12-2017

Demand Side Management in Smart Grid using Big Data Analytics

Sidhant Chatterjee
Utah State University

DEMAND SIDE MANAGEMENT IN SMART GRID USING BIG DATA ANALYTICS

by

Sidhant Chatterjee

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Engineering

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

2017

ABSTRACT

Demand Side Management in Smart Grid using Big Data Analytics

by

Sidhant Chatterjee, Master of Science

Utah State University, 2017

Major Professor: Rose Qingyang Hu, PhD
Department: Electrical and Computer Engineering

The growing demands and rising costs of electricity require a more advanced, secure and reliable grid. With the integration of information technology systems in the current electrical grids, a data feedback system can be established to reduce the grid vulnerabilities. Advanced form of grid management includes Demand Side Management (DSM) which deals with the monitoring and manipulating of peak demands and flattening of the load profile over the day. This project aims at developing a predictive model that can forecast the short-term to medium-term loads of electric utilities. Load forecasting aids in DSM practices by furnishing the time-varying load data and marking the peaks and troughs in load throughout the day.

(70 pages)

PUBLIC ABSTRACT

Demand Side Management in Smart Grid using Big Data Analytics

Sidhant Chatterjee

Smart Grids are the next generation electrical grid system that utilizes smart metering devices and sensors to manage the grid operations. Grid management includes the prediction of load and and classification of the load patterns and consumer usage behaviors. These predictions can be performed using machine learning methods which are often supervised. Supervised machine learning signifies that the algorithm trains the model to efficiently predict decisions based on the previously available data.

Smart grids are employed with numerous smart meters that send user statistics to a central server. The data can be accumulated and processed using data mining and machine learning techniques to extract meaningful insights. Forecasting of future grid load (electricity usage) is an important task for gaining intelligence in the gird. Accurate forecasting will enable a utility provider to plan the resources and also to take controlled actions to balance the supply and the demand of electricity. This forecasting can be achieved using machine learning based predictive models.

In this project, a predictive system is designed that uses data mining and machine learning techniques to process the smart meter data and to use it as training data for the model. The main objective of this project is to forecast short term to mid-term load for the grid entity. The outcomes are backed with visualizations to make the data and results more user readable.

To the supreme power that always guides me out of darkness, and to my family for their unending love and support in all spheres of life.

ACKNOWLEDGMENTS

Every big tree stands on a network of dense roots, that supports and nourishes it. Completing this project also required the support, appraisal and criticism of my advisor, committee members, my family and my peers.

I would like to express my deepest gratitude to my advisor, Dr. Rose Hu, for her untiring guidance and patience, and providing me with an excellent atmosphere to learn and grow as a student.

I would also like to thank Dr. Don Cripps and Dr. Tung Nguyen for accommodating with my schedules and helping me complete all the tasks well in time. A special note of thanks to Dr. Kyumin Lee, who ignited my interest in the field of data science and data mining and helped me develop my background.

Special thanks goes to all my committee members, who were willing to participate in my final defense committee at the last moment.

Finally, I would like to thank my parents and my brother for supporting me all throughout my life, for always motivating and pushing me towards all the big endeavors in my life. Thanks for being by my side and always supporting me regardless of the circumstances.

Sidhant Chatterjee

CONTENTS

LIST OF FIGURES

ACRONYMS

SG	Smart Grids
DSM	Demand Side Management
DR	Demand Response
LC	Load Classification
DEM	Dynamic Energy Management
LF	Load Forecasting
SCADA	Supervisory Control and Data Acquisition
DoE	Department of Energy
PMU	Phasor Management Unit
DHW	Domestic Hot Water
HDD	Heating Degree Days
CDD	Cooling Degree Days
MLR	Multiple Linear Regression
GBR	Gradient Boosting Regression
RTP	Real Time Pricing
CPP	Critical Peak Pricing
LC	Load Classification
ANN	Artificial Neural Networks
FNN	Fuzzy Neural; Networks

CHAPTER 1

INTRODUCTION

1.1 Background

Electricity is one of the most essential components of the modern human life. It is one of the driving forces of the modern life and world. However, electricity is something most of us may take for granted. On one hand, there are almost 1.3 billion people still not having access to electricity [2] and on the other hand, the demand for electricity is expected to increase significantly over the coming years [3]. Fig.1.2 shows the steady rise in the number of customers from the year 2005 to 2015 all over the world.

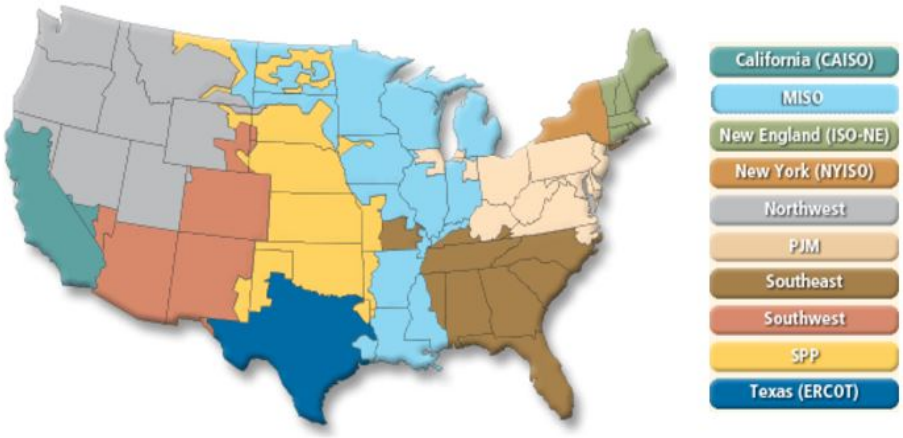

Fig. 1.1: Major electrical markets in the US

Since electricity plays a vital role in the human being society, conservation and appropriate management strategies for the grids is a must. In order to cope with the increasing energy prices and shortages, the need for a smarter approach becomes more and more important. This need of smarter energy management systems has led to the development of technologies like smart grids and demand side management. With the advent of smart

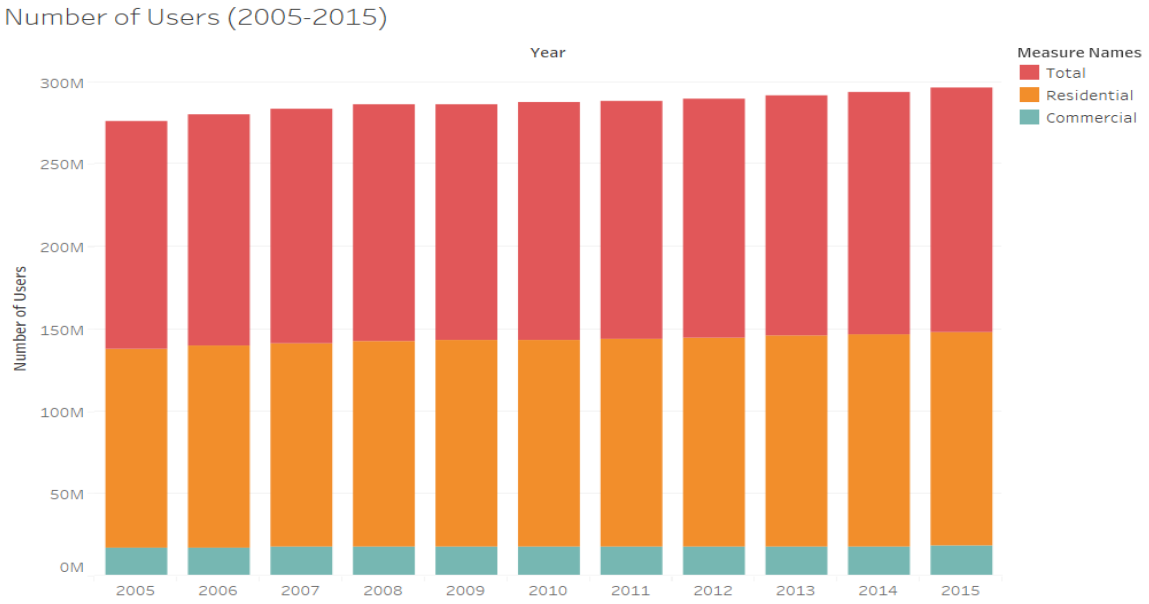

Fig. 1.2: Number of electric grid users from 2005 to 2015

grids, DSM schemes can be implemented both on the utility side and on the customer side. Load forecasting is one of the major tools used in demand side management. For an electric utility, load forecast help planners in making strategic decisions on unit commitment, hydrothermal co-ordination and generation quantities, security assessments and dynamic pricing. [4]

The conventional electrical grids appeared in US around 90 years ago, and have been divided into 10 markets California(CAISO), MISO, New England(ISO-NE), New York (NYISO), Northwest, PJM, Southeast, Southwest, SPP and Texas (ERCOT). Fig 1.1 shows the major US electric markets. These markets are interconnected with a complex network of utility companies that serve as mediators between the generating station and the end-use customers. Since the electric grid has functioned essentially the same way since its inception, it has become vital to integrate information technology to the existing electrical grids to make them more reliable, flexible, and secure. This smart grid would assist consumers to track and manage their energy usage and plan their load and consumption.

There are four major drivers for the development and implementation of smart grids within the existing grid networks. These major drivers, according to a report by VINNOVA

(2011), are growth, sustainability, market, and vulnerability. Growth refers to the exponential growth in the demand of electric power. With the new technological advances and increase in the coverage of electrical network, the net demand keeps increasing exponentially. Market refers to the competition and the market rules in relation to energy management and distribution to the end-use customers. Vulnerability is an important factor in the market, as it deals with the grid vulnerability and the probability of outages, or events like grid failure, overloading or irregularities in supplied voltage.

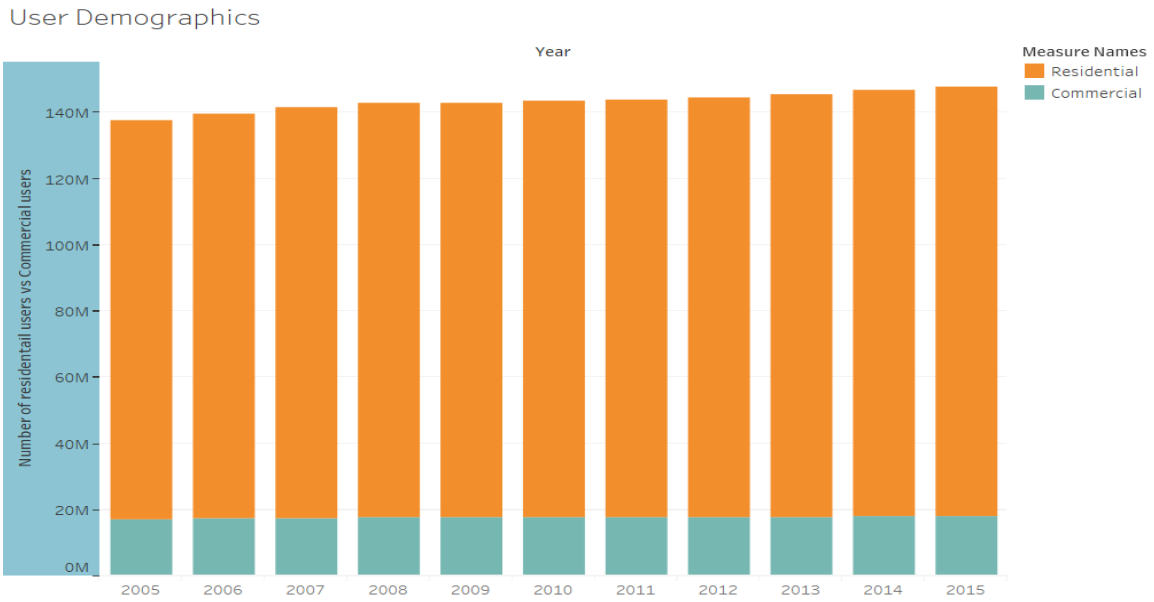

Fig. 1.3: Corporate Vs. Residential users

1.2 Smart Grids

Smart grids are the next generation electrical grids that employs information, communication and control techniques to improve the efficiency and reliability of electrical grids. According to [5], with the exponential growth of smart metering systems in grids, high volumes of data and information are being available form the grid structure [6, 7]. Since the huge amount of data generated from the smart metering systems cannot be processed locally and using normal available methods, big data technologies are employed, which are

designed to extract logical outcomes from very large data volumes. There are several big data sources in smart grids, with the major sources being power consumption data measured in kWh, energy pricing data collected by the automated revenue metering system (ARM), and operational data for grid operation and control.

It is interesting to note that smart grids can hold all types of storage and generation related data, thus increasing the scope for user participation, asset optimization and enabling services that are essential for markets. It is expected that by year 2020, the number of smart meters in the grids would reach around 240 million in Europe, 150 million in North America, and about 400 million in China [8]. Such a large number of smart metering systems will generate petabytes of data in one day. Thus, it is out of the scope of usual storing techniques and tools to derive meaningful insights from the generated data. The variability, variety, and velocity of the data makes it large enough to be called as Big Data.

1.3 Dynamic Energy Management

Dynamic Energy Management (DEM) is an innovative approach to managing load at the demand-side of the grid. The main parts of DEM are Demand Side Management (DSM) and Demand Response (DR). DEM attains long term energy savings through DSM by reducing the peak load occurrences in a utility [9].

1.3.1 Demand Side Management

Authors K.E. Parameter et al. have defined DSM as planning, implementation, and monitoring of the utility activities that influence the customer's use of electricity in ways that changes the utility load shape [9], i.e., changes in the time pattern and magnitude of a utilitys load. One type of demand side management is demand response, which focuses on price signals to handle peak demand. Peak demand can be handled by either using a price based system, where the electricity price fluctuates according to the load, or by using an incentive based system, where incentives are given to customers to reduce load at peak times [10].

Demand response is a key factor of DSM. The Federal Energy Regulatory Commission

defines demand response (DR) as "Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized". In other words, demand response (DR) is the technique to manipulate customer's load during peak demand to the other time, when the demand is less. This helps in reducing the peak demand of the grid, and also in reduction of prices on the customer side. DR can be applied to both residential and industrial loads and includes three different concepts: energy consumption reduction, shifting consumption to periods of low (or high) demand, and efficient utilization of energy storage systems [11] Thus, a crucial issue in Smart Grids (SGs) is to manage DR in order to reduce peak electricity load, utilizing the existing infrastructure more efficiently and in a better planned manner [12].

1.3.2 Load Forecasting and Dynamic Pricing Problem

The functionality of electrical grids depends on the load served for a particular service area. Load is an important aspect of the DR systems and the efficiency of utility companies depends on how effectively the electric load can be managed. Customer load depends on several factors such as temperature, heating/cooling of the property, residency schedule, hot water utilization, and lighting conditions. Thus, load forecasting is essential to perform DSM on the grid. There are three types of load forecasting (LF) - Short term LF used to predict load on an hourly basis, medium term LF used to predict load on a weekly basis and, long term LF to predict load up 50 year ahead [13]. Advanced data mining techniques such as multiple regression, exponential smoothing, time series analysis, and Kalman filtering are used to forecast loads [13]. In this project, a short term to medium term LF is achieved using the electrical data of years 2010 to 2015 and employing the technique of regression based load forecasting proposed in [14–16].

Dynamic pricing is an important part of the problem and, is detrimental in avoiding peak load and high demand situations. In order to curb the problem of peak loads in the grid, Real Time Pricing (RTP) and Critical Peak Pricing (CPP) are employed, which

increase the energy prices during peak demands. According to [13], user load can be classified into either interruptible/movable loads or interruptible/immovable loads. Due to the effect of CPP, users will be forced to avoid the interruptible/movable loads and postpone them, thus bypassing the peak load conditions. In absence of load forecasting , complex Load Classification (LC) algorithms are required to classify the loads and apply RTP to balance loads. LC can be a challenging task, and is often done using complex Artificial Neural Networks (ANN) [16, 17], which create user patterns of consumption and facilitate the classification of loads.

1.4 Proposed Solution

Accurate short-term load forecasting, defined in the hours-to-days time frame, can lead to an efficient and economic system operation. In this project, a short-term to mid-term load forecasting model is developed that can forecast load for hours to days time intervals. A load profile is generated for different utility service areas and a predictive model is formulated. Load profiles are data points that reflect the variations of electrical loads with respect to time, and are generated by cumulative electrical energy consumption of any building. Typically, the load profile varies according to the customer type, and are different for domestic, industrial and agricultural sectors. Load consumption also depends on the temperature, non-working days or holidays and the time of day, which are discussed more in the section **??**. For an accurate load forecasting, an hourly load data is used in this project. As given by Nurettin etinkaya [18], an hourly load profile can be used to generate load forecasts for a weekly, a daily, and an hourly basis.

The project can be divided into four stages, each of which leads to the next stage and helps in the formulation of the final prediction model.

- *Data Collection and Clustering:* In this project, electric profile and load data are used for all the medium and large-scale utilities in the state of Utah. This data is then combined with the regional temperature and set-point data for the different utility service areas to form clusters. Clustering is done for all the residential users with

similar climatic conditions and share the same service area, but fall under different utility areas.

- *Generate load profile:* Load profile is the time-dependent electric consumption data for a specific utility, region or a building. An accurate load profile is needed to train the predictive model in order to perform predictions. Since the features of the prediction model varies with the service area of the utility, a similar but unique load profile is generated for all the individual load clusters.

- *Formulate a predictive model:* For obtaining a short-term - mid-term prediction, a predictive model is formulated, (as given in chapter 3) and is trained with the generated load profile. A multivariate linear regression (MLR) model and a gradient boosting (regressors-GBR) predictive model is formulated in this project. Since separate models are formulated and trained for each cluster, the predictive coefficients are different for each cluster. A model that can predict accurately for a particular cluster, may not predict accurately for a different cluster, given to the fact that the correlation of the features can be different for different clusters.

- *Test the model, visualize results:* Cross Validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. So, the formulated models are evaluated using prediction scores, error estimates and a cross validation technique that randomly samples the load profile to evaluate if overfitting exists. The training load profile data has been extracted for the year 2015, and the model accuracy is evaluated using the load data for 2016 as the test set.

CHAPTER 2

LITERATURE REVIEW

With the growing demand for electric power, the need and demand of proper power management techniques are on a rise. A promising solution to attain grid management and reduce vulnerability in electric distribution systems is Smart Grids (SG). According to the Federal Energy Regulatory Committee, a smart grid is an electrical grid that integrates a variety of operational and energy management measures which includes smart meters, smart appliances, renewable energy resources and, energy efficient resources [19]. To emphasize the modern developments in Load Forecasting (LF), only papers published after 1980 are taken into this review. In the past 40 years, the developments in LF have been multidimensional and researched upon by many researchers. This review also counts works [20–23] that are not based upon any experiments or data, but are surveys and reviews of existing experiments, concepts and literature, in addition to those [24–27] which implement, analyze and evaluate different techniques based on real data.

2.1 Overview of Demand Side Management

Proposed and defined by Gellings in [28], Demand Side Management (DSM)) is the planning and implementation of those electric utility activities designed to influence the customer usage of electricity in ways that will produce desired changes in the utilitys load shape. It is generally most convenient for utilities [28] to look at DSM in terms of broad load shaping objectives. The load shape is the daily and seasonal electricity demand by time of-day, day-of-week, and season. A study based on the significance of DSM application to smart grid based big data sources is done by Keyan Liu et al. in [29]. Authors in [30–35] propose a game theory based approach to achieve demand side management in smart grids. According to [36], game theory is a mathematical tool that analyzes potentially arising conflict of interest among independent and rational agents and seek to maximize their own

benefit when they strategically interact with each other. In other words, the use of game theory can optimize the load profile by manipulating the peak and valley loads uniformly over the day. A significant amount of research is being done in formulating a decentralized approach to achieve DSM in micro-grids. Authors in [37] propose an IoT based approach to perform DSM in micro grids.

2.2 Load Forecasting Techniques

I. Moghram et. al. in his paper [38] evaluated five load forecasting techniques - multiple linear regression, stochastic time series, general exponential smoothing, state space method, and AI or artificial intelligence based approach. The authors implemented the techniques to generate an hourly, load forecast for the next day using the data from a southeastern utility in US. The authors briefly described the implementation of each technique and compared and analyzed the results. Authors K. Liu et.al. in [39] proposed and evaluated three techniques for load forecasting, which were ANN, FL and a time-series auto-regressive (AR) model. Although the conclusion that AR based model is less efficient than the other two models wasn't clearly explained, the load series data was also considered as a stationary data, which contradicts the fact that load profiles are dynamic sources.

Statistical modeling technique were discussed in more recent papers for load forecasting. For example, authors in [40] propose a regression based load forecasting approach using the PG&E dataset. Other regression based approaches proposed by authors in [41–43] deals with the use of weighted least square technique, temperature modeling (implementing various heating and cooling functions), weekday and weekend modeling etc. As a modification to the basic regression technique, Haida et. al [41] proposed a transformation technique to model the nonlinear relationship between load and weather variables. The transformation technique was used with the Tokyo Power Utility Corporation dataset to forecast short term load. A regression based peak forecasting model was proposed by authors in [44]. A unique approach of forecasting a daily cumulative energy consumption forecast before an hourly load forecast was given by Ruzic et. al. in [45], wherein a two-step multiple linear regression (MLR) model was used for prediction. The first step of the MLR was

used to predict the cumulative energy consumption of the day, and the next step predicted the hourly load profile. Works in [46, 47] propose a probability density based estimation of load forecasting. The load forecast was the conditional expectation of the load given the explanatory variables including time, weather conditions, etc. [48].

Time series analysis of load data is another way to forecast load profile. Autoregressive (AR) and Autoregressive Moving Average (ARMA) techniques has been used for load forecasting in recent years [49, 50] A combination of auto-regressive moving average model and regression techniques has been presented in [51, 52]. The regression part is used to predict the peak and valley loads and ARIMA models are applied to the data to make the hourly load forecast [48]. In [52], a 3^{rd} order polynomial for the temperature attribute was proposed to reflect the nonlinear relationship between the load and temperature [48]. A supervised time-series model, that takes a pre-defined manual input as the primary forecast and then formulates a regression model using the available data has been proposed by authors in [53].

Artificial Neural Networks (ANN) are also highly used in performing load forecasting [54–58]. Although models based on statistical methods generally perform well, but, in case of an abrupt change in the model attributes or the presence of statistical glitch in data, deficiencies arise and the prediction accuracy dips. This greatly affects the load patterns and load profile [59]. AI based techniques like ANN and fuzzy logic can cope with this kind of problem, and perform predictions without any loss in accuracy. Authors in [60, 61] propose an ANN based load prediction model using the back propagation model and the radial basis function model respectively, with a performance review of both the models. Authors in [57] propose the use of ANN to perform real-time load forecasting. Real time data from a local utility is used to forecast the load on an hourly basis. To further improve the prediction efficiency, hybrid schemes employing Support Vector Machines (SVM) and ANN has been proposed in [58]. The proposed model consists of two module, the first one is used to predict the peak load and the second module is used to predict the hourly load.

In addition to ANN, Fuzzy Networks or Fuzzy Neural Networks (FNN) also forms one

of the most used load forecasting techniques. Fuzzy logic is an approach to make partial decisions, not a complete 0 or a complete 1, but more of a fraction between 0 and 1. The idea resonates with the idea of likelihood of an event to be true or false, rather than completely true or false. Authors in [50] propose a long-term load forecasting technique using fuzzy logic approach. According to the authors, fuzzy logic outperforms ANN in long term forecasting, due to increased gap between the weather conditions and load profile. Authors in [62] proposed a unique fuzzy network for load prediction for each individual day of a week, which leads to a load forecast model that forecasts the peak and the valley load and calculates the hourly load profile using the available load data.

Despite the advancement in load forecasting techniques, none of the techniques guarantees a 100% prediction accuracy. Also, there is no single benchmark technique that can be used in any case of load forecasting. The desirable model varies with the data availability and the forecast objectives. Although advanced models like ANN and fuzzy logic offer a high degree of accuracy, they also increase the prediction complexity of the whole system.

CHAPTER 3

Project Methodology

This entire project is divided into four steps. The figure 3.1 shows the methodology of the project consisting of four condensed steps.

The first step loads data for the project. Selecting data sources is an essential step towards building an efficient prediction model. All the data sets that are used for the project are discussed in the subsection 3.1.1. The second step processes the data and transform it into usable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. The Data Preprocessing section in 3.2 contains the data mining techniques that prepares the raw data for further processes.

The third step, Load profile generation, is one of the most important steps in this project. Load profile is the hourly data of electrical load consumption for an average user in a defined utility area. This profile takes into account the heating/cooling energy consumption, hot water usage and the set point temperature of the property. Generating a proper and accurate load profile helps in proper training of the predictive model. The formulation of the predictive model is the fourth step and is discussed in the fourth section in 3.4

3.1 Data Loading

3.1.1 Datasets

This project uses a variety of energy & power, weather and residential data sets for generating the required load profile and forecasting model. The datasets used and their descriptions are given below

- EIA (Form 861): The United States Energy Information Administration is a data

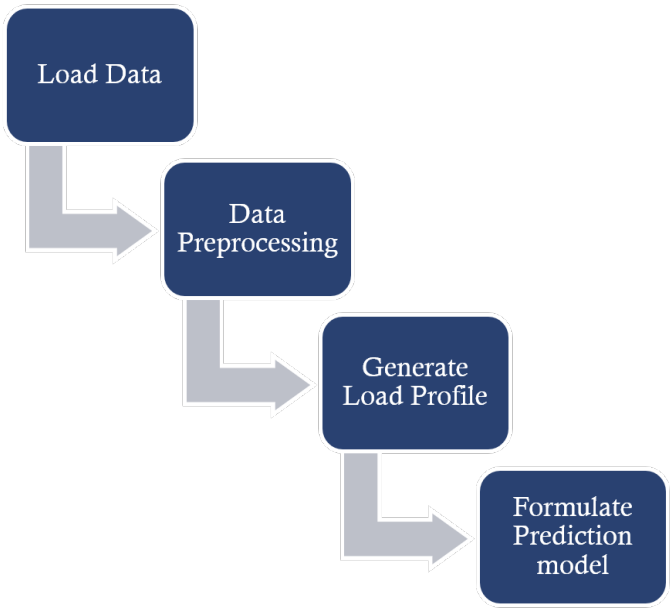

Fig. 3.1: Proposed Algorithm Model

collection and analyzing organization, which collects and hosts energy related data all over the US. The Form EIA-861 and Form EIA-861S (Short Form) data files include information such as peak load, generation, electric purchases, sales, revenues, customer counts and demand-side management programs, green pricing and net metering programs, and distributed generation capacity.

EIA-861S was created in 2012 in an effort to reduce respondent burden and to increase EIA's processing efficiency. Approximately 1,100 utilities completed this form in lieu of the EIA-861. The short form has fewer questions and collects retail sales data as an aggregate and not by customer sector. EIA has estimated the customer sector breakdown for this data and has included it in the file called "Retail Sales". Advanced metering data and time-of-use data are collected on both Form EIA-861 and Form EIA-861S.

- CBECS: Commercial Building Energy Consumption survey. The CBECS is used to determine the end use statistics of the residential water consumption and space heating and cooling.

- US National Observatory: The USNO furnishes data on the daytime hours for each day. It also provides informations about the existing weather stations in the US and the link to fetch the data from those stations.

- USU Utah Climate Center: The USU Utah Climate center has weather stations all over the state of Utah, and furnishes hourly weather data from almost all of its weather stations. Utah's Climate Reference Network (UCRN) is a collection of 16 automated weather stations providing measurements of air temperature, relative humidity, solar radiation, wind and precipitation in near real time. The website also provides hourly and daily summary data tables and graphical displays.

- DOE: The Department of Energy hosts a variety of data related with the energy consumption, sources, end-uses etc. DOE dataset is used in this project for information on hourly hot water usage profiles, probability of each end use and variation in the water usage for different regions.

- NOAA: National Oceanic and Atmospheric Administration provides historical hourly weather data for weather stations globally (NOAA, 2016). Temperature data for weather stations that are not among the 13 weather stations of the USU Utah Climate Center, are fetched from NOAA. The hourly weather data is combined with customer load data to estimate temperature-sensitive loads for residential customers.

3.2 Data Preprocessing

Data preprocessing deals with the formation of usable data sources using the dataset available. Our first task ts to make a consolidated data set of residential users and climatic conditions. The second task is to cluster data points into set of utilities within the same region and having similar climatic conditions. Similar climatic conditions will result in similar Heating Degree Days (HDD) and Cooling Degree Days (CDD) for the selected utilities. The data aggregation for the tasks given above is briefly discussed below –

3.2.1 Residential Data:

The first task is to use EIA form 861 and to load the energy consumption data of all the major utilities over the country. EIA data contains the load characteristics from all over the United States and sectored by the utility names. Since this project performs demand response for the state of Utah, the first level of clustering is to shortlist the utilities of the state of Utah and to cluster the residential properties on basis of utilities.

Fig. 3.2 shows the utility distribution for raw data available before clustering.

```
In [2]: ut_data.head()
```

Out[2]:

	Entity	State	Ownership	Customers	Sales	Revenues	Average Price
0	Alaska Electric Light&Power Co	AK	Investor Owned	14,292	139,475	16,530.00	11.85
1	Alaska Power and Telephone Co	AK	Investor Owned	5,413	25,353	7,478.00	29.5
2	Alaska Village Elec Coop, Inc	AK	Cooperative	7,801	39,493	22,337.10	56.56
3	Anchorage Municipal Light and Power	AK	Municipal	24,555	130,806	21,972.00	16.8
4	Barrow Utils & Elec Coop, Inc	AK	Cooperative	1,500	11,466	1,494.40	13.03

Fig. 3.2: Raw EIA form 861 data - top five rows

The original data set is then filtered out to give the utility areas in the state of Utah, as shown in Fig. 3.3. For clustering purposes, utilities with very small number of users are filtered out. The filtering constant is kept at the first quantile of the data. Fig. 3.4 shows the distribution of customers for different utility areas and the red line shows the first quantile of the distribution. Utilities with the number of customers less than the first quantile limit are marked, to reduce skewness in the training data.

3.2.2 Weather Data:

The weather data is an essential part of this project, as it helps in determining cooling and heating degree days. There are two sources of climatic data for the state of Utah. The first source is the USU Climate center, which hosts data from around 16 weather

```
In [4]: data.head()
```

Out[4]:

	Entity	State	Ownership	Customers	Sales	Revenues	Average Price
1896	Bridger Valley Elec Assn, Inc	UT	Cooperative	1,764	7,372	1,233.90	16.74
1897	Brigham City Corporation	UT	Municipal	6,761	52,586	5,256.00	10
1898	City of Bountiful	UT	Municipal	15,459	148,397	14,486.00	9.76
1899	City of Logan - (UT)	UT	Municipal	17,230	93,585	9,658.70	10.32
1900	City of Murray - (UT)	UT	Municipal	14,615	116,074	10,732.20	9.25

Fig. 3.3: Raw EIA form 861 data - top five rows

stations, and the other is the NOAA, that hosts the weather data for numerous places around the world. Fig 3.5 shows the locations of the utility clusters for this project. Since the USU Climate center provides climatic data for selected cities, in many cases, the exact climatic data of the utility service area was not available. Fig 3.6 shows the locations on the map, for which data was readily available. This issue of absence of data was solved using the technique of interpolation. Weather data from the nearest weather stations were interpolated to fill out the missing data.

For a utility service area, which covers more than one weather stations, the effective hourly temperature T_{eff} is calculated as the weighted average of the hourly temperature and the population served by the weather station [1] –

$$T_{eff} = \frac{\sum_{i=1}^{W} N_i \times t_i}{\sum_{i=1}^{W} N}, \tag{3.1}$$

where N_i is the number of users in the i^{th} weather station and t_i is the temperature of the i^{th} weather station. N is the total combined population of the selected weather stations.

3.2.3 Utility Clustering:

Many big regions of the state, like the Salk Lake City area, Logan City area etc., have more than one utility service organizations. Since the climatic conditions can be assumed

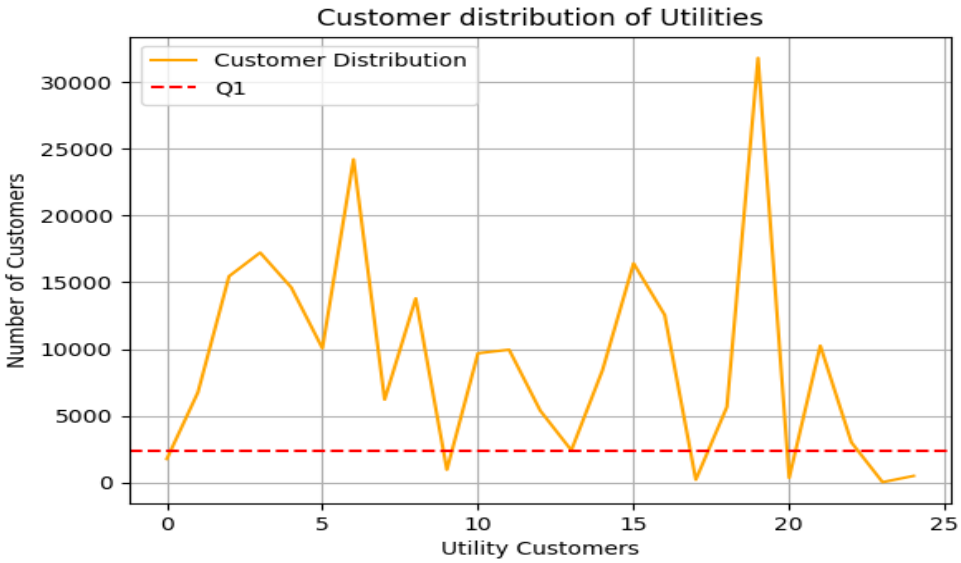

Fig. 3.4: Raw EIA form 861 data - top five rows

to be the same over all the utility areas, and if the utilities serve the same type of end customers, then it is inefficient to formulate different predictive models for these utilities. Following this argument, clusters of similar users are formulated so that the problem of similar models in the project can be reduced. The resulting utilities and their service areas are given in table 3.1 below. It is important to note that each cluster is assigned a letter ID, which represents the cluster wherever it is referenced in the project.

3.2.4 Domestic Hot Water (DHW)

According to Wong Koon Kong [63], the domestic hot water load represents a significant share of the total domestic energy load, ranging from 25% to 40% of the total energy consumption. Thus, proper evaluation of the domestic hot water usage is essential for developing an accurate load profile. According to [64], DHW modeling is quite complex as it involves a wide spectrum of end-uses and applications with varying inlet temperatures, volumes, flow rates, and timing. DHW use can be broken down mainly to five major end uses - showers, baths, sinks, dishwasher, and clothes/washer. The average daily consumption of hot water varies between households by an order of magnitude around 50 liters/day to

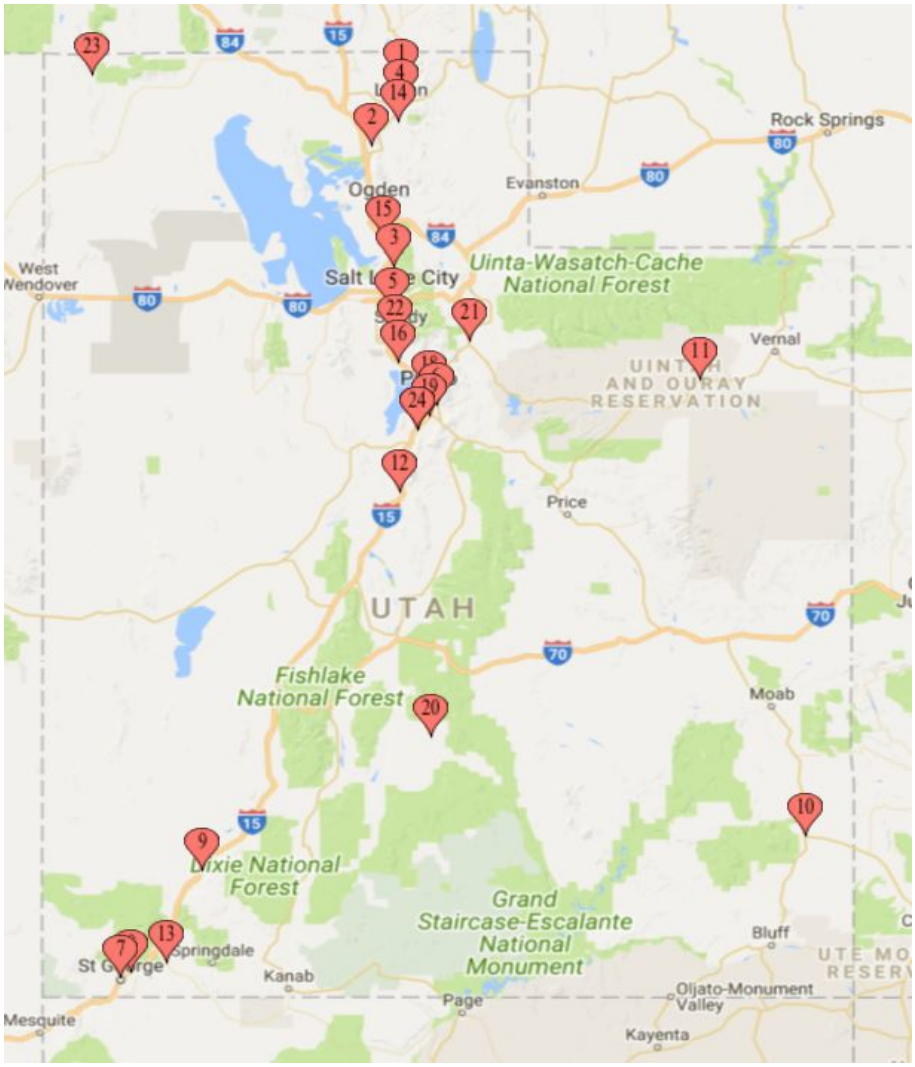

Fig. 3.5: Utility service areas in the original dataset

approximately 500 liters/day, and depends upon the number of people and their way of living [65].

Fig. 3.7 shows the percentage of water consumption with respect to different end-uses. Here Faucet denotes the water consumption from domestic sinks and washer for cloths. It is interesting to note that around 5% of hot water is wasted through leaks and other losses, which also accounts for the total energy consumption.

It is important to note that DHW use varies stochastically with random unoccupied/unused periods, varying number of showers per person per day with varying flow

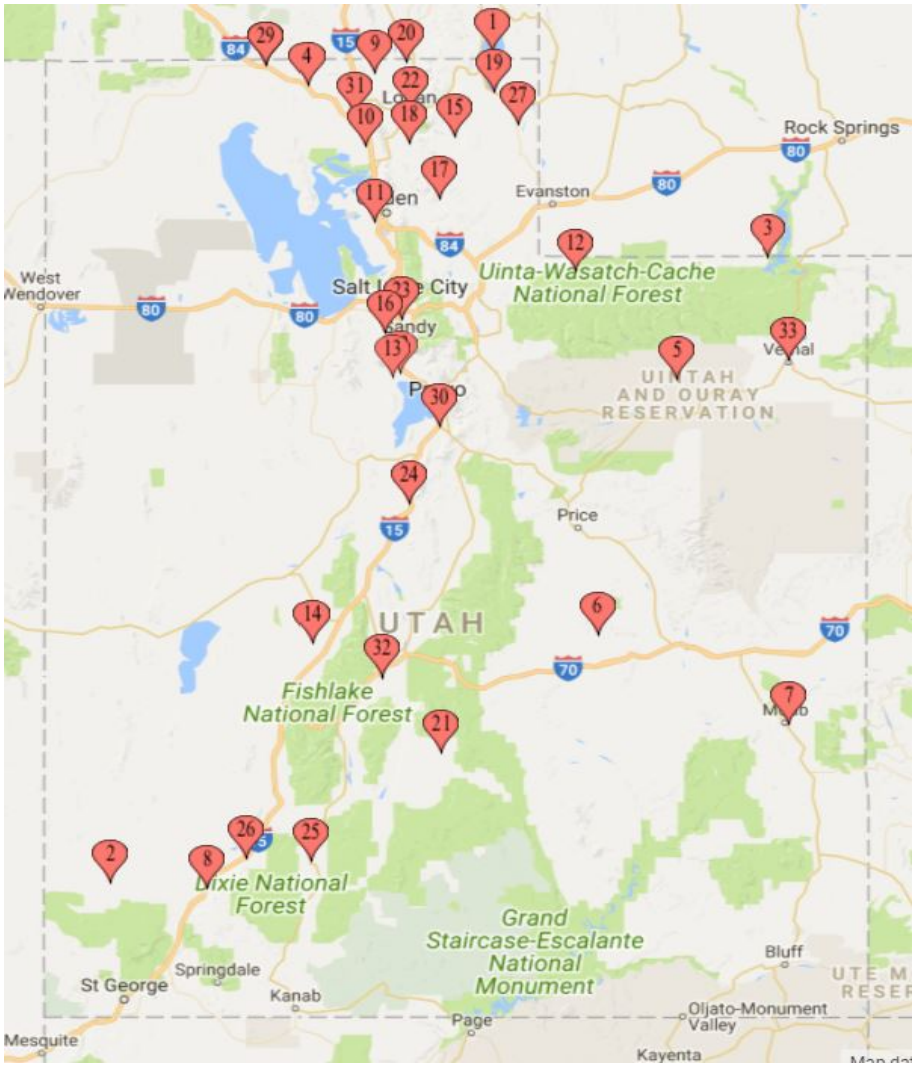

Fig. 3.6: Nearest weather stations to the utility service areas

rates, washer loads and wash temperatures and frequency of dishwasher use [63].

Hot water use profile can be generated using the NREL's hot water calculation equation [66] given in table 3.2. The daily average hot water use equations were established as a linear function of the number bedrooms, acting as a related variable for the number of occupants. The relationship between bedrooms and occupants is based on the DOEs 2001 Residential Energy Consumption Survey (RECS) given in [67].

Table 3.1: Utilities and their corresponding closest locations as per climate data and utility size

Utility Name	Location/Region	Utility Code/ID
Bridger Valley Elec Assn, Inc	Smithfield	a
Brigham City Corporation	Brigham City	b
City of Bountiful Utility	Bountiful	c
City of Logan Utility	Logan City	d
City of Murray Utility	Murray	e
City of Springville Utility	Springville	f
City of St George Utility	St. George	g
City of Washington Utility	Washington City	h
Dixie Escalante R E A, Inc	St. George	i
Empire Electric Assn, Inc	Monticello	j
Garkane Energy Coop, Inc	Loa	k
Heber Light & Power Company	Lehi	l
Hurricane City Power	Hurricane City	m
Hyrum City Power Corporation	Hyrum City	n
Kaysville City (Power) Corporation	Kaysville	o
Lehi City (Power) Corporation	Lehi	p
Moon Lake Electric Assn Inc.	Roosevelt City	q
Mt. Wheeler Power Inc	Nephi	r
PacifiCorp	Draper	s
Payson City (Power) Corporation	Payson City	t
Provo City (Power) Corporation	Provo	u
Raft Rural Elec. Corp. Inc.	Lynn	v
Spanish Fork City (Power) Corporation	Spanish Fork	w
Strawberry Electric Service Dist.	Payson City	x
Vivint Solar, Inc.	Lehi	y
Wells Rural Electric Co.	Wendover	z

Jordan U et. al. in [68] furnishes the mathematical equations to convert the amount of water consumed into electrical load.

$$Q = V_d.t.h_q,$$

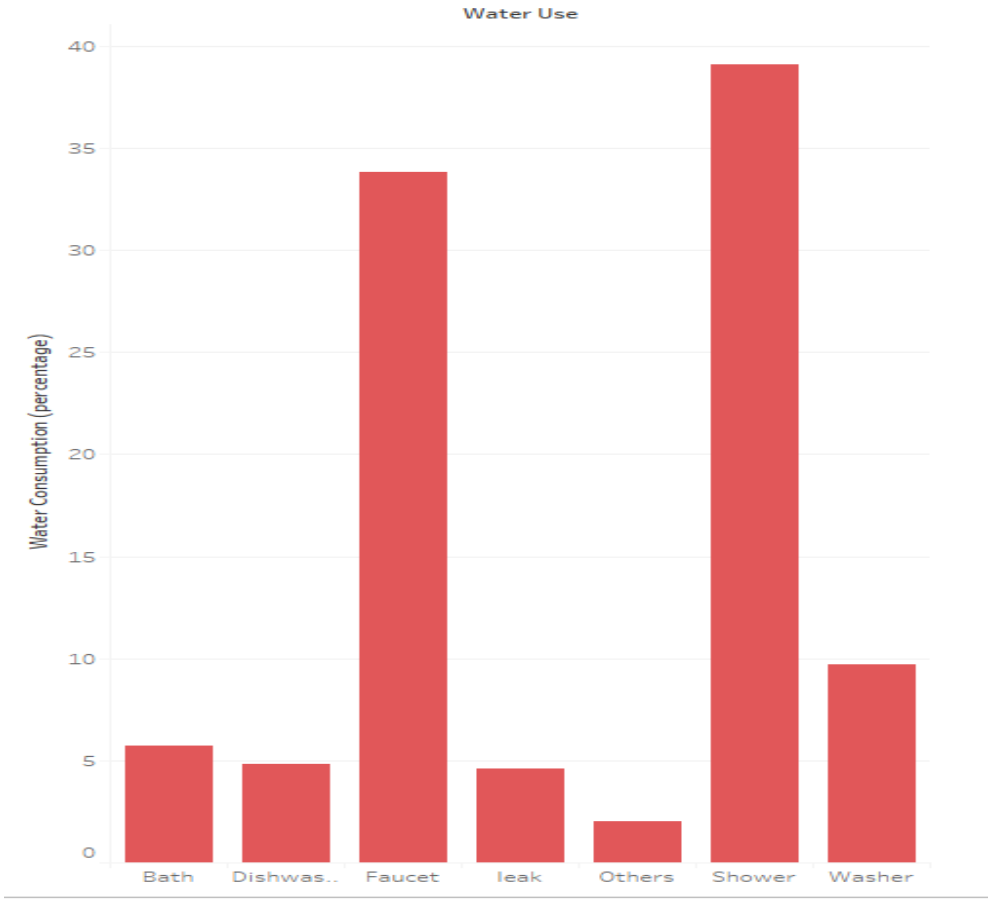

Fig. 3.7: Raw EIA form 861 data - top five rows

Table 3.2: DHW consumption equations

Water End Use	**Volume of Water Used** (gal/day)
Common laundry	2.47 per Building
Washer (Cloths)	$7.25 + 2.5 \times$ No. of room
Washer (Dishes)	$7.25 + 2.5 \times$ No. of room
Shower	$14 + 4.67 \times$ No. of room (incl. cold)
Baths	$3.5 + 1.17 \times$ No. of room (incl. cold)
Sinks	$12.5 + 4.16 \times$ No. of room

where V_d is the flow volume of water in liters per minute, t is the duration of event in minutes and h_q is the heating coefficient of the medium. For the case of water, the heating coefficient h_q is assumed as 40.6 Wh/Kg. Fig. 3.8 shows the probability of DHW use with

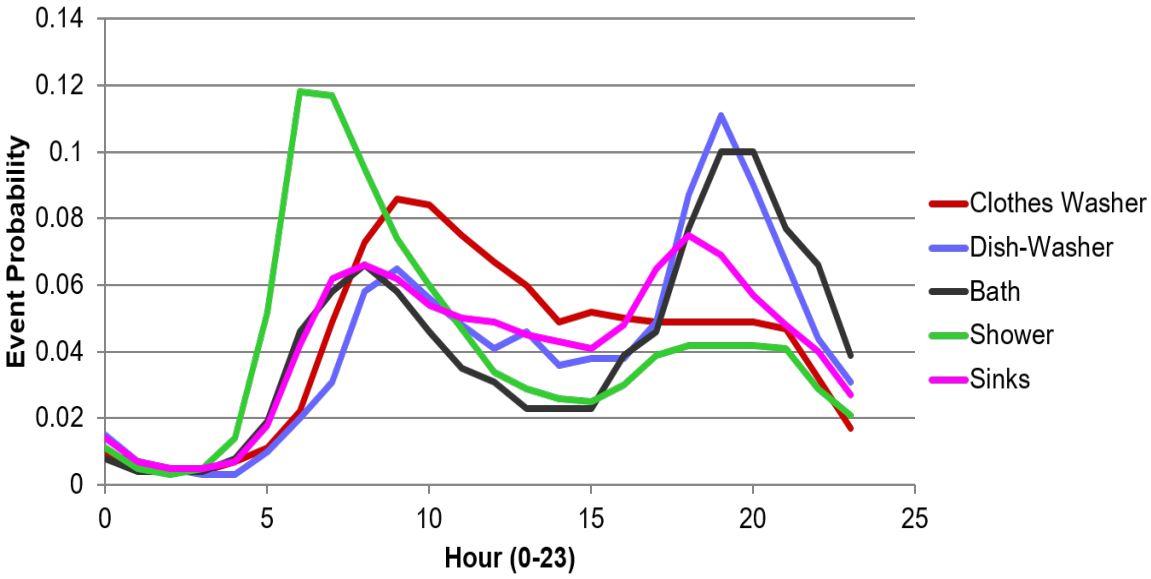

Fig. 3.8: Probability of DHW usage in residential properties by LBNL Labs [1]

respect to the time of the day. This probability distribution, along with the volume of water use of end-water use, is used to determine to the volume flow of DHW. Fig. 3.9 shows the DHW consumption profile for hourly loads.

3.3 Load Profile Generation

After the pre-processing stage, a combined hourly load profile is generated, which is used for training the prediction model. The first stage for generating the load profile is to formulate hourly load data for cooling and heating purposes. Once the load profile for heating/cooling consumption is generated, the final load profile for the building can be generated by taking a time-series summation of heating/cooling data and DHW data.

3.3.1 Heating and Cooling Load

Weather and temperature are important drivers for electricity consumption. More than 40% of end-use energy consumption is related to the heating and cooling needs in the residential and commercial sectors. Electricity consumption forecasting models typically use thresholds for defining when the cooling and heating needs are required. A fairly

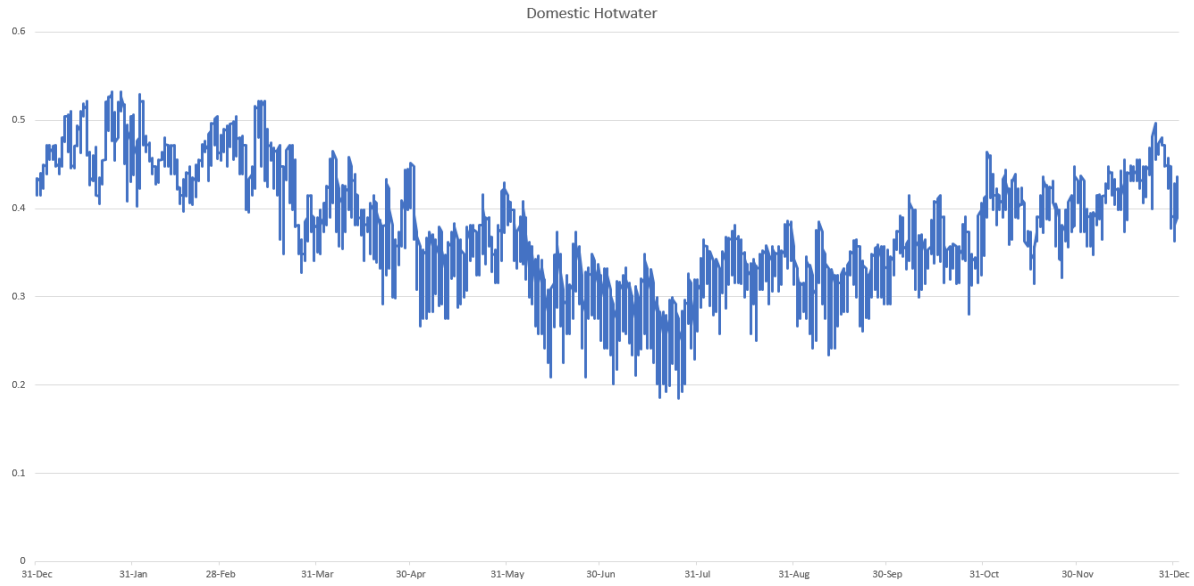

Fig. 3.9: Hourly electric load in KwH for DHW consumption

standard set of thresholds, also called setpoint temperature, are temperatures above 72 degrees Fahrenheit and below 65 degrees Fahrenheit respectively. As the difference between the outside temperature and the setpoint temperatures increase, the cooling/heating load also increases. The setpoint temperature is assumed to be fixed for this project and has been set to be 65 F. Once the setpoint temperature is fixed, we proceed towards calculation of Cooling Degree Days (CDD) and Heating Degree Days (HDD).

3.3.2 Cooling Degree Days (CDD) and Heating Degree Days (HDD)

The cooling degree days are a measure of energy required to cool down a building, when the external temperature is higher than the desired building temperature. On the other hand, Heating degree days are a measure of the heat energy required to heat up a property or a building, when the external temperature is below the desired building temperature. Both measures are complimentary to each other and are visualized later in this subsection.

The daily CDD is calculated using the hourly temperature difference between the setpoint temperature and air temperature, summed over 24 hours. Similarly daily HDD calculated using the hourly temperature difference between the air temperature and setpoint

temperature, summed over 24 hours. This means that the CDD and HDD are directly dependent on the air temperature outside and can drastically vary with the variation in the season.

Mathematically,

$$HDD = \frac{\sum_{i=1}^{t}(T_i^{out} - T_i^{set})}{24}, \tag{3.2}$$

$$CDD = \frac{\sum_{i=1}^{t}(T_i^{set} - T_i^{out})}{24}, \tag{3.3}$$

where T^{out} and T^{set} are the outside temperature and the setpoint temperature respectively.

For a load profile, the degree days need to be transformed into energy consumed for either heating or cooling of the building. The annual energy consumption for heating purpose (kWh) for a given building, is expressed as [69]

$$Q = \frac{U^{'} \times AHDD \times 24}{\eta}, \tag{3.4}$$

where $AHDD$ is the annual HDD of the building, η is the performance efficiency of the heating source and $U^{'}$ is the heat loss coefficient of the building. The building loss coefficient of the building can further be modeled as

$$U^{'} = \frac{A \times U + 0.33 \times N \times V}{1000}, \tag{3.5}$$

where A is the area of the building heated, N is the air filtration rate per hour and V is the volume of space heated. To calculate the approximate value of $U^{'}$, default values for $U^{'}$ equation are given in [69]. Taking the default values, $U^{'}$ can be calculated as $0.350 kW/K$ and the efficiency η can be averaged at 0.9, which means a 90% efficiency is assumed for the heating device.

Thus, the final equation for calculation of the energy consumed for HDD becomes

$$Q = \frac{0.350 \times AHDD \times 24}{0.9}, \tag{3.6}$$

The annual energy consumption for cooling purpose (kWh) for a given building, is expressed as [69]

$$Q = \frac{mC_p \times ACDD \times 24}{\Psi}, \tag{3.7}$$

where $ACDD$ is the annual CDD, Ψ is the coefficient of performance of the cooling source, C_p is the specific heat capacity of air and m denotes the mass flow rate of kilogram of air cooled per second. For a 12000 BTU air conditioning unit, m and Ψ can be assumed to be $0.109Kg/sec^3$ and 3.5 respectively. Also, the specific heat of air is $1.005KJ/Kg/K$. Using these constants, the CDD and HDD values can be readily calculated.

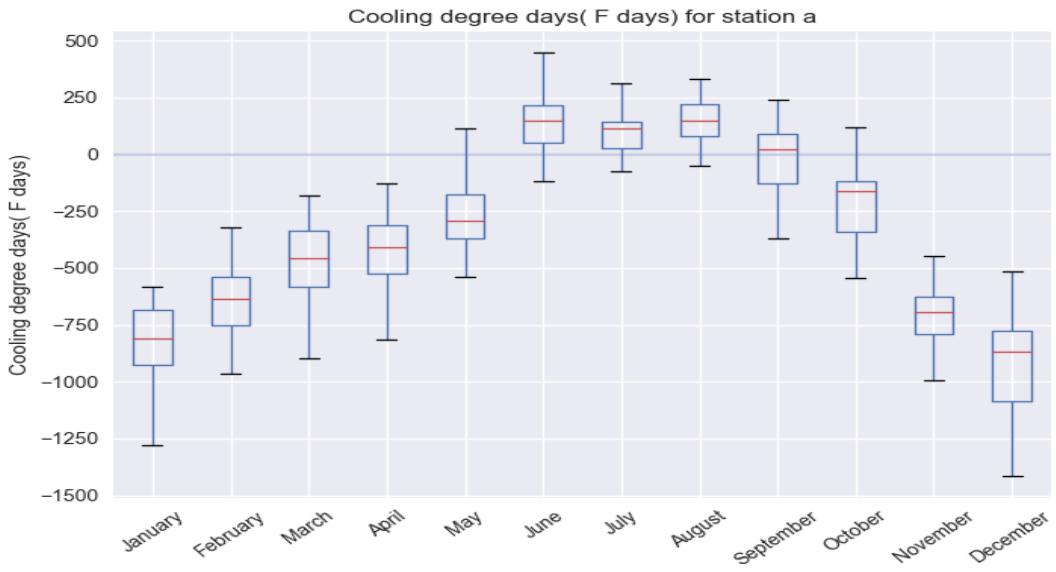

Fig. 3.10: Cooling Degree Days (CDD) for Cluster a (Bridger Valley Elec. Ass. Inc.)

The Figures 3.10 to 3.17 shows the cooling degree days for eight random utility clusters. The CDD box plots shows a common trend for almost all the clusters. The CDD increases in the hotter months and diminishes in the cooler months. A negative value of CDD means that the setpoint temperature is already lower than the outside temperature, and cooling is not required.

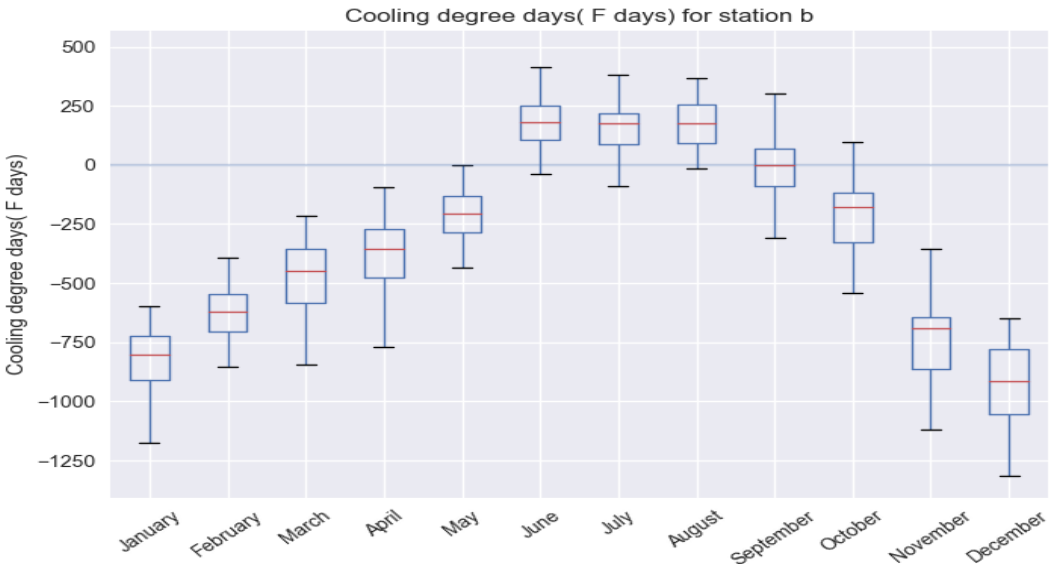

Fig. 3.11: Cooling Degree Days (CDD) for Cluster b (Brigham City Corporation)

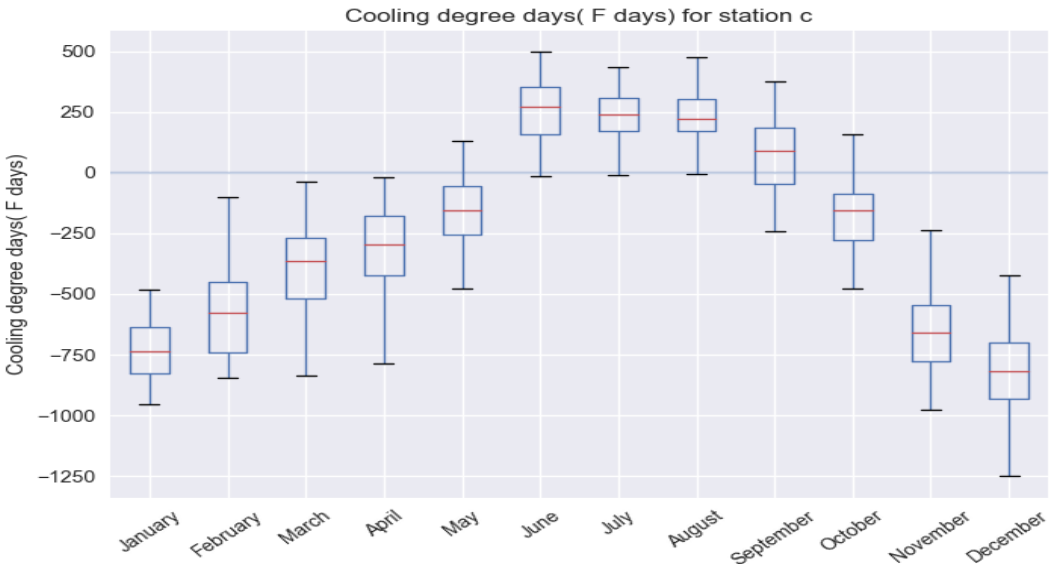

Fig. 3.12: Cooling Degree Days (CDD) for Cluster c (City of Bountiful)

A similar visual expression can be given for HDD or Heating Degree Days. Figures 3.18 to 3.24 gives the box-plots of HDD variation for every month of the year.

The HDD figures also show a similar trend for all the utility clusters. Although the variance of HDD (and CDD too) varies widely over different clusters, the trend is similar. This is due

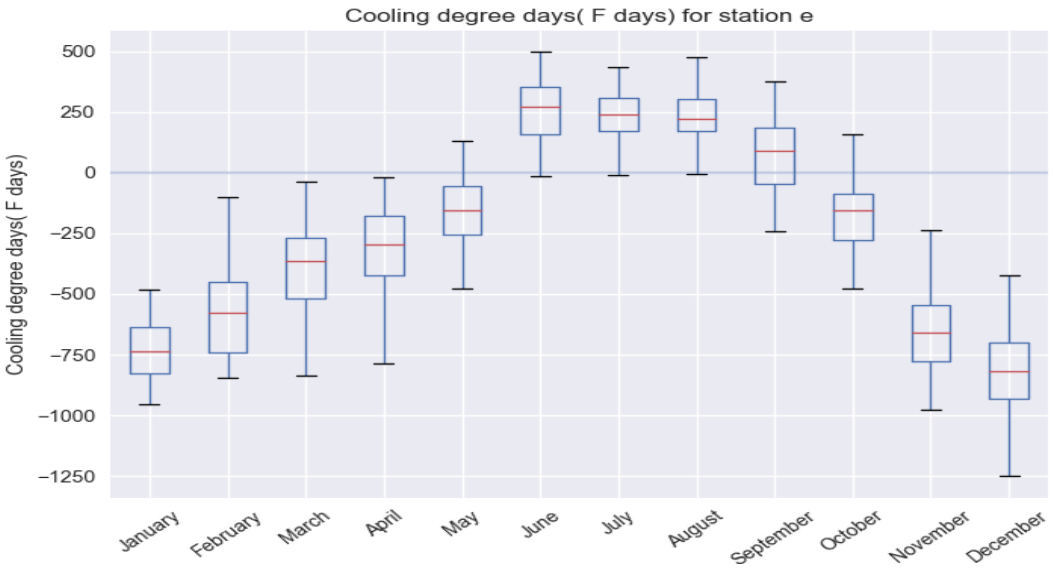

Fig. 3.13: Cooling Degree Days (CDD) for Cluster e (City of Murray Utility)

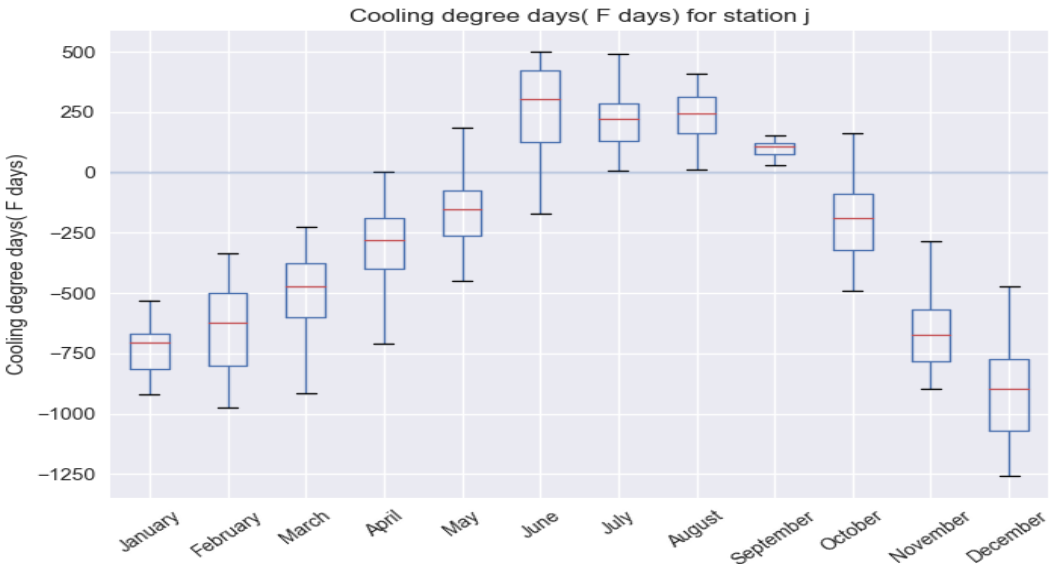

Fig. 3.14: Cooling Degree Days (CDD) for Cluster j (Empire Electric Assn.)

to the fact that HDD is proportional to the difference between the setpoint temperatures and the outside temperature. As the difference is reduced, the HDD is also reduces accordingly. The trend in the variation of box-plots over the year is dependent on the seasonal structure of the cluster region. In the warmer months, heating of buildings is not required, and thus

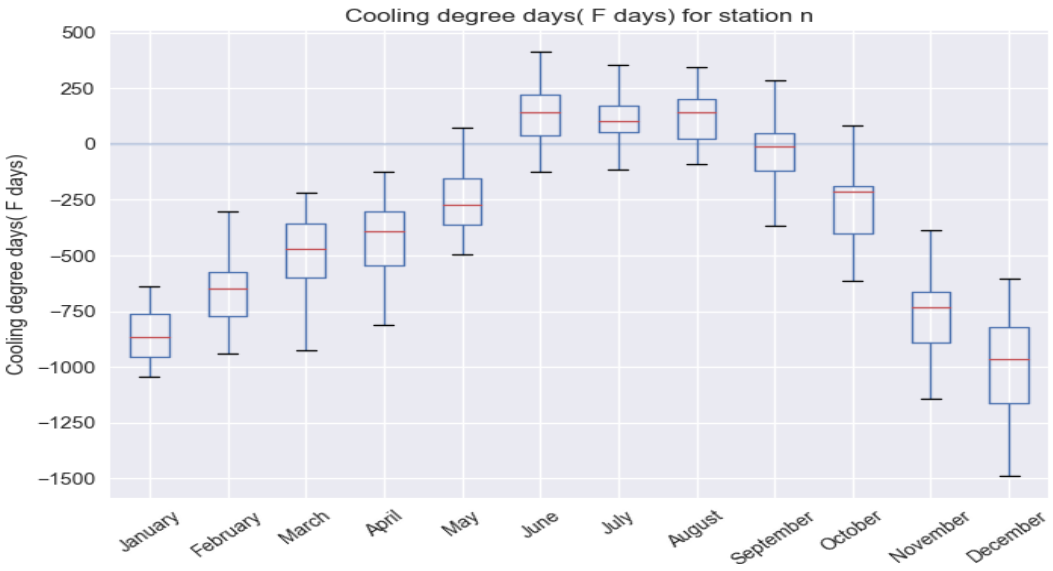

Fig. 3.15: Cooling Degree Days (CDD) for Cluster n (Hyrum City Power Corp.)

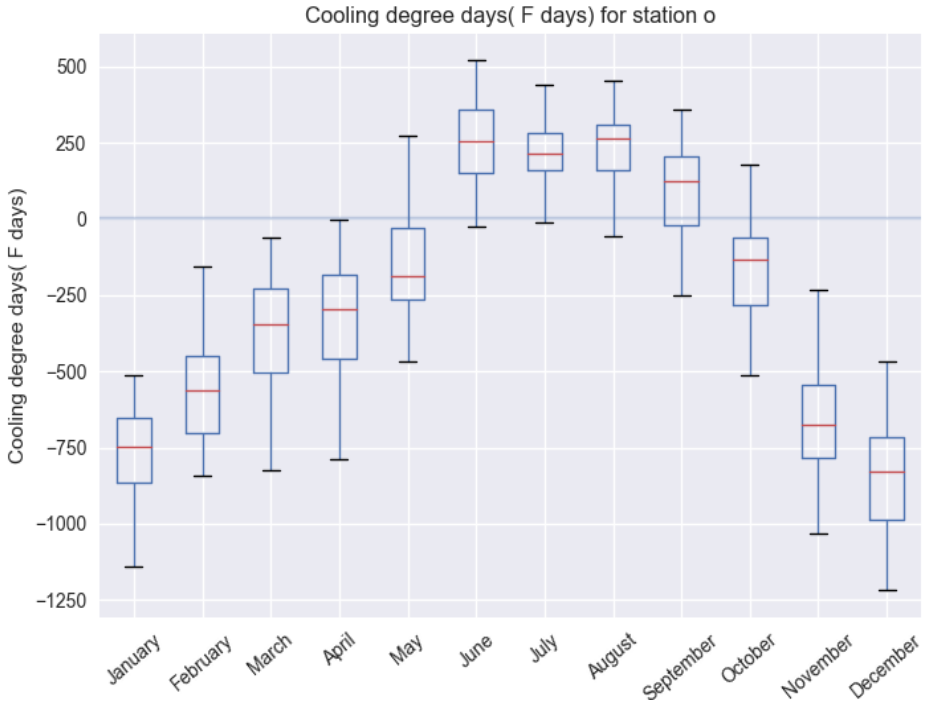

Fig. 3.16: Cooling Degree Days (CDD) for Cluster o (Kaysville City Corp.)

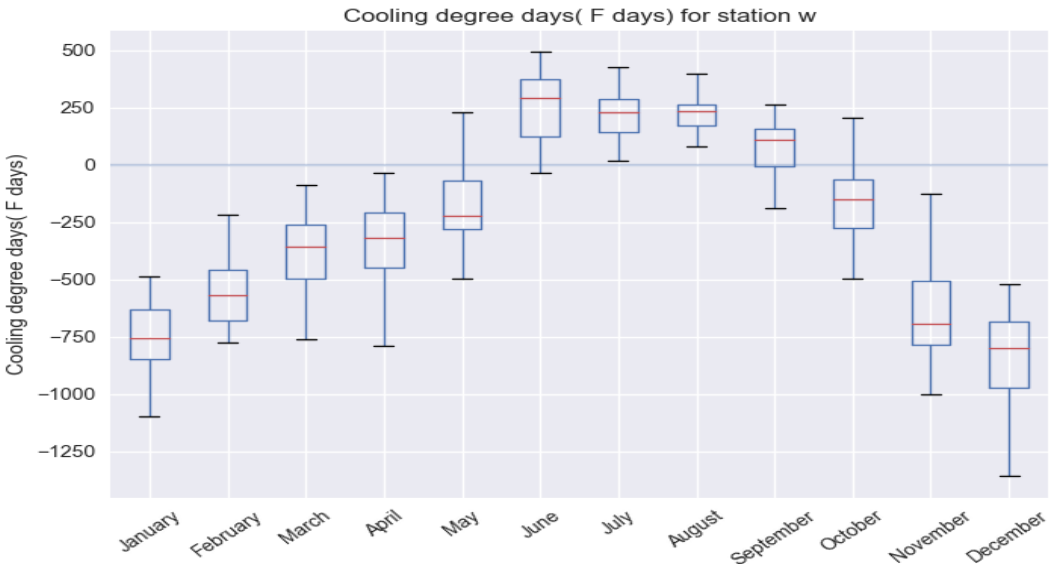

Fig. 3.17: Cooling Degree Days (CDD) for Cluster w (Spanish Fork Power Corp.)

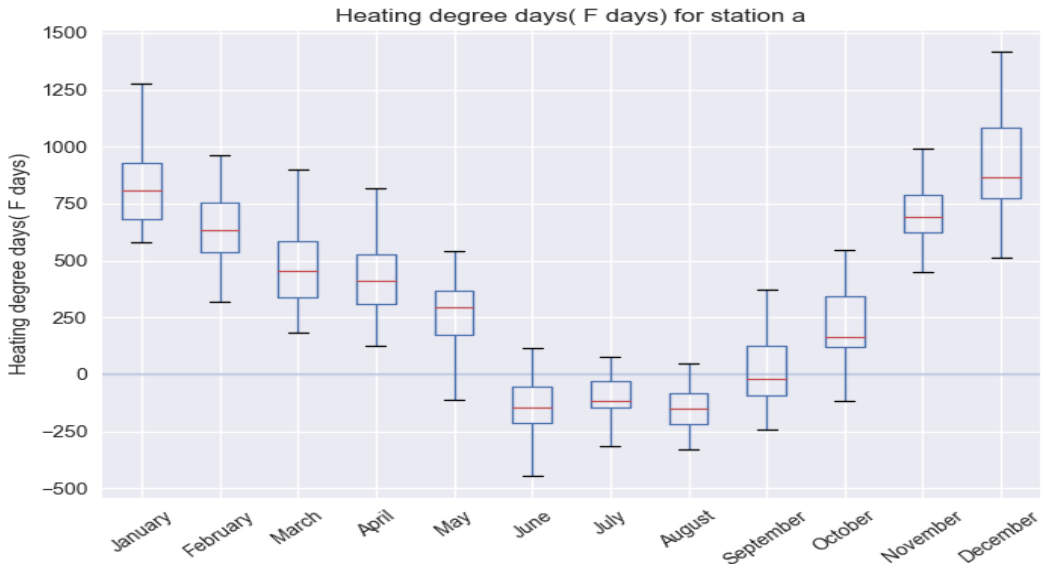

Fig. 3.18: Heating Degree Days (HDD) for Cluster a (Bridger Valley Elec. Ass. Inc.)

the HDD falls below 0; whereas in cooler months or winter months, excessive heating is required, which results in high HDD values in the beginning and ending months of the year.

CDD and HDD are important factors in calculating the load profile of the cluster. Once the energy consumed is generated using CDD/HDD, it is combined with the hourly load

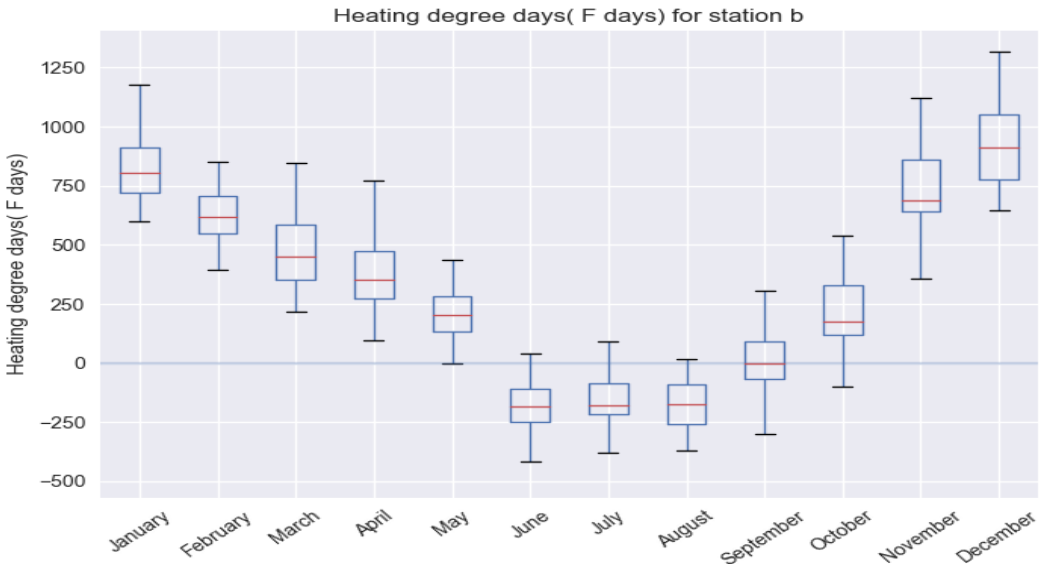

Fig. 3.19: Heating Degree Days (HDD) for Cluster b (Brigham City Corporation)

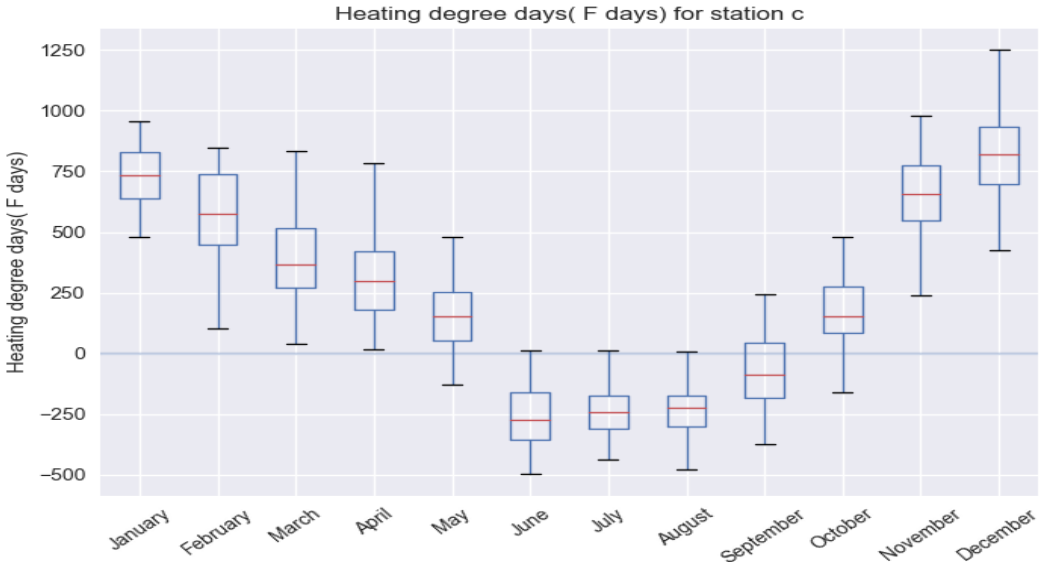

Fig. 3.20: Heating Degree Days (HDD) for Cluster c (City of Bountiful)

data for DHW consumption and the final load profile is calculated for a particular cluster.

3.3.3 Load Profile Data Training

Post generation of CDD/HDD load profile and hourly load of DWH consumptions, a

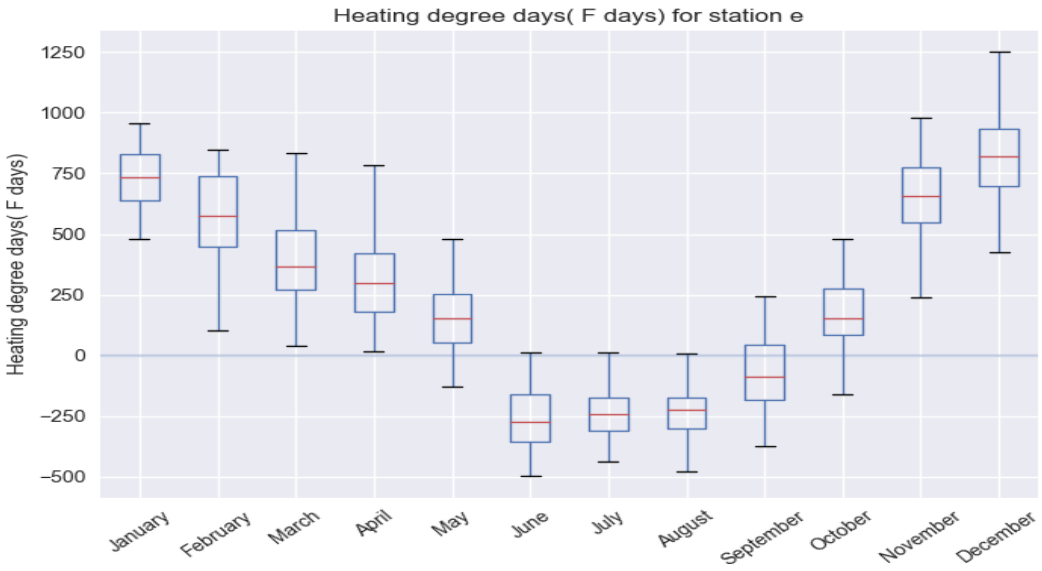

Fig. 3.21: Heating Degree Days (HDD) for Cluster e (City of Murray Utility)

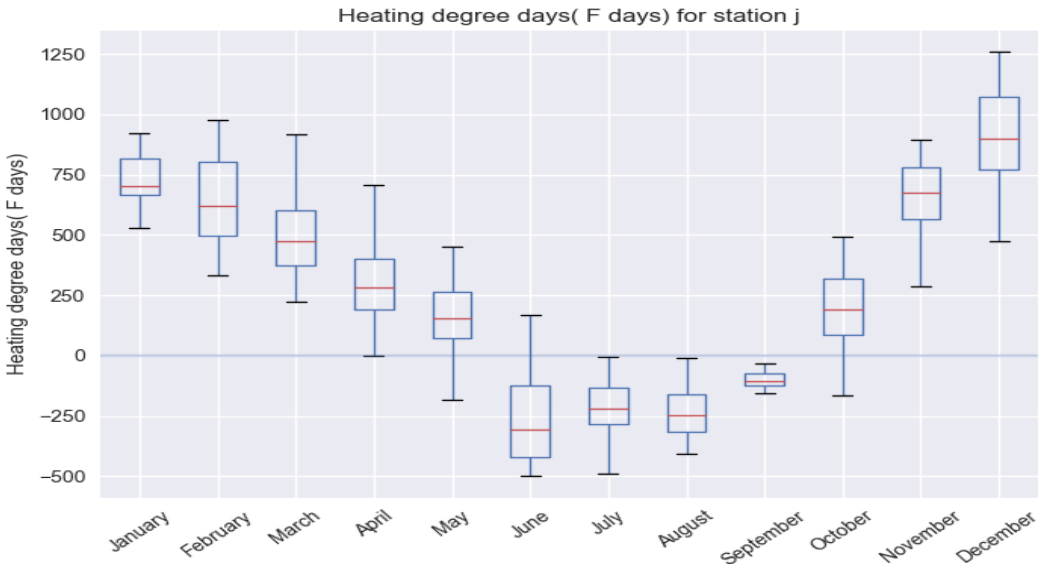

Fig. 3.22: Heating Degree Days (CDD) for Cluster j (Empire Electric Assn.)

combined load profile for the building is formulated. The combined load profile is then used as a training data set for the prediction model. Fig 3.27 gives the hourly load profile of cluster *a*, which is Bridger Valley Electric Association Inc. The hourly load profile can be understood by fig 3.26, which shows that electric loads are dependent on the time of the

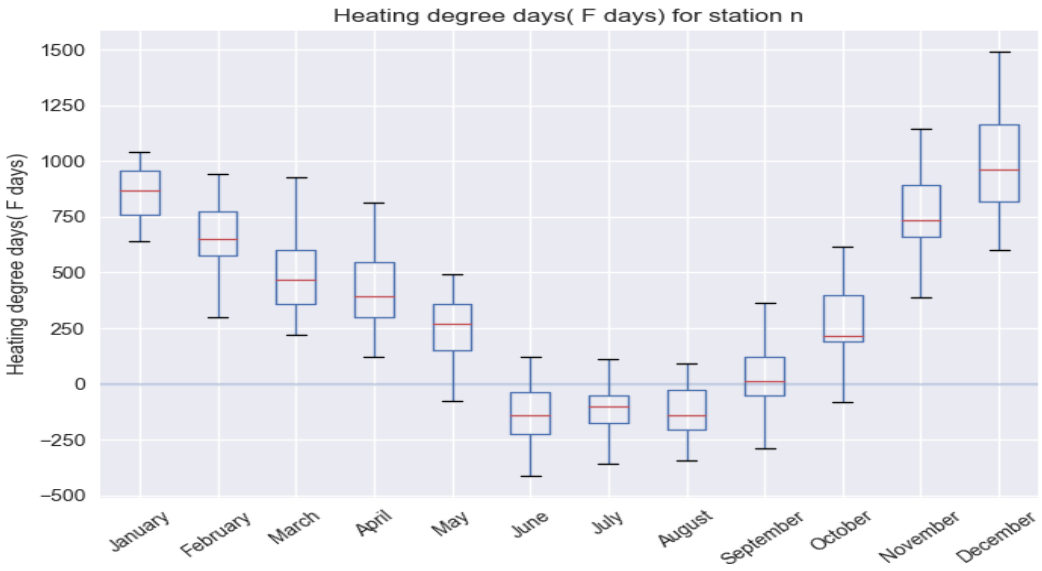

Fig. 3.23: Heating Degree Days (HDD) for Cluster n (Hyrum City Power Corp.)

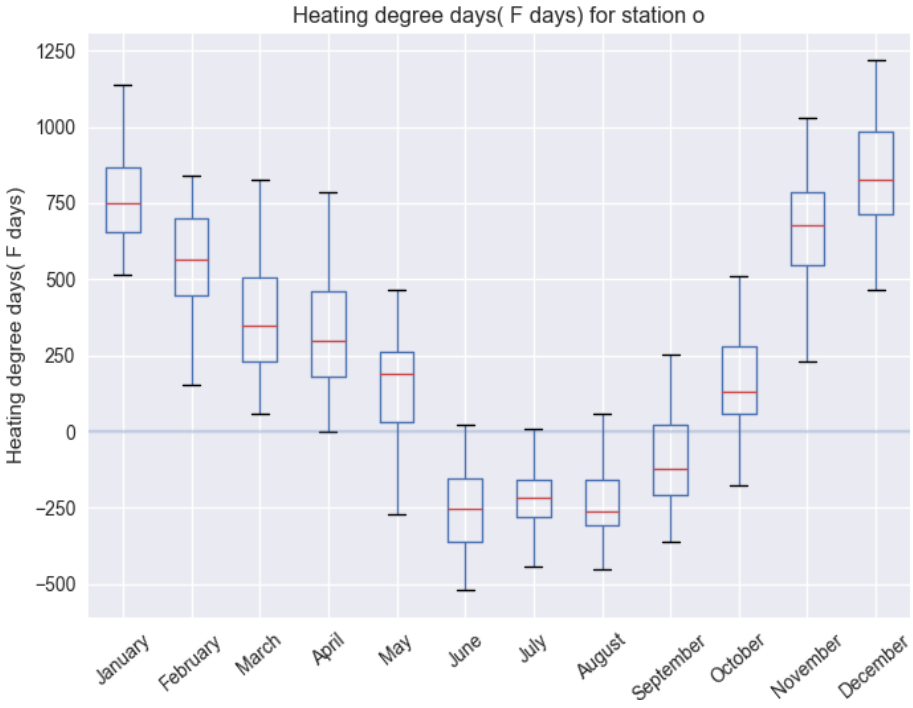

Fig. 3.24: Heating Degree Days (CDD) for Cluster o (Kaysville City Corp.)

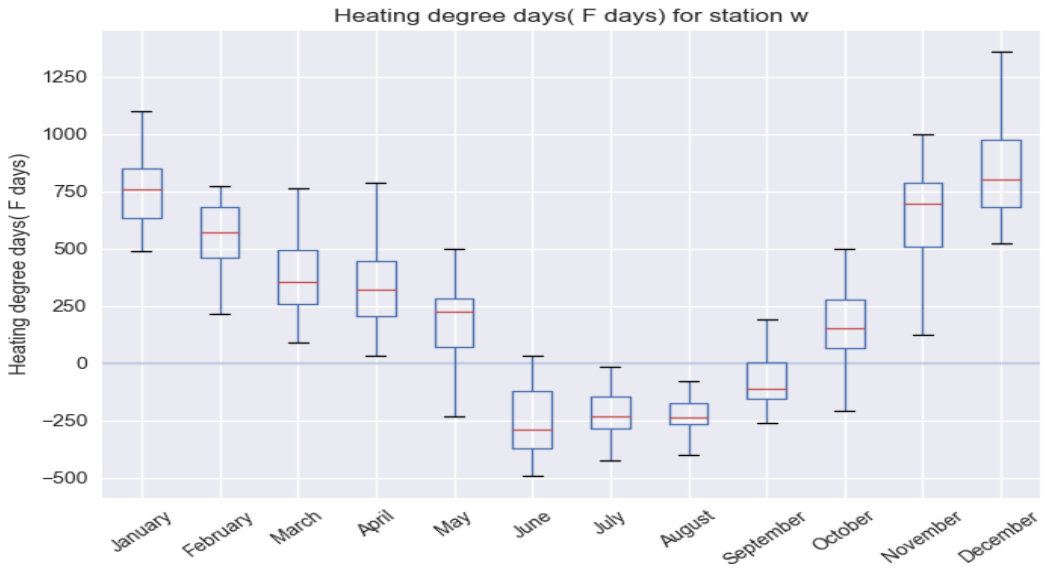

Fig. 3.25: Heating Degree Days (HDD) for Cluster w (Spanish Fork Power Corp.)

day.

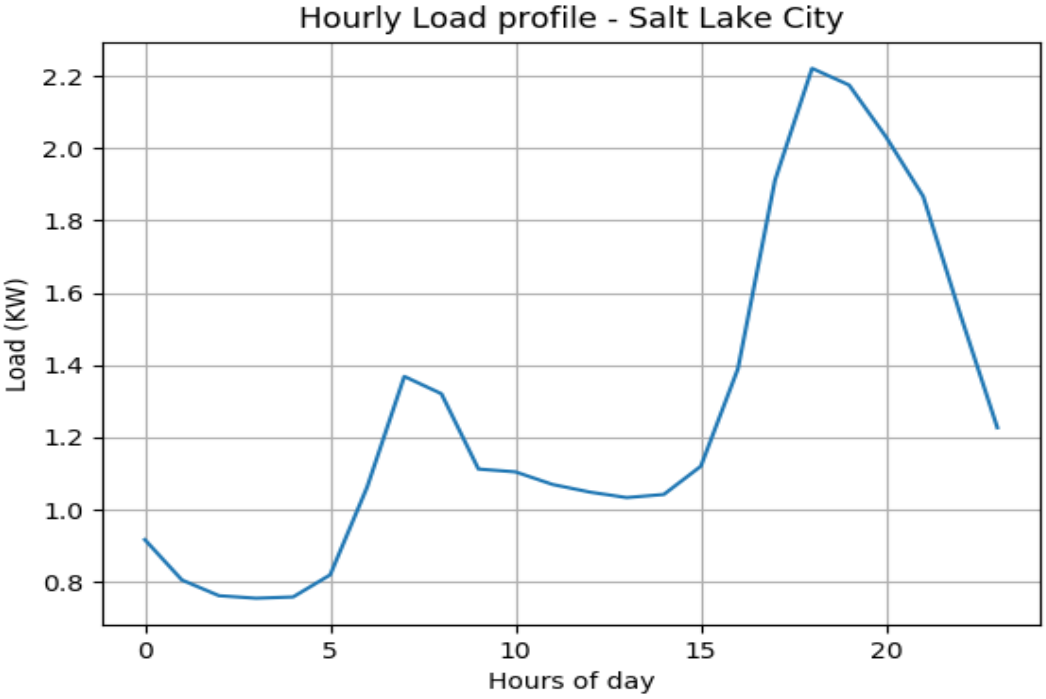

Fig. 3.26: Hourly Load Profile of Salt Lake City Cluster - January 1

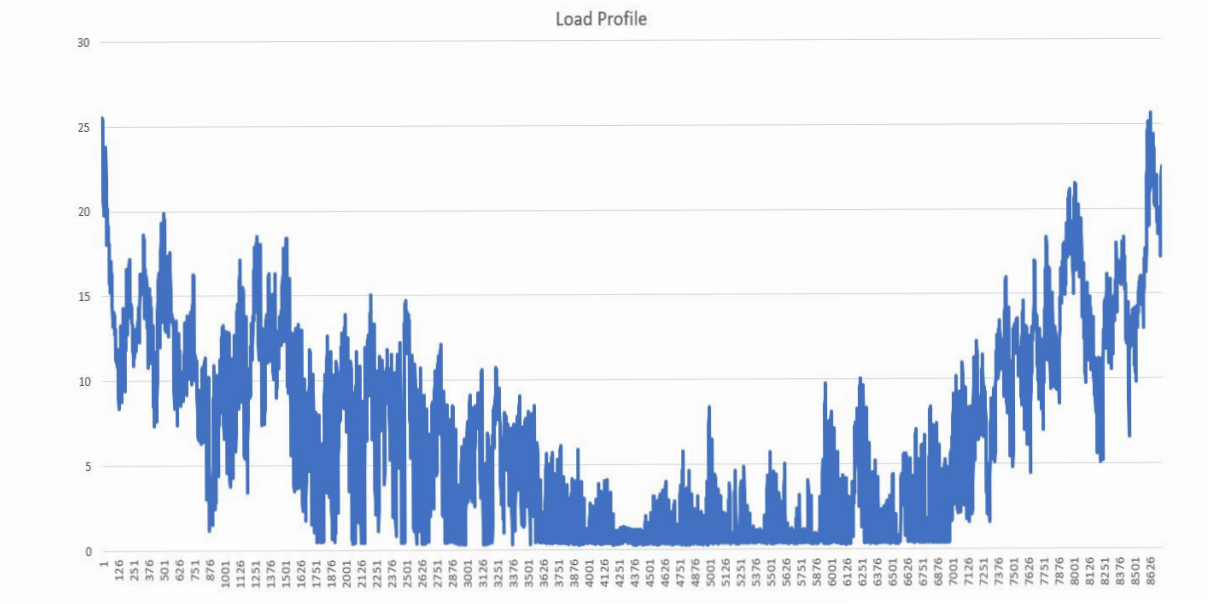

Fig. 3.27: Final Load profile for cluster a

Fig 3.27 contains many mini peaks and troughs with a general trend of a dip in the middle and peaks at the beginning and end of the year. The climatic conditions of the state of Utah can be credited for this trend in the load profile. Being a cold state, at state of Utah the amount of energy spent is usually high in the winter seasons, both for space heating and water heating. Energy consumption reduces during the summer months, as the cooling requirements are not as high as the heating requirements in winter months. The mini peaks and troughs are caused due to the variation in load consumptions over a single day, as evident from Fig. 3.26.

3.4 Predictive Modeling

In this project a predictive model for the short term load forecasting is formulated using the Multiple Linear Regression (MLR) method and Gradient Boosting Technique. In this section, the MLR model and the Gradient boosting regression technique are discussed briefly. The formulation equations are discussed and machine learning techniques for checking the accuracy and score of the predictive model are also discussed.

3.4.1 Multiple Linear Regression (MLR)

Regression analysis aims to find a relation between the dependent variable with the independent variables by calculating the balancing coefficients in the regression equation. A general regression function [70] can be expressed as

$$\mathbb{E}(Y|X) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ...\alpha_n x_n, \tag{3.8}$$

where $\alpha_i, i \in [1, n]$ are the prediction coefficients or slopes and denote the weight of each independent variable on the dependent variable; x_i's are the independent variables; α_0 is called the intercept value of the fitted line.

For each response variable, the equation 3.8 can be modified and expressed as

$$Y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + ...\alpha_n x_{i,n} + \epsilon_i \text{ for } i = 1, 2, ...p, \tag{3.9}$$

which can be reorganized as

$$Y_i = \mathbb{E}(Y|X) + \epsilon_i \text{ for } i = 1, 2, ...p, \tag{3.10}$$

where ϵ_i's are the prediction error, also known as residuals, and can be modeled as a random variable with 0 mean $\mathbb{E}_{\epsilon_i} = 0$ and a constant variance $var_{\epsilon_i} = \sigma^2$. The observed values for Y have the same standard deviation of σ and a mean of μ_y. Using the least square technique for fitting the model, the best fitting line for the given data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the fit line. In other words, the fit line that minimizes error or reduces the distance between the line and data points becomes the best fitted line.

Following the predictor equation 3.9, the values fit by the equation $\alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + ...\alpha_n x_{i,n}$ can be expressed as $\hat{y}_i$. So, the residuals can be expressed as $\epsilon_i = y_i - \hat{y}_i$. For the best fitted line, our aim is to minimize this value of ϵ_i.

The estimated variance, also called the MSE or mean squared error, can be expressed as

$$\sigma^2 = \frac{\sum_{\epsilon_i^2}}{p-n-1}, \tag{3.11}$$

whereas the standard error of the model is the square root of the MSE

$$s_e = \sqrt{\frac{\sum_{\epsilon_i^2}}{p-n-1}}. \tag{3.12}$$

3.4.2 Gradient Boosting Regression

Gradient boosting is a classic machine learning technique that relies on the fact that iterating a simple decision tree over and over again leads to convergence and high accuracy. Like other boosting methods, gradient boosting combines weak learners into a single strong learner in an iterative fashion. A weak learner is defined as the one whose performance is at least slightly better than random chance.

A naive depiction of the gradient boosting technique is given in the following equations.

- Fit an initial model to the data

$$F_1(x) = y. \tag{3.13}$$

- Fit a model to the residual values

$$r_1(X) = Y - F_1(X). \tag{3.14}$$

- Create a new model

$$F_2(x) = F_1(x) + r_1(x). \tag{3.15}$$

- Repeat the process

The above steps can be readily generalized as given by Ben Gorman in [71].

$$F(x) = F_1(x) \Rightarrow F_2(x) = F_1(x) + r_1(x) \Rightarrow F_3(x) = F_2(x) + r_1(x) \cdots \Rightarrow F_n(x) = F_{n-1}(x) + r_{n-2}(x). \tag{3.16}$$

So we keep searching for the next step of residual value $r_i(x)$ at every boosting iteration.

Now, to minimize the squared error (MSE), we initialize F as the mean value of the training set data

$$F_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} (\gamma - y_i)^2, \tag{3.17}$$

$$F_0(x) = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{3.18}$$

Now, all we need is to determine the factor r_m coming from a class of base learners, which in our case are the decision trees. To find the appropriate value of m, the concept of cross-validation is used.

A regression tree, which by default minimizes the square error, focuses heavily on reducing the residual of the first training sample. But if we want to minimize the absolute error, moving each prediction one unit closer to the desired target produces an equal reduction in the cost function. So, accordingly, instead of training r_0 on the residuals of F_0, we can train r_0 on the gradient of the loss function, $\mathcal{L}(y, F_0(x))$ with respect to the prediction values produced by $F_0(x)$. By far, the algorithm [71] can be written in a modified form as

- Initialize the model

$$F_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, \gamma). \tag{3.19}$$

- For $m \in [1, M]$,
  - Calculate residual function

$$R_{i,m} = -\left(\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} \right)_{F(x)=F_{(m-1)}(x)}, \text{ for } i = 1 \cdots n. \tag{3.20}$$

  - Compute $r_m(x)$ by fitting it to the pseudo residuals
  - Update $F_m(x)$

$$F_m(x) = F_{(m-1)}(x) + \gamma_m r_m(x). \tag{3.21}$$

3.5 Model Formulation

3.5.1 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in the model construction and is also called variable selection or attribute selection. Feature selection methods help to create an accurate predictive model as fewer attributes can reduce the complexity of the model and prevent the occurrence of errors. Among many used techniques to select feature for a model, Chi squared test, information gain and correlation coefficient scores are most widely used.

For this project, the number of features are already reduced by the formation of the combined load profile. Since the load profile is used to train the prediction model, a large number of attributes are not needed, as the training set contains most of the required information. It is evident from the section of Data Preprocessing (3.2) that the hourly load profile is heavily dependent on the outside air temperature. Thus, one of the attributes for the model is taken as the hourly temperature, because of its high correlation with the load.

The second and third attributes in this project are the absolute values of CDD and HDD. Only absolute values are taken into consideration because negative values of CDD and HDD are not important in the terms of energy consumed. An important point to note is that, the hourly load profile given in fig 3.27 shows variations each day, with the load dipping during weekends and soaring on Fridays and other days. This concept, proposed by Johanna L. Mathieu in [72], indicates that the load of a residential property is dependent on the day of the week and varies accordingly. Due to this fact, the fourth feature for the prediction model becomes the day of week. This feature adds to the time-dependent nature of the model and shows that the load varies both on the time of the day as well as the day of the week. The regression model now can be expressed as

$$L_{pred} = \alpha_0 + \alpha_1 X_{temperature} + \alpha_2 X_{CDD} + \alpha_3 X_{HDD} + \alpha_4 X_{Day}. \tag{3.22}$$

The day of week feature of the prediction model can be one of the seven days of the

```
print df.head()
```

	Load	Temprature (F)	CDD	HDD	Day of Week
0	23.905050	4.5	0.0	60.5	Thursday
1	24.685159	2.5	0.0	62.5	Thursday
2	24.840359	2.1	0.0	62.9	Thursday
3	24.719850	2.4	0.0	62.6	Thursday
4	25.147472	1.3	0.0	63.7	Thursday

Fig. 3.28: First five rows of the preliminary training dataset

week. To change the categorical nature of the attribute to ordinal, a technique of encoding, called the one-hot encoding, has been used in this project. The One-Hot encoding encodes categorical integer features using a one-hot aka one-of-K scheme. In other words, the categorical attribute is broken into k attributes, where k is the number of responses the original categorical attribute has. Out of these new categorical attributes, only one of of k attribute can be true for a sample case. The true case is given a value of 1, and all other cases are given a value of 0.

For the seven days of the week, seven new columns were added into the final dataset, each for one day. For a given slot in the data frame, only one particular column out of the seven day columns has a value 1 while the values of all other columns are kept 0. Following this development, the regression equation changes to

$$\begin{aligned} L_{pred} = \alpha_0 + \alpha_1 X_{temperature} + \alpha_2 X_{CDD} + \alpha_3 X_{HDD} + \alpha_4 X_{Mon} + \alpha_5 X_{Tue} \\ + \alpha_6 X_{Wed} + \alpha_7 X_{Thu} + \alpha_8 X_{Fri} + \alpha_9 X_{Sat} + \alpha_1 0 X_{Sun}. \end{aligned} \tag{3.23}$$

The gradient boosting regression in the Scikit learn package requires fined tuned parameters. The parameters are determined using the cross-validation toolbox in python's scikit-learn package, and are given in table 3.3.

Table 3.3: Gradient Boosting parameters

Parameter	**Value**
`n_estimators`	500
`max_depth`	4
`min_samples_split`	2
`learning_rate`	0.01
loss	'ls'

Table 3.4: Gradient Boosting parameters definitions and possible values

Parameter	**Possible Values**	**Descriptions**
`n_estimators`	Integer, default = 100	Gives the number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance.
`max_depth`	Integer, default = 3	The maximum depth limits the number of nodes in the tree. Tuning this parameter gives the best performance in predictions
`min_samples_split`	Can be Integer or Float, default = 2	Gives the minimum number of samples required to split an internal node
`learning_rate`	Float, default = 0.1	The learning rate shrinks the contribution of each tree by `learning_rate`
loss	Can be any of 'ls, 'lad, 'huber, 'quantile ,default='ls'	Denotes the loss function to be optimized. 'ls refers to least squares regression, 'lad to least absolute deviation, 'huber as a combination of the two. 'quantile allows quantile regression

The parameters given in Table 3.3 are explained in Table 3.4. The values given in the columns can be understood by looking at the explanations in Table 3.4.

CHAPTER 4

PREDICTIVE OUTCOMES AND PERFORMANCE EVALUATION

This chapter presents the outcomes of predictive models that were modulated in the previous section. The equation given in 3.23 determines the predictive coefficients and the intercept constant for a particular cluster.

As given in table 3.1, there are 26 clusters, each of which is given an alphabetic code like $a,\ b,\ \cdots\ z$. Since each model is based on a separate set of data, and the weather conditions are different for different models, a separate forecasting model and equation are formulated for each cluster. Fig. 4.1 shows the hourly load profile of cluster a, for the year of 2016. Using the training data of year 2015, Fig. 4.1 shows the load plot of the predicted load for the year of 2016.

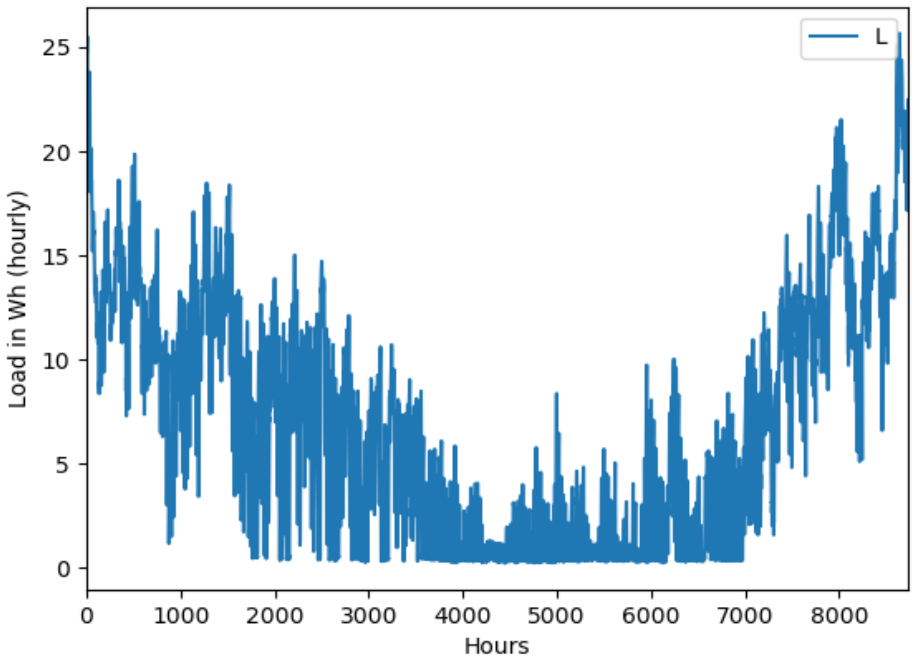

Fig. 4.1: Predicted load profile for cluster 'd' (City of Logan Utility)

Figures 4.2 and 4.4 show the line plot and bar plot representation of the actual vs predicted load profiles for cluster 'd', i.e. for the City of Logan Utility Cluster. As evident from the figure, the predicted load profile closely resembles the true load profile, especially during the winter months. Low energy consumptions for DHW and space heating/cooling systems in the summer months may lead to this kind of anomaly. Figure 4.3 is the magnified version of fig. 4.2, in which the data is sliced to $1/10^{th}$ of the actual data in fig. 4.2. The variations can be clearly seen in the sliced version, caused by load fluctuation due to hour of day.

Figures 4.5, 4.6 also shows the magnified versions of fig. 4.4, sliced to $1/10^{th}$ and $1/369^{th}$ portion of the data shown in fig. 4.4. Fig. 4.6 shows the predicted vs true load data for a span of 24 hours (from the 601th hour to 625th hour), which is for the 26th day of January. Thus, figures 4.4 4.5 and 4.6 shows the barplot of predicted load profile vs the actual load profile, sliced throughout the year, roughly a month and a day.

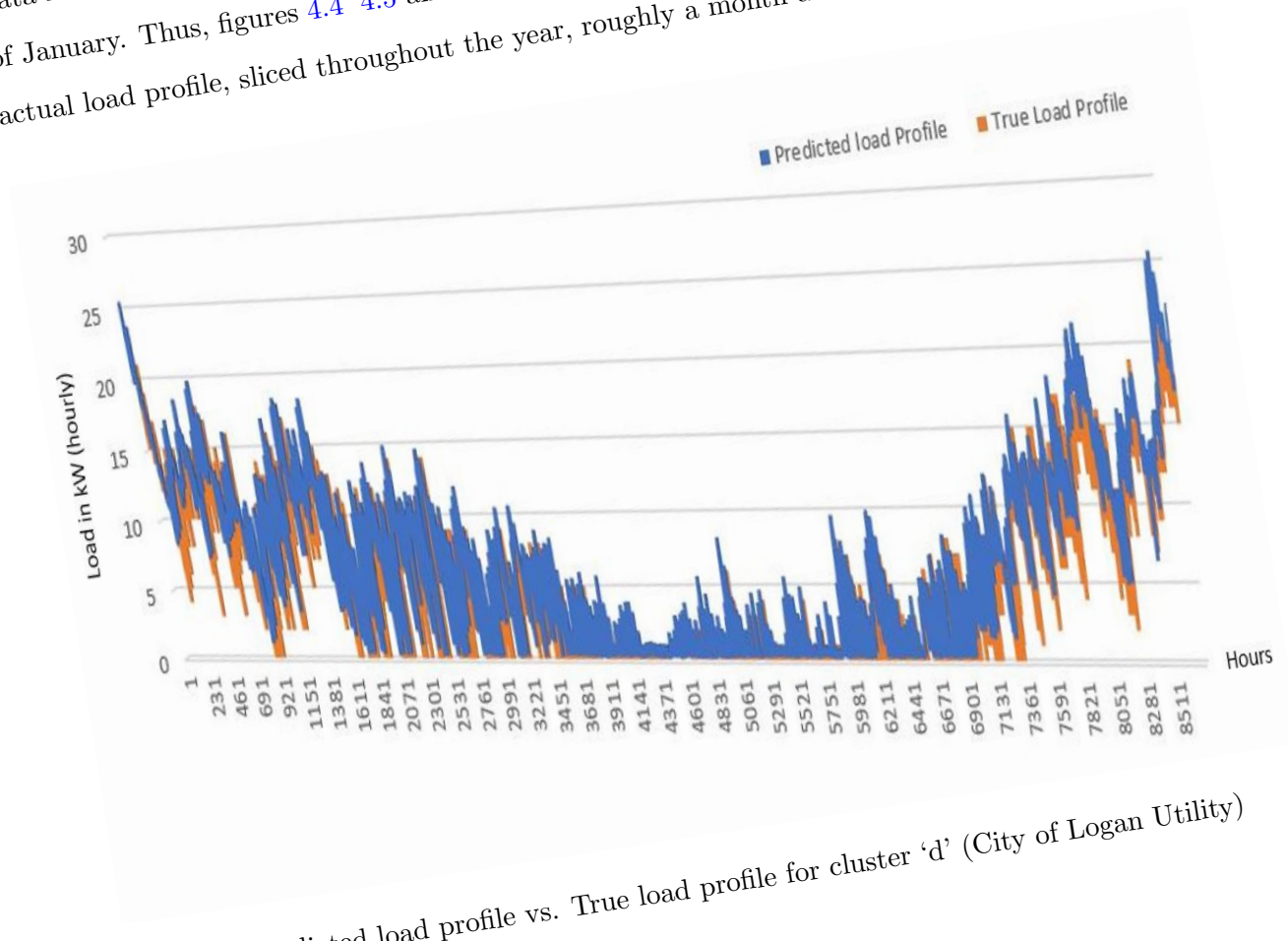

Fig. 4.2: Predicted load profile vs. True load profile for cluster 'd' (City of Logan Utility)

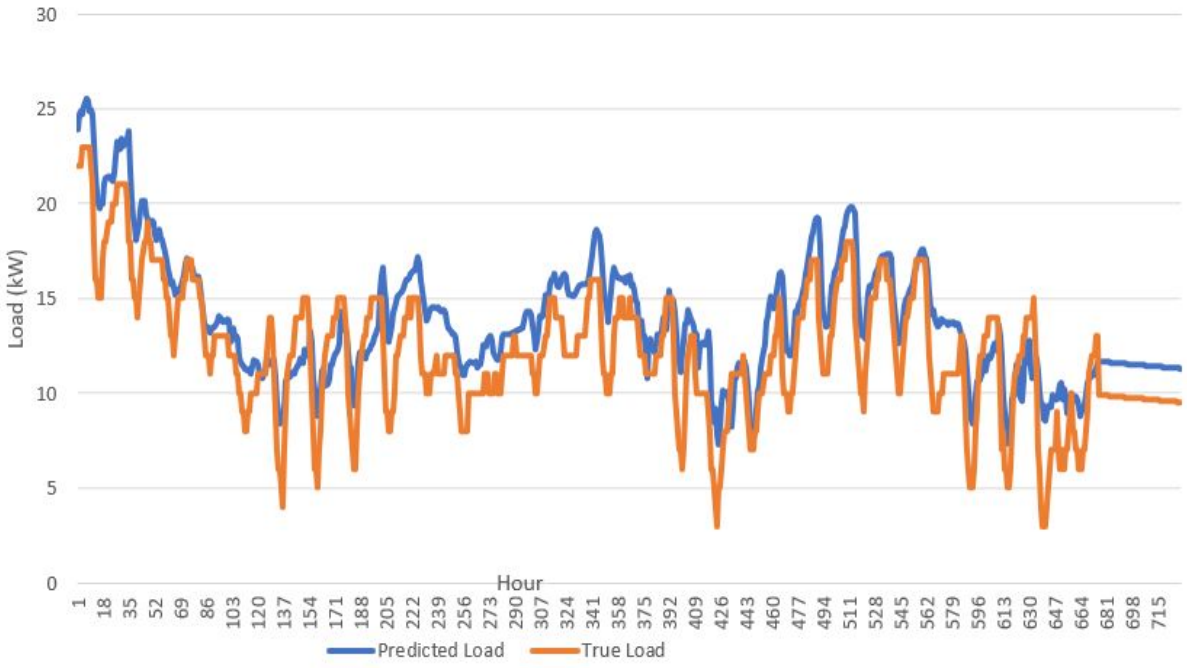

Fig. 4.3: Predicted load profile vs. True load profile for cluster 'd' (City of Logan Utility) - Sliced from 1st hour to 721st hour

To further determine the prediction accuracy, two metrics are defined, namely model accuracy score and the MMSE error of the predicted values. These two metrics are defined as follow.

4.0.1 Accuracy Score

Accuracy score gives the prediction accuracy of a model using the concept of confusion matrix. A confusion matrix is a table that describes the performance of a classification model on a set of test data for which the true values are known. There are four major terms when formulating a confusion matrix - True positive, True negative, False positive, and False negatives. True positive represents the case in which the predicted value and the actual value are same. True negative represents the case in which the predicted value and the real value compliment each other. False positive (Also known as a "Type I error.") occurs when a True is predicted but the real value is actually False. False negative (Also

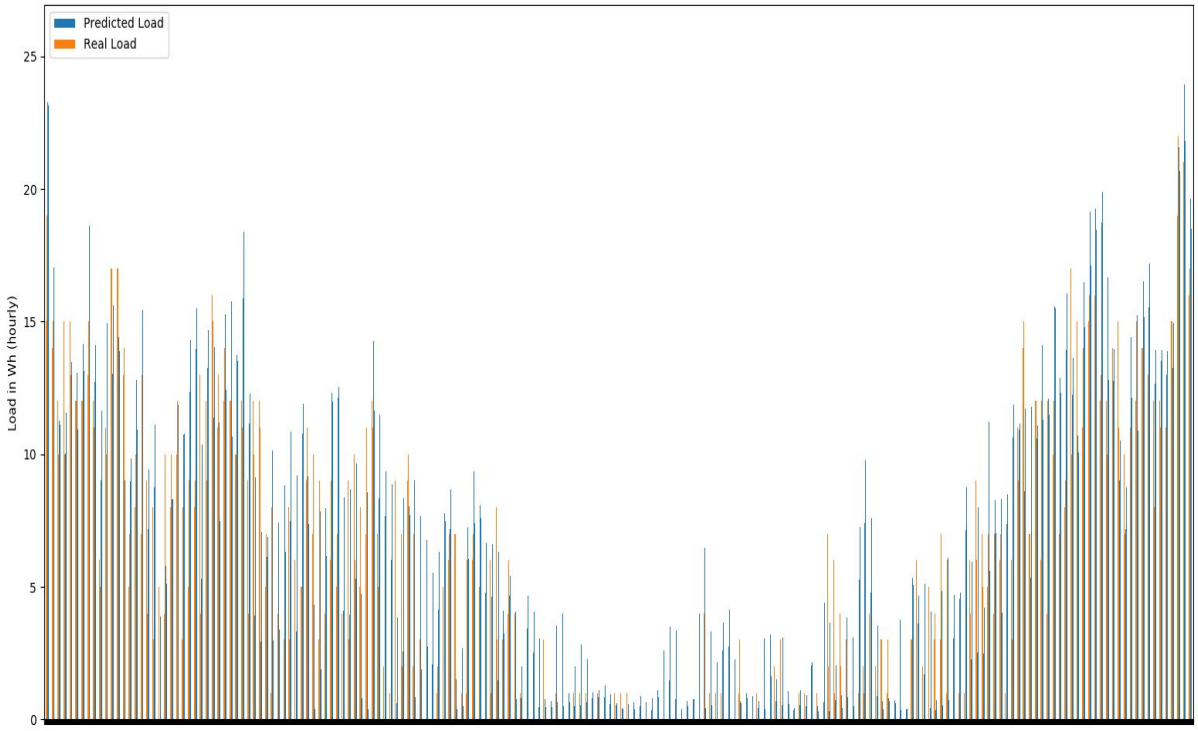

Fig. 4.4: Barplot of predicted load profile Vs. true load profile for cluster 'd' (City of Logan Utility)

known as a "Type II error.") occurs when a False is predicted but the real value is True.

Following these definitions of the confusion matrix, an accuracy score can be defined as

$$S = \frac{(TP + TN)}{Total}. \tag{4.1}$$

In other words, accuracy score is the ratio of successful predictions made with respect to the total number of events. For a perfectly designed ideal model, the accuracy score is 100% or 1.

4.0.2 Minimum Mean Squared Error (MMSE)

The mean squared error is the average of the squares of errors between the predicted value and the true value. The MSE is a measure of the quality of an estimator and is always non-negative. To understand MMSE, MSE should be formulated.

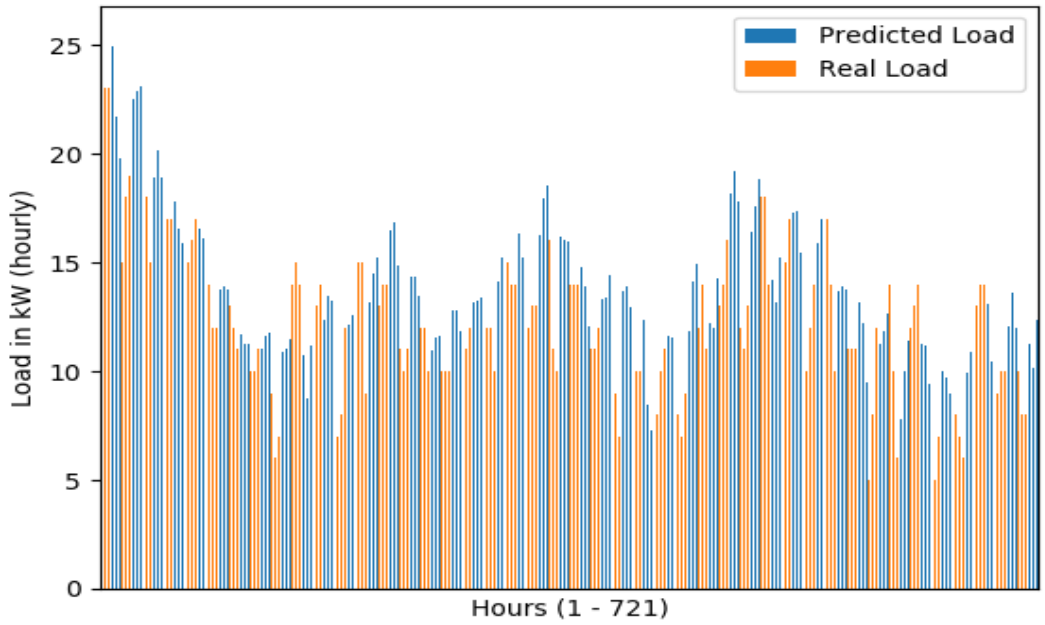

Fig. 4.5: Barplot of predicted load profile Vs. true load profile for cluster 'd' (City of Logan Utility)- Sliced from 1st hour to 721st hour

Consider Y and $\widehat{Y}$ are the real and the predicted vectors of a predictive model. By the definition [73], we can express MSE as

$$\mathbb{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2. \tag{4.2}$$

In terms of expected values, MSE can also be written as

$$\mathbb{MSE}(\hat{\theta}) = \mathbb{E}\left[(\theta - \hat{\theta})^2\right]. \tag{4.3}$$

Coming back to MMSE, we still have Y and $\widehat{Y}$ as the real vector and the predicted vector respectively. The estimation error between these vectors is $\epsilon = Y - \widehat{Y}$. Then MSE

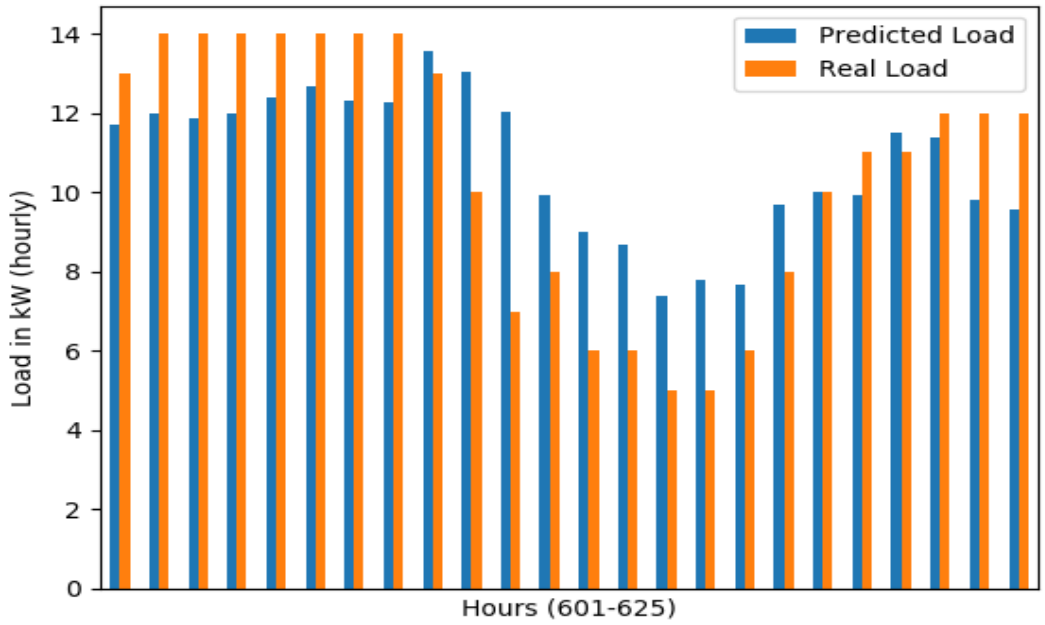

Fig. 4.6: Barplot of predicted load profile Vs. true load profile for cluster 'd' (City of Logan Utility)- Sliced from 601st hour to 625th hour

can also be expressed as the trace of the error covariance matrix, expressed as -

$$\mathbb{MSE} = tr\{\mathbb{E}[(\widehat{Y} - Y)(\widehat{Y} - Y)^T]\} \tag{4.4}$$

$$= \mathbb{E}[(\widehat{Y} - Y)^T(\widehat{Y} - Y)]. \tag{4.5}$$

Following the above equation 4.5, MMSE can be defined as the value of predicted vector that minimizes the MSE, and can be expressed as

$$\widehat{Y}_{MMSE} = arg\min_{\widehat{y}} \mathbb{MSE}. \tag{4.6}$$

The prediction accuracy and the MMSE of the predictive model used in this project are given in Table 4.1. Since the forecasted load is essentially used to plan the grid operations and to manipulate and balance peak loads, the prediction accuracy should be more than

Table 4.1: Model Score and MMSE

Parameter	Value
MLR Accuracy Score	0.962
GBR Accuracy Score	0.9714
MLR MMSE	4
GBR MMSE	2

at least 90%. If the prediction accuracy drops below 90%, the model cannot be used for load forecasting. As evident from table 4.1, the average prediction accuracy of the model developed in this project is around 96.2%, which is more than the minimum required accuracy.

Since the MMSE of MLR algorithm is traditionally poorer than other complex models, the MMSE of Gradient Boost regression model is also given in table 4.1.

4.0.3 Train - Test Split and Random Sampling

Overfitting occurs when a model learns the details and noise in the training data to the extent which negatively impacts the performance of the model on new data. When overfitting occurs, the best-fit curve follows the training data points too closely and thus takes error into its coefficients.

To check for overfitting, a random sampling approach is considered, which takes random samples from the data and compares it with the predicted values. Thus, to check if overfitting exists in the model, the time-dependent continuous data is broken into discrete samples and compared with the predicted samples. Fig 4.7 shows the comparison between the true and predicted data. For a closer inspection, fig. 4.8 shows the sliced version of fig. 4.7, sliced to 24 samples. As evident, the predicted samples are fairly comparable to the true samples, which shows that there is no overfitting of data in the prediction model.

The random splitting is achieved using the scikit-learn machine learning package in Python. The `train_test_split` function takes a data set and splits it into the training data and the test data according to the input parameters.

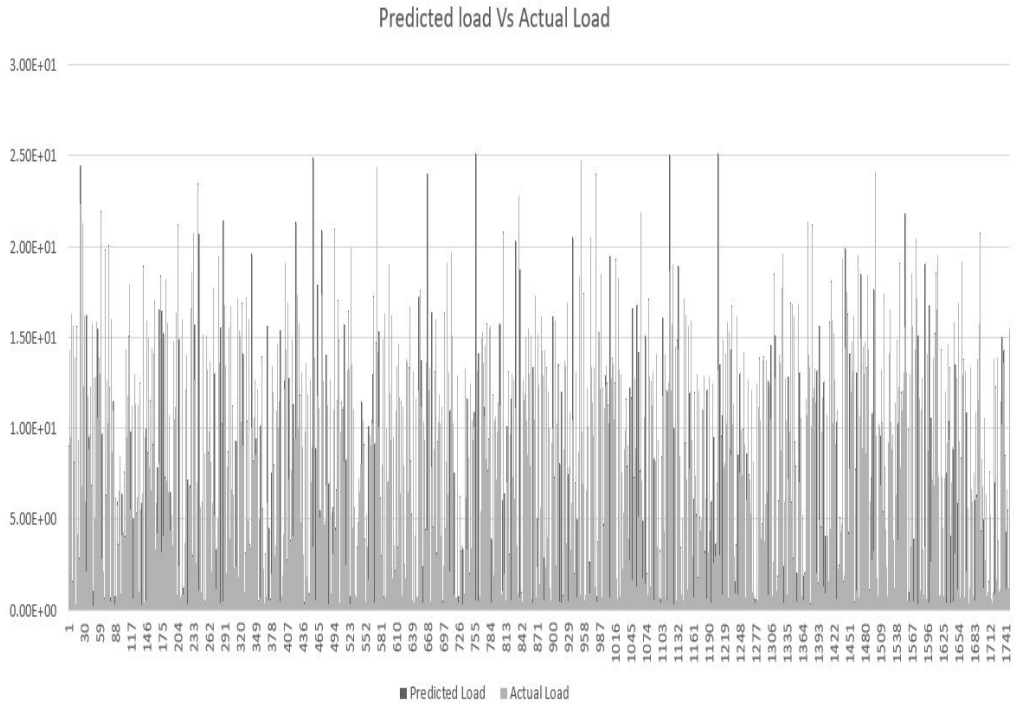

Fig. 4.7: Predicted Vs True values of random samples for cluster 'a'

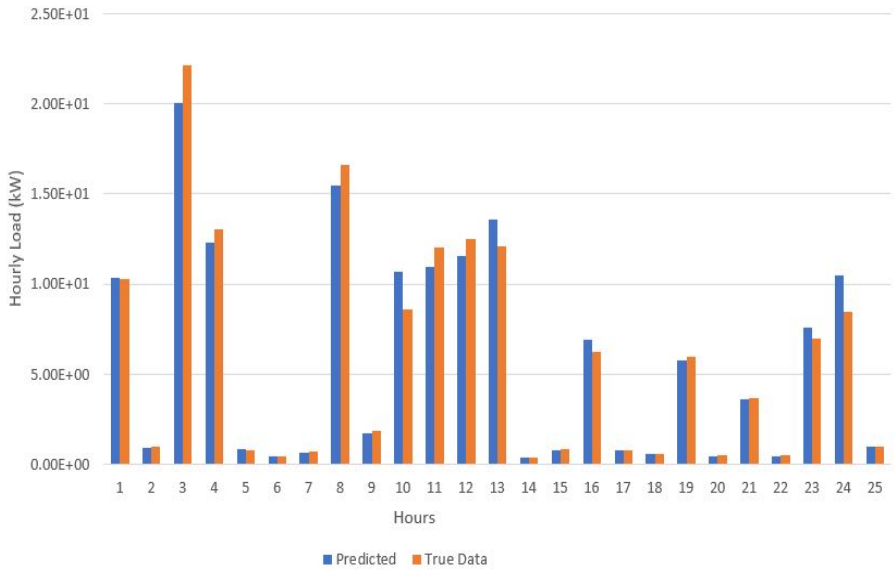

Fig. 4.8: Predicted Vs True values of random samples for cluster 'a' - sliced to 24 samples

CHAPTER 5

CONCLUSION

5.0.1 Project Summary

This project aims at developing a short to medium term load forecasting model, for the electrical grid system in the state of Utah. The forecasting of load can help the utility companies to better manage the grid operations, reduce grid failure occurrences, and operate the grid in a more economical and organized manner. On the other hand, Load forecasting can help the end use customers to estimate their net loads, normalize the load distribution and manage their utility bills by avoiding the effects of Dynamic Pricing. Thus, formulating an accurate load forecasting model would improve and assist the existing DSM techniques on both, provider's and consumer's end. As a part of this project, an hourly load profile is generated, which is the backbone of the predictive model. An accurate and error-free hourly load profile helps to effectively train the model and reduce the probability of error. For a completely digital smart grid network, the need for generating hourly load profile is minimized, thanks to the use of smart metering systems.

The project has been primarily divided into three sections - introduction, literature review, and methodology. First, an introduction on DSM is provided in Chapter 1. The main types of DSM initiatives are described along with the description of each initiative and its importance. At the same time, a literature review of DSM methods and their results are presented. In the subsequent subsections, the types of load forecasting has been discussed, along with all the techniques and algorithm employed by the load forecasting techniques.

In Chapter 3, the project methodology is discussed in detail. In the first subsection, all the required data sets are discussed briefly, and the data extraction sources are listed. In the next subsection, the data preprocessing steps are discussed in detail. Starting from clustering of data into distinct utility clusters, within the same geographical area and similar

climatic conditions to interpolation of the missing weather data using weather data of neighboring stations. The consecutive section deals with the modeling of DHW systems, and establishes mathematical equations for calculating the volume of hot water usage for different end uses. Combined with the probability of end use, the hourly consumption of hot water is translated to the hourly load of DHW systems.

After preprocessing of data, the next subsection discusses the load profile generation techniques in detail. Load profile generation combines the heating/cooling load as well as the DHW system load. Since last subsection discusses the DHW modeling techniques, the modeling of cooling/heating loads is discussed in this subsection. HDD and CDD are the two most important aspects of cooling/heating load profile generation and are given in detail in this subsection. The variation in the CDD/HDD is shown through a series of figures (3.10 to 3.25).

In the fourth subsection, the predictive modeling algorithms are discussed in detail. A mathematical perspective is presented for MLR and GBR machine learning techniques that are used as the forecasting model in this project. Additionally, the peaks and troughs in the load profile shape are explained, as evident from fig 3.27 and fig 3.26. The next subsection deals with the model equations and parameters that are used to define the model. The process of feature selection is discussed and the reasons for selecting particular attributes are explained.

In Chapter 4, the results obtained from the predictive models are briefly discussed. A comparison plot is given (fig 4.2) to compare the true load profile for cluster d (City of Logan) with the predicted load profile. The concept of train-test split is also discussed in this section and the time series data is broken into random samples using the test-train split function. These samples are then compared with their predicted counterparts, as a test for model overfitting. Finally, the techniques to determine the accuracy of the model and its error margin is discussed, followed by the actual average accuracy score and the MMSE error of the predictor models.

5.0.2 Future Prospects

This project aims to provide the short-term load forecasting data to the utilities, to assist in their DSM process. Also, the smallest entity for which a load profile is generated is a utility cluster and not an end use customer. An advanced version of the project can process real-time data from millions of smart metering systems employed in a smart grid and formulate individual load forecasting models for each residential unit. Further more, automatic load adjustment can be achieved for every individual residential unit if the load can be predicted. This would reduce the amount of computation and data storage required by the utility companies and achieve load balancing and demand side management from root level.

Further improvements can be achieved by integrating more advanced machine learning techniques link neural networks and deep learning schemes like recurrent neural network, deep neural nets etc. More efficient modeling of the space cooling and heating and DHW systems can lead to a higher prediction accuracy. The better demand handling becomes in a smart grid, the more stable and secure the grid will be.

REFERENCES

[1] L. N. Dunn, M. A. Berger, and M. D. Sohn, "Demand response forecasting methodology - berkeley lab," Available at https://openei.org/doe-opendata/.../drforecastingmethodology20160624.pptx.

[2] "Global status of modern energy access," in *International Energy Agency - World Energy Outlook*. Academic Press, 2012, pp. 23–52.

[3] V. der Hoeven and M. . Birol, "World energy outlook 2012 presentation," in *International Energy Agency - World Energy Outlook*, 2012.

[4] M. A. Faisal, Z. Aung, J. Williams, and A. Sanchez, "World energy outlook 2012 presentation," in *International Energy Agency - World Energy Outlook*, 2012.

[5] J. Lee, D.-K. Jung, Y. Kim, and Y.-W. . K. Lee.

[6] A. Liotta, D. Geelen, G. V. Kempen, and F. van Hoogstraten, "A survey on networks for smart-metering systems," in *Int. J. Pervasive Computing and Communications*. Academic Press, 2012, pp. 23–52.

[7] D. Geelen, G. V. Kempen, F. V. Hoogstraten, and A. Liotta, "A wireless mesh communication protocol for smart-metering," in *International Conference on Computing Networking and Communications ICNC*. Academic Press, 2012, pp. 343–349.

[8] C. S. Lai and L. L. Lai, "Application of big data in smart grid," in *IEEE International Conference on Systems, Man, and Cybernetics, Kowloon*. Academic Press, 2015, pp. 665–670.

[9] K. E. Parmenter, P. Hurtado, and G. Wikler, "Dynamic energy management," Oct. 2008.

[10] K. D, G. G, I. R, M. C, and M. N. et al, "Assessment of demand response and advanced metering - staff report," 2008.

[11] P. Siano, "Demand response and smart grids ? a survey, renew. sustain," 2014, pp. 461–478.

[12] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," in *Big Data Research*, 2015, pp. 94–101.

[13] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load forecasting, dynamic pricing and dsm in smart grid: A review," in *Renewable and Sustainable Energy Reviews*, 2016, pp. 1311–1322.

[14] E. H. Barakat, M. A. Qayyum, M. N. Hamed, and S. A. A. Rashed, "Short-term peak demand forecasting in fast developing utility with inherit dynamic load characteristics. i. application of classical time-series methods. ii. improved modelling of system dynamic load characteristics," *IEEE Transactions on Power Systems*, vol. 5, no. 3, pp. 813–824, Aug 1990.

[15] M. Macedo, J. Galo, L. de Almeida, and A. de C. Lima, "Demand side management using artificial neural networks in a smart grid environment," in *Renew. Sustain. Energy Rev. 41*, 2010.

[16] Z. H. T, X. F. Y, and Z. L., in *Artificial neural network for load forecasting in smart grid*, 2015, pp. 128–133.

[17] H. T, "Short term electric load forecasting," Ph.D. dissertation, North Carolina State University, 2010.

[18] N. Cetinkaya, "A new mathematical approach and heuristic methods for load forecasting in smart grid," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Aug 2016, pp. 1103–1107.

[19] F. E. R. Commission, "Assessment of demand response and advanced metering - staff report," Tech. Rep., January 2008. [Online]. Available: https://www.ferc.gov/legal/staff-reports/12-08-demand-response.pdf

[20] M. A. Abu-El-Magd and N. K. Sinha, "Short-term load demand modeling and forecasting," in *IEEE Transactions on Systems, Man and Cybernetics*, vol. 12, 1982, pp. 370–382.

[21] G. Gross and F. D. Galiana, "Short-term load forecasting," in *Proceedings of the IEEE*, vol. 75, 1987, pp. 1558–1573.

[22] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short term load forecasting: a review and evaluation," in *IEEE Transactions on Power Systems*, vol. 16, 2001, pp. 44–55.

[23] K. Metaxiotis, A. Kagiannas, D. Askounis, , and J. Psarras, "Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher," in *Energy Conversion and Management*, vol. 44, 2003, pp. 1525–1534.

[24] H. S. Hippert and C. E. Pedreira, "Estimating temperature profiles for short term load forecasting: neural networks compared to linear models," in *IEEE Proceedings - Generation, Transmission and Distribution*, vol. 151, 2004, pp. 543–547.

[25] K. Liu, S. Subbarayan, R. R. Shoults, M. T. Manry, C. Kwan, F. I. Lewis, and J. Naccarino, "Comparison of very short-term load forecasting techniques," in *IEEE Transactions on Power Systems*, vol. 11, 1996, pp. 877–882.

[26] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," in *IEEE Transactions on Power Systems*, vol. 4, 1989, pp. 1484–1491.

[27] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: An evaluation based on european data," in *IEEE Transactions on Power Systems*, vol. 22, 2007, pp. 2213– 2219.

[28] C. W. Gellings, "The concept of demand-side management for electric utilities," *Proceedings of the IEEE*, vol. 73, no. 10, pp. 1468–1470, Oct 1985.

[29] K. Liu, W. Sheng, and D. Zhang, "Big data application requirements and scenario analysis in smart distribution network," in *Proceedings of the CSEE*, vol. 35, 2015, pp. 287–293.

[30] B. Zhu, K. Xia, and X. Xia, "Game-theoretic demand-side management and closed-loop control for a class of networked smart grid," *IET Control Theory Applications*, vol. 11, no. 13, pp. 2170–2176, 2017.

[31] Z. Tu, X. Xia, and B. Zhu, "Demand-side management and control for a class of smart grids based on game theory," in *2017 36th Chinese Control Conference (CCC)*, July 2017, pp. 10 662–10 667.

[32] H. Chen, Y. Li, R. H. Y. Louie, and B. Vucetic, "Autonomous demand side management based on energy consumption scheduling and instantaneous load billing: An aggregative game approach," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1744–1754, July 2014.

[33] E. R. Stephens, D. B. Smith, and A. Mahanti, "Game theoretic model predictive control for distributed energy demand-side management," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1394–1402, May 2015.

[34] X. Liu, B. Gao, C. Wu, and Y. Tang, "Demand-side management with household plug-in electric vehicles: A bayesian game-theoretic approach," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–11, 2017.

[35] K. Ma, C. Wang, J. Yang, Z. Tian, and X. Guan, "Energy management based on demand-side pricing: A supermodular game approach," *IEEE Access*, vol. 5, pp. 18 219–18 228, 2017.

[36] F. B. Saghezchi, F. B. Saghezchi, A. Nascimento, and J. Rodriguez, "Game theory and pricing strategies for demand-side management in the smart grid," in *2014 9th International Symposium on Communication Systems, Networks Digital Sign (CSNDSP)*, July 2014, pp. 883–887.

[37] L. Raju, S. Gokulakrishnan, P. R. Muthukumar, S. Jagannathan, and A. A. Morais, "Iot based autonomous demand side management of a micro-grid using arduino and multi agent system," in *2017 International Conference on Power and Embedded Drive Control (ICPEDC)*, March 2017, pp. 44–49.

[38] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," in *IEEE Transactions on Power Systems*, vol. 4, 1989, pp. 1484–1491.

[39] K. Liu, S. Subbarayan, R. R. Shoults, M. T. Manry, C. Kwan, F. I. Lewis, and J. Naccarino, "Comparison of very short-term load forecasting techniques," in *IEEE Transactions on Power Systems*, vol. 11, 1996, pp. 877–882.

[40] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," in *IEEE Transactions on Power Systems*, vol. 5, 1990, pp. 1535–1547.

[41] T. Haida and S. Muto, ""regression based peak load forecasting using a transformation technique," in *IEEE Transactions on Power Systems*, vol. 9, 1994, pp. 1788–1794.

[42] O. Hyde and P. F. Hodnett, "An adaptable automated procedure for short-term electricity load forecasting," in *IEEE Transactions on Power Systems*, vol. 12, 1997, pp. 84–94.

[43] S. Ruzic, A. Vuckovic, and N. Nikolic, "Weather sensitive method for short term load forecasting in electric power utility of serbia," in *IEEE Transactions on Power Systems*, vol. 18, 2003, pp. 1581–1586.

[44] O. Hyde and P. F. Hodnett, "An adaptable automated procedure for short-term electricity load forecasting," in *IEEE Transactions on Power Systems*, vol. 12, 1997, pp. 84–94.

[45] S. Ruzic, A. Vuckovic, and N. Nikolic, "Weather sensitive method for short term load forecasting in electric power utility of serbia," in *IEEE Transactions on Power Systems*, vol. 18, 2003, pp. 1581 – 1586.

[46] W. Charytoniuk, M. S. Chen, and P. V. Olinda, "Nonparametric regression based short-term load forecasting," in *IEEE Transactions on Power Systems*, vol. 13, 1998, pp. 725–730.

[47] W. Charytoniuk, M. S. Chen, P. Kotas, and P. V. Olinda, "Demand forecasting in power distribution systems using nonparametric probability density estimation," in *IEEE Transactions on Power Systems*, vol. 14, 1999, pp. 1200–1206.

[48] T. Hong, "Short term electric load forecasting," Ph.D. dissertation, Raleigh, North Carolina, 2010.

[49] A. Harvey, "Forecasting, structural time series models, and the kalman filter," 1990.

[50] R. Shumway and D. Stoffer, "Time series analysis and its applications," Springer-Verlag Available at http://www.stat.pitt.edu/stoffer/conpref.pdf.

[51] B. Krogh, E. S. de Llinas, and D. Lesser, "Design and implementation of an on-line load forecasting algorithm," in *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, 1982, pp. 3284 – 3289.

[52] M. T. Hagan and S. M. Behr, "The time series approach to short term load forecasting," in *IEEE Transactions on Power Systems*, vol. 2, 1987, pp. 785–791.

[53] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," in *IEEE Transactions on Power Systems*, vol. 16, 2001, pp. 798–805.

[54] V. Shrivastava and R. B. Misra, "A novel approach of input variable selection for ann based load forecasting," in *2008 Joint International Conference on Power System Technology and IEEE Power India Conference*, Oct 2008, pp. 1–5.

[55] C. Hu and L. Cao, "Ann based load forecasting: a parallel structure," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, vol. 4, Oct 2004, pp. 3594–3598 vol.4.

[56] Y. Rui and A. A. El-Keib, "A review of ann-based short-term load forecasting models," in *Proceedings of the Twenty-Seventh Southeastern Symposium on System Theory*, Mar 1995, pp. 78–82.

[57] S. Khatoon, Ibraheem, A. K. Singh, and Priti, "Ann based electric load forecasting applied to real time data," in *2015 Annual IEEE India Conference (INDICON)*, Dec 2015, pp. 1–5.

[58] A. Selakov, S. Ili, S. Vukmirovi, F. Kuli, A. Erdeljan, Z. Gorecan, and Z. Gorean, "A comparative analysis of svm and ann based hybrid model for short term load forecasting," in *PES T D 2012*, May 2012, pp. 1–5.

[59] Y. Rui and A. A. El-Keib, "A review of ann-based short-term load forecasting models," in *Proceedings of the Twenty-Seventh Southeastern Symposium on System Theory*, Mar 1995, pp. 78–82.

[60] A. D. Papalexopoulos, H. Shangyou, and P. Tie-Mao, "An implementation of a neural network based load forecasting model for the ems," in *IEEE Transactions on Power Systems*, vol. 9, 1994, pp. 1956–1962.

[61] D. K. Ranaweera, N. F. Hubele, and A. D. Papalexopoulos, "Application of radial basis function neural network model for short-term load forecasting," in *IEE Proceedings - Generation, Transmission and Distribution*, vol. 16, 2001, pp. 798–805.

[62] S. E. Papadakis, J. B. Theocharis, S. J. Kiartzis, and A. G. Bakirtzis, "A novel approach to short-term load forecasting using fuzzy neural networks," in *IEEE Transactions on Power Systems*, vol. 13, 1998, pp. 480–492.

[63] W. K. Kong, "Demand-Side Management of Domestic Hot Water Load," Master's thesis, School of Engineering, University of Tasmania, 2014.

[64] R. Hendron and J. Burch, "Development of standardized domestic hot water event schedules for residential buildings," in *Proceedings of the ASME Energy Sustainable Conference*, 2007.

[65] Chapter49, Ed., *American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE)*. HVAC, 2003.

[66]

[67] U. D. of Energy, Ed., *Residential Energy Consumption Survey*. U.S. Department of Energy, 2001.

[68] U. Jordan and K. Vajen, *Realistic Domestic Hot-Water Profiles in Different Time Scales*. Universitt Marburg, 2001.

[69] K. P. Moustris, P. T. Nastos, A. Bartzokas, I. K. Larissi, P. T. Zacharia, and A. G. Paliatsos, "Energy consumption based on heating/cooling degree dayswithin the urban environment of athens, greece," in *Theoretical and Applied Climatology November 2014*, 2014.

[70] K. Shedden, "Multiple linear regression," Stat handout Available at http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/401-multreg.pdf.

[71] G. Ben, "Gradient boosting explained," Available at https://gormanalysis.com/gradient-boosting-explained/ (January, 2017).

[72] J. L. Mathieu, "Modeling, analysis, and control of demand response resources," Ph.D. dissertation, University of California, Berkeley, California, 2012.

[73] Wikipedia, "Minimum mean square error," Available at https://en.wikipedia.org/wiki/Minimum_mean_square_error.