ASSESSING RESIDENTIAL BUILDING ENERGY SIMULATION ACCURACY THROUGH THE USE OF CLUSTERING

by

Sarah Valovcin

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Mathematics and Statistics).

Golden, Colorado

Date ____________________

Signed: ____________________
Sarah Valovcin

Signed: ____________________
Dr. Amanda Hering
Thesis Advisor

Golden, Colorado

Date ____________________

Signed: ____________________
Dr. Willy Hereman
Professor and Head
Department of Applied Mathematics and Statistics

ABSTRACT

Residential building energy simulation software plays an important role in evaluating the energy consumption and efficiency of residential homes. The goal of this project is to analyze the accuracy of the residential building energy simulation (RBES) process involving the Building Energy Optimization (BEopt^{TM}/ EnergyPlus) program developed at the National Renewable Energy Laboratory by modeling the difference between measured energy consumption and predicted energy consumption of residential archetypes. Residential archetypes are defined by clustering 997 homes from the Field Data Repository with categorical and quantitative variables that describe household characteristics. First, highly dependent variables are removed based on each one's variance inflation factor, and a principal component analysis is applied to the remaining variables to produce independent linear combinations for clustering. Both hierarchical clustering and fuzzy clustering are performed with weighted and unweighted Euclidean distances. The number of clusters is selected using the Xie-Beni, C, and Dunn indices, and all three agree on two clusters. Given that the data is not a random sample, the clusters form based on climate, and residential archetypes that apply to the entire United States cannot be defined. However, homes in cluster 1 are generally "warm climate" homes, and homes in cluster 2 are generally "cold climate" homes. Multiple linear regression models are built and compared for homes within each cluster and for all homes, and their predictive capabilities are assessed using leave-one-out cross validation. The following four responses are modeled: measured electricity, delta electricity, measured natural gas, and delta natural gas, where the delta responses are the difference between the RBES predicted and observed energy consumption. The models built for the measured responses indicate which predictors are influential in modeling energy consumption, and the models built for the delta responses indicate where improvements may be made in the simulation process. The models created for all four responses for the warm climate cluster fit the observed data better than the models created for the cold climate

cluster. Geometry, duct, and ventilation variables appear most frequently in the models for the delta responses, so information related to these variables may point to improvements that may be made in the simulation process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Number of observations . . . . n

Number of clusters . . . . K

Distance (dissimilarity) between two observations . . . . d_{ij}

LIST OF ABBREVIATIONS

Building Energy Optimization software . . . . BEopt™

Error Sum of Squares . . . . ESS

Field Data Repository . . . . FDR

Gower's Coefficient of Similarity . . . . GCS

Mean-squared Error . . . . MSE

Multiple Linear Regression . . . . MLR

National Renewable Energy Laboratory . . . . NREL

Principal Component Analysis . . . . PCA

Residential Building Energy Simulation . . . . RBES

Squared Error . . . . SE

Variance Inflation Factor . . . . VIF

ACKNOWLEDGMENTS

I would like to thank Dr. Amanda Hering, Dr. William Navidi, and Dr. Luis Tenorio for all of their guidance and support over the course of my graduate studies. I would also like to thank Ben Polly and Mike Heaney for guiding me through the realm of building science and the National Renewable Energy Laboratory for supporting this research (APUP UGA-0-41025-33).

INTRODUCTION

In 2010, the residential sector accounted for 23% of the United States' total energy consumption [1]. The most common sources of energy in the U.S. are natural gas and electricity, which are predominantly used in homes for space conditioning, water heating, and to run appliances [2]. Given the high amount of energy usage in American homes (4.4% of world energy consumption), energy simulation programs have become an integral tool in evaluating residential energy consumption.

This project focuses on analyzing the accuracy of the simulation process involving the Building Energy Optimization (BEopt™/ EnergyPlus) program developed at the National Renewable Energy Laboratory (NREL) by modeling the difference between measured energy consumption and predicted energy consumption of "residential archetypes." Residential archetypes are defined as a collection of representative residential homes, which in this case will be limited to single-family, detached homes in the United States. The accuracy of building energy simulation tools is crucial when evaluating energy consumption and efficiency of homes, especially when retrofits are under consideration. Retrofit measures in residential homes are changes made to a home that are intended to decrease the amount of energy consumption, and given the costs associated with these measures, it is imperative to quantify the energy consumption reduction expected due to retrofits before they are implemented.

The building energy simulation tool, BEopt, has been developed by NREL as part of the U.S. Department of Energy Building America program. BEopt is an optimization software program that uses one of two simulation engines: DOE-2.2 or EnergyPlus. Currently, EnergyPlus is the primary building energy simulation engine supported by the Department of Energy, and EnergyPlus simulations typically model the energy consumption of a home at hourly or sub-hourly time steps using a physics-based approach. BEopt's capabilities include simulating annual energy consumption and finding the optimal and near-optimal building designs towards zero net energy, where a zero net energy home refers to a home that produces as much energy annually as it consumes. A user can describe a home in BEopt with predefined options in various categories. Then, BEopt determines the retrofit savings or the optimal path to zero net energy of the described home [3]. There are also some limitations to the BEopt simulation tool. A requirement for simulating a home in BEopt is that it must be three-dimensional with closed geometry. In addition, certain household features, like zonal space conditioning and multiple hot water heaters or duct systems, cannot currently be simulated using this software.

The goal of this project is to assess the accuracy of the residential building energy simulation (RBES) process involving BEopt. Hierarchical and fuzzy clustering are first performed to group homes based on a measure of similarity, where the measure of similarity is based on a comparison of the explanatory variable values recorded for each pair of homes in the database. Then, models are built using a multiple linear regression (MLR) within each cluster with one of four energy consumption variables as the response. MLRs using measured energy consumption as the response are performed to determine which variables influence electricity and natural gas consumption. MLRs using the difference between RBES predicted and observed energy consumption as the response are performed to assess the accuracy of the RBES process. Our hypothesis is that this method will allow us to determine which types of homes are predicted well by the simulation process, and will indicate where improvements may be made for the types of homes with energy consumption that is overpredicted or underpredicted in BEopt. While this project focuses on the RBES process involving BEopt, BEopt is only one type of simulation program, and this method can be applied to any simulation tool for validation. Chapter 1 describes the data used in this study and surveys related prior studies; Chapters 2 through 4 describe and apply hierarchical clustering techniques, variable reduction, and clustering using trimmed data; Chapters 5 through 8 outline the principal component analysis and compare the resulting clusters; Chapter 9 describes the MLR models built; Chapter 10 compares the predictive ability of the MLR models fit using homes within each cluster to the MLR models fit with all homes; and Chapter 11 outlines conclusions and potential future work.

CHAPTER 1
AN OVERVIEW

An overview of the data being used and a description of the simulation process involving BEopt is given below. In addition, prior work on residential archetyping and energy related projects is surveyed, and the data used in these studies is compared and contrasted with the type of data that is used in this project.

1.1 Description of Data

The data used in this study is a subset of homes from NREL's Field Data Repository (FDR), which is a database containing energy audit information for approximately 1,400 homes that have been collected from 9 different energy audit studies. The variables recorded for each home include information on actual energy consumption, location, weather, and household characteristics. The household characteristics cover a broad range of variables that can be put into the following categories: general, geometry, attic and roof, foundation, window, infiltration, duct, ventilation, appliances, space conditioning, and water heating.

BEopt is not capable of simulating the energy consumption of certain unique features of homes present in the FDR, such as multiple hot water heaters or duct systems. Therefore, the homes used in this project are the 997 single-family, detached homes that can be successfully simulated in BEopt using EnergyPlus. While BEopt's user interface allows homes to be described using predefined options, the household characteristics of homes from the FDR are generally not described using these options. For many categories, custom BEopt options are developed for these homes based on the information given in the FDR, like in the case of many of the duct and infiltration variables. Furthermore, there is no information on occupants in the FDR data, so electricity and natural gas consumption are simulated assuming "typical occupant behavior." Typical occupant behavior refers to the standardized occupancy defined using the Building America House Simulation Protocols [4].

In order for the energy usage of homes from the FDR to be simulated in BEopt, the FDR data must first be translated from its original form to an input file that BEopt can read. The general simulation process is shown in the flow chart in Figure 1.1. The homes stored in the FDR database are first processed using NREL translation scripts, which are currently under development. These scripts extract information from the FDR data, calculate any additional information required by BEopt, and create an input file that BEopt can read. This simulation process requires certain assumptions to be made and limits the type and characteristics of homes being used. For example, some specific household characteristics, like the insulation levels for rim joists[1], present in the FDR cannot be simulated in BEopt, so these characteristics are not included in the translation. Furthermore, translating the geometry related variables from the original data set into inputs to be used in BEopt is a difficult process, and the physical representation of a home cannot be precisely mapped. One such example is that if a garage is present, the translation script always maps the garage to the left side of the house and redistributes the window area from the left side to the other three sides.

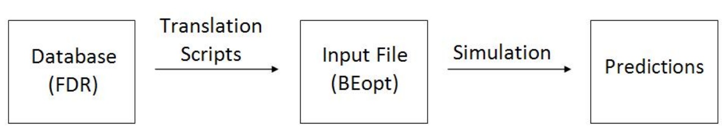


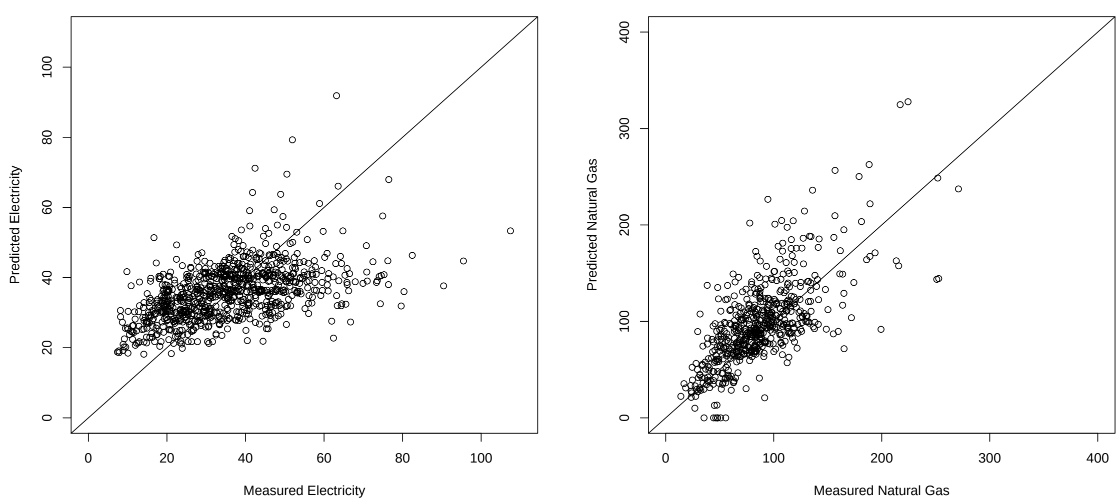
Figure 1.1: Outline of general BEopt simulation process.

Given that there are several steps to the simulation process, the variables used in this study come from multiple points in this process. Specifically, the household characteristics we use come from the FDR data, intermediate steps in the translation scripts, and the BEopt input files. While the majority of the variables are straight from the BEopt input file, many

[1]A rim joist is the final joist in a row of joists that support a floor or ceiling.

of the geometry related variables come from the FDR, and all of the variables that relate to U-values are from the translation scripts. U-values refer to the thermal conductivity of a feature, like above grade walls and windows. For a complete list of variables used, their types, definitions, and summary statistics, see Appendix A.

For the 997 homes used in this project, Figure 1.2 shows the RBES predicted versus measured annual energy consumption[2] for both electricity and natural gas with a reference line of perfect agreement. The observations that fall below the line of perfect agreement are homes for which the simulation process underpredicts energy consumption, and the observations above the line are homes that are overpredicted by the simulation process. The RBES predictions for annual natural gas consumption are generally better than those made for electricity. This is because electricity consumption is more difficult to simulate due to miscellaneous electric loads, such as appliances and electronics, which are driven by occupant behavior.



(a) Electricity

(b) Natural Gas

Figure 1.2: Plot of predicted versus measured annual energy consumption for electricity and natural gas with a reference line of perfect agreement.

[2]Measured energy consumption is weather normalized.

In predicting the energy consumption of a home, there are many sources of potential error. These errors can be attributed to

1. Discrepancies in data collection,

2. Deficiencies in RBES (translation scripts or simulation tool),

3. Inaccuracies in utility processing, and

4. Deviations in occupant behavior [5].

The first potential source of error is in discrepancies in data collection. This refers to errors that can occur during the energy audit of a home, where an energy audit is a formal evaluation of the energy usage and efficiency of a home. An energy audit examines each room in a home and often includes a blower door test and thermographic scan. The blower door test is used to determine the air tightness of a home, and the thermographic scan detects sources of heat loss [6]. Differences between the true values and what is collected during an energy audit can result from inaccurate measurements/observations or incorrect handbook values for material properties. Deficiencies in the RBES process are also a source of potential error. Deficiencies can occur in either the translation scripts or the BEopt simulation tool. It is possible that information is lost when the data is translated into BEopt inputs, or the software tool may use an inappropriate model to simulate the home. It is also possible that the translation scripts or simulation tool contain software coding errors. Another source of potential error is in the way the utility billing data is handled. Energy consumption from utility billing data undergoes weather normalization, which includes disaggregation,[3] in order to be comparable with the predicted consumption, so errors may occur in this process. Errors may also occur due to deviations in occupant behavior, since predicted energy consumption is simulated assuming standardized occupant behavior [5]. While there are several sources of error in predicting the energy consumption of a home, this analysis will only detect errors due deficiencies in the RBES process.

[3] In this context, disaggregation refers to the separation of base load energy use.

1.2 Building Archetypes

Residential building energy simulation software has come to play an important role in evaluating the energy consumption and efficiency of residential homes. These software programs are geared towards providing an analysis of single building designs, and some, like BEopt, are also designed to determine various levels of energy savings efficiency packages that are optimal in cost [7]. Clustering of residential homes is a novel way to approach the subject of accuracy assessment and will offer a framework for residential building archetyping.

In the past, efforts have been made to develop residential building archetypes for different purposes and with different methods. In 2011, Ballarini, et al. developed a methodology to define building archetypes with the goal of improving large-scale building energy modeling. The data used was from the Piedmont Regional Database of Energy Performance Certificates, where a subset of 325 row homes located in the Piedmont region of Italy were used in the analyses. Three different approaches to defining residential archetypes were used. They (1) defined a typical building based on expert choice rules-of-thumb (known as the Example Building); (2) extrapolated a real building with geometrical and thermo-physical characteristics that are similar to the sample mean (Real Building); and (3) identified a typical building as one that is most probable within a group of buildings (Theoretical Building). The Real Building was identified by the intersection of all interquartile ranges, and the Theoretical Building was created with clustering. The clustering analysis was performed using five variables and several distance metrics and hierarchical algorithms. The cophenetic correlation coefficient, which is a measure of the preservation of pair-wise distances between observations that are unmodeled in a given dendrogram, was used to compare distance metrics and the linkage algorithm. These results were used to create a building-types library that had appropriate default values for energy analysis [8].

Famuyibo, et al. (2012) developed archetypes for residential buildings by performing a multiple linear regression and clustering analysis in a case study using data from the Energy Performance Survey of Irish Housing (INSIH) and the Irish National Survey of Housing

Quality (INSHQ). The MLR was performed using all 150 homes from the INSIH to determine energy consumption predictors and rank them, where the full regression model contained 23 variables, some of which were categorical. Then, 9 variables were selected based on the results of the MLR and a ranking of variables based on 17 different studies. The central values of the distributions of these 9 variables were determined, and 81 categories were created by combining these values with one categorical variable of 9 building construction types. Clusters were then defined based on scatter-plots of the pairs of the individual variables and visually identifying clustered values. The values of the clusters were aggregated, and 13 archetypes were created [9].

Parekh, et al. (2005) developed 56 residential archetypes based on the age and climate region/location of 2,930 Canadian homes. Their purpose was to develop libraries of building types for a simple energy analysis software tool. Residential archetypes were designed based on the combination of seven age ranges and eight locations/climate regions. They found that this simple energy analysis was not sufficient for complex homes [10].

Similarly, work has been done to develop archetypes for commercial buildings. Gaitani, et al. (2010) performed K-means clustering and a principal component analysis to determine representative buildings and create energy classes using data from the school authority of Greece. Their data had records for 1,100 schools from all districts of the country and included information on annual consumption for space heating[4]. K-means clustering was performed using only the normalized thermal energy consumption. It was used to rate school buildings, and the number of clusters was determined based on the silhouette value, which measures the similarity of points within the same cluster compared to points in differing clusters. Principal component analysis was then performed on each cluster to determine the "typical" characteristics of school buildings in each cluster. This resulted in forming five balanced energy classes for space heating in schools [11].

These prior works demonstrate the ways in which building archetypes have been formed

[4]The majority of school buildings in Greece use oil for space heating.

in the past for use in building energy analyses, and they are summarized in Table 1.1. Table 1.1 also contains a summary of two additional studies described later in this section. The approach for this project will differ in that residential building archetypes will be defined using two formal types of clustering and will be used to help determine specific areas for improvement of the simulation process. Additionally, the homes are located in the United States, and the number of variables collected for each home is extensive, with both categorical and quantitative variables present.

To illustrate the benefits of clustering, an example given in Sheather (2010) is described. This example uses synthetic data on age and amount of money spent in the past year of customers who have purchased one of two types of tours at a small travel agency (adventure tour or cultural tour) [12]. The regression model using only the age of the customer to model amount spent does not fit the data well, as seen in Figure 1.3. The effect of the observations in the top part of the "X" pattern cancel out the effect of the observations on the bottom. Thus, regressing amount spent on age alone misses the distinct "X" shape. It is clear from Figure 1.3 that there are two separate groups of customers, so a model that captures this pattern is needed.

Figure 1.3: Plot of amount of money spent versus age.

Table 1.1: Summary of prior archetyping work, showing the sample size, the number and types of variables used (categorical, quantitative, and mixed), the type and location of building, the method used for archetyping, and the year of the publication.

First Author	**Year Published**	**Building Type**	**Location of Buildings**	**Sample Size**	**# of Variables**	**Variable Type**	**Method Used**
Ballarini	2011	Residential *(row homes)*	Italy	325	5	Quantitative	Hierarchical Clustering
Famuyibo	2012	Residential *(detached, semi-detached, terraced, purpose built flat, and converted flat)*	Ireland	150	10	Mixed	Clustering based on scatter-plots of variables
Parekh	2005	Residential *(split-level, detached, semi-detached, row homes, and walk-ups)*	Canada	2930	2	Categorical	Combinations of categorical variables
Gaitani	2010	Commercial *(schools)*	Greece	1100	1	Quantitative	K-means clustering
Santamouris	2007	Commercial *(schools)*	Greece	320	1	Quantitative	Intelligent fuzzy clustering
Petcharat	2012	Commercial *(hotel and hospital, department store, and office and school)*	Thailand	8770	1	Quantitative	Expectation-Maximization algorithm
This Project	NA	Residential *(detached)*	United States	997	49	Mixed	Fuzzy and Ward's hierarchical clustering on PCs of variables

In this case, the groupings are based on a simple categorical variable: whether a customer bought an adventure tour or a cultural tour. By breaking the data into adventure tour buyers and cultural tour buyers, the model now fits the data more accurately, shown in Figure 1.4. With high dimensional data, it is impossible to visualize groupings as shown in this example. Clustering offers a more sophisticated way to break homes into groups, which are based on a defined measure of similarity or distance. This can lead to more meaningful groupings for data with large dimensionality. This example will be revisited in Chapter 2 to demonstrate how the observations are grouped automatically based on age and type of tour purchased with select hierarchical clustering methods.

Figure 1.4: Plot of amount of money spent versus age broken into adventure tour buyers and cultural tour buyers.

In two prior studies, grouping the observations was shown to improve the subsequent analysis. Santamouris, et al. (2007) applied intelligent fuzzy clustering to classify the energy performance of approximately 320 school buildings in Greece. With fuzzy clustering, each observation belongs to a cluster with a certain probability that is defined by a membership grade, where the membership grade is an indication of the strength of the association between the observation and a given cluster. They found that classifying energy performance using fuzzy clustering was more advantageous than frequency rating procedures [13]. Moreover, Petcharat et al. (2012) used clustering to estimate potential energy savings in building lighting systems using both simulated and collected data. In this study, the author found that clustering gave more accurate energy savings predictions than averaging across all buildings [14]. These studies are also summarized in Table 1.1.

1.3 Software Accuracy Research

Previously, several approaches have been taken to assess the accuracy of residential building energy simulation programs. Lomas, et al. (1997) performed an empirical exercise to validate 25 dynamic thermal simulation programs (DSPs) of buildings. These programs created heating energy demand and air temperature predictions for three single-zone test rooms located in the United Kingdom, which either had no windows or a single-glazed or double-glazed south-facing window. The DSPs that were likely to have significant internal errors were differentiated from the DSPs that performed well by comparing the predictions to the measured, where experimental uncertainty was taken into account [15].

Tronchin, et al. (2008) calculated energy consumption using three different models to quantify and compare the difference between the calculated and observed energy consumption. The models were based on average observed energy consumption using three years of utility bills, a simulation using DesignBuilder with EnergyPlus, and a simulation using software called BestClass for one single-family home located in central Italy. Intervals of confidence of the models were calculated, which were found experimentally by comparing the results with dynamic numerical methods. It was found that since the intervals of confidence were dependent on the different models, the results were not comparable [16].

Work has also been done to analyze the accuracy of predicted energy savings due to implementing retrofits. Brown (2012) analyzed the relationship between the forecasted energy savings of the EnergyPro simulation tool and the actual energy savings when retrofits were implemented on 51 California homes. The energy simulation tool significantly overestimated relative and absolute net savings due to retrofits for these homes, based on the mean difference between forecasted savings and utility bill savings for one year [17].

Most similar to this project, Roberts, et al. (2012) assessed the accuracy of the Home Energy Scoring Tool (HEST) using 859 homes from the FDR as part of NREL's goal to assess and improve residential building energy simulation accuracy. The HEST simulation tool requires fewer inputs than BEopt and calculates a score that encompasses a range of values for energy consumption. It was found that for these homes, HEST underpredicted electricity and natural gas consumption in over half of the homes. MLR models were also built to determine which factors were significant in contributing to the difference between predicted and measured energy consumption. Then, a Monte Carlo uncertainty analysis was performed, which was based on the range of energy use and a distribution of possible occupant behaviors. The Monte Carlo uncertainty analysis was used to estimate the portion of the total observed variability of the difference between the HEST predicted and measured score that can be explained by variability in occupant behavior [18].

These prior works demonstrate some of the ways that the accuracy of building energy simulation tools has been assessed, and they are summarized in Table 1.2. The approach for this project differs in that first a clustering analysis is performed to group similar types of homes based on multiple characteristics. Then, models will be built within each cluster using a multiple linear regression with one of four energy consumption variables as the response. MLRs using measured energy consumption as the response are performed to determine which variables influence electricity and natural gas consumption, and MLRs using the difference between predicted and measured energy consumption as the response are performed to assess the accuracy of the RBES process. This project differs from the analysis of HEST by Roberts, et al. (2012) in several ways. While the homes in the FDR were used for accuracy assessment of the HEST simulation tool, the data set we use in this project contains 138 additional homes and will be used to assess the RBES process using BEopt. The HEST simulation tool is also very different than BEopt; it requires significantly fewer input variables, most of which are categorical. Furthermore, we include variables from the translation scripts in addition to the required BEopt inputs, while only required HEST inputs were used in the accuracy assessment of the HEST simulation tool. Although MLR analyses are used in both projects, we build additional MLR models for the groups of homes that are determined by clustering to pinpoint areas of improvement in the RBES process that are specific to the homes in each group.

Table 1.2: Summary of prior software accuracy assessment work, showing the software tool, the type and location of building, the method used for assessing the accuracy, and the year of the publication.

First Author	**Year Published**	**Software Tool**	**Building Type**	**Location of Buildings**	**Sample Size**	**Method Used**
Lomas	1997	25 DSPs	Single-zone test room	United Kingdom	3	Empirical validation of predictions
Tronchin	2008	DesignBuilder BestClass	Residential *(detached)*	Italy	1	Intervals of confidence
Brown	NA	EnergyPro	Residential	California	51	Summary statistics and scatter plots
Roberts	2012	HEST	Residential *(detached)*	United States	859	Multiple linear regressions and Monte Carlo uncertainty analysis
This Project	NA	BEopt	Residential *(detached)*	United States	997	Multiple linear regressions of all homes and each cluster

CHAPTER 2

PRELIMINARY CLUSTERING ANALYSIS

Cluster analysis is an exploratory technique used to find "natural" groupings in multidimensional data. The groupings are determined based on a calculated similarity or distance measure computed for each pairwise comparison of individuals, where no assumptions are made about the number of groups or the group structure [19]. The measure of similarity, hierarchical clustering algorithms, and a method for choosing the number of clusters are described below, along with a demonstration using the vacation travel example presented in Section 1.2. In addition, the results of an initial clustering analysis using 93 variables measured on the FDR homes are given.

2.1 Gower's Coefficient of Similarity

When determining the appropriate measure of similarity, several things must be considered: the type of variables, the scales of measurement, and knowledge of the subject matter [19]. Given that the residential building data is a mixture of categorical and quantitative variables and that many have very different units of measurement, Gower's coefficient of similarity (1971) will be used and is defined as follows:

$$s_{ij} = \frac{\sum_{p=1}^{P} w_{ijp} s_{ijp}}{\sum_{p=1}^{P} w_{ijp}}, \tag{2.1}$$

where s_{ijp} is the similarity between the i^{th} and j^{th} observation for variable p, and w_{ijp} is the weight assigned to the p^{th} variable. If x_{ip} or x_{jp} is missing, then $s_{ijp} = \text{NA}$, and $w_{ijp} = 0$. The computation of the similarity, s_{ijp}, is dependent on the type of variable p. If variable p is categorical, then

$$s_{ijp} = \begin{cases} 1, & \text{if } x_{ip} = x_{jp}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

If variable p is quantitative, then

$$s_{ijp} = 1 - \frac{|x_{ip} - x_{jp}|}{r_p}, \tag{2.3}$$

where r_p is the range of observations for the p^{th} variable. The corresponding distance (dissimilarity) is $d_{ij} = \sqrt{1 - s_{ij}}$ [20].

2.2 Hierarchical Clustering

Agglomerative hierarchical clustering, also known as the "bottom-up" method, begins with each observation as its own cluster, and at each step clusters are merged until only one cluster remains. There are several algorithms that may be used for agglomerative hierarchical clustering to combine two clusters. These algorithms are based on different criteria, such as minimum or average distances between the observations in two clusters [21]. The travel example introduced in Section 1.2 will be evaluated using the clustering algorithms described below.

2.2.1 Single-Linkage

Single-linkage is a common agglomerative hierarchical clustering algorithm that is based on the distance (or dissimilarity) between the two closest observations in the two clusters. Mathematically, the distance between cluster I and cluster J can be expressed as

$$D(I, J) = \min_{i \in I,\, j \in J} d_{ij} \ , \tag{2.4}$$

where d_{ij} is the dissimilarity between observations i and j. The two clusters with the smallest distance are then merged. This method of hierarchical clustering is also referred to as the "nearest-neighbor" approach [21].

Revisiting the travel data example from Section 1.2, Figure 2.1 shows the resulting dendrogram (hierarchical tree diagram) using single-linkage to cluster the travel data with Gower's coefficient as the dissimilarity measure. A dendrogram is a visual display of the clustering, where the branches represent the clusters, and the height is the value of the distance, $D(I, J)$, when the two clusters are joined. The dendrogram for this data shows two distinct clusters, which are the adventure and cultural tour buyer groups.

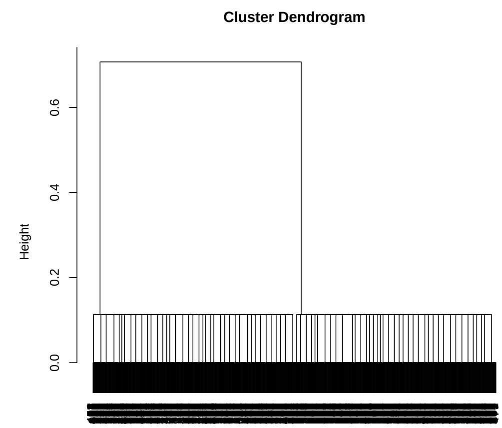


Figure 2.1: Dendrogram of travel data using hierarchical clustering with single-linkage.

2.2.2 Average-Linkage

Another common agglomerative hierarchical clustering algorithm is average-linkage, which is based on the mean distance between observations in two clusters. Mathematically, the distance between cluster I and cluster J can be expressed as

$$D(I, J) = \frac{1}{n_i n_j} \sum_{i \in I} \sum_{j \in J} d_{ij} \ , \tag{2.5}$$

where d_{ij} is the dissimilarity between observations i and j, and n_i and n_j are the number of observations in cluster I and cluster J, respectively. The two clusters with the smallest distance are then merged [21].

Figure 2.2 shows the resulting dendrogram using average-linkage to cluster the travel data with Gower's coefficient as the dissimilarity measure. The dendrogram for this method looks very different compared to the dendrogram using single-linkage; however, the groupings are the same if two clusters are chosen.

2.2.3 Ward's Method

Ward's method is a type of agglomerative hierarchical clustering that is based on minimizing the within cluster variance, which is calculated using the error sum of squares [22].

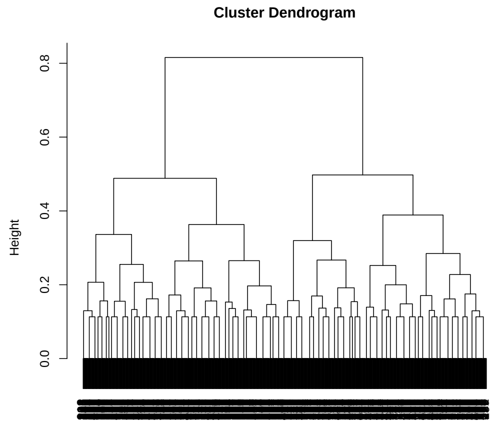


Figure 2.2: Dendrogram of travel data using hierarchical clustering with average-linkage.

The error sum of squares (ESS) is given by

$$ESS = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\mathbf{x_{jk}} - \bar{\mathbf{x}}_\mathbf{k})'(\mathbf{x_{jk}} - \bar{\mathbf{x}}_\mathbf{k}), \tag{2.6}$$

where n_k is the number of observations in the k^{th} cluster, $\mathbf{x_{jk}}$ is the j^{th} observation in cluster k, $\bar{\mathbf{x}}_\mathbf{k}$ is the k^{th} cluster mean, and K is the total number of clusters. As with all agglomerative clustering methods, each cluster is initially made up of one observation. At each step, the ESS is computed for the union of every possible pair of clusters, and the pair with the smallest increase in ESS is joined [19].

Since agglomerative clustering methods use a distance matrix, Lance and Williams (1966, 1967) introduced a recurrence formula to update distances when clusters are merged. This formula gives the updated distance between cluster a and the newly formed cluster (bc), which is created by combining clusters b and c.

Let d_{bc} be the distance between clusters b and c, and let $d_{a(bc)}$ be the updated distance of cluster a to the newly formed cluster (bc). Then, the recurrence formula may be written as

$$d_{a(bc)} = \alpha_b d_{ab} + \alpha_c d_{ac} + \beta d_{bc} + \gamma |d_{ab} - d_{ac}|, \tag{2.7}$$

where α, β, and γ are parameters defined by the linkage process [23]. Milligan (1979) determined that these parameters for Ward's method are

$$\alpha_b = \frac{n_b + n_a}{n_b + n_c + n_a}, \; \alpha_c = \frac{n_c + n_a}{n_b + n_c + n_a}, \; \beta = \frac{-n_a}{n_b + n_c + n_a}, \text{ and } \gamma = 0,$$

where n_i refers to the number of observations in the i^{th} cluster. Thus, the recurrence formula for Ward's method can be written as

$$d_{a(bc)} = \left(\frac{n_b + n_a}{n_b + n_c + n_a}\right) d_{ab} + \left(\frac{n_c + n_a}{n_b + n_c + n_a}\right) d_{ac} - \left(\frac{n_a}{n_b + n_c + n_a}\right) d_{bc}, \tag{2.8}$$

and at each step, the pair of clusters that is merged are the two clusters with the smallest distance [24]. Given the nature of this technique, Ward's method typically forms clusters that are roughly elliptical in shape. In the resulting dendrograms, the "height" is the value of within cluster variance (distance) at which the clusters are combined [19].

Figure 2.3 shows the resulting dendrogram using Ward's method to cluster the travel data with Gower's coefficient as the measure of dissimilarity. The dendrogram for this method shows well-formed clusters, and when two clusters are chosen, they match the adventure and cultural tour buyer groups.

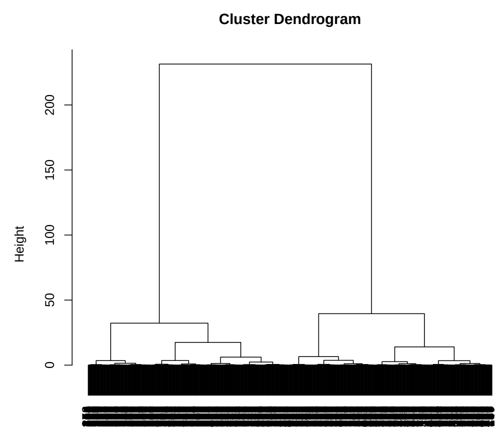


Figure 2.3: Dendrogram of travel data using hierarchical clustering with Ward's method.

Given that Ward's method typically lends itself to well-formed clusters, where well-formed clusters are clusters that are roughly the same size and are not made up of a single observation until they are merged in the final steps, this method will be used in the preliminary clustering analysis.

2.3 Choosing the Number of Clusters

There are many methods to determine the optimal number of clusters, K, and Ward's method lends itself to an intuitive idea. Since Ward's method is based on the error sum of squares, the amount of increase in the error sum of squares when two clusters merge can give a good indication as to the optimal number of clusters. This is known as the merging cost. The idea is to keep reducing K until the merging cost jumps, and choose the K right before the jump [25].

2.4 The Additive Constant Problem

Since Ward's method is based on the error sum of squares, Euclidean distances are needed for this method of hierarchical clustering. The Euclidean distance between two $p \times 1$ vectors, $\mathbf{x}$ and $\mathbf{y}$, is given by

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \tag{2.9}$$

and satisfies the following properties ($\mathbf{z}$ is an intermediate $p \times 1$ vector) [19]:

1. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$,

2. $d(\mathbf{x}, \mathbf{y}) > 0$ if $\mathbf{x} \neq \mathbf{y}$,

3. $d(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} = \mathbf{y}$, and

4. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})$ (triangle inequality).

Unfortunately, Gower's coefficient of similarity gives corresponding distances that do not always satisfy the triangle inequality. Therefore, we face what is known as the additive constant problem, which appears in multidimensional scaling. The problem is set up as follows. Define W_d as

$$W_d = -\frac{1}{2} ADA,$$

with $D = \{d_{ij}^2\}$ and $A = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$, where I is the identity, n is the number of observations, $\mathbf{1}$ is an n-dimensional vector of ones, and d_{ij} is the dissimilarity measure between observations i and j. Since d_{ij} does not have a Euclidean representation, W_d is not positive semi-definite.

Let d_{ij}^* be a measure of dissimilarity defined on a finite set that has no Euclidean representation. The solution to the additive constant problem finds the smallest number m^* such that the dissimilarity measure d_{ij}^* has a Euclidean representation for all $m \geq m^*$, where d_{ij}^* is given by

$$d_{ij}^* = \begin{cases} \sqrt{d_{ij}^2 + m}, & \text{if } i \neq j, \\ 0, & \text{if } i = j. \end{cases} \tag{2.10}$$

The smallest number m^* such that d^* is Euclidean for all $m \geq m^*$ is $-2\lambda_n$, where λ_n is the smallest eigenvalue of W_d. For more detail on the analytical solution of this problem, see Cailliez's "The Analytical Solution of the Additive Constant Problem" [26]. Using the solution to this problem, the matrix of Gower's distances can be transformed into a Euclidean representation to be used in Ward's method.

2.5 Initial Results

First, 93 variables are used to perform an initial clustering analysis. Later, we apply a variable reduction technique and principal component analysis to select a subset of these 93 variables from which we then create independent linear combinations.[5] But first, we examine the larger original set of 107 variables of household characteristics that can be placed into one of the following categories:

1. General,
2. Location,
3. Geometry,
4. Attic and Roof,
5. Foundation,
6. Window,
7. Infiltration, Duct, and Ventilation,
8. Appliances,
9. Space Conditioning, and
10. Water Heating.

[5] For a complete list of variables included in the clustering analysis, see Appendix A.

Before this analysis is performed, variables that are perfect linear combinations of other included variables are removed. Furthermore, if the absolute value of the correlation between two variables is larger than 0.90, one of these variables is omitted from the analysis. This results in excluding 14 variables, leaving 93 variables to be used in the analysis.

First, Gower's coefficient of similarity is calculated using equal weights for all variables. The corresponding distance matrix is not Euclidean, so the smallest number such that Gower's distance has a Euclidean representation must be found. This number is known as the Cailliez constant and is equal to 0.46206 for this analysis. This constant is applied to Gower's distance, and the transformed distance matrix is used for Ward's Method of hierarchical clustering. The resulting dendrogram is shown in Figure 2.4, and it can be seen that the clusters produced with this method are well-formed.

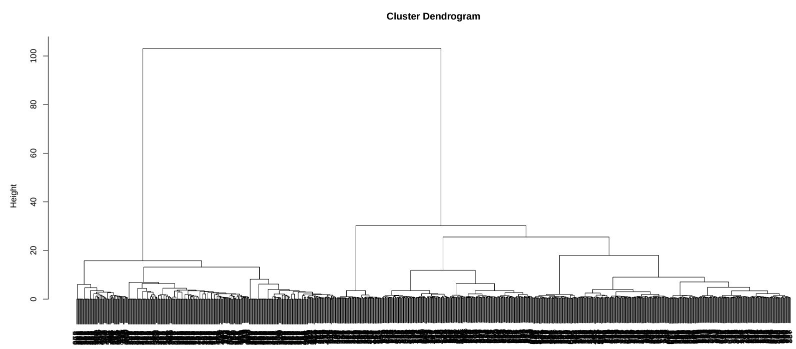


Figure 2.4: Dendrogram plot using Ward's Method of hierarchical clustering.

Next, the optimal number of clusters is determined. Figure 2.5 shows a plot of the cost of merging the clusters. There are several peaks in the merging cost plot, where the first large jump that is furthest to the right is at 7. Since the optimal number of clusters occurs before the jump in merging cost that is furthest to the right, 8 clusters will be used. These clusters are outlined on the dendrogram in Figure 2.6. The number of homes in each cluster is given in Table 2.1, and the clusters range in size from 62 to 298 homes.[6]

[6]Note that the clusters are labeled as seen in the dendrogram in Figure 2.6 from left to right.

Table 2.1: Number of homes in each of the 8 clusters.

Cluster	**# of Homes**	**Cluster**	**# of Homes**
Cluster 1	72	Cluster 5	95
Cluster 2	169	Cluster 6	116
Cluster 3	122	Cluster 7	62
Cluster 4	63	Cluster 8	298

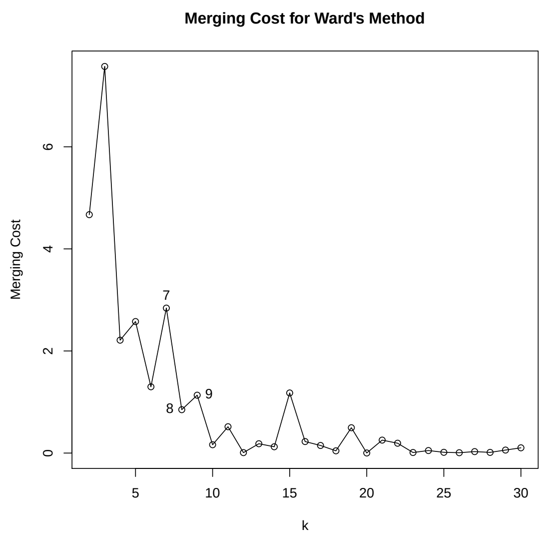


Figure 2.5: The plot of the merging cost, where K is the number of clusters.

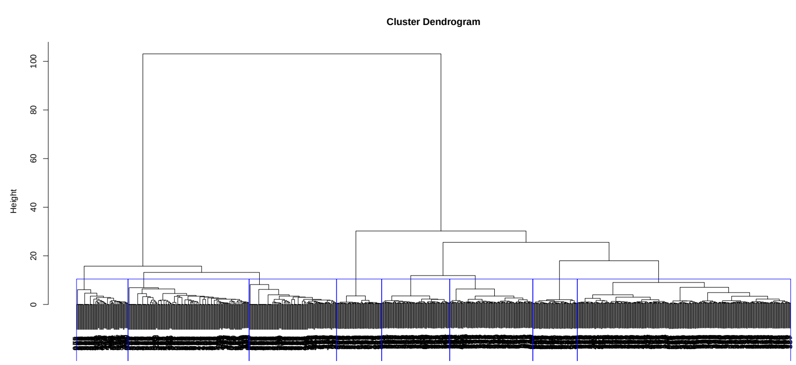


Figure 2.6: Dendrogram plot using Ward's Method of hierarchical clustering with clusters selected using the merging cost outlined in blue.

Characteristics common to the homes within each cluster are explored. First, variables that do not have the same value shared by all homes are investigated. Table 2.2 summarizes a few key numeric variables. Clusters 1, 2, 3, and 5 have a median home age of 3, whereas clusters 7 and 8 have a median home age of 4, cluster 4 has a median age of 7, and cluster 6 has a median age of 51. Looking at conditioned floor area, cluster 7 appears to have more homes that are larger in size, whereas clusters 2 and 3 have smaller sized homes. Additionally, clusters 5, 6, and 8 have a median living height of 16, which implies that they have two stories.

Table 2.2: Mean, median, and standard deviation for a few key numeric variables in each cluster.

	Age of Home			**Living Height**			**Conditioned Floor Area**		
Cluster	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
1	3	3	0	10.75	8	3.95	1306	1222	225.28
2	2.97	3	0.33	9.39	8	2.86	1175	1101	164.04
3	2.91	3	0.60	8.67	8	1.95	1165	1148	171.73
4	6.64	7	1.05	18.86	16.75	4.35	2464	2333	686.19
5	31.93	3	38.16	16.04	16	5.90	2301	2199	794.65
6	53.25	51	34.51	15.73	16	5.36	2102	2064	805.71
7	3.98	4	0.13	18.03	18	3.83	3491	3681	442.99
8	21.96	4	22.74	14.02	16	5.54	3069	3015	934.58
All	17.54	3	26.46	13.29	9	5.60	2190	1995	1073.23

Next, variables that have the same value shared by all homes in a given cluster are investigated. Table 2.3 summarizes these variables.[7] In Table 2.3, a " - " indicates that the variable does not have the same value for all homes in the cluster, and an "NA" indicates that the variable has the value "NA" for all homes in the cluster. Additionally, this table does not include variables that have information for only one home in a given cluster. For example, there is one home in cluster 2 with an ERV Efficiency value of 0.2, and the ERV Efficiency for the rest of the homes in cluster 2 is "NA," so ERV Efficiency is not included in the table. Homes in clusters 1, 2, and 3 are all located in a mixed-humid climate, while

[7] Note that Table 2.3 includes only some variables that are "NA" for all homes in a given cluster. For a complete list of variables that are "NA" for all homes in a cluster, see Table 2.4.

homes in clusters 5, 7, and 8 are all located in a cold climate. All homes in clusters 1, 2, 3, 4, and 7 have a medium roofing material color. Furthermore, all homes in clusters 1, 4, and 7 have air conditioners, while all homes in cluster 3 have air source heat pumps.

In some clusters, there are variable values that are "NA" for all homes. A list of these variables is provided in Table 2.4.[8] Cluster 6 does not have any variables with the value of "NA" for all homes. Clusters 3, 4, and 7 have no information on the R-value of the ceiling cavity or the R-value of the wall insulation for unfinished basements. This is due to the fact that these clusters do not have an unfinished basement (in Table 2.3, the unfinished basement volume is 0 for all three of these clusters). Furthermore, clusters 4 and 7 also do not have information on the R-value of the crawlspace ceiling cavity, the R-value of the crawlspace wall insulation, or the air changes per hour (ACH) of the crawlspace. In Table 2.3, the enclosed crawlspace perimeter is 0 for both clusters 4 and 7, suggesting that none of the homes in these clusters have crawlspaces.

[8] Note that Table 2.4 does not include variables that have information for only one home in a given cluster.

Table 2.3: Variables with the same value for all homes in a given cluster and the corresponding value.

Variable	Cluster 1	2	3	4	5	6	7	8
Unfinished Basement Volume	-	-	0	0	-	-	0	-
Ground Height	-	-	-	0	-	-	-	-
Station	-	-	-	722436	-	-	-	-
Altitude	-	-	-	19.68	-	-	-	-
Climate Region	Mixed-Humid	Mixed-Humid	Mixed-Humid	Hot-Humid	Cold	-	Cold	Cold
Age of Home	3	-	-	-	-	-	-	-
Cooling Degree Days	-	-	-	2817	-	-	-	-
Heating Degree Days	-	-	-	1415	-	-	-	-
Interzonal Floor Cavity R-value	0	0	0	-	0	-	-	-
Roofing Material Color	Medium	Medium	Medium	Medium	-	-	Medium	-
Has Radiant Barrier	F	-	-	-	F	F	F	F
ERV Efficiency	-	-	-	NA	NA	-	0.2	-
AC Cooling SEER	-	-	NA	-	-	-	12	-
AH Leak SA	-	-	-	-	-	-	0	-
Furnace Fuel Type	Gas	-	-	Gas	-	-	Gas	Gas
Unconditioned Duct R-value	-	-	-	6	-	-	0	-
Duct Location	-	-	-	Unfinished Attic	-	-	Living	-
Duct Location Fraction	-	-	-	1	-	-	NA	-
Duct Number of Returns	1	1	-	1	-	-	-	-
Crawlspace ACH	-	-	-	NA	0.2	-	NA	-
Ventilation Type	-	-	-	-	Exhaust	-	-	-
Use Ceiling Fan	F	F	-	-	F	-	F	-
Water Heater Type	-	Tank	-	Tank	-	-	Tank	Tank
Water Heater Fuel Type	-	-	Electric	Gas	-	-	Gas	-
Water Heater Wrap R-value	-	0	-	0	-	-	0	-
Water Heater Recovery Efficiency	-	-	0.99	0.76	-	-	-	-
Dishwasher EF	0.46	0.46	-	0.46	-	-	-	-
Cooking Range Fuel Type	-	-	Electric	-	-	-	Gas	-
Clothes Dryer Fuel Type	Electric	-	Electric	-	-	-	Gas	-
Lighting Values Defaulted	T	-	-	T	-	-	-	-
Has AC	T	-	F	T	-	-	T	-
Has ASHP	F	-	T	F	F	-	F	F
Has Furnace	T	-	F	T	-	-	T	-
Has Boiler	F	F	F	F	-	-	F	-
Perimeter Conditioned Bsmt	-	0	0	0	-	-	-	-
Perimeter Unconditioned Bsmt	-	-	0	0	-	-	0	-

Continued on next page

Table 2.3 continued: Variables with the same value for all homes in a given cluster and the corresponding value.

Variable	**Cluster**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Perimeter Enclosed Crawlspace	-	-	-	0	-	-	0	-
Has Garage	-	-	-	-	-	-	T	-
Watts Per CFM Defaulted	T	-	T	F	-	-	F	-
Has Cooking Range	-	-	-	T	T	-	T	-

Table 2.4: Variables in each cluster with "NA" values for all homes in the cluster.

Cluster 1	**Cluster 2**
HP Cooling SEER	Finished Bsmt Wall R-value
Cluster 3	**Cluster 4**
Unfinished Bsmt Ceiling Cavity R-value Unfinished Bsmt Wall R-value Finished Bsmt Wall R-value AC Cooling SEER Furnace AFUE	Unfinished Bsmt Ceiling Cavity R-value Unfinished Bsmt Wall R-value Finished Bsmt Wall R-value Crawlspace Ceiling Cavity R-value Crawlspace Wall R-value Crawlspace ACH ERV Efficiency Refrigerator Annual Energy HP Cooling SEER
Cluster 5	**Cluster 6**
ERV Efficiency HP Cooling SEER	
Cluster 7	**Cluster 8**
Unfinished Bsmt Ceiling Cavity R-value Unfinished Bsmt Wall R-value Crawlspace Ceiling Cavity R-value Crawlspace Wall R-value Crawlspace ACH HP Cooling SEER Duct Location Fraction	HP Cooling SEER

CHAPTER 3

REMOVING MULTICOLLINEARITY

Multicollinearity occurs when one predictor variable is a linear combination of the remaining predictors in the set. In this project, the presence of multicollinearity presents the problem of overstating the similarities and differences between observations, since the predictor variables are used for clustering. Furthermore, multicollinearity can have a large impact on fitting MLR models. Let the predictor matrix of a MLR model be denoted as $\mathbf{X}$, and let $\mathbf{Y}$ be the response. Then, the estimate of the model coefficients is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If there is multicollinearity in the columns of $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$ may be nearly singular. This causes the estimate of the model coefficients to be unstable, with large standard errors [12].

Multicollinearity can be detected using variance inflation factors (VIFs). The variance inflation factor of variable x_j is given by

$$VIF_j = \frac{1}{1 - R_j^2} \ , \tag{3.1}$$

where R_j^2 is the coefficient of multiple determination when regressing x_j on the remaining predictor variables. The rule of thumb is that variables with VIFs larger than 5 or 10 are removed [12].

Given that many of the variables present in the FDR dataset are related, the VIFs are used to identify redundant variables for subsequent analysis. See Table B.1 in Appendix B for a list of the VIF values for all of the variables present in this dataset and for the subset of variables being used in the analysis. The subset of variables was chosen by removing variables based on large VIFs and expert opinion. This process was continued until all but one variable had VIFs less than 10. The one variable that has a VIF value slightly above the cutoff is Heating Degree Days. Even though the VIF for Heating Degree Days is very close to the cutoff, it is considered to be important from a building engineering standpoint, so it is retained. Additionally, the VIFs for variables that have more than 33% NAs present were calculated only using the variables with less than 33% NAs present and were not included in the calculation of the VIFs of the other variables. Lists of the subset of 52 quantitative and 14 categorical variables that are used in the analysis are shown in Table 3.1 and Table 3.2, respectively. From here on, only the subset of variables given in these tables is used.

Table 3.1: Table of the subset of quantitative variables used in the analysis and the number assigned to each variable.

#	**Variable**	#	**Variable**
1	Number of Bedrooms	27	North Window Area
2	Heating Degree Days	28	West Window Area
3	Fixed Heating Capacity	29	East Window Area
4	Fixed Cooling Capacity	30	Dishwasher EF
5	Age of Home	31	Conditioned Floor Area
6	Interzonal Floor Cavity R-value	32	Number of Stories
7	Front Window Percent of Total	33	Attic Area
8	Left Window Percent of Total	34	Cathedral Ceiling Area
9	Right Window Percent of Total	35	Net Above Grade Wall Area
10	Front Window U-value	36	Unconditioned Basement Perimeter
11	Back Window U-value	37	Enclosed Crawlspace Perimeter
12	Back Window SHGC	38	Wall Difference
13	Left Window U-value	39	Unfinished Bsmt Ceiling Cavity R-value
14	Right Window U-value	40	Unfinished Bsmt Wall R-value
15	Living Space ACH50	41	Finished Bsmt Wall R-value
16	Fraction of ASHRAE	42	Crawlspace Ceiling Cavity R-value
17	House Vent Fan Power	43	Crawlspace Wall R-value
18	AH Leak RA	44	Crawlspace ACH
19	Unconditioned Duct R-value	45	ERV Efficiency
20	Duct Number of Returns	46	Refrigerator Annual Energy
21	Water Heater Energy Factor	47	AC Cooling SEER
22	Water Heater Wrap R-value	48	Furnace AFUE
23	Water Heater Tank Volume	49	Boiler AFUE
24	Above Grade Wall U-value	50	Duct Location Fraction
25	Unfinished Attic U-value	51	Interzonal Wall U-value
26	South Window Area	52	Finished Roof U-value

Table 3.2: Table of the subset of categorical/binary variables used, each one's variable values, and the number assigned to each variable.

#	**Variable**	**Categories**
53	Azimuth	0°, 90°, 180°, 270°
54	Roof Color	Dark, Light, Medium
55	Has Radiant Barrier	0, 1
56	Ventilation Type	Balanced, Supply, Exhaust
57	Duct Location	Garage, Unfinished Attic, Unfinished Bsmt, Other
58	Use Ceiling Fan	0, 1
59	Cooking Range Fuel Type	Gas, Electric
60	Clothes Dryer Fuel Type	Gas, Electric
61	Lighting Values Not Defaulted	0, 1
62	No Furnace	0, 1
63	No Garage	0, 1
64	Slab Option	4, 6, 7, 8, 12, Other
65	Unfinished Bsmt Ceiling Not Insulated	0, 1
66	Furnace Fuel Type	Gas, Electric

CHAPTER 4

CLUSTERING WITH TRIMMED DATA

Gower's coefficient of similarity is sensitive to outliers because the range of the quantitative variables is used in the denominator, so the distributions of all the quantitative variables given in Table 3.1 are explored (see Appendix C). Many of the variables in the data set have distributions that are skewed right. Given that many of the variables have apparent outlying values (i.e. Crawlspace Ceiling Cavity R-value), the upper 10% and lower 10% of the data for quantitative variables with more than 10 unique values are trimmed for the clustering analysis (see Appendix D for the number of unique values observed for each quantiative variable). Hierarchical and fuzzy clustering are performed on the subset of variables for the trimmed and untrimmed data, and the results are discussed.

4.1 Hierarchical Clustering

Gower's coefficient of similarity is calculated using the trimmed and untrimmed data, and the corresponding distance matrices are both not Euclidean. Therefore, a transformation is needed in order to use Ward's method of hierarchical clustering. The calculated Cailliez constant for the untrimmed data is 1.3028, and the properties of a Euclidean distance are satisfied with this transformation. The Cailliez constant given for the trimmed data is 2.2221; however, even with this transformation the distances still do not satisfy all of the properties of a Euclidean distance. Figure 4.1 shows the resulting dendrogram for the untrimmed data using Gower's coefficient of similarity and Ward's method. Compared to the clusters in the preliminary clustering analysis seen in Figure 2.4, these clusters are not as well-formed.

Since Ward's method requires Euclidean distances, this method cannot be used to cluster the trimmed data because a constant cannot be found such that the distances are Euclidean. Therefore, single-linkage and average-linkage will be used. Figure 4.2 shows the resulting dendrograms for the trimmed data using Gower's coefficient of similarity and single-linkage and average-linkage hierarchical clustering algorithms. These clusters are very poorly-formed.

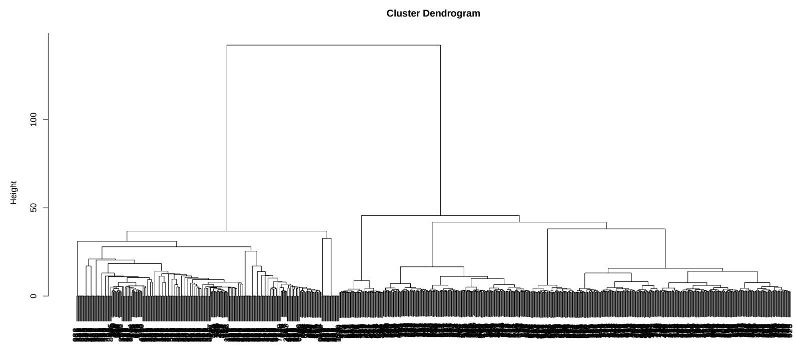


Figure 4.1: Dendrogram for untrimmed data using Ward's method with Gower's coefficient of similarity.

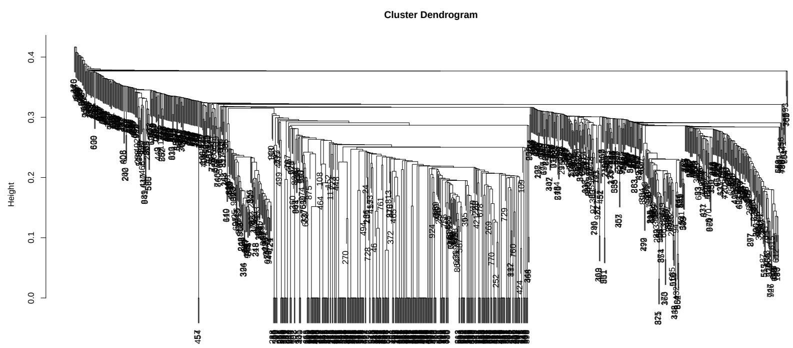
4.2 Fuzzy Clustering

In addition to hierarchical clustering, the uncertainty of cluster membership is explored through fuzzy clustering. Unlike hierarchical clustering, the number of clusters, K, in fuzzy clustering must be preselected. Fuzzy clustering assigns each observation a probability of belonging to each cluster, known as the "strength of membership" [21]. The strength of membership, u_{ik}, of the i^{th} observation for the k^{th} cluster thus has two constraints: $u_{ik} \geq 0$ for $k = 1, 2, ..., K$ and all i, and $\sum_{k=1}^{K} u_{ik} = 1$ for all i. The objective function

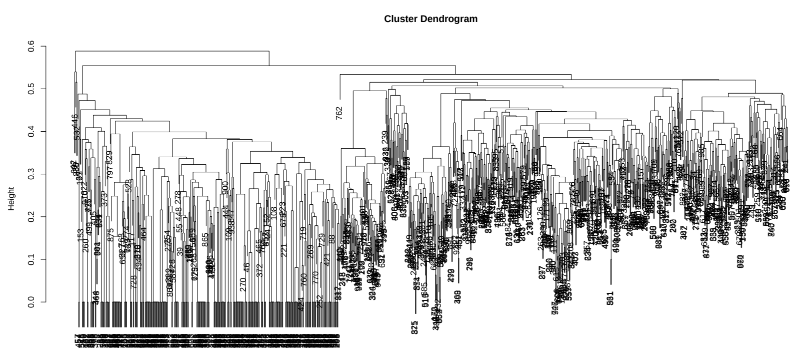
$$\sum_{k=1}^{K} \frac{\sum_i \sum_j u_{ik}^r u_{jk}^r d_{ij}}{2 \sum_l u_{lk}^r} \tag{4.1}$$

is minimized to find these strength of memberships, where d_{ij} is the distance between the i^{th} and j^{th} observation, and r is the membership exponent. The clusters are considered to be more crisp as r goes to 1, and as r goes to infinity, the clustering moves towards complete fuzzyness, which is when observations are equally likely to be in each cluster [27].

The same distance matrices used in Section 4.1 are used in fuzzy clustering for the trimmed and untrimmed data. Regardless of the number of clusters or membership exponent chosen for both the trimmed and untrimmed data, the strength of membership is $\frac{1}{k}$ for all homes and all clusters $k = 1, 2, ..., K$. In other words, all homes are equally likely to be in



(a) Single-Linkage



(b) Average-Linkage

Figure 4.2: Dendrograms for trimmed data using single-linkage and average-linkage with Gower's coefficient of similarity.

any of the K clusters. These results paired with the results of the hierarchical clustering imply that the homes are not suitable for clustering when using the subset of 66 variables, regardless of whether or not the data has been trimmed.

CHAPTER 5

PRINCIPAL COMPONENT ANALYSIS

Given that the subset of variables is not forming good clusters of homes, a principal component analysis will be performed on the untrimmed data, and the clustering analysis will be done on the resulting principal components. The hypothesis is that rotating the coordinate plane will allow distinct clusters to form by creating more evident similarities and differences between the homes.

Principal Component Analysis (PCA) is a multivariate statistical technique that transforms the original data into a new set of variables through an eigendecomposition of the covariance matrix (or correlation matrix). Let $\mathbf{X}' = [X_1, X_2, ..., X_p]$ have covariance matrix $\mathbf{\Sigma}$ (or correlation matrix $\boldsymbol{\rho}$) with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e_1}), (\lambda_2, \mathbf{e_2}), ..., (\lambda_p, \mathbf{e_p})$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Then, the i^{th} principal component is given by

$$Z_i = \mathbf{e_i'X} = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, \tag{5.1}$$

for $i = 1, 2, ..., p$. The variance of Z_i is $\mathbf{e_i'\Sigma e_i}$, which is equal to λ_i. Furthermore, if the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ are distinct, then the eigenvectors are orthogonal. Therefore, the principal components are uncorrelated and have maximum variance with decreasing magnitude (Var$(Z_1) \geq$Var$(Z_2) \geq ... \geq$Var(Z_p)) [19].

Given that the variables in the FDR data set have very different units, the correlation matrix will be used for the principal component analysis. Pearson's product-moment correlation coefficient is used, which is given as

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}, \tag{5.2}$$

where cov(X_i, X_j) is the covariance of X_i and X_j, and σ_{X_i} and σ_{X_j} are the standard deviations of X_i and X_j, respectively. Pearson's correlation coefficient is chosen since it is mathematically equivalent to the phi coefficient, which is used to measure the association between two binary variables [28], and the point-biserial correlation coefficient, which is used to measure association between continuous and binary variables [29]. The variables with more than 33% missing values are removed in order to calculate the correlation, so the final data set includes 49 variables, which are variables 1 through 38 in Table 3.1 and 53 through 63 in Table 3.2. There are 55 variables after the categorical variables are coded using dummy variables for the principal component analysis.

The first few principal components of the 55 variables are examined. Table 5.1 shows the weighting coefficients for the first four principal components. The first principal component can be interpreted as a factor representing the size of the home. The second principal component is primarily influenced by the age of the home. The third principal component can be described as heating and cooling systems, and the fourth principal component is mostly impacted by window variables. The scree plot, which is a plot of the ordered eigenvalues, can be seen in Figure 5.1. This plot shows the proportion of variability explained by each principal component. Only 38% of the variability present in the data is explained by the first four principal components, so all of the principal components will be used to capture the total variability.

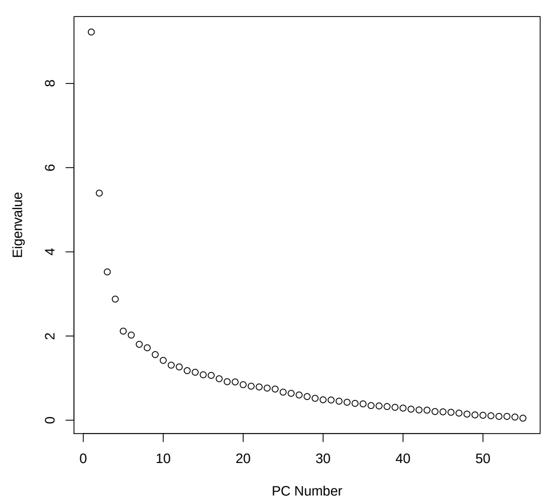


Figure 5.1: Scree plot from principal component analysis.

Table 5.1: The weighting coefficients for the first four principal components, with variable coefficients with large weights in bold.

Variable	**PC 1**	**PC 2**	**PC 3**	**PC 4**
Number of Bedrooms	0.12	-0.14	-0.15	-0.05
Heating Degree Days	**0.21**	0.03	**0.34**	-0.02
Fixed Heating Capacity	**0.26**	0.07	-0.10	-0.02
Fixed Cooling Capacity	0.16	-0.07	**-0.34**	0.03
Age of Home	0.11	**0.27**	0.04	-0.08
Interzonal Floor Cavity R-value	0.13	-0.15	-0.07	-0.09
Front Window Percent of Total	0.10	-0.06	-0.01	**0.41**
Left Window Percent of Total	-0.18	0.16	-0.01	**-0.28**
Right Window Percent of Total	-0.03	0.02	0.05	**-0.35**
Front Window U-value	0.07	**0.25**	-0.13	0.04
Back Window U-value	0.11	**0.28**	-0.17	0.13
Back Window SHGC	0.09	**0.32**	0.02	0.08
Left Window U-value	0.07	0.13	-0.05	**0.30**
Right Window U-value	0.05	0.14	-0.05	**0.34**
Living Space ACH50	-0.08	**0.32**	-0.10	-0.04
Fraction of ASHRAE	-0.01	**-0.23**	-0.17	-0.13
House Vent Fan Power	0.12	0.01	0.19	0.06
AH Leak RA	-0.01	0.12	**-0.33**	-0.01
Unconditioned Duct R-value	**-0.20**	-0.06	**-0.23**	-0.02
Duct Number of Returns	**0.23**	-0.06	**0.20**	-0.05
Water Heater Energy Factor	**-0.25**	-0.04	-0.03	-0.07
Water Heater Wrap R-value	-0.03	-0.01	0.05	0.04
Water Heater Tank Volume	0.13	-0.03	-0.13	-0.15
Above Grade Wall U-value	0.04	**0.28**	-0.04	-0.07
Unfinished Attic U-value	0.04	0.17	-0.06	-0.04
South Window Area	0.19	0.01	0.01	-0.15
North Window Area	0.18	-0.01	0.04	-0.13
West Window Area	0.12	-0.10	**-0.30**	-0.01
East Window Area	0.11	-0.09	**-0.21**	-0.01
Dishwasher EF	0.06	0.02	0.01	0.03
Conditioned Floor Area	**0.28**	-0.11	0.02	-0.00
Number of Stories	0.19	0.05	-0.10	-0.13
Attic Area	0.09	-0.19	-0.11	0.11
Cathedral Ceiling Area	0.08	0.08	-0.13	-0.02
Net Above Grade Wall Area	**0.25**	-0.13	-0.19	-0.07
Unconditioned Bsmt Perimeter	0.00	0.17	-0.08	-0.11
Enclosed Crawlspace Perimeter	-0.19	-0.02	-0.08	0.01
Wall Difference	0.05	0.08	0.01	-0.07
Azimuth: 0 Degrees	0.06	0.05	0.11	-0.00
Azimuth: 90 Degrees	-0.13	-0.05	0.01	0.08

Continued on next page

Table 5.1 continued: The weighting coefficients for the first four principal components, with variable coefficients with large weights in bold.

Variable	**PC 1**	**PC 2**	**PC 3**	**PC 4**
Azimuth: 180 Degrees	0.07	0.06	0.11	-0.14
Roof Color: Dark	0.06	-0.03	0.10	0.06
Roof Color: Light	0.00	0.06	-0.01	0.03
Has Radiant Barrier	-0.10	-0.06	0.02	-0.01
Ventilation Type: Balanced	0.10	-0.17	0.01	**-0.22**
Ventilation Type: Supply	-0.17	-0.10	-0.03	-0.03
Duct Location: Garage	0.01	0.05	-0.01	0.03
Duct Location: Unfinished Attic	-0.07	-0.03	**-0.29**	-0.01
Duct Location: Unfinished Bsmt	-0.00	0.19	-0.09	-0.13
Use Ceiling Fan	0.08	-0.00	0.05	0.02
Cooking Range Fuel Type: Gas	0.13	0.07	-0.00	**-0.25**
Clothes Dryer Fuel Type: Gas	0.13	-0.04	0.10	-0.19
Lighting Values Not Defaulted	0.05	0.17	0.06	-0.00
No Furnace	**-0.23**	-0.05	0.01	-0.02
No Garage	**-0.21**	0.15	-0.03	-0.18

CHAPTER 6

CLUSTERING USING PRINCIPAL COMPONENTS

Clustering is now performed using the principal components. Given that the principal components are numeric, Euclidean distances can and will be used instead of Gower's coefficient. Clustering is performed on the principal components using unweighted and weighted Euclidean distances, where the weights are the corresponding eigenvalues. Both fuzzy and hierarchical clustering methods are applied, and potential outliers are identified and investigated.

6.1 Hierarchical Clustering

Single-linkage, average-linkage, and Ward's method are used to cluster the PCs, and the resulting dendrograms are shown in Figure 6.1, Figure 6.2, and Figure 6.3, respectively. The clusters formed using single and average-linkage both have one home that is in its own cluster until the last few steps, and these homes are identified as potential outliers in the data set.

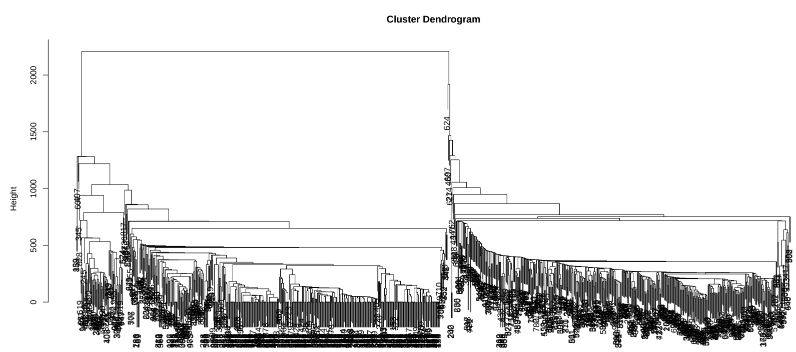


Figure 6.1: Dendrogram for single linkage hierarchical clustering using principal components.

6.1.1 Identifying Potential Outliers

The two homes in Figure 6.1 and Figure 6.2 that are not clustered until the last steps are potential outliers in the data set. In single linkage, this is home #624 and is joined with

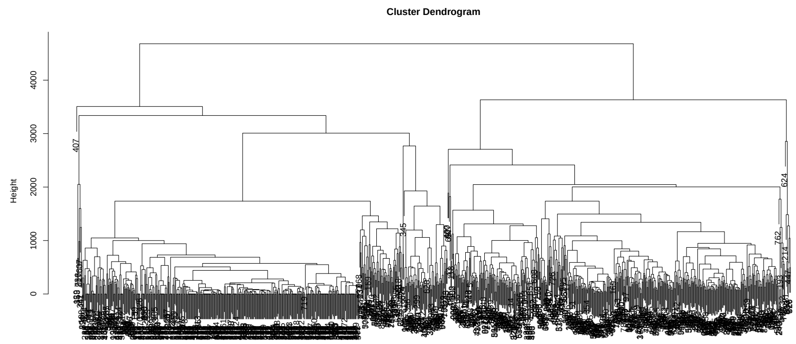


Figure 6.2: Dendrogram for average linkage hierarchical clustering using principal components.

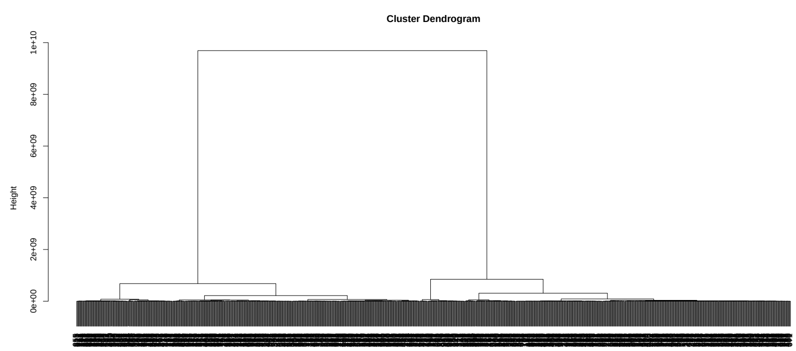


Figure 6.3: Dendrogram for Ward's method of hierarchical clustering using principal components.

a cluster in the second to last step. In average linkage, it is home #407, and this home is joined with a cluster in the third to last step. To determine whether or not these homes are outliers, their characteristics are explored. Table 6.1 shows particular variable values of these homes that differ significantly from the median. Home #407 is an older home with a larger conditioned floor area than the median. It also has a large amount of cathedral ceiling area, when the median value is 0. On the other hand, home #624 has a larger number of heating degree days than the median but less cathedral ceiling area than home #407. Furthermore, both homes do not have a garage, which is not common for their size.

Table 6.1: Table of potential outlying homes and their values for a few variables that differ from the median.

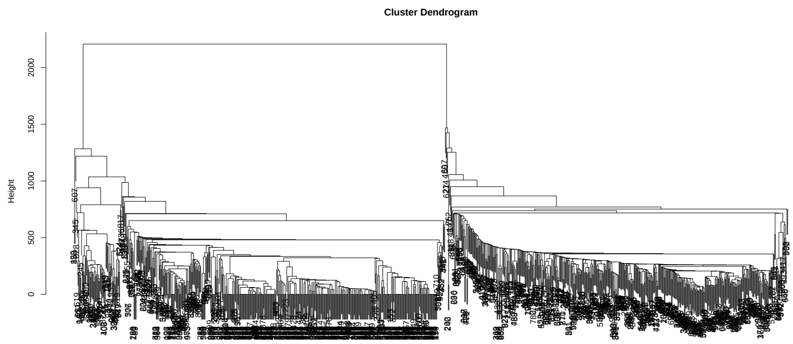
Home	**Heating Degree Days**	**Age of Home**	**Conditioned Floor Area**	**Cathedral Ceiling Area**	**No Garage**
Median	4340	3	1995	0	No
407	4301	50	3218	2417	Yes
624	7281	3	7648	570	Yes

To explore their impact on the clustering analysis, these two homes are removed and clustering using single-linkage, average-linkage, and Ward's method are recalculated. Removing these homes has little impact on the resulting clusters, which can be seen in the dendrograms in Figure 6.4 for single-linkage, average-linkage, and Ward's method. Therefore, neither home will be removed.

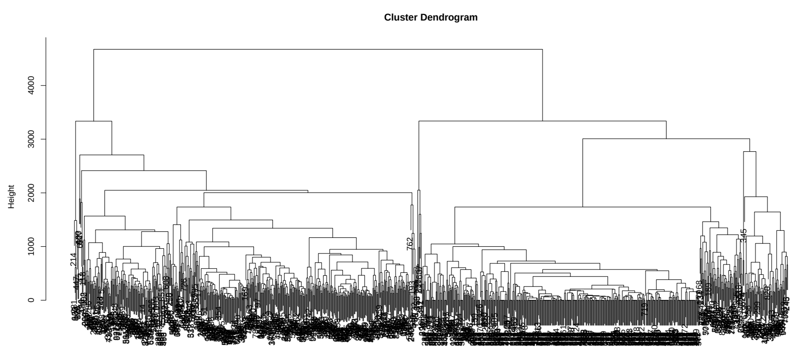
6.1.2 Weighted Clustering

Next, the distances between homes based on the principal components are weighted by the eigenvalues since the eigenvalues represent the portion of variability attributable to the corresponding principal component. The weighted Euclidean distance between two $p \times 1$ vectors of principal components, $\mathbf{y}_i$ and $\mathbf{y}_j$, is given by

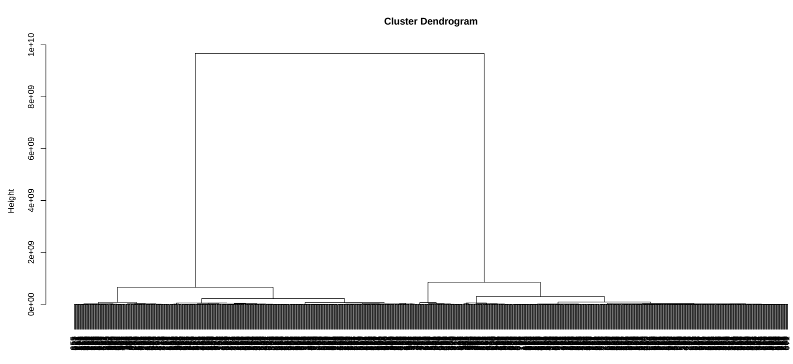
$$d_{ij} = \sqrt{\sum_{m=1}^{p} w_m (y_{im} - y_{jm})^2}, \tag{6.1}$$



(a) Single-Linkage



(b) Average-Linkage



(c) Ward

Figure 6.4: Dendrograms for single-linkage, average-linkage, and Ward's method using principal components with potential outliers removed.

where w_m is equal to $\frac{\lambda_m}{\sum_{m=1}^{p} \lambda_m}$. The weighted distances are then used to cluster the homes. Figure 6.5 is the resulting dendrogram for Ward's method, and the structure is very similar to that of Ward's method with unweighted distances.

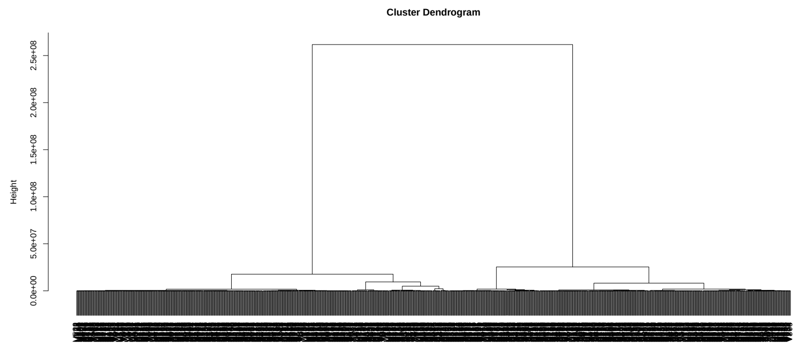


Figure 6.5: Dendrogram for Ward's method using weighted Euclidean distances of the principal components.

6.2 Fuzzy Clustering

Using the unweighted and weighted Euclidean distance matrices and a membership exponent of 1.5, the clusters resulting from fuzzy clustering no longer have equal strength of membership for all homes and all clusters $k = 1, 2, ..., K$ for varying K. The membership exponent was chosen to be 1.5 to ensure that the clusters are clearly defined. These results imply that the data is more suitable for clustering after rotating the coordinate plane. Table 6.2 has a sample of some membership coefficients when $K = 2$ clusters are chosen. Most homes have one cluster membership that strongly suggests the home belongs in that cluster.

Table 6.2: Sample of membership probabilities for 2 fuzzy clusters.

	Strength of Membership	
Home	**Cluster 1**	**Cluster 2**
1	0.81	0.19
2	0.13	0.87
3	0.16	0.84
4	0.75	0.25
5	0.20	0.80
6	0.11	0.89
7	0.89	0.11
8	0.15	0.85
9	0.75	0.25
10	0.16	0.84

CHAPTER 7

DETERMINING THE NUMBER OF CLUSTERS

Methods for determining the number of appropriate clusters are examined. The approaches considered are the Xie-Beni Index, C Index, and Dunn Index. These are selected due to their interpretable definitions and common use. Also, a simulation study is done using both nonparametric and parametric approaches, and the number of clusters is chosen based on the results.

7.1 Xie-Beni Index

The Xie-Beni index [30] is commonly used to determine the number of clusters for fuzzy clustering, as well as for other clustering techniques, and is defined as the quotient between the mean quadratic error and the minimum of the minimal squared distances between the points in the clusters. Let I_k be the set of indices of the observations belonging to cluster C_k, V_j be the j^{th} column of the data matrix, and the center vectors, v_j, be given by $v_j = V_j - \mu_j$. Then, the within group scatter matrix, $WG_{\{I_k\}}$, is

$$WG_{\{I_k\}} = \mathbf{A}'_{\{I_k\}}\mathbf{A}_{\{I_k\}}, \tag{7.1}$$

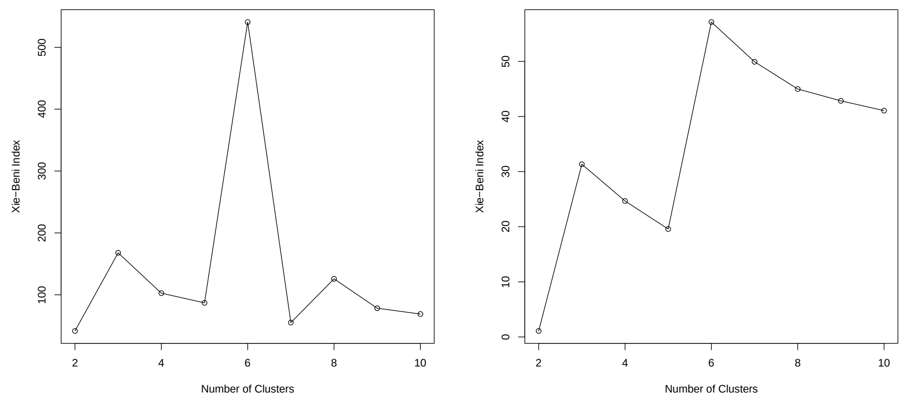
where $\mathbf{A}_{\{I_k\}}$ is the matrix formed by the center vectors for the observations belonging to cluster C_k. The pooled within-cluster sum of squares, denoted $WGSS$, is thus given by

$$WGSS = \sum_{k=0}^{K} \mathrm{Tr}(WG_{\{I_k\}}), \tag{7.2}$$

and the Xie-Beni index can be written as

$$C = \frac{1}{N}\frac{WGSS}{\min_{k<k'} \delta_1(C_k, C_{k'})^2}, \tag{7.3}$$

where $\delta_1(C_k, C_{k'})$ is the minimum distance between two points, one located in cluster C_k and the other in $C_{k'}$, and N is the total number of observations. The minimum value of the Xie-Beni index indicates the best number of partitions [31]. Figure 7.1 is a plot of the Xie-Beni index for fuzzy clustering with a membership exponent of 1.5 and for Ward's method, both using weighted Euclidean distances. The minimum value of the Xie-Beni index is 41.5024 for fuzzy clustering and 1.0836 for Ward's method, and both occur at $K = 2$ clusters. Results are similar using unweighted Euclidean distances, with a minimum value of 0.3434 for both fuzzy clustering and for Ward's method. Like with the weighted Euclidean distances, both occur at $K = 2$ clusters.



(a) Fuzzy (b) Ward

Figure 7.1: Plot of Xie-Beni index for fuzzy clustering and Ward's method.

7.2 C Index

Another measure used to determine the best number of partitions is the C index [32]. Let N be the total number of observations and n_k be the number of observations in cluster C_K. N_W is the total number of distinct pairs of points in C_k, calculated as

$$N_W = \frac{1}{2}\left(\sum_{i=1}^{k} n_i^2 - N\right), \tag{7.4}$$

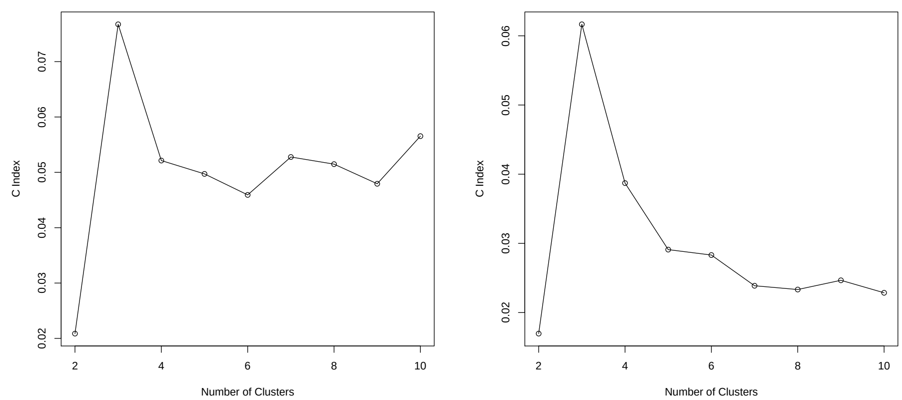
and N_T is the total number of distinct pairs of points in the data set, given by

$$N_T = \frac{N(N-1)}{2}. \tag{7.5}$$

The C index is defined as

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}}, \tag{7.6}$$

where S_W is the sum of the N_W distances between all the pairs of points in each cluster, S_{min} is the sum of the N_W smallest distances between all the pairs of points in the entire dataset, and S_{max} is the sum of the N_W largest distances between all the pairs of points in the entire dataset. Note that there are N_T such pairs, so S_{max} and S_{min} are the sums of the N_W largest or smallest values [31]. Figure 7.2 is a plot of the C index for fuzzy clustering with a membership exponent of 1.5 and for Ward's method, both using weighted Euclidean distances. The minimum value is 0.0209 for fuzzy clustering and 0.0170 for Ward's method, and both occur at $K = 2$ clusters. Results are similar using unweighted Euclidean distances, with a minimum value of 0.0146 for both fuzzy clustering and Ward's method. Like with the weighted Euclidean distances, both occur at $K = 2$ clusters.



(a) Fuzzy (b) Ward

Figure 7.2: Plot of C index for fuzzy clustering and Ward's method.

7.3 Dunn Index

The Dunn index [33] uses the maximum value of its criteria to determine the optimal number of partitions. Let M_j be the vector of values for observation j, and let I_k be the indices of the observations belonging to cluster C_k. The distance between the closest points in clusters C_k and $C_{k'}$ is given as

$$d_{kk'} = \min_{i \in I_k,\, j \in I_{k'}} \|M_i^{\{k\}} - M_j^{\{k'\}}\|, \tag{7.7}$$

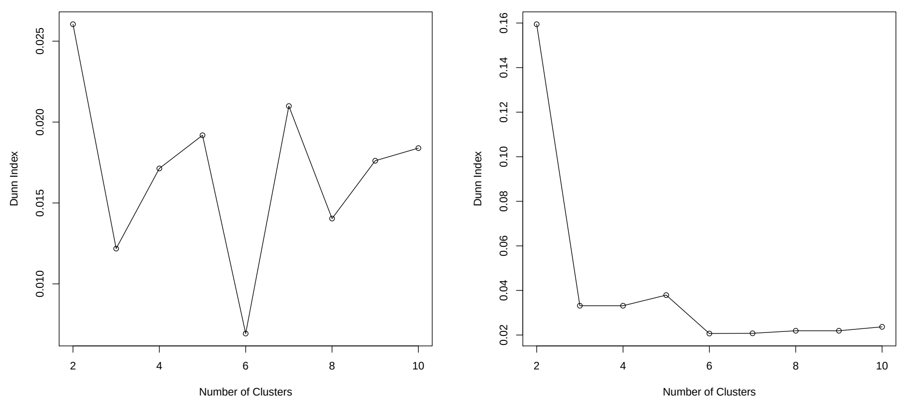
and the largest distance separating two distinct points in cluster C_k is

$$D_k = \max_{i,j \in I_k,\, i \neq j} \|M_i^{\{k\}} - M_j^{\{k'\}}\|. \tag{7.8}$$

Then, the Dunn index is given by

$$C = \frac{\min_{k \neq k'} d_{kk'}}{\max_{1 \leq k \leq K} D_k} = \frac{d_{min}}{d_{max}}. \tag{7.9}$$

Unlike the Xie-Beni and C indices, the maximum value of the Dunn index indicates the best partition [31]. Figure 7.3 is a plot of the Dunn index for fuzzy clustering with a membership exponent of 1.5 and for Ward's method, both using weighted Euclidean distances. The maximum value of the Dunn index is 0.0260 for fuzzy clustering and 0.1595 for Ward's method, and both occur at $K = 2$ clusters. Results are similar using unweighted Euclidean distances, with a maximum value of 0.2810 for both fuzzy clustering and Ward's method. Like with the weighted Euclidean distances, both occur at $K = 2$ clusters.



(a) Fuzzy (b) Ward

Figure 7.3: Plot of Dunn index for fuzzy clustering and Ward's method.

7.4 Simulation Study

Next, a simulation study is performed using bootstrapping to estimate the number of clusters. Bootstrapping is a technique used to make inferences about a given population characteristic based on a sample from the population. Both nonparametric and parametric bootstrapping approaches are taken to find the distribution of the number of clusters chosen based on the Xie-Beni, C, and Dunn indices.

7.4.1 Nonparametric Approach

Resampling is a nonparametric bootstrapping technique that can be used to estimate the shape of the sampling distributions of the optimal number of clusters chosen based on each of the three indices. One benefit to nonparametric bootstrapping is that it makes no assumptions about the underlying distribution of the data [34]. The steps we take to nonparametrically generate the sampling distribution of the minimum Xie-Beni, minimum C, and maximum Dunn indices are as follows:

1. A bootstrap sample, $(\mathbf{x_1^*}, \mathbf{x_2^*}, ..., \mathbf{x_{997}^*})$, is chosen with replacement from the original sample of 997 observations $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_{997}})$.

2. Then, clustering is performed using the PCs of the bootstrapped sample (for both fuzzy clustering and Ward's method of hierarchical clustering).

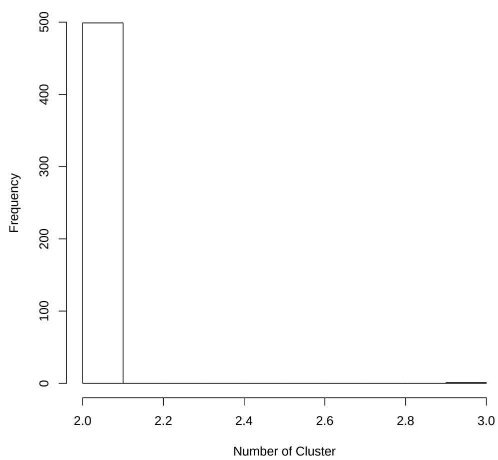
3. The Xie-Beni, C, and Dunn indices are computed for $K = 2, ..., 10$.

4. The minimum values for the Xie-Beni and C indices and the maximum value of the Dunn index are saved, denoted as $\hat{C}_{xie}$, $\hat{C}_c$, and $\hat{C}_{dunn}$, respectively.

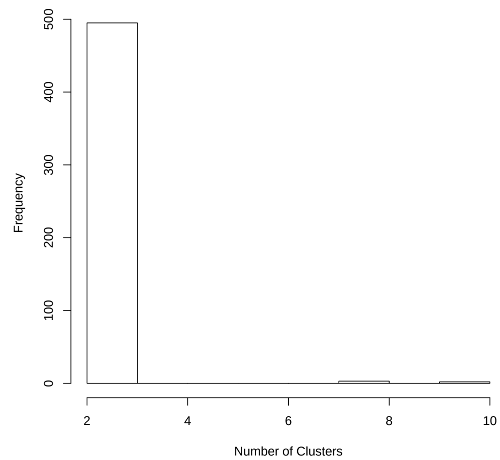
5. Steps 1-4 are repeated 500 times.

6. The distributions of $\hat{C}_{xie}$, $\hat{C}_c$, and $\hat{C}_{dunn}$ are plotted with histograms.

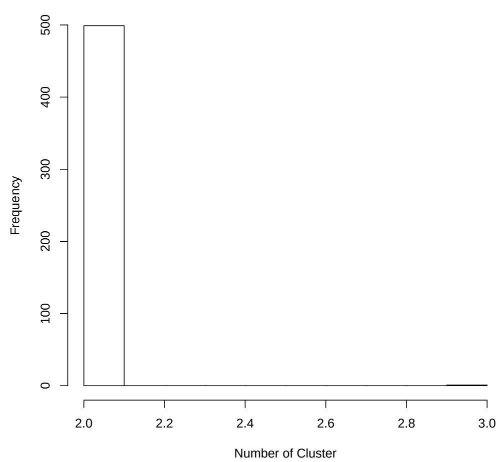
Given that calculating weighted Euclidean distances is computationally expensive, unweighted distances were used in both clustering algorithms. Figure 7.4 shows the resulting histograms of the best number of clusters chosen under Ward's method. For the Xie-Beni index, only one simulation out of 500 did not select two as the best number of clusters. For the C index, of the 500 simulations, there were 17 instances where three was determined to be the best number of partitions; in three instances 8 was calculated as the best number; and in two instances 10 was chosen as the best number of clusters. Like the Xie-Beni index, only one simulation out of 500 did not give two as the best number of clusters for the Dunn index. The results were similar for bootstrapping using fuzzy clustering. In that case, all 500 bootstrap samples selected two as the best number of clusters for the Xie-Beni and Dunn indices, and only three bootstrap samples did not select two as the best number of clusters for the C index.



(a) Xie-Beni Index



(b) C Index



(c) Dunn Index

Figure 7.4: Nonparametric bootstrapped distributions of the cluster selection indices for Ward's method.

7.4.2 Parametric Approach

Next, a parametric bootstrapping approach is used to estimate the shape of the sampling distribution of the number of clusters selected by the Xie-Beni, C, and Dunn indices. Unlike nonparametric bootstrapping, this method requires assumptions about the underlying distribution of the data. Parametric bootstrap samples are simulated using a parametric model fitted to the data [34]. A commonly used distribution for modeling multivariate data is the multivariate normal distribution, but this data is not multivariate normal. Therefore, a multivariate Box-Cox transformation is used to transform the data to multivariate normality [35]. Then, a parametric bootstrap sample is drawn from a multivariate normal distribution with mean and covariance based on the transformed data. The parametric bootstrap sample is then transformed using the inverse Box-Cox transformations so that the simulated data has a distribution similar to that of the original. Using the inversely transformed bootstrap samples, clustering is done for both Ward's method and fuzzy clustering. The basic steps in parametric bootstrap simulation are as follows:

1. A bootstrap sample, $(\mathbf{x_1^*}, \mathbf{x_2^*}, ..., \mathbf{x_{997}^*})$, is simulated from a multivariate normal distribution with mean and covariance matrix based on the 997 transformed observations.

2. The bootstrap sample is transformed using the inverse Box-Cox transformations.

3. Then, clustering is performed using the PCs of the transformed bootstrapped sample (for either fuzzy clustering or Ward's method of hierarchical clustering).

4. The Xie-Beni, C, and Dunn indices are computed for $K = 2, ..., 10$.

5. The minimum values for the Xie-Beni and C indices and the maximum value of the Dunn index are saved, denoted as $\hat{C}_{xie}$, $\hat{C}_c$, and $\hat{C}_{dunn}$, respectively.

6. Steps 1-5 are repeated 500 times.

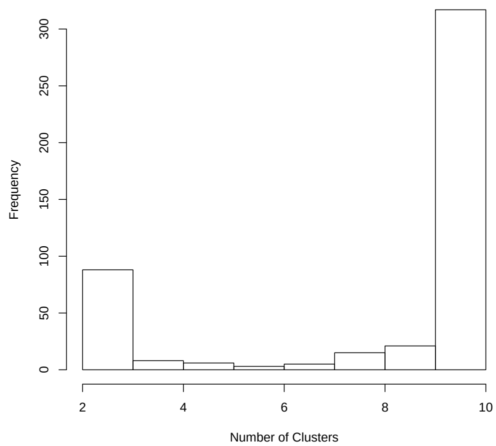
7. The distributions of $\hat{C}_{xie}$, $\hat{C}_c$, and $\hat{C}_{dunn}$ are used to estimate the shape of the sampling distributions of the Xie-Beni, C, and Dunn indices.

Given that calculating weighted Euclidean distances is computationally expensive, unweighted distances were used in both clustering algorithms. Figure 7.5 shows the resulting histograms of the best number of clusters for parametric bootstrapping with Ward's method for 500 simulations. Based on the distribution of the Xie-Beni and C indices, the best number of clusters is 10; whereas two is the best number of clusters for the Dunn index. Figure 7.6 shows the resulting histograms for the number of clusters for parametric bootstrapping using fuzzy clustering for 500 simulations. Like with Ward's method, the best number of clusters is 10 according to the Xie-Beni and C indices, and the best number of clusters is two for the Dunn index.

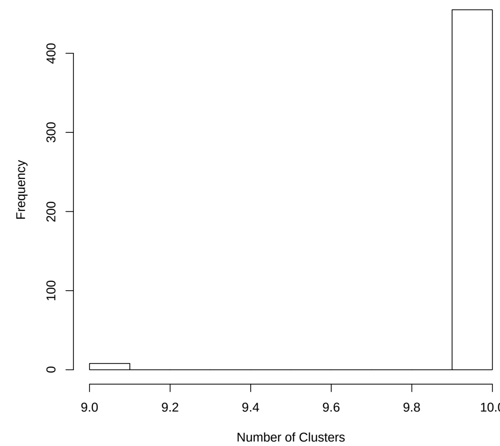
Since these results differ from the results in Section 7.4.1 and the only assumption made was multivariate normality, this assumption is evaluated. This is done through the use of a chi-square plot. This method is based on evaluating the joint normality of the data based on the Mahalanobis distance, which is given by

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}), \tag{7.10}$$

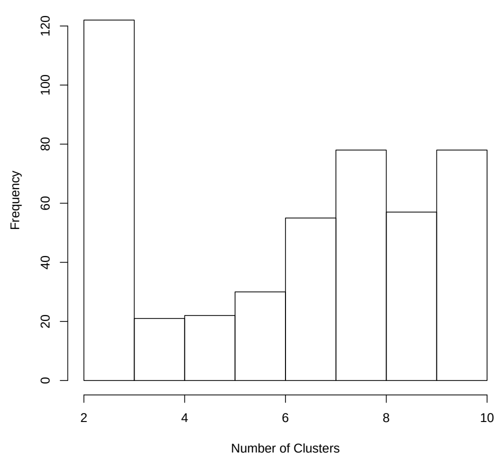
where $\bar{\mathbf{x}}$ is the vector of sample means, $\mathbf{S}$ is the sample covariance matrix, and $\mathbf{x}_j$ is a $p \times 1$ vector of observations for $j = 1, 2, ..., n$. When the underlying distribution of the data is multivariate normal, each of the Mahalanobis distances should behave like a chi-square random variable. Therefore, the chi-square plot, which is a plot of the pairs of the sorted Mahalanobis distances and the $\frac{100(j-0.5)}{n}$ quantile of the chi-square distribution with p degrees of freedom, should look like a straight line with a slope of one [19]. Figure 7.7 shows the chi-square plot for the original and transformed data, along with one sample from the multivariate normal distribution with mean and covariance matrix based on the transformed observations. The curved pattern in both the original and transformed data plots suggests a lack of multivariate normality even after the transformation. Therefore, the simulated data does not even reflect the transformed data, and this explains why the results differ so much from the nonparametric bootstrapping experiment. Since the assumption of multivariate normality used in the parametric bootstrapping does not accurately reflect the distribution



(a) Xie-Beni Index

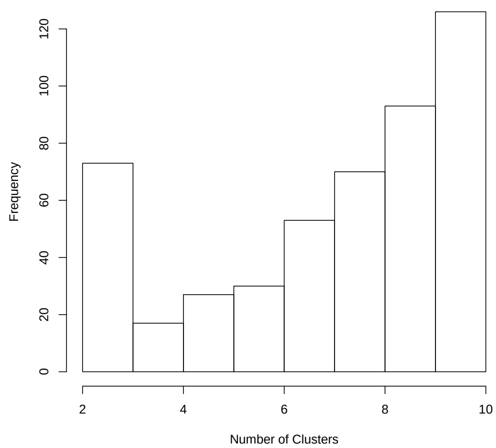


(b) C Index

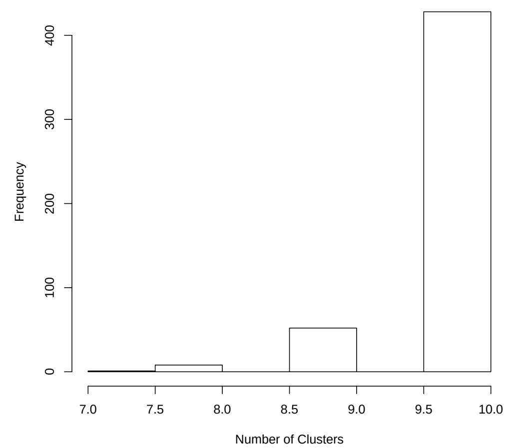


(c) Dunn Index

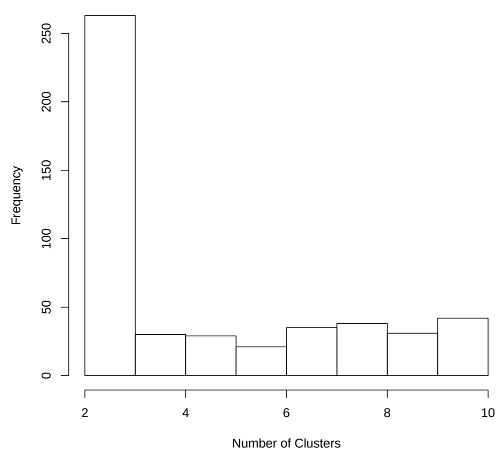
Figure 7.5: Parametric bootstrapped distributions of the cluster selection indices for Ward's method.



(a) Xie-Beni Index



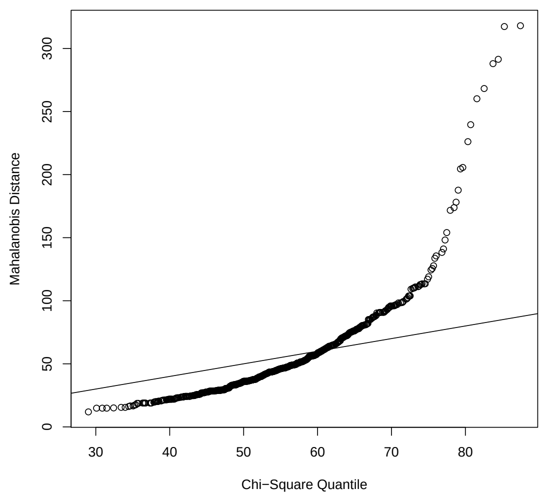
(b) C Index



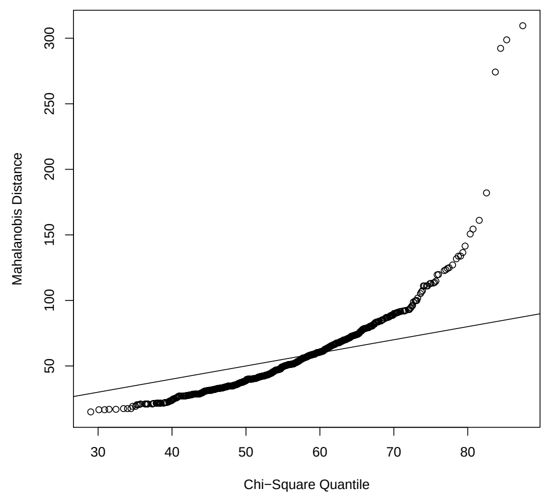
(c) Dunn Index

Figure 7.6: Parametric bootstrapped distributions of the cluster selection indices for fuzzy clustering.

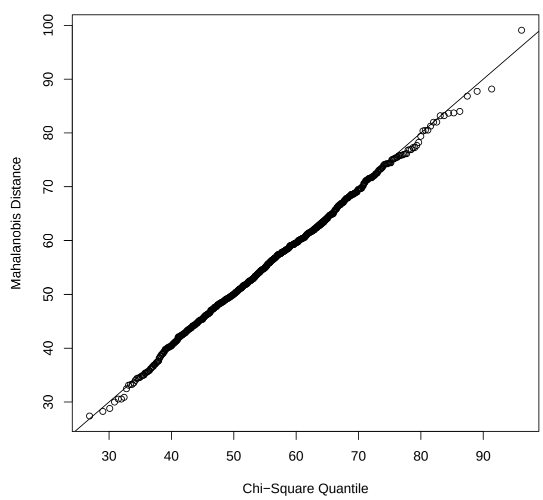
of the transformed data, only the results from the nonparametric bootstrapping will be used to determine the best number of clusters. The nonparametric bootstrapped distributions of the minimum Xie-Beni, minimum C, and maximum Dunn indices for both Ward's method and fuzzy clustering provide strong evidence for $K = 2$, so two clusters is chosen.



(a) Original Data



(b) Transformed Data



(c) Bootstrapped Sample

Figure 7.7: Chi-squared plot for original data, transformed data, and one sample from the multivariate normal distribution with mean and covariance matrix based on the transformed observations.

CHAPTER 8

COMPARISON OF CLUSTERS

Regardless of the method used and whether the distances are weighted, the best partition is two clusters according to the Xie-Beni, C, and Dunn indices, and confirmed by the nonparametric bootstrap simulation study. The number of homes present in each cluster for Ward's method and fuzzy clustering using weighted and unweighted distance metrics is shown in Table 8.1.

Table 8.1: Unweighted and weighted cluster sizes.

	Unweighted		**Weighted**	
Cluster	**Ward**	**Fuzzy**	**Ward**	**Fuzzy**
1	518	518	520	507
2	479	479	477	490

The clusters formed using unweighted distances in Ward's method are the same as the clusters formed when using unweighted distances in fuzzy clustering. The clusters formed using Ward's method with weighted distances are two homes different than the clusters formed using Ward's method with unweighted distances. The clusters formed using fuzzy clustering with weighted distances have 13 homes that are different than the clusters formed using fuzzy clustering with unweighted distances. Additionally, there are 14 homes that are in cluster 1 for the weighted Ward's method that are in cluster 2 for weighted fuzzy clustering, and there is one home that is in cluster 2 for the weighted Ward's method that is in cluster 1 for weighted fuzzy clustering. An outline of a few variables for these 15 homes is shown in Table 8.2. Most of these homes are located in the same area, are older, and have an enclosed crawlspace.

Given that a very small number of homes change clusters when using weighted and unweighted distance metrics for the different clustering methods, the clusters formed using the weighted distance metric and Ward's method will be used for the regression analysis, since Ward's method is less computationally expensive. The weighted distance metric was

Table 8.2: Homes that switch clusters for weighted Ward's method and weighted fuzzy clustering.

Home	Cluster Fuzzy	Cluster Ward	Heating Degree Days	Age of Home	Enclosed Crawlspace Perimeter
33	1	2	4301	24	206
139	1	2	4301	65	168
342	1	2	4301	50	0
355	1	2	4301	67	57
407	1	2	4301	50	193
441	1	2	4301	86	0
460	1	2	6542	4	349.5
532	2	1	6542	89	137.5
607	1	2	4301	70	12
627	1	2	6542	12	264
648	1	2	4301	5	179
736	1	2	4301	4	90
817	1	2	4301	5	197
858	1	2	4301	4	232
943	1	2	4301	101	0
Median (15 homes)	-	-	4301	50	168
Median (all homes)	-	-	4340	3	0

chosen because the weights emphasize the variables that explain most of the variability present in the data.

Now, the within cluster similarities are explored. In Table 8.3 are the mean, median, and standard deviations of select numeric variables for both clusters. The mean heating degree days for the homes in cluster 1 is smaller than the mean heating degree days for cluster 2, which suggests that the homes in each cluster are located in different climate regions. Moreover, the mean unconditioned duct R-value is higher for cluster 1 than for cluster 2, while the mean duct number of returns is bigger for cluster 2. Homes in cluster 2 also have larger conditioned floor areas on average than homes in cluster 1. The most significant difference between clusters is the mean number of heating degree days. The homes in cluster 1 are located in North Carolina, Oregon, and Texas, and the homes in cluster 2 are located in Minnesota, Oregon, and Wisconsin, shown in Table 8.4.

Table 8.3: Mean, median, and standard deviation of a few numeric variables for both clusters.

	Heating Degree Days			**Fixed Heating Capacity**			**Unconditioned Duct R-value**		
Grouping	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Cluster 1	3418	3481	877.04	46.58	26	31.20	5.08	6	2.44
Cluster 2	7576	7718	298.84	77.81	70	27.86	0.66	0	2.96
All Homes	5408	4340	2182.40	61.52	70	33.50	3.03	0	3.48

	Conditioned Floor Area			**South Window Area**			**Duct Number of Returns**		
Grouping	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Cluster 1	1525	1234	670.44	40.28	27.60	39.60	1.19	1	0.53
Cluster 2	2915	2869	951.99	74.78	62.80	56.68	3.51	4	2.11
All Homes	2190	1995	1073.23	56.79	44.50	51.47	2.27	1	1.89

Table 8.4: Number of homes in each state for both clusters.

State	**Cluster 1**	**State**	**Cluster 2**
NC	363	MN	89
OR	94	OR	22
TX	63	WI	366

While the similarities of each cluster were intended to be used to define residential archetypes, given that the data is not a random sample, residential archetypes cannot accurately be defined for all homes in the United States. The homes in this data set can, however, be categorized by climate region. Although Oregon is considered to have a marine climate, the homes in cluster 1 will be labeled as "warm climate" homes, and the homes in cluster 2 will be labeled as "cold climate" homes. The location of the homes at the city level can be seen in Figure 8.1. Cities with red circles indicate that the homes are located in cluster 1, while cities with blue circles indicate that the homes are located in cluster 2. Moreover, the more homes located in a particular city, the darker the color. While the clusters will no longer be used to define residential archetypes, they still provide information regarding the accuracy of the BEopt simulation process and where improvements may be made for the two types of homes.

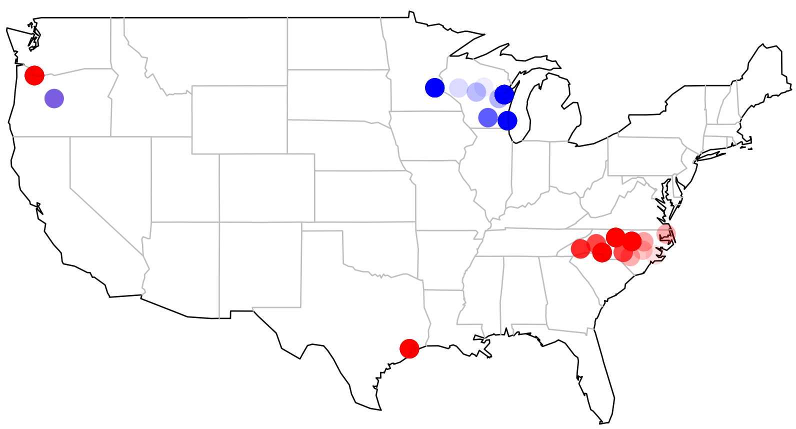


Figure 8.1: Plot of the location of homes at city level for both clusters. Red cities indicate homes are in cluster 1, while blue cities indicate homes are in cluster 2. The blue city in Oregon contains homes in both clusters.

CHAPTER 9

MULTIPLE LINEAR REGRESSION

Regression analysis is a technique used to assess the relationship of a response variable to a set of predictors. The goal of regression analysis is to fit an empirical model that explains the effects of the predictors, or explanatory variables, on the response [19]. Regression models have a variety of uses, which include predictive capabilities. They are defined as follows: Define the $n \times 1$ vector of responses, $\mathbf{y}$, the $n \times (p+1)$ matrix of predictors, $\mathbf{X}$, the $(p+1) \times 1$ vector of unknown regression parameters, $\boldsymbol{\beta}$, and the $n \times 1$ vector of random errors, $\mathbf{e}$, as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \text{and } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

The multiple linear regression model can then be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{9.1}$$

and the assumptions that must be satisfied for this model to be valid are [12]:

1. The observations are independent.
2. There is a linear relationship between the response and predictors.
3. The variance is constant.
4. The errors are normally distributed.

For this project, models will be created for several different responses, $\mathbf{y}$. These responses will be

- measured electricity,
- delta electricity (RBES predicted – measured),
- measured natural gas, and

• delta natural gas (RBES predicted – measured).

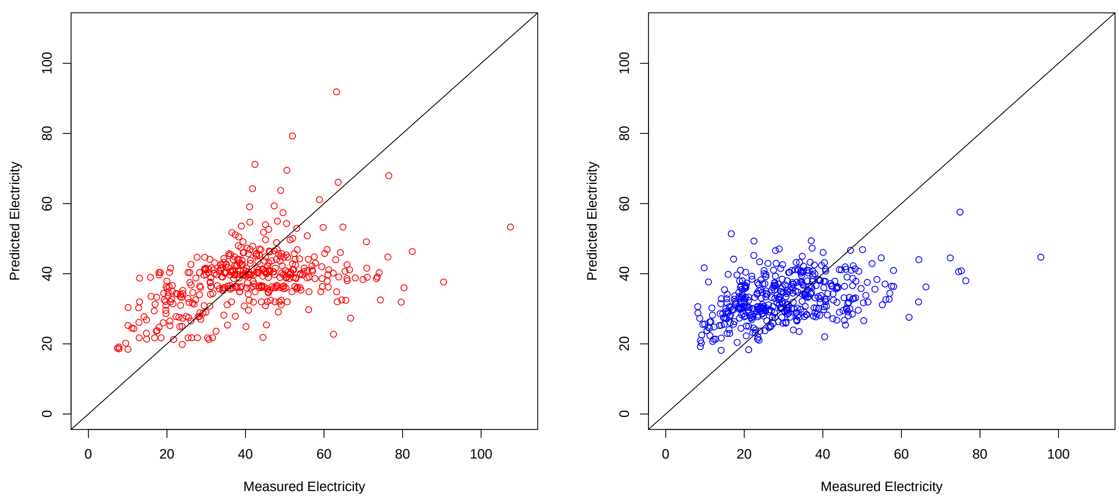
The regressions with measured responses (electricity and natural gas) will give an indication as to which predictors are important in modeling electricity and natural gas consumption for all homes and for each cluster. The regressions with delta responses (again, electricity and natural gas) will give an indication as to how well the building energy simulation process predicts electricity and natural gas consumption for all homes and for each cluster. The regressions with delta responses will also point to the predictors that are significant in explaining the difference between predicted and measured energy consumption. This indicates that the simulation process may need revisions to better model these particular predictors. Table 9.1 shows the mean, median, and standard deviations of the four dependent variables for both clusters and all homes. The mean measured electricity consumption for the cold climate homes (cluster 2) is smaller than for the warm climate homes (cluster 1), with a smaller standard deviation as well. However, the mean measured natural gas consumption for cold climate homes is larger than the mean measured natural gas consumption for warm climate homes. Additionally, electricity consumption is underpredicted by 4.9% on average for warm climate homes and overpredicted by 8% on average for cold climate homes. Natural gas consumption is overpredicted by 8.6% on average for warm climate homes and 12.7% on average for cold climate homes. When all homes are considered, electricity consumption is overpredicted by 0.8% on average, and natural gas consumption is overpredicted by 10.1% on average.

Figure 9.1 shows the predicted versus measured electricity consumption for each cluster, and Figure 9.2 shows the predicted versus measured natural gas consumption for each cluster. The observations that fall below the line of perfect agreement are homes that the simulation process underpredicts energy consumption, and the observations above the line are homes that are overpredicted by the simulation process. The warm climate homes have more extreme differences between predicted and measured electricity consumption and less extreme differences for natural gas than the cold climate homes.

Table 9.1: Mean, median, and standard deviation of the responses for the groupings.

Grouping	Measured Electricity Mean	Median	SD	Delta Electricity Mean	Median	SD
Warm Climate	40.37	40.41	15.00	-1.96	-0.19	13.50
Cold Climate	30.63	29.22	13.03	2.46	4.23	11.63
All Homes	35.41	34.49	14.85	0.29	2.13	12.77

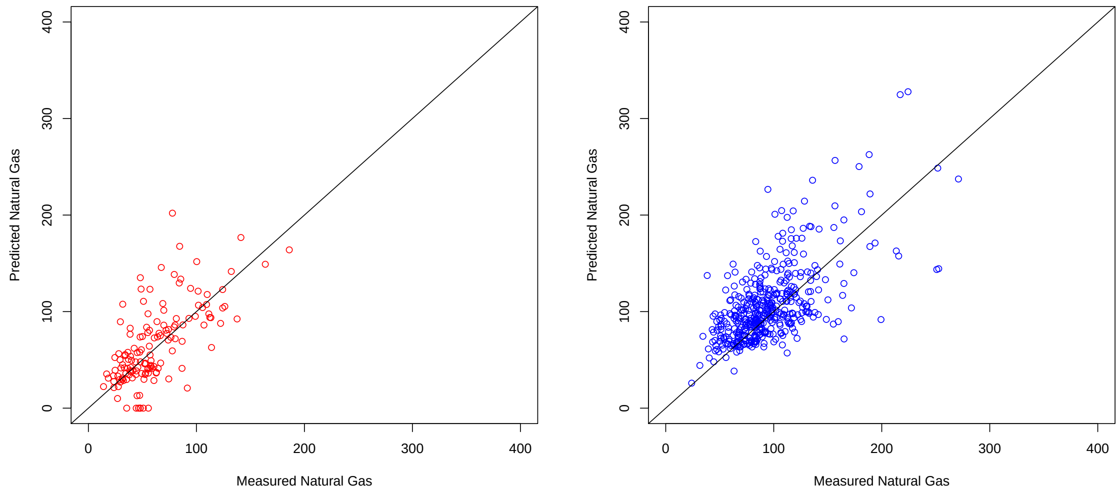
Grouping	Measured Natural Gas Mean	Median	SD	Delta Natural Gas Mean	Median	SD
Warm Climate	62.31	55.20	31.34	5.34	0.00	20.66
Cold Climate	95.02	87.88	35.48	12.05	9.13	33.38
All Homes	87.00	83.27	37.25	8.75	0.00	28.04



(a) Warm Climate

(b) Cold Climate

Figure 9.1: Plot of predicted versus measured electricity consumption for warm and cold climate homes with a reference line of perfect agreement.



(a) Warm Climate

(b) Cold Climate

Figure 9.2: Plot of predicted versus measured natural gas consumption for warm and cold climate homes with a reference line of perfect agreement.

In the remainder of this section, backward elimination for variable selection is described, and models are built and interpreted for all four responses using all homes and the subset of homes in each cluster. Then, the models are compared for each of the responses, and the variables that lead to overprediction and underprediction in BEopt are investigated.

9.1 Backward Elimination

Backward elimination is a variable selection technique based on the p-values of the predictor variables in a model. This method starts with all potential predictor variables in the model, and at each step, the variable with the largest p-value is removed. This process is repeated until all variables have a p-value significant for level α or until all variables have been removed from the model [12].

9.2 Regressions with All Homes

MLRs are performed using all homes ($n = 997$) in the FDR dataset for the electricity responses. For the natural gas responses, only homes with non-zero natural gas usage will be used in the regressions ($n = 709$). Backwards elimination is used as the variable selection method, where the reduced model will have coefficients that are all significant at level $\alpha = 0.05$. Assumptions of the model for the measured electricity response are checked in this section as an example, and the assumptions are checked for the remaining models in Appendix E. Moreover, the number of variables that are significant in each model is extensive, so only key variables will be discussed.

9.2.1 Measured Electricity Response

First, a multiple linear regression is performed for the measured electricity response. Table 9.2 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 11.70 on 710 degrees of freedom, and the adjusted R-squared is 0.3378. In this model, 273 observations are not used due to the presence of NAs.

Table 9.2: MLR results for all homes and the measured electricity response.

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-14.560	4.198	-3.468	0.000556
Number of Bedrooms	5.435	0.722	7.524	1.61e-13
Water Heater Energy Factor	26.420	4.463	5.920	5.02e-09
Net Above Grade Wall Area	0.005	0.001	4.346	1.59e-05
East Window Area	0.035	0.010	3.523	0.0005
Age of Home	-0.064	0.019	-3.426	0.0006
Clothes Dryer Fuel Type: Gas	-3.654	1.132	-3.227	0.0013
No Furnace	5.335	1.724	3.094	0.0021
Duct Number of Returns	-0.880	0.325	-2.711	0.0069
Interzonal Floor Cavity R-value	-0.107	0.040	-2.648	0.0083
Front Window U-value	7.220	3.004	2.404	0.0165
House Vent Fan Power	1.622	0.713	2.275	0.0232
Has Radiant Barrier	-5.708	2.525	-2.261	0.0241
Attic Area	0.002	0.0009	2.126	0.0339

The assumptions of this model are checked to determine whether or not this model is valid. Figure 9.3 shows some diagnostic plots. The observations in this dataset are independent, and from the plot of the standardized residuals versus fitted values, there is no distinct nonlinearity or non-constant variance. Given that electricity use is bounded by zero, small outliers are not expected, which is why there are more standardized residuals above zero than below. Furthermore, the errors appear to be approximately normally distributed as seen in the normal Q-Q plot. While the upper tail does not follow the Q-Q line, this is expected given that there is a lower bound but no upper bound to electricity consumption. There are also some points of high leverage, but all of the values of Cook's distance are less than one. Cook [36] suggests a cut-off value of one as an indication of highly influential observations. Since all of the values of Cook's distance are less than the cutoff, no outliers need to be removed.

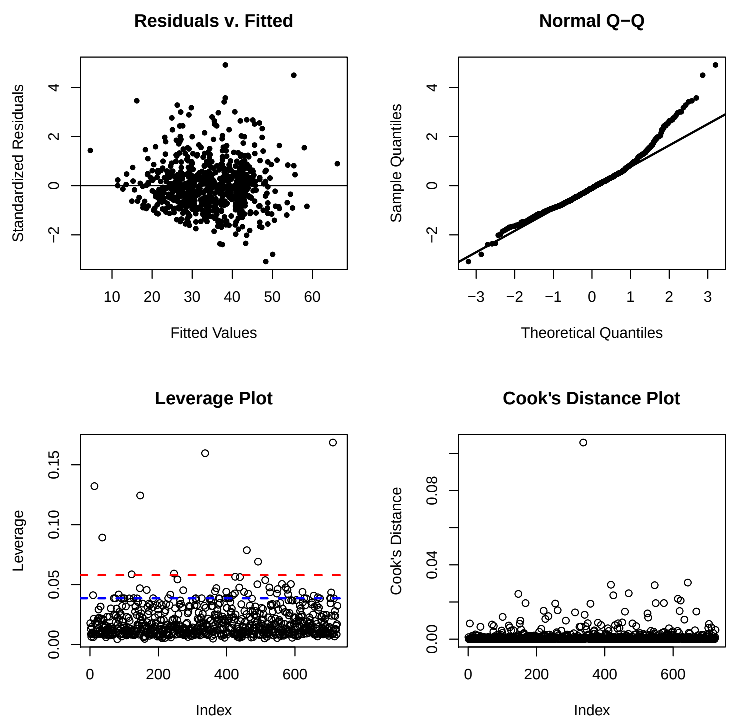


Figure 9.3: Diagnostic plots for the reduced multiple linear regression model with all homes and measured electricity as the response.

There are several variables in this model related to home geometry, including East Window Area, Attic Area, and Net Above Grade Wall Area. The coefficients for these variables are positive, indicating that an increase in any of these variables will increase electricity consumption when all other variables in the model are held constant. Typically, larger homes consume more electricity, so this agrees with what is expected. The most statistically significant variable in this model is Number of Bedrooms since its p-value is the smallest. Increasing Number of Bedrooms by one unit is predicted to increase average measured electricity consumption by 5.435 MBtu when all other variables are held constant. Furthermore, Has Radiant Barrier has a negative coefficient, which complies with what is expected from a building science perspective, since radiant barriers are designed to reduce the cooling load and thus reduce electricity consumption. A home with a radiant barrier is predicted to consume an average 5.708 MBtu less in electricity than if a radiant barrier was not present, given that all other variables in the model are held constant.

9.2.2 Delta Electricity Response

Next, a multiple linear regression is performed for the delta electricity response, where delta electricity is the difference between measured electricity consumption and the RBES predicted electricity consumption (predicted - measured). Table 9.3 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 11.71 on 824 degrees of freedom, and in this model, 157 observations are not used due to the presence of NAs. Furthermore, the adjusted R-squared is 0.1675.

Many of the significant variables in this model are also significant variables in the model for measured electricity, and these appear in bold in Table 9.3. Enclosed Crawlspace Perimeter, No Garage, East Window Area, and Conditioned Floor Area are all geometry related variables. These five variables are processed to be used in BEopt, since BEopt requires a 3-dimensional closed geometry, which indicates that information from these variables has potentially been lost in that process. Like in the model for measured electricity, the most statistically significant variable is Number of Bedrooms since its p-value is the smallest. Increasing Number of Bedrooms by one unit is predicted to decrease average delta electricity consumption by 2.874 MBtu when all other variables are held constant. This suggests that

Table 9.3: MLR results for all homes and the delta electricity response.

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	19.520	4.164	4.687	3.24e-06
Number of Bedrooms	-2.874	0.643	-4.471	8.88e-06
Age of Home	0.076	0.018	4.152	3.64e-05
Enclosed Crawlspace Perimeter	-0.031	0.008	-4.046	5.70e-05
AH Leak RA	61.320	19.060	3.217	0.0014
Duct Location: Garage	-22.360	6.953	-3.216	0.0014
Roof Color: Dark	7.922	2.617	3.027	0.0025
Left Window U-value	-4.819	1.624	-2.967	0.0031
No Garage	-3.221	1.100	-2.929	0.0035
East Window Area	-0.027	0.009	-2.887	0.0040
House Vent Fan Power	-1.767	0.651	-2.716	0.0068
Unconditioned Duct R-value	0.414	0.153	2.713	0.0068
Ventilation Type: Balanced	4.544	1.826	2.488	0.0130
Dishwasher EF	-16.520	6.703	-2.464	0.0140
Conditioned Floor Area	0.001	0.0006	2.293	0.0221
No Furnace	-2.725	1.322	-2.061	0.0396

increasing the number of bedrooms contributes to underpredicting electricity consumption in BEopt. Additionally, there are five variables in this model that relate to ducts and ventilation, e.g. AH Leak RA, Ducts Located in the Garage, House Vent Van Power, Unconditioned Duct R-value, and Balanced Ventilation Type. Duct location and duct leakage are difficult to translate into BEopt inputs, so it is likely that the error is primarily introduced by the translation process.

9.2.3 Measured Natural Gas Response

A multiple linear regression is fit for the measured natural gas response. Table 9.4 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 23.95 on 558 degrees of freedom. In this model, 130 of the 709 observations are not used due to the presence of NAs. The adjusted R-squared is 0.5879, which is much higher than for the electricity response.

The majority of the variables in this model impact space heating, such as Heating Degree Days and the geometry related variables, which is expected given that space heating is a

Table 9.4: MLR results for all homes and the measured natural gas response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	-38.649	14.142	-2.733	0.006477
Heating Degree Days	0.009	0.0007	11.699	$<$ 2e-16
No Furnace	26.150	3.709	7.050	5.29e-12
Living Space ACH50	2.367	0.379	6.254	7.98e-10
East Window Area	0.109	0.022	5.065	5.55e-07
Water Heater Energy Factor	-57.472	11.616	-4.948	9.95e-07
Conditioned Floor Area	0.009	0.002	4.873	1.44e-06
Age of Home	-0.195	0.048	-4.073	5.31e-05
Fixed Heating Capacity	0.218	0.059	3.678	0.0003
Water Heater Tank Volume	0.605	0.168	3.596	0.0004
Fraction of ASHRAE	-13.125	3.683	-3.564	0.0004
Lighting Values Not Defaulted	-10.025	2.814	-3.563	0.0004
South Window Area	0.068	0.021	3.189	0.0015
Ventilation Type: Supply	26.822	9.009	2.977	0.0030
Number of Stories	7.580	2.740	2.767	0.0059
Use Ceiling Fan	-8.665	3.145	-2.755	0.0061
No Garage	14.214	5.581	2.547	0.0111
Above Grade Wall U-value	82.492	32.604	2.530	0.0117
Net Above Grade Wall Area	0.007	0.004	2.079	0.0381
Left Window Percent of Total	-51.196	24.900	-2.056	0.0402
Dishwasher EF	-33.392	16.907	-1.975	0.0488

major end use for natural gas. Heating Degree Days is the most statistically significant variable in this model since its p-value is the smallest. Given all other variables in the model are held constant, increasing Heating Degree Days by one unit is predicted to increase average measured natural gas consumption by 0.009 MBtu. One interesting and unexpected variable present in this model is Lighting Values Not Defaulted. Not defaulting lighting values in BEopt is expected to decrease average measured natural gas consumption by 10.025 MBtu when all other variables in the model are held constant. It is not obvious why this variable is included in this model since it relates to BEopt, so it may be an indication of the completeness of the audits performed on the homes in the dataset, where homes that do not have defaulted lighting values have more complete audits. Another interesting variable present in this model is the variable that denotes the use of a ceiling fan. Given all other variables in the model are held constant, a home that uses ceiling fans is expected to consume 10.025 MBtu less natural gas on average than if the home did not use ceiling fans. While ceiling fans are powered by electricity, reversing the direction of ceiling fan blades in the winter is expected to reduce energy consumption related to heating, and thus decrease natural gas consumption.

9.2.4 Delta Natural Gas Response

Lastly, a multiple linear regression is performed for the response delta natural gas, where delta natural gas is the difference between measured natural gas consumption and the RBES predicted natural gas consumption. Table 9.4 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 23.06 on 540 degrees of freedom. In this model, 155 of the 709 observations are not used due to the presence of NAs, and the adjusted R-squared is 0.4458.

Many of the significant variables in this model are also significant variables in the model for measured natural gas, and these eight variables appear in bold in Table 9.5. There are six geometry related variables present, namely Above Grade Wall U-value, Cathedral Ceiling Area, East Window Area, Conditioned Floor Area, South Window Area, and Unfinished

Table 9.5: MLR results for all homes and the delta natural gas response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	-63.410	10.570	-6.001	3.59e-09
Above Grade Wall U-value	235.400	31.700	7.424	4.44e-13
Heating Degree Days	0.006	0.0008	7.284	1.16e-12
Cathedral Ceiling Area	0.024	0.003	6.820	2.44e-11
Duct Location: Unfinished Attic	19.390	3.351	5.785	1.23e-08
Living Space ACH50	1.661	0.360	4.610	5.03e-06
East Window Area	-0.076	0.020	-3.770	0.0002
Conditioned Floor Area	0.005	0.002	3.445	0.0006
South Window Area	-0.062	0.020	-3.125	0.0019
Water Heater Energy Factor	-30.990	10.340	-2.998	0.0028
Age of Home	0.132	0.045	2.953	0.0033
Unfinished Attic U-value	55.070	19.390	2.840	0.0047
Duct Number of Returns	1.756	0.655	2.683	0.0075
AH Leak RA	97.980	40.690	2.408	0.0164

Attic U-value. Since translating the geometry variables from the original data set into inputs to be used in BEopt is a difficult process that alters the physical representation of a home, information from these variables is likely lost in the process. Above Grade Wall U-value is the most statistically significant variable in this model since its p-value is the smallest. Increasing Above Grade Wall U-value by one unit is expected to increase the average delta natural gas 235.4 MBtu, given all other variables in the model are held constant. Above Grade Wall U-value is an intermediate variable calculated in the translation scripts and is not used directly in BEopt, so it is possible that the variables created from Above Grade Wall U-value for use in BEopt mishandle the information present in this variable. Similar to the model for response delta electricity, there are several variables in this model that relate to ducts and ventilation. Since duct and ventilation variables are significant in both delta models, it is probable that information contained in these variables is lost or misrepresented in the simulation process.

9.3 Cluster Regressions

Next, multiple linear regressions are performed on the 520 warm climate homes (cluster 1) and the 477 cold climate homes (cluster 2) for the dependent variables: measured electricity, delta electricity, measured natural gas, and delta natural gas. Backwards elimination is used as the variable selection method, where the reduced models have coefficients that are all significant at the $\alpha = 0.05$ level. Diagnostic plots to check the assumptions are given in Appendix E for both clusters and all four responses.

9.3.1 Measured Electricity Response

First, a regression analysis is performed on the warm climate cluster, where the reduced model has coefficients that are significant for $\alpha = 0.05$. Table 9.6 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 10.99 on 404 degrees of freedom. In this model, 99 of the 520 observations are not used due to the presence of NAs. The adjusted R-squared is 0.4567.

Table 9.6: MLR results for the warm climate cluster and the measured electricity response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	-19.106	7.816	-2.444	0.014936
Net Above Grade Wall Area	0.011	0.002	5.988	4.70e-09
Number of Bedrooms	5.615	0.982	5.720	2.07e-08
No Furnace	10.942	2.550	4.290	2.24e-05
Water Heater Tank Volume	0.459	0.112	4.081	5.41e-05
Front Window U-value	17.879	4.906	3.644	0.0003
Water Heater Energy Factor	21.469	6.767	3.173	0.0016
Age of Home	-0.106	0.034	-3.133	0.0019
Right Window Percent of Total	-24.463	8.640	-2.831	0.0049
South Window Area	-0.047	0.018	-2.611	0.0094
Use Ceiling Fan	-7.225	2.789	-2.590	0.0099
Lighting Values Not Defaulted	-4.813	1.905	-2.527	0.0119
Right Window U-value	-7.437	3.080	-2.415	0.0162
Ventilation Type: Supply	-3.267	1.394	-2.344	0.0196
Unconditioned Duct R-value	-0.683	0.298	-2.293	0.0224
Back Window U-value	-15.508	6.906	-2.246	0.0253
Interzonal Floor Cavity R-value	-0.250	0.121	-2.060	0.0401

The variables present in this model are similar to the variables present in the model for measured electricity with all homes. Net Above Grade Wall Area is the most statistically significant variable present in this model since its p-value is the smallest. Increasing Net Above Grade Wall Area by one unit is predicted to increase average electricity consumption by 0.011 MBtu given all other variables in the model are held constant. In modeling measured electricity consumption for homes in the warm climate cluster, there are five window variables present. These variables are related to window area, window U-values, and the percentage of total window area. The coefficients for most of these variables are negative, except for the coefficient for Front Window U-value. Thus, when all other variables are held constant, increasing Front Window U-value by one unit is expected to increase average measured electricity consumption by 17.879 MBtu, and increasing any of the other window variables is predicted to decrease average measured electricity consumption. The U-value of a window is an indication of the insulation present, where a smaller value corresponds to more resistance to heat transfer. Therefore, it is expected that decreasing window U-values decreases electricity consumption, so it is unclear why the coefficients for Right Window U-value and Back Window U-value are negative. There is an inverse relationship between the use of a ceiling fan and electricity consumption. When all other variables in the model are held constant, a home that uses ceiling fans is predicted to consume an average of 7.225 MBtu less electricity than if the home did not use ceiling fans.

Next, a regression analysis is performed on the cold climate cluster. Table 9.7 shows the estimated coefficient, standard errors, t-values, and p-values for the reduced model with coefficients significant for $\alpha = 0.05$. The residual standard error for this model is 11.51 on 435 degrees of freedom. In this model, 31 of the 477 observations are not used due to the presence of NAs. Furthermore, the adjusted R-squared is 0.2203, which is smaller than the adjusted R-squared for the model for the warm climate cluster.

There are fewer significant variables present in the model for the cold climate homes than there are in the model for the warm climate homes. Like in the warm climate cluster, several

Table 9.7: MLR results for the cold climate cluster and the measured electricity response.

Variable	**Estimate**	**Std. Error**	**t-value**	p**-value**
(Intercept)	-6.988	4.751	-1.471	0.1421
East Window Area	0.040	0.011	3.590	0.0004
Number of Bedrooms	2.953	0.851	3.472	0.0006
Age of Home	-0.072	0.021	-3.368	0.0008
South Window Area	0.036	0.011	3.310	0.0010
Front Window U-value	14.053	4.593	3.060	0.0024
Conditioned Floor Area	0.003	0.001	3.002	0.0028
Azimuth: 0	4.238	1.440	2.942	0.0034
Front Window Percent of Total	16.701	6.554	2.548	0.0112
West Window Area	0.030	0.012	2.492	0.0131
Living Space ACH50	0.447	0.191	2.347	0.0194

window related variables appear in this model. East Window Area is the most statistically significant variable in this model since its p-value is the smallest. Electricity consumption is expected to increase an average of 0.040 MBtu when East Window Area increases by one unit, given all other variables in the model are held constant. The only variable with a negative coefficient in this model is Age of Home. So, when all other variables are held constant, increasing Age of Home one unit is predicted to decrease average electricity consumption 0.072 MBtu. Older homes typically use less electricity on average, since fewer outlets are present, so this agrees with what is expected.

The variables that are bold in Table 9.6 and Table 9.7 are variables that appear in both models for the measured electricity response. These variables are Number of Bedrooms, Age of Home, South Window Area, and Front Window U-value. The sign of the coefficient for South Window Area in the model for the warm climate homes is negative, while the sign of the coefficient for South Window Area in the model for the cold climate homes is positive. This is the only variable that appears in both models for the measured electricity response that has a coefficient that changes sign. So, increasing South Window Area is expected to decrease average electricity consumption for warm climate homes and increase average electricity consumption for cold climate homes. Window area is related to size of home, where larger homes typically have more window area, and an increase in the size of a home generally leads to an increase in electricity consumption. One possible explanation that may account for the decrease in average electricity consumption for warm climate homes is the effect that window area has on lighting. The warm climate homes are generally located further south than the cold climate homes, so they typically experience more natural sunlight (especially during the winter months). With that being the case, less lighting may be needed as south window area increases for these warm climate homes, which would reduce their electricity consumption.

While there are variables common to both models for the measured electricity response, there are also several variables that are unique to the model for one of the clusters. There are eight variables that are unique to the warm climate cluster model, and five variables that are unique to the cold climate cluster model. For the warm climate homes, these variables are: Water Heater Tank Volume, Right Window Percent of Total, Use Ceiling Fan, Lighting Values Not Defaulted, Right Window U-value, Supply Ventilation Type, Unconditioned Duct R-value, and Back Window U-value. For the cold climate homes, these variables are: Conditioned Floor Area, Azimuth of 0 Degrees, Front Window Percent of Total, and West Window Area. These variables indicate differences between the two types of homes in modeling electricity consumption, and many of these variables relate to the orientation of window characteristics. For example, Right Window Percent of Total is significant in modeling electricity consumption for warm climate homes, while Front Window Percent of Total is significant in modeling electricity consumption for cold climate homes.

9.3.2 Delta Electricity Response

Regressions within each cluster using delta electricity as the response are now performed. First, a regression analysis is performed on the warm climate cluster, where the reduced model has coefficients that are significant for $\alpha = 0.05$. Table 9.8 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 11.29 on 402 degrees of freedom. In this model, 101 of the

520 observations are not used due to the presence of NAs, and the adjusted R-squared is 0.2922.

Table 9.8: MLR results for the warm climate cluster and the delta electricity response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	38.883	8.506	4.571	6.46e-06
Number of Bedrooms	-4.567	0.929	-4.914	1.30e-06
Duct Location: Garage	-37.713	8.364	-4.509	8.55e-06
Above Grade Wall U-value	99.019	22.773	4.348	1.74e-05
Age of Home	0.171	0.040	4.261	2.54e-05
South Window Area	0.071	0.018	3.905	0.0001
Left Window U-value	-10.324	2.833	-3.645	0.0003
No Garage	-5.493	1.574	-3.491	0.0005
Unconditioned Duct R-value	1.056	0.324	3.265	0.0012
Water Heater Tank Volume	-0.377	0.121	-3.105	0.0020
Dishwasher EF	-31.187	10.099	-3.088	0.0022
Right Window U-value	7.879	2.672	2.948	0.0034
AH Leak RA	77.436	26.977	2.870	0.0043
Living Space ACH50	-0.622	0.273	-2.276	0.0234
Duct Number of Returns	-3.296	1.450	-2.273	0.0236
Duct Location: Unfinished Bsmt	-6.788	3.365	-2.017	0.0443
No Furnace	-3.174	1.583	-2.005	0.0457

Many of the significant variables in this model are also significant variables in the model for delta electricity using all homes. There are 9 variables that appear in both models: Number of Bedrooms, Ducts Located in the Garage, Age of Home, Left Window U-value, No Garage, Unconditioned Duct R-value, Dishwasher EF, AH Leak RA, and No Furnace. Number of Bedrooms is the most statistically significant variable present in this model since its p-value is the smallest. Increasing Number of Bedrooms by one unit is expected to decrease average delta electricity by 4.567 MBtu when all other variables in the model are held constant. In the RBES process, miscellaneous electric loads are calculated based on the number of bedrooms and floor area of a home, so the presence of Number of Bedrooms in this model could be a proxy for occupant behavior. Furthermore, there are variables present in this model that are from the intermediate steps in the translation scripts, so information for these variables has potentially been lost in that process. These variables include Above

Grade Wall U-value, South Window Area, and Dishwasher EF. Increasing Above Grade Wall U-value by one unit is expected to increase average delta electricity by 99.019 MBtu for warm climate homes, given all other variables in the model are held constant.

Next, a regression model is fit to the homes in the cold climate cluster. Table 9.9 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model with coefficients significant for $\alpha = 0.05$. The residual standard error for this model is 11.14 on 430 degrees of freedom. In this model, 37 of the 477 observations are not used due to the presence of NAs. Furthermore, the adjusted R-squared is 0.0824, which is smaller than the adjusted R-squared for the model for the warm climate cluster.

Table 9.9: MLR results for the cold climate cluster and the delta electricity response.

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-4.608	5.418	-0.850	0.3955
Age of Home	0.069	0.022	3.174	0.0016
East Window Area	-0.031	0.011	-2.934	0.0035
Azimuth: 0	-3.776	1.349	-2.800	0.0054
Ventilation Type: Balanced	5.561	2.003	2.777	0.0057
Roof Color: Dark	6.300	2.472	2.549	0.0112
Conditioned Floor Area	0.002	0.001	2.463	0.0142
Number of Bedrooms	-2.064	0.862	-2.394	0.0171
House Vent Fan Power	-1.411	0.603	-2.342	0.0197
Water Heater Energy Factor	17.414	7.593	2.293	0.0223

Many of the significant variables in this model are also significant variables in the model for delta electricity using all homes. There are seven variables that appear in both models: Age of Home, East Window Area, Balanced Ventilation Type, Dark Roof Color, Conditioned Floor Area, Number of Bedrooms, and House Vent Fan Power. The most statistically significant variable in this model is Age of Home since its p-value is the smallest. Increasing Age of Home by one unit is expected to increase delta electricity by 0.069 MBtu on average for cold climate homes. Age of Home is not used directly in simulating electricity consumption, so one possible explanation for its presence in the model is the effect of defaults on older homes. Particular characteristics, like AC set points, are defaulted in BEopt because no information on these characteristics is given in the FDR. These defaults are based on benchmark values given by the Building America House Simulation Protocols, and it is possible that they are not representative for older homes. For example, older homes are typically less energy efficient than newer homes, which could lead to a higher AC set point than the benchmark value since they would be more expensive to cool.

The variables that are bold in Table 9.8 and Table 9.9 are variables that appear in both models for the delta electricity response. These variables are Number of Bedrooms and Age of Home. Number of Bedrooms is the most significant variable in modeling delta electricity for warm climate homes, and Age of Home is the most significant variable in modeling delta electricity for cold climate homes.

9.3.3 Measured Natural Gas Response

A regression analysis is then performed on the warm climate cluster for the dependent variable measured natural gas, where the reduced model has coefficients that are significant for $\alpha = 0.05$. Of the 520 warm climate homes, only 144 have non-zero natural gas usage. Table 9.10 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 14.67 on 66 degrees of freedom. In this model, 59 of the 144 observations are not used due to the presence of NAs. Furthermore, the adjusted R-squared is 0.8007, which is larger than the adjusted R-squared for the model with all homes.

Most of the variables present in this model relate to space heating, which is the major end use for natural gas. The most statistically significant variable present in this model is Duct Number of Returns since its p-value is the smallest. Increasing Duct Number of Returns by one unit is expected to increase average natural gas consumption by 26.840 MBtu, given all other variables in the model are held constant. Furthermore, there are several window variables present. These variables are related to the percentage of total window area and window solar heat gain coefficients. For example, increasing Back Window SHGC by one

Table 9.10: MLR results for the warm climate cluster and the measured natural gas response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	-130.100	22.820	-5.703	3.00e-07
Duct Number of Returns	26.840	3.511	7.646	1.14e-10
Back Window SHGC	144.300	24.890	5.798	2.06e-07
Fixed Heating Capacity	-0.734	0.142	-5.153	2.52e-06
Right Window Percent of Total	119.900	29.960	4.000	0.0002
Front Window Percent of Total	118.800	32.690	3.634	0.0006
Attic Area	0.016	0.004	3.609	0.0006
Azimuth: 0	21.020	5.835	3.603	0.0006
Water Heater Tank Volume	1.024	0.291	3.516	0.0008
Unconditioned Duct R-value	-3.250	0.948	-3.428	0.0011
AH Leak RA	-165.200	55.480	-2.977	0.0041
Unfinished Attic U-value	-101.800	34.790	-2.926	0.0047
No Garage	-23.040	8.076	-2.853	0.0058
Clothes Dryer Fuel Type: Gas	-20.830	7.477	-2.786	0.0070
Number of Stories	13.680	5.288	2.587	0.0119
Net Above Grade Wall Area	0.010	0.004	2.342	0.0222
Left Window Percent of Total	78.480	34.750	2.259	0.0272
Fixed Cooling Capacity	5.338	2.368	2.254	0.0275
Above Grade Wall U-value	-143.500	65.210	-2.201	0.0312

unit is predicted to increase measured natural gas by 144.300 MBtu on average. Unlike the model for measured natural gas using all homes, the coefficient for Fixed Heating Capacity is negative in this model. When all other variables in the model are held constant, increasing Fixed Heating Capacity by one unit is expected to decrease average natural gas consumption by 0.734 MBtu for the warm climate homes.

Next, a regression analysis is performed on the cold climate cluster for the measured natural gas response. Of the 477 cold climate homes, 443 have non-zero natural gas usage. Table 9.11 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model with coefficients significant for $\alpha = 0.05$. The residual standard error for this model is 22.93 on 313 degrees of freedom. In this model, 117 of the 443 observations are not used due to the presence of NAs. The adjusted R-squared is 0.4936, which is smaller than the adjusted R-squared for the model with all homes.

Table 9.11: MLR results for cold climate cluster and the measured natural gas response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	2.788	18.277	0.153	0.878851
Fixed Heating Capacity	0.381	0.079	4.850	1.95e-06
Conditioned Floor Area	0.011	0.002	4.740	3.25e-06
Living Space ACH50	2.282	0.552	4.131	4.64e-05
Duct Number of Returns	-2.518	0.677	-3.718	0.0002
East Window Area	0.090	0.026	3.528	0.0005
Back Window U-value	48.454	14.085	3.440	0.0007
Water Heater Energy Factor	-77.123	24.191	-3.188	0.0016
Net Above Grade Wall Area	0.013	0.004	2.955	0.0034
South Window Area	0.067	0.025	2.726	0.0068
Number of Stories	9.140	3.582	2.552	0.0112
Fixed Cooling Capacity	-6.752	2.869	-2.353	0.0192
Water Heater Tank Volume	0.469	0.231	2.034	0.0428

In modeling measured natural gas consumption for homes in the cold climate cluster, half of the twelve variables present are geometry related. These variables are Conditioned Floor Area, East Window Area, Back Window U-value, Net Above Grade Wall Area, South Window Area, and Number of Stories. The coefficients for all the geometry related variables are positive, which means that when all other variables in this model are held constant, average measured natural gas increases as any of these variables increases. For example, a one unit increase in Conditioned Floor Area corresponds to an expected average increase of 0.011 MBtu in measured natural gas consumption when all other variables in the model are held constant. This is generally expected given that a larger home typically requires more heating system usage, which is the major driver of natural gas consumption. Furthermore, the most statistically significant variable in this model is Fixed Heating Capacity since its p-value is the smallest. Increasing Fixed Heating Capacity by one unit is predicted to increase average measured natural gas consumption by 0.381 MBtu for the cold climate homes, given all other variables present in the model are held constant.

The six variables that are bold in Table 9.10 and Table 9.11 are variables that appear in both models for the measured natural gas response. These variables are Duct Number of Returns, Fixed Heating Capacity, Water Heater Tank Volume, Number of Stories, Net Above Grade Wall Area, and Fixed Cooling Capacity. Duct Number of Returns, Fixed Heating Capacity, and Fixed Cooling Capacity have different signs in the two models. The sign of the coefficient for Duct Number of Returns in the model for the warm climate cluster is positive, while the sign of the coefficient is negative in the model for the cold climate cluster. Therefore, increasing Duct Number of Returns is expected to increase average natural gas consumption for warm climate homes and decrease average natural gas consumption for cold climate homes. Increasing Fixed Heating Capacity is predicted to decrease average natural gas consumption for warm climate homes and increase average natural gas consumption for cold climate homes, and the reverse is true for an increase in Fixed Cooling Capacity. It is unclear why the signs for Fixed Heating and Cooling Capacity change between models, since these variables relate to the rate of heat addition and removal per hour to a home.

While there are several variables common to modeling measured natural gas in each cluster, there are also many variables that are unique to the model for one of the clusters. There are ten variables that are unique to modeling natural gas consumption in warm climate homes, and four variables that are unique to modeling natural gas consumption in cold climate homes. For the warm climate homes, these variables are: Back Window SHGC, Right Window Percent of Total, Front Window Percent of Total, Azimuth of 0 Degrees, Unconditioned Duct R-value, AH Leak RA, Unfinished Attic U-value, No Garage, Left Window Percent of Total, and Above Grade Wall U-value. For the cold climate homes, these variables are: Conditioned Floor Area, Living Space ACH50, Back Window U-value, and South Window Area. These variables provide additional information on which predictors are influential in modeling natural gas consumption for each type of home, and many relate to different window characteristics. For example, while Back Window SHGC is significant in modeling natural gas consumption for warm climate homes, Back Window U-value is significant in modeling natural gas consumption for cold climate homes.

9.3.4 Delta Natural Gas Response

Regressions within each cluster using delta natural gas as the response are now performed. First, a regression analysis is performed on the warm climate cluster, where the reduced model has coefficients that are significant for $\alpha = 0.05$. Of the 520 warm climate homes, only 144 have non-zero natural gas usage. Table 9.12 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model. The residual standard error for this model is 22.28 on 136 degrees of freedom. In this model, 2 of the 144 observations are not used due to the presence of NAs. Furthermore, the adjusted R-squared is 0.4734.

There are only five variables in this model, and they all relate to space heating. Furthermore, the variables Living Space ACH50, Unfinished Attic U-value, Fixed Heating Capacity, and Cathedral Ceiling Area also appear in the model with all homes for the delta natural

Table 9.12: MLR results for the warm climate cluster and the delta natural gas response.

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	-54.292	6.579	-8.252	1.20e-13
Living Space ACH50	2.493	0.438	5.685	7.67e-08
Unfinished Attic U-value	131.649	28.820	4.568	1.09e-05
Fixed Heating Capacity	0.323	0.080	4.027	9.35e-05
Cathedral Ceiling Area	0.019	0.006	3.384	0.0009
Azimuth: 0	-11.020	4.917	-2.241	0.0266

gas response. While these variables all have positive coefficients, the coefficient for Azimuth of 0 degrees is negative. Azimuth of 0 degrees indicates the orientation of the front of a home is at 0 degrees and is calculated based on the distribution of window area. Given all other variables are constant, a home with an azimuth of 0 degrees is expected to have an average 11.020 MBtu less delta natural gas than if the home did not have an azimuth of 0 degrees. The most statistically significant variable in this model is Living Space ACH50 since its p-value is the smallest, where Living Space ACH50 indicates the air changes per hour at 50 Pascals of the home's living space. Increasing Living Space ACH50 by one unit is predicted to increase average delta natural gas by 2.493 MBtu when all other variables present in the model are held constant.

Next, a regression analysis is performed on the cold climate cluster for the delta natural gas response. Of the 477 cold climate homes, 443 have non-zero natural gas usage. Table 9.13 shows the estimated coefficients, standard errors, t-values, and p-values for the reduced model with coefficients significant for $\alpha = 0.05$. The residual standard error for this model is 23.32 on 400 degrees of freedom, and the adjusted R-squared is 0.4290. In this model, 29 of the 443 observations are not used due to the presence of NAs.

As with the model using all homes for the delta natural gas response, there are several geometry related variables present. These variables are Above Grade Wall U-value, Cathedral Ceiling Area, East Window Area, Conditioned Floor Area, South Window Area, Left Window Percent of Total, and Enclosed Crawlspace Perimeter. These variables are processed to be used in BEopt, so their presence in this model indicates that information is poten-

Table 9.13: MLR results for the cold climate cluster and the delta natural gas response.

Variable	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	-190.171	48.04	-3.959	8.92e-05
Above Grade Wall U-value	284.366	42.750	6.652	9.54e-11
Cathedral Ceiling Area	0.027	0.005	5.971	5.21e-09
Living Space ACH50	2.332	0.460	5.072	6.04e-07
Duct Location: Unfinished Attic	21.009	4.982	4.217	3.06e-05
East Window Area	-0.088	0.023	-3.854	0.0001
Conditioned Floor Area	0.006	0.002	3.373	0.0008
South Window Area	-0.073	0.022	-3.264	0.0012
Duct Number of Returns	2.204	0.688	3.202	0.0015
Heating Degree Days	0.016	0.006	2.713	0.0069
Use Ceiling Fan	8.244	3.471	2.375	0.0180
Dishwasher EF	42.121	18.300	2.302	0.0219
Left Window Percent of Total	38.722	16.94	2.286	0.0228
Enclosed Crawlspace Perimeter	0.088	0.044	1.990	0.0473

tially lost in the translation process for these variables for homes in the cold climate cluster. Above Grade Wall U-value is the most statistically significant variable since its p-value is the smallest. Increasing Above Grade Wall U-value by one unit is predicted to increase average delta natural gas by 284.400 MBtu for cold climate homes, given all other variables present in the model are held constant.

The two variables that are bold in Table 9.10 and Table 9.11 are variables that appear in both models for the delta natural gas response. These variables are Living Space ACH50 and Cathedral Ceiling Area. Living Space ACH50 is the most significant variable in modeling delta natural gas for warm climate homes, and Above Grade Wall U-value is the most significant variable in modeling delta natural gas for cold climate homes. Living Space ACH50 is also a significant variable in the model with all homes for the delta natural gas response. One possible explanation for the presence of Living Space ACH50 in all models for the delta natural gas response is due to an error that has since been corrected in the translation scripts. Given that the scripts are currently under development, improvements have been made since the data for this project was generated. One such improvement was related to leakage area, which is calculated using Living Space ACH50.

9.4 Comparison of Models

The models created for all homes and for each cluster are now compared. Table 9.14 is a table of the adjusted R-squared values from the models for all responses. The warm climate cluster has the largest adjusted R-squared values for all responses, whereas the adjusted R-squared values for the cold climate cluster are always the smallest. While the highest adjusted R-squared value for natural gas is for the warm climate cluster, the number of homes used in this model is also the smallest.

Table 9.14: Table of Adjusted R-squared values for all responses and the number of homes, n, present in each regression.

	Measured Electricity		**Delta Electricity**	
Grouping	n	Adj. R^2	n	Adj. R^2
All Homes	724	33.8 %	840	16.8 %
Warm Climate	421	45.7 %	419	29.2 %
Cold Climate	446	22.0 %	440	8.2 %

	Measured Natural Gas		**Delta Natural Gas**	
Grouping	n	Adj. R^2	n	Adj. R^2
All Homes	579	58.8 %	554	44.6 %
Warm Climate	85	80.1 %	142	47.3 %
Cold Climate	326	49.4 %	414	42.9 %

There are several variables common to all of the models (overall, warm climate cluster, and cold climate cluster) for a given response. For the measured electricity response, these variables are Number of Bedrooms, Age of Home, and Front Window U-value. For the delta electricity response, only Number of Bedrooms and Age of Home appear in all models. For the measured natural gas response, Fixed Heating Capacity, Water Heater Tank Volume, Number of Stories, and Net Above Grade Wall Area are common to all models, and the only variables that appear in all models for delta natural gas are Living Space ACH50 and Cathedral Ceiling Area. These variables are significant in modeling the responses regardless of the type of home.

The variables in the delta electricity models are explored and interpreted. Table 9.15 lists the variables that lead to an increase or decrease in average delta electricity consumption in the models for all homes, the warm climate cluster, and the cold climate cluster. The variables in Table 9.15 are listed in order of increasing p-value, which indicates their statistical significance in the model since a smaller p-value indicates the variable is more significant. When referring to these variables, the terms "overpredicting" and "underpredicting" are used in a relative sense. Variables with a positive estimated coefficient lead in the direction of overpredicting electricity consumption, but not necessarily a positive delta electricity, and variables with a negative estimated coefficient lead in the direction of underpredicting electricity consumption, but not always a negative delta electricity. For example, Age of Home is a variable that contributes to overpredicting electricity for the model with all homes since its estimated coefficient is positive, but a one unit increase in Age of Home could still result in a negative average delta electricity. In the model for all homes, Ducts Located in the Garage is a variable that contributes to underpredicting electricity, so a home with ducts that are located in the garage is expected to have 22.36 MBtu less delta electricity on average than if ducts were not located in the garage. This means that for homes with ducts that are located in the garage, the average RBES predicted electricity consumption is less than for homes that do not have ducts in the garage.

Next, the variables present in the delta natural gas models are investigated. Table 9.16 lists the variables that lead to an increase or decrease in average delta natural gas consumption in the models for all homes, the warm climate cluster, and the cold climate cluster, and variables are listed in order of increasing p-value. In the model with all homes, Above Grade Wall U-value is a variable that contributes to overpredicting natural gas, and a one unit increase in this variable is expected to increase average delta natural gas by 235.4 MBtu. Furthermore, East Window Area contributes to underpredicting natural gas in this model, and a one unit increase in East Window Area is predicted to decrease average delta natural gas by 0.076 MBtu. In the model for warm climate homes, Azimuth of 0 degrees is the only variable that leads to underpredicting natural gas, so a home with an azimuth of 0 degrees is expected to have 11.02 MBtu less delta natural gas on average than if the home did not

Table 9.15: List of variables that lead to increasing and decreasing delta electricity in the model for all homes, the warm climate cluster, and the cold climate cluster and their absolute estimated coefficients.

Increasing Delta Electricity		**Decreasing Delta Electricity**	
Variables	**|Est.|**	**Variables**	**|Est.|**
Model with All Homes			
Age of Home	0.076	Number of Bedrooms	2.874
AH Leak RA	61.320	Enclosed Crawlspace Perimeter	0.031
Roof Color: Dark	7.922	Duct Location: Garage	22.360
Unconditioned Duct R-value	0.414	Left Window U-value	4.819
Ventilation Type: Balanced	4.544	No Garage	3.221
Conditioned Floor Area	0.001	East Window Area	0.027
		House Vent Fan Power	1.767
		Dishwasher EF	16.520
		No Furnace	2.715
Model for Warm Climate Cluster			
Above Grade Wall U-value	99.019	Number of Bedrooms	4.567
Age of Home	0.171	Duct Location: Garage	37.713
South Window Area	0.071	Left Window U-value	10.324
Unconditioned Duct R-value	1.056	No Garage	5.493
Right Window U-value	7.879	Water Heater Tank Volume	0.377
AH Leak RA	77.436	Dishwasher EF	31.187
		Living Space ACH50	0.622
		Duct Number of Returns	3.296
		Duct Location: Unfinished Bsmt	6.788
		No Furnace	3.174
Model for Cold Climate Cluster			
Age of Home	0.069	East Window Area	0.031
Ventilation Type: Balanced	5.561	Azimuth: 0	3.776
Roof Color: Dark	6.300	Number of Bedrooms	2.064
Conditioned Floor Area	0.002	House Vent Fan Power	1.411
Water Heater Energy Factor	17.414		

have an azimuth of 0 degrees. This means that for homes with an azimuth of 0 degrees, the average RBES predicted natural gas consumption is less than for homes that do not have an azimuth of 0 degrees.

There are several variables that appear in the model for the warm climate cluster or the model for the cold climate cluster that do not appear in the model with all homes for both of the delta responses. There are nine variables that appear in the model for the warm or cold climate cluster but not the model with all homes for delta electricity. The seven variables from the warm climate cluster are Above Grade Wall U-value, South Window Area, Water Heater Tank Volume, Right Window U-value, Living Space ACH50, Duct Number of Returns, and Ducts Located in the Unfinished Basement. The two variables from the cold climate are Azimuth of 0 Degrees and Water Heater Energy Factor. These variables provide additional information as to where improvements may be made in the RBES process for both groups of homes when simulating electricity consumption. For example, an Azimuth of 0 degrees is significant in explaining the difference between RBES predicted and measured electricity for cold climate homes but not for warm climate homes.

Like in the models for delta electricity, there are several variables that appear in either of the cluster models but not the model with all homes for the delta natural gas response. There are two variables from the model for the warm climate homes and four variables from the model for the cold climate homes. The two variables from the warm climate cluster are Fixed Heating Capacity and Azimuth of 0 Degrees. The four variables from the cold climate cluster are Use Ceiling Fan, Dishwasher EF, Left Window Percent of Total, and Enclosed Crawlspace Perimeter. These variables provide additional information as to where improvements may be made in the RBES process for both groups of homes when simulating natural gas consumption. For example, Fixed Heating Capacity is significant in explaining the difference between RBES predicted and measured natural gas for warm climate homes but not for cold climate homes. These results indicate that while similar variables appear in all models for each of the delta responses, when the homes are broken into clusters, additional

Table 9.16: List of variables that lead to increasing and decreasing delta natural gas consumption in the models for all homes, the warm climate cluster, and the cold climate cluster and their absolute estimated coefficients.

Increasing Delta Natural Gas Variables	**|Est.|**	**Decreasing Delta Natural Gas Variables**	**|Est.|**
Model with All Homes			
Above Grade Wall U-value	235.400	East Window Area	0.076
Heating Degree Days	0.006	South Window Area	0.062
Cathedral Ceiling Area	0.024	Water Heater Energy Factor	30.990
Duct Location: Unfinished Attic	19.390		
Living Space ACH50	1.661		
Conditioned Floor Area	0.005		
Age of Home	0.132		
Unfinished Attic U-value	55.070		
Duct Number of Returns	1.756		
AH Leak RA	97.980		
Model for Warm Climate Cluster			
Living Space ACH50	2.493	Azimuth: 0	11.020
Unfinished Attic U-value	131.649		
Fixed Heating Capacity	0.323		
Cathedral Ceiling Area	0.019		
Model for Cold Climate Cluster			
Above Grade Wall U-value	284.400	East Window Area	0.088
Cathedral Ceiling Area	0.027	South Window Area	0.073
Living Space ACH50	2.332		
Duct Location: Unfinished Attic	21.010		
Conditioned Floor Area	0.006		
Duct Number of Returns	2.204		
Heating Degree Days	0.016		
Use Ceiling Fan	8.244		
Dishwasher EF	42.120		
Left Window Percent of Total	38.720		
Enclosed Crawlspace Perimeter	0.088		

information is gained on which areas may need improvement in the simulation process for homes of these particular types.

CHAPTER 10

COMPARISON OF METHODS

A common technique used to determine the predictive capabilities of a method is cross validation. Cross validation predicts the value of a reserved observation by repeating the entire modeling process on the remainder of data (clustering and model fitting) and comparing the prediction made for the observation left out with its actual observed value. One type of cross validation, known as leave-one-out cross validation, removes a single observation at a time, repeating for all observations. While this method is computationally expensive, it is commonly used when the sample size is small [37].

Given that the data set used is relatively small ($n = 997$), leave-one-out cross validation is employed to compare the predictive capabilities of the methods used. Since clustering is performed using weighted Euclidean distances, it is assumed that removing one home does not significantly affect the correlation matrix. Therefore, the principal components will not change, and new weighted distances do not need to be calculated. It is also assumed that the number of clusters is not affected by the removal of one home. The basic steps for this application of leave-one-out cross validation are:

1. Remove observation i, where $i = 1, 2, ..., n$.
2. Cluster the remaining $n - 1$ homes into 2 clusters using Ward's method.
3. Calculate the distance between observation i and the mean of each cluster.
4. Assign observation i to the cluster, C, with the smallest distance.
5. Fit a model using observations in C (not including observation i).
6. Calculate the predicted value, $\hat{y}_{-i}$, for obs i.

Once the predicted values are computed for all observations, the cross validation mean-squared error (MSE) is calculated as follows:

$$CV_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2, \tag{10.1}$$

where y_i is the observed value for observation i.

In 48 cases, one or two homes shift from the warm climate cluster to the cold climate cluster. In the case when home #858 is removed, 11 homes shift from the warm to the cold climate cluster. However, there are no cases where homes shift from the cold climate cluster to the warm climate cluster when a single observation is removed. Furthermore, only 2 homes were not assigned to their original cluster. These homes are #460 and #627, and both are from the warm climate cluster but were assigned to the cold climate cluster in step 4. The cross validation MSE estimates for all homes, both clusters, the warm climate cluster, and the cold climate cluster are given in Table 10.1. The value for "Both Clusters" is a weighted average of the cross validation MSE estimates for the warm and cold climate clusters, weighted based on the number of observations in each cluster. The cross-validation MSE for the method using all homes for the regression (and no clustering) is generally smaller than the MSE for the methods using clustering, except for the delta electricity response. Moreover, in all cases except measured electricity, the MSE for the warm climate cluster is significantly larger than the MSE for the cold climate cluster. These results indicate that overall, the method that does not involve clustering and uses all homes in the regression has better predictive capabilities than when the homes are broken into clusters. However, there are occasions where individual clusters perform better than when all homes are used. In the cases of delta electricity and measured natural gas, the models created when homes are assigned to the cold climate cluster have the best predictive capabilities.

Since it was expected that the cross validation MSE values would be smaller for the clusters, the cluster cross validation squared errors (SEs) are investigated. The histogram for the cluster cross validation SE values for measured electricity is shown in Figure 10.1. The distribution is skewed-right, and 19 SE values are greater than 1000 and are considered to be unusual. The observations with these high squared errors have measured electricity

Table 10.1: Table of cross validation MSE values for all responses.

Grouping	Measured Electricity	Delta Electricity	Measured Natural Gas	Delta Natural Gas
All Homes	**148.53**	151.99	**679.93**	**595.37**
Both Clusters	151.41	**148.51**	684.65	683.86
Warm Climate	**149.47**	163.52	817.26	800.47
Cold Climate	153.22	**134.29**	**650.58**	**644.62**

values that fall in the tails of the distribution, shown in Figure 10.2. Two of the 19 homes have measured electricity values below 16 MBtu, and the rest have a measured electricity consumption greater than 55 MBtu. The histogram for the cluster cross validation SE values for delta electricity is shown in Figure 10.3, and shows that the distribution is skewed-right. Twenty SE values are greater than 1000 and are considered to be unusual. Most of the observations with these high squared errors have delta electricity values that fall in the tails of the distribution, shown in Figure 10.4. The observations either have delta electricity values less than -24 MBtu or greater than 11 MBtu.

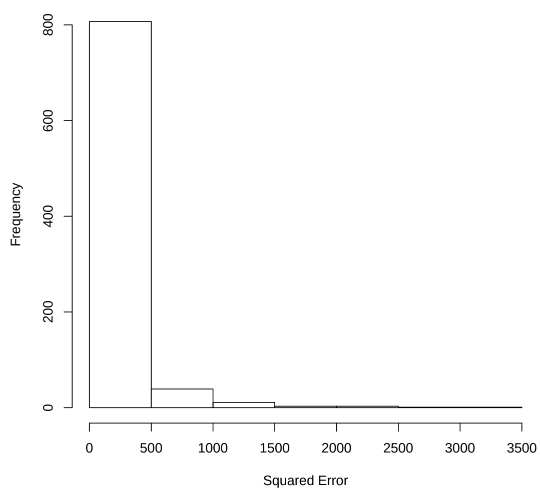


Figure 10.1: Histogram of cluster cross validation SE values for measured electricity.

Next, the cluster cross validation SE values for the natural gas responses are explored. The histogram for the cluster cross validation SE values for measured natural gas is shown in Figure 10.5. The distribution is skewed-right, and eight SE values are greater than 5000 and

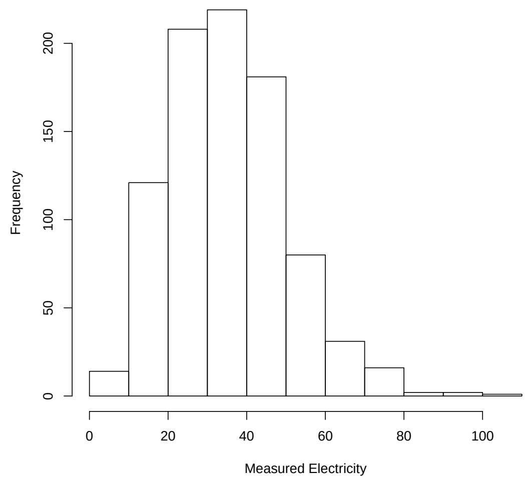


Figure 10.2: Histogram of measured electricity.

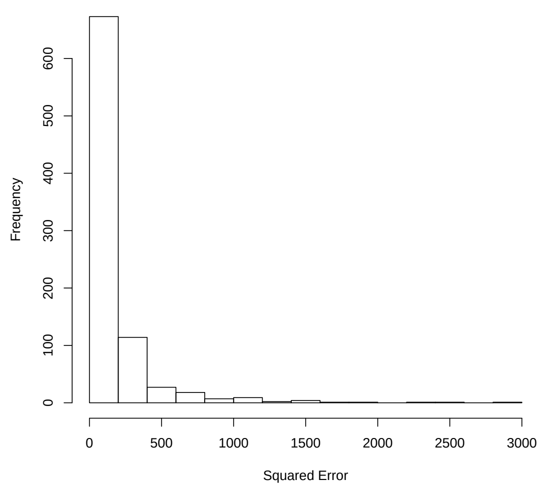


Figure 10.3: Histogram of cluster cross validation SE values for delta electricity.

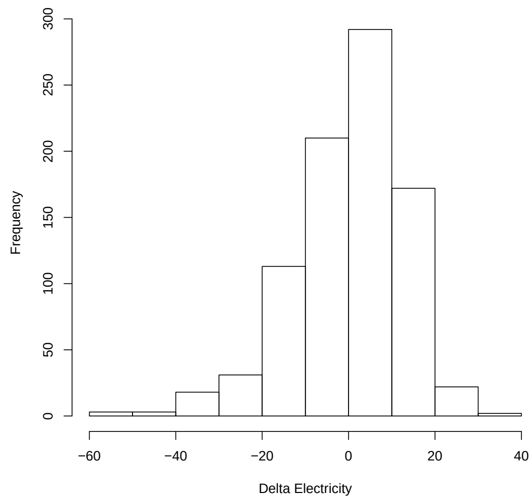


Figure 10.4: Histogram of delta electricity.

are considered to be unusual. Most of the observations with these high SEs have measured natural gas values that fall in the right tail of the distribution, shown in Figure 10.6. One observation has a measured natural gas consumption of 77.8 MBtu, but the rest have a measured natural gas consumption greater than 163 MBtu. The histogram for the cluster cross validation SEs for delta natural gas is shown in Figure 10.7. The distribution is skewed-right, and the eleven SE values that are greater than 5000 are considered to be unusual. Most of the observations with these high squared errors have delta natural gas values that fall in the tails of the distribution, shown in Figure 10.8. Two observations have delta natural gas values that are not in the tails of the distribution, but the rest have values that are either less than -44 MBtu or greater than 85 MBtu.

The majority of observations with unusual cross validation squared errors have values in the tails of the response distributions, so it appears as though a small number of homes with unusual consumption values are fit better with the overall model than the cluster models. Therefore, the largest 5% of cross validation squared errors are removed, and the cross validation MSE is recomputed for all four responses. Table 10.2 gives the cross validation MSE estimates for all homes, both clusters, the warm climate cluster, and the cold climate

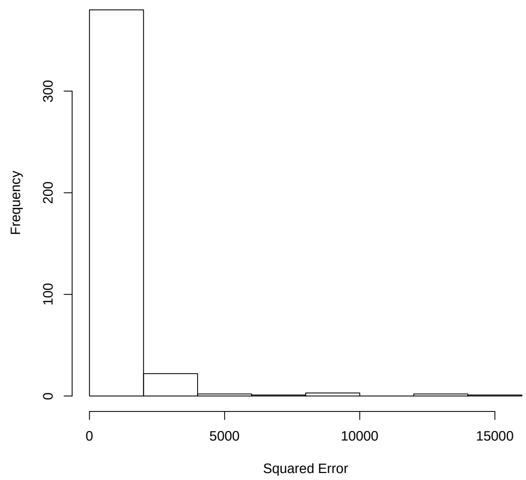


Figure 10.5: Histogram of cluster cross validation SE values for measured natural gas.

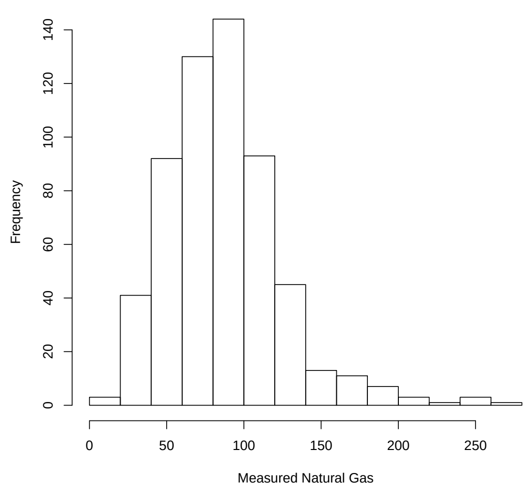


Figure 10.6: Histogram of measured natural gas.

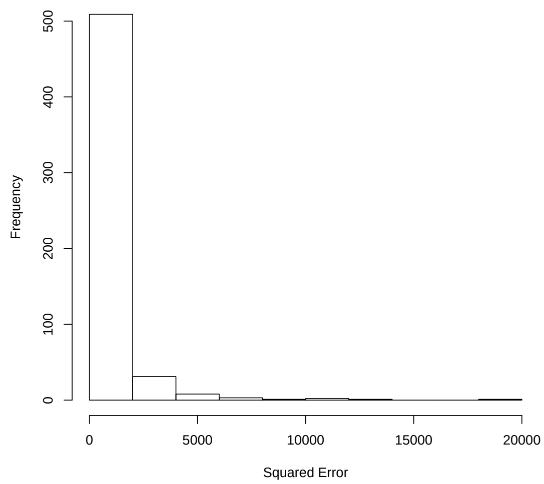


Figure 10.7: Histogram of cluster cross validation SE values for delta natural gas.

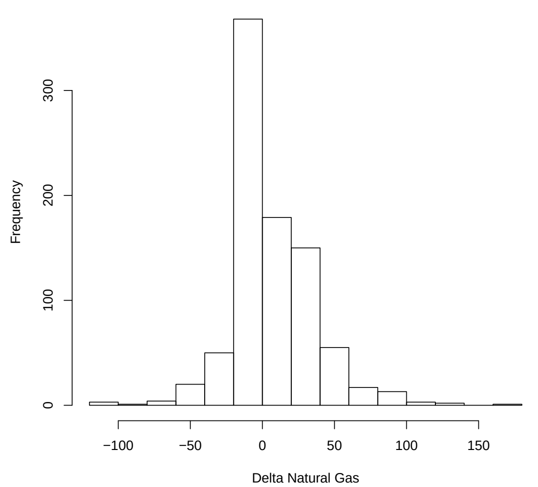


Figure 10.8: Histogram of delta natural gas.

cluster with the largest 5% of the squared errors removed, where "Both Clusters" denotes the weighted average of the warm and cold climate clusters. Measured electricity is generally more difficult to model than natural gas due to the variability in miscellaneous electric loads, so it is not surprising that in this case, the model with all homes has the smallest cross validation MSE. However, for all other responses, the weighted average cross validation MSE for the warm and cold climate clusters is smaller than the cross validation MSE for all homes. Therefore, after removing the unusually large cross validation squared errors, breaking the homes into clusters generally creates models with better predictive capabilities than when using all homes.

Table 10.2: Table of cross validation MSE values for all responses with the largest 5% of SE values removed.

Grouping	**Measured Electricity**	**Delta Electricity**	**Measured Natural Gas**	**Delta Natural Gas**
All Homes	**91.79**	99.64	431.05	470.63
Both Clusters	99.22	**97.59**	**398.51**	**408.68**
Warm Climate	**94.68**	102.85	431.11	**329.79**
Cold Climate	103.43	**92.69**	**390.34**	433.47

Comparing mean squared errors is just one way to compare the methods used. Table 10.3 is a table of the average adjusted R-squared for the models created without observation i in the leave-one-out cross validation, where "Both Clusters" denotes the weighted average adjusted R-squared for the warm and cold climate clusters. The average adjusted R-squared value for both clusters is smaller than the adjusted R-squared value for all homes for both measured responses. However, the adjusted R-squared value for both clusters is larger than the adjusted R-squared value for all homes for both delta responses. Therefore, breaking the homes into clusters gives a better fitting model for the difference between measured energy consumption and RBES predicted energy consumption than the model with all homes.

Using the results in Table 10.2 and Table 10.3, it appears as though breaking the homes into clusters generally creates models that have better predictive capabilities than the models using all homes and also fits the data better for certain responses.

Table 10.3: Table of average adjusted R-squared values for models created when the i^{th} observation is removed.

Grouping	**Measured Electricity**	**Delta Electricity**	**Measured Natural Gas**	**Delta Natural Gas**
All Homes	**33.7 %**	16.7 %	**58.7 %**	44.6 %
Both Clusters	33.5 %	**18.3 %**	55.6 %	**45.1 %**
Warm Climate	**45.5 %**	**28.9 %**	**74.5 %**	**52.4 %**
Cold Climate	22.1 %	8.2 %	49.6 %	42.8 %

CHAPTER 11

CONCLUSIONS

The goal of this project was to assess the accuracy of the simulation process involving BEopt by modeling the difference between measured energy consumption and RBES predicted energy consumption, as well as to determine which variables have a strong impact on electricity and natural gas consumption, for different types of homes. The types of homes were defined by a formal clustering analysis based on variables that describe household characteristics. First, highly dependent variables were removed using variance inflation factors, and a principal component analysis was performed on the remaining variables to produce independent linear combinations for clustering. Both hierarchical clustering and fuzzy clustering were applied with weighted and unweighted Euclidean distances. The number of clusters was selected using the Xie-Beni, C, and Dunn indices, and all three agreed on two clusters. The clusters formed largely based on climate, but residential archetypes that applied to the entire United States could not be defined since the FDR homes are not a representative sample of all homes in the U.S. Homes in cluster 1 were generally "warm climate" homes, and homes in cluster 2 were generally "cold climate" homes. Multiple linear regression models were then built and compared for each cluster and for all homes for four responses: measured electricity, delta electricity, measured natural gas, and delta natural gas. The models built for the measured responses indicated which predictors have a significant impact on energy consumption, and the models built for the delta responses indicated where improvements may be made in the simulation process.

The results of this project suggest that there are some benefits to clustering the FDR data. The regression models for all responses using the warm climate homes (cluster 1) fit better than the models for all homes and for the cold climate homes (cluster 2). After removing the unusually large cross validation squared errors, the weighted average cross validation MSE for both clusters is generally smaller than for all homes. Additionally, the homes in the warm climate cluster can be predicted better than the homes in the cold climate cluster for the measured electricity and delta natural gas responses. The regression results for the delta responses also indicate that while similar variables appear in all models for each of the delta responses, when the homes are broken into clusters, additional information is gained on factors that could be improved in the simulation process. Given that the models for delta responses are primarily used to pinpoint areas for improvement in the simulation process, this is a major benefit to implementing clustering techniques in the future when different types of homes are added to the data set. Moreover, as improvements are made to the simulation process, the effect these changes have on each type of home in the current data set can be determined. This provides a way to analyze how specific changes in the simulation process impact the predicted energy consumption for warm climate and cold climate homes.

Another benefit to clustering these homes is its potential use in assessing the accuracy of other residential building energy software. The Software Accuracy Working Group has been established at NREL to develop an empirical data-based test suite for residential energy simulation tools. This test suite would include supplying a subset of homes from the FDR to industry software developers that would be used to analyze and improve the accuracy of other residential building energy simulation programs. Providing a set of different types of homes, which have been identified through clustering, would allow these developers to analyze the accuracy of their results and pinpoint specific areas of improvement in their simulation process for each type of home.

With the addition of more homes to the FDR data set, potential future steps include building more representative residential archetypes and assessing the energy consumption and prediction accuracy within these archetypes. To develop meaningful residential archetypes through clustering, a random sample of homes located throughout the United States would be ideal for unbiased results, but a representative sample would also suffice. Furthermore, once the translation scripts are capable of translating attached single-family homes and/or multi-family units into BEopt input files, the prediction accuracy of the simulation process can be compared and similar statistical techniques may be implemented to model energy consumption and investigate areas of improvement in the simulation process for these different types of residential buildings.

REFERENCES CITED

[1] "Annual Energy Review 2010." U.S. Energy Information Administration. (2011). Accessed (Aug 2012): `http://www.eia.gov/totalenergy/data/annual/pdf/aer.pdf`.

[2] "Energy Use in Homes." U.S. Energy Information Administration. (2011). Accessed (Aug 2012): `http://www.eia.gov/energyexplained/index.cfm?page=us_energy_homes#tab1>`.

[3] Christensen, C., R. Anderson, S. Horowitz, A. Courtney, and J. Spencer. "BEopt™ Software for Building Energy Optimization: Features and Capabilities." *National Renewable Energy Laboratory*, (2006).

[4] Hendron, R., and C. Engebrecht. "Building America House Simulation Protocols." *U.S. Department of Energy*, Building Technologies Program, (2010).

[5] Polly, B., N. Kruis, and D. Roberts. "Assessing and Improving the Accuracy of Energy Analysis for Residential Buildings." *National Renewable Energy Laboratory*, (2011).

[6] "Energy 101: Home Energy Checkup." U.S. Department of Energy. (2012). Accessed (Mar 2013): `http://energy.gov/energysaver/articles/energy-101-home-energy-checkup-text-version`.

[7] "Energy Analysis and Tools," National Renewable Energy Laboratory. Accessed (Aug 2012): `http://www.nrel.gov/buildings/energy_analysis.html`.

[8] Ballarini, I., S. Corgnati, V. Corrado, and N. Talà. "Improving Energy Modeling of Large Building Stock Through the Development of Archetype Buildings." *Proceedings of Building Simulation: 12th Conference of International Building Performance Simulation Association*, Sydney, (2011).

[9] Famuyibo, A., A. Duffy, and P. Strachan. "Developing Archetypes for Domestic Dwellings-An Irish Case Study." *Energy and Buildings*, **50**, (2012), 150–157.

[10] Parekh, A., M. A. Sc., and P. Eng. "Development of Archetypes of Building Characteristics Libraries for Simplified Energy Use Evaluation of Houses." *9th International IBPSA Conference*, Montreal, (2005).

[11] Gaitani, N., C. Lehmann, M. Santamouris, G. Mihalakakou, and P. Patargias. "Using Principal Component and Cluster Analysis in the Heating Evaluation of the School Building Sector." *Applied Energy*, **87**, no. 6, (2010), 2079–2086.

[12] Sheather, S. *A Modern Approach to Regression with R*. New York: Springer Science+Business, (2010).

[13] Santamouris, M., G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, M.N. Assimakopoulos, R. Mitoula, and S. Zerefos. "Using Intelligent Clustering Techniques to Classify the Energy Performance of School Buildings." *Energy and Buildings*, **39**, no. 1, (2007), 45–51.

[14] Petcharat, S., S. Chungpaibulpatana, and P. Rakkwamsuk. "Assessment of Potential Energy Saving Using Cluster Analysis: A Case Study of Lighting Systems in Buildings." *Energy and Buildings*, **52**, (2012), 145–152.

[15] Lomas, K., H. Eppel, C. Martin, and D. Bloomfield. "Empirical Validation of Building Energy Simulation Programs." *Energy and Buildings*, **26**, (1997), 253–275.

[16] Tronchin, L., and K. Fabbri. "Energy Performance Building Evaluation in Mediterranean Countries: Comparison Between Software Simulations and Operating Rating Simulation." *Energy and Buildings*, **40**, (2008), 1176–1187.

[17] Brown, R. "Modeled vs. Actual Energy Savings for Energy Upgrade California Home Retrofits." PIER Project PIR-08-018, working paper 3.1. (2012).

[18] Roberts, D., N. Merket, B. Polly, S. Casey, and J. Robertson. "Assessment of the U.S. Department of Energy's Home Energy Scoring Tool." *National Renewable Energy Laboratory*, (2012).

[19] Johnson, R., and D. Wichern. *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River: Pearson Education, (2007).

[20] Gower, J. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics*, **27**, no. 4, (1971), 857–871.

[21] Izenman, A. *Modern Multivariate Statistical techniques*. New York: Springer Science+Business, (2008).

[22] Ward, J. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, **48**, (1963), 236-244.

[23] Lance, G., and W. Williams. "A General Theory of Classificatory Sorting Strategies." *Computer Journal*, **9**, (1967), 373-380.

[24] Milligan, G. "Ultrametric Hierarchical Clustering Algorithms." *Psychometrika*, **44**, no. 3, (1979), 343-346.

[25] Shalizi, C. "Distances between Clustering, Hierarchical Clustering." Course notes. (2009). Accessed (Aug 2012): `http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf`.

[26] Cailliez, F. "The Analytical Solution of the Additive Constant Problem." *Psychometrika*, **48**, no. 2, (1983), 305–308.

[27] Maechler, M. "Cluster Analysis Extended Rousseeuw et. al." *CRAN*, (2012).

[28] Guilford, J. *Psychometric Methods.* New York: McGrawHill Book Company, Inc, (1936).

[29] Glass, G. and K. Hopkins. *Statistical Methods in Education and Psychology* (3rd ed.). Allyn & Bacon, (1995).

[30] Xie, L, and G. Beni. "Validity Measure for Fuzzy Clustering." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **3**, no. 8, (1991),841–847.

[31] Desgraupes, B. "Clustering Indices." (2012). Accessed (Feb 2013): `http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf`.

[32] Hubert, L, and J. Schultz. "Quadratic Assignments as a General Data-Analysis Strategy." *British Journal of Mathematical and Statistical Psychologie*, **29**, (1976) 190–241.

[33] Dunn, J. "Well Separated Clusters and Optimal Fuzzy Partitions." *Journal of Cybernetics*, **4**, (1974), 95–104.

[34] Efron, B, and R. Tibshirani. *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman & Hall/CRC, (1993).

[35] Box, G, and D. Cox. "An Analysis of Transformations." *Journal of the Royal Statistical Society, Series B.* **26**, (1964), 211–246.

[36] Cook, R. D, and Weisberg, S. *Residuals and Influence in Regression.* New York: Chapman and Hall. (1982).

[37] Wilks, D. *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Oxford, UK: Elsevier Inc, (2011).

APPENDIX A - VARIABLE DESCRIPTION TABLES

Table A.1 through Table A.10 are tables describing the variables in the original data set.

Table A.1: Description of general variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

General Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Number of Bedrooms	Integer	Number of bedrooms present in home	Mean: 3.28 Median: 3 Range: 1-7	-
Story Height	Binary	Height of stories (ft)	8 9	-
Age of Home	Integer	Age of home (yr)	Mean: 17.54 Median: 3 Range: 1-118	-
Azimuth	Factor (4)	Orientation of home, calculated based on the distribution of window area	0° 90° 180° 270°	-
Number of Stories	Binary	Number of stories in home given in the FDR	1 2	-

Table A.2: Description of location variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Location Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Station	Factor (22)	Station corresponding to the city where home is located	726450 723035 ...	-
Altitude	Numeric	Altitude of home (ft)	Mean: 689.5 Median: 685.7 Range: 13.12-3061	-
Climate Region	Factor (4)	Climate region where home is located	Cold Hot-Humid ...	-
Cooling Degree Days	Integer	Cooling degree day (CDD) with base temperature of 65°F	Mean: 983.4 Median: 695 Range: 248-2817	-
Heating Degree Days	Integer	Heating degree day (HDD) with base temperature of 65°F	Mean: 5408 Median: 4340 Range: 1415-8331	-

Table A.3: Description of geometry variables in FDR/BEopt dataset.

Geometry Variables	Type	Description	Variable Stats	# NAs
Living Volume	Integer	Volume of living space (ft^3)	Mean: 14600 Median: 12540 Range: 3996-50900	-
Living Height	Numeric	Height of living space (ft)	Mean: 13.29 Median: 9 Range: 8-26	-
Garage Volume	Integer	Volume of garage (ft^3)	Mean: 2652 Median: 3456 Range: 0-8856	-
Ground Height	Integer	Height of basement wall, calculated in BEopt (ft)	Mean: 3.17 Median: 0 Range: 0-10	-
Outside Height	Integer	Height of exterior or cantilevered wall, calculated in BEopt (ft)	Mean: 8.48 Median: 6.75 Range: 3-18.5	-
Interzonal Floor Cavity R-value	Numeric	Cavity insulation R-value of interzonal floors	Mean: 3.29 Median: 0.00 Range: 0.00-80.79	-
Above Grade Wall U-value	Numeric	Equivalent U-value for above grade walls	Mean: 0.082 Median: 0.075 Range: 0.037-0.22	-
Interzonal Wall U-value	Numeric	Equivalent U-value for interzonal walls	Mean: 0.080 Median: 0.078 Range: 0.032-0.204	397
Conditioned Floor Area	Integer	Conditioned floor area given in the FDR (ft^2)	Mean: 2190 Median: 1995 Range: 715-7648	-
Net Above Grade Wall Area	Numeric	Net above grade wall area given in the FDR (ft^2)	Mean: 1440 Median: 1275 Range: 601-4423	-
Has Garage	Logical	Indicates whether a garage is present	T/F	-
Aspect Ratio	Numeric	Ratio between conditioned floor area and wall area	Mean: 2.02 Median: 2.00 Range: 1.00-3.09	-
Wall Difference	Numeric	Measure of difference between FDR and BEopt walls	Mean: 0.051 Median: 0.005 Range: 0-1	-

Table A.4: Description of attic and roof variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Attic/Roof Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Unfinished Attic Volume	Integer	Volume of unfinished attic (ft^3)	Mean: 3617 Median: 3380 Range: 0-14370	-
Unfinished Attic Height	Numeric	Height of unfinished attic (ft)	Mean: 6.16 Median: 6 Range: 0-10.5	-
Roofing Material Color	Factor (3)	Color of roofing material	Dark Light Medium	-
Has Radiant Barrier	Binary	Indicates whether home has radiant barrier	True False	-
Unfinished Attic U-value	Numeric	Equivalent U-value for unfinished attic	Mean: 0.040 Median: 0.031 Range: 0.00-0.478	-
Finished Roof U-value	Numeric	Equivalent U-value for finished roof	Mean: 0.071 Median: 0.051 Range: 0.016-0.302	763
Attic Area	Integer	Attic area given in the FDR (ft^2)	Mean: 1220 Median: 1174 Range: 0-3245	-
Cathedral Ceiling Area	Numeric	Cathedral ceiling area given in the FDR (ft^2)	Mean: 105.6 Median: 0 Range: 0-2417	-

Table A.5: Description of foundation (slab, crawlspace, and basement) variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Foundation Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Unfinished Bsmt Volume	Integer	Volume of unfinished basement (ft^3)	Mean: 149.3 Median: 0 Range: 0-9870	-
Unfinished Bsmt Ceiling Insulated	Logical	Indicates whether unfinished basement ceiling is insulated	TRUE FALSE	964
Unfinished Bsmt Ceiling Cavity R-value	Numeric	Cavity insulation in unfinished basement ceiling	Mean: 8.96 Median: 1.32 Range: 0-41.06	964
Unfinished Bsmt Wall R-value	Numeric	Unfinished basement wall insulation R-value	Mean: 0.73 Median: 0 Range: 0-8.47	965
Finished Bsmt Wall R-value	Numeric	Finished basement wall insulation R-value	Mean: 2.81 Median: 0.45 Range: 0-21.4	502
Crawlspace Ceiling Cavity R-value	Numeric	Cavity insulation in crawlspace ceiling	Mean: 18.99 Median: 23.94 Range: 0-117.70	605
Crawlspace Wall R-value	Numeric	Rigid insulation in crawlspace wall	Mean: 1.28 Median: 0 Range: 0-19	606
Crawlspace ACH	Numeric	Air changes per hour of crawlspace	Mean: 1.04 Median: 0.20 Range: 0.20-2.00	606
Slab Insulation Options	Factor (8)	BEopt insulation options for slab foundation	1 4 ...	781
Conditioned Bsmt Perimeter	Numeric	Perimeter length of foundation walls in conditioned basement given in the FDR (ft)	Mean: 74.39 Median: 0 Range: 0-345	-
Unconditioned Bsmt Perimeter	Numeric	Perimeter length of foundation walls in unconditioned basement given in the FDR (ft)	Mean: 3.12 Median: 0 Range: 0-164	-
Enclosed Crawlspace Perimeter	Numeric	Perimeter length of foundation walls in enclosed crawlspaces given in the FDR (ft)	Mean: 54.44 Median: 0.00 Range: 0.00-384.00	-
Slab Perimeter	Numeric	Perimeter length of slab given in the FDR (ft)	Mean: 26.43 Median: 0 Range: 0-315	-

Table A.6: Description of window variables in FDR/BEopt dataset.

Window Variables	Type	Description	Variable Stats	# NAs
Window Area Percent of Wall Area	Numeric	Percent of wall area that is window	Mean: 0.144 Median: 0.140 Range: 0.060-0.450	-
Front Window Percent Of Total	Numeric	Percent of window area located on front of home	Mean: 0.285 Median: 0.280 Range: 0.000-0.500	-
Back Window Percent Of Total	Numeric	Percent of window area located on back of home	Mean: 0.459 Median: 0.450 Range: 0.250-0.890	-
Right Window Percent Of Total	Numeric	Percent of window area located on right side of home	Mean: 0.167 Median: 0.170 Range: 0.000-0.420	-
Front Window U-value	Numeric	Window U-value for front of home	Mean: 0.456 Median: 0.400 Range: 0.2448-1.160	-
Front Window SHGC	Numeric	Window solar heat gain coefficient (SHGC) for front of home	Mean: 0.463 Median: 0.450 Range: 0.270-0.760	-
Back Window U-value	Numeric	Window U-value for back of home	Mean: 0.442 Median: 0.400 Range: 0.285-1.160	-
Back Window SHGC	Numeric	Window SHGC for back of home	Mean: 0.457 Median: 0.450 Range: 0.250-0.760	-
Left Window U-value	Numeric	Window U-value for left side of home	Mean: 0.531 Median: 0.419 Range: 0.284-1.160	-
Left Window SHGC	Numeric	Window SHGC for left side of home	Mean: 0.490 Median: 0.461 Range: 0.250-0.760	-
Right Window U-value	Numeric	Window U-value for right side of home	Mean: 0.538 Median: 0.426 Range: 0.170-1.160	-
Right Window SHGC	Numeric	Window SHGC for right side of home	Mean: 0.494 Median: 0.480 Range: 0.190-0.760	-
South Window Area	Numeric	South window area given in FDR (ft^2)	Mean: 56.8 Median: 44.5 Range: 0.0-328.0	-
North Window Area	Numeric	North window area given in FDR (ft^2)	Mean: 54.7 Median: 51.3 Range: 0.0-378.0	-
West Window Area	Numeric	West window area given in FDR (ft^2)	Mean: 60.9 Median: 47.9 Range: 0.0-504.0	-
East Window Area	Numeric	East window area given in FDR (ft^2)	Mean: 59.8 Median: 48.1 Range: 0.0-517.0	-

Table A.7: Description of infiltration, duct, and ventilation variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Infiltration, Duct, and Ventilation Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Living Space ACH50	Numeric	Air changes per hour at 50 Pascals of living space	Mean: 6.522 Median: 6.526 Range: 0.494-31.140	-
Ventilation Type	Factor (3)	Type of ventilation in home	Exhaust Supply Balanced	-
Fraction of ASHRAE	Numeric	Fraction of the ventilation rate (including any infiltration credit) specified by ASHRAE 62.2 that is desired in the building	Mean: 0.299 Median: 0.031 Range: 0.000-2.081	-
House Vent Fan Power	Numeric	Fan power per delivered airflow rate of fan(s) providing whole house ventilation (W/cfm)	Mean: 0.546 Median: 0.300 Range: 0.157-8.000	-
ERV Efficiency	Numeric	Energy recovery ventilator efficiency, applicable only for balanced ventilation type	Mean: 0.222 Median: 0.200 Range: 0.200-0.700	942
AH Leak SA	Numeric	Fraction of air-handler flow rate that is lost on the supply-side of the air-handler	Mean: 0.031 Median: 0.019 Range: 0.000-0.330	30
Unconditioned Duct R-value	Numeric	R-value of insulation around ducts in unconditioned space	Mean: 3.03 Median: 0.00 Range: 0.00-38.00	30
Duct Location	Factor (5)	Location of ducts in home	Living Crawlspace ...	30
Duct Location Fraction	Numeric	Fraction of ducts in specified BEopt location	Mean: 0.933 Median: 1.000 Range: 0.030-1.000	479
Duct Number of Returns	Integer	Number of return registers	Mean: 2.3 Median: 1 Range: 1-8	30

Table A.8: Description of appliance variables in FDR/BEopt dataset.

Appliance Variables	Type	Description	Variable Stats	# NAs
Refrigerator Annual Energy	Integer	Annual energy used by refrigerator in home (kWh)	Mean: 583.9 Median: 445 Range: 400-2931	773
Has Cooking Range	Logical	Indicates whether cooking range present in home	T/F	-
Cooking Range Fuel Type	Binary	Fuel type of cooking range	Electric Gas	140
Dishwasher EF	Numeric	Dishwasher energy factor given in FDR	Mean: 0.472 Median: 0.460 Range: 0.000-0.850	-
Clothes Dryer Fuel Type	Binary	Fuel type of clothes dryer	Electric Gas	142
Lighting Values Not Defaulted	Logical	Indicates whether lighting values were defaulted or not in BEopt	T/F	-
Watts Per CFM Defaulted	Logical	Indicates whether watts per CFM were defaulted or not in BEopt	T/F	-

Table A.9: Description of space conditioning variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Space Conditioning Variables	Type	Description	Variable Stats	# NAs
Fixed Heating Capacity	Numeric	Measures the ability of heating equipment to add heat to an enclosed space (Btu/h)	Mean: 61.52 Median: 70.00 Range: 18.00-90.00	494
Fixed Cooling Capacity	Numeric	Measures the ability of cooling equipment to add heat to an enclosed space (Btu/h)	Mean: 2.58 Median: 2.00 Range: 1.42-7.00	183
AC Cooling SEER	Integer	SEER value for AC	Mean: 11.2 Median: 12 Range: 8-17	473
Furnace Fuel Type	Factor (3)	Fuel type of furnace	Gas Electric	-
Furnace AFUE	Numeric	Annual Fuel Utilization Efficiency (AFUE) of furnace	Mean: 0.863 Median: 0.900 Range: 0.600-0.980	339
HP Cooling SEER	Integer	SEER value for heat pump	Mean: 13.45 Median: 13 Range: 13-17	707
Use Ceiling Fan	Logical	Indicates use of ceiling fan	T/F	-
Has AC	Logical	Indicates presence of AC in home	T/F	-
Has ASHP	Logical	Indicates presence of air source heat pump in home	T/F	-
Has Furnace	Logical	Indicates presence of furnace in home	T/F	-
Has Boiler	Logical	Indicates presence of boiler in home	T/F	-

Table A.10: Description of water heating variables in FDR/BEopt dataset. The number in parentheses next to type Factor is the number of categories.

Water Heating Variables	**Type**	**Description**	**Variable Stats**	**# NAs**
Water Heater Type	Factor (3)	Type of water heater in home	Tank Tankless Heat Pump	-
Water Heater Fuel Type	Binary	Fuel type of water heater	Gas Electric	-
Water Heater Wrap R-value	Numeric	R-value of any additional insulation wrapped around the water heater tank	Mean: 0.32 Median: 0.00 Range: 0.00-11.00	2
Water Heater Tank Volume	Integer	Volume of water heater tank	Mean: 43.1 Median: 40 Range: 30-82	11
Water Heater Recovery Efficiency	Numeric	Ratio of energy delivered to the water to the energy content of the fuel used by the water heater	Mean: 0.838 Median: 0.760 Range: 0.540-0.990	13

APPENDIX B - VARIANCE INFLATION FACTORS

Table B.1 is a table of the variance inflation factors for the original variables present in the dataset and for the subset of variables being used in the analysis.

Table B.1: Table of the variance inflation factors.

		VIF	
#	**Variable**	**Original**	**Subset**
1	Finished Floor Area	Inf	
2	Above Grade Finished Floor Area	Inf	
3	Number of Bedrooms	2.46	1.91
4	Building Height	341.87	
5	Story Height	15.21	
6	Crawlspace Volume	23.59	
7	Crawlspace Height	9.26	
8	Crawlspace Floor Area	90.72	
9	Unfinished Bsmt Volume	556.76	
10	Unfinished Bsmt Floor Area	1246.04	
11	Finished Bsmt Volume	81.78	
12	Finished Bsmt Height	29.27	
13	Finished Bsmt Floor Area	Inf	
14	Living Volume	244.16	
15	Living Height	16.92	
16	Garage Volume	2824.75	
17	Garage Height	13917.43	
18	Unfinished Attic Volume	25.89	
19	Unfinished Attic Height	11.09	
20	Altitude	Inf	
21	Cooling Degree Days	Inf	
22	Heating Degree Days	Inf	10.23
23	Fixed Heating Capacity	7.63	5.51
24	Fixed Cooling Capacity	5.68	4.31
25	Age Of Home	5.00	3.76
26	Interzonal Floor Cavity R-value	1.91	1.63
27	Window Area Percent Of Wall Area	14.47	
28	Front Percent Of Total	1165.27	2.91
29	Back Percent Of Total	1518.37	
30	Left Percent Of Total	1634.05	7.82
31	Right Percent Of Total	1200.58	2.48
32	Front Window U-value	9.30	2.53
33	Front Window SHGC	17.77	
34	Back Window U-value	9.00	5.05

Continued on next page

Table B.1 continued: Table of the variance inflation factors.

#	Variable	VIF Original	VIF Subset
35	Back Window SHGC	21.06	4.92
36	Left Window U-value	11.17	1.60
37	Left Window SHGC	20.32	
38	Right Window U-value	13.60	2.13
39	Right Window SHGC	20.05	
40	Living Space ACH50	3.63	3.12
41	Fraction of ASHRAE	5.93	3.27
42	House Vent Fan Power	2.14	1.85
43	AH Leak SA	13618.62	
44	AH Leak RA	13555.42	2.83
45	Unconditioned Duct R-value	4.81	3.33
46	Duct Number of Returns	7.27	5.37
47	Water Heater Energy Factor	57.02	4.93
48	Water Heater Wrap R-value	1.54	1.25
49	Water Heater Tank Volume	1.93	1.63
50	Water Heater Recovery Efficiency	30.47	
51	Above Grade Wall U-value	2.97	2.40
52	Unfinished Attic U-value	1.60	1.38
53	Total Window Area	Inf	
54	South Window Area	Inf	3.30
55	North Window Area	Inf	3.23
56	West Window Area	Inf	3.50
57	East Window Area	Inf	2.53
58	Dishwasher EF	1.87	1.37
59	Conditioned Floor Area	10165.44	9.17
60	Number of Stories	301.05	4.11
61	Attic Area	10.83	4.82
62	Cathedral Ceiling Area	13.66	2.05
63	Net Above Grade Wall Area	21.38	7.83
64	Conditioned Bsmt Perimeter	40.15	
65	Unconditioned Bsmt Perimeter	219.97	6.67
66	Enclosed Crawlspace Perimeter	59.82	4.74
67	Slab Perimeter	40.40	
68	Aspect Ratio	27.35	
69	Wall Difference	1.62	1.28
70	State: MN	Inf	
71	State: NC	Inf	
72	State: OR	Inf	
73	State: TX	Inf	
74	Station: 722436	Inf	

Continued on next page

Table B.1 continued: Table of the variance inflation factors.

#	Variable	VIF Original	VIF Subset
75	Station: 723035	Inf	
76	Station: 723060	Inf	
77	Station: 723066	Inf	
78	Station: 723068	NaN	
79	Station: 723095	Inf	
80	Station: 723140	Inf	
81	Station: 723143	Inf	
82	Station: 723145	Inf	
83	Station: 723150	Inf	
84	Station: 723170	Inf	
85	Station: 726400	Inf	
86	Station: 726410	Inf	
87	Station: 726435	Inf	
88	Station: 726457	Inf	
89	Station: 726463	NaN	
90	Station: 726574	Inf	
91	Station: 726580	Inf	
92	Station: 726835	Inf	
93	Station: 726980	Inf	
94	Station: 746943	Inf	
95	Climate: Hot Humid	Inf	
96	Climate: Marine	Inf	
97	Climate: Mixed Humid	Inf	
98	Azimuth: 0°	3.75	3.21
99	Azimuth: 90°	3.80	3.20
100	Azimuth: 180°	4.58	3.54
101	Roof Color: Dark	2.62	1.37
102	Roof Color: Light	1.46	1.16
103	Has Radiant Barrier	2.76	1.82
104	Ventilation Type: Balanced	3.02	2.53
105	Ventilation Type: Supply	3.45	2.52
106	Duct Location: Crawlspace	17.36	
107	Duct Location: Garage	1.39	1.20
108	Duct Location: Unfinished Attic	15.32	3.63
109	Duct Location: Unfinished Bsmt	12.18	6.95
110	Use Ceiling Fan	2.44	1.68
111	Water Heater Type: HeatPump	NaN	
112	Water Heater Type: Tankless	NaN	
113	Water Heater Fuel Type: Electric	87.39	
114	Cooking Range Fuel Type: Gas	3.17	2.33

Continued on next page

Table B.1 continued: Table of the variance inflation factors.

#	Variable	VIF Original	VIF Subset
115	Clothes Dryer Fuel Type: Gas	2.98	2.26
116	Lighting Values Not Defaulted	3.03	2.37
117	No AC	Inf	
118	Has ASHP	Inf	
119	No Furnace	Inf	3.80
120	Has Boiler	Inf	
121	No Garage	4446.42	7.35
122	No Cooking Range	NaN	
123	No Clothes Dryer	NaN	
124	Unfinished Bsmt Ceiling Cavity R-value	Inf	Inf
125	Unfinished Bsmt Wall R-value	Inf	Inf
126	Finished Bsmt Wall R-value	3.33	2.45
127	Crawlspace Ceiling Cavity R-value	8.60	3.34
128	Crawlspace Wall R-value	4.81	2.11
129	Crawlspace ACH	7.01	2.57
130	ERV Efficiency	Inf	Inf
131	Refrigerator Annual Energy	37.38	5.59
132	AC Cooling SEER	3.61	2.80
133	Furnace AFUE	2.30	1.94
134	Boiler AFUE	Inf	Inf
135	HP Cooling SEER	72.13	
136	Duct Location Fraction	10.04	3.78
137	Interzonal Wall U-value	2.97	2.29
138	Finished Roof U-value	7.00	2.90
139	Slab Option: 4	82.17	8.02
140	Slab Option: 5	NaN	
141	Slab Option: 6	36.49	4.39
142	Slab Option: 7	3.99	2.04
143	Slab Option: 8	19.51	6.52
144	Slab Option: 9	32.69	
145	Slab Option: 12	11.39	3.32
146	Watts Per CFM Defaulted	38.79	
147	Unfinished Bsmt Ceiling Not Insulated	Inf	Inf
148	Furnace Fuel Type: Electric	1.34	1.21

APPENDIX C - VARIABLE HISTOGRAMS

Figure C.1-Figure C.6 are the histograms for the quantitative variables.

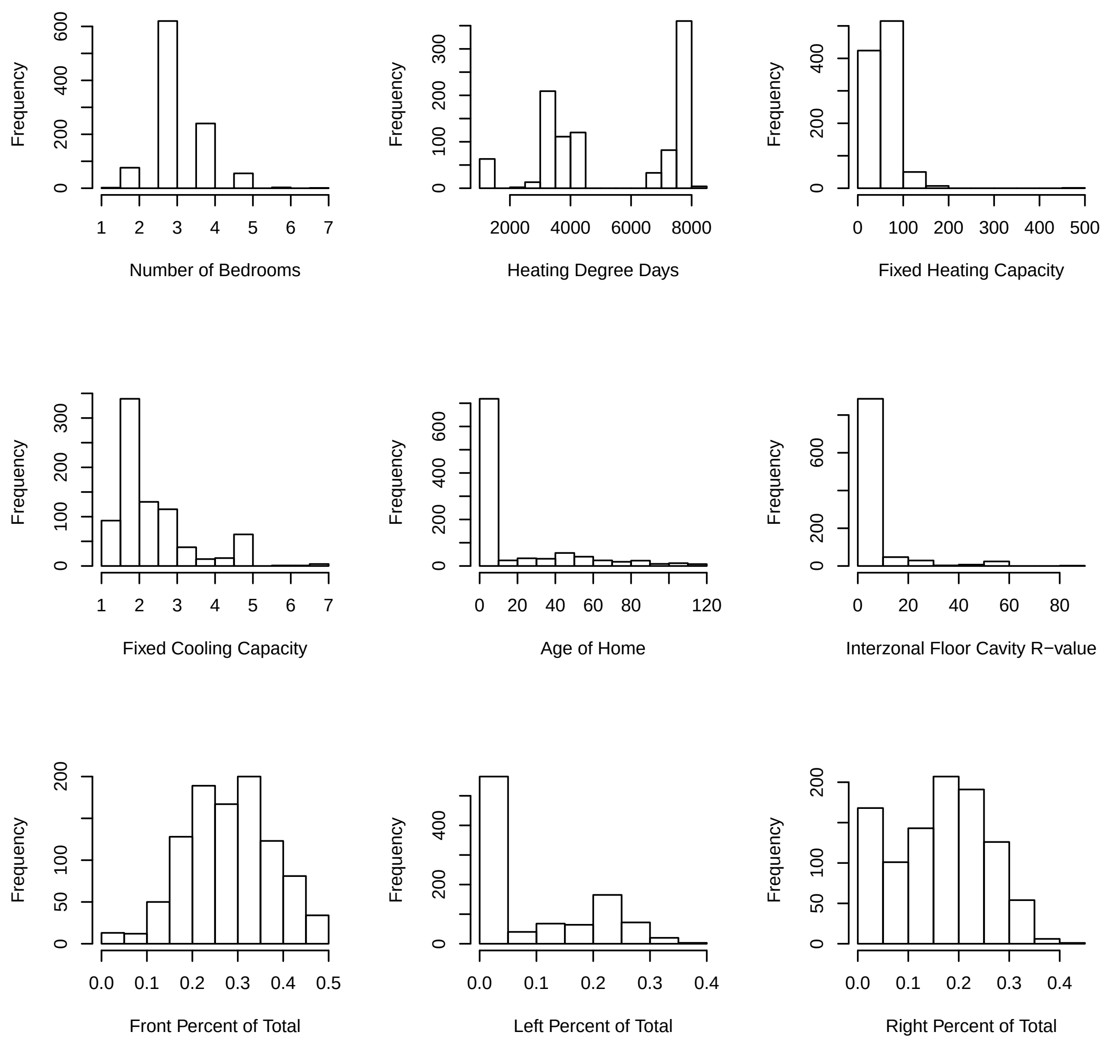


Figure C.1: Histograms for Number of Bedrooms, Heating Degree Days, Fixed Heating Capacity, Fixed Cooling Capacity, Age of Home, Interzonal Floor Cavity R-value, Front Window Percent of Total, Left Window Percent of Total, and Right Window Percent of Total.

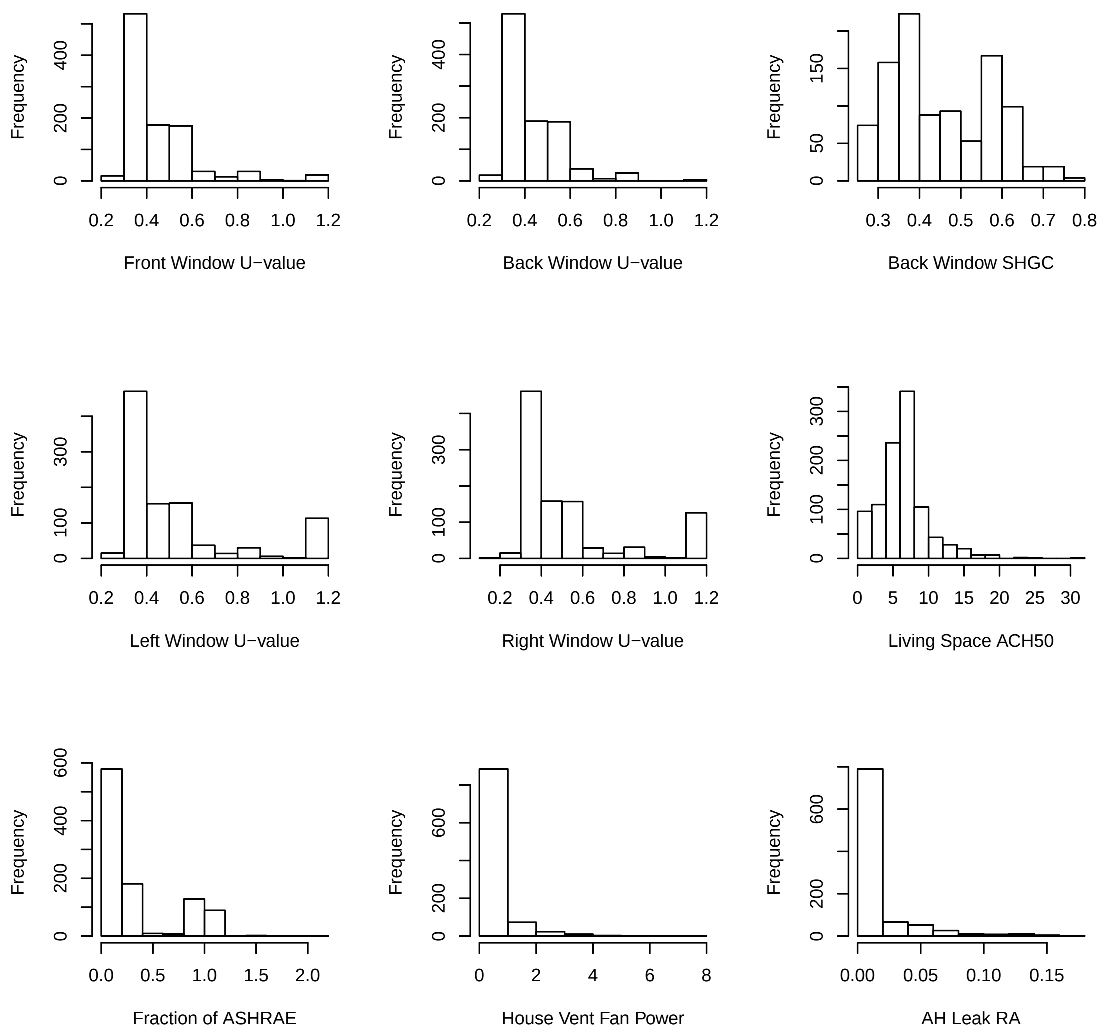


Figure C.2: Histograms for Front Window U-value, Back Window U-value, Back Window SHGC, Left Window U-value, Right Window U-value, Living Space ACH50, Fraction of ASHRAE, House Vent Fan Power, and AH Leak RA.

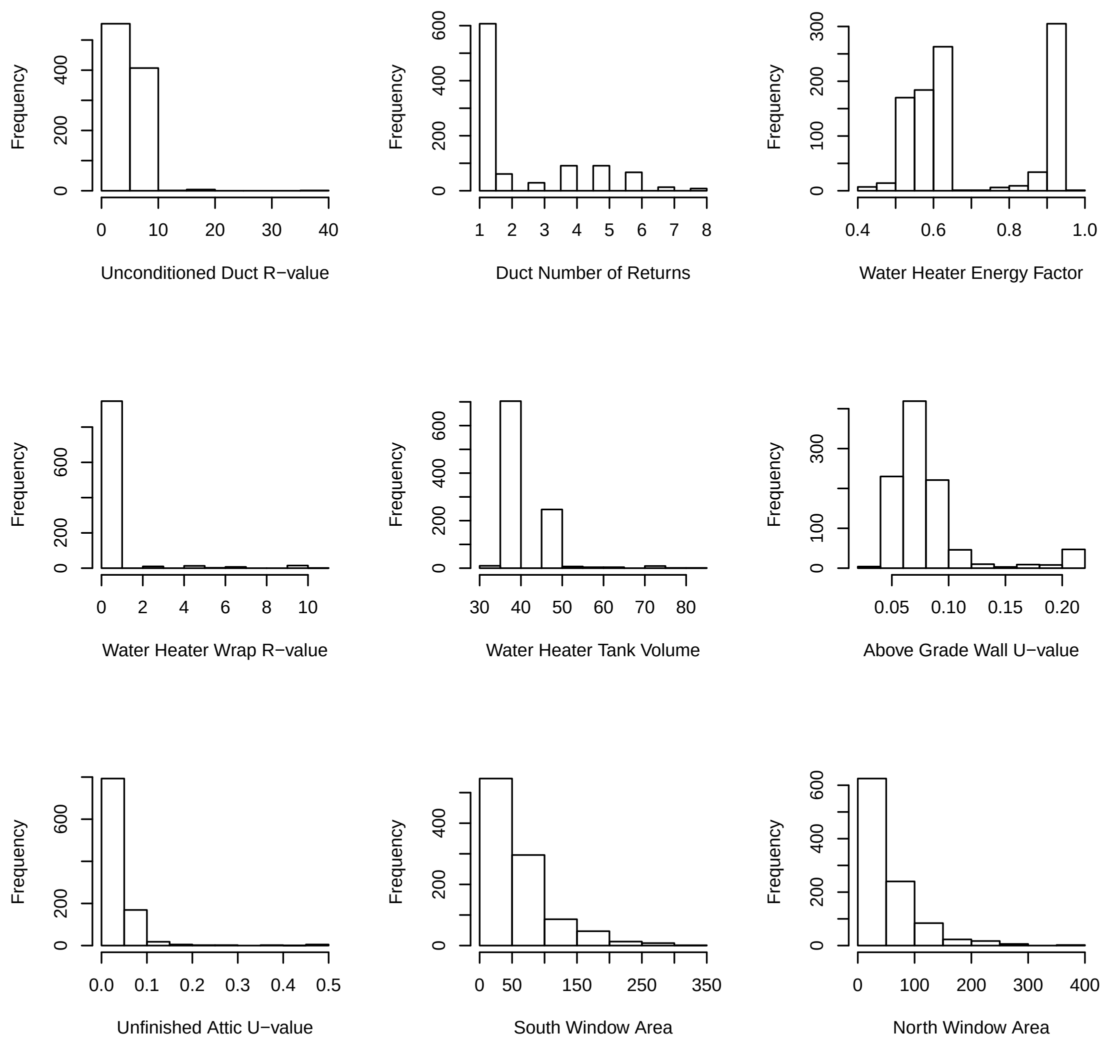


Figure C.3: Histograms for Unconditioned Duct R-value, Duct Number of Returns, Water Heater Energy Factor, Water Heater Wrap R-value, Water Heater Tank Volume, Above Grade Wall U-value, Unfinished Attic U-value, South Window Area, and North Window Area.

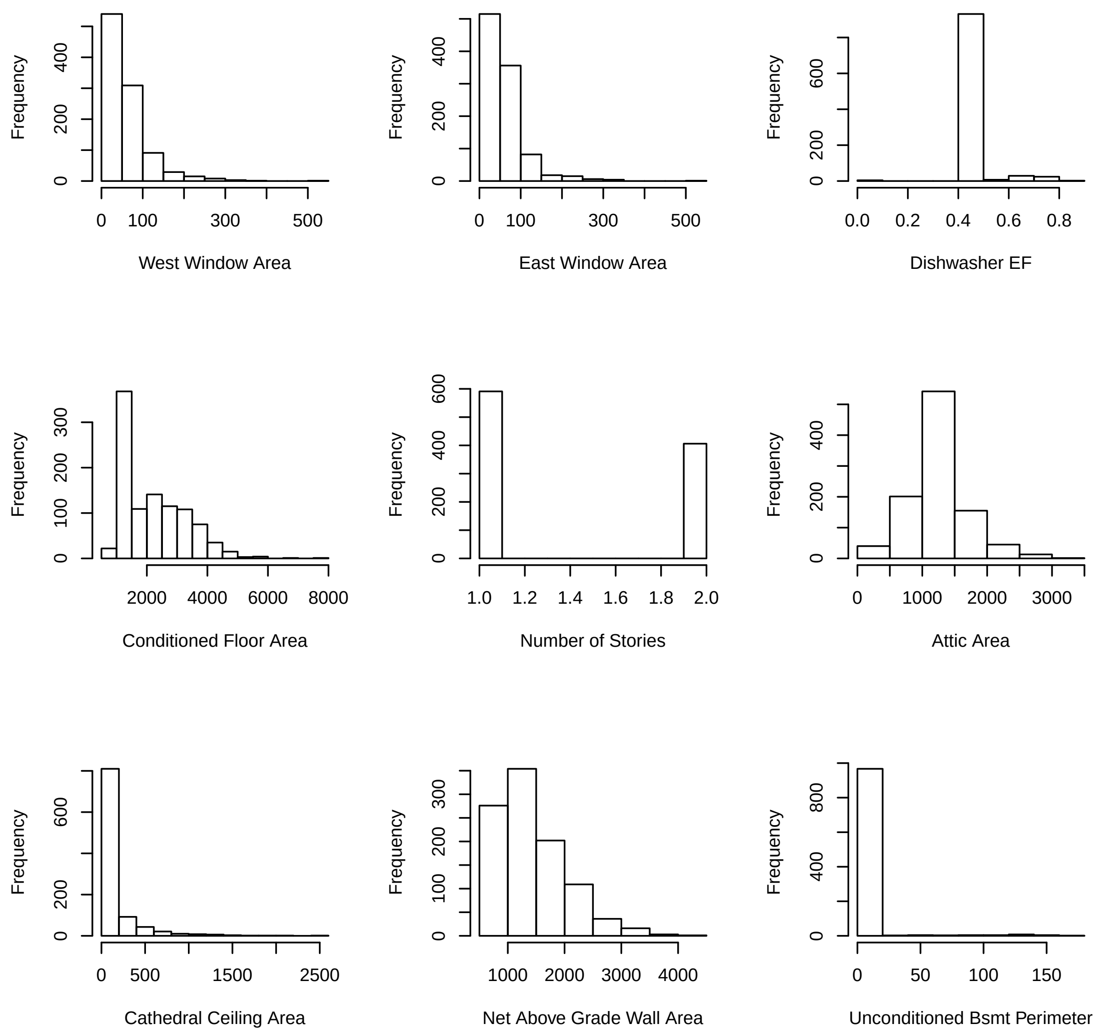


Figure C.4: Histograms for West Window Area, East Window Area, Dishwasher EF, Conditioned Floor Area, Number of Stories, Attic Area, Cathedral Ceiling Area, Net Above Grade Wall Area, and Unconditioned Bsmt Perimeter.

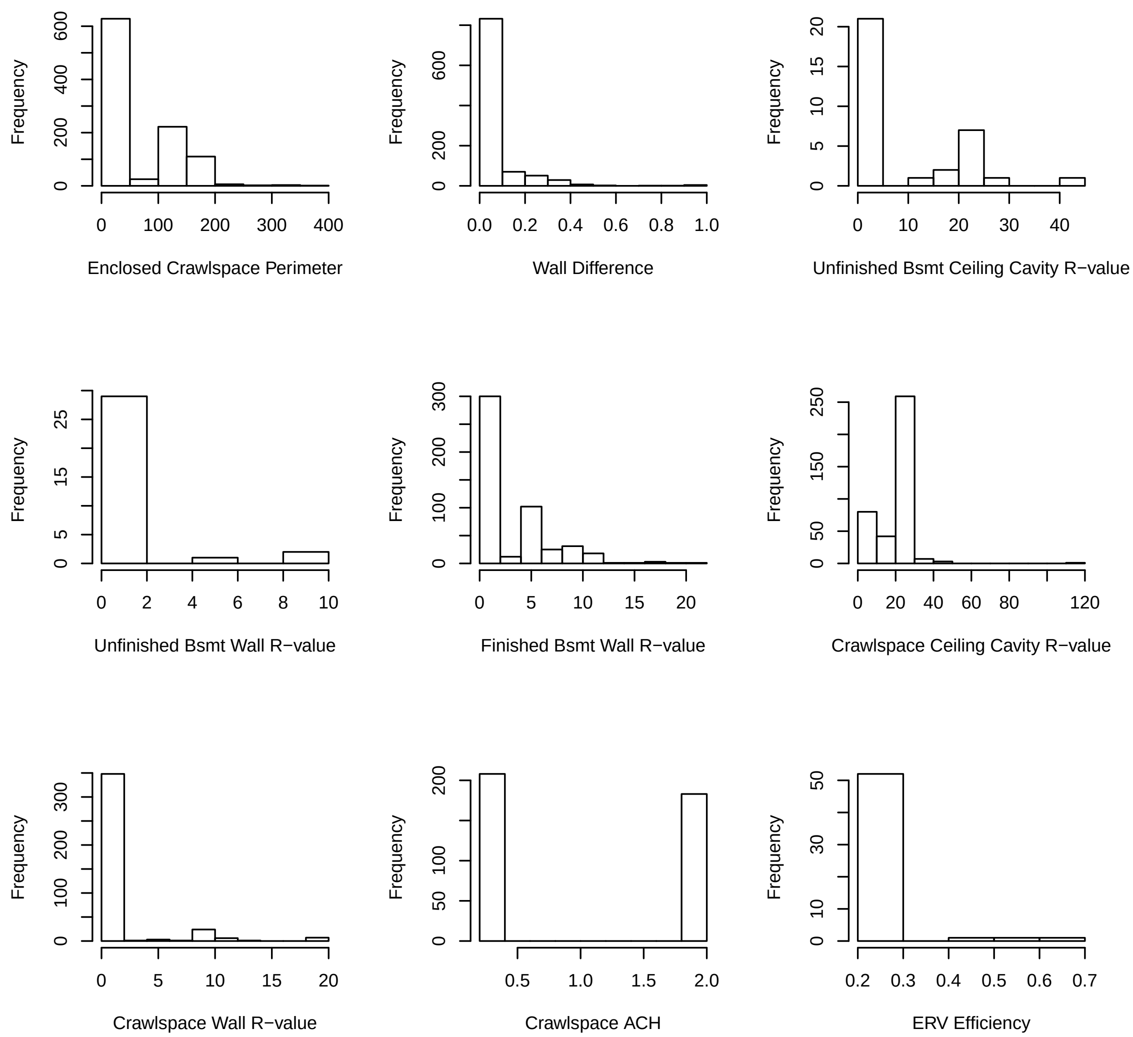


Figure C.5: Histograms for Enclosed Crawlspace Perimeter, Wall Difference, Unfinished Bsmt Ceiling Cavity R-value, Unfinished Bsmt Wall R-value, Finished Bsmt Wall R-value, Crawlspace Ceiling Cavity R-value, Crawlspace Wall R-value, Crawlspace ACH, and ERV Efficiency.

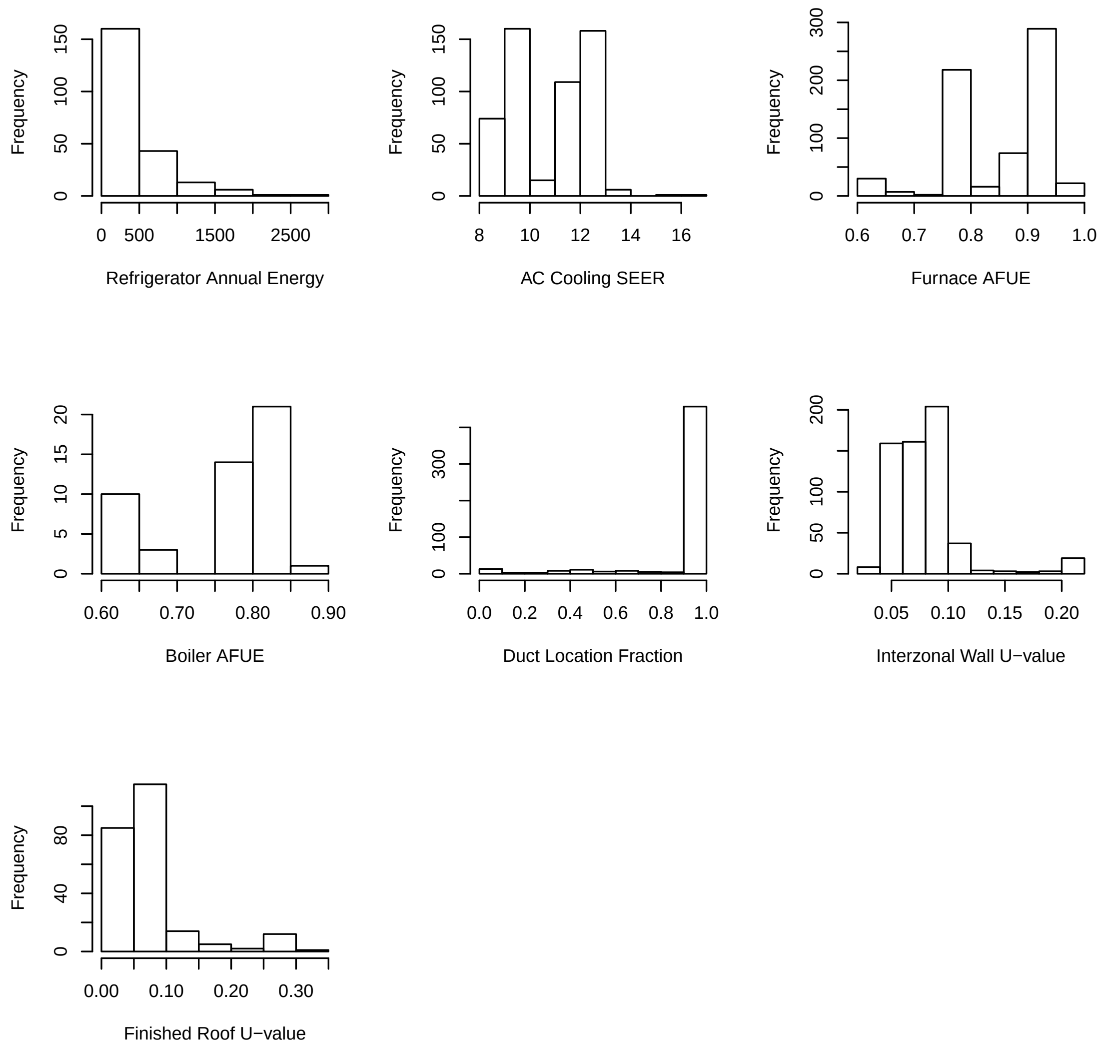


Figure C.6: Histograms for Refrigerator Annual Energy, AC Cooling SEER, Furnace AFUE, Boiler AFUE, Duct Location Fraction, Interzonal Wall U-value, and Finished Roof U-value.

APPENDIX D - DISCRETE VALUES OF QUANTITATIVE VARIABLES

Table D.1 is a list of the number of unique values in each quantitative variable.

Table D.1: Table of quantitative variables and the number of unique values in each variable.

Variable	**Discrete Values**	**Variable**	**Discrete Values**
Number of Bedrooms	7	North Window Area	505
Heating Degree Days	22	West Window Area	544
Fixed Heating Capacity	57	East Window Area	544
Fixed Cooling Capacity	20	Dishwasher EF	19
Age of Home	97	Conditioned Floor Area	628
Interzonal Floor Cavity R-value	45	Number of Stories	2
Front Window Percent of Total	44	Attic Area	566
Left Window Percent of Total	34	Cathedral Ceiling Area	200
Right Window Percent of Total	40	Net Above Grade Wall Area	753
Front Window U-value	333	Unconditioned Basement Perimeter	29
Back Window U-value	391	Enclosed Crawlspace Perimeter	122
Back Window SHGC	410	Wall Difference	735
Left Window U-value	291	Unfinished Bsmt Ceiling Cavity R-value	11
Right Window U-value	266	Unfinished Bsmt Wall R-value	5
Living Space ACH50	782	Finished Bsmt Wall R-value	143
Fraction of ASHRAE	389	Crawlspace Ceiling Cavity R-value	43
House Vent Fan Power	204	Crawlspace Wall R-value	19
AH Leak RA	79	Crawlspace ACH	2
Unconditioned Duct R-value	33	ERV Efficiency	4
Duct Number of Returns	8	Refrigerator Annual Energy	69
Water Heater Energy Factor	43	AC Cooling SEER	9
Water Heater Wrap R-value	7	Furnace AFUE	28
Water Heater Tank Volume	10	Boiler AFUE	9
Above Grade Wall U-value	148	Duct Location Fraction	35
Unfinished Attic U-value	236	Interzonal Wall U-value	129
South Window Area	537	Finished Roof U-value	83

The diagnostic plots used to validate the model assumptions for all homes, the warm climate homes, and the cold climate homes are given below for all four responses.

E.1 Measured Electricity

The diagnostic plots corresponding to the model for the warm climate homes (cluster 1) and the measured electricity response are seen in Figure E.1. The standardized residuals versus fitted values shows no distinct nonlinearity or non-constant variance. The errors appear to be approximately normally distributed as seen in the normal Q-Q plot. While there are also some points of high leverage, all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

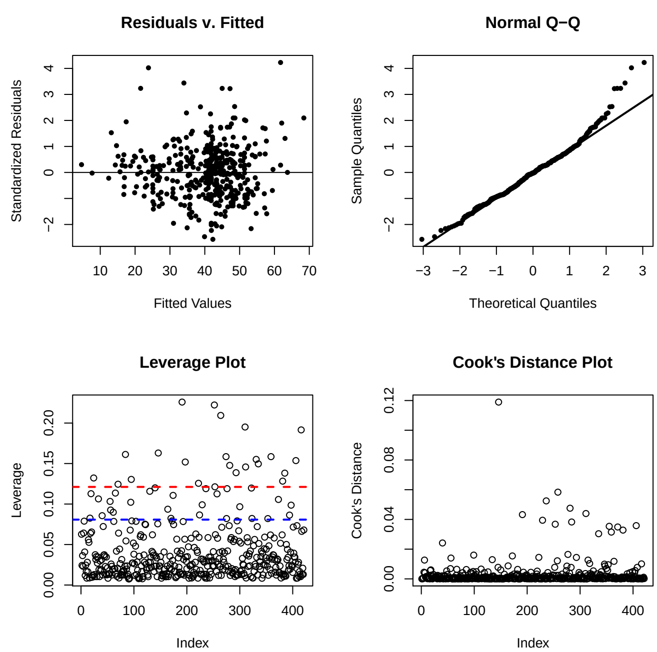


Figure E.1: Diagnostic plots for model with measured electricity as the response for the warm climate homes.

The diagnostic plots corresponding to the model for the cold climate homes (cluster 2) are seen in Figure E.2, which will be used to check the assumptions of the model. The standardized residuals versus fitted values shows no non-constant variance or distinct non-linearity. The normal Q-Q plot shows that the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of Cook's distance are less than one. Therefore, no observations need to be removed as outliers.

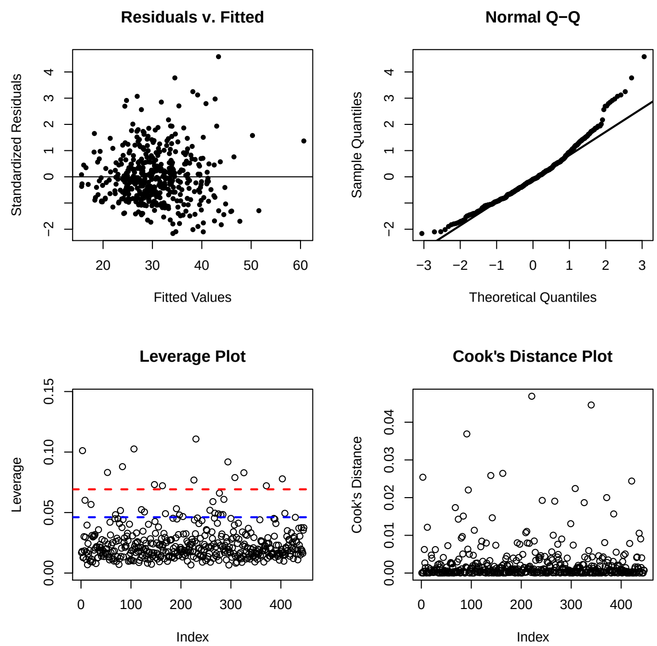


Figure E.2: Diagnostic plots for model with measured electricity as the response for the cold climate homes.

E.2 Delta Electricity

The diagnostic plots corresponding to the model for all homes and the delta electricity response are seen in Figure E.3. From the plot of the standardized residual versus fitted values, no nonlinearities are present, and the variance is constant. Moreover, the Q-Q plot shows that the errors generally look normally distributed. The lower tail does not follow the Q-Q line, but this is expected since the distribution of delta electricity is skewed left. There are some points of high leverage, but all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

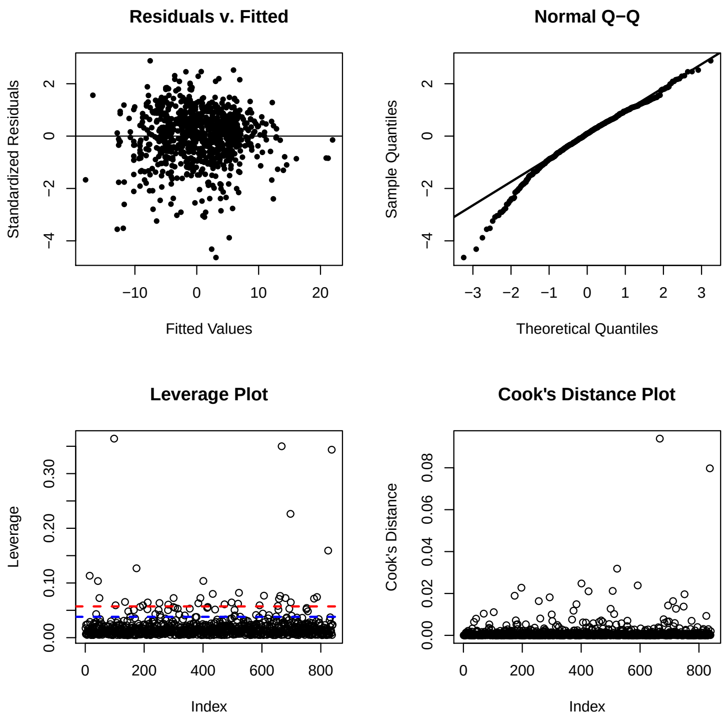


Figure E.3: Diagnostic plots for the reduced multiple linear regression model with delta electricity as the response.

The diagnostic plots corresponding to the model for the warm climate homes are seen in Figure E.4, which will be used to check the assumptions of the model. The standardized residuals versus fitted plot shows no distinct pattern, so the assumption of linearity is met. From the Normal Q-Q plot, the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

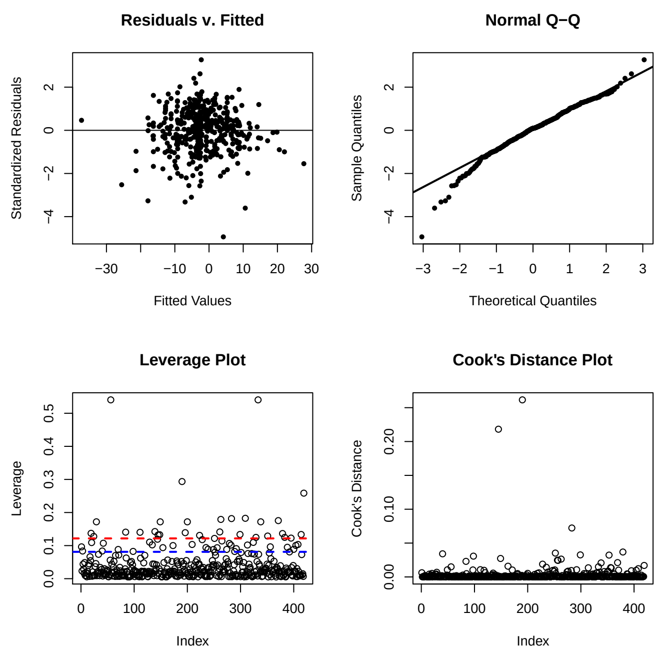


Figure E.4: Diagnostic plots for model with delta electricity as the response for the warm climate homes.

The diagnostic plots corresponding to the model for the cold climate homes are seen in Figure E.5 and will be used to check the assumptions of the model. The standardized residual versus fitted plot shows no non-constant variance or distinct nonlinearity. The normal Q-Q plot shows that the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of cook's distance are less than one, so there are no outliers that need to be removed.

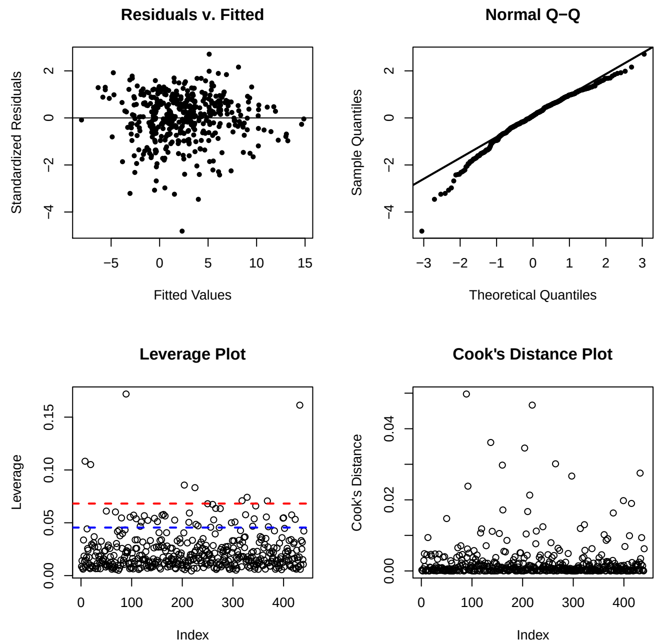


Figure E.5: Diagnostic plots for model with delta electricity as the response for the cold climate homes.

E.3 Measured Natural Gas

The diagnostic plots corresponding to the model for all homes and the measured natural gas response are seen in Figure E.6. From the plot of the standardized residual versus fitted values, the variance is constant, and no nonlinearities are present. Moreover, the Q-Q plot shows that the errors generally look normally distributed. The upper tail does not follow the Q-Q line, but this is expected since there is a lower bound but no upper bound for natural gas consumption. There are some points of high leverage, but all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

The diagnostic plots corresponding to the model for the warm climate homes are seen in Figure E.7 and will be used to check the assumptions of the model. The standardized residual versus fitted plot shows no distinct pattern, therefore the assumption of linearity is met. The normal Q-Q plot shows that the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of Cook's distance are less than one. Therefore, no observations need to be removed as outliers.

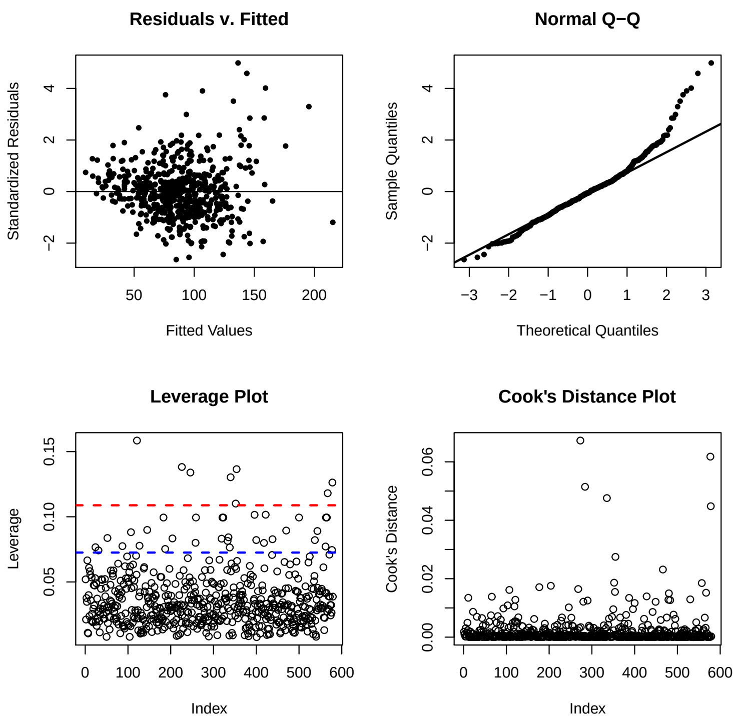


Figure E.6: Diagnostic plots for the model with all homes and measured natural gas as the response.

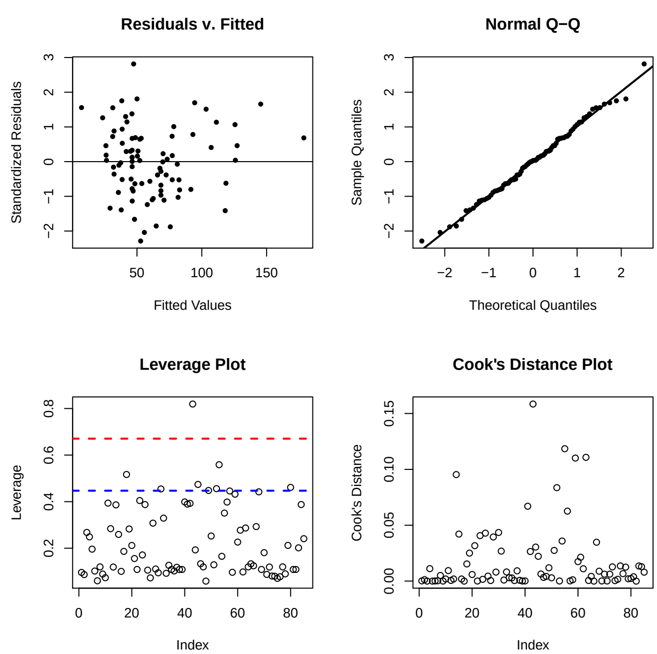


Figure E.7: Diagnostic plots for reduced model with measured natural gas as the response for the warm climate homes.

The diagnostic plots corresponding to the model for the cold climate homes are seen in Figure E.8 and will be used to check the assumptions of the model. The standardized residual vs. fitted plot shows no distinct nonlinearity or non-constant variance. The normal Q-Q plot shows that the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of cook's distance are less than one, so there are no outliers that need to be removed.

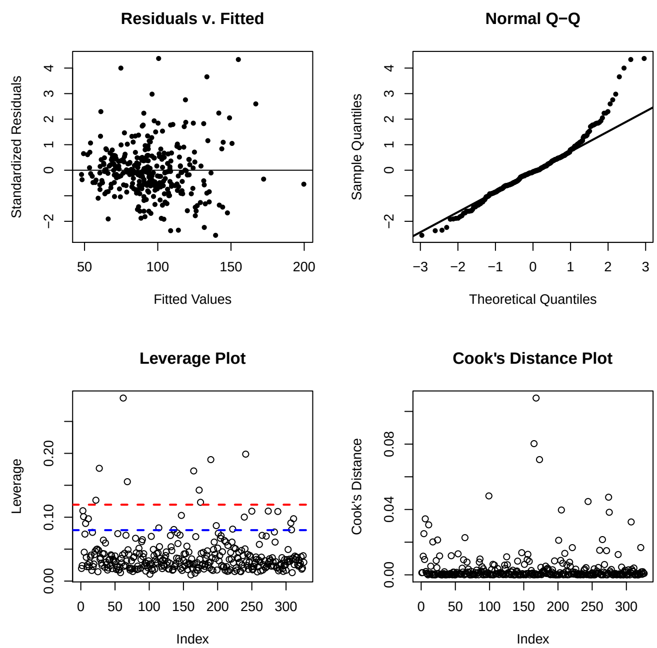


Figure E.8: Diagnostic plots for reduced model with measured natural gas as the response for the cold climate homes.

E.4 Delta Natural Gas

The diagnostic plots corresponding to the model for all homes and the delta natural gas response are seen in Figure E.9. From the plot of the standardized residual versus fitted values, no nonlinearities are present, and the variance is constant. The Q-Q plot shows that the errors generally look normally distributed. The lower tail does not follow the Q-Q line, but this is expected since the distribution of delta natural gas is skewed left. There are some points of high leverage, but all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

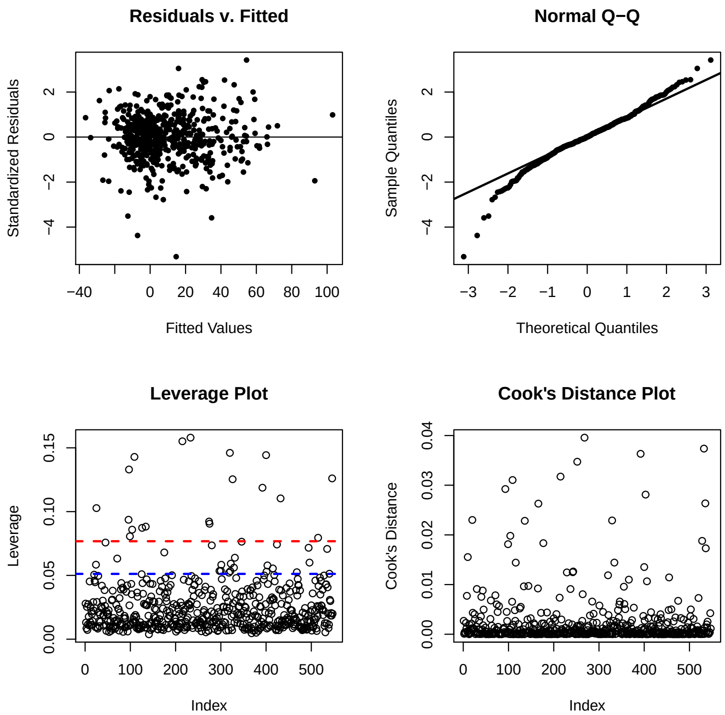


Figure E.9: Diagnostic plots for the reduced multiple linear regression model with delta natural gas as the response.

The diagnostic plots corresponding to the model for the warm climate homes are seen in Figure E.10 and will be used to check the assumptions of the model. The standardized residual vs. fitted plot shows no evidence of non-constant variance or nonlinearity. From the Normal Q-Q plot, the errors appear to be approximately normally distributed. While there are some points of high leverage, all of the values of Cook's distance are less than one. Therefore, no observations need to be removed as outliers.

The diagnostic plots corresponding to the model for the cold climate homes are seen in Figure E.11 and will be used to check the assumptions of the model. The standardized residuals versus fitted values shows no distinct nonlinearity or non-constant variance. The errors appear to be approximately normally distributed as seen in the normal Q-Q plot. While there are also some points of high leverage, all of the values of Cook's distance are less than one, so there are no outliers that need to be removed.

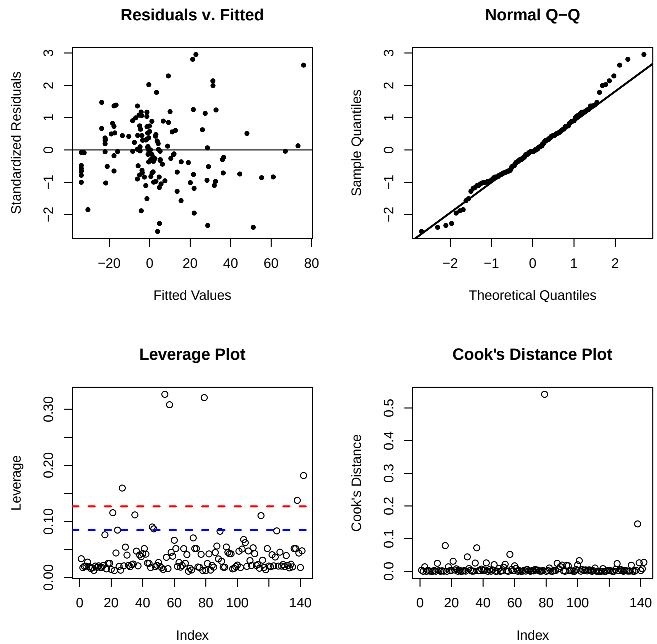


Figure E.10: Diagnostic plots for model with delta natural gas as the response for the warm climate homes.

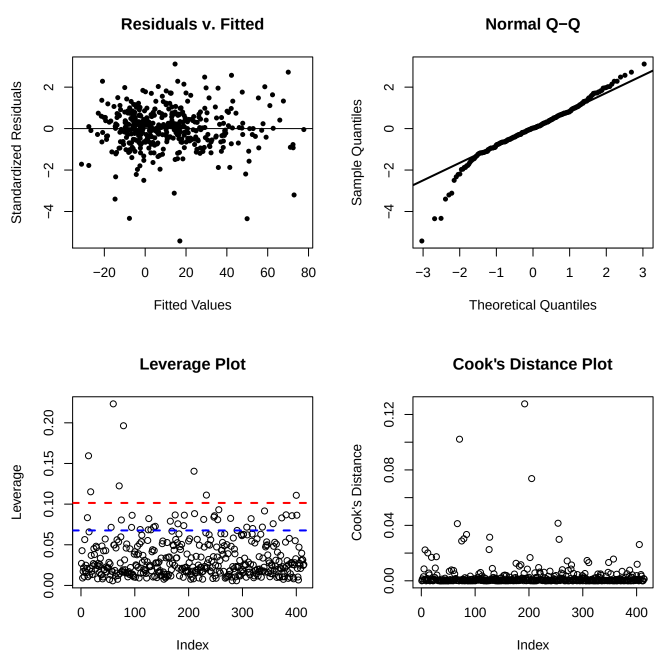


Figure E.11: Diagnostic plots for model with delta natural gas as the response for the cold climate homes.