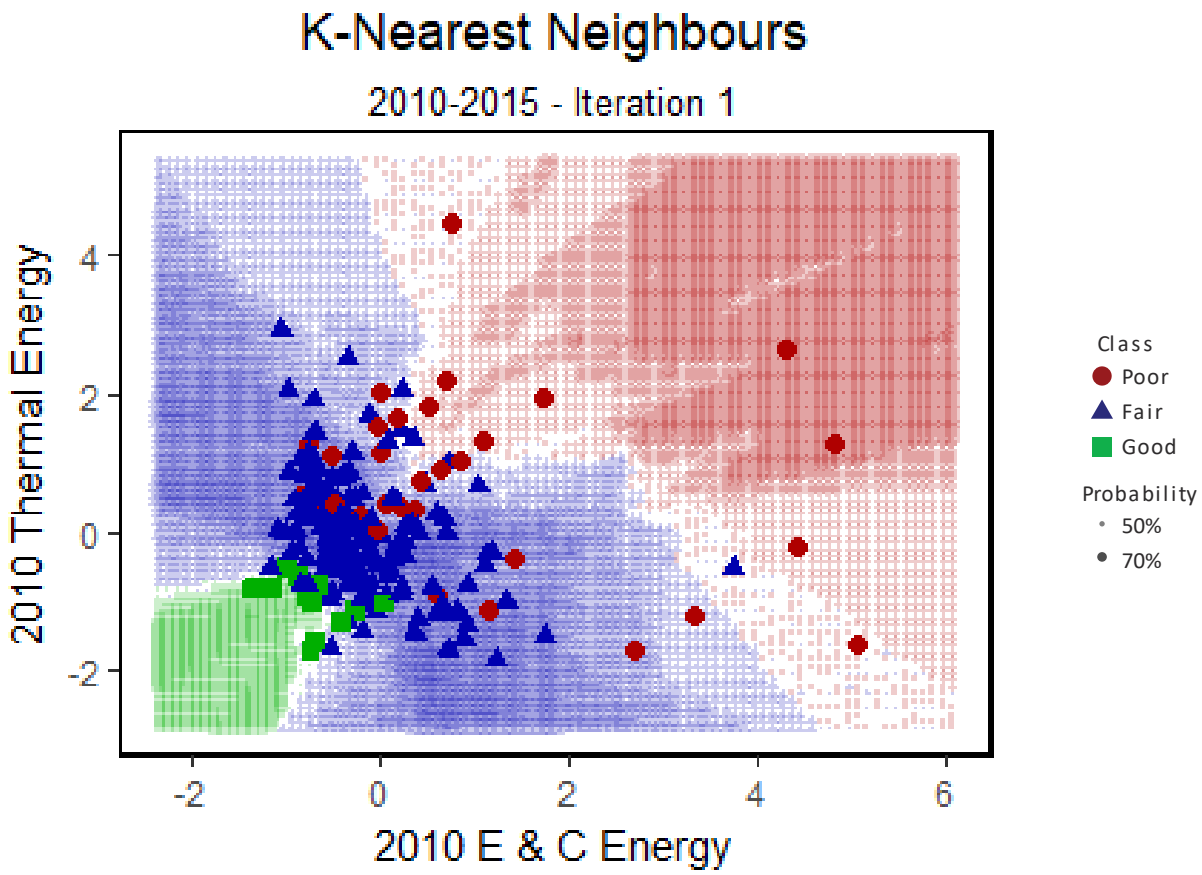


Figure 4-22 Estimated probabilities from KNN Model vs Actual individuals for 2010 Thermal Energy and 2010 E&C classified by 2010 Qualitative EUI. Moderate separability is observed.

errorThe same two variables were explored in iteration 2 - 2010 E&C Energy vs 2010 Thermal Energy (Figure 4-22), displayed very weak separability of the 2015 Qualitative EUI classes. This implies that the 2015 Qualitative EUI is not explained well by the 2010 E&C Energy vs 2010 Thermal Energy.

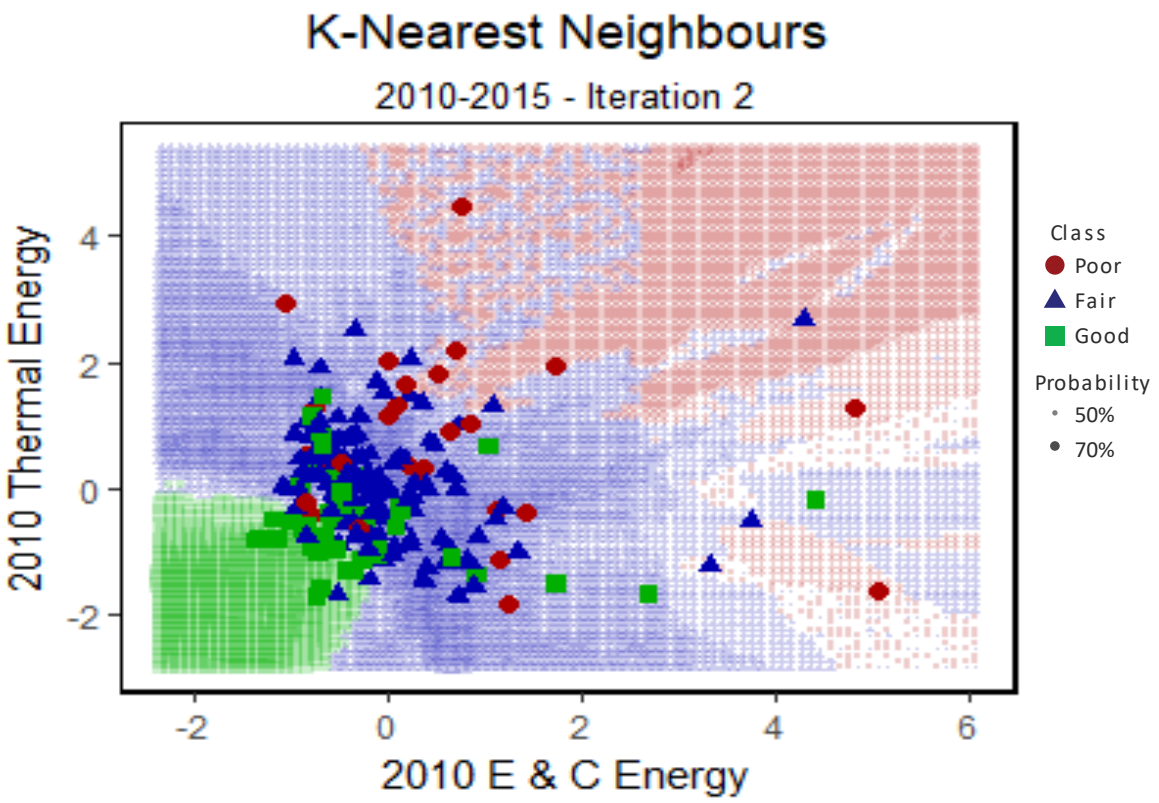


Figure 4-23 Estimated probabilities from KNN Model vs Actual individuals for 2010 Thermal Energy and 2010 E&C classified by 2015 Qualitative EUI. Weak separability is observed.

Across all iterations, the resulting graphs are inadequate in their class separability. An example of a typical poorly clustered graph can be seen in Figure 4-24. Datapoints of different classes can be seen overlapping each other and the KNN probabilities do not correlate to the actual individuals.

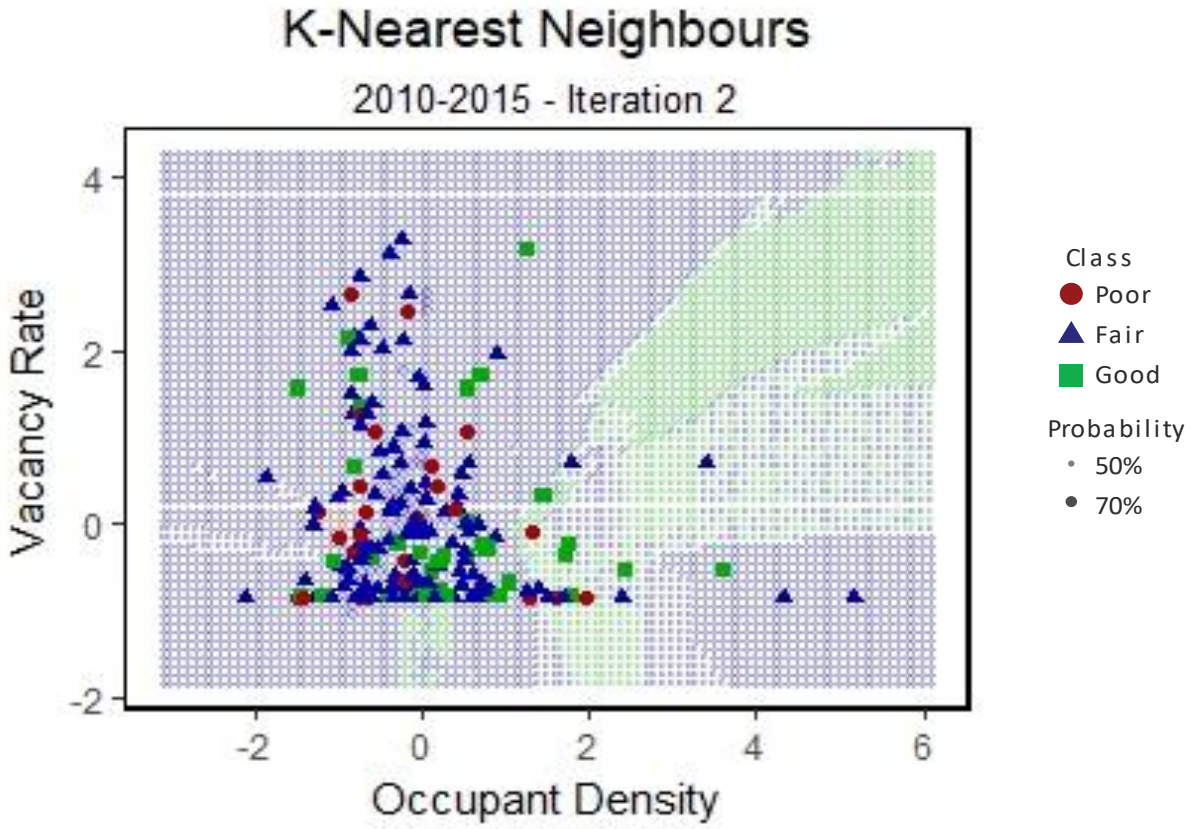


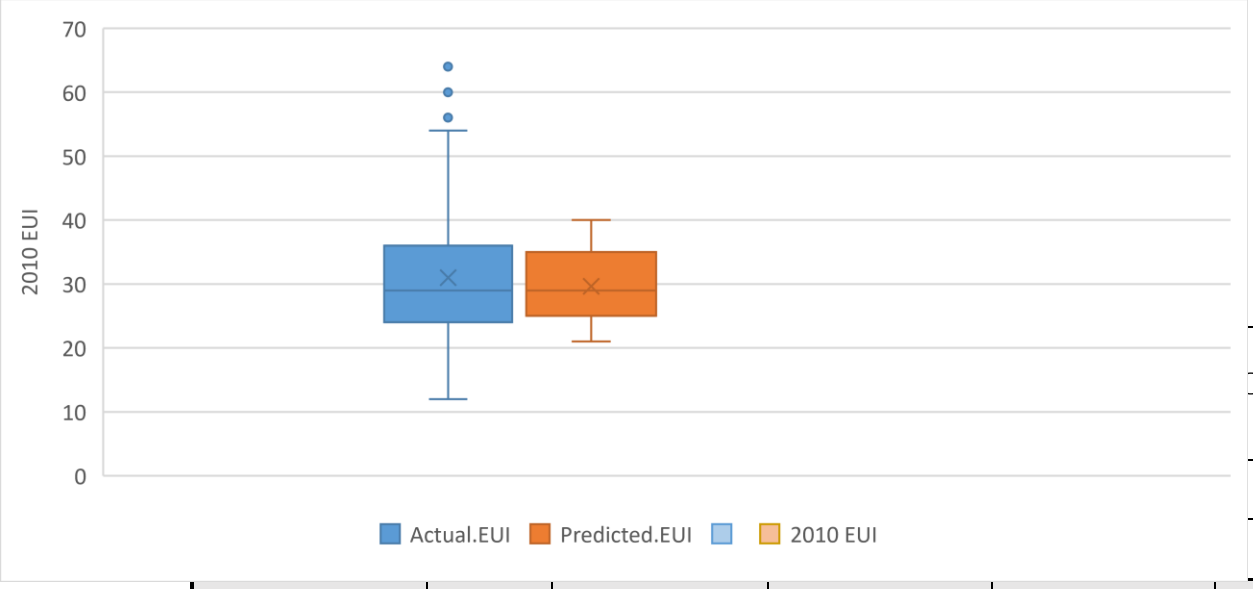
Figure 4-24 Estimated probabilities from KNN Model vs Actual individuals for Vacancy Rate and Occupant Density classified by 2015 Qualitative EUI. Poor separability is observed.

Figure 4-24 estimated probabilities from KNN Model vs Actual individuals for Vacancy Rate and Occupant Density classified by 2015 Qualitative EUI. Poor separability is observed.

In conclusion, the KNN algorithm was not successful at classifying new samples, even when cross validation was used. Furthermore, it is difficult to use KNN models to highlight clusters within the 20 by '15 dataset when plotting the datapoints in a two-dimensional space and using independent input variables.

4.5 Multiple Linear Regression

For Multiple Linear Regression, three outcomes were anticipated. First, that the MLR algorithm would be able to correctly classify the new samples. Second, the model would illustrate which features contributed significantly to each dependent variable. Third, that MLR graphs would illustrate a strong, positive correlation between the predicted and actual values. As illustrated in Table 4-8, the first expectation proved to be reasonable for the overall classification and true for



					'Good' Classification Accuracy
					0.0%
					0.0%
2 (2015 EUI)	Train	47.8%	20.0%	75.0%	5.1%
2 (2015 EUI)	Test	67.8%	14.3%	90.2%	18.2%
3 (*EUI Change*)	Train	68.3%	**86.2%**	5.0%	0.0%
3 (*EUI Change*)	Test	6.7%	0.0%	0.0%	**100.0%**

**note that the null accuracy (accuracy if the model predicted that all elements belonged to the dominant class) is 71%, which skews the 'Fair' (dominant class) accuracy upwards.*

'Poor' Region

'Fair' Region

'Good' Region

Figure 4-25 Distributions within the Actual and Predicted values for the training and test data using 2010-2015 dataset

The highest overall classification accuracy rates were observed when predicting 2010 EUI in both the test (78%) and training (68.4%) dataset in iteration 1. Unfortunately, the results are deceptive as the iteration 1 MLR model performed poorly at classifying buildings with *'Poor'* and *'Good'* 2010 EUI values. Iteration 2 was also heavily biased to classify a building as *'Fair'* despite predicting 2015 EUI instead. This same bias was observed with the PCA, LDA and KNN

prediction results and is likely because the distribution of building classes is more heavily weighted to *'Fair'* buildings. In Figure 4-25, the distributions can be observed to lie dominantly in the *'Fair'* Region. It would could be assumed that the fault lies with the chosen boundaries which result in 12% of 2010 buildings being classified as *'Poor'* and 10% being classified as *'Good'* however the classification boundaries were defined to identify the best and worst performing buildings.

Iteration 2 had low accuracy rates for classifying *'Good'* and *'Poor'* Buildings, but the model still performed slightly better than both iterations 1 and 2, indicating that it is easier for an MLR model to correctly classify a low frequency 2015 EUI classes given 2010 data.

The poorest performing overall accuracy was from iteration 3, though it performed the best in classifying *'Poor'* buildings in the training set and *'Good'* buildings in the test set. The amount of buildings falling into each class for the Qualitative EUI Change variable had the least accurate distribution. Overall the results indicate that 2010 data cannot be used to successfully predict the EUI Change.

Next, the statistical summaries output by each model were analyzed for features which contributed significantly to the dependent variable. Significance was defined as a $p<0.05$, indicating less than 5% chance variability. Despite each model targeting different dependent variables, they mostly shared the same significant features (see Table 4-9). Longitude, Latitude, Construction Year and Occupant Density were identified as significant in two out of three MLR models.

Table 4-9 Statistically significant features identified by the three MLR models

p<0.05	Iteration 1 (2010 EUI)	Iteration 2 (2015 EUI)	Iteration 3 (*EUI Change*)
Longitude	X	X	X
Latitude		X	X
Construction Year	X	X	
Occupant Density		X	X

Lastly, the actual values were plotted against the predicted values output by the MLR models. A successful graph would display a strong, linear correlation between the two variables. Generally, the graphs showed low to moderate correlation, though iteration 2 training data (Figure 4-26)

resulted in a graph with moderately successful class separability. The same model applied to new test data yielded a graph demonstrating poor linear correlation (see Figure 4-27). This outcome indicates the MLR model is not robust when confronted with new observations.

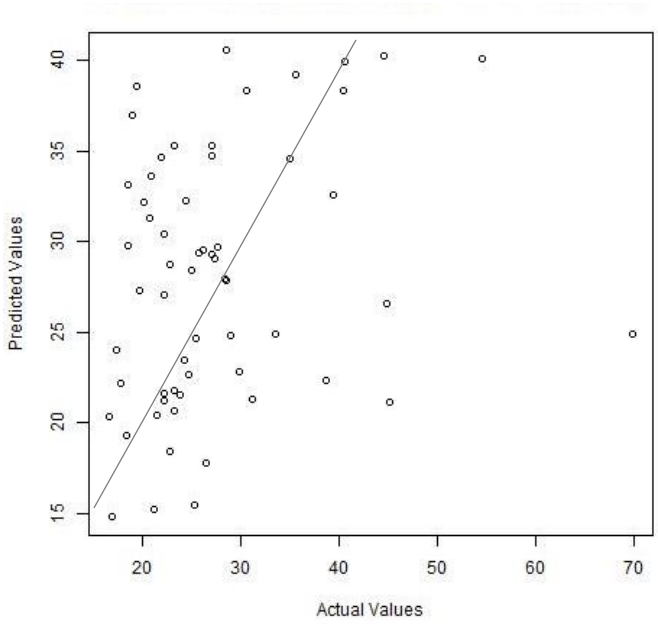


Figure 4-26 Actual vs Predicted 2015 EUI using 2010-2015 training dataset. Moderate linear correlation is observed. (Predicted = actual curve shown for reference)

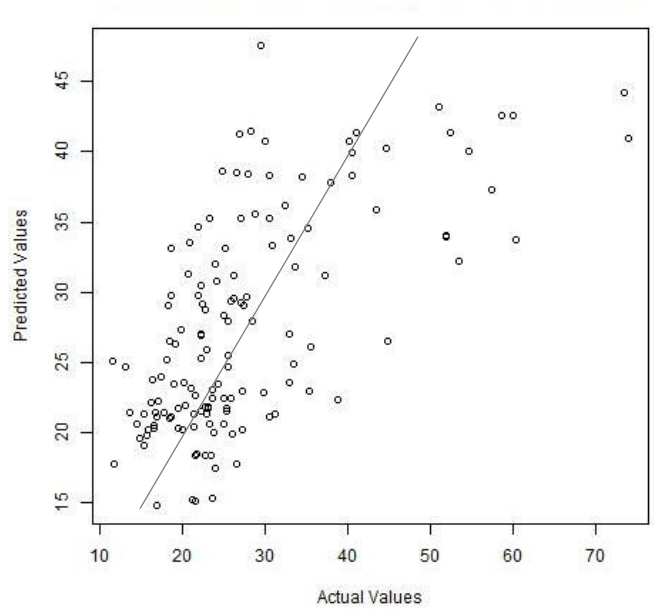


Figure 4-27 Actual vs Predicted 2015 EUI using 2010-2015 test dataset. Poor linear correlation is observed. (Predicted =. actual curve shown for reference)

In conclusion, the MLR models are not successful at predicting new observations, as evidenced by the classification accuracy rates at the poor correlation observed in the graphs. The MLR models are useful at highlighting Longitude, Latitude, Construction Year and Occupant Density as overall significant variables.

4.6 Support Vector Machines

For Support Vector Machines, two outcomes were anticipated. First, that the SVM algorithm would be able to correctly classify the new samples. Second, that MLR graphs would illustrate a strong, positive correlation between the predicted and actual values.

Iterations 1 and 2 were run using regression analysis. As summarized in Table 4-10, the RMSE for each regression iteration and kernel were calculated. It is interesting to note that within each iteration, the changing kernels altered each RMSE by less than 0.02 (Note the normalized numbers are unitless absolute values). It appears that the choice of kernel had little effect on predictive accuracy. No iteration achieved the threshold of 5 ekWh/sf/yr to be considered successful. The RMSE Scores were on average six times higher than the threshold for denormalized EUI predictions.

Table 4-10 RMSE for iterations 1 and 2

	Iteration 1 (2010 EUI)		Iteration 2 (2015 EUI)	
Kernel	Train	Test	Train	Test
Linear	**32.27**	29.85	29.86	29.11
Polynomial	32.28	**29.83**	**29.85**	29.08
Sigmoid	**32.27**	**29.83**	29.86	**29.07**
Radial	32.28	29.85	**29.85**	29.09
Success Threshold: 5 Ideal Threshold: 2				

As illustrated in Table 4-11, the first expectation proved to be reasonable for the overall classification and true for *'Fair'* class accuracy, but untrue for *'Poor'* and *'Good'* classes. The models performed poorly at classifying the *'Poor'* and *'Good'* classes; there was a heavy bias for the model to predict only one class when confronted with new observations. The Linear and Polynomial kernels predicted all new observations to be *'Poor'* while the Sigmoid and Radial kernels predicted all new observations as *'Fair'*.

The Radial kernel performed best on the iteration 3 (2010 Qualitative EUI) training dataset with only one misclassified observation, therefore the same kernel was used for iteration 4. The

iteration 4 radial kernel model perfectly classified the training data, but proved to not be able to predict more than one class when applied to new data.

Table 4-11 Classification Accuracy rates for iterations 3 and 4

Iteration	Kernel	Dataset	Overall Class. Accuracy	*'Poor'* Class. Rate	*'Fair'* Class. Rate	*'Good'* Class. Rate
3 (2010 Qualitative EUI)	Linear	Train	77%	0%	77%	0%
		Test	10%	10%	0%	0%
	Polynomial	Train	92%	9%	77%	6%
		Test	10%	10%	0%	0%
	Sigmoid	Train	77%	0%	77%	0%
		Test	85%	0%	85%	0%
	Radial	Train	99%	15%	77%	7%
		Test	85%	0%	85%	0%
4 (2015 Qualitative EUI)	Radial	Train	100%	14%	63%	23%
		Test	71%	0%	71%	0%

Overall, the SVM models are not successful at predicting new observations, either using regression or classification models, as evidenced by the RMSEs and the classification accuracy rates. The SVMs models are useful at highlighting that the data may be dispersed in a radial fashion and would require a radial kernel to become linearly separable.

4.7 Decision Trees

It was anticipated the Decision Tree algorithm would have a low RMSE when using regression algorithms and a high accuracy rate when classifying new observations using the classification algorithm.

As summarized in Table 4-12, the RMSEs for each iteration using regression analysis were calculated. The RMSEs were calculated using the normalized actual and predicted values, so that each variable may be compared against the other. There is little variation in RMSE as the dataset and outcome variable changed within each iteration; all results have an RMSE between 0.09 and 0.16, a range of only 0.07. The lowest RMSE was attained using the 2013 dataset to predict 2013

Actual Electricity and Cooling Energy. The highest RMSE resulted from using the 2015 dataset to predict 2015 EUI.

Overall the RMSE is not low enough to confidently predict the energy performance of buildings, however it does provide a rough estimate. It appears that the choice of outcome variable or year of data has little effect on predictive accuracy.

Table 4-12 Normalized RMSE for predicting each Energy Performance Measure with each Dataset

Outcome Variable	Dataset						Success Threshold	Ideal Threshold
	2011	2012	2013	2014	2015	2015|2010		
EUI	0.15	0.15	0.14	0.15	0.16	0.15	.09	.04
Actual Total	0.14	0.14	0.12	0.14	0.13	0.14	.08	.03
Actual E&C	0.14	0.13	0.09	0.11	0.11	0.14	.05	.02
Actual Thermal	0.14	0.15	0.14	0.14	0.15	0.14	.03	.01

As seen in Figure 4-28, the classification accuracy rate of Random Forest Decision Trees was evaluated for each year and energy performance measure with error margins varying between 0% and 25%.

As seen with the RMSE calculations, the year and energy performance measure appeared to have little effect on the classification accuracy. Naturally, increasing the error margin led to an increase in the classification accuracy. Most iterations, 22 out of 24, required an error margin of 25% to attain a classification accuracy of 90%, though 10 iterations achieved 90% with an error margin of 20%.

It is interesting to note that the graphed results in Figure 4-28 (f) vary little from Figure 4-28 (a-e), considering they are measuring the ability of the model at classifying 2015 energy measures using 2010 data, while the other iterations are predicting current year energy data. This indicates that the accuracy of decision tree models will not improve significantly when calculating current year energy use as opposed to future energy use.

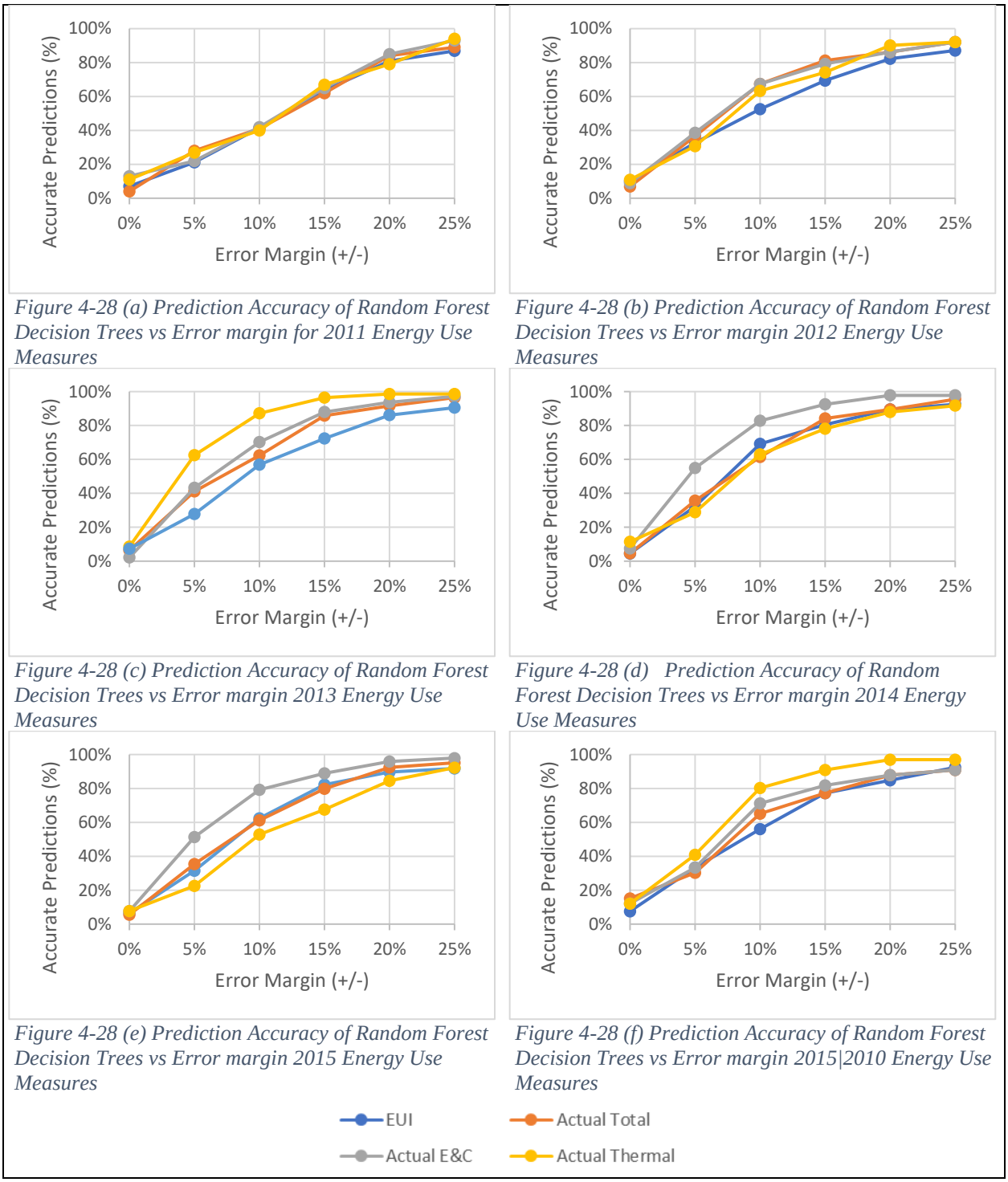


Figure 4-28 (a) Prediction Accuracy of Random Forest Decision Trees vs Error margin for 2011 Energy Use Measures

Figure 4-28 (b) Prediction Accuracy of Random Forest Decision Trees vs Error margin 2012 Energy Use Measures

Figure 4-28 (c) Prediction Accuracy of Random Forest Decision Trees vs Error margin 2013 Energy Use Measures

Figure 4-28 (d) Prediction Accuracy of Random Forest Decision Trees vs Error margin 2014 Energy Use Measures

Figure 4-28 (e) Prediction Accuracy of Random Forest Decision Trees vs Error margin 2015 Energy Use Measures

Figure 4-28 (f) Prediction Accuracy of Random Forest Decision Trees vs Error margin 2015|2010 Energy Use Measures

Figure 4-28 Prediction Accuracy of Random Forest Decision Trees vs Error margin for Energy Use Measures

The results of the random forest decision tree classification algorithm are shown below in Figure 4-29. The algorithm was trained to predict current year qualitative energy use as well as 2015 qualitative energy use using 2010 data. The 2014 iteration had the worst predictive accuracy

while 2011 iteration displayed the best. When classifying using three classes, the accuracy rate varied between 54% and 77% and when using two classes, *'Good'* or '*Not Good'* (i.e. *'Poor'* and *'Fair'* together) the rate was between 81% and 95%.

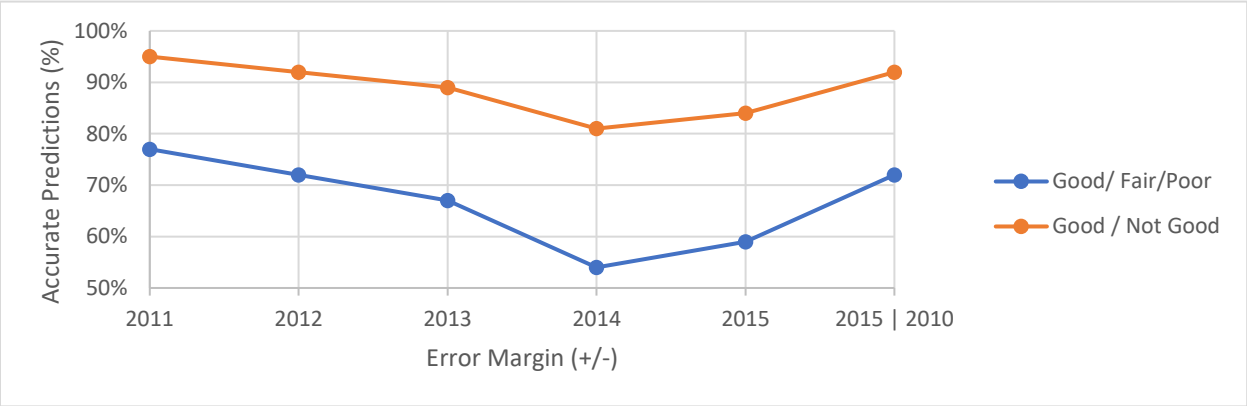


Figure 4-29 Accuracy of Random Forest Decision Trees for Predicting Qualitative Energy Use for Current and Future Years

These accuracy rates appear to be high except when the within-class classification accuracy rates are analysed, the same previously witnessed bias to *'Fair'* results were observed. Refer to Table 4-13 to see the summary of the overall and within-class classification accuracy rates for predicting Qualitative EUI with each dataset. Conversely, the correct classification of *'Good'* classes only occurs between 7% and 29% of the time. The highest classification accuracy rate was achieved using the 2013 dataset and the lowest was attained using the 2011 Dataset. Finally, the correct classification of *'Poor'* classes only occurs between 0% and 29% of the time. The highest classification accuracy rate was achieved using the 2011 dataset and the lowest was attained using the 2013 Dataset; this observation is the inverse of the *'Good'* Classification Accuracy Rate.

Table 4-13 Classification Accuracy rate for Decision Tree Classification Algorithm

Dataset	Overall Classification Accuracy Rate	*'Poor'* Classification Accuracy Rate	*'Fair'* Classification Accuracy Rate	*'Good'* Classification Accuracy Rate
2011	77.00%	28.57%	94.87%	6.67%
2012	71.57%	11.11%	90.54%	26.32%
2013	66.90%	0.00%	84.47%	29.63%
2014	54.48%	14.18%	73.88%	11.94%
2015	58.62%	18.18%	90.00%	29.03%
2015|2010	71.57%	11.11%	90.54%	26.32%

Overall the random forest decision tree models were not able to accurately predict the class without a bias towards *'Fair'* results or the energy performance measure within a meaningful range. The results did indicate that the dataset and energy performance measure chosen have little impact on the quality of the prediction.

4.8 Artificial Neural Networks

It was anticipated the Artificial Neural Networks algorithm would have a low RMSE and high accuracy rate when classifying new observations

Seven iterations were run, exploring the retention of a varying quantity of variables with a different quantity of layers in the ANN chart when predicting Actual EUI Change. Ten-fold cross validation was used. As the 3-layer ANN algorithms often failed to converge, only 1-layer and 2-layer ANN models were used to predict the % EUI Change.

The results of the best tuned models within each iteration are summarized in Figure 4-30. The models trained to predict the percent change in EUI from 2010 to 2015 performed significantly worse than the models predicting Actual EUI Change. There was little variation in the predictive accuracy of the latter models, which averaged at 62.5%. Generally, the more layers, the better the model performed however the improvement was not more than 5%.

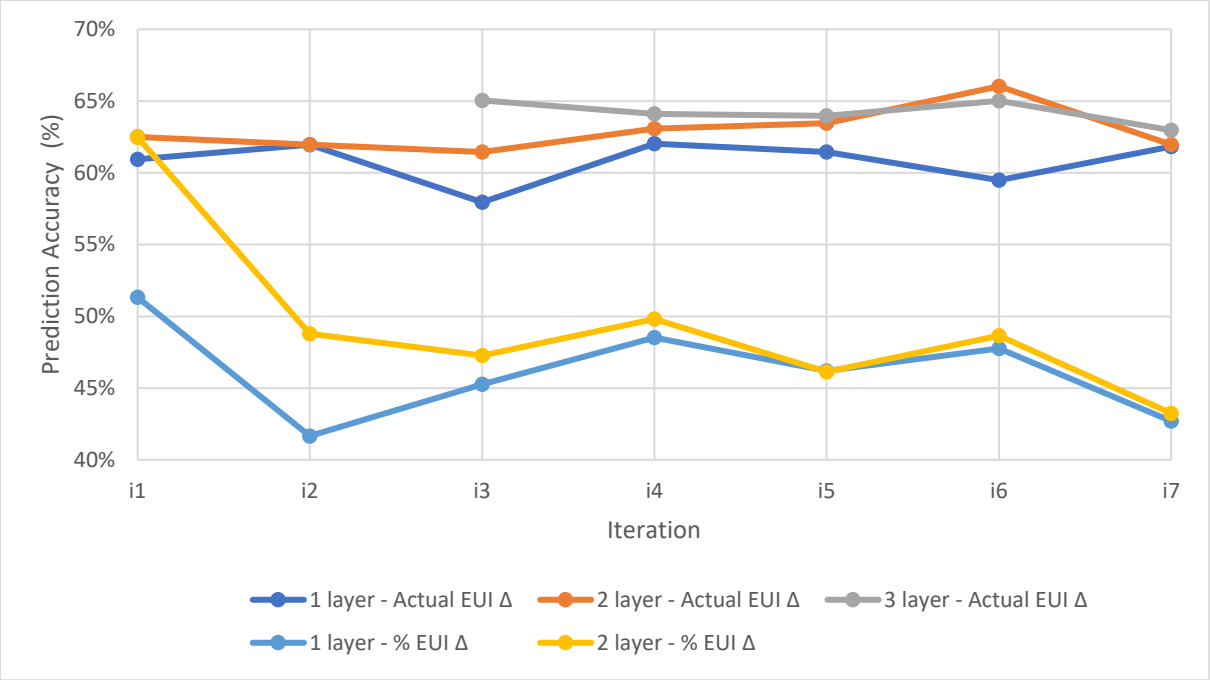


Figure 4-30 Predicting 2010 - 2015 EUI Change ANN - Prediction Accuracy vs Iteration

The best model was from the iteration with two hidden layers which used the set six input variables. The model only achieved an accuracy of 66.03% with eight nodes in each of its two hidden layers. A visualization of this ANN model is presented in Figure 4-31.

Overall the ANN model did not produce strong predictive accuracy rates. The overall classification accuracy rates were so low that the within class accuracy rates were not analysed. It is worth noting that the ANN algorithm models were the most time intensive as the models took many days to complete, if they did converge at all. For the aforementioned reasons ANN is considered the worst performing algorithms explored in this paper and they were not able to predict the class of a new buildings 2015 EUI using its 2010 data.

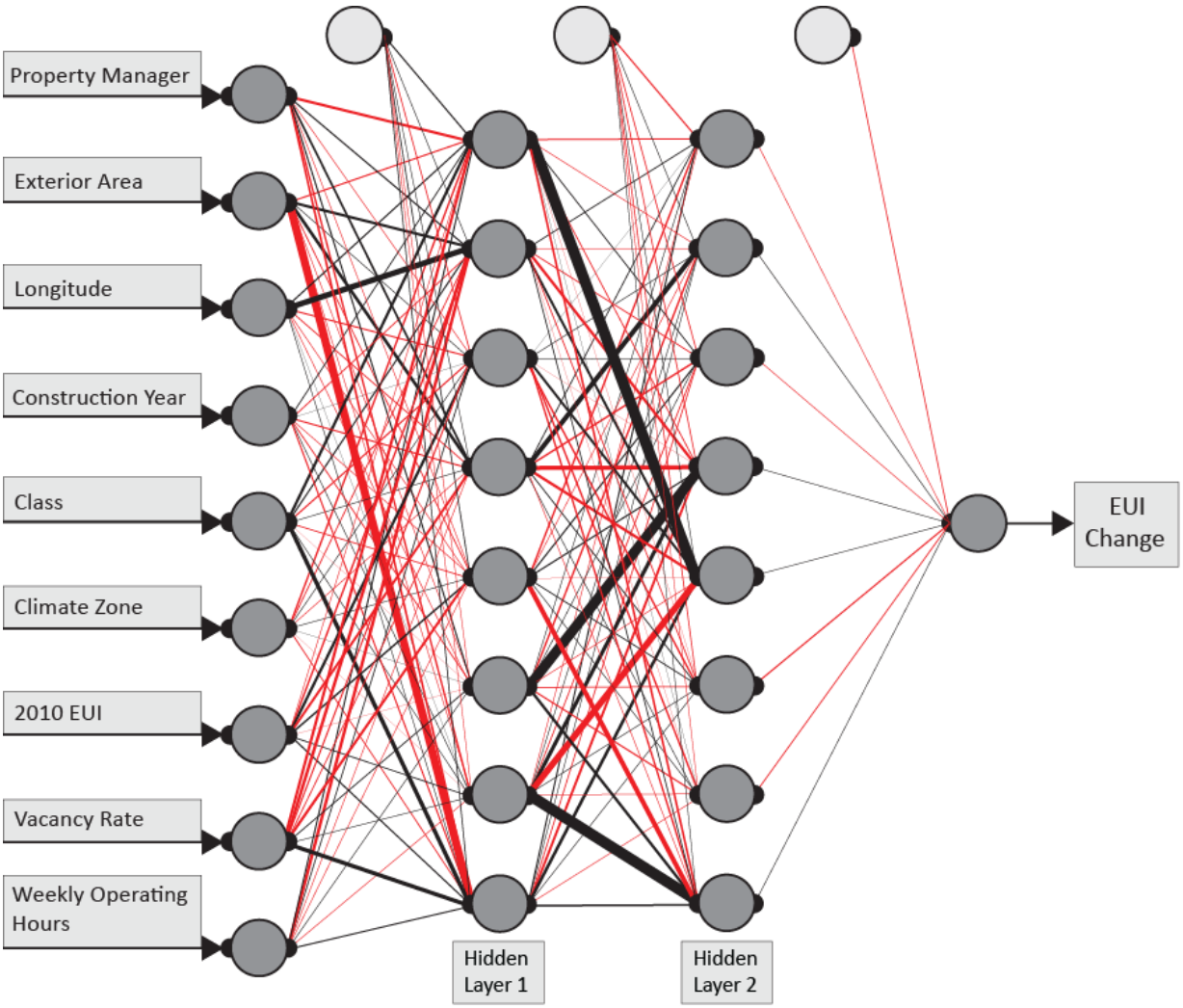


Figure 4-31 Best Fit ANN Model (Iteration 6, 2 hidden layers, 8 nodes in each hidden layer)

5 Discussion

The objectives of the research were to determine the most significant predictors of EUI and to accurately predict a building's energy usage. The following sections discusses the implications of the findings and potential sources of error.

5.1 What are the most significant EUI predictors?

Data visualization, PCA, LDA and MLR were explored to identify trends and important variables which explain energy use intensity. The data visualization was the least helpful, although some useful observations were made. The linear regression plots between the variables of the 2015 data do not highlight any important relationships. When mapping the longitudes and latitudes of each building, it was found that most buildings are located in major cities. The geographical distributions of each class are relatively proportional and a building's location is not indicative of its class.

It was observed that smaller buildings are more likely to be classified as *'Poor'*, this indicated that when being normalized for size, small buildings are punished more aggressively than larger building which have their energy usage distributed over a greater area.

It was noted that Vancouver had the highest average EUI out of all the major cities for each survey year. This indicates that though Vancouver may have lower actual energy usage due to the climate and when the energy use is normalized based on a Toronto climate, the buildings are revealed to actually be performing poorly. The cities' buildings have also been performing worse on average since 2012.

When PCA, LDA, and MLR were used to highlight the most important variables for predicting future energy use, LDA identified the most variables (Table 5-1). Although MLR identified less variables, the results supporting with the LDA findings. PCA was not relatively effective at identifying significant variables, though it did determine that Latitude and Longitude were more useful than Climate Zone.

Table 5-1Most important variables as determined by PCA, LDA and MLR algorithms.

	PCA	LDA	MLR
2010 Actual E&C		X	
2010 Actual Thermal		X	
2010 Actual Total Energy		X	
Occupant Density		X	X
Weekly Operating Hours		X	
Asset Manager		X	
Longitude	X	X	X
Latitude	X	X	X
Construction Year		X	X

Feature reduction is favourable as it informs building energy surveyors which building characteristics are not worth collecting, thereby saving time. This paper would recommend collecting Occupant Density, Weekly Operating Hours, Asset Manager, Longitude, Latitude, and Construction Year when predicting current EUI. When predicting EUI change or future EUI, it would likely be helpful to collect the Actual E & C and Actual Thermal of the start year.

5.2 How accurately can one predict a building's EUI given the available data?

Two types of evaluations were used to access the accuracy of the models. RMSE evaluated the numerical outputs from regression functions. Overall classification and within-class accuracy rates were used to assess classification function. As the results for both were unsuccessful, the regression results were sometimes classified during post processing to explore if the regression outputs perhaps still fell into appropriate class categories. This aim was that the increased flexibility and error margins might result in a successful classification result.

5.2.1 RMSE Evaluation

The RMSE was calculated for the SVM, Decision Tree, and ANN models. Overall, none achieved the set thresholds for a successful model. The best performance was observed with the Decision Tree model when predicting the Actual Total energy on the 2013 dataset. The threshold was set at 0.08 for the particular outcome variable and the model had an RMSE of 0.12 which was only 1.5 times greater. For decision trees the worst RMSE were returned when the outcome variable was Actual E&C and Actual Thermal.

Table 5-2 lowest RMSE values for each regression algorithm

Algorithm	Outcome Variable	RMSE	Success Threshold	Ideal Threshold
SVM	2015 EUI	29.07	5	2
DT	Actual Total	0.12	0.08	0.03
ANN	Percent EUI Change	25	5	2

Although the ANN model returned a dismal RMSE, the SVM model performed the worst. The threshold was set at 5 for the particular outcome variable and the model had an RMSE of 29.07 which was almost six times greater.

Overall, none of the models were determined to be successful when their RMSEs were evaluated.

5.2.2 Classification Accuracy Evaluation

It was expected that the more complex algorithms such as decision trees and ANNs would be able develop more comprehensive models and thereby have better classification accuracy rates. It was interesting to note that the ANN model returned the lowest classification accuracy and it was the simpler LDA algorithm that performed better than any of the others.

The overall LDA accuracy was 85.4%. The SVM model may initially appear as a close second as it was only 0.4% lower than the LDA model. However, the SVM model was only able to output one class prediction, which renders it unhelpful.

LDA is the better model as each within-class accuracy rate were also the highest. Although LDA performed the best, it cannot be used to develop a predictive tool as classification accuracy rates of 51.6% for *'Poor'* and 42.9% for *'Good'* are not reliable enough. It is far below the 90% threshold required.

Table 5-3 Best classification accuracy rates for each classification algorithm

	Overall Classification Accuracy	Class *'Poor'* Accuracy	Class *'Fair'* Accuracy	Class *'Good'* Accuracy
LDA	**85.4**	**51.6%**	**96.1%**	**42.9%**
KNN	71.7%	20.0%	92.9%	23.1%
MLR (Test)	67.8%	14.3%	90.2%	18.2%
SVM	85%	0%	85%	0%
DT	77.00%	28.57%	94.87%	6.67%
ANN	66.03%	*	*	*
*Not recorded due to poor overall classification accuracy				

Overall the worst performing models are SVM and ANN, as discussed earlier, as well as MLR. The MLR model had the second lowest overall, 'Poor' and "Good' Classification accuracy rates.

It is assumed that the LDA model performed best as the data clusters in an adequate and useful manner when the axes are rotated to maximize class separability. The remaining models likely could not manipulate the variables in a way to determine comprehensive rules about class.

The greatest problem was the larger proportion of 'Fair' buildings, which on average made up 71% of the data set. The chosen boundaries result in an average of 11% of buildings classified as *'Poor'* and 17% classified as *'Good'*. The models appeared to over predict the 'Fair' class as it returned the best overall classification accuracy rate. This perhaps resulted in too few 'Poor' and "Good' buildings for the model to successfully train on and develop a basis for selecting the less populated classes. However, moving the boundaries would diminish the usefulness of the results as the goal is to identify the best and worst performing buildings.

In addition, it is likely the models were not successful because the collected variables did little to explain energy usage within a building. For example, thermal energy consumption is highly dependent on the efficiency of the building envelope and the mechanical systems within the buildings, both of which are data not collected in the survey.

Overall, none of the models were determined to be successful when their classification accuracies were evaluated, though LDA shows the most promise for future research.

5.3 Sources of Prediction Error

The following are the potential sources of prediction error which possibly lead to the unsuccessful machine learning models:

- The data appeared to be too complex and diverse. There did not appear to be linear or other simple geometric relationships present within the data, inhibiting the success of the data visualization and MLR models.
- Many of the input variables are interdependent with one another and confounded the models. For example, the normalized EUI is the actual energy with the Gross Floor Area, Enclosed Parking, Occupant Density, Vacancy Rate, Weekly Operating Hours and Closest Major City variables factored out. The Occupant Density is a function of the building's Gross Floor Area.
- The REALPAC normalization Database may have removed a necessary information from the final normalized EUI such as the influence of building characteristics.
- The dataset consisted a high dimensionality with a relatively low number of individuals. The large amount of feature implied strain the model's ability to draw connection and to accurately produce energy use. The dataset likely required more individuals so the model could have a better chance at identifying patterns in the structure of the dataset.
- There was an inability to test more specific predictions on subsets due to limited sample size within some subsets. For example, the ten-fold cross validation subsets only contained about 30 buildings in each, making training and testing more difficult.
- Energy consumption is heavily influenced by the type and quality of a building's constructions. The dataset contained limited information on building construction.

6 Conclusions and Recommendations

The key findings of this research are summarized below:

1. Poor buildings are on average smaller than 300,000 sf.
2. Vancouver buildings on average perform the worst and have not improved overall since 2012. It is suggested for Vancouver to revaluate its energy efficiency in comparison to the performance of buildings with less ideal climates.
3. The LDA model was the best at identifying significant variables and predicting future and current energy usage.
4. The outcome variable (i.e. magnitude of energy use or energy reduction) has little effect on the accuracy of the model.
5. The chosen area input variable has little effect on the accuracy of the model, as it is an interdependent variable are may reduce the accuracy of a model.
6. Evaluating by regression or classification accuracy did not improve the number of successful models.

Further, as noted in the discussion, the initial survey design did not collect the necessary data to predict energy performance and there are numerous interdependent variables that render the analysis difficult. The inclusion of the following information would have significantly improved the usefulness of the dataset and is thus recommended for future studies:

- The types of building materials in wall, roof and foundation assemblies: these affect the R value and therefore the thermal energy consumption.
- The window to wall ratio: this corresponds strongly to heat transfer.
- The building volume to surface area ratio: this influences heat demand on a site and the potential for heat loss.
- The level of thermal bridging at material junctions: contact between highly conductive materials will lower the effective R value of an assembly and increase the thermal energy used by a building.
- The presence of leaks: water penetration is damaging to the building envelope and reduces the effectiveness of the insulation within increasing unwanted heat transfer.
- Completed energy efficiency projects. Presently, it is unknown if drastic changes in EUI are the result of user input error or design changes. Knowledge of these projects, and

their scope, would likely allow the model to better identify buildings that improved and perhaps have less potential for future improvement.

- The efficiency, age, and condition of the mechanical systems: these will have a large impact on electricity use.
- The efficiency and density of the lighting systems: these will have a large impact on electricity use.

To address the above-mentioned limitations, it is recommended that more information is collected regarding the energy systems, sustainability practices, building construction, condition of assemblies, thermal bridge reduction strategies, building surface area to volume ratio and window to wall ratio. Future data collecting initiative are advised mandate a third-party energy audit to be performed on each building entered into the benchmarking survey. It will encourage the gathering of actionable data, reduce user-input errors and produce customized strategies that will better facilitate a building's energy efficiency improvement. It is anticipated that these additional factors will lead to a dataset with detectable relationships and give building managers the proper tools to increase occupant comfort, decrease operating costs and drive the reduction of green house gas emissions.

7 References

City of Toronto, 1986. *Zoning By-Law 438-86.* Toronto: City of Toronto.

Pérez-Lombard, L., Ortiz, J. & Pout, C., 2007. A review on buildings energy consumption information. *Elsevier,* Volume 40, pp. 394-398.

The R Foundation, 2018. *The Comprehensive R Archive Network.* s.l.:s.n.

Abdi, H. & Williams, L. J., 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics [1939-5108],* 2(4), pp. 433-459.

Ahmad, A. S. et al., 2014. A reviewon applications of ANN and SVM for building electrical energy. *Renewable and Sustainable Energy Reviews,* Volume 33, pp. 102-109.

Akinduko, A. A. & Gorban, A. N., 2013. Multiscale principal component analysis. *Journal of Physics,* p. 10.

Allen, M. et al., 2018. Framing and Context. . In: *Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change.* s.l.:[V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P.R. Shukla,A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, T. Waterfield (eds.)].

Ami, Y. & Geman, D., 1997. Shape Quantization and Recognition with Randomized Trees. *Neural Computation,* Volume 9, pp. 1545-1588.

ASHRAE, 2013. Canada climate zones map. In: *ASHRAE Standard-169.*. s.l.:ASHRAE.

Aydinalp, M., Ismet Ugursal, V. & Fung, A. S., 2004. Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. *Applied Energy,* 79(2), pp. 159-178.

Aydinalp, M., Ugursal, V. I. & Fung, A. S., 2002. Modeling of the appliance, lighting, and space cooling energy consumptions in the residential sector using neural networks. *Applied Energy,* Volume 71, pp. 87-110.

Bernoulli, 1773. *Law of Large Numbers.* s.l.:s.n.

BOMA International, 2009. *The Gross Areas of a Building: Methods of Measurement,* Washington, D.C: Building Owners and Managers Association (BOMA) International.

Bonhomme, V. & Claude, J., 2018. *Package 'Momocs',* s.l.: CRAN-R.

Breiman, L., 1994. *Bagging Predictors,* Berkeley: University of California.

Breiman, L., 2001. Random forests. *Machine Learning,* 45(1), pp. 5-32.

Bro, R. & Smilde, A. K., 2014. Analytical Methods - Principal component analysis. *The Royal Society of Chemistry,* 6(9), pp. 2812-2831.

Campbell, C. & Ying, Y., 2011. *Learning with Support Vector Machines.* s.l.:Morgan & Claypool.

Capozzoli, A. et al., 2016. A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres. *Applied Energy,* Volume 171, pp. 592-607.

Charnes, A., Cooper, W. W. & Rhodes, E., 1978. Measuring the efficiency of decision-making units. *European Journal of Operational Research,* 2(6), pp. 429-444.

Cooper, W. W., Charnes, A. & Rhodes, E., 1978. Measuring the efficiency of decision making units. *Add this article to my list,* 2(6), pp. 429-444.

Duda, R., Hart, P. & Stork, D., 2012. *Pattern classification.* s.l.:John Wiley & Sons.

Eichholtz, P., Kok, N. & Yonder, E., 2012. Portfolio greenness and the financial performance of REITs. *Journal of International Money and Finance,* 31(7), pp. 1911-1929.

ENERGY STAR, n.d. *Energy Star Portfolio Manager.* [Online]
Available at: https://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager
[Accessed 14 February 2018].

Environment and Climate Change Canada, 2016. *Canadian Environmental Sustainability Indicators - Greenhouse Gas Emissions,* s.l.: s.n.

Fix, E. & Hodges, J. L., 1951. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*, s.l.: USAF School of Aviation Medicine.

Fritsch, S., Guenther, F., Suling, M. & Mueller, S. M., 2016. *Package 'neuralnet',* s.l.: s.n.

Fuerst, F., 2009. Building momentum: An analysis of investment trends in LEED and Energy Star-certified properties.. *J Retail Leisure Property,* 8(4), pp. 285-297.

Haykin, S., 1998. *Neural Netwrorks: A Comrpehensive Foundation.* Second Edition ed. Hamilton: Pearson Education.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology,* 24(6), p. 417.

Joliffe, I. T., 1986. Principal Component Analysis. In: *Springer Series in Statistics.* New York: Springer-Verlag, pp. 66-136.

Knox, S. W., 2018. Survey of Classification Techniques. In: *Machine Learning : A Concise Introduction.* Hoboken: John Wiley & Sons, Incorporated,, pp. 83-85.

Kok, N., McGraw, M. & Quigley, J., 2011. *The Diffusion of Energy Efficiency in Building.* s.l., s.n., p. 7.

Kotsiantis, S. B., 2013. Decision trees: a recent overview. *Artificial Intelligence Review,* 39(4), pp. 261-283.

Lee, W.-S. & Lee, K.-P., 2009. Benchmarking the performance of building energy management using data envelopment analysis. *Applied Energy,* Volume 171, pp. 592-607.

Meese, G. B., Kok, R., Lewis, M. I. & Wyon, D. P., 1984. A laboratory study of the effects of moderate thermal stress on the performance of factory workers.. *Ergonomics,* 27(1), pp. 19-43.

Meyer, D. et al., 2018. *Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien,* s.l.: CRAN-R.

Microsoft, 2018. *Microsoft Excel.* s.l.:s.n.

Mills, E., 2016. Action-Oriented Energy Benchmarking for Nonresidential Buildings. *Proceedings of the IEEE ,* 104(4), pp. 697-712.

Mucherino, A., Papajorgji, P. J. & Pardalos, P. M., 2009. k-Nearest Neighbor Classification. In: *Data Mining in Agriculture.* New York: Springer, pp. 83-106.

National Energy Board Canada, n.d. *Provincial and Territorial Energy Profiles - Ontario,* s.l.: s.n.

NRCan, 2009. *Survey of Commercial and Institutional Energy Use (SCIEU),* Ottawa: Government of Canada (Natural Resources Canada).

Ontario Ministry of Energy, 2017. *Energy and Water Reporting - Regulation Guidelines,* s.l.: s.n.

Ontario, 2017. *O. Reg. 20/17: REPORTING OF ENERGY CONSUMPTION AND WATER USE.* Ontario: Ontario.

Pérez-Lombard, L., Ortiz, J., Gonzáles, R. & Maestre, I. R., 2009. A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes. *Energy and Buildings,* 41(3), pp. 271-278.

Qiu, Y., Tiwari, A. & Wang, Y., 2015. The diffusion of voluntary green building certification: a spatial approach.. *Energy Efficiency,* 8(3), pp. 449-471.

REALpac, 2015. *REALpac Energy Benchmarking Database Input Guidelines v1.0,* Toronto: REALpac.

REALPAC, 2017. *Energy Benchmarking Report: 2010 to 2015 Results,* Toronto: REALPAC.

REALPAC, 2018. *20 by '15.* [Online]
Available at: www.20x15.ca
[Accessed 2018].

Riedmiller, M. & Braun, H., 1993. *A direct adaptive method for faster backpropagation learning: the RPROP algorithm.* San Francisco, IEEE International Conference on Neural Networks.

Ringnér, M., 2008. What is principal component analysis?. *Nature Biotechnology,* March, 26(3), pp. 303-304.

Ripley, B. D., 1996. *Pattern Recognition and Neural Networks.*, Cambridge: s.n.

Ripley, B. D. & Veneables, W. N., 2002. *Modern Applied Statistics with S.* fourth edition ed. s.l.:Springer.

Ripley, B. & Venables, W., 2015. *Package 'class',* s.l.: CRAN-R.

Robbins, H. & Hsu, P. L., 1947. Complete Convergence and the Law of Large Numbers. *Proceedings of the National Academy of Sciences of the United States of America,* 33(2), pp. 25-31.

RStudio, 2018. *RStudio.* s.l.:s.n.

Shumeli, G., 2017. In: P. E. Central, ed. *Data Mining for Business Analytics: Concepts, Techniques and Applications in R.* s.l.:John Wiley & Sons, pp. 174-176.

Stamp, M., 2017. A reassuring Introduction to Support Vector Machines. In: R. Herbrich & T. Graepel, eds. *Introduction to Machine Learning with Applications in Information Security.* Boca Raton: CRC Press, pp. 96-120.

Stamp, M., 2017. k-Nearest Neighbors. In: R. Herbrich & T. Graepel, eds. *Introduction to Machine Learning with Applications in Information Security.* Boca Raton(Florida): CRC Press, pp. 177-179.

Statistics Canada, 2012. *Canada Year Book, 2012.* Minister of Industry ed. Ottawa: Minister of Industry.

Statistics Canada, 2017. *Table 25-10-0017-01 Electric power generation, annual fuel consumed by electric utility thermal plants,* s.l.: s.n.

StatSoft Inc, 2013. Electronic Statistics Textbook. In: Tulsa: StatSoft.

Steinwart, I. & Christmann, A., 2014. *Support Vector Machines.* s.l.:Springer.

Tharwat, A., Gaber, T., Ibrahim, A. & Aboul Ella, H., 2017. Linear discriminant analysis: A detailed tutorial. *Ai Communications,* 30(2), pp. 169-190.

Therneau, T., Atkinson, B. & Ripley, B., 2018. *Package 'rpart',* s.l.: s.n.

UNEP-SBCI, 2009. *Common Carbon Metric for Measuring Energy Use & Reporting Greenhouse Gas Emissions from Building Operations,* Geneva: United Nations Energy Program - Sustainable Building and Climate Initiative.

USGBC; UL Environment, 2016. *Part A Enhancement: Requirements for Product Category Rules and Environmental Product Declarations,* s.l.: U.S. Green Building Council and UL Environment.

Vaissie, P., Monge, A. & Husson, F., 2016. *Package 'Factoshiny'*. [Online]
Available at: https://cran.r-project.org/web/packages/Factoshiny/Factoshiny.pdf
[Accessed 1 11 2017].

Vaissie, P., Monge, A. & Husson, F., 2018. *PCAshiny.* [Online]
[Accessed 2018].

Warren, F. J. & Lemmen, D. S., 2014. *Canada in a Changing Climate: Sector Perspectives on Impacts and Adaptation.* Ottawa: Government of Canada.

Werbos, P., 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* Cambridge(MA): Harvard University.

Wickham, H., 2016. *Package 'plyr,* s.l.: CRAN-R.

Wright, M., 2017. *A Fast Implementation of Random Forests.* [Online]
Available at: https://cran.r-project.org/web/packages/ranger/ranger.pdf
[Accessed 21 12 2017].

Wright, M. N., Wager, S. & Probst, P., 2018. *Package 'ranger',* s.l.: CRAN-R.

Wu, Y.-c. & Feng, J.-w., 2017. Development and Application of Artificial Neural Network. *Wireless Personal Communications,* pp. 1-12.

Appendix I: REALPAC 20 by ’15 Data Collection

Data for a single building was added by selecting the New Building Data Input tab, followed by Energy Data Input as seen in Figure 7-1 and Figure 7-2.

Figure 7-1 REALPAC 20' by '15-Start Page

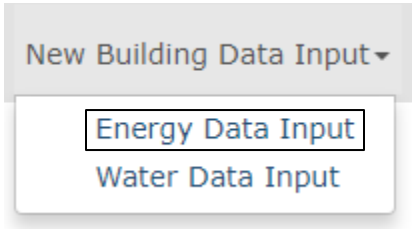


Figure 7-2 REALPAC 20by15 - Energy Data Input Tab

Step 1 involved the entry of contact and building information. The user was permitted to leave most of the fields blank, however the *Building Name* entry was mandatory to proceed to the next step. If the *City*, *Year of Construction*, *Number of Structures* and *Building Owner* fields were not also completed, the website returned a message warning “Your results will be more accurate if you complete the required fields” but users were still allowed to proceed to the next step without entering more information. Refer to Figure 7-3, Figure 7-4 and Figure 7-5

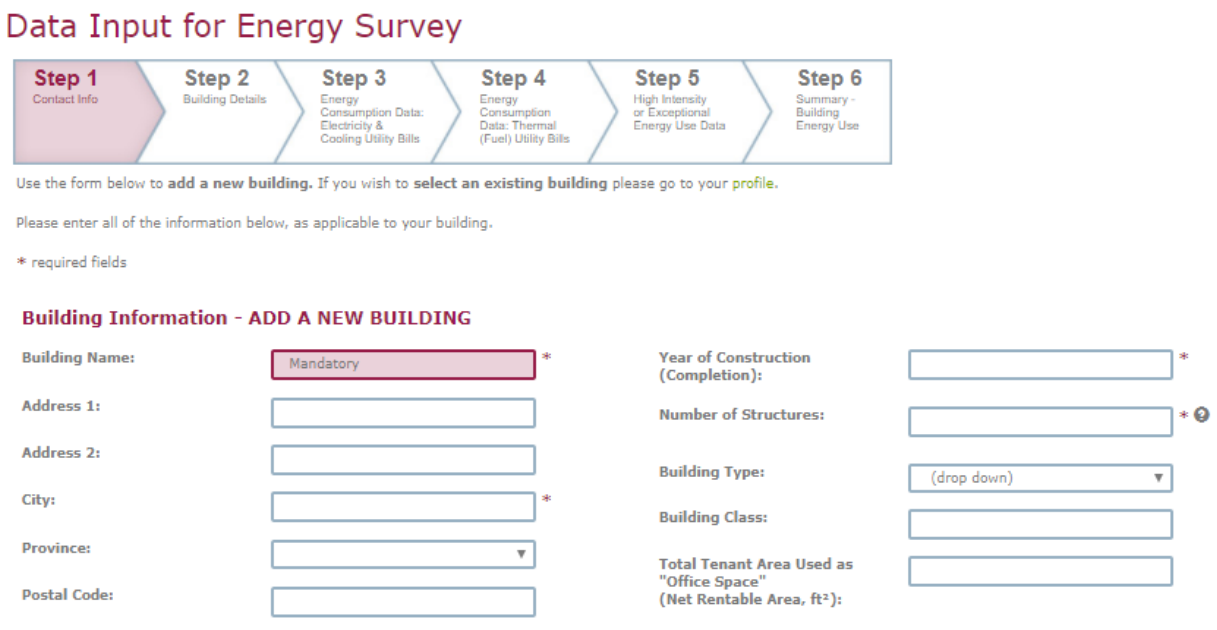


Figure 7-3 REALPAC 20by15 - Step 1 - Building Information

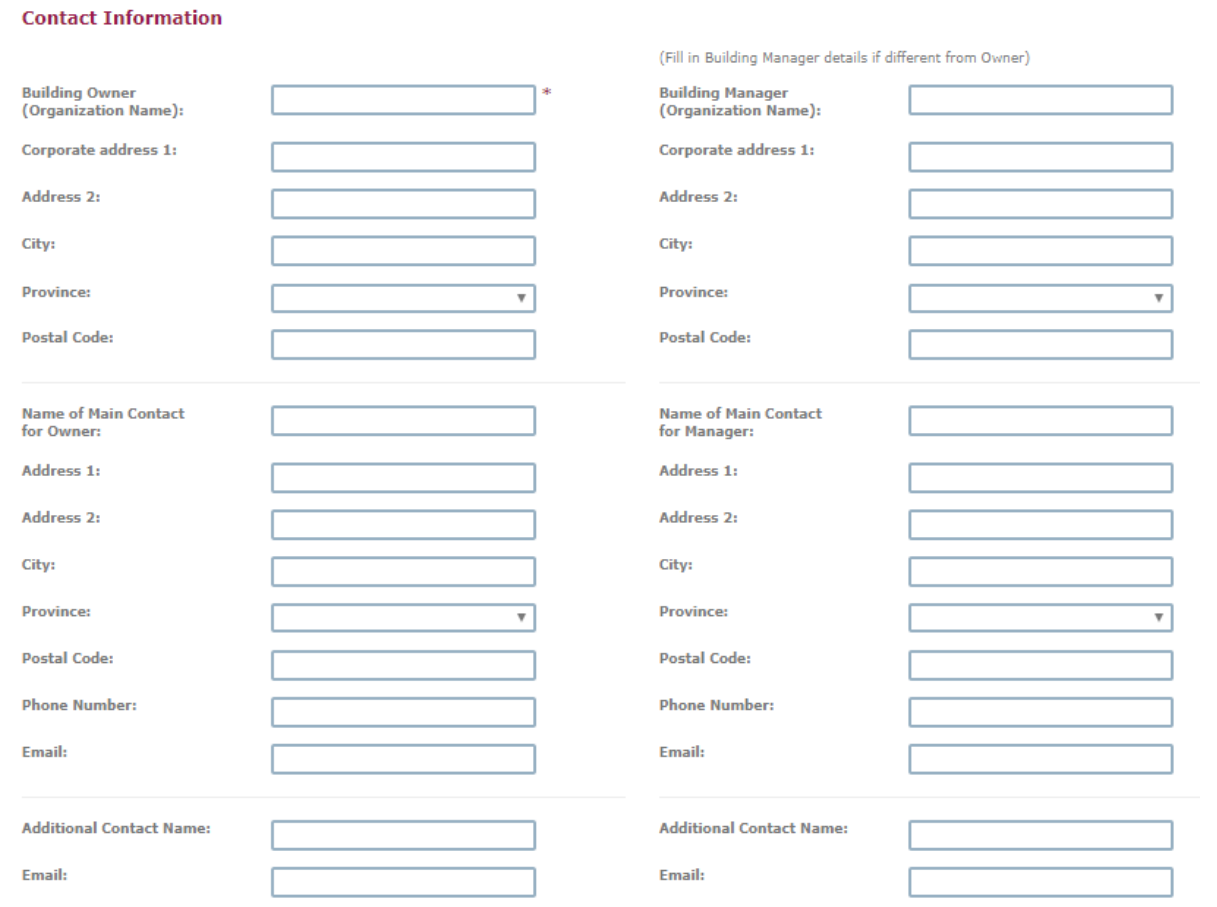


Figure 7-4 REALPAC 20by15 - Step 1 -Contact Information

Figure 7-5 REALPAC 20by15 - Step 1 - Engineer Information

Step 2 involved the entry of building characteristics and environmental certifications. Again, some fields were allowed to be left blank, however *Annual Year of Utility Data* and *Annual Year of Utility Data* were required to be entered to proceed to the next step. If the user does not complete the *Exterior Gross Area, Gross Floor Area, Number of Occupants, Energy Occupant Density, Average Annual Vacancy, Weekly Operating Hours, Energy Specific, Water Specific, LEED Rating System, LEED Certification Achieved* and *BOMA BESt Certification Level Achieved* fields, the website returned a message warning "Your results will be more accurate if you complete/correct the required fields. Click to continue." but users were still allowed to proceed to the next step without entering more information. Refer to Figure 7-6.

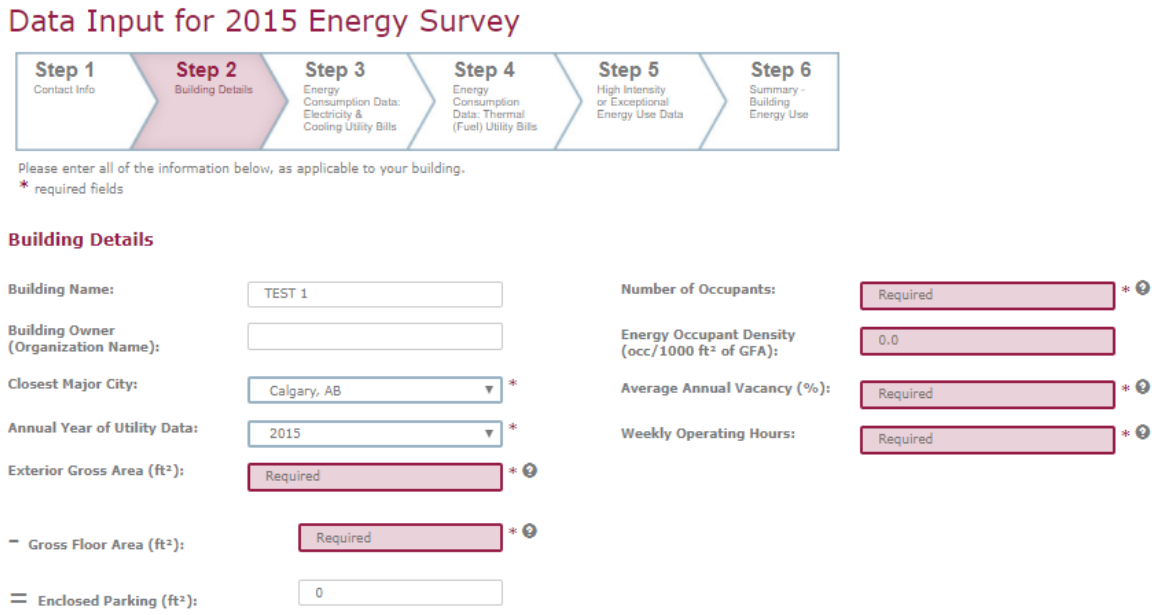


Figure 7-6 REALPAC 20by15 - Step 2 - Building Details

In Step 3, Electricity & Cooling usage data was input. There were no mandatory fields were required to be completed before continuing. The web page prompted for energy type and energy

units, as well as the billing period dates and energy consumption, for each month. A warning message reading “Your results will be more accurate if you complete/correct the required fields. Click to continue without correcting.” would appear if the field were left incomplete but users were still permitted to proceed to the next step.

This step was set up to reflect the layout of a standard utility bill. The dates covered by the billing period were required to be input so the amount of billing days were recorded alongside energy consumption. The total amount of billing days resulted in a number above or below 365, the amount of days in a typical year. To adjust for the difference, the daily consumption rate for December was calculated. The difference was multiplied by the daily consumption rate and added or subtracted from the consumption total as appropriate. See Equation 6 for clarification.

$$365\ day\ normalization = 365 \pm \left[(Billing\ Days - 365) \times \frac{December\ Consumption}{December\ Billing\ Days}\right]$$

Equation 6 - 365 Day Energy Consumption Normalization

If the specified units were not kWh, the webpage converted the total consumption to equivalent kWh. Refer to Figure 7-7 REALPAC 20by15 - Step 3 Energy Consumption DataFigure 7-7 and Figure 7-8.

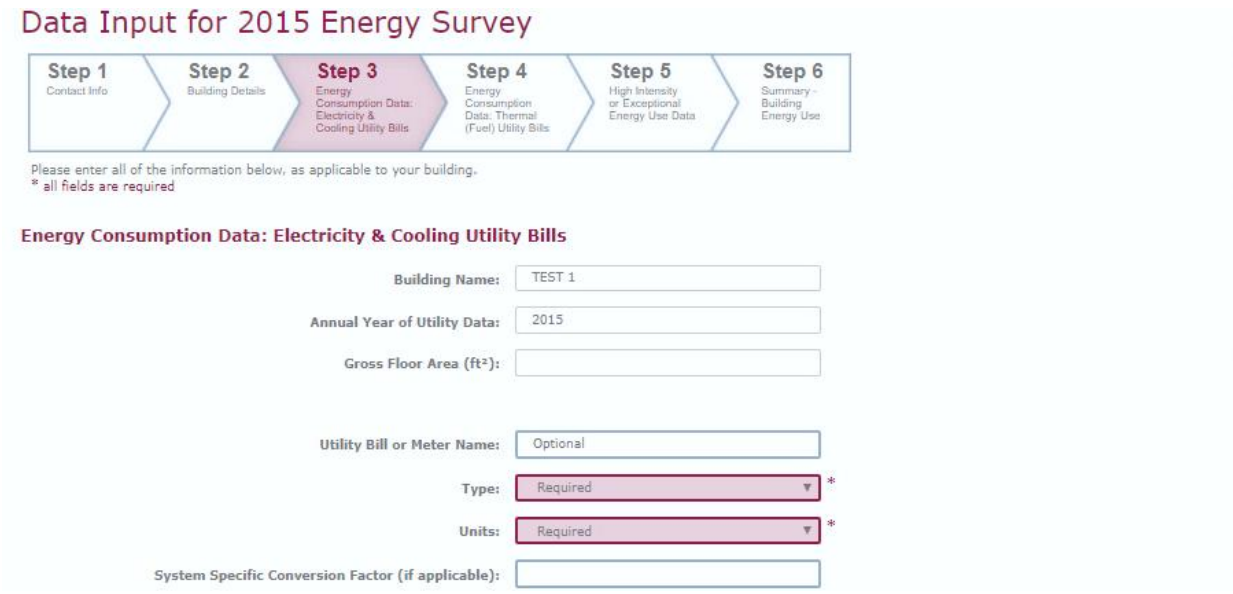


Figure 7-7 REALPAC 20by15 - Step 3 Energy Consumption Data

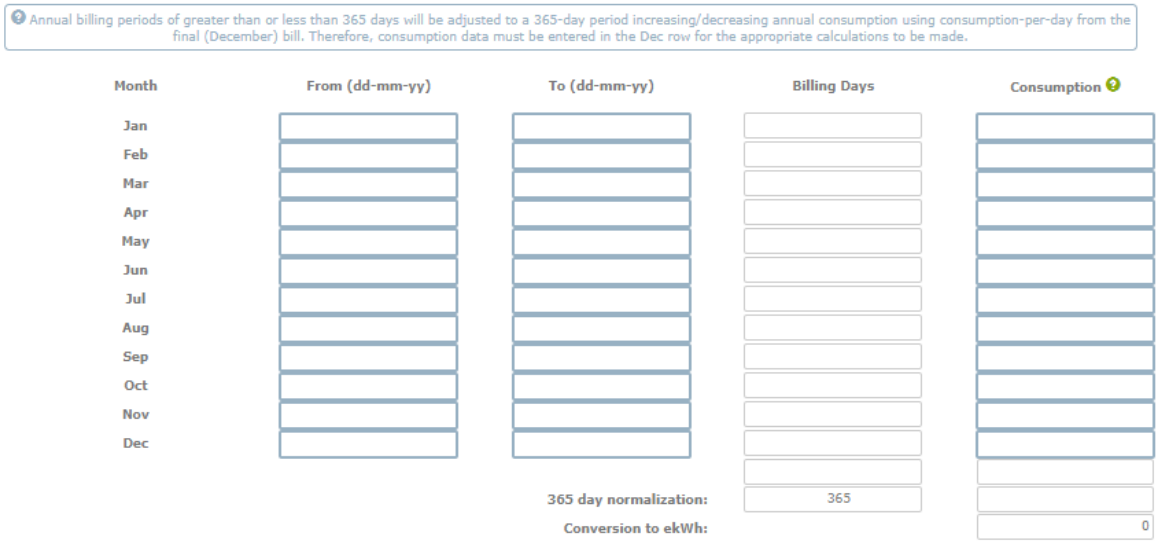


Figure 7-8 REALPAC 20by15 - Step 3 Energy Consumption Data Utility Bills

If there were multiple electricity sources or meters on site, there was the option to add an addition meter, as seen in Figure 7-9. Before proceeding to step 4, the Actual Energy Use (Electricity and Cooling) was normalized for Gross Floor Area.

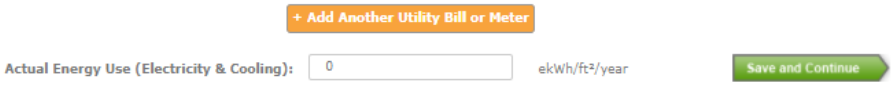


Figure 7-9 REALPAC 20by15 - Step 3 - Area normalized Actual Energy Use (Electricity and Cooling)

Step 4 was not necessary for buildings exclusively heated by electricity. It displayed the same layout as step 4 but was intended for thermal fuel utility data. The following fuel types could be input: district heating, natural gas, fuel oil, propane and steam (onsite). The ensuing units were accepted by the algorithm: cubic meters, cubic feet, contained cubic feet, liters of propane, litres of oil, kilopounds, megapounds, gigajoules, megajoules, British thermal units and therms. The energy types were converted to equivalent kilowatts hours (ekWh) using the conversion factors listed in Table 7-1 Energy Conversion Factors to allow comparison between different types and measures of thermal energy.

Table 7-1 Energy Conversion Factors

	Divide the number of:	By:	To Obtain:
Quantity of heat energy	megajoules (MJ)	3.6*	ekWh
	gigajoules (GJ)	0.0036	ekWh
	British thermal units (Btu)	3412*	ekWh
	Therms (thm)	0.03412	ekWh
	Multiply the number of:	**By:**	**To Obtain:**
Quantity of natural gas	Cubic feet (cf)	0.2931*	ekWh
	100 Cubic feet (Ccf)	29.31	ekWh
	Cubic meters (m3)	10.35*	ekWh
Quantity of propane	Litres (L)	7.028*	ekWh
Quantity of fuel oil (#2)	Litres (L)	10.611*	ekWh
Quantity of steam (onsite generation)	kilo-pounds (klbs)	351.41[£]	ekWh
	Million pounds (Mlbs)	351406.8[£]	ekWh
Quantity of District Heating (steam)	kilo-pounds (klbs (Dist))	468.54[†]	ekWh
	Million pounds (Mlbs (Dist))	468542.4[†]	ekWh
Quantity of District Cooling (chilled water)	ton-hours (ton*h)	0.9[¥]	ekWh
Quantity of Deep Lake Water Cooling	ton-hours (ton*h)	0.285[£]	ekWh

*Conversion factors as per NRCan.

[†]District steam has beenfurther adjusted from onsite generated steam by a factor of 1.333 to account for an average efficiency of 75%. This factor is an average reported efficiency factor of three district energy providers in Toronto.

[¥]Conversion factor represents an average reported kW/ton value of two district cooling providerd in Toronto.

[£]Conversion factors as per Enwave.

Note. Reprinted from REALPAC Energy Benchmarking Database Input Guidelines v1.0 2015 by REALPAC

If a building had certain areas which used energy differently from a typical office, it was referred to as Exceptional Energy Use. If these areas were sub metered, Step 5 allowed the user to input the Exceptional Energy Use type (i.e. Retail, Data Centers, Call Centers, Enclosed Parking and Other), associated area (sf) and Annual Electricity and Annual Natural Gas consumption (kWh). The Database then normalized the buildings energy use so that it is portrayed as if it were only comprised of office space. To do this, the Database subtracted the Exceptional Energy Use from the building's total energy use, multiplied the buildings office area EUI by the Exception use area and added this to the office area energy consumption. The exceptional use normalized EUI was the adjusted energy consumption divided by the total Gross Floor Area. See Figure 7-10 for the webpage illustrating the data collection form for step 5.

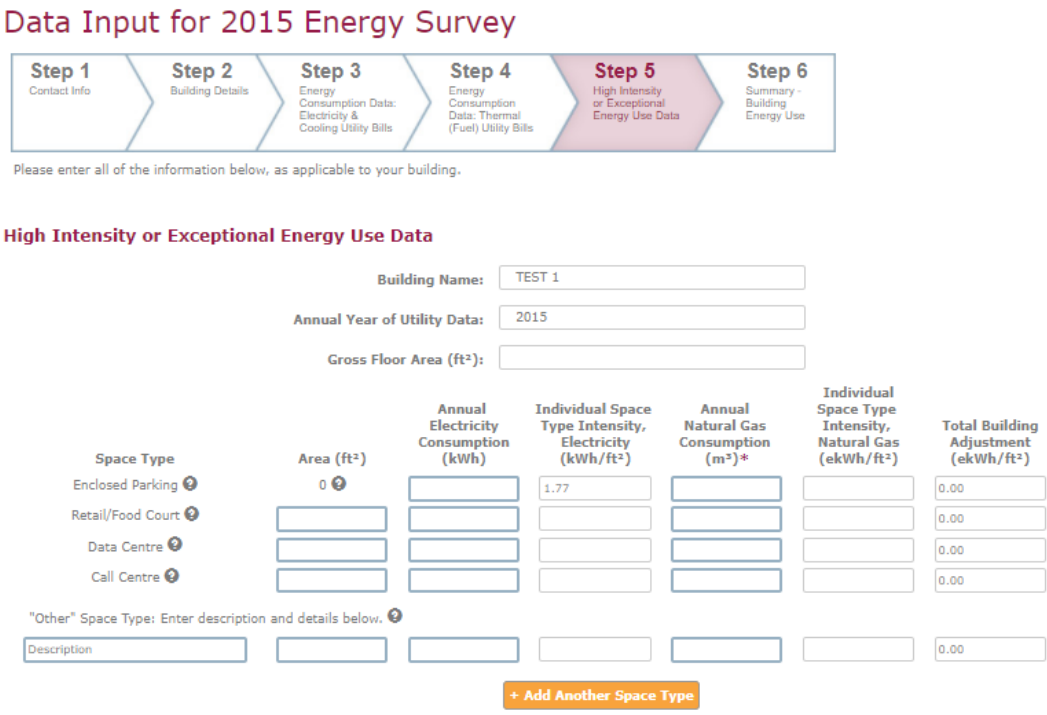


Figure 7-10 REALPAC 20by15 - Step 5 - High Intensity or Exceptional Energy Use Data

In regards to Enclosed Parking, the Database was still able to provide an EUI adjustment without sub metred data by assuming the parking lighting energy consumption was 0.17 W/sf, operating at 24 hours a day for 7 days a week. The parking ventilation was assumed to use 0.15 W/sf, operating at 6 hours a day for 6 days a week.

Step 6 displays the final summary of all the building characteristics, the actual (pre-normalized) building energy use, the building characteristic energy use, weather normalized energy use to base year 2009 and the location normalized energy use which was the considered the final EUI of which all buildings would eventually be compared to each other (See Figure 7-11). If the Database flagged any errors in the data entry, the user was be notified and asked to address them before continuing and submitting the data for a final review by REALPAC. REALPAC contacted building managers to adjust to their data entries if obvious errors, atypical values or missing data were present. If the errors are not resolved the building was discarded from the survey.

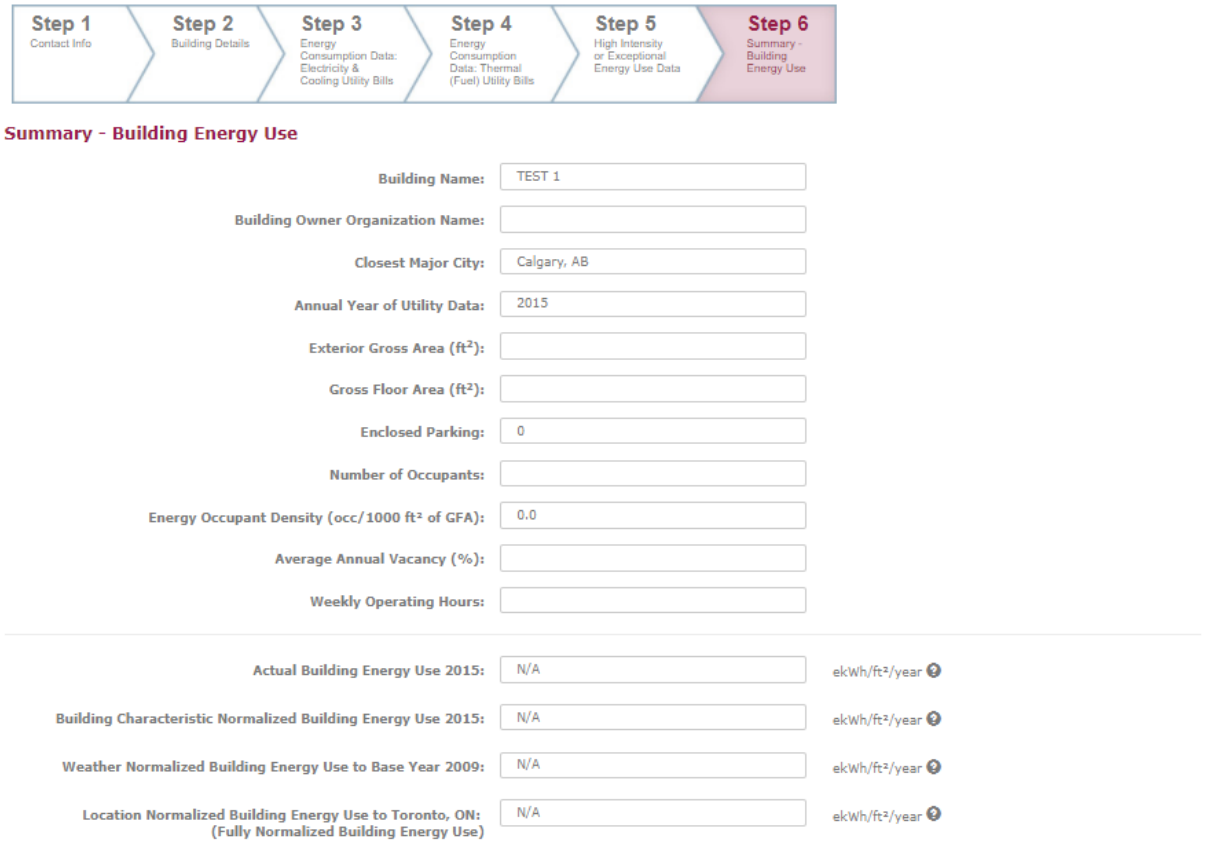
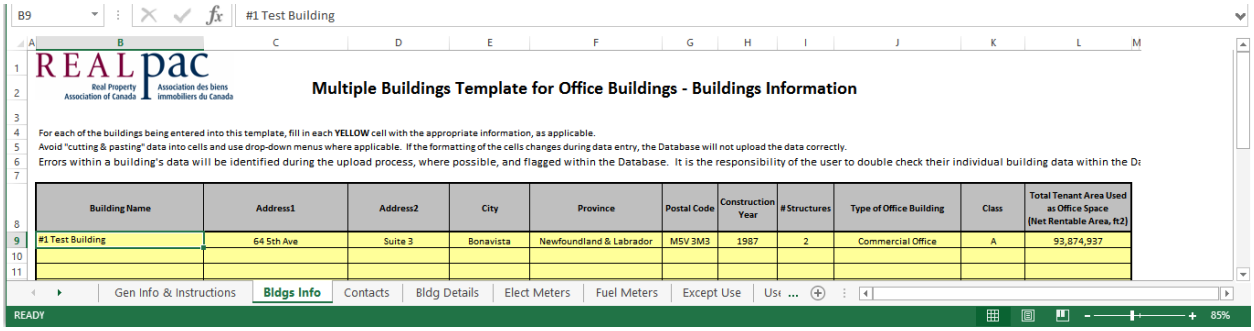


Figure 7-11 REALpac 20by15 Step 6 – Building Data Summary

If a building manager needed to input several buildings at once, they could download the excel workbook titled 'Multiple Buildings Template (MBT) for Office Buildings' from the Multiple Building Input Tab on the 20by15 online portal as seen in Figure 7-12.

The workbook included sheets pertaining to the following: *General Information & Instructions, Buildings Information, Contacts for Buildings, Building Details, Electrical Energy Data Input, Fuel Data Input, High Intensity* or *Exceptional Energy Use Data Input* and *User Emails Associated* with *Buildings*. Refer to Figure 7-12 and Figure 7-13 for the typical layout of the MBT excel work sheets.

The MBT could only be used for up to 40 buildings and another template was needed if a user wished to input more buildings. After the user completed the templates, they could be uploaded into the Database through the 20by15 online portal. The Database flagged the same errors that would be marked if the user had employed the single building input method.



REALpac — Real Property Association of Canada / Association des biens immobiliers du Canada

Multiple Buildings Template for Office Buildings - Buildings Information

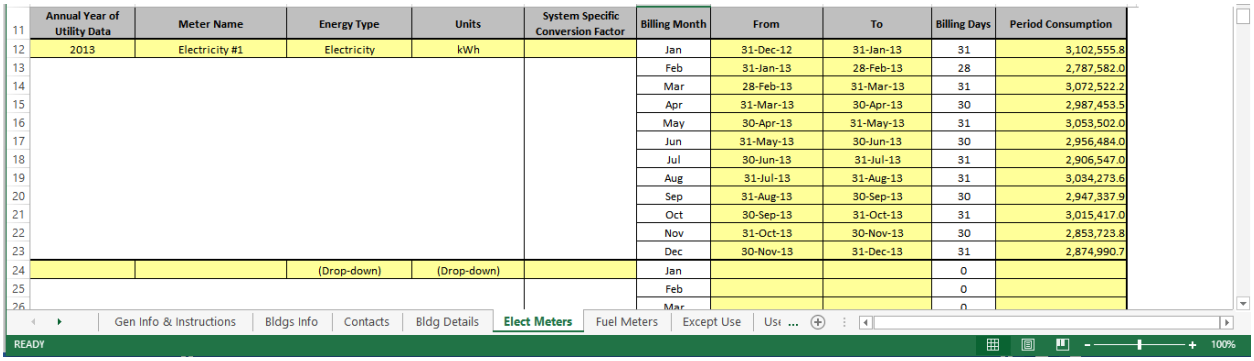
For each of the buildings being entered into this template, fill in each **YELLOW** cell with the appropriate information, as applicable.
Avoid "cutting & pasting" data into cells and use drop-down menus where applicable. If the formatting of the cells changes during data entry, the Database will not upload the data correctly.
Errors within a building's data will be identified during the upload process, where possible, and flagged within the Database. It is the responsibility of the user to double check their individual building data within the Da

Building Name	Address1	Address2	City	Province	Postal Code	Construction Year	# Structures	Type of Office Building	Class	Total Tenant Area Used as Office Space (Net Rentable Area, ft2)
#1 Test Building	64 5th Ave	Suite 3	Bonavista	Newfoundland & Labrador	M5V 3M3	1987	2	Commercial Office	A	93,874,937

Gen Info & Instructions | Bldgs Info | Contacts | Bldg Details | Elect Meters | Fuel Meters | Except Use | Use ...

READY 85%

Figure 7-12 REALpac 20by15 – Multiple Building Template – Buildings Information Sheet



Annual Year of Utility Data	Meter Name	Energy Type	Units	System Specific Conversion Factor	Billing Month	From	To	Billing Days	Period Consumption
2013	Electricity #1	Electricity	kWh		Jan	31-Dec-12	31-Jan-13	31	3,102,555.8
					Feb	31-Jan-13	28-Feb-13	28	2,787,582.0
					Mar	28-Feb-13	31-Mar-13	31	3,072,522.2
					Apr	31-Mar-13	30-Apr-13	30	2,987,453.5
					May	30-Apr-13	31-May-13	31	3,053,502.0
					Jun	31-May-13	30-Jun-13	30	2,956,484.0
					Jul	30-Jun-13	31-Jul-13	31	2,906,547.0
					Aug	31-Jul-13	31-Aug-13	31	3,034,273.6
					Sep	31-Aug-13	30-Sep-13	30	2,947,337.9
					Oct	30-Sep-13	31-Oct-13	31	3,015,417.0
					Nov	31-Oct-13	30-Nov-13	30	2,853,723.8
					Dec	30-Nov-13	31-Dec-13	31	2,874,990.7
		(Drop-down)	(Drop-down)		Jan			0	
					Feb			0	

Gen Info & Instructions | Bldgs Info | Contacts | Bldg Details | Elect Meters | Fuel Meters | Except Use | Use ...

READY 100%

Figure 7-13 REALpac 20by15 – Multiple Building Template – Electricity Energy Data Input Sheet

Appendix II – Acronyms, Abbreviations, Dataset Titles and Definitions

General Acronyms and Abbreviations

ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CaGBC	Canadian Green Building Council
ekWh	Equivalent kilowatt hours
EUI	Energy Use Intensity
kW	kilowatt
kWh	kilowatt-hour
NRCan	Natural Resources Canada
REALPAC	Real Property Association of Canada

Algorithm Acronyms and Abbreviations

ANN	Artificial Neural Networks
DT	Decision Trees
KNN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
SVM	Support Vector Machines

Variable Acronyms and Abbreviations

CZ	Climate Zone
E&C	Electricity and Cooling
EA	Exterior Area
EH	Electrically Heated
EP	Enclosed Parking
EUI	Energy Use Intensity
GFA	Gross Floor Area
L&L	Latitude and Longitude
NRA	Net Rentable Area
OD	Occupant Density
VR	Vacancy Rate
WOH	Weekly Operating Hours

Dataset Titles

2010 Dataset	Cleaned dataset containing only data extracted from the 2010 survey year. This dataset is used to predict a building's 2010 energy use by training the algorithms using non energy variables.
2011 Dataset	Cleaned dataset containing only data extracted from the 2011 survey year. This dataset is used to predict a building's 2011 energy use by training the algorithms using non energy variables.
2012 Dataset	Cleaned dataset containing only data extracted from the 2012 survey year. This dataset is used to predict a building's 2012 energy use by training the algorithms using non energy variables.
2013 Dataset	Cleaned dataset containing only data extracted from the 2013 survey year. This dataset is used to predict a building's 2013 energy use by training the algorithms using non energy variables.
2014 Dataset	Cleaned dataset containing only data extracted from the 2014 survey year. This dataset is used to predict a building's 2014 energy use by training the algorithms using non energy variables.
2015 Dataset	Cleaned dataset containing only data extracted from the 2015 survey year. This dataset is used to predict a building's 2015 energy use by training the algorithms using non energy variables.
2010-2015 Dataset	Cleaned Dataset containing only data with buildings that appear in both 2010 and 2015. Energy use variables from both 2010 and 2015 are included however the remaining variables are solely from the 2010 survey year. This dataset is used to predict a building's 2015 EUI by training the algorithms on the 2010 data.

Definitions

Building	-	"A contiguous and undivided shelter comprising a partially or totally enclosed space, erected by a means of a planned process of forming and combining materials (BOMA International, 2009)."
Climate Zone	-	Defined according to ASHRAE-169 "Canada climate zones map" (ASHRAE, 2013).
ekWh	-	"Equivalent amount of kilowatt hours of energy from different fuel sources (REALpac, 2015)."
Enclose(d)	-	"To separate the inside of a building from the outside, affording protection from the elements appropriate to the occupancy and the local climate. All enclosed space must have a roof (BOMA International, 2009)."
Exterior enclosure	-	"The wall, roof or soffit that constitutes the envelope necessary to enclose a building. The exterior enclosure generally determines the location of the measure line (BOMA International, 2009)."
Exterior gross area (EGA)	-	"The total of all the horizontal floor areas (as viewed on a floor plan) of all floors of a building contained within their measure lines, excluding voids (with the exception of occupant voids), interstitial space, unexcavated space, and crawl space. This includes the exterior gross area of every floor in the building including basements, mechanical floors, mezzanines, penthouses, and structured parking without the removal of column area or other structural elements within the measure line (BOMA International, 2009)."
External circulation	-	"unenclosed pedestrian circulation providing the minimum path for access to tenant suites, egress stairs, elevators, refuge areas, toilets, and building entrances, and required by local building code to meet egress requirements, only when there are no fully enclosed pedestrian corridors serving a floor or portion (such as a wing) thereof (BOMA International, 2009)."
Floor	-	"a normally horizontal, load bearing structure and constituting the bottom level of each story in a building including its associated permanent mezzanine, if any exists (BOMA International, 2009)."

Gross floor area (GFA) - "The exterior gross area of a building minus the enclosed parking area (REALpac, 2015)."

Measure line - "A horizontal line on the outermost structural or architectural surface of the exterior face of the exterior enclosure, or at the exterior edge of any external circulation of a given floor of a building. In determining the measure line, do not consider overhangs, pilasters, columns, awnings, eaves, cornices, sills, ledges, casing, wainscoting, gutters, downspouts, chimneys, signs, shutters, attached electrical or mechanical systems, decorative projections and the like that protrude beyond such surface or edge (BOMA International, 2009)."

Occupant Density - "The number of occupants is defined as the number of workers who are present during the main shift [per 1000sf of GFA] (REALpac, 2015)."

Occupant void - "a floor opening between two or more adjacent floors created by removal of floor area by or for the occupant that would otherwise be included in the exterior gross area or construction gross area of the floor (BOMA International, 2009)."

Parking - "enclosed structured floor area used for transient storage of motor vehicles, including associated circulation and building services (such as exhaust fans and ducts that serve the parking area) but not including the loading docks, sally ports and building service areas such as enclosed auxiliary lobbies used to enter a building from parking areas (BOMA International, 2009)."

Penthouse - "Fully enclosed floor area located on the roof level of a building that occupies less than all of the roof (BOMA International, 2009)."

Restricted headroom - "For occupiable space: Space that does not meet the requirement of the International Building Code section 1208.2 Minimum Ceiling Heights, including subsections thereof. For all other space: Space that has a clear ceiling height of less than 7'-0" (approximately 213 cm) (BOMA International, 2009)."

Soft Landscaping - "An open, unobstructed area that supports the growth of vegetation such as grass, trees, shrubs flowers or other plants and permits infiltration into the

ground. Soft landscaping must allow for the planting of, and sustaining of plant material (City of Toronto, 1986)."

Vacancy Rate - The percentage of total Gross Floor Area that is under-utilized (REALpac, 2015).

Vault space - "Sub-grade space that is enclosed and contiguous to a basement that extends below the adjacent ground plane past the property line, often under a public right-of-way, such as a sidewalk or alley (BOMA International, 2009)."

Void - "absence of a floor within the exterior enclosure of a building in excess of ten square feet (1 square meter) where a floor might otherwise be expected or measured, that is typically in the plane of the upper floors adjacent to multi-story atria or lobbies, light wells, auditoria or the area adjacent to a partial floor, permanent mezzanine or unclassified mezzanine at a given floor level. Only the lowest floor of a multi-story space, such as an atrium, or a well, or lobby, is included in construction gross area and exterior gross area (BOMA International, 2009)."

Weekly Operating Hours - "Number of hours per week that a building (or space within a building) is occupied by at least 75% of the tenant employees averaged over the year under review (REALpac, 2015)".