University of Cincinnati

Date: 3/25/2019

I, Sai Santosh Yakkali, hereby submit this original work as part of the requirements for the degree of Master of Science in Civil Engineering.

It is entitled:

Decomposing Residential Monthly Electric Utility Bill Into HVAC Energy Use Using Machine Learning

Student's name: **Sai Santosh Yakkali**

This work and its defense approved by:

Committee chair: Julian Wang, Ph.D.

Committee member: Hazem Elzarka, Ph.D.

Committee member: Jiaqi Ma, Ph.D.

32839

Decomposing Residential Monthly Electric Utility Bill Into HVAC Energy Use Using Machine Learning

A thesis submitted to the
Graduate school
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Master of Science in Civil Engineering
in the Dept. of Civil & Architectural Engineering &Construction Management
of College of Engineering and Applied Science
by

Sai Santosh Yakkali
Bachelor of Technology in Civil Engineering,
Jawaharlal Nehru Technological University, Hyderabad, 2015

Committee Chair:
Julian Wang, Ph.D.

Committee Members:
Hazem Elzarka, Ph.D.
Jiaqi Ma, Ph.D.

March 2019

ABSTRACT

About 38% of total energy consumption in the US can be attributed to residential usage[1], 48% of which is consumed by Heating, Ventilation and Air Conditioning (HVAC) systems. Inefficient operation of energy systems in residential sector motivates many researchers to develop an easy and affective method to educate consumers and reduce inefficient usage. A detailed energy bill is proven to motivate users to reduce energy consumption by 6-20% [9]. Further, a system or device level energy consumption data can be used to propose energy saving practices. Information of HVAC usage alone can trigger a big saving, as about half of total consumption is HVAC. However, existing methods to disaggregate usage rely on sensors or meters at the either device or central power-level, which hinders the utilization for home owners. Alternatively, information about monthly electric utility is normally accessible for households, which may be utilized to attain HVAC energy use through data mining techniques. In this study, machine learning is used to construct a regression model to accurately estimate HVAC energy used based on monthly electricity used (from utility bill), home profiles, and monthly weather data. The main dataset used for training and testing the model is from the Pecan Street home energy use dataset.

ACKNOWLEDGEMENTS

Firstly, I would like to express my heartfelt gratitude to my advisor – Dr. Julian Wang and thesis committee members – Dr. Hazem Elzarka, and Dr. Jiaqi Ma, for their patient guidance and help throughout this project. I am grateful to each one of them. Thanks are due to all the professors of CAECM department of the University of Cincinnati. I feel humbled and fortunate to have support of such distinguished team and constant mentoring from my advisor Dr Wang.

I would like to thank Praneeth, Raghuram, Anjana, Azharuddin for providing their inputs throughout my thesis. I would like to thank all my close friends – Kavya, Pavan, Pradyumna, Siddharth, Srichand, Hemanth, and Harshith, whose support and encouragement is invaluable and has helped me complete my thesis.

Most importantly, I would like to thank my parents – Uma and Koteswara Rao whose sacrifices and unconditional love helped me in pursuit of my dreams. I would also like to thank my sisters Bhavani, Sneha and their family members for their love, encouragement and immense support.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND NOTATIONS

A	Area of house

E_c	Monthly cooling usage per Area of the house
E_h	Monthly heating usage per Area
E_t	Monthly total energy consumption per Area
E_i	Electricity consumption at ith minute
Y_r	Age of house
h_{hd}	Heating Degree Hours with temperature less than 65^{o} F
h_{cd}	Cooling Degree Hours with temperature greater than 75^{o} F
h_{hh}	High Humidity Hours with relative humidity greater than 60%
h_{lh}	Low Humidity Hours with relative humidity less than 30%
h_{hw}	*High Wind Hours with wind speed greater than 7.5 mph*
h_{lw}	Low Wind Hours with wind speed less than 7.5 mph
x	Humidity Ratio: Mass of water vapor in air per mass of dry air (Monthly)
T_{wb}	Wet bulb Temperature (^{o}C)
T_{dew}	Dew Point Temperature (^{o}C)
T_{db}	Dry bulb Temperature (^{o}C)
ΔT	Cumulative of difference between dry bulb and wet bulb temperatures
P	Atmospheric Pressure (kPa)
P_{ws}	Average Saturated Vapor Pressure

$\mathit{\Upsilon}$	Psychrometric Constant
$\mathit{\Delta}$	Delta
$Avg\ T_{wb}\ T_{db}$	Average of Dry-bulb and Wet-bulb Temperature (^{o}C)

1 INTRODUCTION

1.1 Background

Today energy conservation is a major problem due to exponential increase in energy demand. With a rapid increase of 4% in total US households from 2009 to 2015, today residential sector consumes about 22% of total energy more than commercial buildings' (18%) according to an analysis results released by the U.S. Energy Information Administration (EIA). As seen in figure (1) from the analysis of EIA, nearly 47% of the energy consumed in households is used by HVAC systems [1, 2]. HVAC in commercial sector takes only 34% of its total share which is further less than residential usage.

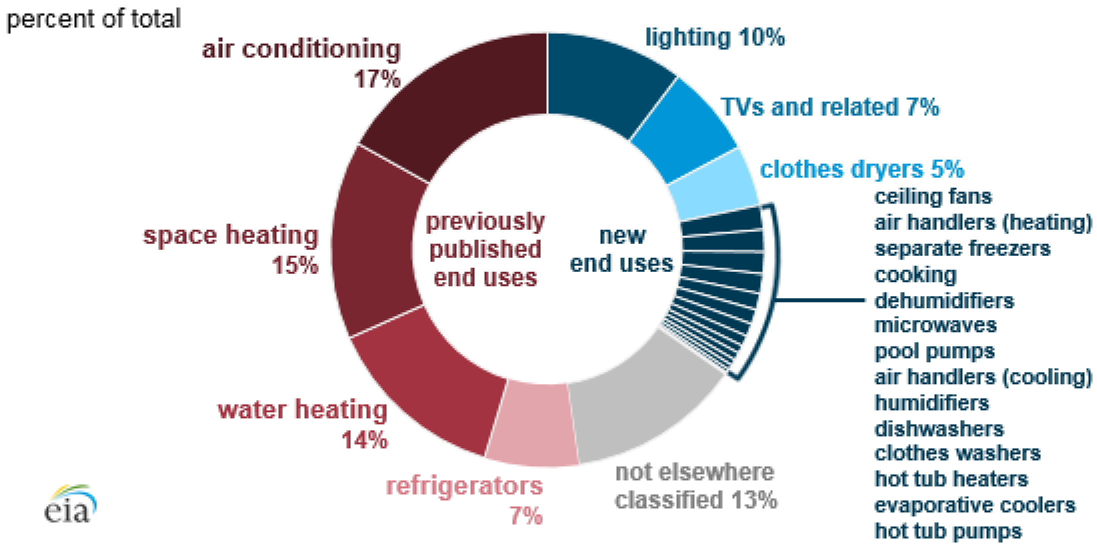


Figure 1: Pie chart of energy consuming classification in Residential sector

© U.S. Energy Information Administration, 2015 Residential Energy Consumption Survey

Numerous factors influence the power use in residential sector, including climate conditions, building size, building properties, for example, window and building envelope properties, infiltration and ventilation, occupant behavior, and so on. However, in the U.S., the HVAC is among the biggest energy consuming end-use in a home; its energy demand is related with the measure of heat gains and losses in a building, just as its size, efficiency, indoor thermostat setpoints, and the nearby weather conditions in which it works. Here comes the biggest problem faced in residential sector consumption that most of them do not operate in an energy efficient manner, such as utilizing natural ventilation, adjusting thermostat according to weather conditions, etc., thus leading to a high use than required.

Further classification shows air-conditioning relatively takes less share compared to space heating depending on the climate zone. On average air-conditioning accounts for 17% of the total usage, but the space heating takes over 30% of it.

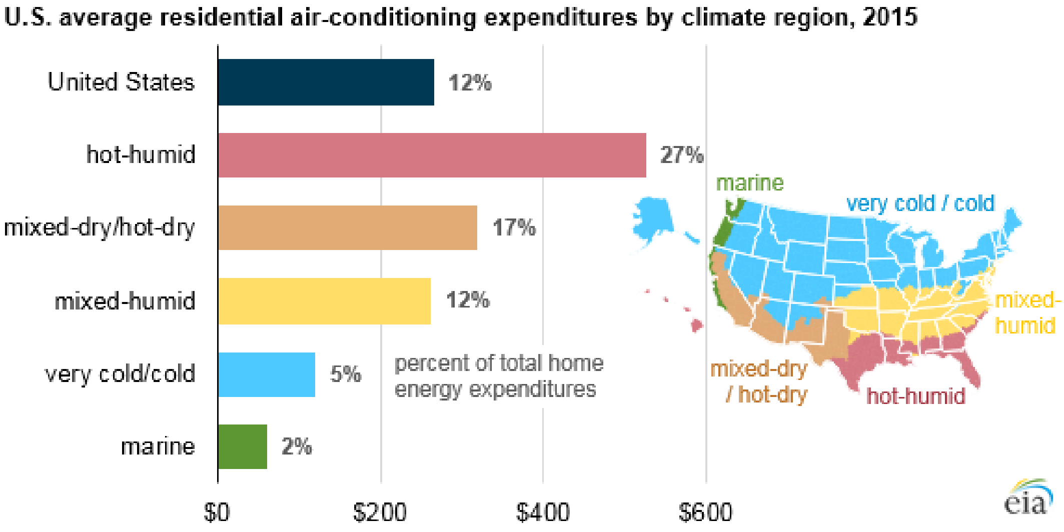


Figure 2: Percentage share of air-conditioning by zone in USA

© U.S. Energy Information Administration, 2015 Residential Energy Consumption Survey

A study has shown that improper HVAC system usage and maintenance lead to 35% of the houses using ventilation in an inefficient manner [20]. In addition to the purposes of facilitating energy saving behaviors, one needs information of HVAC energy usage to make decisions of home energy-efficiency upgrades related to system replacement, enclosure retrofits, etc. One way to educate household and create awareness is by comparing their HVAC usage with optimum usage for the location's weather and building profile or through normative energy use feedbacks [10,11]. A study has proven to reduce 6-20% of HVAC consumption by just providing a detailed energy bill. Therefore, before conducting such energy saving strategies or actions, it is indispensable to attain the HVAC energy use.

1.2 Research Overview

Numerous researchers have studied this problem which generally falls into the field of energy disaggregation. The traditional method being Intrusive Load Monitoring [3] (ILM), where target appliance systems are fixed with sensors. Other and most popular method used for energy disaggregation is Non-Intrusive Load Monitoring [4,5,6](NILM) where a sensor is fixed at a signal monitoring point, generally at central power supply meter. Energy signal data coming from the sensor is fed into energy disaggregation algorithms. Nonetheless, all these existing studies need hardware installations to obtain data [4,12,32] and then decompose energy signals, making the solution of energy disaggregation costly and invasive. There are few statistical models that can predict total energy consumption without complex hardware installation, it still needs lot of minute details [7,8], which is either impractical to find a valid equation for large number of houses or susceptible for inaccuracy as the data is coming from relatively small and highly controlled sample size [3,8].

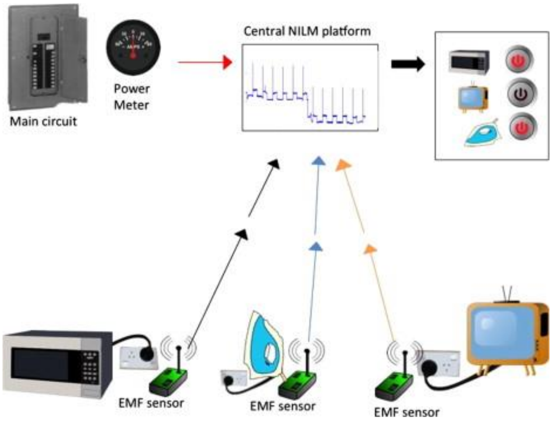


Figure 3: Describing the basic methodology of NILM

The question that comes immediately to mind is whether we need such complex machine learning-based models, specifically do we need deep-disaggregated HVAC use, like minute time-series data? In our hypothesis, for the purposes of energy-efficient upgrades or strategies rather than system faulty detection or monitoring, monthly information of HVAC energy use is more useful and effective. The reason is on three-fold. First, from the practical perspective, minute or hourly HVAC energy use is mainly for fault detection or smart grid design [14,15,16].

Second, the monthly utility data is readily accessible for most households, which informs home owners of much related information including utility rate changes, expenses, and energy use percentiles in that district. The information has been widely used in other studies for facilitating energy savings in homes [8,22].

Third, building energy use is normally associated with outdoor weather conditions which also often in monthly formats and patterns. Because this project is under a large research scope for the specific goal - facilitating household energy saving activities, in this study, we aim to develop a simple predictive model for the air conditioning energy used upon the monthly electric utility data.

From basic concepts of psychrometry, building properties and heat transfer we aim to develop new variables that explains HVAC usage more effectively much closer to actual building behavior.

Previous research on this topic have only considered pre-existing parameters which we observed did not perform well to explain cooling or heating usage. The implication is that we do not need hardware installation or other complex models to estimate cooling & heating energy use and in turn affect household energy efficiency.

1.3 Research Objectives

In this research we aim to propose an easy methodology for energy disaggregation into HVAC energy usage using monthly electricity utility bill. The existing model are either too complex implement in a regular household or the implementation itself requires many features like percentage of wall to window ratio, etc. We aim to develop a statistical model where data collection from user end is as simple and minimum scope of human error as possible. This approach will integrate understanding of building physical principles with statistical model to make an educated prediction which can be generalized to most of the houses and climate zones in the USA. By the end of this research, a clear understanding between different weather features and their influence on energy usage. By making a statistically correct model we can generalize the findings to all the houses and regions that comply with the assumptions this research made.

1.4 Significance

Many research studies indicate energy saving behaviors may lead upto 20% energy use reduction just by educating consumers with a more detailed appliance-level utility bill [18,19].

HVAC being a major part of energy used in residential sector, we hypothesize more tailored energy feedback messages (i.e. potential energy savings and associated utility bill savings by actions) for individual household will stimulate the household to adjust their behaviors and habits to be energy efficient.

In order to perform the energy saving calculations by energy saving activities, such as thermostat set points adjustment, using natural ventilation, it is imperative to know the information or performance of existing HVAC energy use. This was the initial motivation of our research on energy disaggregation. The significance of this thesis can be summarized into:

- Unique approach to disaggregate utility bill into HVAC using weather and building physics
- Developed and have adopted new weather features like Degree Hours, Humidity Ratio, etc.
- Robust filtering and selection of features for model using statistics and building behavior concepts
- Model to use simple, easy to access, open source features so that it can widely be adopted
- Be able to calculate percentage use of HVAC in utility bill and thereby educating households on what percentile they stand in consumption

To achieve a convenient and user-friendly method, the goal of developing this energy disaggregation technique is to estimate the HVAC energy use upon a few input data from home owners. It is expected that the input variable should be easily and readily accessible, such as location, home size, home age, monthly utility bill, and weather data among which the basic home profiles such as home size and age can be obtained through public construction information or public record.

2 LITERATURE REVIEW

Many research papers have proposed several methods in approaching this issue of HVAC prediction / disaggregation. Statistical models are generally used to estimate total energy consumption, but none have used its applications to disaggregate energy to estimate HVAC usage. With that, energy disaggregation methods can broadly be grouped into three (Figure 4).

1) ILM (Intrusive load Monitoring)

2) NILM (Non-Intrusive Load Monitoring)

3) Statistical Model

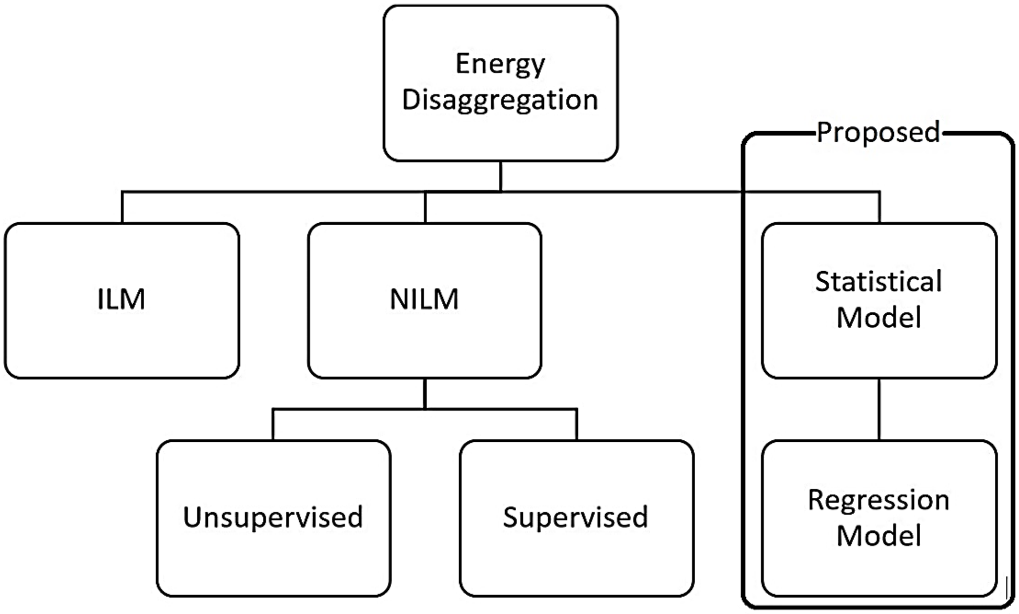


Figure 4: Tree diagram outlining different approaches available and proposed for energy disaggregation

2.1 Intrusive Load Monitoring

The first method, commonly known as ALM (Appliance Load Monitoring), is a technique for appliance-specific energy consumption statistics that can further be used to devise load scheduling strategies for optimal energy utilization[4,31,32]. This primarily focuses on classifying electrical events. There has not been much research or practical implementation of this approach, at least after NILM was introduced in 1991. The drawback of this method is that it requires many sensors to monitor the energy usage. Further, devices with low power consumption cannot brand significant spike in electrical signal. Some researches came with very high frequency sampling solutions to overcome the drawback[33,34,35], but the hardware to record such low time interval high frequency sampling is very expensive making this an unaffordable solution. Sala, Enric, et al.[4], took a direct approach by installing sensors to all the HVAC equipment and tracking their load profiles for further study. There is no generalized solution that applies to all houses in this method.

2.2 Non-Intrusive Load Monitoring

Prominent method being NILM, where a single sensor is installed at main supply unit. Each appliance has a unique energy signature, this technique takes advantage of the periodical yet unique energy signal to classify and disaggregate the electrical loads to appliance specific power consumption. As the data is acquired only at single point without installing sensors for each appliance inside customer's house, this method is called Non-Intrusive Load Monitoring Method. The problem with these approaches is the load profile will vary depending on the manufacturer and the technology used, both of which are very dynamic. Which means, most of the time the model needs calibrations for new settings, neither are they less expensive methods. NILM can further be classified based on how the aggregate energy signal obtained from the sensor is disaggregated – Supervised and Unsupervised.

2.2.1 Supervised Algorithms

In a supervised technique, labelled energy signal (the appliance source of the signal) is fed into a machine learning algorithm, using pattern recognition techniques the algorithm is trained and further optimized to disaggregate similar patterns from aggregate energy signal[24,25,26]. This requires large training data to calculate the coefficients and to capture possible times that the same device may operate. Supervised approach is a bit laborious work, but it has some obvious advantages over unsupervised method. Unsupervised sometimes cannot pick appliance signal with small energy signature which is very much manageable in case of labeled data input.

2.2.2 Unsupervised Algorithms

Unsupervised method on the other hand is not trained but the algorithm is developed to automatically determine the total number of appliance clusters from the signal data and each cluster is assumed to be a linear combination of multiple appliance sources, which are further broken down into individual sources[4,30,35]. Some research papers were able to predict HVAC usage up to daily, weekly, and monthly usage with 90% accuracy. For instance, Li, Nan, et al.[3] developed a predictive model of HVAC energy consumption in commercial buildings using multiagent systems, which took a similar approach by observing a commercial space for 267 days and used Multi-Agent System (MAS) tool to simulate HVAC usage. It is noted that this duration of observation is very small and with only one building, the study lacks generalizing the model to buildings with other weather and building type situations.

Perhaps, the most common method of disaggregation is proposed by Perez et al [23], where 19 newly constructed houses where equipped with smart sensors and a sub-meter for HVAC cooling to record consumption with one-minute interval. Training periods are designed where cooling is dominant load to acquire enough data for parameters estimation. Once the parameters are developed, they are likely to remain same unless the cooling system undergoes significant change. This works on a simple formula of –

$$\Delta \boldsymbol{E_i} = \boldsymbol{E_{i+1}} - \boldsymbol{E_i} \tag{1}$$

Where ΔE_i is the change in electricity consumption, E_i is the consumption at i^{th} minute.

Using the algorithm showed below in Figure (5), they were able to identify and differentiate the signal from HVAC and hence develop the parameters that can be used for disaggregation.

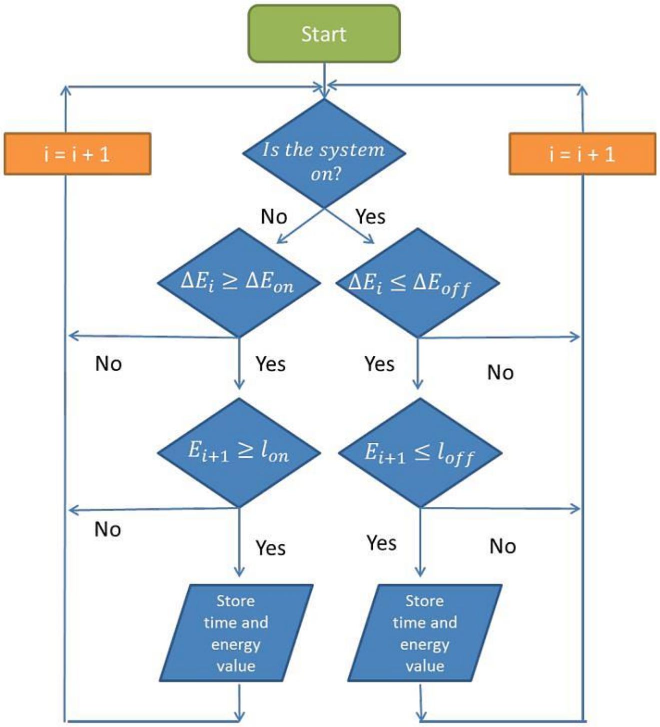

Start
i = i + 1
Is the system on?
i = i + 1
No
Yes
$\Delta E_i \geq \Delta E_{on}$
$\Delta E_i \leq \Delta E_{off}$
No
Yes
Yes
No
$E_{i+1} \geq l_{on}$
$E_{i+1} \leq l_{off}$
No
Yes
Yes
No
Store time and energy value
Store time and energy value

Figure 5: Algorithm proposed by Perez et al 23

© *Nonintrusive Disaggregation of Residential Air-Conditioning Loads From Sub-Hourly Smart Meter Data*

Eq. (1) is quite easy approach. Despite many research advancements in the field of energy disaggregation, the requirement of new hardware installation, expert involvement has not been compensated. This makes consumer to pay more initially to have a clear knowledge on their energy using habits, moreover this still keeps the consumer dependent on expert consultation every now often.

Ansari, et al.[7] & Yan, Xiao et al.[8] had more technical approach dealing with physics and electricity property, they proposed to estimate the load on cooling system & to find out energy performance assessment of an existing building, nonetheless these methods have very complex input variables like asking consumer for window properties, orientation, insulation properties. This makes it hard for applications to be user-friendly.

Moreover, these papers were based on data with very few buildings. Sonderegger, et al.[9] proposed a modified model from other researches by adding weather related components to the equation. This is based on a statistical model where the equation relays on large previous data. The model classifies buildings into conventional, weather sensitive, & utility dominated buildings to better understand the building's behavior. But the only drawback of this research is it requires detailed properties of the building such as occupancy, meals served during the month, widget produced.

Thus, several researches made great contributions in the topic of energy disaggregation some of which are more technical and precise, but lack simplicity. Most of the projects still haven't reached the consumer market and the projects available in market are complex which require great investment making it harder to educate public regarding system-wise energy usage.

2.3 Statistical model

The third method, we adopted for our research is statistical approach. A large data is used to develop a regression model, generally with one unknown variable with two or more independent variables. There are many successful efforts in estimating total electricity consumption using Regression model and Genetic Algorithm techniques. The main backdrop of this approach is difficulty is obtaining large historical data, except that the accuracy of this methods is on par with other NILM methods which require initial investment. Hausfather, et al[28] proposed a novel method to predict energy consumption by developing a statistical model for residential houses in the USA. This model predicts usage based on ZIP code. However, the accuracy of the model is very less compared to other models with R^2 ranging from 0.3 to 0.59.

Amber, et al[29] developed a multiple regression equation for various building types to predict energy usage with 6 independent variables – Temperature, solar radiation, humidity, windspeed, weekday index, building type with 5 years data. Among these, three variables showed significant weightage in regression model and achieved 12% NRMSE (Normalized Root Mean Square Error). Amber, et al[29] developed this model using SPSS software. H Alton et al. (2014) adopted a similar technique to predict future energy consumption of a supermarket in the UK.

MR Braun, et al.[30] estimated the gas and electricity consumption for future years based on supermarkets' physical properties, weather variables and previous years consumption. Apart from main result of usage estimation, investigation also found out that energy will be more sensitive with respect to outside temperature than relative humidity. A similar result was found later during model development where humidity had very less correlation with both energy and HVAC usage.

Similarly, there are many other researches to estimate the total energy consumption of building using statistical models[36,37] with good accuracy. In many cases the accuracy of statistical models is higher compared to NILM techniques, however all these statistical models are used to predict the total energy consumption and NILMs were used to disaggregate the total energy consumption. Nonetheless, this research tries to take advantage of high accuracy of statistical models over NILM by collecting detailed weather data coupled with previous energy consumption to develop a model that estimates HVAC usage from utility bill. We aim to integrate thermodynamics of weather and building physical principles with statistical model to make an educated prediction which holds accurate for any house and climate zones in the USA.

3 COUPLED MODELING METHODOLOGY OF STATISTICS AND BUILDING PHYSICS

3.1 Motivation

A model purely based on building physics cannot explain cooling or heating usage no matter how accurate the features be selected. There are many features to be taken into consideration like radiation from sun, building direction, window-wall ratio, occupant behavior, insulation by walls, temperature, humidity, and many more. There is no exact math or equation to translate all these features to predict HVAC usage tailor made for a house, because there are too many anomalies which can never be fit into an equation.

Similarly, a simple statistical model obtained by dumping all readily available data cannot explain the physics behind the nature of HVAC. For example, a simple input of average temperature of the month is not enough to extract hourly fluctuations in temperatures which trigger HVAC usage. It requires deeper understanding of both physical and mathematical concepts to make a valid model suitable to predict the desired.

This brings us to the necessity of combining both building physics and statistical modeling. Statistical methods are highly reliable to predict an information given right feature that affect the outcome. Choosing accurate features from concepts of building physics that really affect the outcome and build a model will result in predictions with great precision. The features like HR, Age, CDH instead of CDD, considering humidity, and wind speed have a crucial role in building physics to estimate cooling or heating load, but were never considered to develop a statistical model with.

3.2 Psychrometry

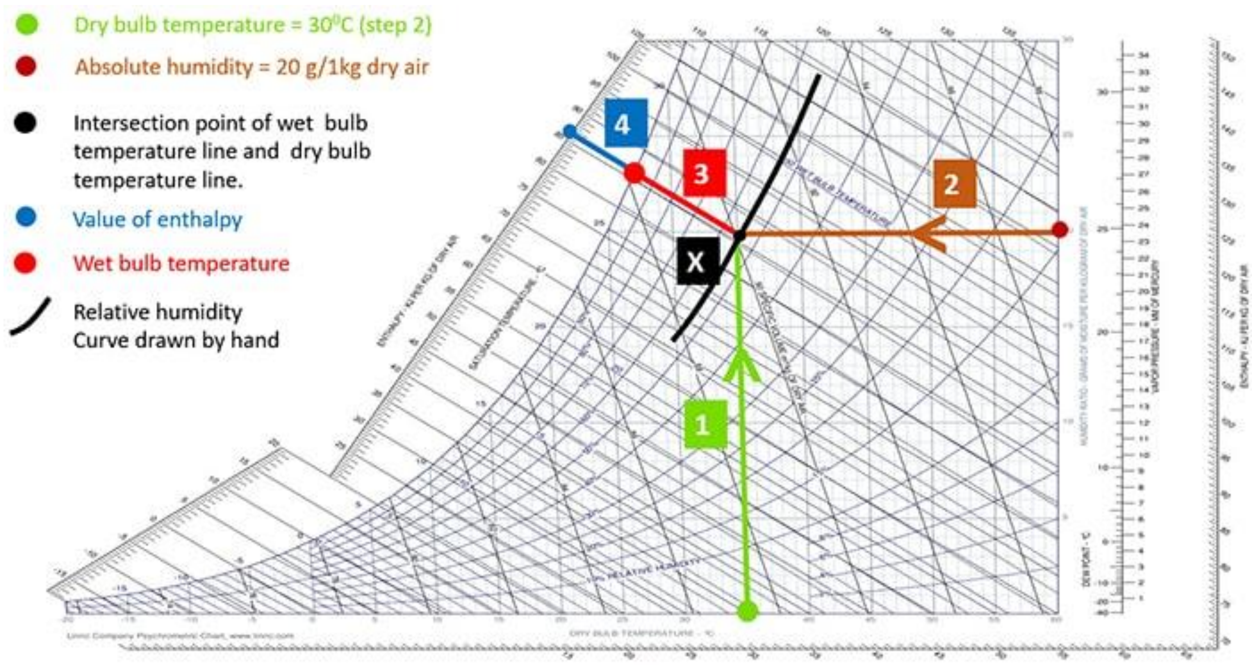


Figure 6: Psychrometric chart

The behavior of buildings based on weather can be understood using psychrometry. Figure (6) shows the psychrometric chart where many of the variables and their ripple effect can be explained. The team understood Humidity ratio plays an important role in human comfort in the building using psychrometric chart and have adopted into the study. Some other features and their importance in regards with human comfort are discussed below.

Dry-Bulb Temperature: The dry-bulb temperature is the temperature indicated by a thermometer exposed to the air in a place sheltered from direct solar radiation. The term dry-bulb is customarily added to temperature to distinguish it from wet-bulb and dewpoint temperature.

Wet-Bulb Temperature: The thermodynamic wet-bulb temperature is a thermodynamic property of a mixture of air and water vapor. The value indicated by a wet-bulb thermometer often provides an adequate approximation of the thermodynamic wet-bulb temperature.

Dew Point Temperature: The saturation temperature of the moisture present in the sample of air, it can also be defined as the temperature at which the vapor changes into liquid (condensation).

Relative Humidity: The ratio of the vapor pressure of moisture in the sample to the saturation pressure at the dry bulb temperature of the sample.

Temperature is not the only climate variable that affects energy/HVAC consumption. Generally, in case of cooling load it is the rate at which sensible and latent heat must be removed from the space to maintain a constant space dry-bulb air temperature and humidity. Sensible heat in the space causes its air temperature to rise while latent heat is responsible for the rise of moisture in the space. The building design, internal equipment, occupants, and outdoor weather conditions may affect the cooling load in a building using different heat transfer mechanisms. Relative and total humidity have a significant impact on energy consumption because humidity is altered as a part of HVAC and energy is needed to extract moisture from the air. This is because moisture is capable of holding heat and high humidity during summers cause suffocation.

Heat and moisture inside buildings come from sources like buildings' occupants, equipment, and appliances, and HVAC systems work harder to remove both heat and moisture from the air. Although tightly sealed buildings would contain conditioned air better, showed that decreases in natural ventilation would increase in indoor moisture levels. Thus, buildings without operable windows are better samples to study the impact of weather on energy demand. It is observed in some cases that managing humidity will be more important than temperature because humidity has greater impact on the thermal comfort for people in buildings.

Other variables considered to have effect on heating and cooling load combined are relative humidity, difference between dry-bulb and wet-bulb temperature, wet-bulb temperature alone.

Relative humidity (RH) also has substantial effect on human comfort levels and thus on HVAC load. RH is a measure of air's ability to absorb moisture and thus it affects the amount of heat a body can dissipate by evaporation. High RH slows down heat rejection by evaporation, especially at high temperatures, and low RH speeds it up. The desirable level of RH is the broad range of 0.30 to 0.70, with 0.50 being the optimum. At this desired RH, most occupants feel neither hot nor cold, and does not need to activate any of the defense mechanisms to maintain the normal body temperature.

In addition, many of the factors are correlated with one another. For example, occupants' ambient temperature is influenced by outdoor temperature and energy rate, and real-time energy demand and energy supply is often disconnected.

3.3 Building Physics Principles

The principles of building physics combine the concepts of psychrometry, applied physics and building construction engineering to investigate the energy efficiency of old and new buildings. The application of building physics allows the construction and renovation of high performance, energy efficient buildings, while minimizing their environmental and energy impacts.

Building physics addresses several different areas in building performance including air movement, thermal performance, control of moisture, ambient energy, acoustics, light, climate and biology. This field helps us understand the principal aspects of a building's indoor and outdoor environments so that an eco-friendlier standard of living is obtained. Building physics by combining the sciences of architecture, engineering and human biology and physiology helps us estimate and predict energy loads and how change in one aspect affect others. Building physics not only addresses energy efficiency and building sustainability, but also a building's internal environment conditions that affect the comfort and performance levels of its occupants.

Although the principles of psychrometry apply to any physical system consisting of gas-vapor mixture, the most common system with high importance is the mixture of water vapor and air because of its application in HVAC. In occupants' terms, our thermal comfort is in large part a consequence of not just the temperature of the surrounding air, but the extent to which that air is saturated with water vapor and also the rate of fresh air replacing the air around us. Variables of psychrometry are the ideal example to explain building physics and how change of one weather component affects the other. Here is the psychrometry chart and brief concepts behind the chart to better understand the principles of building physics.

3.4 Variables and Significance

The modeling approach in this study represents a combination of physical functions derived from first principles and statistical analysis on experimental data and their structure and features. For instance, to increase the sensitivity of model to weather, this research tested out a new concept of Heating Degree Hours (HDH), Cooling Degree Hours (CDH), and extended it to humidity, and wind speed: High Humid Hours (HHH), Low Humid Hours (LHH), High Wind Hours (HWH), Low Wind Hours (LWH). The resultant model proved that even strong correlation and harmony with total energy consumed.

The more common form of finding how much cooling/heating required is by looking at Cooling Degree Days (CDD) where the number of degrees that a day's temperature is average temperature is less that 65ºF, similarly Heating Degree Days (HDD). But this shrinks the scope of 24 hours usage to one average value. This hinders the possibility to find whether few hours of day being required to have cooling if the average value is more than 65°. Thus, a variable sensitive to hourly weather conditions has been adopted to calculate cooling degrees (CDH). This value was much closely related to cooling usage as hypothesized. Here is the table explaining all the variables considered for model in both cooling and heating dataset.

After deciding upon to continue with degree hours instead of degree days, considering the sensitivity of degree hours to accurately represent HVAC demand, the next step was to test and verify between extracting the degree hours with dry-bulb or wet-bulb temperatures. Preliminary model with features extracted from both the temperature measurements have been developed and tested. The accuracy of the models was very close with dry-bulb temperatures having more coherent and least error amongst both. The features selected for dry-bulb model covered all aspects of weather and building characteristics, making it the best fit for the data obtained.

The initial set-points (Threshold points) for variables are adopted from ***"Efficient Comfort Conditioning"*** [1], but the temperatures obtained from Pecan are outside temperature. As shown in Table (1) the team tested various adjacent setpoints and concluded with optimum values where the correlation with HVAC usage was maximum, indicating the inside house temperature for comfort.

Table 1: Testing different threshold points for variables against HVAC usage

	Set-points	Correlation with HVAC
CDH **(tested against Cooling Usage)**	$T_{db} > 65$ °F	53%
	$T_{db} > 75$°F	**61%**
	$T_{db} > 85$°F	34%
HDH **(tested against Heating Usage)**	$T_{db} < 55$°F	40%
	$T_{db} < 65$°F	**56%**
	$T_{db} < 75$°F	33%

During the hours of high wind (greater than 7.5 mph), the building has more infiltration from outside. This higher infiltration rates can decrease the necessity of cooling due to natural air circulation from outside. However, Wind speed alone cannot be considered, as higher infiltration of outside air with less humidity doesn't have much influence compared to a high humid air. Thus, both variables with High Wind Hours and High Humidity Hours have been adopted aside from temperature as they highly influence the thermal comfort inside the house.

Similarly, during winters infiltration due to higher windspeed will cause more heating use and lower windspeed causes stagnation inside the house. This stagnation will keep the house warmer thereby settling down with a low heat requirement. Likewise, higher or lower infiltration combined with humidity causes different comfort levels within the house. This made HHH, LHH, HWH, LWH important variables to describe human comfort levels inside the house along with CDH and HDH.

Further, the cooling model needed deeper experimentation to improve the results like involving new variables such as Humidity ratio (HR), Wet bulb temperature (Twb), and aggregate of difference between Dry bulb and Wet bulb temperatures. Wet bulb temperature accurately represents humans comfort levels as it takes humidity into consideration. And combined with humidity ratio, which is responsible for holding either heat or making it feel much cold than it actually is, the model can make a much better prediction on cooling usage.

In the case of heating, apart from electric heaters most of the houses rely on gas heating. For this problem we have converted the gas consumption into equivalent electricity usage. The small backdrop by just representing gas usage to electricity is, the amount of heat produced by an electric heater for 1 kW-h is not equal to the heat produced by a gas heater. Being it too complex to convert to an equivalent amount of heat produced, the electricity units have been adopted. Several studies facing the same problem are observed to adopt the same system. The final variables with their threshold points are listed below in Table (2)

Table 2: List of variables used for analysis

Variable	Symbol	Meaning	unit
Area	A	Area of house	Sq. Ft
Cooling/sq.ft	E_c	Monthly cooling usage per Area of the house	kWh/sq.ft
Heating/sq.ft	E_h	Monthly heating usage per Area	kWh/sq.ft
Use/sq.ft	E_t	Monthly total energy consumption per Area	kWh/sq.ft
Age	Y_r	Age of house	Year
HDH (T < 65)	h_{hd}	Heating Degree Hours	Hrs
CDH (T >75)	h_{cd}	Cooling Degree Hours	Hrs
HHH (RH > 60%)	h_{hh}	High Humidity Hours	Hrs
LHH (RH < 30%	h_{lh}	Low Humidity Hours	Hrs
HWH (speed >7.5 mph)	h_{hw}	High Wind Hours	Hrs
LWH (speed < 7.5mph)	h_{lw}	Low Wind Hours	Hrs
Humidity Ratio (HR)	x	Mass of water vapor in air per mass of dry air (Monthly)	Kg_w/kg_a
T_{wb}	T_{wb}	Average wet bulb temperature (Monthly)	°F
T_{db-wb}	ΔT	Cumulative of difference between dry bulb and wet bulb	°F

3.5 Predictive Modeling Methods

The basic principle of machine learning is to build predictive models that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. The model describes the relationship between a dependent variable Y (in this case cooling or heating use) as a function of one or more independent variables Xi (called the predictors).

The resultant model developed after this study is aimed to involve only readily accessible data, which differs to most other works that require either hardware installations for sensing and monitoring or complex computations on much more input variables related to power signals, hourly data, and others. Even though the hourly data of energy use is available in the Pecan Street database, accumulated monthly data upon the given dataset is been used in the modeling procedure. This is to make a simple method where the households do not need any changes to their current electricity metering, zero investment, and to get HVAC energy use as precise as possible by asking minimal information like geographical information, electricity bill information, area of the building, and its age.

As mentioned before, to predict the HVAC energy consumption we adopted linear regression modeling technique. Where previous energy data of houses and other parameters are used to predict cooling and heating usage. This method falls under supervised data analysis in machine learning. Statistical correctness of such model ensures reliability of model to predict the necessary on new data, thus generalizing the findings to all suitable scenarios.

Some of the other prominent methods of machine learning which suits the predictive purpose, their methodology will be briefly discussed in coming sub-sections:

3.5.1 Regression Model

In statistical modeling, regression analysis is a group of statistical processes for estimating the relationships between dependent and independent variables. It includes many procedures for modeling and analyzing several variables, when the attention is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are kept unchanged.

In polynomial regression, the relationship between independent and dependent variables is assumed to have nonlinear relation to n^{th} degree. A nonlinear relationship between the value of x_i

and the corresponding conditional mean of y, denoted f(y/x), and has been used to describe nonlinear phenomena.

Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function f(y/x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

Among many classifications with in regression based on forms of variables adopted, having one dependent variable and multiple predictor variables with quadratic terms – the current considered form of model for this research falls under multiple polynomial model with second degree or simple multiple quadratic model.

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In simple linear regression, the model

$$y = \beta_0 + \beta_i x_i + \beta_{ii} x_i^2 + \beta_{ki} x_k x_i + \varepsilon \qquad \textit{where } i=1, 2, \ldots, n \textit{ and } k = 1, 2, \ldots, n-1 \qquad \textbf{(2)}$$

Underlying Assumptions of Regression Analysis:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The independent variables (predictors) are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.

- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

3.5.2 Random Forest

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

The general method of random decision forests was first proposed by Ho in 1995. Ho established that forests of trees splitting with oblique hyperplanes can gain accuracy as they grow without suffering from overtraining, as long as the forests are randomly restricted to be sensitive to only selected feature dimensions. A subsequent work along the same lines concluded that other splitting methods, as long as they are randomly forced to be insensitive to some feature dimensions, behave similarly. Note that this observation of a more complex classifier (a larger forest) getting more accurate nearly monotonically is in sharp contrast to the common belief that the complexity of a classifier can only grow to a certain level of accuracy before being hurt by overfitting.

Decision trees are simple but intuitive models that utilize a top-down approach in which the root node creates binary splits until a certain criterion is met. This binary splitting of nodes provides a predicted value based on the interior nodes leading to the terminal (final) nodes. In a classification context, a decision tree will output a predicted target class for each terminal node produced.

Although intuitive, decision trees have limitations that prevent them from being useful in machine learning applications. Decision trees tend to have high variance when they utilize different training and test sets of the same data, since they tend to overfit on training data. This leads to poor performance on unseen data. Unfortunately, this limits the usage of decision trees in predictive modeling. However, using ensemble methods, we can create models that utilize underlying decision trees as a foundation for producing powerful results.

Regression trees are used as base learners in the RF regression algorithm. After selecting the number of trees in the forest, each regression tree is grown on a separate bootstrap sample derived from the initial training data. Each node in a tree represents a binary test against the selected predictor variable. The variable is selected to minimize the residual sum of squares for the examples flowing down both branches. Terminal nodes contain no more than the specified maximal number of examples from which the target value is obtained by averaging. In order to avoid high correlations among the trees in a forest, a procedure of selecting the best splitting predictor in each node of a tree is modified to choose between only m randomly selected predictors—selecting a random subspace of the original n-dimensional problem.

RF does not require an external cross-validation procedure to calculate the model accuracy. The example in the training data is not in the bootstrap sample containing about one third of the trees (the example is “out of bag”— OOB). Averaging the predictions of these trees produces the RF prediction on the example. The mean squared error (MSE) on the OOB samples gives the estimate for the general MSE on a separate test set. Additionally, OOB can be used to estimate the i^{th} predictor variable’s importance as follows: (1) make a random permutation of values for the variable concerning the examples from the OOB, and then (2) note down the increase in the MSE for the permuted OOB comparing to the original one. Here, an assumption is made that a more important variable, when permuted, will produce a greater increase in MSE when using the same regression model.

3.5.3 Artificial Neural Network

A neural network (ANN) may provide a superior solution over a traditional approach for certain classes of problems (Burke, 1991). These classes include problems in where outliers may exist, and where noise is present in the data. These are common conditions in forest inventory data. The hidden layer is the key component of a ANN because of the neurons it contains; they work together to do the major calculations and produce the output.

Each neuron takes a set of input values; each is associated with a weight (more about that in a moment) and a numerical value known as bias. The output of each neuron is a function of the output of the weighted sum of each input plus the bias.

Most ANN use mathematical functions to activate the neurons. A function in math is a relation between a set of inputs and a set of outputs, with the rule that each input corresponds to an output. Neurons in a ANN can use sigmoid functions to match inputs to outputs. When used that way, a sigmoid function is called a logistic function and its formula looks like this:

$$f(input) = 1/(1+e^{output}) \tag{3}$$

Here f is the activation function that activates the neuron, and e is a widely used mathematical constant that has the approximate value of 2.718. Most sigmoid functions have derivatives that are positive and easy to calculate. They're continuous, can serve as types of smoothing functions, and are also bounded functions. This combination of characteristics, unique to sigmoid functions, is vital to the workings of an ANN algorithm — especially when a derivative calculation — such as the weight associated with each input to a neuron — is needed.

The weight for each neuron is a numerical value that can be derived using either supervised training or unsupervised training. In the case of supervised training, weights are derived by feeding sample inputs and outputs to the algorithm until the weights are tuned (that is, there's a near-perfect match between inputs and outputs).

Neural Networks can in principle model nonlinearities automatically (see the universal approximation theorem), which you would need to explicitly model using transformations (splines etc.) in linear regression.

The ANNs are highly vulnerable to overfit as adding hidden layers or neurons looks harmless. So being extra careful to look at out-of-sample prediction performance is much needed trait.

4 METHODOLOGY

A multiple linear model has been adopted to predict the cooling and heating usage. Two separate models are generated for datasets with cooling and heating data. The variables with high correlation were observed to have a linear or quadratic relation with dependent variable, hence a multiple polynomial model as decided has a best fit for the model. A comparative study was made later to compare the results and accuracy of linear model to a random forest (RF) method. The comparison will be laid out in results section of the report, for this section a detailed procedure and methodology adopted will be discussed.

As illustrated in figure (7), after collecting data, the desired variables were derived and calculated. The pre-processing of data played a vital role in execution, apart from understanding concepts of building physics and implementing. A detailed variable testing and variables selection was conducted using statistics and evaluating their performance in view of psychrometry. The final combinations of variables were then used to develop a regression model in R programming platform. These combinations are tested for prediction, accuracy, and validation.

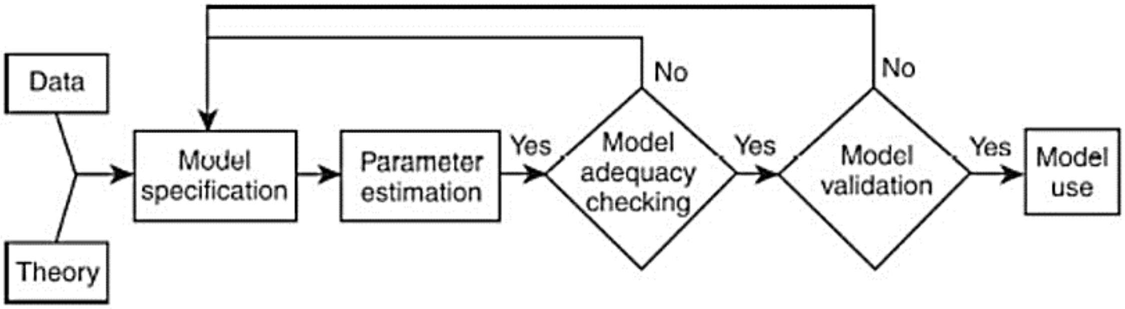


Figure 7: Flowchart illustrating general model methodology

4.1 Data

The data is collected from an open-source database, Pecan street. Pecan street measures circuit-level electricity use and generation from nearly 1000 volunteer homes located in Austin, Boulder, San Diego.The anonymized energy and respective hourly weather data are open-sourced to research. Pecan street has encouraged several researchers with a chance to work on real world energy data since its start on 2009. There were very few small open source datasets on energy consumption before with limited scope before like REDD. Pecan street provides various forms of data from 1minute, 15-minute, 1-hour time interval datasets with detailed equipment level energy consumption. Pecan street also provides the gas usage, water usages of some of the houses. This study was fortunate to find around 600 useful house energy usage statistics with the help of Pecan street, Austin-TX.

The data used comprised of around 600 houses data for cooling model and around 350 houses for heating model, along with hourly weather data of Austin, Boulder, San Diego. The metadata gives the general and building characteristics of the houses like year of construction, area, location, etc. The data needed several processing procedures in order to get a tailored data useful for analysis. Table 2 mentions column names extracted from database using SQL.

Table 3: Description of columns names used from Pecan database

Column Name	Description
dataid	The unique identifier for the home.
city	The city in which this residence is located.
state	The state in which this residence is located
house_construction_year	The year in which this home was constructed
total_square_footage	The total square footage of the home
air1	Air compressor circuit eGauge data present
air2	Second air compressor circuit eGauge data present
air3	Third air compressor circuit eGauge data present
airwindowunit1	Window unit air conditioner circuit eGauge data present
heater1	Stand-alone heater circuit eGauge data present
housefan1	Whole home fan circuit eGauge data present

4.2 Data Processing

The initially acquired data was in raw hourly format for all the houses during the span of 2013 to 2017 in cooling and in case of heating it was from 2015 to 2017. It is collected in hourly format in the first place to check detailed entries and verify for data accuracy. Some of the hourly data with faulty entries like cooling or heating data being greater than total energy usage have been filtered out. Later, the data was converted from hourly data to monthly data. Some missing data in case of year of construction was imputed with median of all the houses age.

Metadata, energy and weather datasets were merged based on year, month labels. Some houses with missing year of construction has been imputed with median age of all the houses. New variables were derived from existing raw and metadata, some of which are, from cooling to cooling/sq.ft, Age of the construction, all weather variables like CDH & HDH from temperature, etcetera.

Apart from minor derivations with available data to come up with desired variables, Humidity Ratio (HR) and wet-bulb temperature (T_{wb}) was not readily available to use. To calculate both, several other variables like saturation pressure (P_{ws}), psychrometric constant (ϒ), Delta(Δ) were calculated. The formula for wet-bulb temperature (T_{wb}) itself was an iterative one.

$$\boldsymbol{x = \frac{0-62198 \times P_{ws}}{P-P_{ws}}} \tag{4}$$

$$\boldsymbol{P_{ws} = 0.61164 \times 10^{\left(\frac{7.591 * T_{db}}{T_{db}+240.97}\right)}} \tag{5}$$

$$\boldsymbol{T_{\omega b} = \frac{T_{db} * \Upsilon + \Delta * T_{dew}}{\Delta + \Upsilon}} \tag{6}$$

$$\Upsilon = \boldsymbol{0.000665 * P} \tag{7}$$

$$\boldsymbol{\Delta = \frac{17.502 * 240.97 * P_{ws}}{(240.97 + Avg\ T_{db}T_{wb})^2}} \tag{8}$$

Where, P is the atmospheric pressure in kPa, avg $T_{db}T_{wb}$ is the average of wet bulb and dry bulb temperatures in ^{o}C.

The data some issues with outliers, especially heating dataset, where the outliers were effectively removed by Inter Quartile Range (IQR) method. Instead of just finding the extreme observations for entire data, category-wise IQR outlier was observed to be more effective and logically fitting. The heating dataset had very few observations in zones of San Diego and Boulder, so a simple month-wise categorical split was made to detect the outliers. Observations outside 1.5*IQR range in each category were considered outliers and removed. Figure (8) and (9) displays the category-wise outliers for both cooling and heating datasets using IQR concept.

$$\textit{IQR} = 3^{rd}\ \textit{Quartile} - 1^{st}\ \textit{Quartile} \qquad (9)$$

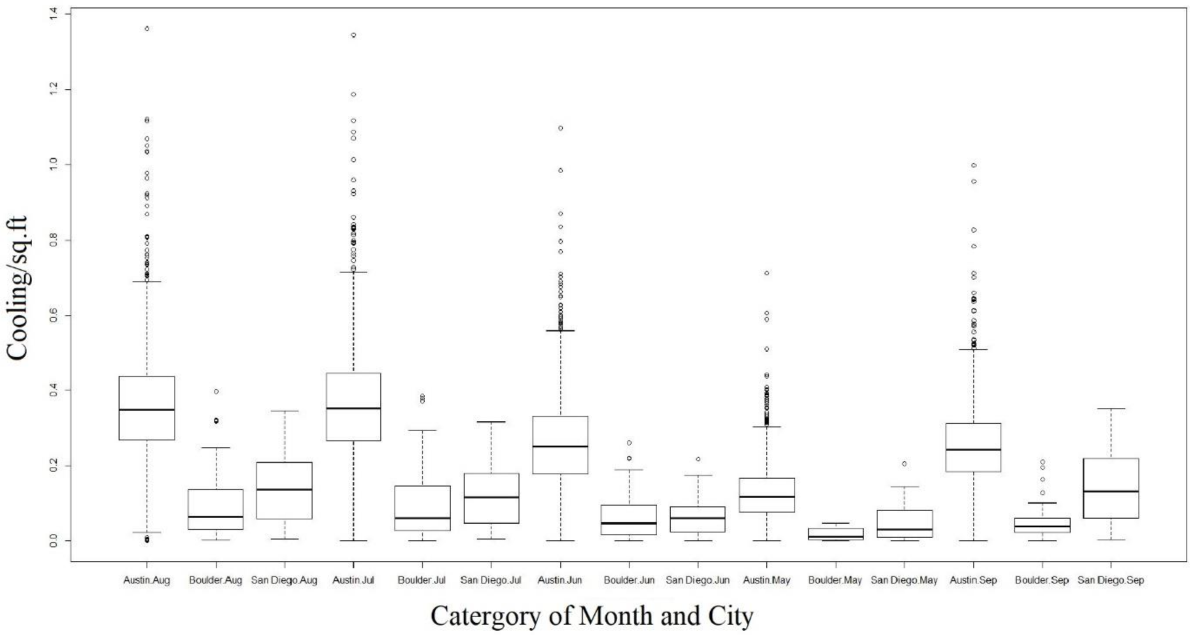


Figure 8: Boxplot representing the outliers for cooling data based on each category of City and Month

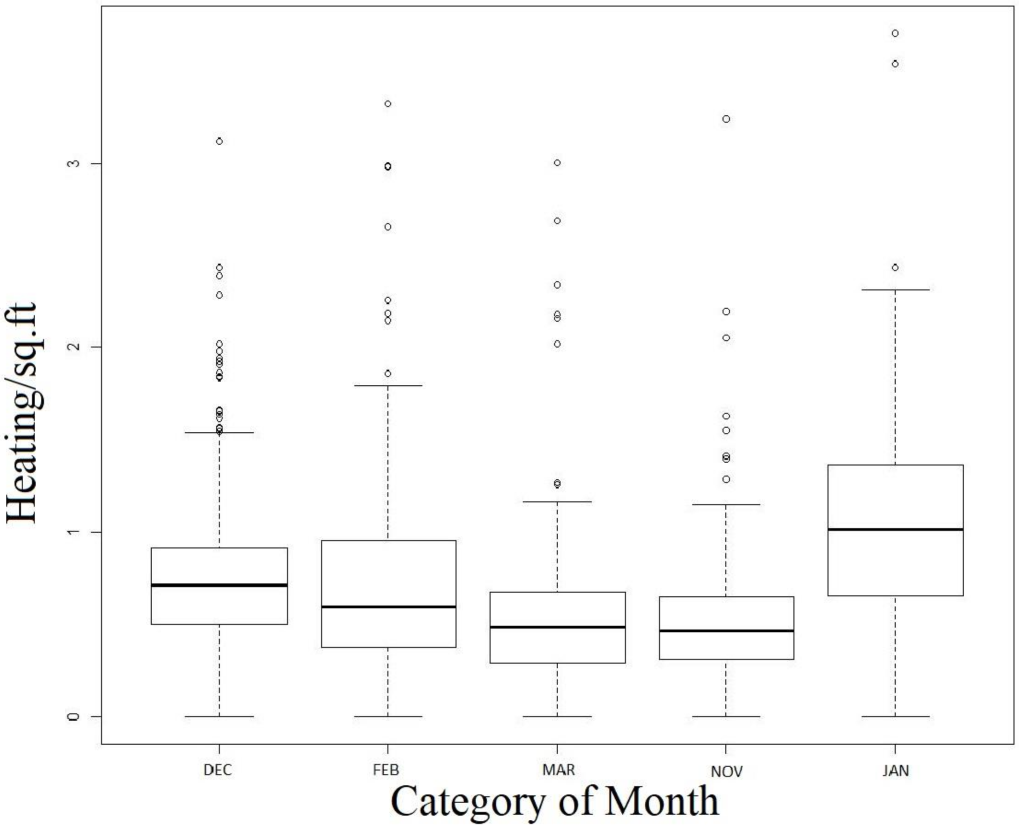


Figure 9: Boxplot representing the outliers for heating data based on each category of Month

Both the datasets have a huge difference in ranges of each variable. For example, the electricity usages per square feet have range of 0-1.5 kW-hr/sq.ft, but CDH, HDH and others have data ranging from 0-6000. These dissimilarities between the ranges did not cause any mathematical problem, neither did we have any reason to transform the data to achieve a specific distribution suitable for modeling, but due to this the coefficients of some variables in the model had a complex value like 10^{-12}. To solve this issue the data has been normalized to the common scale of 0-1. The function used to normalize the complete data is as follows.

$$\boldsymbol{x'} = \frac{\boldsymbol{x - Min(x)}}{\boldsymbol{Max\ (x) - Min(x)}} \qquad (10)$$

Categorical variables like Month and Zone were considered in model anticipating having a significance in the value of cooling or heating usage. The variables were assigned with specific values and kept aside during data normalization. The distributions of the data as per assumptions of linear model needs to have a normal distribution, after initial processing data looked fairly close enough. After data processing, the dependent variables (cooling and heating consumptions) were tested for normality. Normality of dependent variables is an important underlying assumption for regression analysis. Figure (10) and (12) shows the distribution of cooling and heating observations respectively over the range of 0 to 1, the y axis shows the frequency of values encountered. Cullen and Frey is a normality test to get a clear picture of distribution in regards with kurtosis and square of skewness, both the dependent variables were found to have minimum skewness acceptable to run regression analysis. The pictorial representation of Cullen and Frey analysis are displayed in figure (11) and figure (13) for cooling and heating variables respectively.

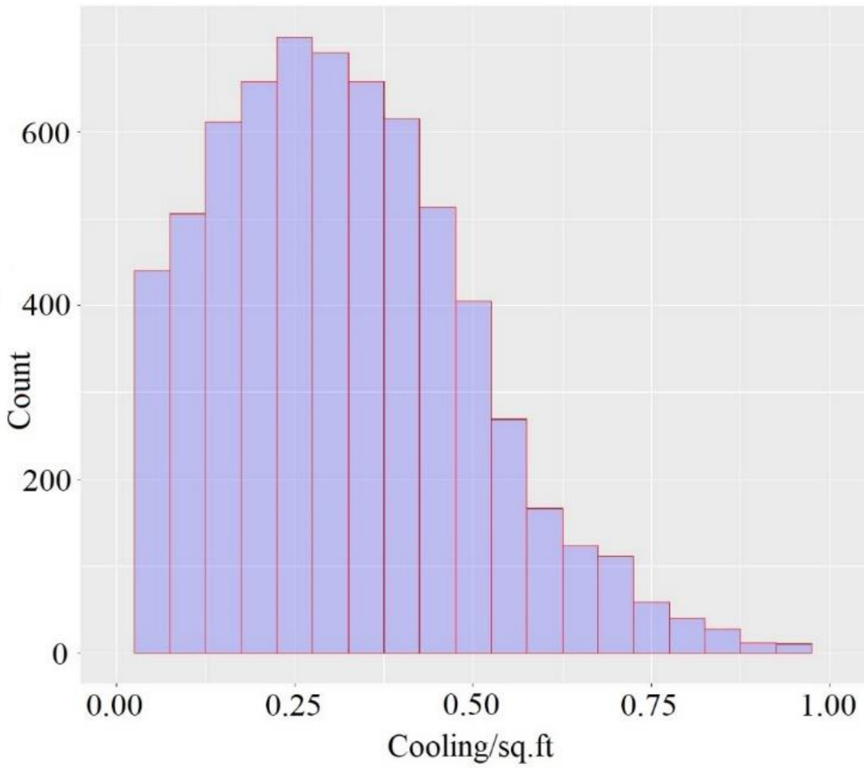


Figure 10: Distribution of Cooling/sqft after data cleaning

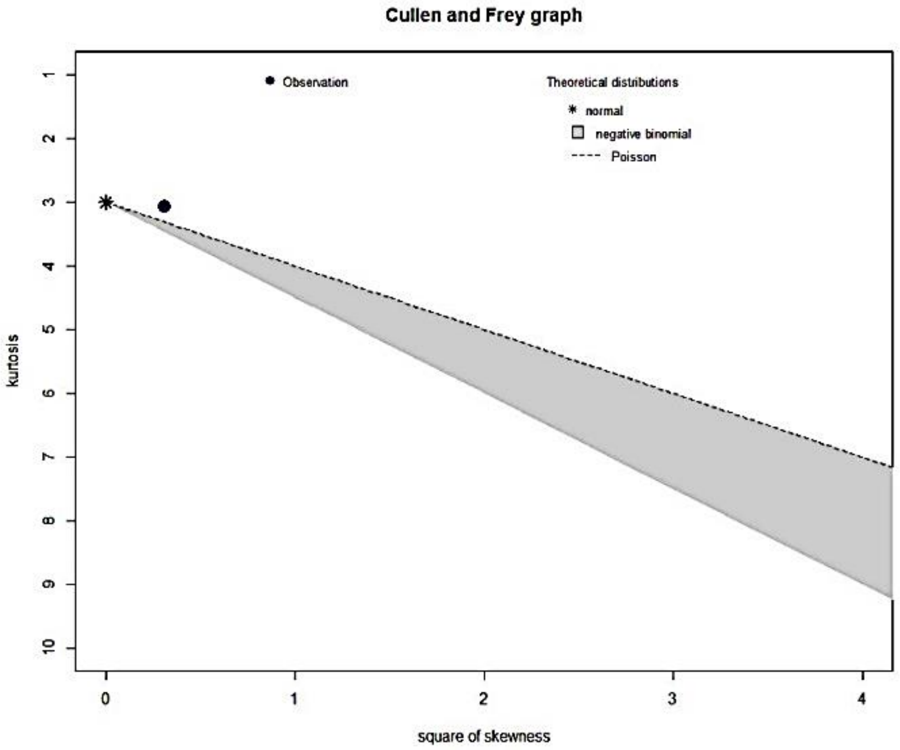


Figure 11: Normality test result of Cooling/sq.ft distribution

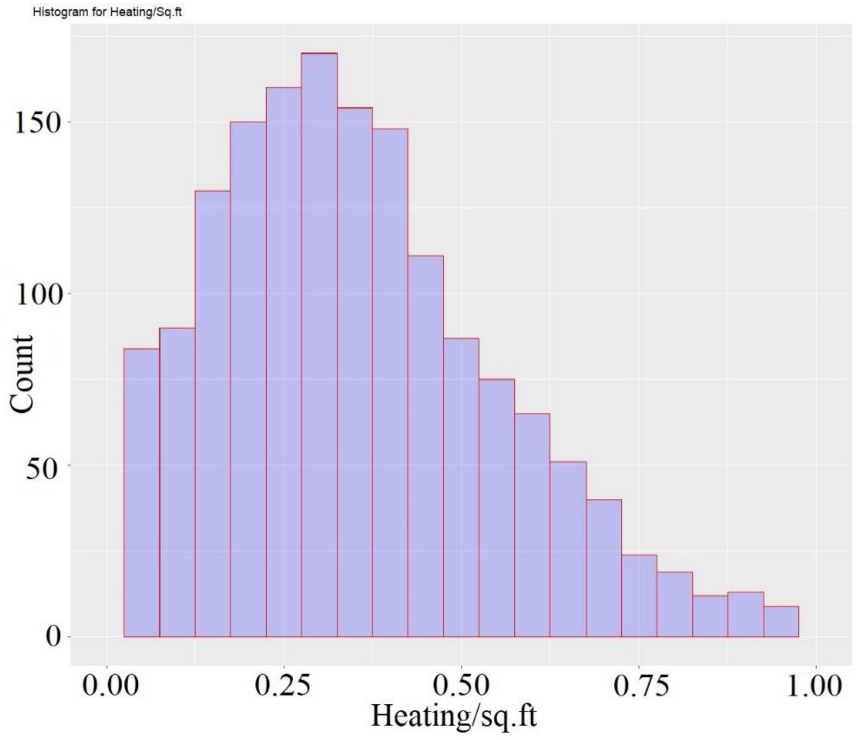


Figure 12: Distribution of Heating/sq.ft after data cleaning

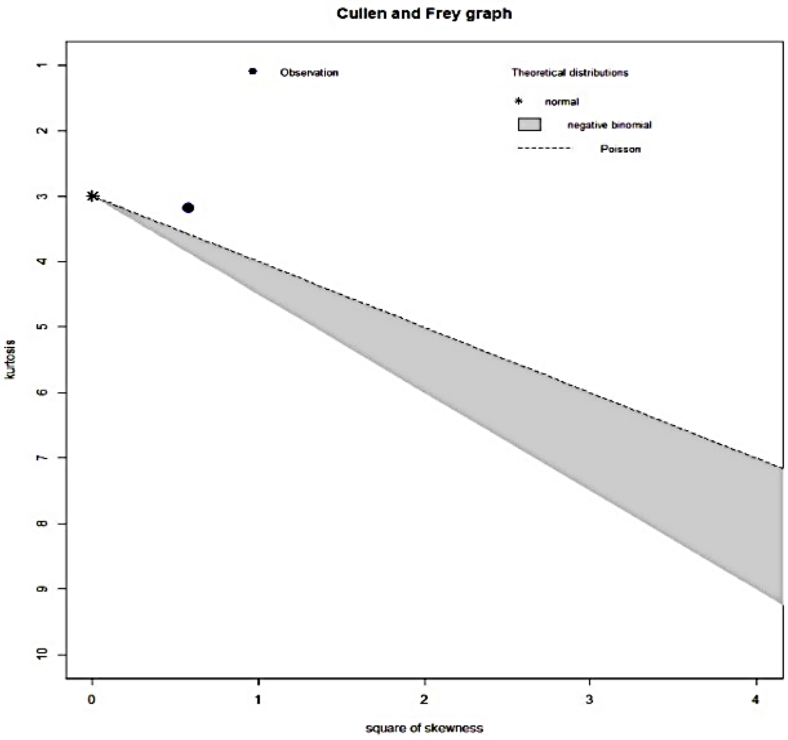


Figure 13: Normality test result of Heating/sq.ft distribution

4.3 Feature Selection

In machine learning and statistics, feature selection is the process of selecting a subset of important features to build a model. Feature selection is performed mainly for four reasons

a. Simplification of model to make them easier to interpret by users

b. Shorter training time

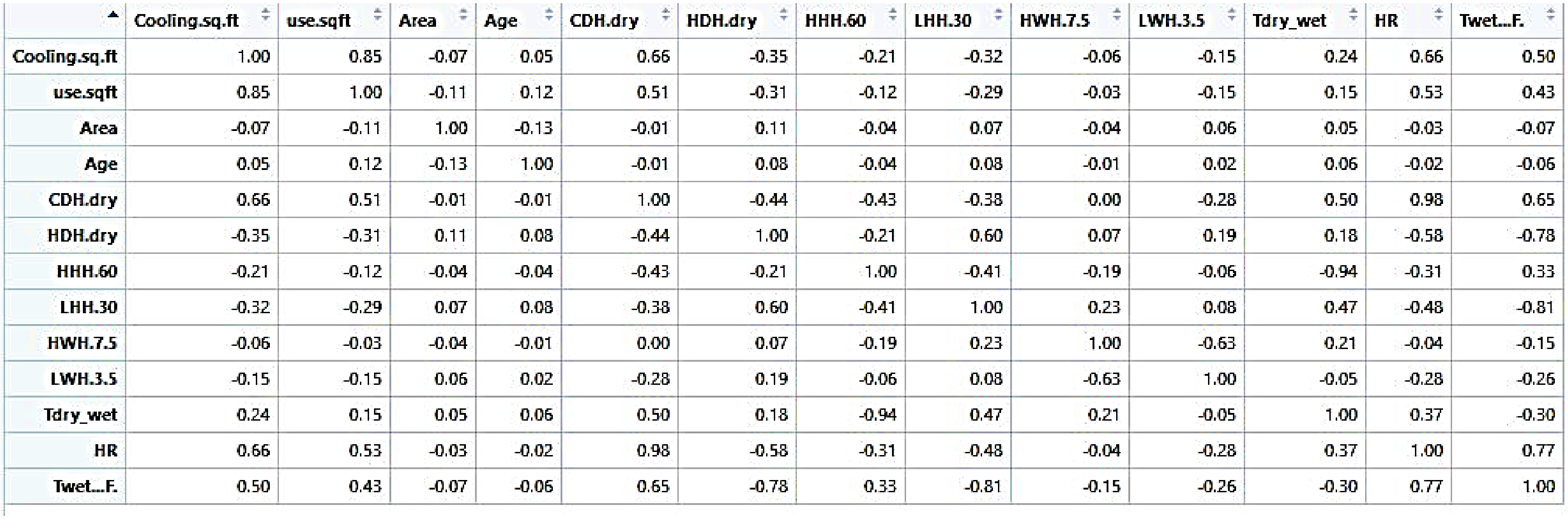
c. Avoid dimensionality

d. Enhance generalization by reducing overfitting; also characterized as variance

Among many ways to select a subset of features, in this study we adopted a filtering based on correlation and AIC. The features having high correlation (>±30%) with dependent variable and relatively lower correlation (collinearity) with other features. Being a basic filtering processing, a minimum 30% has been adopted.

Multicollinearity is a phenomenon where a variable in a multiple regression can be predicted from another with a substantial degree of accuracy. Due to this, coefficients of features are most effected responding to any small change in the model data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multiple polynomial regression model with collinear predictors can indicate how well the entire group of predictors estimates the outcome variable, but it might fail to give a valid result about a individual predictor, or about which predictors are redundant with respect to others.

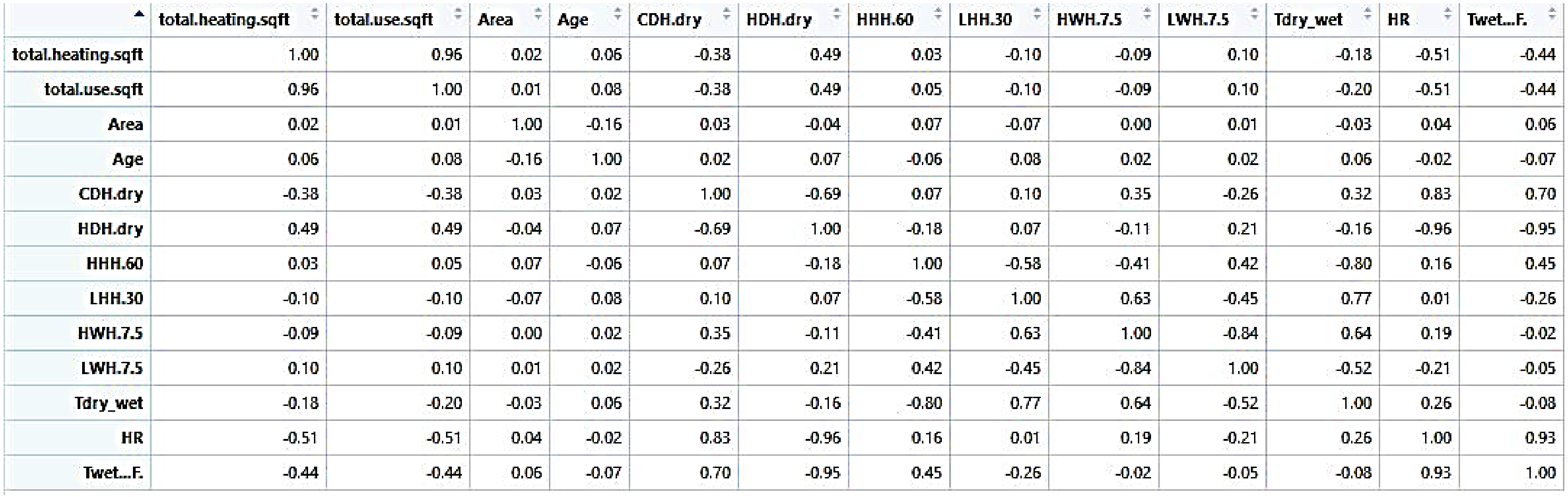
Table 4: Correlation matrix with dependent and predictor variables of Cooling usage dataset



▲	Cooling.sq.ft	use.sqft	Area	Age	CDH.dry	HDH.dry	HHH.60	LHH.30	HWH.7.5	LWH.3.5	Tdry_wet	HR	Twet...F.
Cooling.sq.ft	1.00	0.85	-0.07	0.05	0.66	-0.35	-0.21	-0.32	-0.06	-0.15	0.24	0.66	0.50
use.sqft	0.85	1.00	-0.11	0.12	0.51	-0.31	-0.12	-0.29	-0.03	-0.15	0.15	0.53	0.43
Area	-0.07	-0.11	1.00	-0.13	-0.01	0.11	-0.04	0.07	-0.04	0.06	0.05	-0.03	-0.07
Age	0.05	0.12	-0.13	1.00	-0.01	0.08	-0.04	0.08	-0.01	0.02	0.06	-0.02	-0.06
CDH.dry	0.66	0.51	-0.01	-0.01	1.00	-0.44	-0.43	-0.38	0.00	-0.28	0.50	0.98	0.65
HDH.dry	-0.35	-0.31	0.11	0.08	-0.44	1.00	-0.21	0.60	0.07	0.19	0.18	-0.58	-0.78
HHH.60	-0.21	-0.12	-0.04	-0.04	-0.43	-0.21	1.00	-0.41	-0.19	-0.06	-0.94	-0.31	0.33
LHH.30	-0.32	-0.29	0.07	0.08	-0.38	0.60	-0.41	1.00	0.23	0.08	0.47	-0.48	-0.81
HWH.7.5	-0.06	-0.03	-0.04	-0.01	0.00	0.07	-0.19	0.23	1.00	-0.63	0.21	-0.04	-0.15
LWH.3.5	-0.15	-0.15	0.06	0.02	-0.28	0.19	-0.06	0.08	-0.63	1.00	-0.05	-0.28	-0.26
Tdry_wet	0.24	0.15	0.05	0.06	0.50	0.18	-0.94	0.47	0.21	-0.05	1.00	0.37	-0.30
HR	0.66	0.53	-0.03	-0.02	0.98	-0.58	-0.31	-0.48	-0.04	-0.28	0.37	1.00	0.77
Twet...F.	0.50	0.43	-0.07	-0.06	0.65	-0.78	0.33	-0.81	-0.15	-0.26	-0.30	0.77	1.00

The above chart (Table 4) illustrates the correlation matrix between all the variables of cooling dataset after data processing. Variable selection is performed based on these correlations and preparing second-degree variables. The strong correlation of cooling with total electricity usage and weather variables like CDH, HHH and other features and a linear relation with total electricity usage suggests a polynomial model. Age alone did not have much correlation with cooling usage, but combined with other variables as a polynomial, it showed significantly better correlation than the other variable alone.

Table 5: Correlation matrix with dependent and predictor variables of Heating usage dataset



	total.heating.sqft	total.use.sqft	Area	Age	CDH.dry	HDH.dry	HHH.60	LHH.30	HWH.7.5	LWH.7.5	Tdry_wet	HR	Twet...F.
total.heating.sqft	1.00	0.96	0.02	0.06	-0.38	0.49	0.03	-0.10	-0.09	0.10	-0.18	-0.51	-0.44
total.use.sqft	0.96	1.00	0.01	0.08	-0.38	0.49	0.05	-0.10	-0.09	0.10	-0.20	-0.51	-0.44
Area	0.02	0.01	1.00	-0.16	0.03	-0.04	0.07	-0.07	0.00	0.01	-0.03	0.04	0.06
Age	0.06	0.08	-0.16	1.00	0.02	0.07	-0.06	0.08	0.02	0.02	0.06	-0.02	-0.07
CDH.dry	-0.38	-0.38	0.03	0.02	1.00	-0.69	0.07	0.10	0.35	-0.26	0.32	0.83	0.70
HDH.dry	0.49	0.49	-0.04	0.07	-0.69	1.00	-0.18	0.07	-0.11	0.21	-0.16	-0.96	-0.95
HHH.60	0.03	0.05	0.07	-0.06	0.07	-0.18	1.00	-0.58	-0.41	0.42	-0.80	0.16	0.45
LHH.30	-0.10	-0.10	-0.07	0.08	0.10	0.07	-0.58	1.00	0.63	-0.45	0.77	0.01	-0.26
HWH.7.5	-0.09	-0.09	0.00	0.02	0.35	-0.11	-0.41	0.63	1.00	-0.84	0.64	0.19	-0.02
LWH.7.5	0.10	0.10	0.01	0.02	-0.26	0.21	0.42	-0.45	-0.84	1.00	-0.52	-0.21	-0.05
Tdry_wet	-0.18	-0.20	-0.03	0.06	0.32	-0.16	-0.80	0.77	0.64	-0.52	1.00	0.26	-0.08
HR	-0.51	-0.51	0.04	-0.02	0.83	-0.96	0.16	0.01	0.19	-0.21	0.26	1.00	0.93
Twet...F.	-0.44	-0.44	0.06	-0.07	0.70	-0.95	0.45	-0.26	-0.02	-0.05	-0.08	0.93	1.00

Despite being a three-year data from over 500 houses a large part of data had either improper data entries, or zero values. A large amount of data was well below 0.5 kW-hr/sq.ft of total usage and even more insignificant total heat usage thus over emphasizing a certain value range. After a careful data processing using distribution plots and scatterplots with total electricity usage, a clear data with nearly gaussian distribution has been obtained with around 1829 observations. The above chart (Table 5) illustrates the correlation matrix between all the variables of Heating dataset after data processing. Variable selection is performed based on these correlations and step-wise method, where a combination of variables with low Akaike Information Criteria (AIC) are used to develop the linear predictive model.

Among varies methods based on AIC, we have used is a step-wise variable selection. There are several other robust methods than step-wise AIC based variable selection, but to the data present and its complexity, this method is enough to lead to an accurate result. AIC is based on information theory and is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

4.4 Modeling

Compared to neural networks, regression analysis could be an easier and more practical solution to different problems. Regression analysis is one of the most used statistical tools to describe the variation of a dependent variable y (cooling or heating usage) to explanatory variables (weather, building characteristics, and energy usage) used as inputs in the function. The aim of the regression analysis is to find an appropriate mathematical model and to determine the best fitting coefficients of the model that make sense from building physics point of view. Given the fact that the output variable spans a continuous range of values and that the pattern of inputs influence on the output is known, regression technique is a clear viable solution to develop a predictive model.

When fitting a regression model to a given set of data, we begin with simple linear regression model. Based on the result, later we decide to change it to a quadratic or wish to increase the order from quadratic to a cubic model etc. In each case, we must begin the modeling from scratch. It is preferable to have a situation in which adding an extra term merely refine the model in the sense that by increasing the order and redoing all the calculations from the scratch is not needed. This aspect was of more importance in pre-computer era when all the calculations were done manually. This cannot be achieved by using the powers in succession. But it can be achieved by a system of orthogonal polynomials. The th k orthogonal polynomial has degree k. Such polynomials may be constructed by using Gram-Schmidt orthogonalization. Another issue in fitting the polynomials in one variable is ill conditioning. An assumption in usual multiple linear regression analysis is that all the independent variables are independent. In polynomial regression model, with real world data this assumption is not satisfied. Even if the ill-conditioning is removed by centering, there may exist still high levels of multicollinearity.

Linear regression is a linear approach to modelling the relationship between a dependent variable (scalar response) and one or more independent variables (or explanatory variables). The case of one independent variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multiple polynomial regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. General form of MP regression is,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_I \qquad (11)$$

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are es9timated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all these variables, which is the domain of multiple polynomial analysis.

In this study, a linear model was first developed with decent R^2 of 76%, the results will be discussed further in benchmarking section of chapter 5. To further generate a robust model, polynomial variables were calculated. After data cleaning and variable selection, a polynomial model with order two was developed with the selected combination of features. Datasets of both heating and cooling into 65% of training and 35% of testing subsets. Training set was used in model developing. Cross validation seemed unnecessary for the linear model, so a random sample of 35% was used for testing the model accuracy.

5 MODEL DEVELOPMENT

5.1 Model and Coefficients

Several preliminary models were developed with different combinations of weather variables that makes sense with thermodynamics of the house envelope, like weather degree hours related to wet bulb temperature, and other model with dry bulb temperature along with HR and monthly average wet bulb temperature. The final model with better results in regards with residuals, coefficients and error rate is with the following variables.

Table 6: Cooling model variables and coefficients

Variable	*Symbol*	*Coefficient*
Intercept	β	***-0.080275***
Use	E_t*A	***-0.332802***
Use/sqft^2	E_t^2	***-0.456094***
*Use/sqft*HR*	E_t*x	***1.52714***
*CDH**$T_{db\text{-}wb}$	$h_{cd}*\Delta T$	***-0.089191***
Area	A	***0.591695***
Area^2	A^2	***-0.405048***
*Age*HR*	Y_r*x	***0.168819***
*Area*HHH*	$A*h_{hh}$	***-0.154574***
*Age*Area*	Y_r*A	***-0.320375***

The above table(6) shows the list of variables and their coefficients for the cooling model. This model gave an adjusted R2 of 81.8% and was consistent over other testing parameters compared to other preliminary models. Other models adopted will be discussed in the upcoming benchmarking section. Month is the categorical variable which takes a specific value based on which months' data being entered along with the coefficient.

Table 7: Heating model variables and coefficients

Variable	*Symbol*	*Coefficient*
Intercept	β	***-0.0669***
Use/sqft	E_t	***1.35899***
*Use/sqft*HR*	E_t*x	***-0.33349***
HDH^2	h_{hd}^2	***-0.098847***
Area^2	A^2	***-0.397590***
Area	A	***0.385338***
*Age*Area*	Y_r*A	***-0.08106***
Age	Y_r	***0.070974***

Table (7) is the list of variables with coefficients for final heating model. The Positive coffiecient of Use/sq.ft shows the coefficients are properly evaluated. Strong correlation between total energy usage and combined heat and gas usage accounted for superior R^2 of 92.77%. Tough high initial noises were encountered in heating dataset, most of which was a lump of data concentrated below 0.3 kwh/sq.ft of heating usage. After filtering out highly repeated values, the linear relation between the Y and X values was dominant.

5.2 Model Validation

A detailed analysis of residuals was done due to its importance to validate the model and generalize for similar houses. The residuals from the fitted model represent the differences between the responses observed at each combination values of the independent variables and the respective prediction of the response calculated using the regression function. Histograms and scatter plots used to investigate the distribution of residuals.

In figure (14) & (15), distribution of residuals scattered around zero error were observed for both heating and cooling model and did not exhibit any specific pattern in relationship to the value of dependent or independent variables. This proves that the model's prediction has no significant drift and does not lack any significant phenomenon that could generate such an overall drift in the prediction.

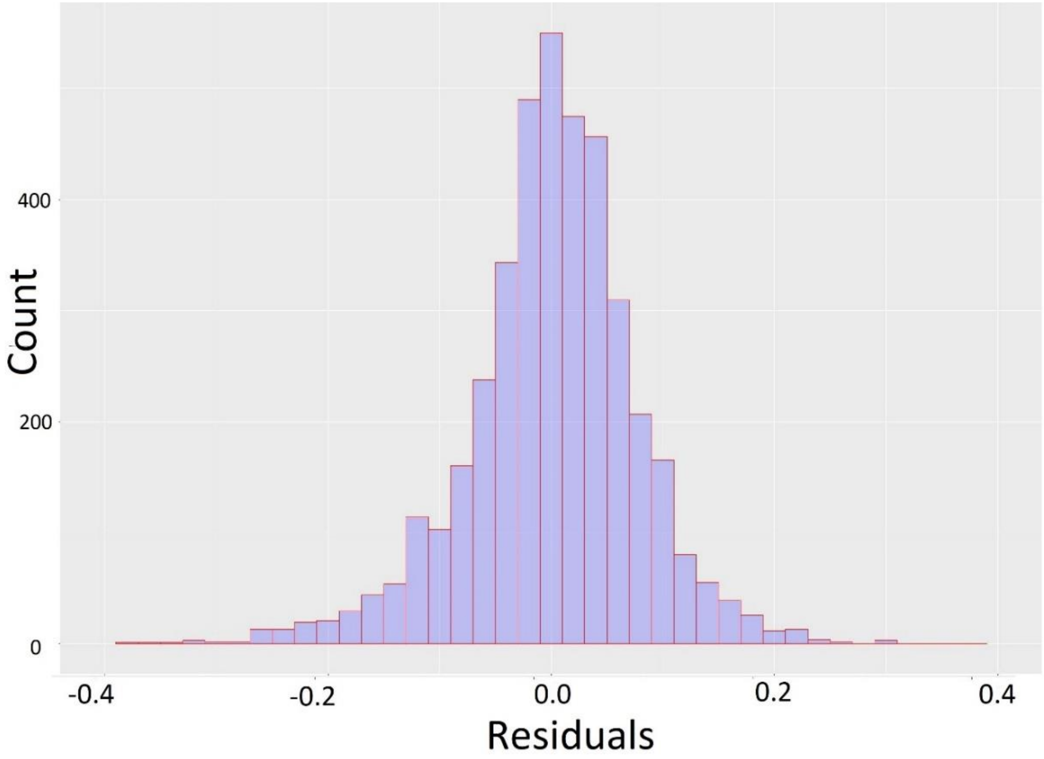


Figure 14: Distribution of residuals for cooling model

Although the histograms are sometime considered too simple for advanced statistical modeling, they give a clear scenario of the errors and suitability of the mathematical model. The proximity of the error value to zero means that the output intended by the model fits with anticipated outputs.

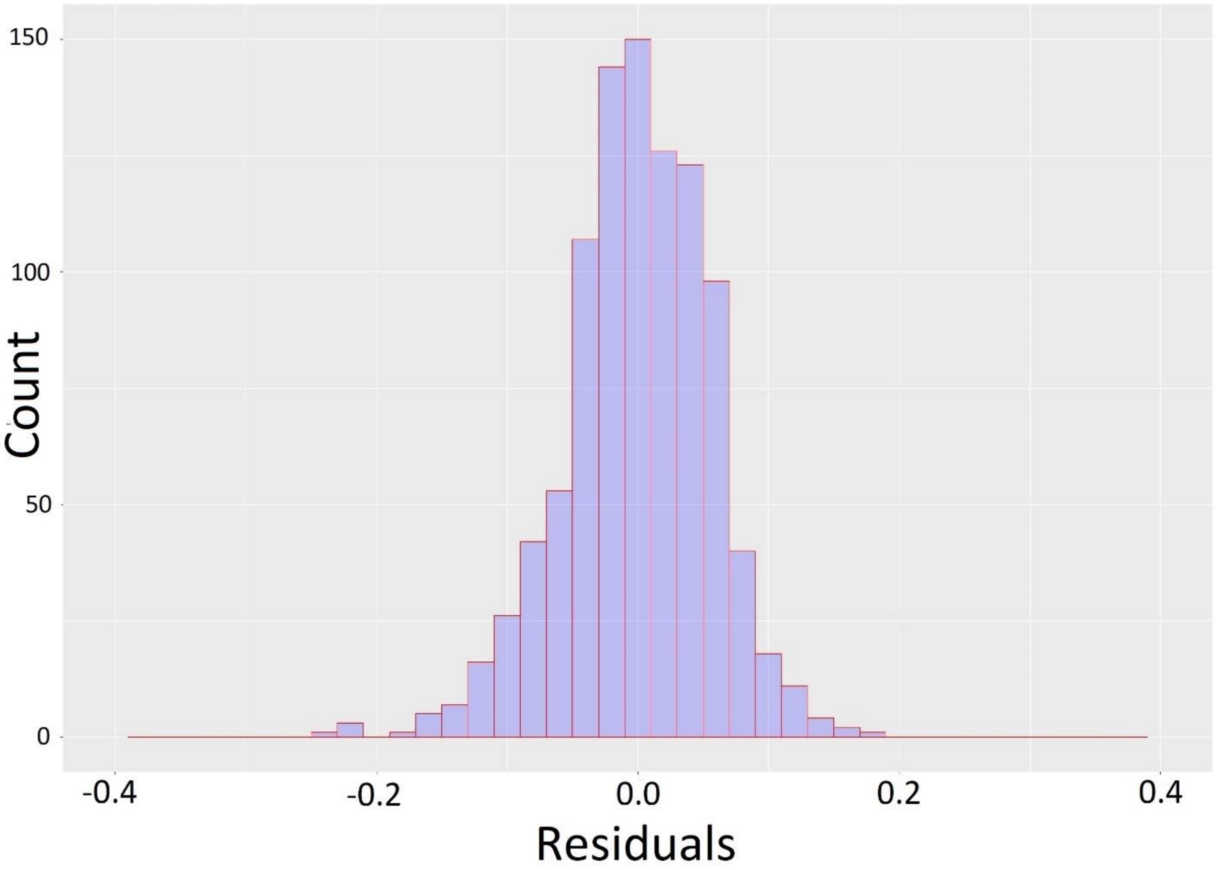


Figure 15: Distribution of residuals for heating model

Except for the residual analysis and the error histograms any measure of error should meet five basic criteria — measurement validity, reliability, ease of interpretation, clarity of presentation, and support of statistical evaluation. In order to respond to these five criteria, we have calculated the mean absolute error (MAE), the mean square error (MSE), the root mean square error (RMSE) and finally the multiple determination coefficient (R^2).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \bar{y}_i|}{n} \quad (12)$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \bar{y_i})^2}{n} \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}{n}} \quad (14)$$

Where y_i is the actual value from the test dataset, $\bar{y}_i$ is the predicted value using the model developed, and n is the number of observations. These test variables are tested against both the final models and the results are elaborated in table (8).

Table 8: Model Error Statistics

	Cooling Model	Heating Model
MAE	0.0567254	0.04261294
MSE	0.00611237	0.00308150
RMSE	0.07818166	0.05551266
R^2	81.8%	92.43%
Minimum Residual	-0.530647	-0.1854097
Maximum Residual	0.3146	0.2307808
Average Residual	-0.00176	0.000734

A plot between the predicted values using the model developed and actual values gives a better picture of the good correlation between these two variables. Compared to other studies related to this field, this regression model is simpler and explains the cooling and heating usage with much easily obtainable variables. Unlike many industrial methods which require additional equipment with high cost, these methods yield a decent accurate prediction at practically no cost once developed. In the figures (16) and (17) are the predicted vs actual values for both cooling and heating datasets.

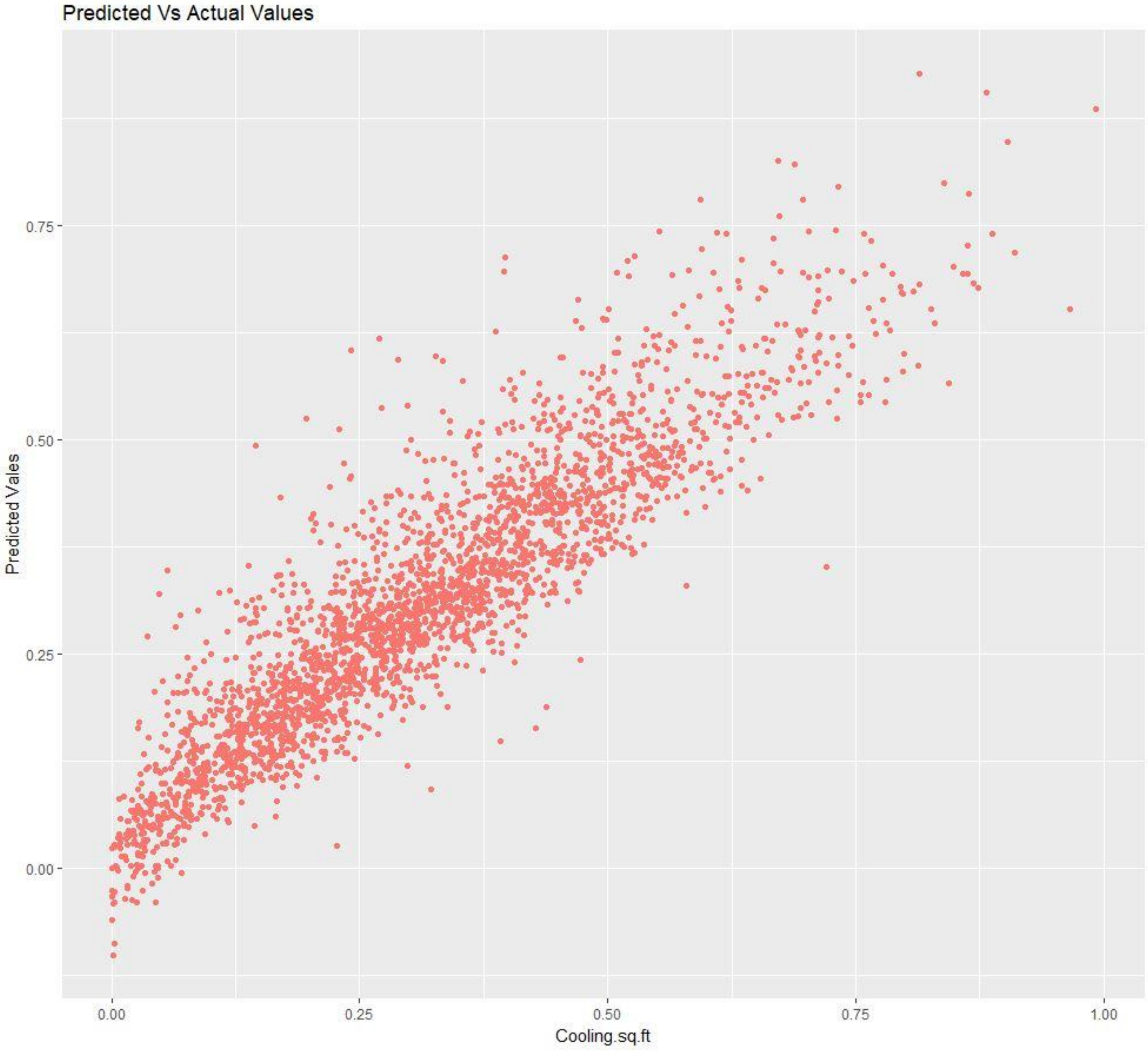


Figure 16: Predicted vs Actual values of Cooling test dataset using the model developed

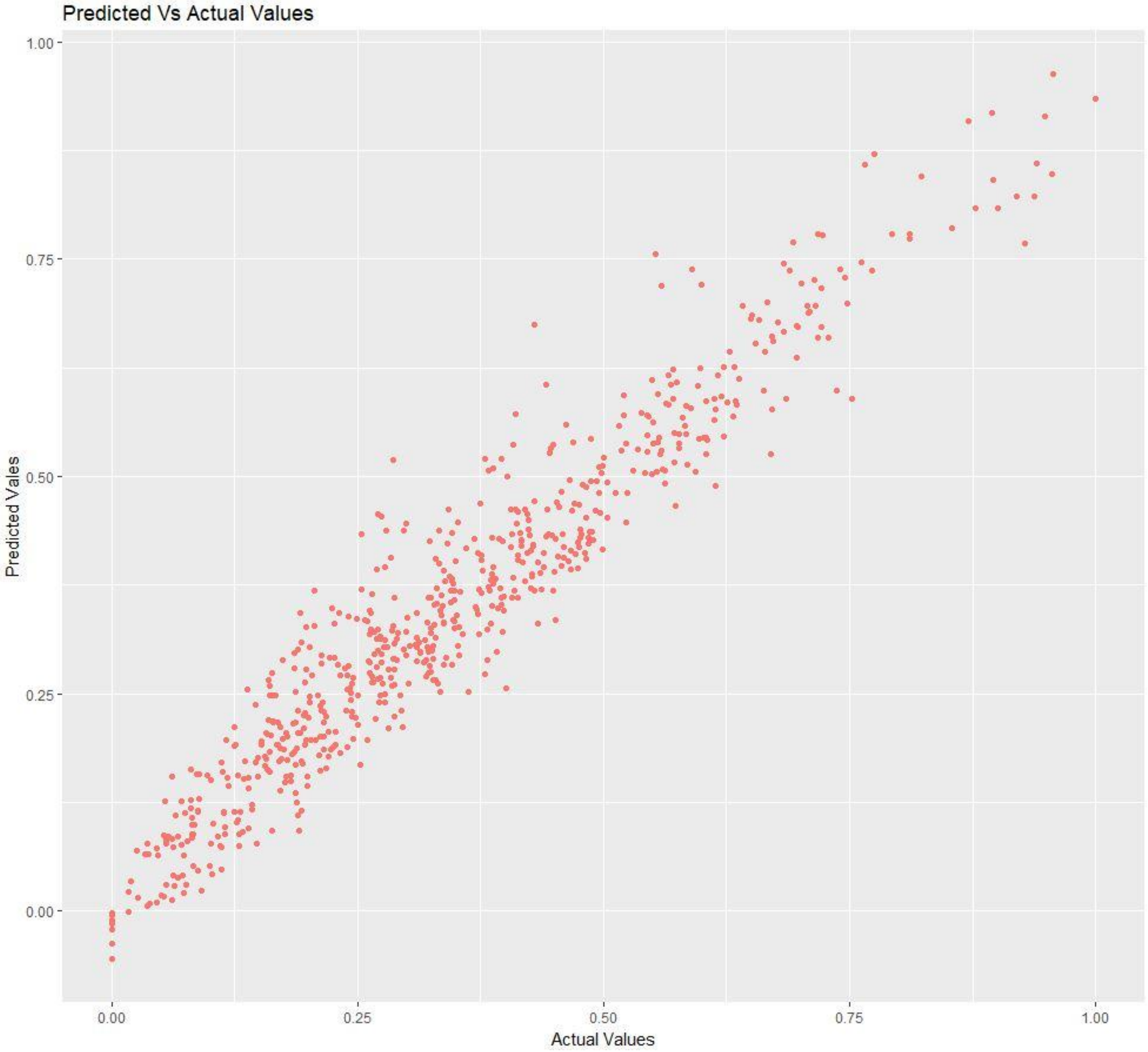


Figure 17: Predicted vs Actual values of Heating test dataset using the model developed

The plot show a clear 45^o angle between actual and predicted values suggesting the prediction is made fairly accurate and no significant imperfection can be observed with either models related actual value.

5.3 Benchmark Models

Several models were tried to compare with the decisive findings. Significant benchmarks were linear model with first degree variables, and Random Forest regressor machine learning techniques. The current method adopted was a polynomial with second-degree model to increase the accuracy over a traditional model with first degree features.

5.3.1 Linear Model

While developing a regression model, we start from the basic linear model, that is with single order variables. If the result is unsatisfactory or if there is a significant improvement by adopting any complex form of regression, we implement. We have followed the same procedure by developing a simple linear regression. The resultant linear model for cooling data had an adjusted R^2 of 76% which is close compared to final model. But the residuals have a range of -0.49 to 0.35, very significant compared to the actual range of dependent variable (0 to 1). The error statistics of cooling simple linear regression are as follows:

MSE = 0.0065

RMSE = 0.081

MAE = 0.0601

Mean of residuals = -0.0027

Maximum residual = 0.35

Minimum Residual = -0.49

Due to high correlation between heating use and total energy use, the linear model showed every other variable insignificant though the R^2 was close to 92.77%. Due to this the significance of other factors over the use of heating is completely foreshadowed. As a result, the team had to consider next complex form of regression, polynomial with second degree.

5.3.2 Random Forest

To compare the result of polynomial regression, a random forest model was developed. We hoped the resultant model to have more coherent and concrete variables using RF feature selection. The outcome was similar to polynomial model, further the feature selection did not take collinearity into account while cutting down the features. As a result, all the features had high correlation with dependent variable and also collinear to each other, resulting in much of the variance between features left unexplained. The research team tried to implement AIC based feature selection for developing RF model, leading to a very similar result as polynomial model.

The cooling RF model had a percentage variance explained of 80.9, and MSE 0.00612, which is very similar to the final model of the study. The optimum trees generated to look for least MSE is calculated to be 100 with a random combination of 3 variables at a time. That is, 4057 observations in the cooling were split into 100 random trees with a combination of 3 random variables after feature selection.

The heating RF had a similar issue as linear model, were the highly correlated total energy usage was repeatedly selected during feature selection. The model had a percentage variance explained of 91.34 with MSE 0.00344, outperforming polynomial model. The optimum number of trees for 1300 observations of training data was calculated to be 80 with 3 random variables together.

5.4 Conclusions

During this study, we statistically explored the open source energy dataset in detail and gained important insights of the underlying influence and correlations between different features and justified with concepts of building physics. We used regression technique to build a working model that works with similar building type and weather conditions and large data with wide range of building and weather properties justifies it to generalize the findings. The benchmarking study with a much simpler linear model and complex model like RF suggests the method adopted is optimum for the data.

The study group has successfully suggested a mathematical model to predict both cooling and heating usage out of utility and weather data, where previously a physical hardware installation was required. We also have adopted a new weather-related feature like cooling/heating degree hours (similarly for humidity and wind speed) where it was never before taking for experimental analysis and concluded with a better result compared to existing features, CDD & HDD. The simple mathematical model with readily available features overcomes the inconvenient of complex models like random forest and neural networks. For the final model of cooling, an R^2 value of 81.8% was achieved with MSE 0.0061 and a 92.77% R^2 is achieved for heating with MSE of 0.003. A simpler linear model needed improvements in range of residuals, furthermore a complex model like RF could not find a better combination of variables to improve the result.

5.4.1 Application Strategy

The regression model developed can be used to accurately predict the HVAC usage and percentage of HVAC in total energy consumption only by using utility bill, weather features and building information. The model can easily be retrained and applied to new data easily. The importance of each variable, procedure to extract them and concepts behind adopting them for this study are explained in-detailed, this approach can be generalized to address similar applications or improvising the current model.

5.5 Future Work

As for the result to given data, the categorical variables like zone was found not significant, the HVAC was only highly correlated with cooling degree hours. The data has building observations from mainly 3 climatic zones in the US, further research can test the model against houses from other climatic zones and re-train the model to generalize for all the residential buildings in the US, if need be. And, a dynamic application can be developed using the models, where users can input the location, basic building information, and utility bill details to find out HVAC usage, thereby educating the users about their usage. With the available data, research group was able to make two separate model covering eight months of a year where the usage of HVAC is predominant. Further studies and development of this concept can attempt to make a single, more robust model to conquer both cooling and heating using a single model.

REFERENCES

1. Efficient comfort conditioning: the heating and cooling of buildings

   Walter Berl - W. Powell - Published by Westview Press for the American Association for the Advancement of Science – 1979

2. Heating & Cooling – https://energy.gov/public-services/homes/heating-cooling

3. Li, Nan, et al. "***Predicting HVAC energy consumption in commercial buildings using multiagent systems***" Proceedings of the 30th International Symposium on Automation and Robotics in Construction and Mining, ISARC. 2013.

4. Sala, Enric, et al. "***Disaggregation of HVAC load profiles for the monitoring of individual equipment***" Emerging Technologies and Factory Automation (ETFA), 2016 IEEE 21st International Conference on. IEEE, 2016

5. Anderson, Kyle D. Non-intrusive load monitoring: ***"Disaggregation of energy by unsupervised power consumption clustering"***. Diss. Carnegie Mellon University, 2014.

6. Zoha, Ahmed, et al. "***Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey***" Sensors 12.12 (2012): 16838-16866.

7. Ansari, F. A., et al. "***A simple approach for building cooling load estimation***" American Journal of Environmental Sciences 1.3 (2005): 209-212.

8. Yan, Chengchu, Shengwei Wang, and Fu Xiao., et al. "***A simplified energy performance assessment method for existing buildings based on energy bill disaggregation***" Energy and buildings 55 (2012): 563-574.

9. Sonderegger, Robert C. "***A baseline model for utility bill analysis using both weather and non-weather-related variables***" ASHRAE transactions 104 (1998): 859.

10. Fischer, Corinna. , et al. "***Feedback on household electricity consumption: a tool for saving energy?.***" Energy efficiency 1.1 (2008): 79-104.

11. Ayres, Ian, Sophie Raseman, and Alice Shih. , et al. "***Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage.***" The Journal of Law, Economics, and Organization 29.5 (2013): 992-1022.

12. Hart, George William., et al. "***Nonintrusive appliance load monitoring.***" Proceedings of the IEEE 80.12 (1992): 1870-1891.

13. Papakostas, Konstantinos, and Nicolaos Kyriakis. "***Heating and cooling degree-hours for Athens and Thessaloniki, Greece***." Renewable Energy 30.12 (2005): 1873-1880.

14. Han, Chia Y., Yunfeng Xiao, and Carl J. Ruther. ***"Fault detection and diagnosis of HVAC systems."*** Ashrae Transactions 105 (1999): 568.

15. Seung Uk, Lee, et al. "***Whole-Building Commercial HVAC System Simulation for Use in Energy Consumption Fault Detection.***" ASHRAE Transactions, vol. 113, no. 2, Oct. 2007, pp. 52-61. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=aph&AN=28452216&site=eds-live.

16. Shaw, S. R., et al. "***Detection and diagnosis of HVAC faults via electrical load monitoring."*** HVAC&R Research 8.1 (2002): 13-40.

17. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis https://www.eia.gov/consumption/commercial/maps.php

18. Santin, O. G., Itard, L., & Visscher, H. (2009). "***The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. Energy and buildings***", 41(11), 1223-1232.

19. Yun, G. Y., & Steemers, K. (2011). "***Behavioural, physical and socio-economic factors in household cooling energy consumption***". Applied Energy, 88(6), 2191-2200.

20. Kelly, Jack, and William Knottenbelt. "***Does disaggregated electricity feedback reduce domestic electricity consumption? A systematic review of the literature.***" arXiv preprint arXiv:1605.00962 (2016)..

21. Street, Pecan. ***"The pecan street project."*** Austin, TX: Working Group Report (2010).

22. Wilhite, Harold, and Rich Ling. "***Measured energy savings from a more informative energy bill***." Energy and buildings 22.2 (1995): 145-155.

23. Perez, Krystian X., et al. "***Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data***" Energy and Buildings 81 (2014): 316-325.

24. Kolter, J. Zico, Siddharth Batra, and Andrew Y. Ng. ***"Energy disaggregation via discriminative sparse coding."*** Advances in Neural Information Processing Systems. 2010.

25. Wytock, Matt, and J. Zico Kolter. ***"Contextually Supervised Source Separation with Application to Energy Disaggregation."*** AAAI. 2014.

26. Mavrokefalidis, Christos, et al. ***"Supervised energy disaggregation using dictionary—based modelling of appliance states."*** PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2016 IEEE. IEEE, 2016.

27. Kim, Hyungsul, et al. ***"Unsupervised disaggregation of low frequency power measurements."*** Proceedings of the 2011 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2011.

28. Min, Jihoon, Zeke Hausfather, and Qi Feng Lin. ***"A High-Resolution Statistical Model of Residential Energy End Use Characteristics for the United States."*** Journal of Industrial Ecology 14.5 (2010): 791-807.

29. Amber, K. P., M. W. Aslam, and S. K. Hussain. ***"Electricity consumption forecasting models for administration buildings of the UK higher education sector."*** Energy and Buildings 90 (2015): 127-136.Braun, M. R., H. Altan, and S. B. M. Beck. ***"Using regression analysis to predict the future energy consumption of a supermarket in the UK."*** Applied Energy 130 (2014): 305-313.

30. Lifton, Joshua, et al. ***"A platform for ubiquitous sensor deployment in occupational and domestic environments."*** Proceedings of the 6th international conference on Information processing in sensor networks. ACM, 2007.

31. Gupta, Sidhant, Matthew S. Reynolds, and Shwetak N. Patel. ***"ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home."*** Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010.

32. Berges, Mario E., et al. ***"Enhancing electricity audits in residential buildings with nonintrusive load monitoring."*** Journal of industrial ecology 14.5 (2010): 844-858.

33. Kolter, J. Zico, and Tommi Jaakkola. ***"Approximate inference in additive factorial hmms with application to energy disaggregation."*** Artificial Intelligence and Statistics. 2012.

34. Laughman, Christopher, et al. ***"Power signature analysis."*** IEEE power and energy magazine 99.2 (2003): 56-63.

35. Tapia, Emmanuel Munguia, et al. ***"The design of a portable kit of wireless sensors for naturalistic data collection."*** International Conference on Pervasive Computing. Springer, Berlin, Heidelberg, 2006.

36. Lam, Joseph C., et al. ***"Multiple regression models for energy use in air-conditioned office buildings in different climates."*** Energy Conversion and Management 51.12 (2010): 2692-2697.

37. Westergren, Karl-Erik, Hans Högberg, and Urban Norlén. ***"Monitoring energy consumption in single-family houses."*** Energy and buildings 29.3 (1999): 247-257.

38. K. Papakostas-N. Kyriakis - Renewable Energy – 2005 ***“Heating and Cooling Degree-hours For Athens and Thessaloniki, Greece”*** Renewable Energy 30 (2005) 1873–1880

39. Kriengkrai Assawamartbunlue - ***“An Investigation Of Cooling and Heating Degree-hours in Thailand”*** - Journal Of Clean Energy Technologies - 2013