Energy Analytics for Infrastructure:

An Application to Institutional Buildings

by

Hariharan Naganathan

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2017 by the
Graduate Supervisory Committee:

Oswald W. Chong, Chair
Samuel T. Ariaratnam
Kristen Parrish

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

Commercial buildings in the United States account for 19% of the total energy consumption annually. Commercial Building Energy Consumption Survey (CBECS), which serves as the benchmark for all the commercial buildings provides critical input for EnergyStar models. Smart energy management technologies, sensors, innovative demand response programs, and updated versions of certification programs elevate the opportunity to mitigate energy-related problems (blackouts and overproduction) and guides energy managers to optimize the consumption characteristics. With increasing advancements in technologies relying on the ‘Big Data,' codes and certification programs such as the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), and the Leadership in Energy and Environmental Design (LEED) evaluates during the pre-construction phase. It is mostly carried out with the assumed quantitative and qualitative values calculated from energy models such as Energy Plus and E-quest. However, the energy consumption analysis through Knowledge Discovery in Databases (KDD) is not commonly used by energy managers to perform complete implementation, causing the need for better energy analytic framework.

The dissertation utilizes Interval Data (ID), and establishes three different frameworks to identify electricity losses, predict electricity consumption and detect anomalies using data mining, deep learning, and mathematical models. The process of energy analytics integrates with the computational science and contributes to several objectives which are to

1. Develop a framework to identify both technical and non-technical losses using clustering and semi-supervised learning techniques.
2. Develop an integrated framework to predict electricity consumption using wavelet based data transformation model and deep learning algorithms.
3. Develop a framework to detect anomalies using ensemble empirical mode decomposition and isolation forest algorithms.

With a thorough research background, the first phase details on performing data analytics on the demand-supply database to determine the potential energy loss reduction potentials. Data preprocessing and electricity prediction framework in the second phase integrates mathematical models and deep learning algorithms to accurately predict consumption. The third phase employs data decomposition model and data mining techniques to detect the anomalies of institutional buildings.

I dedicate this dissertation to my loving parents Saraswathy (Kuttima) and Naganathan (Apputi). I am eternally grateful for your unconditional love, unwavering support, and continuing motivation. Without you, this would not have been possible.

ACKNOWLEDGMENTS

I would like to thank Dr. Oswald Chong who have provided me an unwavering support in all aspects for last five years. Your guidance, patience, and motivation along with financial assistance helped me to be a person who I am today. I would like to thank you for instigating my thought to pursue Ph.D., and you have guided me through every single hurdles and milestone. I am very grateful, and I appreciate your effort and the opportunity you provided me to work with you over last five years.

I would like to thank Dr. Samuel Ariaratnam, who is my committee member. You have provided me lots of support and guidance whenever I required. I would like to thank you for giving me opportunities to learn about trenchless technologies through your conferences, research projects, and seminars. I appreciate your time, effort and consideration whenever I asked you for suggestions and guidance.

I would like to thank my other committee member, Dr. Kristen Parrish for her unwavering support during last two semesters. Working with you gave me a valuable insight on approaching different research studies and helped me to improve my writing skills through your guidance. I appreciate your time and financial support you provided me to continue pursuing my Ph.D. for last two semesters.

I would like to thank Dr. Jonghoon Kim, my friend who has been unconditionally helping me throughout my career by various ways. Your industry experience, people management skill, and your positivity inspired me to cross every milestone with confidence. I wish you all success in all your endeavors and would like to collaborate and work together in opportunities that suit both of us.

I also would like to thank my friends who provided unconditional love and support whenever I needed without having a second thought. I would like to Aravind Kumar (Boss), Goutham (Gori), Bala, Shilpa, Jamuna, Vadiraj and all other friends from Kansas who motivated me to pursue Ph.D. right after my Masters. I also would like to thank Aditya, Niranjan (Katta), Prashanth (Gunda), Pradeep (Jii), Srivatsan (Hotel), and all others in Arizona who have been with me through all my ups and downs in this journey. I would like to thank my colleagues Vamsi, Nanda, and Zia for all your support and positivity for last three and half years of my Ph.D. journey.

I would like to thank Mrs. Susan Garrison for her unconditional love and support. I thank you for all your financial assistance, care, motivation and all kinds of help you provided with lots of smiles every time. I would also like to thank Dr. Wylie Bearup and Dr. Matt Eicher for their valuable guidance and insights.

Finally, I would like to thank my beloved parents (Naganathan and Sarawathy), without whom all this wouldn’t have been possible. Your support and unconditional love helped me gain this experience in the United States. You both were always praying for my goodwill, and stood by me on all my decisions and supported me in all good and bad times. I thank you both for being such a supportive parent, and I will make sure to make you proud soon.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

1. INTRODUCTION

Energy Analytics is a process of developing better-sophisticated tools for the energy managers to audit energy consumption, and manage buildings' requirements through real-time information. Smart meters that can provide every minute data on different energy units helps in visualizing the energy performance of individual buildings. The humongous database from these smart meters often contains data quality issues due to wrong manipulation, meter readings, unexpected events and unknown errors. While ASHRAE project committee is still debating on data standards on all analytical system (ASHRAE, 2017), it is important to understand the need for data treatment before implementing analytics. By this way, the comprehensive database can improve the accuracy of prediction and anomaly detection. It drives the motivation of this research to utilize different mathematical concepts to preprocess the selected consumption data before performing supply-demand loss analysis, electricity prediction and anomaly detection. Thus, the proposed energy analytic framework has outputs from the well-treated database, which can reduce the computational time and increase the efficiency of different computerized models.

This chapter details on the overview of energy consumption in the United States, various statistical and computational energy models, and recent trends in energy analytics. In addition, the chapter includes the objectives of this research and the format of the entire dissertation.

1.1 Overview

The United States consumes 18% of total primary energy consumed globally (EIA, 2017). It is an indication of the energy demand requirements in the developed countries. Commercial buildings in the United States account for 19% of the total energy consumption annually (EIA,2015). Commercial Building Energy Consumption Survey (CBECS), which serves as the benchmark for all the commercial buildings provides critical input for EnergyStar models and other codes and standards. Smart energy management technologies, sensors, innovative demand response programs from the industries, and the updated versions of certification programs elevate the opportunity to mitigate energy-related problems (blackouts and overproduction) and guides energy managers to optimize the consumption characteristics. With increasing advancements in technologies relying on ‘Big Data,' the standards and certification programs such as the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), and the Leadership in Energy and Environmental Design (LEED) evaluates the building during pre-construction phase. It is mostly carried out with assumed quantitative and qualitative values calculated from energy models such as Energy Plus and E-quest. However, the robust methods for tracking and analyzing the energy consumption through Knowledge Discovery in Databases (KDD) are not commonly used by energy managers to perform complete implementation, causing the need for better and innovative energy analytic framework. The following section indicates prior studies and research on how energy analytical models evolved over decades and presented the importance of recent “Internet of Things (IoT)” developments and their utilization.

1.2 Research Background

Commercial and residential buildings in the U.S consumes 40% of total energy utilized every year. In specific, commercial buildings consume 19% of the building energy consumption and 36% of the electricity (EIA,2015). Most of the electricity generated is from nuclear and fossil fuel sources such as coal and oil. According to EIA (2017), by 2040, commercial buildings are projected to consume the highest among all the infrastructure (transportation, residential, etc.). The rate of growth of energy demand is alarming and can contribute to adverse impacts on the environment, resource depletion and the climate change.

The Department of Energy (DOE) claims that 75% of electricity generated is for Heating, Ventilation, and Air Conditioning (HVAC) systems (Kelso, 2012). With unexpected events, occupancy variations and different building types, commercial buildings play a critical role in the electricity consumption. The research identifies potential opportunities to reduce the losses from the supply (substation totals) to the demand (building totals), to improve the accuracy of the electricity prediction through deep learning algorithms and, to detect anomalies from a preprocessed database using mathematical and data mining models.

1.2.1 Energy Management Strategies

Facility managers control their building energy demands and costs through building energy management system. According to ASHRAE 90.1, it is essential to maintain indoor environmental comfort, which meets occupancy needs. To maintain better comfort levels and to reduce the consumption of technologies like lighting, plug loads, water use, HVAC, governments make numerous efforts to drive the sustainability and promote energy efficiency strategies. Public entities such as Energy Information Administration (EIA), U.S Green Building Council (USGBC) and DOE established numerous programs and certifications such as EnergyStar and LEED to optimize and design systems like HVAC using calculated numbers from models like the EQuest. The programs and standards have improved the energy management strategies for both commercial and residential buildings. According to the EnergyStar, almost one billion office equipment is EnergyStar certified and thus, promotes energy conservation through its voluntary programs. Similarly, LEED from USGBC has updated versions of their different levels of certification with the latest program including evaluation of building energy post construction. LEED is a point-based system that offers a validation of energy efficiency through different topics from the material to HVAC systems. Assessing the performance of the building and establishing goals are the most important steps in elevating the energy conservation.

Several methods have been developed by the researchers and scholars to optimize, predict, analyze, and visualize energy demand using various mathematical, statistical and computational models. These models evolved from the concept of basic mathematics to the most complex neural networks to provide improvements to the building energy management system. The adverse effect of the increasing population and their demand pave the way to several companies and researchers to develop software and tools for energy optimization. The following section explains how energy models evolved over decades and later into analytical models in recent times.

1.2.2 Evolution of Energy Models

This section details on different energy models from previous literature. These models are utilized for various purposes combining requirements for improving energy conservations using different external and internal factors.

1.2.2.1 Energy Planning Models

In the early 70s, researchers in building energy modeling had their keen interest in developing models that help on energy planning and policy developments. Jebaraj & Iniyan (2006) did an extensive study on different energy models from 1977 and hence, published detailed information on various energy models. According to their study, energy models can be divided into six primary types that include planning, supply-demand, forecasting, optimization, neural networks and emission reduction models. Landsberg (1976) formulated the first model in which the author used the economic reasons such as inflation rate and rate of interest to understand the conversion efficiency of the solar energy. Marchetti (1977) had developed a synthetic model of primary energy substitution where literacy and resources were used as variables for analysis.

Ambrosone et. al (1983) developed a mathematical model to manage thermal energy characteristics of the building and established 95% of accuracy in his results. Similarly, Fawkes (1987) developed a checklist for the energy managers using a soft systems methodology. The energy economy model was developed with input as GDP and investments to perform long-term energy prediction for global countries (Jebaraj & Iniyan, 2006). Thus, external factors and economic variables were predominantly utilized in energy models during early developments. In addition to the change in the variables, many mathematical and statistical concepts were derived by mathematicians and researchers to handle data of different quality and type. These models included social, economic and environmental variables as their essential inputs since researchers focused on establishing a sustainable future, for elevating the standards of fellow humans.

1.2.2.2 Statistical Energy Models

The use of statistics in energy models raised the accuracy of energy forecasting, anomaly detection and consumption detailing. Various energy models are developed using different statistical and mathematical concepts for building energy demand management. While statistics still prevails to be the basics of any high-level, complex modeling, most of the models are innovated by combining with concepts such as time series, regression, econometric, decomposition, unit root test, and ARIMA (Suganthi & Samuel, 2012). These models were commonly implemented to predict consumption of the building over short, medium and long terms. For instance, time series models which are the simplest of models for analyzing energy trends are utilized by several researchers in the past decades. Grey-Model predicts coal, and electricity consumption (Kumar & Jain, 2010). Several researchers examine load forecasting using time series models (Nima Amjady, 2001; Espinoza, Joye, Belmans, & De Moor, 2005; Gross & Galiana, 1987; Hagan & Behr, 1987; Yalcinoz & Eminoglu, 2005).

In the energy industry, most of the regression models are utilized to perform load forecasting (Charytoniuk, Chen, & Van Olinda, 1998; Dudek, 2016; Papalexopoulos & Hesterberg, 1990; J. Wu, Wang, Lu, Dong, & Lu, 2013). In addition, Suganthi et al. (2012) detail decomposition models into two different approaches such as energy consumption approach and energy intensity approach. In consumption approach, specific effects impact production values, however, in intensity approach, the specific effects had no impact.

Cointegration models were integrated with multivariate models to examine the GDP and energy demand (Dincer & Dost, 1997). Similarly, ARIMA models are utilized in understanding the supply-demand characteristics, and to predict the power consumption for the future. Many cointegrated models were developed in the late 90s for better results and outputs.

The evolution of integrated models helped many researchers to integrate different techniques to improve accuracy and processing time. The advantages of these statistical techniques are that it is easily implementable and can be utilized with less computational effort. However, statistical models cannot handle large sets of data without integrating with data mining methods because of its complexity. Using advancements in data storage, the availability of different forms of data has become readily available giving importance to the computer-based data mining models. These data mining concepts possess statistics as one of their processes to clean or filter data based on the model requirements.

1.2.2.3 Computer-Based Energy Models

This dissertation "*Energy analytics for infrastructure: An application to institutional buildings"* provides extensive studies on computer-based energy prediction models, and outlier detection models through various chapters. High-speed computers, data storage capacity, inefficiencies of statistical models have paved the way for data mining and machine learning models, which can be automated and adjusted for speeding up the analysis. Several researchers adopted data mining techniques to optimize energy characteristics of building (Cappers, Goldman, & Kathan, 2010; Chicco, Napoli, & Piglione, 2006; Figueiredo, Rodrigues, Vale, & Gouveia, 2005; Nizar, Dong, & Zhao, 2006; Rodrigues, Duarte, Figueiredo, Vale, & Cordeiro, 2003; Silva & Yu, 2011). Artificial

Neural Network (ANN), Support Vector Machines (SVM), Clustering techniques, and other data mining methods have been highly utilized to improve the computational time and to provide faster and more accurate outputs. In addition, machine learning algorithms with supervised and unsupervised techniques are employed in recent research works to automate the process of data analytics and to elevate the knowledge discovery from the database (KDD).

Many recent publications suggest that some of today's artificial neural networks can be trained to recognize complex patterns. One important feature of future power grids is the ability to predict the energy consumption over a wide range of time horizons (Mocanu, Nguyen, Gibescu, & Kling, 2016a). It is important to forecast not only aggregated demand but also to go deep into the individual building so that distributed generation resources can be deployed based on the local consumption, especially due to large appliances (Mocanu, Nguyen, Gibescu, et al., 2016a).

Real-time information has become a major contributor to the advancements in data analytics(Chou, Telaga, Chong, & Jr, 2017). Most of the studies focus on the energy consumption patterns and characterization of loads for the consumers, and very few of them attempted to diagnose several losses on energy from the substation to the building itself. **Chapter 2** of this research addresses this gap in knowledge by developing a framework to identify both technical and non-technical losses using clustering algorithms and semi-supervised learning techniques. The results of this part of research identify a 15% loss reduction potential through semi-supervised machine learning technique.

A successful energy forecasting model can be combined with other building simulation models to generate useful operations. Nguyen et. al (2010) adds that electricity demand forecasting provides necessary information on elevating pricing strategies, supply-demand characteristics, and marketing to maximize their benefits. Universities make constant efforts to ameliorate sustainability and conserve energy by better demand management strategies. An accurate energy predictive model is essential to facilitate better energy demand management system. **Chapter 3** presents a novel integrated technique to predict energy utilization using data-driven methodology at a sustainably-elevated institution in Arizona. The results of this part of the research demonstrate the energy prediction of the institutional buildings at different campuses and examine the efficiency of the proposed infused Wavelet-Deep learning technique using mean absolute percentage error.

A modern building energy system is a complex dynamical system that is interconnected and influenced by external (weather) and internal (system efficiency) factors. Energy design involves connecting the intimate lifecycle relationships between energy demand and supply, and the successful connection would propel energy efficiency to the next level. Smart meters generate dynamic, diverse and large dataset and signals that have the potential to transform the management of buildings. Such data has the possibilities to detect anomalies, identify consumption patterns, determine supply-demand characteristics, and analyze peak loads. Traditional Fourier transforms based analyses are limited because, fundamentally, they are designed for linear and stationary systems.

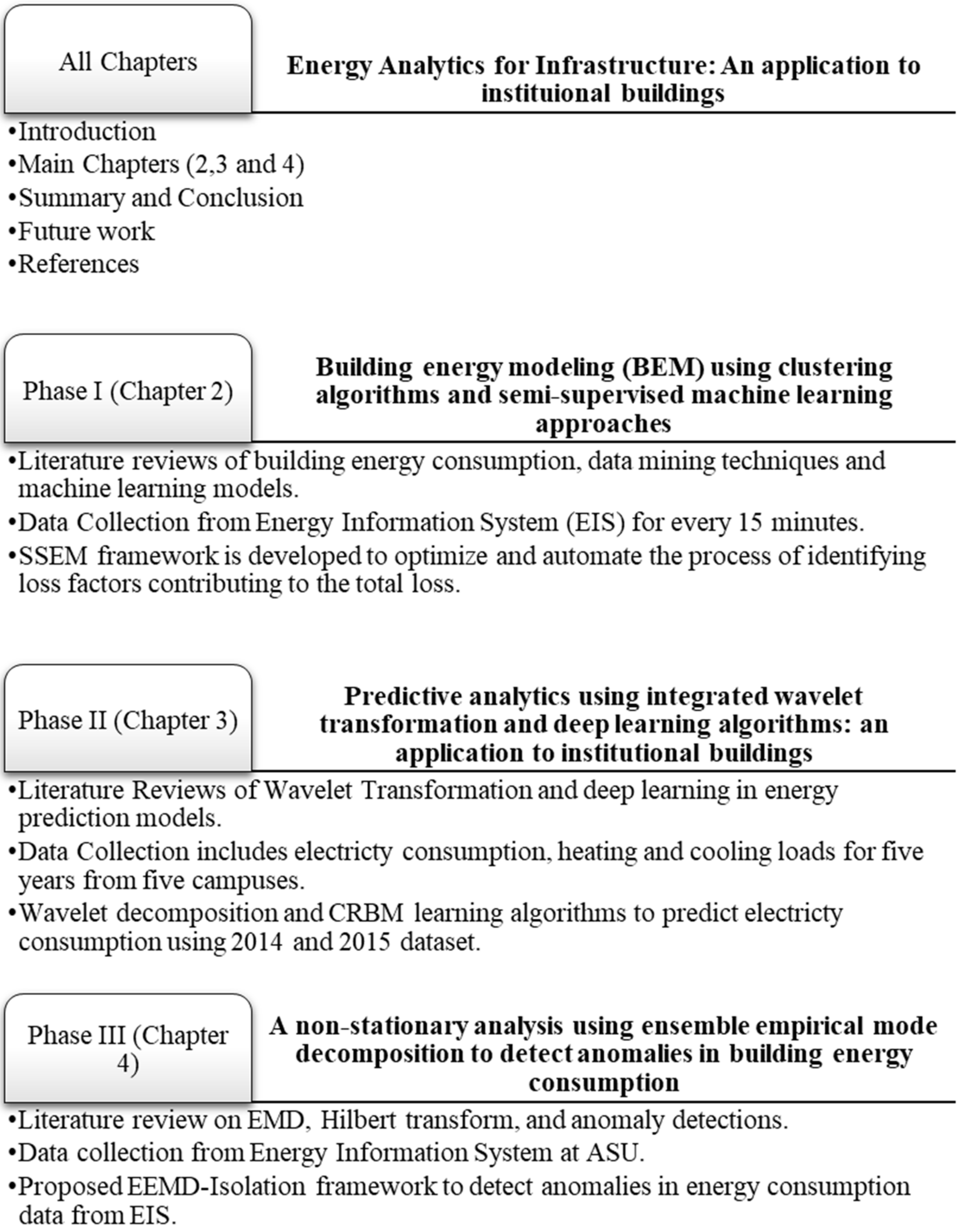
Chapter 4 of this research details on a non-stationary analysis of institutional buildings' consumption data to detect anomalies using ensemble empirical mode decomposition method and Isolation forest algorithm. The finding of this study includes the percentage of anomalous points of seven institutional buildings from 5 different campuses.

1.3 Research Objectives and Methods

The dissertation utilizes Interval Data (ID) from the Energy Information System (EIS), and establishes three different frameworks to identify energy losses, predict consumption and detect anomalies using data mining, deep learning, and mathematical models. The dissertation details three various processes of energy analytics such as electricity prediction, anomaly detection and identifying energy losses, which contributes to the development of overall energy analytic framework in the future. The process of energy analytics embeds with the data mining processes that include data preprocessing, data processing and data visualization and thus, contributes to the overall objectives which are to

1. Develop a framework to identify both technical and non-technical losses using clustering algorithms and semi-supervised learning techniques.
2. Develop an integrated framework to predict electricity consumption using wavelet based data transformation model and deep learning algorithms.
3. Develop a framework to detect anomalies using ensemble empirical mode decomposition and isolation forest algorithms from the institutional data set.

The objectives of this research disaggregate into three essential phases of this dissertation. With a thorough research background, the first phase details on performing data analytics on the demand-supply database to determine the potential energy loss reduction potentials. It involves developing semi-supervised energy model to identify and optimize both technical and non-technical losses using clustering algorithms and semi-supervised learning techniques. Data preprocessing and electricity prediction framework in the second phase integrates mathematical models and deep learning algorithms to accurately predict consumption values. The third and final phase employs data decomposition model and machine learning algorithms to detect anomalies from an extensive database of different institutional buildings. Figure 1 indicates the research flow of this dissertation.



All Chapters

Energy Analytics for Infrastructure: An application to instituional buildings

- Introduction
- Main Chapters (2,3 and 4)
- Summary and Conclusion
- Future work
- References

Phase I (Chapter 2)

Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches

- Literature reviews of building energy consumption, data mining techniques and machine learning models.
- Data Collection from Energy Information System (EIS) for every 15 minutes.
- SSEM framework is developed to optimize and automate the process of identifying loss factors contributing to the total loss.

Phase II (Chapter 3)

Predictive analytics using integrated wavelet transformation and deep learning algorithms: an application to institutional buildings

- Literature Reviews of Wavelet Transformation and deep learning in energy prediction models.
- Data Collection includes electricty consumption, heating and cooling loads for five years from five campuses.
- Wavelet decomposition and CRBM learning algorithms to predict electricty consumption using 2014 and 2015 dataset.

Phase III (Chapter 4)

A non-stationary analysis using ensemble empirical mode decomposition to detect anomalies in building energy consumption

- Literature review on EMD, Hilbert transform, and anomaly detections.
- Data collection from Energy Information System at ASU.
- Proposed EEMD-Isolation framework to detect anomalies in energy consumption data from EIS.

Figure 1. Research Study Flow Chart

1.3.1 Chapter 2: Clustering Algorithm and Semi-Supervised Learning to Identify Energy Losses

Chapter 2 details on clustering technique and deep learning algorithms implemented to determine the energy losses. A set of clustering algorithms is proposed to model the supply-demand characterization of different sub-stations clusters. SSEM framework is developed to optimize and automate the process of identifying loss factors contributing to the total loss. SSEM trains machine through a certain set of labeled and unlabeled data. The model showed a 15% loss reduction potential with the data collected.

1.3.2 Chapter 3: Predictive Analytics Using Integrated Wavelet Transformation and Deep Learning Algorithms

Chapter 3 discussed various methods used in predicting energy consumption and developed a novel WT-DL framework to predict electricity consumption. The proposed framework pre-processes the data using wavelet decomposition, and then the decomposed data is utilized to perform Conditional Restricted Boltzmann Machine (CRBM) learning predictions. The findings of the study show an accuracy more than 90% of all chosen buildings.

1.3.3 Chapter 4 Anomaly Detection using Empirical Mode Decomposition and Isolation Forest Algorithms

Chapter 4 discusses a novel EEMD-Isolation Forest framework to detect anomalies from the electricity consumption at Arizona State University. The data is preprocessed using Ensemble Empirical Mode Decomposition (EEMD), and later the anomalies are identified using Isolation forest algorithms. The objective is to develop an automated computational method to detect, characterize, and understand abnormal dynamical behaviors from big energy data sets. The steps include: (1) perform EEMD and calculate distinct IMFs, (2) and determine anomalies using isolation forest algorithms.

1.4 Dissertation Format

This dissertation is organized into three journal article formats. Each of the three subsequent chapters represents an independent journal article that has been accepted or in the review. Therefore, each section will have its own abstract, introduction, objectives, methodology, results, and conclusion. Chapter 3 and Chapter 4 are being prepared for publishing in a journal.

Chapter 1 presents the basis of the current body of knowledge related to this research study, including the research background, methodology, objectives, and scopes and format. Chapter 2 provides a model using clustering and semi-supervised learning framework that can automate the process of identifying energy losses. Chapter 3 presents the novel preprocessing and deep learning prediction algorithms for predicting electricity consumption of the buildings. Finally, Chapter 4 details on anomaly detection using Isolation forest and EEMD framework. Chapter 5 includes the research summary and conclusions based on the research study from Chapters 2 through 4 as well as the research limitations and contributions of the dissertation and future research studies. References and appendices are included at the end of this dissertation.

2. BUILDING ENERGY MODELING (BEM) USING CLUSTERING ALGORITHMS AND SEMI-SUPERVISED MACHINE LEARNING APPROACHES

Summary

The chapter details on the implementation of K-means clustering and Semi-Supervised Learning (SSL) algorithms to determine the electricity losses of 105 buildings of ASU Tempe campus. The findings of this section were published in the Elsevier *Journal of Automation in Construction.*

2.1 Abstract

Energy efficiency is a critical element of building energy conservation. Energy Information Administration (EIA) and International Electrotechnical Commission (IEC) estimated that over 6% of electrical power was lost during transmission and distribution. Sensing and tracking technologies, and data-mining offer new windows to better understanding these losses in real-time. Recent developments in energy optimization computational methods also allow engineers to characterize power consumption load profiles better. The research study focuses on developing new and robust data-mining techniques to explore vast and complex data generated by sensing and tracking technologies. These methods would potentially offer new avenues to understand and prevent energy losses during transmission. The research study presents two new concepts: First, a set of clustering algorithms that model the supply-demand characterization of four different substations clusters, and second, a semi-supervised machine learning and clustering technique are developed to optimize the losses and automate the process of identifying loss factors contributing to the total loss. This three-step process uses real-time data from buildings and the substations that supply electricity to the buildings to develop the proposed technique. The preliminary findings of this research study help the utility service providers to understand the energy supply-demand requirements.

Keywords: Building Clustering, Electricity losses, Data Mining, Semi-supervised learning, deep learning framework.

2.2 Introduction

The residential and commercial sectors in the United States consumed more than 40% of all energy produced in the country, and most of this energy is generated by different forms of fossil fuels (EIA, 2015). The improvement of building energy efficiency has not reduced the demand for energy, but the has increased with increase in renewable consumption. Several models were developed to optimize energy use through specific input parameters, and they also laid out the framework for energy consumption reduction (Dong, Cao, & Lee, 2005). Computer-based simulation models have also been used for building energy simulations. Developing energy models (e.g. DOE2 and E-quest) are often time-consuming and resource-intensive exercises, and the complexity and demand increase with the size and complexity of projects (Y. Zhu, 2006). Also, inaccurate and irrelevant input parameters and design assumptions would compromise the accuracy of the models (Y. Zhu, 2006). Simulated data may not reflect the reality better than the actual data collected during a building's lifecycle (Crawley et al., 2001). As a result, many energy models do not reflect the actual performance in reality (Ryan & Sanquist, 2012).

Electricity distribution regulations are developed to encourage competition among utility suppliers so that balance between energy efficiency and customer satisfaction can be struck (Figueiredo et al., 2005). It is a common practice for power utility companies to record customer data, billing procedures and consumption recordings in various databases to support their billing activity (Nizar et al., 2006). Liberalization of the power supply sectors allows customers the freedom of choosing their suppliers of their choices (Figueiredo et al., 2005), and creates competition among the utility providers and elevate their focus on consumers' satisfaction.

Consumers' satisfaction is translated into safe supply, peak loads and sudden change in peak load (to eliminate blackout), and overcoming gaps in energy supply between clusters. Utility suppliers strive to reduce cost by aligning their supply as close to the demand as possible without creating a condition for blackouts. A common strategy is to generate most of the demand for coal or oil and generates sudden peak loads with more expensive (but efficient) energy sources (e.g. natural gas and renewables).

Large and complex dataset collected from the consumers (e.g. their energy demand profiles) offers plenty of opportunities to align energy demand and supply (Nizar et al., 2006). Utility suppliers can now rely on many data collected by sensors and tracking devices that were previously not available, and new methods should be developed to replace the traditional energy prediction methods. One obvious issue with traditional prediction methods is the general design assumptions based on the annual energy consumption patterns. Data mining techniques had also been used to build models to identify patterns and factors, and handle large complex dataset from highly complex systems (Edwards, New, & Parker, 2012). EIA estimated that losses through transmission and distribution accounted for 6% every year (EIA, 2014) and emitted over 1000 pounds per MWh of carbon emission (ranks 33rd among all carbon emitters in the United States)

(EIA, 2014). Apart from line losses, there were also several other technical and non-technical losses which in turn contributed to financial loss (that could be used for facility investment (Nizar et al., 2006).

Semi-Supervised Energy Model (SSEM) is a real-time energy demand and supply framework that would accurately estimate the energy consumption of building clusters by predicting the energy demand and supply for every cluster through the extensive implementation of semi-supervised learning techniques (Tarca et al., 2007). SSEM trains machine (in computer science terminology, machines are referred to computers that understand the pattern) through a definite set of labeled data and integrates with reliable unlabeled data (different loss factors) to determine energy loss values. Through the learning process, the machine could predict energy loss percentage more accurately by analyzing both labeled and unlabeled factors that account for total energy loss. This model would be developed into a dynamic model that could be a significant decision-making and strategic business tool. With the large volume of labeled and unlabeled data, this research aims to develop a modeling technique to reduce the energy losses the electricity losses between the supply and demand sources.

The research study is organized into sections that include the objective of this study, a review of energy models, data mining, and machine learning techniques, methodology, and frameworks involved in this research study, clustering and semi-supervised learning modules and finally, the results and discussions to conclude the study.

2.3 Research objective

The study utilizes the Knowledge Discovery in Databases (KDD) (Figueiredo et al., 2005) procedure using data mining and deep machine learning framework. The primary objective of this study is to develop a framework to identify both technical and non-technical losses using clustering algorithms and semi-supervised learning techniques. The study utilizes data from a research university in Arizona serving approximately 80,000 students. The university has four campuses comprising of more than 400 buildings altogether. The data used in this study are the consumption data of electricity, heating, cooling and outside temperature collected from 105 buildings (at the 15-minute interval) in 2013 from one of the campuses. The large dataset requires preprocessing and clustering before analysis. The motivation of this research focuses on linking the effect of energy losses to the environment.

There are several studies on utilizing data mining techniques in energy optimization and characterization (Cappers et al., 2010; Chicco et al., 2006; Figueiredo et al., 2005; Nizar et al., 2006; Rodrigues et al., 2003; Silva & Yu, 2011). Most of the studies focus on the energy consumption patterns and characterization of loads for the consumers, and none of them attempted to diagnose several losses on energy. The major contribution of this research is on the implementation of a semi-supervised learning technique on identifying several losses through training, testing and, validating the data, and deriving the loss reduction techniques. This knowledge would elevate the opportunity for suppliers to detect both technical and non-technical losses. Also, the consumption pattern of each building connected to the respective cluster would also be modeled using this technique.

2.4 Relevant works

This section reviews the literature on, first, existing energy models and their issues, second, data mining techniques and their utilization on energy characterization and load profiling, and third, various losses on distribution and transmission and existing frameworks as well as their drawbacks.

2.4.1 Energy models and their validation

Existing energy models approximate the baseline energy use of building stocks to predict future energy demand. The accuracy of these models depends on information quality and types of input parameters (Foucquier et al., 2013; Fumo et al., 2010; Schlueter & Thesseling, 2009). The evaluation of building energy consumption requires building energy profiles on an hourly basis (Fumo et al., 2010). The building energy analysis using energy models consists of documentations, simulations, assumptions, and statistical analysis. These energy models provide a valuable projection of energy through design assumptions according to the requirements (Example: HVAC). The key factor for estimating energy efficiency in a building is to set up a minimum energy efficiency for the Heating, Ventilation, and Air-conditioning (HVAC) systems (Perez-Lombard et al., 2011). According to Perez-Lombard et al., (2011), HVAC systems account for 10-20 percent of the energy used in a building. To optimize the HVAC energy utilization, energy models would predict before and simulate the data for better understanding the consumption patterns (Vasan & Sivasubramaniam, 2015).

Zhu (2006) compared the results between the simulated and actual energy data to identify the differences and similarities between data. The process of developing simulation models with predefined templates is time-consuming and resource-demanding

(Fumo et al., 2010; Toftum et al., 2009). Data availability also plays a significant role in the success of a simulation model. All these limitations of the existing tools will eventually contribute to the disparity of gaps between the designed models and actual electricity consumption during operation (Y. Zhu, 2006). As a result, better modeling methods are needed to capture energy consumption in reality better.

Model validation is another important factor characterizing energy consumption patterns. The impact of real-time factors over the models is an important part of the validation process. Building simulation environments such as Energy Plus, TRNSYS, and ESP-r use the design assumptions when predicting building performance, and numerous studies exist to reinforce the validity of these simulation environments as well as the assumptions for predicting building heat transfer (Georgescu & Mezić, 2015). Newsham et al. (2009) investigated 100 LEED buildings and compared them with Commercial Building Energy Consumption Survey (CBECS) from 2003, a national sample survey that collects information on the U.S. commercial buildings, including their energy-related building characteristics and energy usage data (EIA, 2013). The results of their statistical analysis showed that the average LEED buildings consumed 18–39% less energy than CBECS buildings. CBECS conducts regular surveys of buildings to determine their energy consumption per square foot for different types of buildings. However, 28–35% of the LEED buildings were found to use more energy per floor than the sampled CBECS buildings (Scofield, 2009). It is because of the variations in occupancy rates, design conditions, and lack of technological improvements to handle data and consumption characteristics.

Poor commissioning is another important reason for such inefficiency (Newsham, 2009). Consequently, the design energy models compared with actual or real-time demand data on energy consumption patterns would provide different outputs. Thus, the better modeling approach is needed, which should include general, simple or localized based factors that elevate the accuracy of analysis and provides better estimates of energy consumption (relying on demographic considerations) (Pérez-Lombard et al., 2011).

2.4.2 Data mining and building energy simulations

Energy consumption analysis is a primary research area in power systems planning and management (Silva & Yu, 2011). However, the rapid increase in data availability and its dynamic growth bring new challenges to load profiling and consumer characterization for building energy modeling (Figueiredo et al., 2005). It is important to create multi-tier energy models since they will help in utilizing the real-time data and can estimate the actual time projection better than traditional energy models. Multi-tier models are cluster-based models that utilize data mining and machine learning techniques to identify consumer patterns, consumption analysis on industries and load profiling for different requirements. These methods either use data from the utility services such as electric bills, gas bills, and internet usage or collect historical data (meter data and data from energy bills) from reliable sources such as EIA, DOE, and EPA.

Azadeh & Sohrabkhani (2006) proposed the use of Artificial Neural Network (ANN) to improve consumption modeling and analysis for industries. Their study proved that ANN offers high potentials for long-term energy predictions and analysis. Pitt & Kitschen (1999) addressed the use of data mining on load profiling that provides seasonal variations of consumption for predicting future building energy performance. Support Vector Machine (SVM) is another technique used for consumption analysis although its prediction accuracy is high with only small set data (Dong et al., 2005; Li et al., 2009). Figueiredo et al. (2005) identified the electricity classification modules and characterization modules using data mining techniques. Similarly, Nizar et al. (2006) used clustering and data mining techniques to determine load profiles that help producers to prevent non-technical losses. Overall, data mining techniques and other computer-based simulations had been shown effective for improving the accuracy of determining the consumption patterns while optimizing the consumer's utilization.

Though there are numerous research on optimizing energy consumption, no prior works have been applied for driving energy supply-demand strategies. Reducing the losses of power in distribution systems is a much-needed area of research, which will lead to the enhanced management of the utility. It is of particular importance for countries facing a shortfall in energy supply (Ibrahim, 2000). Primary identification of losses and their causes and type of losses and strategies to optimize them are imperative to reduce excess power generation and divert them to more useful consumption (Ibrahim, 2000; Nagi et al., 2010). This study develops novel methods using semi-supervised machine learning to enhance supply-demand strategies.

Semi-supervised learning models have attracted increasing interest (Zhu et al., 2003). While traditional learning methods are either fully supervised or fully unsupervised (Gibson et al. ,2013), there are also many real-world situations where a small set of data are labeled, and most of the data are unlabeled. Semi-supervised learning methods exploit both labeled and unlabeled data and have shown good performance in many applications (Gibson et al., 2013; Goldberg et al., 2011; Xiaojin Zhu & Goldberg, 2009). In this study, we introduce this technique to develop a framework called the *"Semi-supervised Energy*

Modeling (SSEM)” that would identify the losses of different energy clusters and predict the loss reduction percentage using highly confident unlabeled data.

2.5 Research Methodology

The clustering technique utilized in this study has been applied to consumer load profiling and electricity characterization (Figueiredo et al., 2005; Nizar et al., 2006). The proposed framework has three essential but interdependent processes. The first phase of the Semi-Supervised Learning (SSL) framework is data collection and preprocessing. Data is collected from 105 buildings from a university campus on several factors that include electricity, heating, cooling and outside temperature. The data collected is for every fifteen minutes in 2013. The collected data is extensive and includes several factors, and thus, cleaning and preprocessing are necessary. The second phase involves clustering of data on the patterns identified by the K-means algorithms, one of the most efficient clustering techniques for this type of data and research (Chicco et al., 2006; Figueiredo et al., 2005; Nizar et al., 2006). The next phase is knowledge mining using semi-supervised learning techniques, which elevate the use of confident unlabeled data for more accurate outputs. Figure 2 describes the stepwise procedures on the methodology framework.

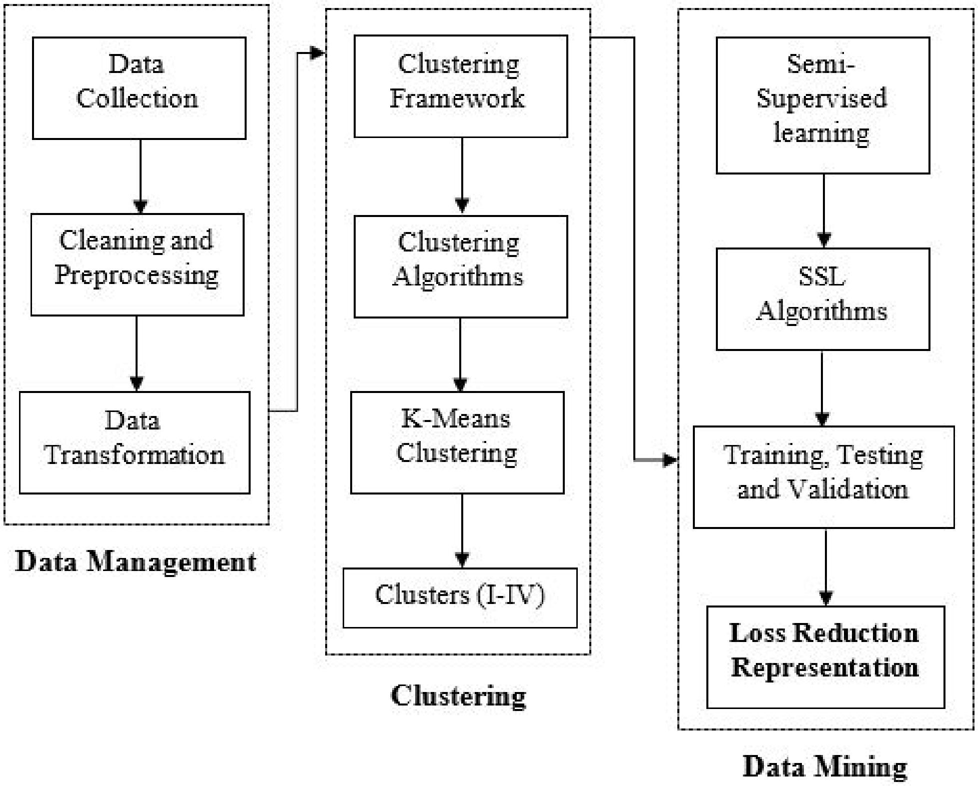


Figure 2. Methodology framework for SSEM

2.5.1 Data management

The initial database containing over five million data points is collected from Energy Information System (EIS) of the institutional buildings from one of the universities in Arizona. The data is large (by the millions) and has involved many associated factors. The first step involved the selection of data from the campus. These data are sets at 15-minute intervals. The study focuses only on the overall electricity consumption and supply from these buildings. The supply data from the substations are the total electricity supplied from all the four substations. It is then transmitted to the respective buildings connected to each substation.

2.5.2 Data Preprocessing

Analyzing data that has not been carefully screened for such problems can produce misleading results. The initial monthly data sets are collated, and the noisy samples are identified with significant variations in the consumption. These noise samples were ignored from further analysis, as they possess low confidence level. The validation is performed using the dataset that has been grouped under different substations. Thus, data preprocessing minimizes the biased data and makes data clean and complete by removing noises, neglecting wrong and biased date from the dataset.

2.5.3 Clustering framework

Discovering knowledge from data involves data partition into identifiable groups before performing any analysis. The research involves millions of data points. According to Rodrigues et al. (2003), good clustering criteria include two parameters: compactness and separation. Compactness describes how short samples within the same cluster are to each other while separation measures the distance between different clusters. K-means algorithm minimizes the mean square errors between each sample and their associated cluster center, where k refers to the number of clusters pre-specified (Rodrigues et al., 2003). The algorithms have the advantage of clear geometrical and statistical explanation and work conveniently with numerical attributes (Chicco et al., 2006).

K-means algorithms take the input parameter, *k*, and partition a set of *n* objects into k clusters so that the resulting intra-cluster similarity is high and at the same time the inter-cluster similarity is low (Han & Kamber, 2006). It has been shown that the k-means algorithms perform better than another commonly used a clustering algorithm, Kohonen Self-Organized Maps (SOM), on electrical consumer load profiles (Rodrigues et al. 2003).

Indeed, SOM performs better when the dimensionality is high (Jain et al., 1999), but K-means suits better when the dimensionality is relatively low (four clusters from four substations), which is the case for our application. In this study, we set k to be equal to four.

2.5.4 Data Mining

Deep learning is considered as one of the most reliable techniques for semi-supervised machine learning. One of the advantages of deep learning over traditional neural networks is the ability to utilize unlabeled data for unsupervised pre-training (Han & Kamber, 2006). This technique discovers the inner data structure by exploring unlabeled data and use labeled data for fine-tuning for improved discrimination power and classification accuracy.

Recent research demonstrated that unlabeled data could be used differently to improve the reliability of data analysis (Weston et al., 2012), i.e. unlabeled data and labeled data can be learned simultaneously in a semi-supervised manner. Compared to other semi-supervised approaches, which are usually based on Support Vector Machine (SVM) methods, deep learning based approaches are expected to be more reliable. Deep learning has become the new state-of-the-art technique for many difficult artificial intelligence tasks. Also, the learning process is less complicated since both labeled and unlabeled data can be learned simultaneously. Using labeled and unlabeled data, the study develops a machine learning technique to estimate energy loss between the supply and demand sources. The study also lays out a novel approach to a semi-supervised learning based on the deep learning framework. The approach carefully selects part of unlabeled data with a high confidence interval that will be integrated with the supervised learning process.

2.6 Model Validation

Semi-Supervised Energy modeling is a cybernetic approach to identify and reduce the energy losses. It uses unlabeled data with high confidence. This self-learning approach estimates the reduction of losses. Before using semi-supervised learning technique, the 15 minutes' data for the year 2013 must be clustered. Multi-tier energy demand-supply characteristics help to improve the results in energy savings and provide more reliable recommendations when clustered and classified. These classifications are based on data using different factors such as consumption, heating, cooling, and watts per square feet. The validation of this model includes three different types as follows:

1. Clustering module
2. Semi-supervised learning module
3. Result validations

2.6.1 Clustering Module

The segmentation of millions of data points and the data set is done after cleaning the data. Figure 3 explains the algorithm framework of K-means clustering technique.

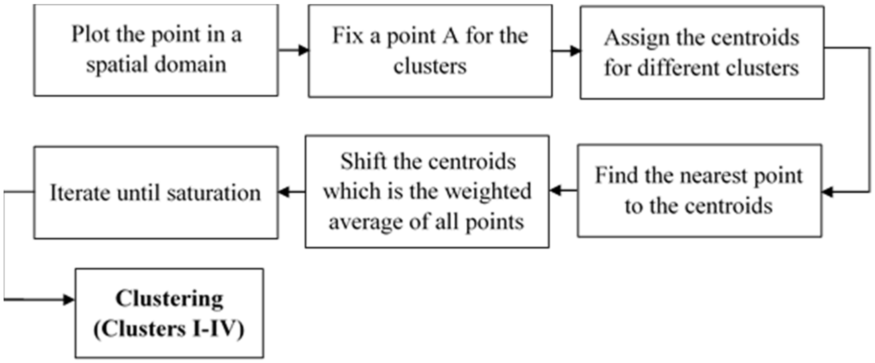

Plot the point in a spatial domain
Fix a point A for the clusters
Assign the centroids for different clusters
Find the nearest point to the centroids
Shift the centroids which is the weighted average of all points
Iterate until saturation
Clustering (Clusters I-IV)

Figure 3. Structure of K-means Algorithm

The first step for K-mean algorithm is to identify the points for all samples in a spatial domain. Second, the point is fixed, and the centroid is plotted for all clusters. Third, the nearest points to the centroids are identified, and the centroids are recalculated and shifted. This step gives the weighted averages of all points, and finally, iteration is continued until saturation. The cluster analysis is a bottom-up approach as statistical analysis is involved. Table 1 shows the various factors involved in determining the different clusters.

Table 1. Factors involving clustering framework

Factors	Units	Data Frequency	Type of Factor	Total Number of Buildings
Electricity	kWh	15 min	Equipment/System	105
Solar	kWh	15 min	Equipment/System	105
Heating load	BTU	15 min	Equipment/System	105
Outside Temperature	Fahrenheit	15 min	Environmental	105
Heat index	Fahrenheit	15 min	Environmental	105
Cooling load	Ton hours	15 min	Equipment/System	105
Watts/ sqft	Watts	15 min	Equipment/System	105

Using the K-means clustering algorithms, data are plotted to visualize the clusters of buildings as shown in figure 3. The value of k is assumed as 4 to perform clustering analysis since each sample (building) is segmented with one of the four different substations of the University. These substations are located on the North, South, West, and Central ends of the University. Figure 4 represents the clustering scatter plots. Different colors in the scatter plots represent different clusters connected to the substation with discriminant coordinates on x and y-axes, respectively.

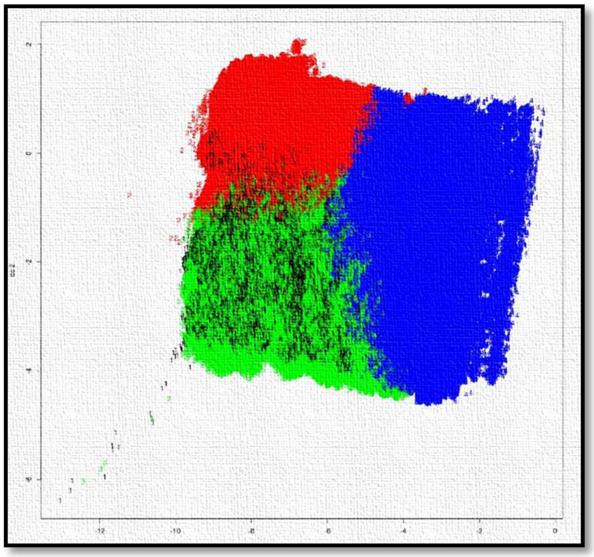


Figure 4. Different clusters of the sample size using Rattle Programming

The scatter plot describes four different clusters (Red, Blue, Green, and Black), and each cluster represents a substation. One cluster (colored black) overlaps the green and red clusters. The reason for this is that the buildings connected to this cluster do have similar consumption and supply characteristics. Both the clusters have similar building types such as classrooms and administrative offices. The clusters contain some errors, but their impact is reduced due to the large data set.

2.6.2 Semi-Supervised Learning module

Building energy models can be broadly grouped into two approaches: top-down and bottom-up. Top-down models focus on econometric data while bottom-up involves engineering and statistics (Swan & Ugursal, 2009). The proposed model is a simpler yet more efficient method using the deep learning approach where pseudo-labels of unlabeled data are calculated during every update based on current parameters (Mann & McCallum, 2007). Also, the pseudo-labels are treated as the original labels so that unlabeled data can be learned as if they are labeled data. Figure 5 explains the stepwise structure of SSL. The uniqueness of SSL is using confident unlabeled data, which is the loss factors in this research. Labeled data in this study refers to data such as electricity, time, and a total number of buildings, whereas unlabeled data relates to the different loss factors.

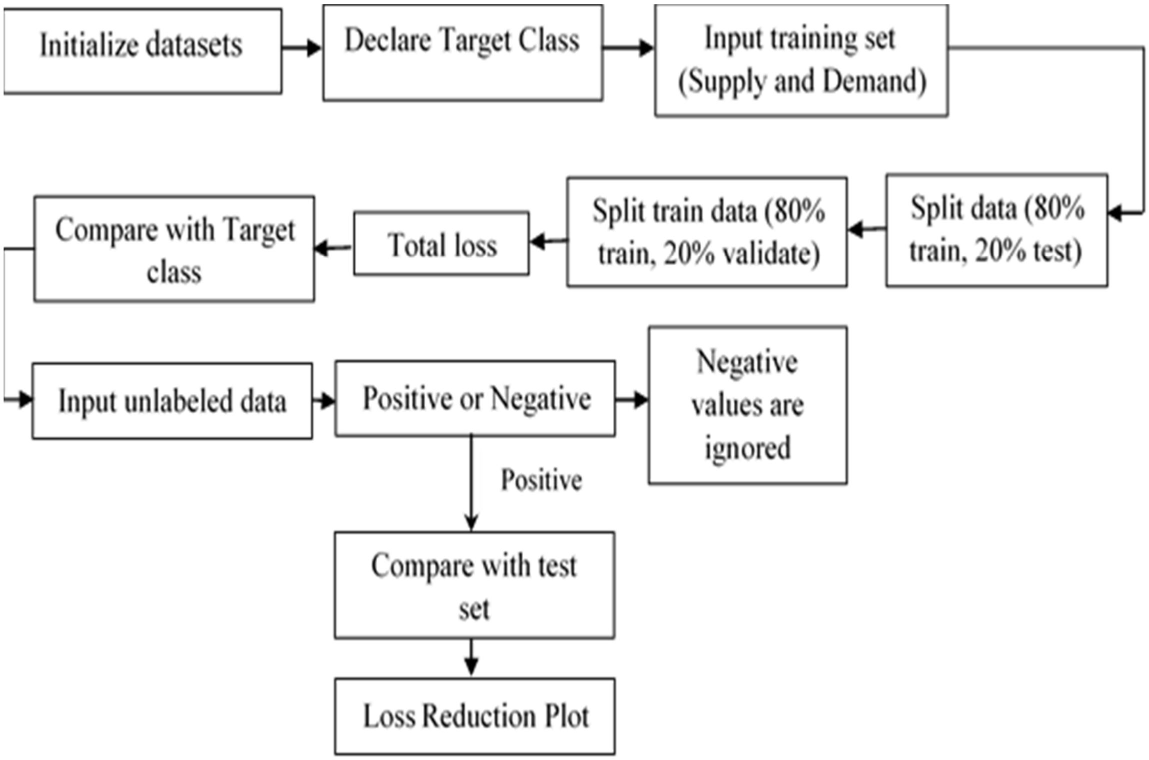

Initialize datasets
Declare Target Class
Input training set (Supply and Demand)
Split data (80% train, 20% test)
Split train data (80% train, 20% validate)
Total loss
Compare with Target class
Input unlabeled data
Positive or Negative
Negative values are ignored
Positive
Compare with test set
Loss Reduction Plot

Figure 5. Structural representation of Semi-Supervised Learning

SSL adopts a two-step approach for treating and testing data. After clustering, the data sets are initialized where the demand and supply data are fed into the machine for the four different clusters. The data fed as input are extremely reliable after the K-means algorithm treatment. Once the data is initialized, the target class and the intended output are defined. The next step is to input the training data, which is the labeled supply and demand data. The training data is split into 80 percent of training data set and 20 percent of validating data. In any data mining and machine learning research, validation plays a vital role in the testing of the accuracy of the models. After training with millions of samples, the machine will learn the pattern of supply and demand from each cluster and generate the total losses. The total losses can be compared with the target class values by validating the pattern.

The second part of the SSL is to input the treated unlabeled data. The percentage of loss factors contributes to the total loss is identified. The machine calculates the actual and desired values. The machine learned to take only the positive percentage values and compares them with the target class to determine the percentage of loss reduction and would determine through this research. The overall loss function is shown in the following equation:

$$L = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f_i\right) + \alpha(t)\frac{1}{n'}\sum_{j=1}^{n'} L\left(y_j, f_j\right)$$

where L represents loss value between demand-supply of building clusters of “n” labeled data with y_i, (the desired output vector for energy supply) for sample x_i (identity of each building cluster) and f_i (actual energy demand output). The second term represents “n’” unlabeled data with y_j being the pseudo-label for sample x_j and f_j, the actual production. The difference between the desired output and the actual output gives the loss value of the demand-supply curve.

The second term in the equation includes all loss factors contributing to the total losses. These loss factors include both technical and non-technical losses. Technical losses include losses through circuits, meters, transformers, and distribution. Each factor has a threshold value, which is the base loss percentage. The overall losses between 8% and 15 %, which suggests that there is potential to reduce CO2 emissions (Iec, 2007). The non-technical loss includes time switch errors, theft, metering and recording errors, and unmetered supplies. This study focuses only on the technical losses at this point since the data collection on non-technical losses requires further processing (and thus more time to model). Also, the actual loss percentage is compared with the standard loss value from EIA and IEC, which is on average 20 %, and this percentage is used as desired value for all the clusters.

Lee (2013) treated unlabeled data as equal for the loss function, even though, they are not. Treating them equally will result in inaccurately labeling them in the wrong pseudo-labels. In other words, those wrongly predicted unlabeled data (loss factors) might be playing a misleading role, and result in inconsistent generalization performance. The key difference (unique to this research) between the proposed method and the actual method proposed by Lee (2013) is that, instead of taking all the unlabeled data into the training process and gradually increasing the importance for each set of data, the research team selected the database on their expected confidence interval and treated them as labeled data. Since the proposed method relies on the prediction confidence, consequently, it is called confidence-based semi-supervised learning (CSL). CSL approach may have an issue with the threshold values that define the confidence level of each data point. A reasonable choice of the threshold can overcome the issue by guaranteeing that majority of the evaluation samples are correctly classified.

2.6.3 Selection of Loss Factors

To select reliable unlabeled data, the differences between the desired and the actual loss factor percentages are observed. The percentage is used to determine the confidence level of the data. Table 2 represents how the loss factors contribute to the total energy savings.

Table 2. Loss factors and its selection

Energy data and Loss factors		**Instance 1**		**Instance 2**		**Instance 3**	
Factors	Outputs	Loss (%)	Inference	Loss (%)	Inference	Loss (%)	Inference
L_1	Desired	a_1		a_2		a_3	
	Actual	b_1		b_2		b_3	
	Difference	$a_1- b_1$	Positive	$a_2- b_2$	Positive	$a_3- b_3$	Positive
L_2	Desired	p_1		p_1		p_3	
	Actual	k_1		k_1		k_3	
	Difference	p_1-k_1	Negative	p_1-k_1	Positive	p_3-k_3	Negative
L_3	Desired	z_1		z_1		z_3	
	Actual	c_1		c_1		c_3	
	Difference	z_1-c_1	Positive	z_1-c_1	Negative	z_3-c_3	Negative
	Total Loss %		**L_1+L_3**		**L_1+L_2**		**L_1**

The loss factors in Table 1, L_1, L_2, and L_3, are the desired output percentage calculated from the treated data. The actual output (the base threshold value) is selected based on specific parameters. The threshold plays another role to indicate the confidence level of the selected unlabeled data. The difference between the outputs is observed to determine whether the percentage is positive or negative. The machine is trained to understand that if the desired output is lesser than the actual output (i.e. negative), it should be omitted from the cumulative loss percentage. This process is repeated for all loss factors, and the total loss percentage is calculated only from the positive values.

The cumulative total loss percentage is then compared with the output from the first term (in Equation 1) to determine the amount of potential reduction in total losses. Users of the model should consider the timing of the unlabeled data that are included in the model (referred to as the transition point).

2.7 Results Validation

The clusters are connected to each substation and thus, have its supply and demand data on the substations. When analyzed, the central cluster showed vague results of an average of more than 60% losses. The reason that the research found is that the central plant by itself is building, and the consumption of this building is not included in the data set. Hence, to improve the accuracy of the model, the central cluster should be taken out from the analysis. Compactness within the cluster shows whether clustering results are better and has more bonding among the factors.

It is important in clustering to have factors or features that are individually and independently distributed (IID) without creating greater impacts or dependency on the other factors within a cluster (Jain et al., 1999). Figure 6 demonstrates the correlation graph with different circle sizes that indicate that the factors of the cluster are independent except the dark blue circles that are same factor correlation values. The Pearson correlation is performed to determine the relationship between the factors within a cluster. The dark blue circle from the figure indicates higher correlations whereas the faded circle with smaller sizes indicate lower correlations.

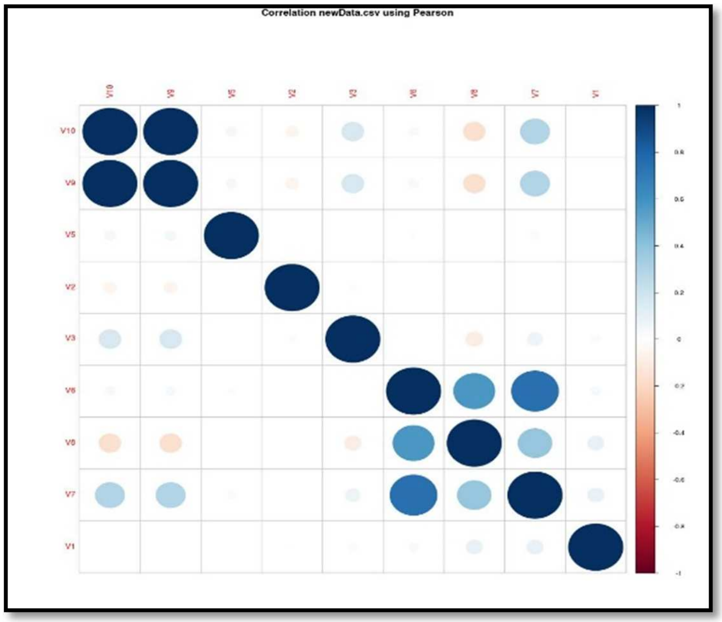


Figure 6. Correlation representation of factors within a cluster (X and Y axis indicates various factors)

The x-axis and y-axis show different factors such as building number, consumption, production, outside temperature, heat index, and watts/Sqft. From the figure 6, most of the correlations are low, or no correlation exists between other factors. Thus, clustering is validated using correlation analysis. For example, it is commonly known that the outside temperature and the day of the year are highly dependent thus; such correlation is weak in the study. However, the data set and the figure shows a moderate correlation between them, since the dataset is large and complex and is not linked which is useful to the clustering exercise. The results and plots from the first part of SSL show the total loss values of each cluster. These values from the model after the training and testing procedures demonstrate that the machine learned the pattern, and the results are comparable to the target class, and the validation showed a similar trend of machine learning. It is found that the total losses from the cluster and the target class output are very close and results nearly matched. Figure 6 shows the loss curves from all four clusters. The research team developed a website to visualize and to integrate the automated SSEM model into the site.

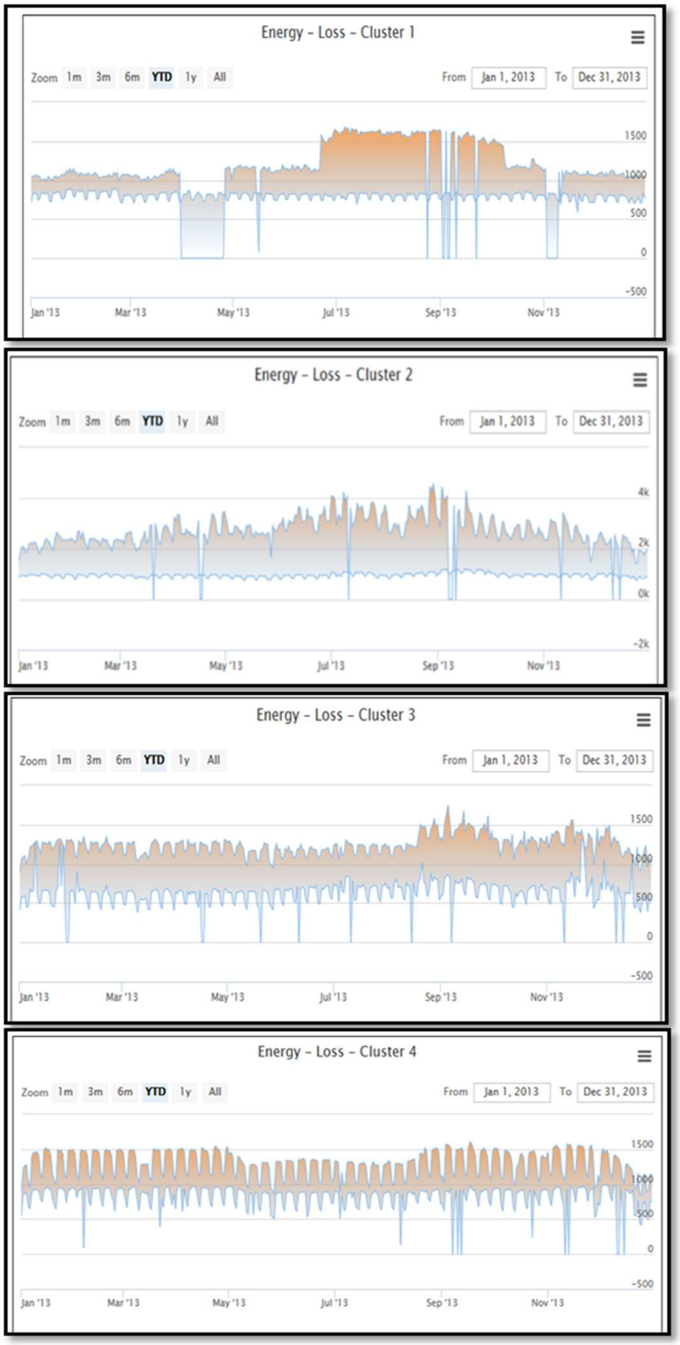


Figure 7. Total losses on each cluster for the year 2013

The data sample includes consumption data of 105 buildings on a University campus. From the figure, each of the clusters contains different loss values, and they are highly dependent on several factors such as transmission lines, technical and non-technical losses, the distance between step-down transformers and the buildings, the age of the transformer and the efficiency of the transformer, and external temperature. The energy losses estimated by the Energy Information Administration is around 10% - 15% in Arizona (EIA, 2014), whereas the data from the 105 buildings and four substations averaged nearly 30 %. An optimum value of 20% is taken as a constant and standard permissible loss on all four clusters, and the results are compared with actual loss percentage. It is done after the machine is supervised and the data validated.

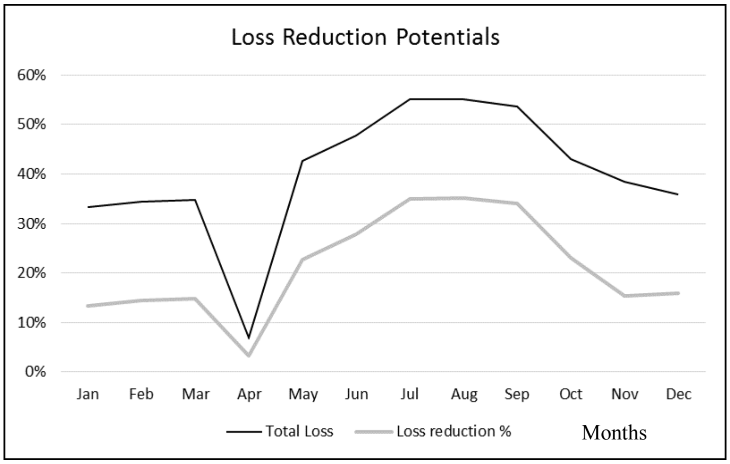


Figure 8. Total Loss versus Loss reduction potential (Cluster III)

The constant unlabeled data of loss is used to develop the model after several supervised training and validation, as these procedures will enhance the identity and clarity of the data. The unlabeled data in this study includes only the loss factors contributing to the total energy loss. This change lowers the density of overlapping at the target class boundary, thus explain how the unlabeled data reduce the loss reduction percentage.

Figure 8 showed a cluster's loss values and predicted saving values modeled using SSL. The model is initially trained with 80% of the data set and then validated with the rest of the dataset. It is done to determine if the model works efficiently. The data is integrated with the labeled total loss data to identify the potential reduction in loss percentage as shown in Figure 7 after that. The results are then compared using the losses percentages from the other studies and the EIA and IEC (Iec, 2007). Consequently, the proposed method minimizes the conditional entropy for unlabeled data to lower the density of class overlapping.

2.8 Conclusion

The study showed that SSL is a reliable technique to estimate energy efficiencies and losses on substations and buildings. In the proposed method, the first step involves preprocessing and cleaning the data using k-means algorithms, followed by integrating both labeled and unlabeled data in a semi-supervised manner to identify the loss reduction percentages. In the proposed model, semi-supervised learning can predict the classes of unlabeled data using labeled data in the first stage, and then select only the reliable sets of unlabeled data to be included in the semi-supervised learning stage. Instead of utilizing all the unlabeled data indiscriminately, the proposed method measured the confidence level of the data before using them. It helps to improve the accuracy of the output resulting from loss prevention. The proposed method also identifies the contributions of the positive loss factors toward energy savings. The proposed concept can be extended to incorporate different clusters and identified and non-identified loss factors, which will improve the reliability of the outputs. This technique helps suppliers understand the underlying reasons behind the losses, by integrating the expertise of facility managers, engineers, and architects, with the power of computing.

The study proposes a preliminary framework for the Semi-Supervised Energy Model (SSEM). Although the reliability and accuracy of this model have been demonstrated to be acceptable, more works are still needed. The effort is necessary to overcome the complexity embedded in both models and data, and extensive knowledge on machine learning and cybernetic concepts, and power generations are also required. Training the machine with algorithms is the highly complex exercise after which the machine learns the pattern and automate the model for greater accuracy. Future research should focus on, first, who should be involved in supervising the learning, second, methods to eliminate such complexities, and develop better and refined procedures and human experience (thus semi-supervised learning) into the procedures. Computers today are powerful enough to handle such complexity too.

The next step will involve another half a million data points using both labeled and unlabeled data. The research will be directed towards utilizing the model to analyze the impacts of different factors e.g. heating, cooling to understand their patterns and automate their effective strategies. It can ameliorate the energy savings and provide more insights to the decision makers on the important factors to promote sustainability through their production strategies.

3. PREDICTIVE ANALYTICS USING INTEGRATED WAVELET TRANSFORMATION AND DEEP LEARNING ALGORITHMS: AN APPLICATION TO INSTITUTIONAL BUILDINGS

Summary

The chapter details on implementing a mathematical model called Wavelet Transformation to decompose or smoothen the consumption data and later, use the decomposed data to predict electricity consumption of the selected buildings by deep learning algorithms. The findings of this section indicate an accuracy of more than 90% in most of the selected buildings. The findings of this study are being prepared for the *ASCE Journal of Energy Engineering.*

3.1 Abstract

The study presents a novel integrated technique to predict energy utilization using data-driven methodology at a sustainably-elevated institution in Arizona. An improved forecasting technique based on Wavelet Transform (WT) and Deep Learning (DL) algorithms contribute to the method of this study. The study tests the robustness of the integrated WT-DL algorithms to predict the electricity consumption of institutional buildings. The model is tested and evaluated using readily available data from Energy Information System (EIS) of the institution. The cases of 10 institutional building that possess similar utilization pattern are examined. Four campuses' buildings' data has been collected for three consecutive years on their electricity consumption. Discrete WT is used to decompose the original signal into several frequency components, and then PL algorithms are employed to provide electricity forecasting of the test case buildings. The approach provides a better demand-side management strategy and facilitates the regulatory authorities, energy managers and decision makers with a simplified yet accurate forecasting technique. Findings of the study demonstrate the energy forecasting of the buildings at different campuses and examine the efficiency of the proposed infused WT-DL technique using Mean Absolute Percentage Error (MAPE).

Keywords: Predictive Analytics, Deep learning, CRBM, Electricity Consumption, Institutional buildings, Wavelet Transformation.

3.2 Introduction

According to Energy Information Administration, buildings account for 40% of the total primary energy consumption with 18% consumed by the commercial sector (Lanzisera et al., 2013). The world energy consumption is predicted to increase by 50% in a decade if the current consumption pattern prevails (Kafaie, Kashefi, & Sharifi, 2011). With elevating economy and expanding the population, the growth rates of energy use are expected to continue further (Harish & Kumar, 2016). The improvement of building energy efficiency has not reduced the demand for energy but has increased with increase in renewable consumption (Naganathan, Chong, & Chen, 2016). Commercial building consumes the energy of more than 200 kWh per square meter of the floor size. Electricity consumption in commercial sectors accounts for 36 % of the total electricty consumption in the U.S., and it is expected to encounter an increase of 40 % from 2010 to 2030 (Kelso, 2012).

With 61 commercially operating nuclear power plants for electricity production in the United States, Arizona's Palo Verde nuclear power station is the largest net generator of electricity in the nation and is the second biggest power plant by capacity (EIA, 2016). Sustainability has changed into a major component of the university campuses because of their increased environmental impact. With different types of buildings and their utilization, college campuses are considered as a medium to large size cities (Deb, Eang, Yang, & Santamouris, 2016).

Load forecasting helps utility companies to target markets with better pricing and commitments. It is also valuable for power generators to schedule operations to match the demand (Nguyen & Nabney, 2010). Nguyen et. al (2010) adds that electricity demand forecasting provides basic information on elevating pricing strategies, supply-demand characteristics, and marketing to maximize their benefits. Electrical load characterization and predictions are performed using many mathematical and computerized models in the past few decades (Bahrami, Hooshmand, & Parastegari, 2014; Conejo, Plazas, Espínola, Member, & Molina, 2005; Eynard, Grieu, & Polit, 2011; Frimpong & Okyere, 2010; Le Cam, Daoud, & Zmeureanu, 2016; Tan, Zhang, Wang, & Xu, 2010; Tso & Yau, 2007; J. Zhang & Tan, 2013; P. Zhang & Wang, 2012; Zhao, Liu, Zhao, & Fan, 2011). Computer-based simulation models have also been used for building energy simulations (Naganathan, Chong et al., 2016). Because of the expansion of smart grid infrastructures, a lot of new prospects have come up recently, such as new data-driven methods for their flexibility and their ability to automatic fit new datasets (Cugliari, Goude, & Poggi, 2016).

A successful energy forecasting model can be combined with other building simulation models to generate useful operations. Universities make constant efforts to ameliorate sustainability and conserve energy by better demand management strategies, automating building management system, promoting the need for sustainability to the younger citizens through education. Chung and Rhee investigated on the potential opportunities for energy conservation in university buildings and proposed several strategies. An accurate energy predictive model is essential to facilitate better energy demand management system. The study presents a novel integrated technique to predict energy utilization using data-driven methodology at a sustainably-elevated institution in Arizona. An improved forecasting technique based on Wavelet Transform (WT) and Deep Learning (DL) algorithms contribute to the methodology of this study. The study tests the robustness of the integrated WT-DL algorithms to predict the electricity consumption of institutional buildings. The model is tested and evaluated using readily available data from EIS of the institution. The case study of 10 institutional building that possesses similar utilization pattern is examined. Findings of the study demonstrate the energy forecasting of the buildings at different campuses and examine the efficiency of the proposed infused WT-DL technique.

The study is organized into sections that include the objective of this study, a review of energy existing forecasting models, the relevance of wavelet transforms and deep learning concepts, research methods, WT-DL framework, predictive analytics and finally, the results and discussions to conclude the study.

3.3 Research Objective

The study utilizes the Interval Data (ID) to predict the electricity consumption of institutional campus buildings using wavelet transform and deep learning algorithms. Interval data are data that are available at a regular interval of time (ex: 1 minute, 10 minutes or 15 minutes) over several years. The primary objective of this study is to test the robustness of the integrated WT-DL algorithms to predict the consumption characteristics of institutional buildings at Arizona. The study utilizes data from Energy Information System (EIS) of Arizona State University (ASU), which has five campuses and more than 420 buildings altogether. The data used in this study are the consumption data of electricity, heating, cooling collected from 10 buildings (at the 15-minute interval) for five years from all five campuses. The extensive database over five years of 15-minute interval data requires a thorough preprocessing technique that can elevate the data quality before performing predictive analytics using deep learning algorithm.

There are several studies on preprocessing and energy prediction, and researchers have developed several modified algorithms to improve accuracy, reduce computing time and to provide the building safety managers a futuristic view on consumption characteristics (N. Amjady & Keynia, 2009; Benaouda, Murtagh, Starck, & Renaud, 2006; Dong, Oneill, Luo, & Bailey, 2014; Khoa, Phuong, Binh, & Lien, 2004; Kim, Yu, & Song, 2002; Xiaoxia Li, Zhang, & Cai, 2008; Moreno-Chaparro, Salcedo-Lagos, Rivas, & Canon, 2012; Platon, Dehkordi, & Martel, 2015; L. Tang, Yu, Wang, Li, & Wang, 2012; Tso & Yau, 2007; Wood & Newborough, 2003; Yalcinoz & Eminoglu, 2005; Yao, Song, Zhang, & Cheng, 2000; B.-L. Zhang & Dong, 2001; Zhao et al., 2011). Also, it can provide insights for the energy managers to understand the ideal consumption of buildings based on their consumption values (Deb et al., 2016).

Data management by Energy Information System (EIS) also requires better prediction models to be aware of the future energy demand and to check on the data quality through outliers of predicted values. The major contribution of this study is on integrating a framework of wavelet transform and deep learning algorithms that would preprocess the data into comprehensive dataset and provide predictions of buildings through modified Boltzmann deep learning technique. The knowledge discovery in databases performed in this study is a novel integrated technique, thus providing useful information to the energy managers with higher accuracy than other traditional computational techniques.

3.4 Relevant Work

This section reviews the literature first, on existing forecasting methodologies using different computational methods, second, wavelet analysis and their utilization in forecasting electric loads, and, third, the role of deep learning in energy forecasting and their efficiencies. The motivation of this study relies on understanding the utilization and accuracy of the novel WT-DL technique in the energy demand management. A word cloud representation of the studies identified in three different section of the relevant work is represented in figure 9 below.

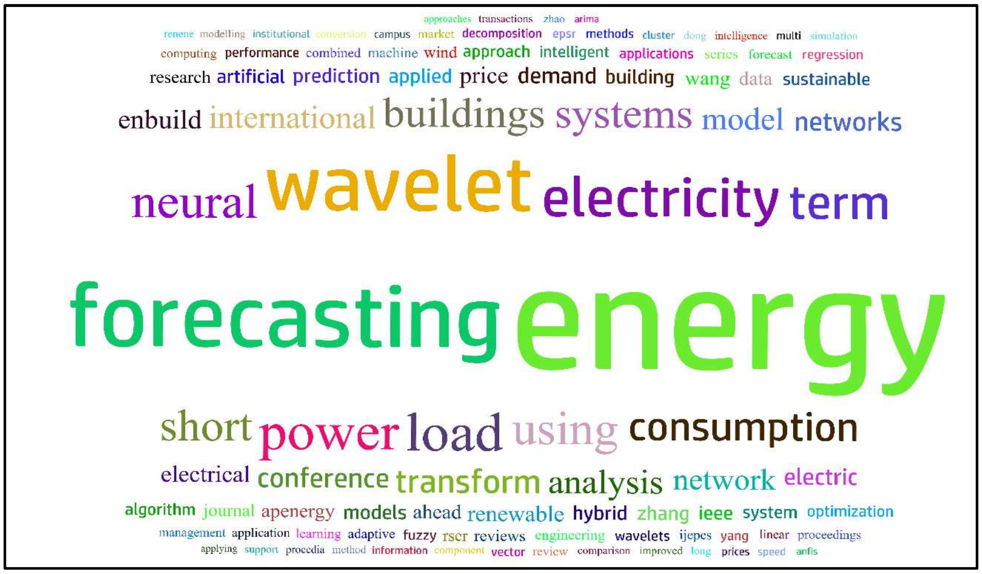

approaches transactions zhao arima
renene modelling institutional conversion campus market decomposition epsr methods cluster dong intelligence multi simulation
computing performance combined machine wind approach intelligent applications series forecast regression
research artificial prediction applied price demand building wang data sustainable
enbuild international buildings systems model networks
neural wavelet electricity term
forecasting energy
short power load using consumption
electrical conference transform analysis network electric
algorithm journal apenergy models ahead renewable hybrid zhang ieee system optimization
management application learning adaptive fuzzy rser reviews engineering wavelets ijepes yang linear proceedings
applying support procedia method information component vector review comparison improved long prices speed anfis

Figure 9. Word cloud representing prior studies of wavelet transform and energy forecasting

3.4.1 Existing forecasting Methods

Predicting electricity consumption is a challenging task since it is a complex, time series values with nonlinear dependencies and possess both periodic and random components (Rana & Koprinska, 2016). Load forecasting started as early as the 1950s (Hu, Wen, Zeng,

& Huang, 2017), and researchers developed many mathematical and statistical models (Al-Hamadi & Soliman, 2004; Norford & Leeb, 1996; Song, Baek, Hong, & Jang, 2005; Z. Wang et al., 2008). The methods include regression analysis, ARIMA, time series methods, wavelet analysis and other mathematical methods (Tan et al., 2010). Load forecasts divide into three types that include short-term (an hour to a week), medium term (week to a year), and long-term (more than a year) (Frimpong & Okyere, 2010).

Short term and medium term forecasts are done using various statistical and neural network techniques that includes regression, smoothing, Kalman filter, space modeling, pattern recognition, fuzzy logics and expert system (Al-Hamadi & Soliman, 2004; N. Amjady & Keynia, 2009; Bahrami et al., 2014; Chen et al., 2010; Chitsaz, Shaker, Zareipour, Wood, & Amjady, 2015; He, Liu, Li, Wang, & Lu, 2017; Monteiro, Ramirez-Rosado, Fernandez-Jimenez, & Conde, 2016; Nguyen & Nabney, 2010; Osório, Matias, & Catalão, 2015; Song et al., 2005; Sudheer & Suseelatha, 2015; Jie-sheng Wang & Zhu, 2015; Xu & Niimura, 2004; Yalcinoz & Eminoglu, 2005).

Similarly, long-term forecasting is developed by several researchers (Citroen & Ouassaid, 2015; Deo, Wen, & Qi, 2016; Hong, Wilson, & Xie, 2014; Khoa et al., 2004; Xiwang Li, Tan, & Rackes, 2015). Suganthi & Samuel (2012) presented a thorough research on different energy models for demand forecasting that includes time series, regression, econometrics, decomposition, cointegration, ARIMA, expert systems, gray predictions, input-output models, integrated models, and bottom-up models. According to Suganthi & Samuel (2012), each of these models has different requirements and challenges and indicates that a model that can integrate energy, economy, and environment will have real accuracy.

The neural network is used to model the energy consumption of both commercial and residential buildings (Biswas, Robinson, & Fumo, 2016; P. Zhang & Wang, 2012). Pao (2009) developed a hybrid model that includes an exponential form of autoregressive model that can predict consumption of electricity and petroleum. Yokoyama, Wakui, & Satake (2009) identified model trimming method to remove noises and periodic change in the time series data and later introduced the treated or preprocessed data into the neural network algorithm to predict the input values. In addition, the predicted input variables are validated and used for accurate prediction of energy demand.

Perception models are developed for predicting long-term energy forecasting using ANN algorithms (Suganthi & Samuel, 2012). Researchers develop several integrated and hybrid electricity forecasting models using NN. Also, ANN models are used for medium term forecasting by several investigators (Ghiassi, Zimbra, & Saidane, 2006; Xia, Wang, & McMenemy, 2010; Yalcinoz & Eminoglu, 2005). Azadeh et al. (2006) forecasted the annual energy consumption of the commercial industries using ANN and regression models and the accuracy is validated using ANOVA test. Thus, ANN has been extensively used in electricity forecasting. The study focuses on deep learning methods inspired by the structure of the artificial neural network integrating Boltzmann algorithms.

3.4.2 Wavelet-based approaches for energy prediction

Wavelet transformation and energy forecasting using different statistical and computational models have been developed by various researchers for over past two decades. With real world applications on noise suppression, fingerprint detection, seismic analysis and medical signals such as ECG, the wavelet transform is one of the most efficient methods for fault detection, data enhancements, and image recognition. In this study, the wavelet transform is utilized for enhancing the data by preprocessing using discrete wavelet transform.

Several studies explain the use of wavelet transform in short-term, medium-term and long-term electricity prediction (Benaouda & Murtagh, 2006; Benaouda et al., 2006; Catalão, Pousinho, & Mendes, 2009, 2011; Citroen & Ouassaid, 2015; Cugliari et al., 2016; Frimpong & Okyere, 2010; Kim et al., 2002; Moreno-Chaparro et al., 2012; Mourad, Bouzid, & Mohamed, 2012; Pandey, Singh, & Sinha, 2010; Sinha, Lai, Ghosh, & Ma, 2007; Vu, 2014; Jujie Wang, Wang, Li, Zhu, & Zhao, 2014; P. Zhang & Wang, 2012). In (Catalão et al., 2009), the authors have developed an NNWT framework and forecasted the electricity pricing using the test case dataset of Spain. Also, the proposed NNWT approach is compared with ARIMA, mixed-model, NN, wavelet-ARIMA, WNN, FNN, HIS, and AWNN approaches, to demonstrate its effectiveness and computation time (Catalão et al., 2009). Rana & Koprinska (2016) developed an advanced wavelet algorithm for load decomposition, Mutual Information (MI) for feature selection and Neural Networks (NNs) as prediction algorithm. The accuracy results of the research outperformed several other research and industry models (Rana & Koprinska, 2016).

Frimpong & Okyere (2010) developed a wavelet-based monthly energy consumption forecasting using radial basis function. A mean absolute percentage error of 7.94% was obtained when the model was tested over a 5-year period when the actual load was used for the forecast model (Frimpong & Okyere, 2010). Vu (2014) utilized wavelet to decompose both demand data and temperature data into low and high-frequency components and then, the Fourier transform is adopted to identify the demand patterns.

The autocorrelation shell representation based wavelet transform is used to approximate short term load at different levels of resolution (Sinha et al., 2007). Wavelet analysis has additional advantages of compressing and de-noising a signal without appreciable degradation (Frimpong & Okyere, 2010). Benaouda & Murtagh (2006) utilized Haar wavelet transform, and a nonlinear multi-resolution autoregressive method to forecast one-hour ahead electricity load of the New South Wales electricity market. The results have shown that the wavelet-based nonlinear model was performing better than the wavelet based multi resolution linear model (Benaouda & Murtagh, 2006). Also, discrete wavelet transform has also been used as a process of wind power forecasting and is integrated with HANTS method to have a deep learning of NN (Azimi, Ghofrani, & Ghayekhloo, 2016). Thus, wavelet transform has been extensively used in the preprocessing the data for greater accuracy of the electricity prediction.

3.4.3 Deep learning and Energy Predictions

It is important to note that the most widely used machine learning techniques for energy prediction are ANN and Support Vector Machines (SVM) (Fan & Hyndman, 2012). Mocanu et al. (2016a) investigates the application of Conditional Restricted Boltzmann Machines (CRBM) and Factored Conditional Restricted Boltzmann Machines (FCRBM) to understand the prediction accuracy of these latest deep learning concepts. These concepts of deep learning are considered to be the future of computational intelligence by few researchers because of its ability to resemble human brain networks better than traditional machine learning techniques (Mocanu, Nguyen, Gibescu, et al., 2016a).

Deep Learning algorithms are one promising avenue of research into the automated extraction of complex data representations (features) at high levels of abstraction (Najafabadi et al., 2015). These algorithms are widely motivated by the field of artificial intelligence, which mimics human brain's ability to observe, analyze, learn, and make decisions (Najafabadi et al., 2015). The data retrieved from deep learning possess better accuracy than other traditional machine learning algorithms which is proven by results from (Mocanu, Nguyen, Gibescu, et al., 2016a). Literature suggests that the implementation of deep learning algorithms into energy consumption analysis and prediction is limited (Mocanu, Nguyen, Gibescu, et al., 2016a; Mocanu, Nguyen, Kling, & Gibescu, 2016), and thus the motivation of this study relies on introducing a novel integrated WT-DL technique to preprocess large sets of data and to predict using comprehensive dataset.

Deep learning has become the new state-of-the-art technique for many difficult artificial intelligence tasks (Naganathan, Chong et al., 2016). This technique discovers the inner data structure by exploring unlabeled data and use labeled data for fine-tuning for improved discrimination power and classification accuracy. There is almost nil or limited prior studies that integrate wavelet transform and deep learning algorithms. Wavelet transformation has been highly utilized in preprocessing data to have better accuracy in the process of analytics. Similarly, study suggests that deep learning algorithms are the future of artificial intelligence. The study aims at integrating these two techniques that can elevate the prediction accuracy of energy consumption. Thus, a novel framework with a combined algorithm of preprocessing and predictive analytics can be utilized by the energy managers and decision makers to get more insights on building consumption characteristics and ameliorate sustainability through optimizing their production rates.

3.5 Research Methodology

The proposed WT-DL framework had three essential processes. The first phase includes data representation and collection, while the second phase details on data preprocessing. During this phase, the process of Wavelet transformation is explained, and the data preprocessing takes place. The third and the final phase includes the predictive analytics using deep learning CRBM technique that predicts the energy using the reconstructed data from wavelet transformation. Data collected from the campuses include electricity consumption, heating, and cooling loads. The time interval of the data is 15 minutes and is collected over five years at five different campuses. Figure 10 describes the stepwise procedure of the methodology.

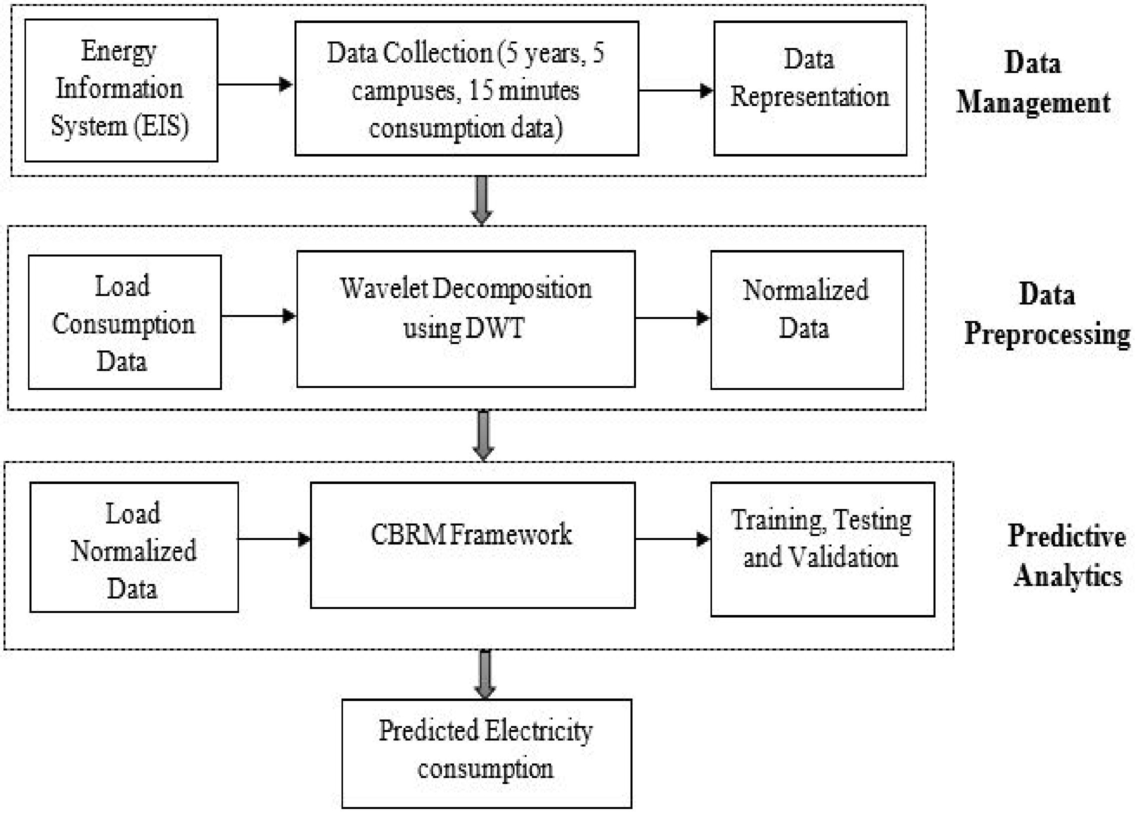
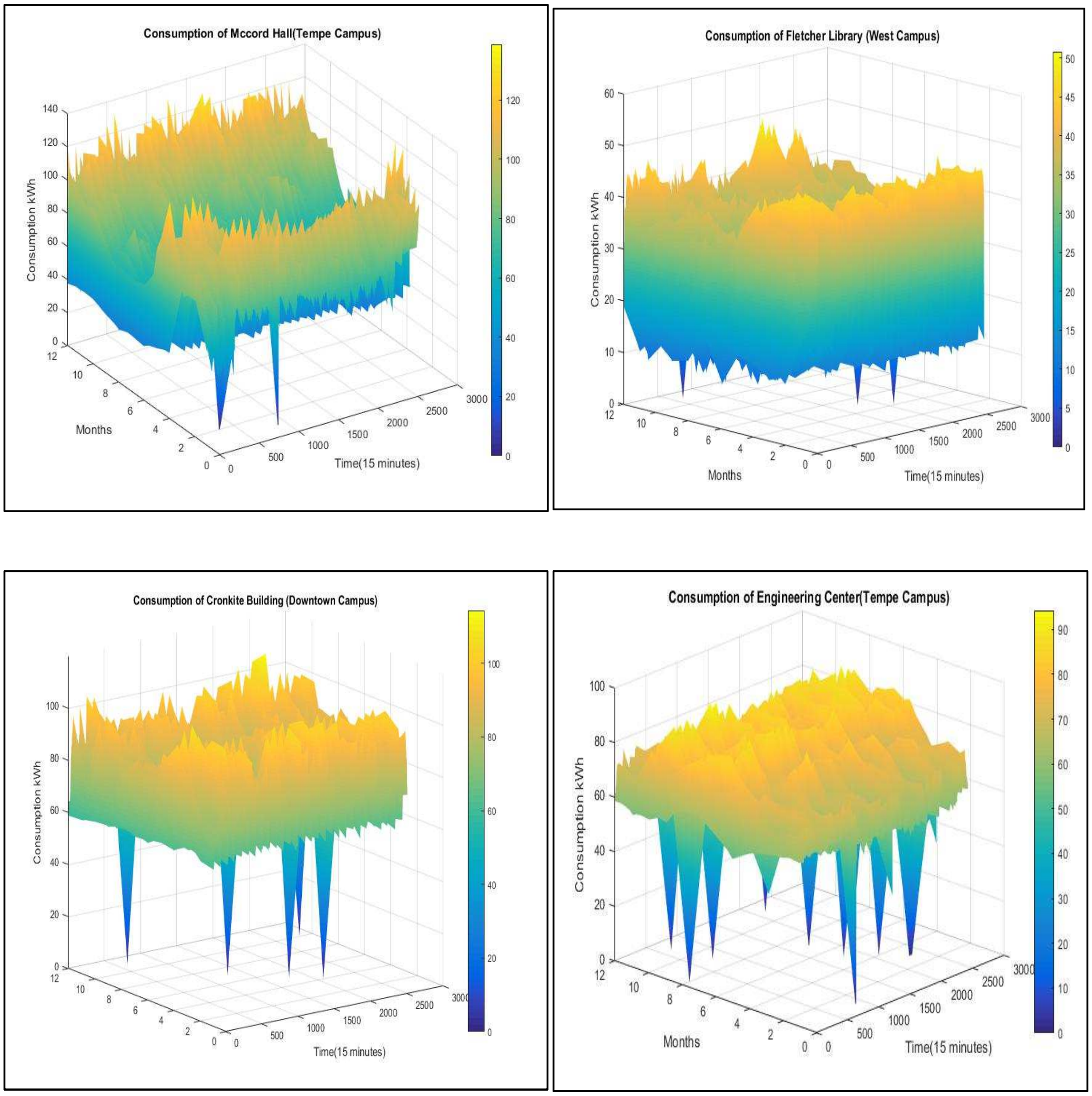

Energy Information System (EIS)
Data Collection (5 years, 5 campuses, 15 minutes consumption data)
Data Representation
Data Management
Load Consumption Data
Wavelet Decomposition using DWT
Normalized Data
Data Preprocessing
Load Normalized Data
CBRM Framework
Training, Testing and Validation
Predictive Analytics
Predicted Electricity consumption

Figure 10. Research Methodology

3.5.1 Data Management

Data collection from EIS includes 15 minutes data on electricity consumption, heating and cooling loads for all the buildings at ASU. Data is collected for last five years from all five campuses of ASU. The figure below indicates the raw energy consumption representation of one building from each campus. The figure is three dimensional with x-axis indicating time (15minutes), y-axis indication consumption and z-axis indicating years



Consumption of Mccord Hall(Tempe Campus)

Consumption of Fletcher Library (West Campus)

Consumption of Cronkite Building (Downtown Campus)

Consumption of Engineering Center(Tempe Campus)

Figure 11. Data representation of Campus Buildings' electricity consumption

Figure 11 shows that the raw data from the campus has many high and low spikes indicating data quality issues during various times of the day over five years. Hence the figure shows the need for data preprocessing and cleaning before doing predictive analytics using these institutional buildings.

3.5.2 Data Preprocessing

Analyzing data that has not been carefully screened for such problems can produce misleading results (Naganathan et al., 2016). The dataset representation from figure 3 indicates noisy samples and has significant variations over the period. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, redundant information, noisy and unreliable data (Naganathan, Seshasayee, Kim, Chong, & Chou, 2016). Thus, the preprocessing technique is required to have a more comprehensive dataset for better prediction. The following section explains wavelet transformation and how decomposition and reconstruction process helps in removing the noises from the data.

3.5.3 Wavelet Transformation

It is essential to develop a constructive algorithm that can facilitate noise filtering and load forecasting and can handle large volumes of data for analysis (Sinha et al., 2007).Wavelet analysis is highly used to reveal discontinuities and provide constitutive series that can be predicted more accurately than the original dataset(Mourad et al., 2012). The process of wavelet transformation includes two different steps. The first step is to decompose the raw data, which is passed through high pass and low pass filters. The outputs are approximate and detailed components. The summation of these two components will provide a clearer representation of the original datasets. Wavelet analysis can help decompose the actual electricity demand into different components belonging to different bands of frequency based on a mother wavelet function (Vu, 2014). During this process, the level of decomposition is selected based on the requirement and selection methods. In this study, discrete wavelet transforms of level 3 has been chosen for the decomposition. Once the data is decomposed, the components derived are introduced to deep learning CRBM framework for training and testing and predicting the decomposed components.

3.5.4 Deep learning framework

Many aspects of the modern age that include internet of things,social network and online shopping websites highly depend on machine learning technologies. Machine-learning helps in identifying images, speech to text (Siri in iPhone), and provide advertisements based on the consumer's interest (Najafabadi et al., 2015). Conventional machine-learning techniques were limited in their ability to process natural data in their raw form (LeCun, Bengio, & Hinton, 2015). Deep learning is considered as one of the most reliable techniques for semi-supervised machine learning (Naganathan, Chong, et al., 2016). Deep learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but nonlinear modules (LeCun et al., 2015)

Future power grids need a system that can monitor, predict, schedule, learn and make decisions regarding local energy consumption and production (Mocanu, Nguyen, Gibescu, et al., 2016a). In this study, a deep learning algorithm called Conditional Restricted Boltzmann Machines (CRBM) is employed, which was highly successful in modeling nonlinear time series. CRBM is a set of algorithms in machine learning that attempt to learn at multiple levels of representation, corresponding to different levels of abstraction (Mocanu, Nguyen, Gibescu, et al., 2016a).

While CRBM has been implemented in one-week predictions (Mocanu, Nguyen, Kling, et al., 2016), it has not been integrated with wavelet transform in the context of energy forecasting. While CRBM is the extension of RBM, the process includes the adding a conditional layer to improve the accuracy of the framework. A precise detail of the working process of CRBM is explained in the framework validation section.

3.6 Framework Validation

The novel WT-DL framework aims at performing predictive analytics of institutional buildings' electricity consumption, which can provide predictions of consumptions that are more accurate than other ANNs, SVMs, and RNNs. The content of the validation section is organized into three sections. The first section explains the process of wavelet transformation and how wavelet decomposition helps in reducing noises. The second section describes the formulation of CRBM algorithm and how the transformed data is utilized to predict the electricity consumption of the buildings. The last part validates the results with actual data from EIS of ASU.

3.6.1 Development of Wavelet transformation

As discussed through prior studies, the discrete wavelet transforms (DWT) decomposes the data into low-frequency components called approximate coefficients and high-frequency components called detail coefficients. The detail coefficients indicate the irregularities and fluctuations whereas the approximate coefficient indicates the regular signals. Wavelet has two different types of wavelet decomposition that include Continuous wavelet transform (CWT), which is used in theoretical research whereas DWT is utilized in engineering (Frimpong & Okyere, 2010). Also, wavelet decomposes a signal into multiple signals of differing frequencies.

In this study, ten buildings have been selected as a test case for implementing WT-DL framework. The selection of buildings was based on the availability of valuable data for all three factors which are electricity consumption (kWh), heating (mmBtu), and cooling (tonHr). Table 3 shows the buildings, their campus, square feet and their average electricity consumption, heating and cooling loads.

Table 3. Selected buildings and their characteristics

Buildings	GSF	Electricity kWh/day	Heating mmBTU	Cooling TonHr	Campus
Computing commons	132518	5598	1.75	1678	Tempe
Engineering Center A	137040	5032	7.78	2060	Tempe
McCord Hall	140092	6935	NA	2232	Tempe
Old main	45017	3195	1.90	970	Tempe
Cronkite	244964	6965	1.75	2915	Downtown
Nursing	183435	3722	NA	NA	Downtown
Academic Center	52581	1923	NA	NA	Poly
Peralta Hall	88456	2590	1.32	840	Poly
Fletcher	102856	2630	NA	1335	West
Sands	75650	1355	NA	3430	West

Table 3 shows the buildings with their average energy consumption per day. In addition, the campus location and size of these buildings are included in the table. It is notable that many buildings have less or no information on their heating and cooling loads. This lack of data prevails almost in most of the buildings at ASU. Since Arizona is under the warmer climate zone, the heating consumption at these institutions are considerably lower than other states and thus leading to data inefficiency.

DWT uses mother wavelet types such as Haar, Daubechies, and Coffman in its analysis. In this study, Daubechies db_3 is utilized to decompose the data. The raw data from different buildings are first decomposed using two level decomposition. The total of the approximate and detailed coefficients describes the original signal data more accurately. Mathematically, this can be expressed by

$$\boldsymbol{y_h\,\{k\} = \sum\nolimits_n x\,(n)\;g(2k-n)}$$

$$\boldsymbol{y_l\,\{k\} = \sum_n x(n)\;g\;(2k-n)}$$

where y_h {k} is the output of high pass filter and y_l {k} is the output of lowpass filters. Figure 12 explains the decomposition and reconstruction of a standard wavelet transform.

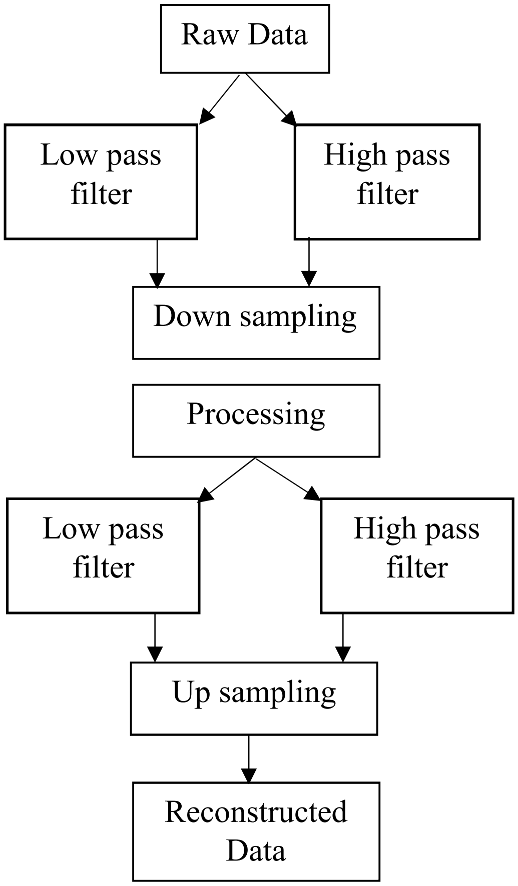


Figure 12. Process of a decomposition and reconstruction of a raw data through wavelet transform

Thus, the data is preprocessed using the wavelet transform for each building, and their respective approximate and detail coefficients are obtained. It is carried out on all the buildings that have enough information on all three factors demonstrated in Table 1. Finally, the combination of the decomposed components will be employed to determine the prediction of electricity consumption.

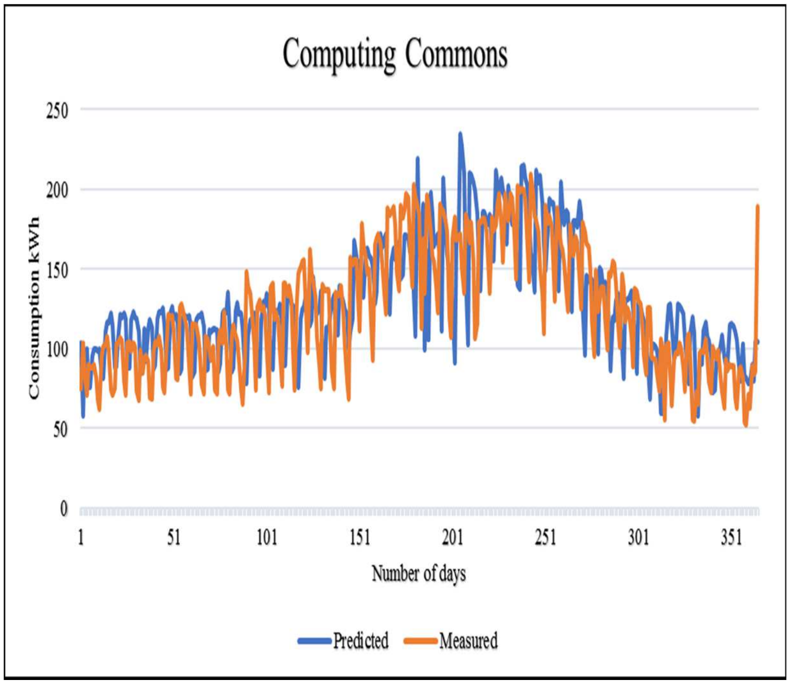
3.6.2 Development of CRBM learning algorithm

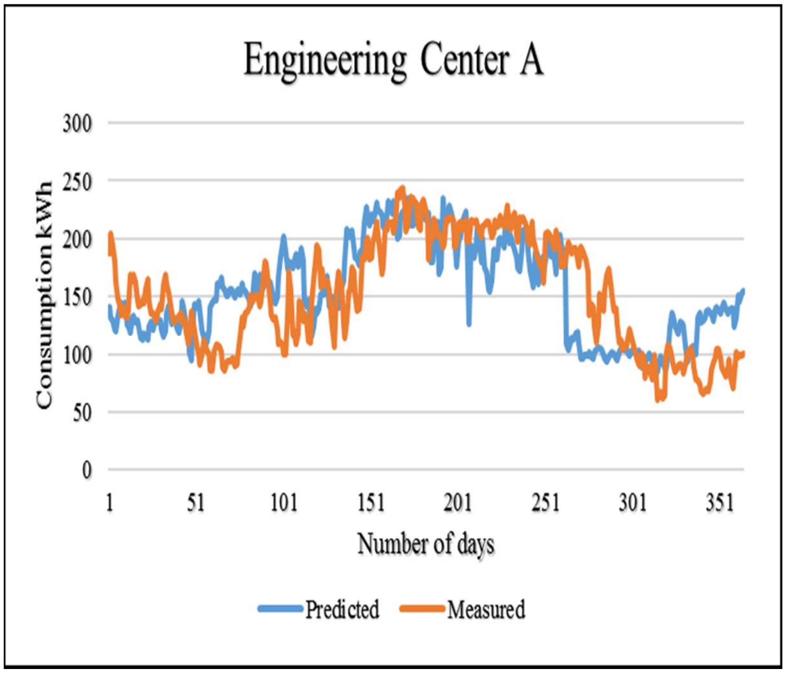
The section introduces the deep learning technique used in this study, which is CRBM. CRBM is an extension over the RBM with a conditional layer added to it for better outputs (Mocanu, Nguyen, Gibescu, et al., 2016a). The neurons in the artificial neural network recognize patterns and serve as a processing unit of the network (Hrasko, Pacheco, & Krohling, 2015). CRBMs are suitable for high dimensional time series data and are commonly called feedforward-based networks. The feedforward networks possess an input layer, a hidden layer, and an output layer. The input layer is visible and is where all the input data is obtained. The hidden layer is where the learning happens through training and testing, and finally, the output layer is what visualizes the testing results of the hidden layer.

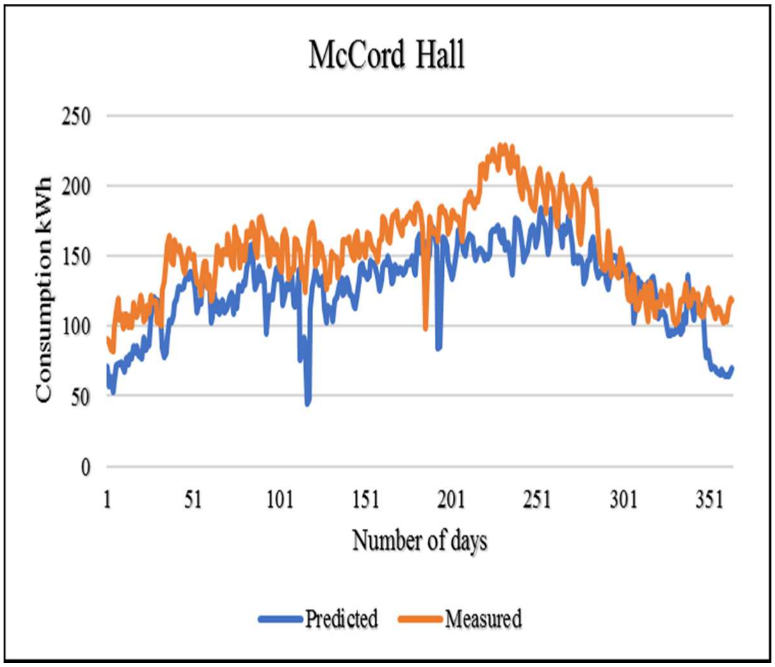
In this study, the input data is labeled, and thus the learning algorithm for this article is supervised learning. In other words, the learning algorithm for this section is determined as back propagation method. In this study, the collected data highlights the electricity consumption, heating and cooling loads of ten buildings. In all ten buildings, 2014 consumption values are used for learning (training and testing), and 2015 values are used to validate the prediction values. The means absolute error percentage is identified for all the buildings to understand the accuracy percentage error of the actual versus predicted values.

3.7 Results and Validation

The process of WT-DL includes wavelet transformation and then utilizing deep learning CRBM algorithms to predict the electricity consumption of ASU buildings. The first step, which is wavelet decomposition enhances the data quality by reducing noises and removing instabilities. After decomposition, the decomposed data utilized with Boltzmann's learning algorithm to predict the decomposed data for predicting the consumption values of 2015.Initially, the machine was trained with all 2014 data using learning algorithms. Figure 5 shows the buildings from Tempe Campus and their actual and predicted values.




Engineering Center A
300
250
200
150
100
50
0
Consumption kWh
1
51
101
151
201
251
301
351
Number of days
Predicted
Measured


McCord Hall
250
200
150
100
50
0
Consumption kWh
1
51
101
151
201
251
301
351
Number of days
Predicted
Measured

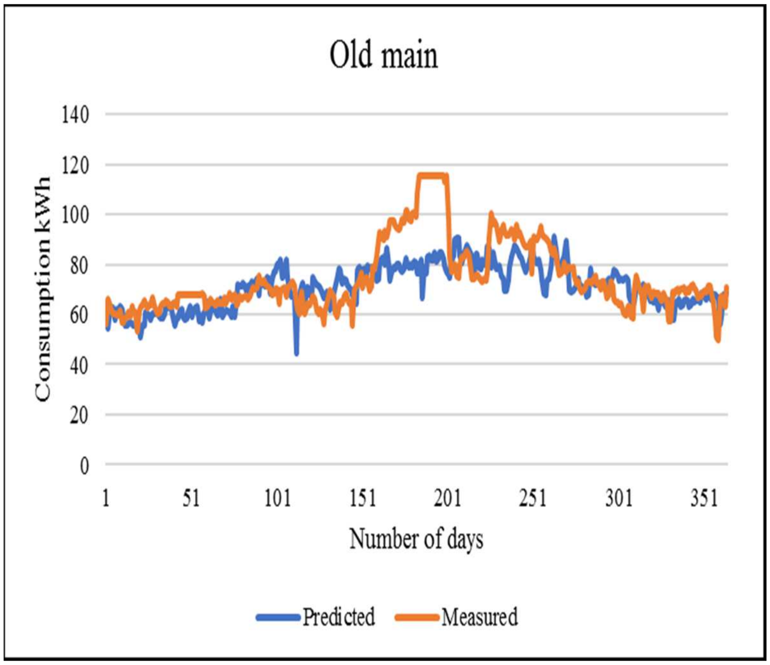
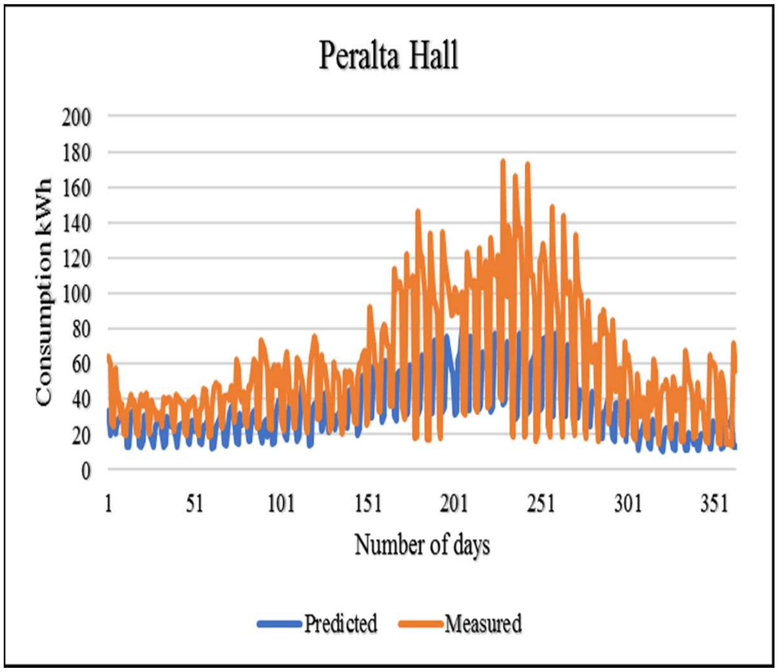
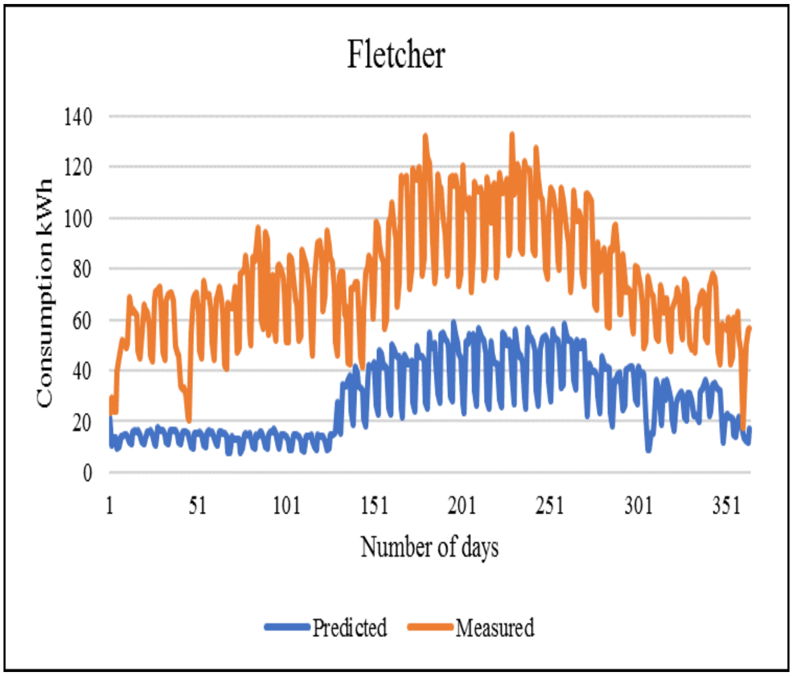


Figure 13. Predicted versus Measured Consumption values of Tempe Campus buildings (2015)

Figure 13 shows four different buildings and their consumption prediction. It is notable that the values and predictions from McCord Hall have a strong deviation from measured consumption (in kWh). While training machine using CRBM learning algorithm for McCord Hall, the input data was not comprehensive enough to train the machine because McCord Hall had no heating values (Table 1). Hence, deep learning requires many attributes to train the machine to its best and to have greater accuracy. The prediction graph of other buildings looks very close to the actual values, which indicates the framework successfulness and robustness.


Peralta Hall
200
180
160
140
120
100
80
60
40
20
0
Consumption kWh
1
51
101
151
201
251
301
351
Number of days
Predicted
Measured


Fletcher
140
120
100
80
60
40
20
0
Consumption kWh
1
51
101
151
201
251
301
351
Number of days
Predicted
Measured

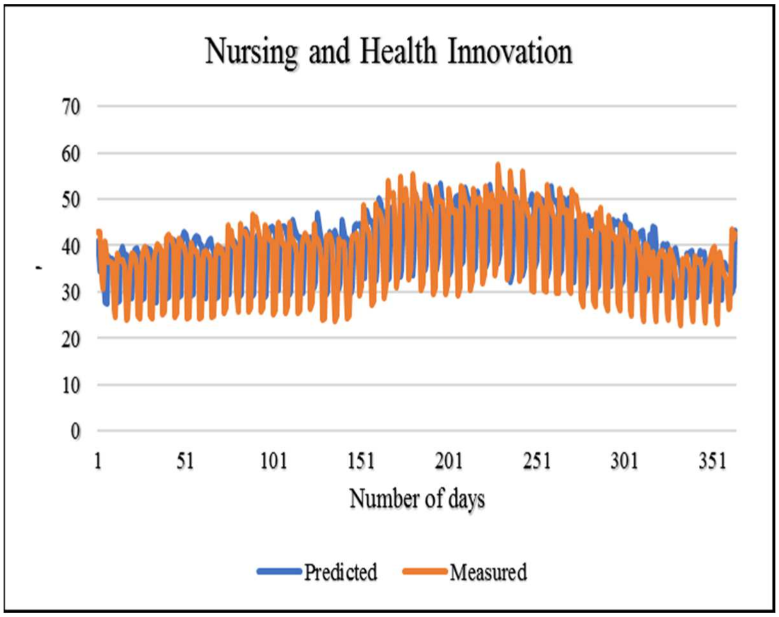


Figure 14. Predicted versus Measured Consumption values of other Campus buildings (2015)

Figure 14 shows the predicted and measured values of other three campuses at ASU. The other three buildings have data ambiguity, and the predicted values widely deviate from the measured value. The authors omitted those buildings to remove any wrong representations of prediction. Figure 14 also indicates that the prediction deviation of Fletcher Building is high. To better understand the prediction errors, the authors performed MAPE to identify and rectify errors on predictive analytics.

3.8 Discussion

Table 4 provides information on buildings and their MAPE values. It is notable that Fletcher and Peralta's Hall has a very larger deviation from the original value. However, Old Main and McCord perform better regarding the mean absolute percentage error. The percentage errors of other buildings range from 2.3%-2.7%. Institutional buildings have more uncertainties on their energy consumption characteristics because of sudden events, occupancy and another unexpected surge.

Table 4. Mean Absolute Percentage Error (MAPE) of predictive values

Buildings	Campus	MAPE
Nursing and Health	Downtown	2.7%
Computing commons	Tempe	2.5%
Engineering Center A	Tempe	2.3%
McCord Hall	Tempe	7.9%
Old Main	Tempe	3.4%
Fletcher	West	32.2%
Peralta Hall	Poly	15.6%

Thus, electricity prediction for an institutional building is a challenging task since the prediction accuracy can always be questionable. With latest smart grids, the solution is to integrate real-time data with more accurate techniques. The proposed method helps in predicting electricity consumption, which can help the building managers to understand the total demand required and can coordinate with the utility companies to avoid over-production or blackout. Also, the building managers can track the consumption of abnormal buildings (on specific days) to understand and implement better energy saving procedures. The study proposes a new integrated framework of WT and DL to preprocess and predict data. It is the first step towards developing a whole model that track, preprocess, detect anomalies, predict and finally visualize the data. Also, the process of automation is one of the future scopes of this study. Though the results have been demonstrated and validated, there are some limitations to this article. The concepts involved in this section are claimed to be future of Artificial Intelligence(AI) by few of the researchers in AI. Hence, the complexities must be eliminated to give better human experience. The other limitation is the data quality. Information is available in abundance, but it is important to have a comprehensive dataset for better outputs. The focus of this research will be by adding more attributes to the deep learning algorithm to improve predictive learning strategies and to elevate accuracy. It can help the industry practitioners, and any infrastructure developers to know and understand their energy strategies beforehand by using predictive analytic model.

4. A NON-STATIONARY ANALYSIS USING ENSEMBLE EMPIRICAL MODE DECOMPOSITION AND ISOLATION FOREST ALGORITHMS TO DETECT ANOMALIES IN BUILDING ENERGY CONSUMPTION

Summary

Chapter 4 details on anomaly detection using Isolation forest algorithms and Ensemble Empirical Mode Decomposition (EEMD). EEMD is utilized to smoothen the data to remove the negatives and noises in the consumption data. Later, Isolation forest algorithms are implemented to identify the anomalies from the selected buildings and to determine the anomalies based on the results from iForest algorithms. A part of this chapter has been published in conference proceedings of International Conference on Sustainable Design, Engineering and Construction (ICSDEC). The findings of this article is prepared for submitting to the *ASCE Journal of Energy and Buildings.*

4.1 Abstract

Commercial buildings' consumption is driven by multiple factors that include occupancy, system and equipment efficiency, thermal heat transfer, consumption loads, maintenance and operational procedures, outside temperature, heat index, etc. A modern building energy system can be regarded as a complex dynamical system that is interconnected and influenced by both external and internal factors. The modern large-scale sensor measures the physical signals to monitor real-time system behaviors and exhibits the potentials to detect anomalies, identify consumption patterns, and analyze peak energy loads. This paper proposes a novel data mining method for the detection of hidden anomalies in the commercial building energy consumption system. The framework is based on the Hilbert-Huang transform and instantaneous frequency analysis. The primary focus of this study is to detect anomalies from a preprocessed dataset and provide solutions with real-time consumption database using Ensemble Empirical Mode Decomposition (EEMD) and isolation Forest algorithms (iForest). The finding of this paper will include the comparisons of Empirical mode decomposition and Ensemble empirical mode decomposition and anomalous points of seven buildings over a daily dataset from selected buildings at the Arizona State University.

Keywords: Empirical mode decomposition; Anomaly Detection; Commercial building; Hilbert Transform; Supply-Demand Characteristics

4.2 Introduction

A modern building energy system can be viewed as a complex dynamical system that is interconnected and thus influenced by both external (weather) and internal (system efficiency) factors. CBECS (2016) statistics suggest that the floor space of commercial buildings have grown by up to 21% over the past decade. This growth in floor space increases the energy consumption even as energy efficiency improves. Humans in developed countries spend 90% of their time indoor (Deguen & Zmirou-Navier, 2010; Gee & Payne-Sturges, 2004). EIA (2016) found that 76% of energy consumed by the building sector was generated from different types of fossil fuels. Sartori et al. (2012) & Torcellini et al. (2006) suggested that new technologies, integrated building design and fault detection systems are needed to optimize energy production strategies and promote environment-friendly sustainable designs.

Modern sensors and tracking devices collect, measure and analyze physical signals to monitor real-time system behaviors. These devices generate dynamic, diverse and large dataset and signals and offer the potential to transform how buildings are managed. Large, continuous and real-time data could potentially be used to detect anomalies, identify patterns, determine characteristics, and analyze peak loads of energy demand, supply and consumption. This paper proposes a framework for detecting hidden anomalous consumption behaviors of buildings using pre-processed and smoothened dataset, that can aid energy managers in the design and management of building energy system. The concepts and data mining techniques employed in this paper include the Ensemble Empirical Mode Decomposition (EEMD) and iForest. These are used to detect the anomalies in electricity consumption of seven Arizona State University (ASU) buildings in the Phoenix Metropolitan Area.

The mathematical foundation of the proposed framework is based on the Hilbert-Huang transform and instantaneous frequency analysis. Hilbert-Huang transform and instantaneous frequency analysis is chosen as it is the leading approach to analyze nonlinear and non-stationary complex infrastructure systems (N. E. Huang et al., 1998; Yalçınkaya & Lai, 1997). Competing approaches, such as the traditional Fourier transform-based analysis, are extremely limited as they are designed for linear and stationary systems. This paper documents the implementation of a preprocessing technique that, first, enhances data comprehensiveness, and second, determine the anomalies of energy consumption patterns using iForest algorithms.

4.3 Research Objectives

The paper utilizes Interval Data (ID) from the Energy Information System (EIS) at ASU to detect, determine and quantify the anomalies. The data is treated to identify abnormal events and came from buildings of significant size. The research objective is to examine the accuracy of the proposed integrated EEMD-Isolation framework. The framework includes both the process of data treatment and anomaly detection. The data used in this article are generated from the daily electricity consumption data (kWh), from seven selected buildings on ASU five campuses. These buildings contain the most complete (and thus reliable) data set, are of mixed-used, and large size. While the data was collected from the ASU smart meters and systems, data quality is enhanced through different data mining methods. The paper includes, first, an extensive literature study on the empirical mode decomposition and the applications and roles of iForest algorithms in different fields, second the research methods, third develop a framework for EEMD and iForest pertaining to the detection of building energy consumption anomalies, and finally, the results and discussions of the anomalous points determined through the iForest algorithms.

4.4 Review of Relevant works

This section contains extensive review of relevant works pertaining to the empirical model decomposition, and applications and roles of iForest algorithms. The first part of the review discusses the gap in existing methods to preprocessing data and the advantages of EEMD over traditional methods. The second part details various anomaly detection methods used by the energy sector, and the role of new iForest techniques that would isolate anomalies better than other data mining methods.

4.4.1 Ensemble Empirical mode decomposition

Post-Occupancy Evaluations (POE) (Majcen, Itard, & Visscher, 2013; Newsham, 2009), data-mining (Ahmed, Korres, Ploennigs, Elhadi, & Menzel, 2011), model calibration

(O’Neill et al., 2011; Petersen & Svendsen, 2011; Raftery, Keane, & Costa, 2011), statistical analysis (Djuric & Novakovic, 2012; Ghiaus, 2006), and investment analysis (Kavgic et al., 2010; Koopmans & te Velde, 2001) were commonly used to narrow the gaps between designs and operations in building energy design and operation through building energy models. However, these methods do not generate sufficient information to connect existing design and operational performances (De Wilde, 2014), and thus were unable to predict actual performance accurately using historical data.

Energy design involves connecting the lifecycle relationships between energy demand and supply, and the successful connection would propel energy efficiency to the next level where energy losses would be accurately estimated, and integrated into energy design. The feedback from operation and factors identification is critical in closing the design-operation gap (Dodoo, Gustavsson, & Sathre, 2011; Pérez-Lombard, Ortiz, & Pout, 2008; Raftery, Keane, & O’Donnell, 2011; Ryghaug & Sørensen, 2009; L. Wang, Mathew, & Pang, 2012). The significance of the relationships between factors and time vary, for example, occupancy rate is highly dependent on time while humidity does not. These factors, however, affect energy system performances indirectly and directly.

Traditional methods, like the Fourier transform, assume stationarity and approximate the physical phenomena with linear models. These approximations may lead to spurious components in their time-frequency distribution diagrams if the underlying signal is nonstationary and nonlinear. The Empirical Mode Decomposition (EMD) is a technique (N. E. Huang et al., 1998) to deal specifically with non-stationary and nonlinear signals. EMD decomposes signals into distinct modes, identified as the intrinsic mode functions (IMFs), giving each signal a distinct time or frequency scale while preserving the amplitude of the oscillations in a frequency range. The decomposed modes are orthogonal to each other, and the sum of all modes becomes the original data. The ease and accuracy with which one uses the EMD method to process non-stationary and nonlinear signals have led to its widespread use in many applications such as seismic data analysis (N. E. Huang et al., 1998), chaotic systems analysis (Lai, 1998; Yalçınkaya & Lai, 1997), neural signal processing in biomedical science and engineering, meteorological data analysis (Ghiaus, 2006), and image processing (Nunes, Bouaoune, Delechelle, Niang, & Bunel, 2003).

While these methods have been applied successfully for different types of analyses, the signals generated by the EMD process have large oscillations and difficult to reiterate IMFs at a different levels (Wu, Zhaohua and Huang, 2009). To overcome such difficulty, the paper proposes Ensemble Empirical mode decomposition (EEMD) method instead of EMD method. EEMD flattens signal oscillation when data discontinues, and smoothens both the oscillation and data by canceling the noise within the data. EEMD was integrated into existing models for long term load forecasting and also used to enhance existing forecasting techniques (Ghelardoni, Ghio, & Anguita, 2013). This research paper only uses EEMD to reduce the data noise and thus smoothen the data. The iForest algorithms developed in this paper are used to detect the anomalies of building energy consumption, and thus to use mathematical and data mining model to elevate the prediction reliability of energy consumption.

4.4.2 Anomaly Detection and Isolation framework

Beniger, Barnett, & Lewis (1980) defined outliers as an observation that is inconsistent with the remainder dataset. Aggarwal & Yu (2001) indicated outliers as the points that stays outside a data cluster that share the similar pattern and are also separated from noises. Outliers or anomalies are errors that occur in a system due to, faults embedded in the mechanical systems, human errors, erroneous meters, data deviations, or unexplained sudden surge in energy consumption. Detecting and removing anomalies is an important process to remove the anomalous behavior of the system before the "cleaned" data is used for analysis and then make decision. Outliers identify the presence of situations and scenarios that deviates data from the norm, and such data needs to be removed from the dataset. The paper discusses the approaches to identify such outliers beyond the noises and negative values, and this will enhance the quality of such analyses. Removing outliers is the first step in the anomaly detection process.

Anomaly detection is a critical task in detecting outliers (and noises) from dataset. As outliers indicate abnormal conditions, removing them would enhance the quality of prediction models (Hodge & Austin, 2004). Mathematical, statistical and data mining models are commonly used to detect anomalies (Aggarwal & Yu, 2001; Beniger et al., 1980; Bhuyan, Bhattacharyya, & Kalita, 2014; Ghosh & Vogt, 2012; Gogoi, Bhattacharyya, Borah, & Kalita, 2011; Langford & Lewis, 1998). Existing models identify instances that do not conform to the standard profile, and thus an isolation-based model is required to detect anomalies (F. T. Liu, Ting, & Zhou, 2010).

Vengertsev & Thakkar (2005) detected anomalies in graphs using unsupervised learning, graph-based features, and deep architecture. In their work, iForest was utilized to construct the ground truth labels (Vengertsev & Thakkar, 2005). Most anomaly detection methods, like the one-class SVM, determine points by assigning labels for each test instance (Abe, Zadrozny, & Langford, 2006; Vengertsev & Thakkar, 2005). Isolation-based anomaly detector is an emerging technique that does not rely on density or distance measurement (F. T. Liu et al., 2010).

iForest is utilized in various fields to detect anomalies in their systems and dataset. iForest was utilized to detect water leaks in pipelines (Begovich & Valdovinos-Villalobos, 2010; Province, 2011). Similarly, Bandaragoda et al., (2014) developed an efficient method using isolation forest and nearest neighbors ensemble to detect anomalies various human activities captured using sensor readings. W. Liu & Hwang (2011) developed a fault detection system for air traffic control using iForest algorithm. iForest requires far less computational time, and the analysis process is far more effective than other methods. Time-saving and efficiency become increasingly prominent as the quantity of algorithms, factors and data increases (Carrasquilla, 2010). iForest technique has become increasingly used in many scientific fields; however, it is still relatively new in the area of energy. The process of EEMD and iForest will be explained in the following sections using the analyses from seven buildings.

4.5 Research Methodology

The proposed EEMD-Isolation framework includes two essential steps. The first step is to implement EEMD concept to smoothen the data, and remove noises and negatives from the dataset. During this process, the intrinsic mode functions (IMF) are created for each building and the smoothened IMF are reconstructed for detecting anomalies. In addition, the advantages of EEMD over EMD is explained by the following graphical representation. The second step is to implement iForest algorithms which have two other sub-steps that include training and testing using iTree and iForest algorithms.

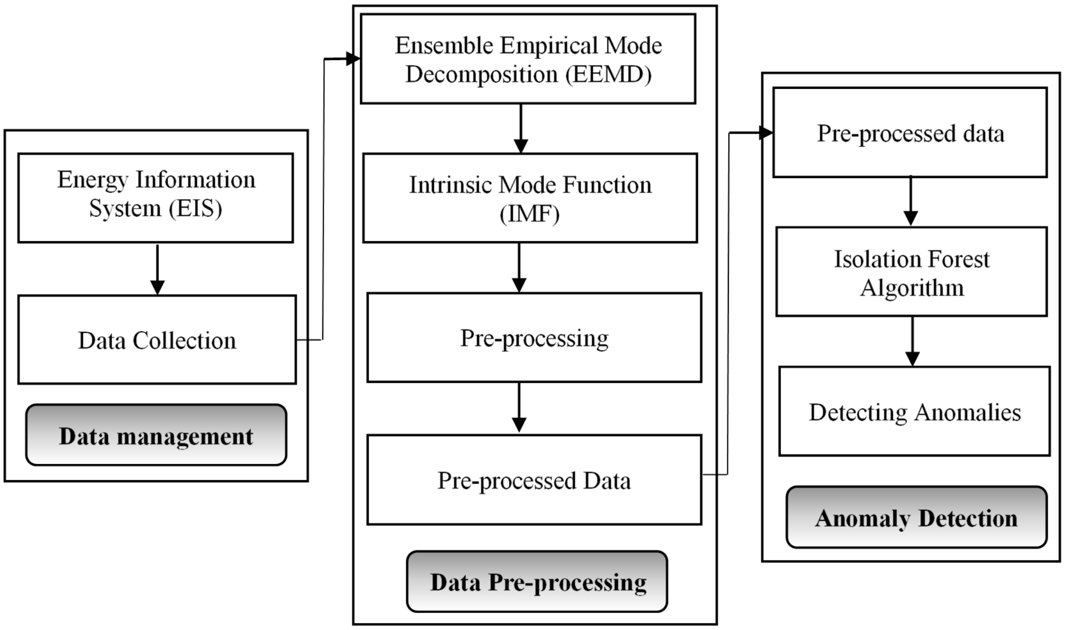


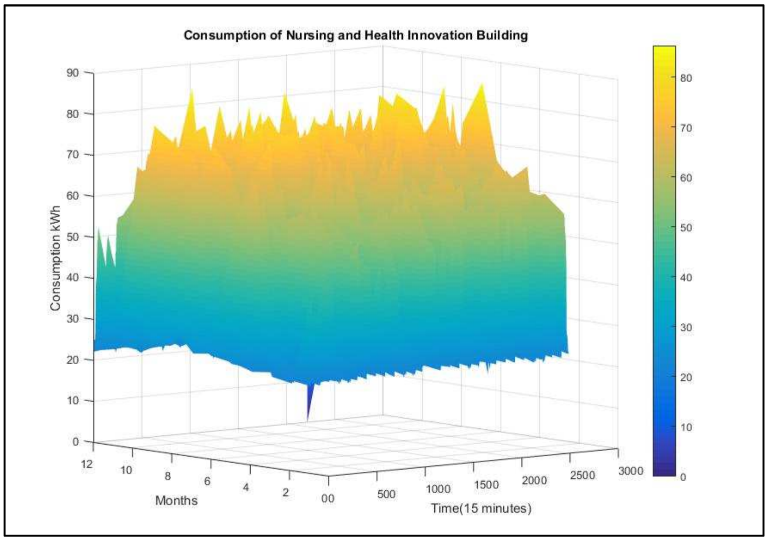
Figure 15. Research Methodology

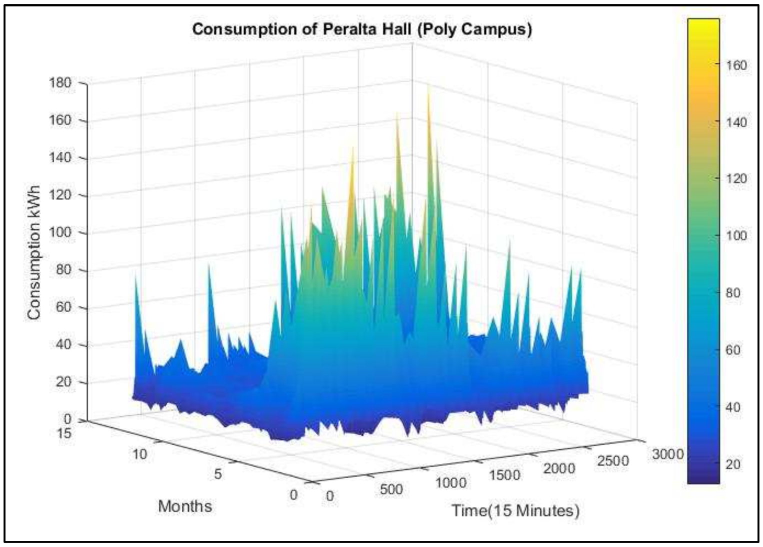
The anomaly score, path length, and the anomalous points are identified during this process of EEMD-Isolation framework. Figure 15 describes the stepwise procedure of the methodology.

4.5.1 Data Collection

Data collection from the EIS includes data on electricity consumption for all seven buildings at ASU. Data is collected from selected buildings on the ASU five campuses.

Each figure in Figure 16 indicates the raw energy consumption representation of a selected building. The figure is a three-dimensional graphical representation with an x-axis (represents time at 15-minute intervals), y-axis (represents energy consumption) and z-axis (represents year).





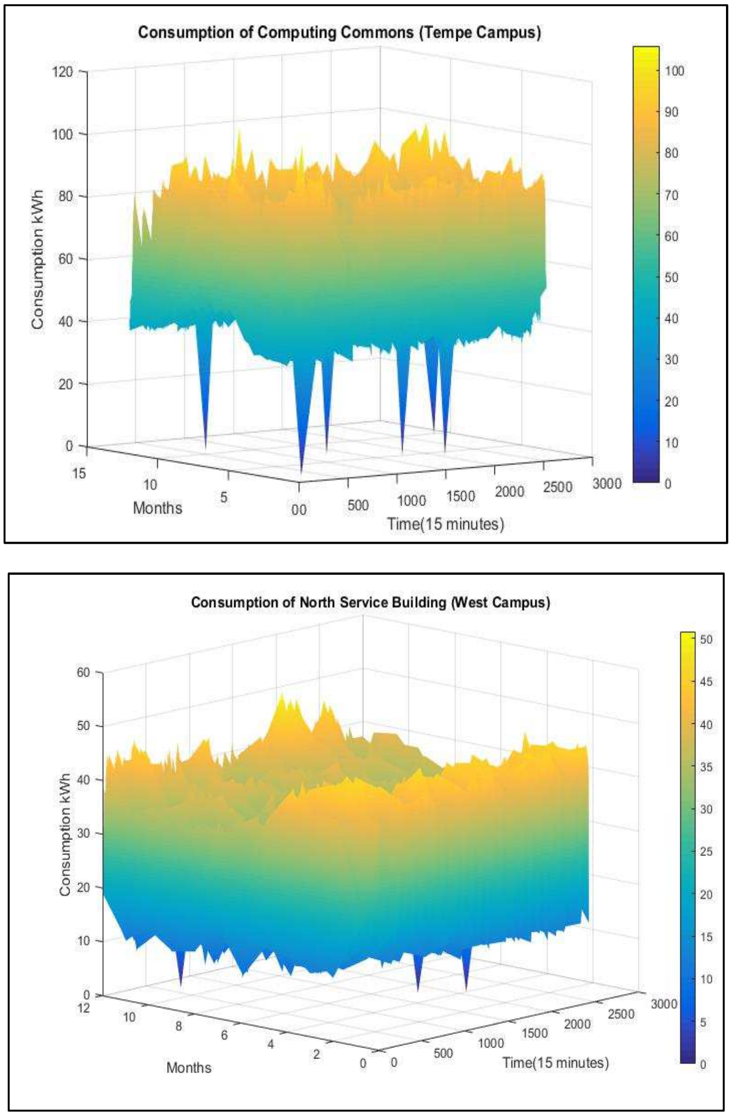


Figure 16. Non-treated Consumption data of Buildings with noises

4.5.2 Data Preprocessing

It is not possible to preserve the integrity of broad and complex dataset, especially for the data recorded continuously and over an extended period of time. The chance of disturbance during the data collection process due to detector/sensor malfunctioning increases as the amount of data and time increase. All dataset contains disturbed and interrupted data segments and such data must be removed from the dataset. Such disturbances and interruptions would affect the analyses and outcomes. Pre-treating dataset is the necessary step in removing the disturbances and interruption, and thus repair the “damaged” segments. The treatment would improve the quality of the dataset to better reflect the actual conditions.

4.5.3 Data Preprocessing techniques

There are different types of data treatment methods, such as the Fourier Transform. However, most of these methods are unable to overcome the limitations when signal frequency changes with time in a time-series analysis. Ensemble Empirical Mode Decomposition (EEMD) is more suitable for generating IMFs where frequencies vary with time when the IMF period is a function of time: T = T (t).

The first step of this research is to apply EEMD to generate the Intrinsic Mode Functions (IMFs) using available dataset. Using IMFs frequency signals and standard mathematical concepts (Hilbert Transform in this research), the data from selected buildings was first preprocessed to remove noises and errors (such as negative values) from the dataset. The frequency signals from the iterated IMFs were then reconstructed into the original database. The reconstructed data was analyzed using iForest algorithms to detect anomalies.

4.5.4 Isolation Forest algorithms

The framework developed in this paper targets the detection of anomalies from the energy consumption data of selected buildings. In this paper, iForest algorithms were utilized to detect anomalies as it is more reliable and accurate than other fault detection methods. Novel iForest algorithm is faster and more accurate than the original developed by the Oak Ridge Cyber Analytics (ORCA is a set of tools used in analytics of information security issues) and random forest. (F. T. Liu, Ting, & Zhou, 2008). Liu et al.'s (2008) iForest algorithms assumed that anomaly represented a small part of the whole dataset and the attributes were different from one other. iForest is best suited for training dataset that does not contain noises, while EEMD smoothens data before iForest implementation.

iForest has three algorithms combined: The first algorithm is where the anomalies stay at the root nodes of the tree. The second algorithm is where the oscillation height limit is specified, while the anomaly scores and anomalous points are determined. The third algorithm specifies the path length of each oscillation for each dataset. The value of the path length determines the accuracy of the anomalous points.

4.6 Framework Validation

EEMD would generate a set of IMFs in different frequency ranges for a given dataset. As the dataset is too large to be processed efficiently, the data was divided into smaller segments to enhance computational efficiency. Each data segment would include a much smaller subset of data points from neighboring and both ends of each segment to form a corresponding boundary sets, to eliminate potential boundary effects between different dataset.

Only the IMFs of the original data segment were kept after performing the EMD calculations, while the association between data within the boundary sets were disregarded. The IMFs resulting from the analysis depend on the sizes of each segment and the boundary sets. Increasing the size of the boundary sets would increase both the IMFs accuracy and the computational time. This research proposes an analytical procedure for large dataset that consist of EEMD analysis to obtain the dataset IMFs, amplitudes and frequencies calculations of the IMFs (for revealing the dynamical evolution of the underlying system), and relevant statistical analyses to support EEMD.

4.6.1 EEMD module

Seven buildings from across five ASU campuses were selected for the project. These buildings shared similar energy consumption and utilization patterns, and are mixed-used buildings. Energy information system (EIS) is installed in these buildings, and data was collected from the EIS and converted into daily values for a year. The above-mentioned data preprocessing technique was used to treat the data. EEMD's robustness is also used to address the problem associating with the consistently flat portion of the dataset. Flat portion of the dataset means that the values continue to remain relatively similar over a long period of time, and data cannot effectively be converted into IMFs and oscillations (i.e. wave). EEMD is a self-adaptive algorithm that could easily be performed on time-domain data. During data decomposition, the sifting process to decompose the data into *n* units of h_j to present the IMFs (Ren, Wang, Huang, Chang, & Kao, 2014). The equation for the EEMD process of EEMD is given in the following equation:

$$x(t) = \sum\nolimits_{j=1}^{n} h_j + r_n$$

Where *n* denotes the number of iterations, h_j denotes the IMF number and r_n denotes the residual values. Figure 17 details on the step by step process of EEMD in denoising the data and reducing the chances of different mode mixing.

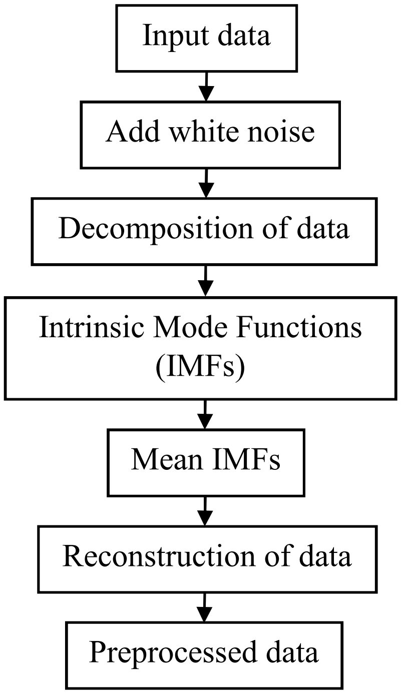


Figure 17. Process of Ensemble Empirical Mode Decomposition (EEMD)

EEMD decomposition would result in the added white noise series canceling one other, and the mean IMFs would stay within the natural dyadic filter windows to reduce the chance of mode mixing thus preserving the dyadic property (Z. Wu & Huang, 2009). Figures 18 depicts the results of the EMD and EEMD decomposition process using a one-year electricity consumption dataset (365 data points) from a sample building. The EEMD process flattens the data from the extremely noisy oscillation to a flatter wave at the bottom of the figure. The figure shows that EEMD yield better results than EMD while handling oscillations. Figures18 show the IMFs (IMF-1 at the second while IMF-5 at the bottom) with raw signals at the top and the residuals at the bottom. EEMD iteration process would result in more intrinsic signals than EMD.

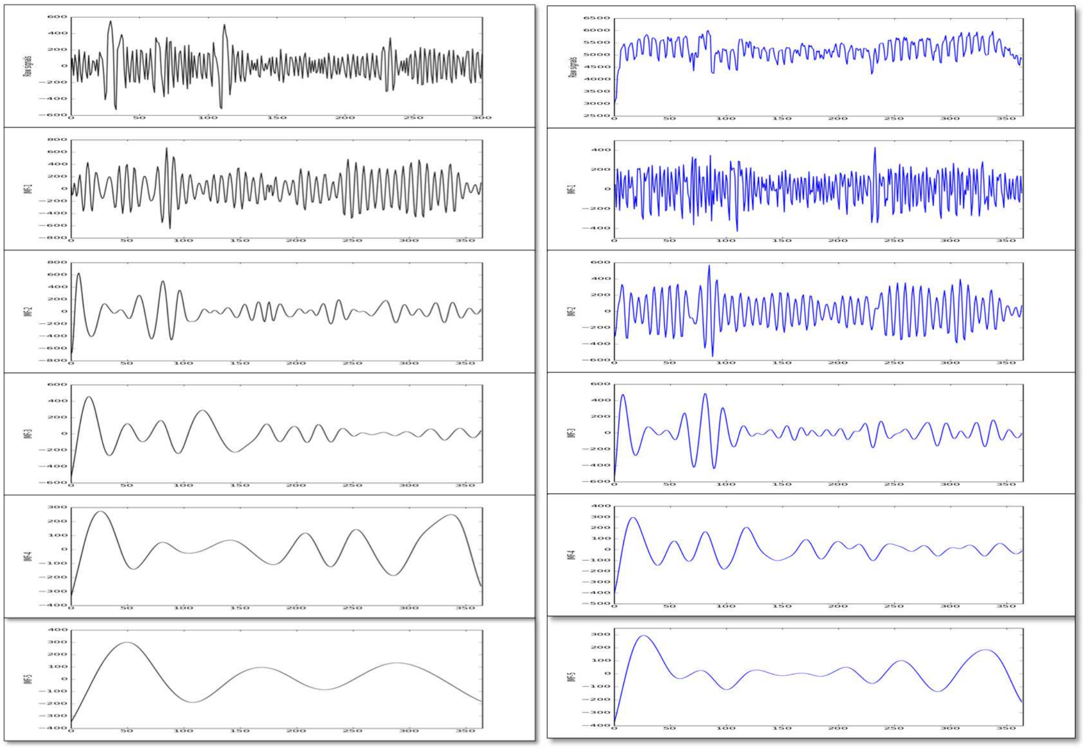


Figure 18. IMFs of EMD and EEMD for a selected building

Note: X axis is the number of days in a year (365) and Y-axis is the IMF signals (from 1 to 5)

The IMF at the bottom of Figure 18 shows how EEMD prevents oscillation variations and the lower value data points are mostly flattened. EEMD thus overcomes the difficulties of "flat" data that the EMD algorithm would deviate to higher oscillations (thus leading to great waves and complicate analysis). EEMD also has the capability of converting unlabeled or unstructured data to reflect the patterns, behavioral changes and intrinsic relationships between devices (J.-P. Tang et al., 2011). EEMD is thus a good method for detecting anomalies of energy optimization.

The consumption data utilized in Figure 3 were from the electricity consumption data of Building A at ASU Downtown campus. Four buildings on the ASU main campus

(Tempe) were selected as most of the ASU buildings are in Tempe, and most of the Tempe buildings are connected to EIS. In addition, data quality at Tempe campus is far more superior and comprehensive than buildings on other campuses. Table 1 shows the selected buildings, their location and square footage.

Table 5 Selected buildings and their characteristics

Buildings	**Campus**	**Gross Square Footage (ft^2)**
Building A	Downtown	183,435
Building B	Tempe	132,518
Building C	Tempe	137,040
Building D	Tempe	140,092
Building E	Tempe	45,017
Building F	West	102,856
Building G	Polytechnic	88,456

After performing EEMD and decomposing the data, the mean IMFs were then calculated to reconstruct the data. The data is further treated by Hilbert transform. Hilbert transform is a unique harmonic analysis approach where data undergoes convolution base on *u(t)* [where *u* is the data which is time dependent – reflected by *(t)*]. Hilbert transform would produce discrete data in the frequency domain to undergo further study in EEMD. Alternatively, the frequency domain displaced data can be retracted using Inverse Hilbert Transform (IHT) (J.-P. Tang et al., 2010). Thus, it is necessary to use the original data on the iForest algorithm for anomaly detection.

4.6.2 EEMD-Isolation Framework

As mentioned before, iForest algorithms are used to detect anomalies from the reconstructed energy data. iForest was recognized as an efficient anomaly detection method due to its processing speed and enhanced reliability than other random forest methods (Liu et al., 2010). Figure 19 illustrates the working process of the iForest algorithms.

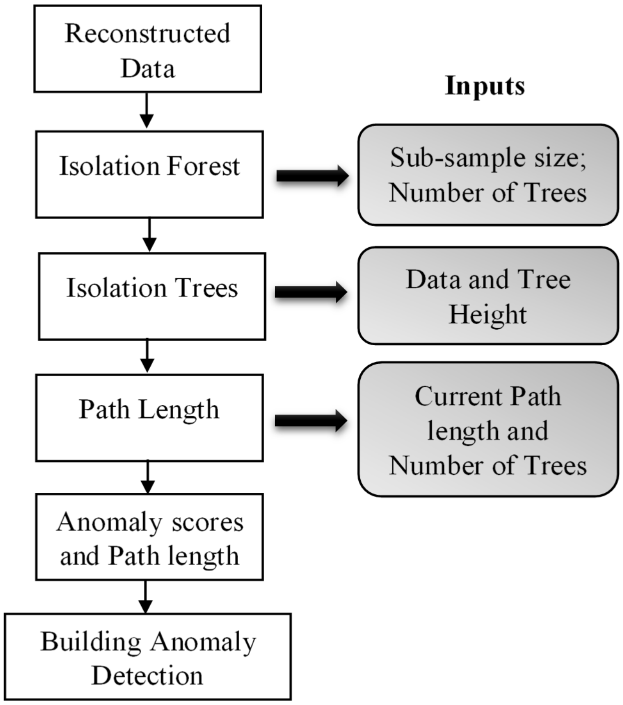

Reconstructed Data
Isolation Forest
Isolation Trees
Path Length
Anomaly scores and Path length
Building Anomaly Detection
Inputs
Sub-sample size; Number of Trees
Data and Tree Height
Current Path length and Number of Trees

Figure 19. Working Process of iForest Framework

4.6.3 iForest Module

iForest algorithms require lesser quantity of subsampling data for the anomaly detection process than other algorithms. As shown in Figure 19, the iForest algorithms and process require several inputs before analysis. The first step is to execute the iForest algorithm in which the subsample size and number of trees must be inserted as inputs. Since the available dataset were converted from 15 minutes data into daily values to enhance the results, the subsample size for running algorithm is taken from the whole dataset for each building. Liu et al., (2008) suggested that these subsamples and the path length will converge completely before 100 iterations, and they recommended to assume the input number of trees as 100. The height limit of the trees to converge was selected as ten (10) before the path length was converged providing enough information on anomalous points.

4.6.4 Path lengths

Path lengths are numeric that indicates the anomalous points are closer to the root of the tree. The points near the root are anomalies and those away from the root are common points. Path length closer to a value of 1 would indicate a greater accuracy of identifying anomaly. The anomaly scores would normally range from 0.5 to 1, and the average value of a building closer to 1 would indicate there are anomalous points in a dataset.

4.6.5 Buildings and their Anomaly detection

After the preprocessing, iForest and iTree were then implemented on the treated data with 100 iterations and 100% subsampling. Figure 20 below illustrates the number of anomalies in Building A and their consumption on the y-axis. Data for Building A contained 34 anomalous points with an average anomaly score of 0.71.

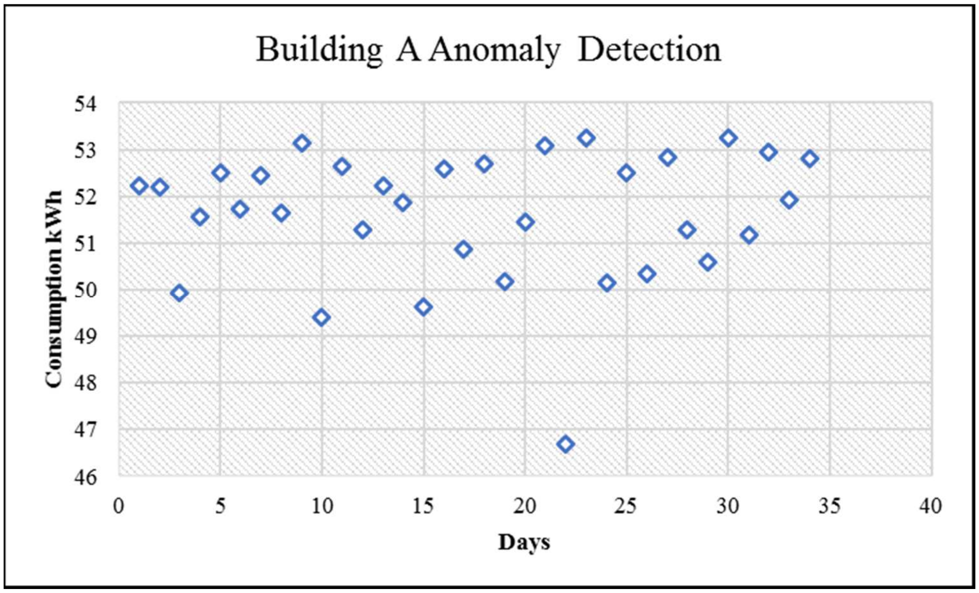


Figure 20. Building A Anomalous Points

In addition, the average path length for Building A converging in the trees is 3.23, and this indicates the accuracy of the anomalies identified. Almost 10% of the total number of data points were determined to be anomalous. The average path length and anomaly score of Building B were 3.74 and 0.73 respectively. Though the anomaly score was close to 1, the average path length indicated a higher divergence. The data was processed to identify the number of anomalies with a height limit of 10, the number of iterations is 100. Figure 21 indicates the anomaly detection graph of Building B.

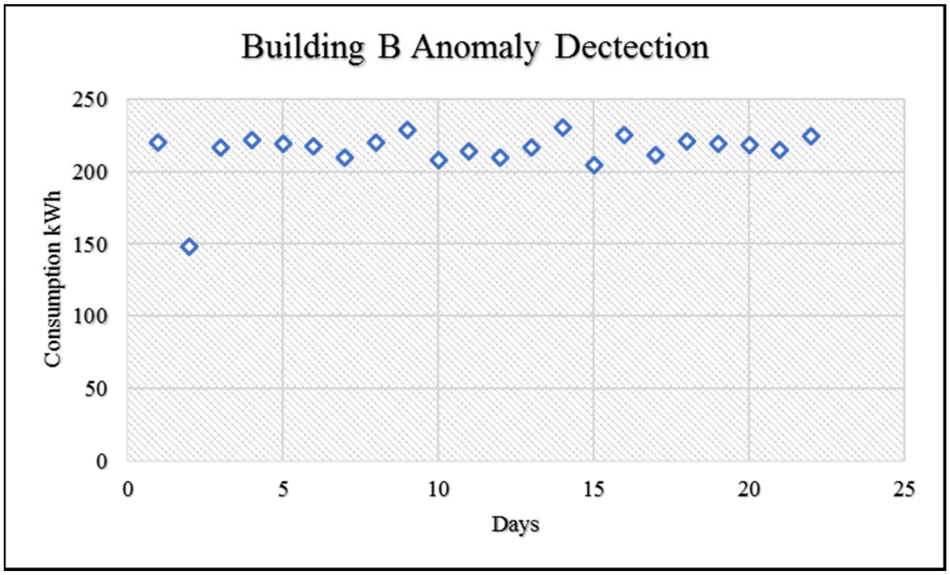
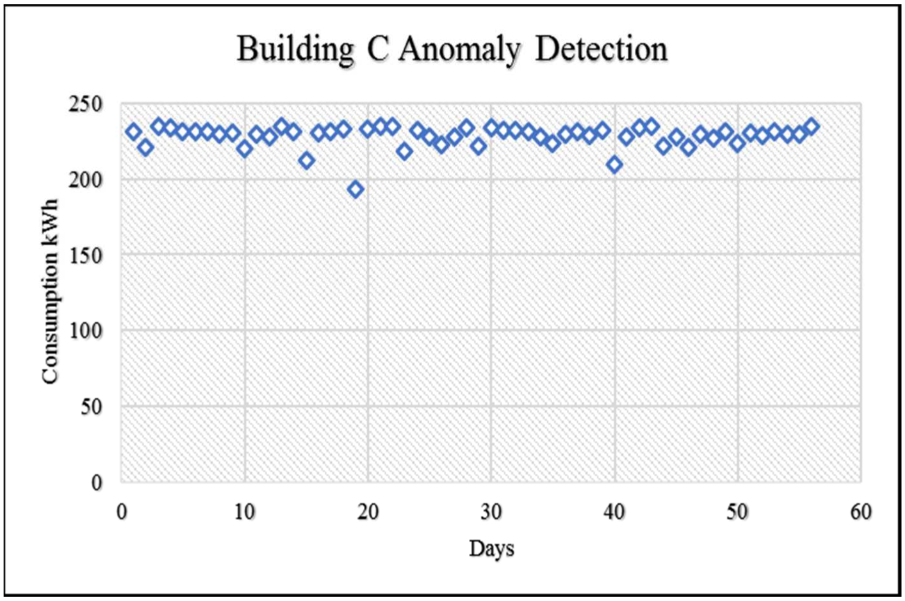


Figure 21. Anomaly detection of Building B

The anomalous points of Building B were 24 and they were lesser than Building A. Figure 8 indicates the anomaly detected on Buildings C and D, and the anomaly points of were 56 and 50 respectively. The average path length of Buildings C and D are 4.56 and 4.41 respectively, and these highlighted that anomalies were detected accurately. There were significant anomalies as there might be a lack of consistent data from the buildings that prevented the algorithm from calculating more accurate results. The anomaly scores of these buildings are 0.69 and 0.67, and these indicate that the numbers were moving towards 0.5. The result proved that the data quality was compromised.



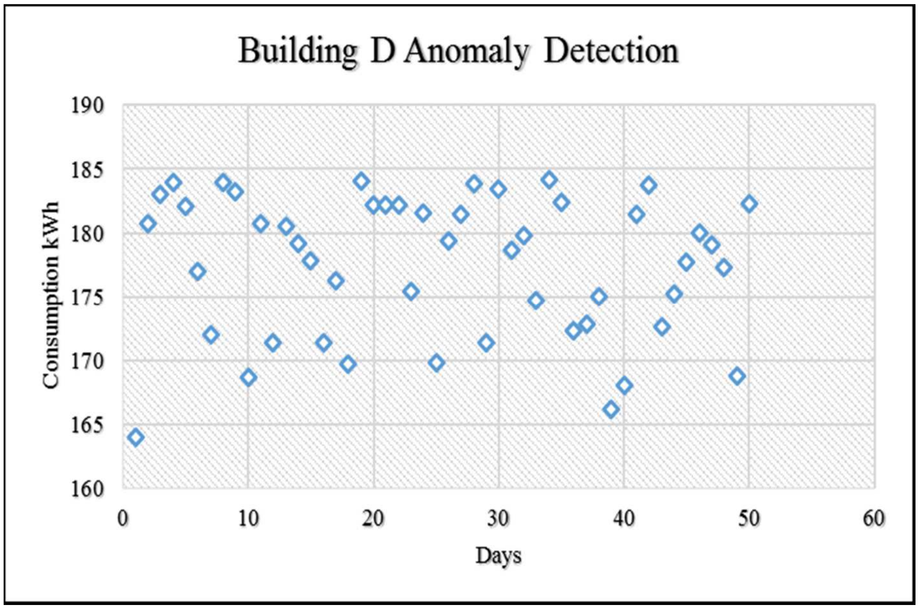
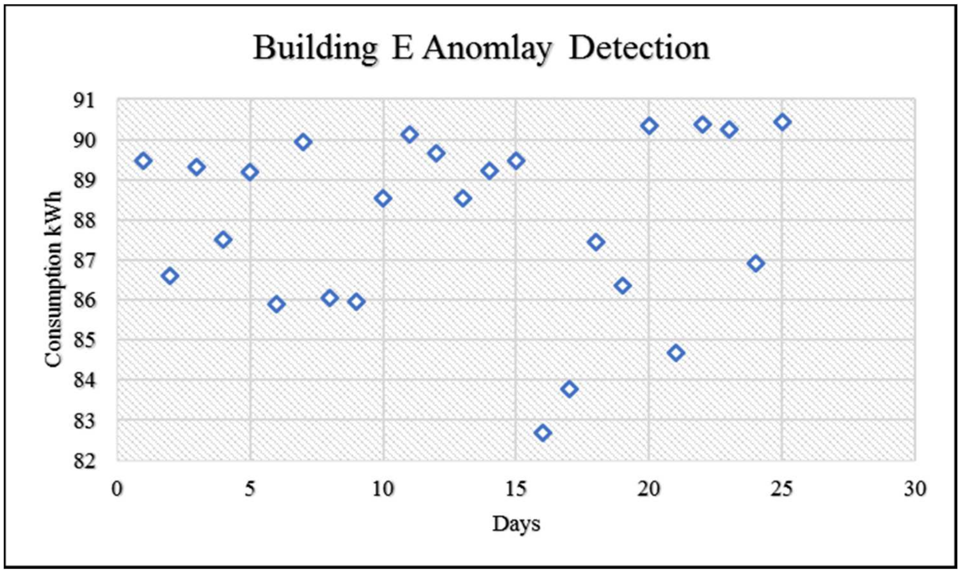
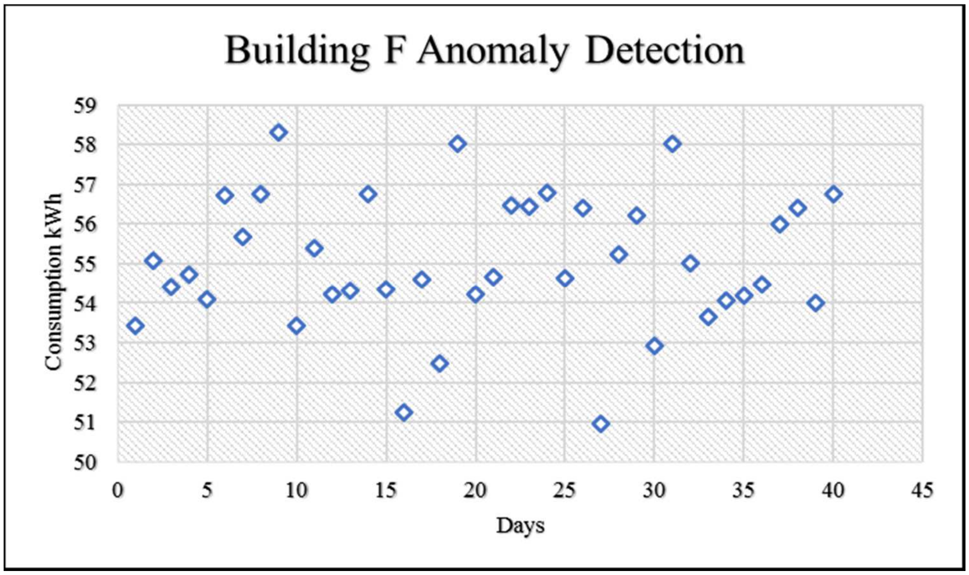


Figure 22. Anomaly Detection of Buildings C and D

Figure 22 indicates that Building C’s anomaly shows some discrepancies within the data. However, the analysis for Building D indicated sufficient variations for the consumption values and thus this indicates more accurate anomaly points. Figure 23 illustrates Buildings

E, F, and G having average path length of 3.12 and contain 25 anomalous points. The anomaly score of Building E was 0.79.





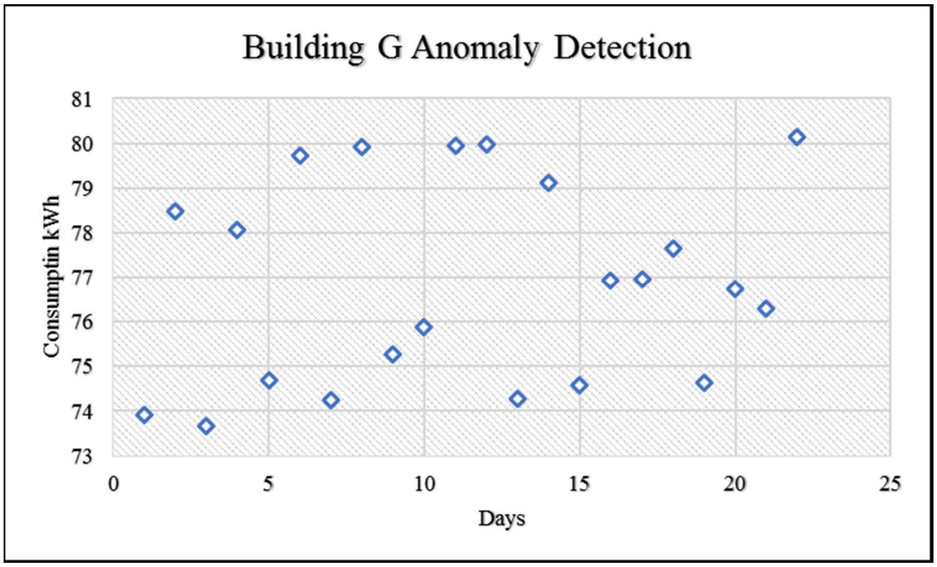


Figure 23. Anomaly Detection of Buildings E, F, and G

Similarly, Building F has 40 points as the anomalies have an average anomaly score and path length of 2.87 and 0.82. These indicate a higher degree of accuracy of anomalies. Building G has 22 anomalies and has an average path length and anomaly score of 3.23 and 0.71 respectively. All the seven buildings' energy consumption values were processed through iForest framework and their respective anomaly points, path length and anomaly score are identified.

The anomalous points of all the buildings are identified using iForest algorithm. These points indicate that there are considerable number of anomalies in each building over a period of one year. These points can help in identifying the reasons such as sudden events, unexpected surge, or high occupancy rates in the building.

Table 6 Buildings and their anomaly percentages

Buildings	Anomalous Points	Total Data Points	Percentage of Anomalies
Building A	34	365	9.3 %
Building B	22	365	6.0 %
Building C	56	365	15 %
Building D	50	365	13.6 %
Building E	25	365	6.8 %
Building F	40	365	10.9 %
Building G	22	365	6.0 %

Table 6 shows the buildings and their respective anomaly percentages. It is evident that Building D has the highest number of anomalies. The objective of this paper is to develop a framework that can detect anomalies and it is important to present these results to the building managers or energy analyst to identify the concrete reasons for these anomalous behaviors in the consumption.

4.7 Discussion

The EEMD-Isolation based algorithms that the paper proposed has the potential to optimize energy fault detection and align energy design and operation. The algorithms will create a platform that leads to a fully automated method to detect dynamical anomalies from broad and complex data sets. It is anticipated that the proposed method would detect many anomalies from generic vast and complex data sets, which are not detectable using traditional methods.

The results from all seven buildings indicate that the model is successful in determining the anomalies using EEMD processed dataset and iForest algorithm. However, some buildings' results (such as Building B and C) indicates the need to improve the model through supervised training and testing with a comprehensive database. It also provides an excellent test ground for probing into the emergence and evolution of anomalies through detailed analysis using methods from nonlinear dynamics, statistics, and statistical physics.

Detecting anomalies in the consumption helps in identifying the abnormality in the consumption pattern. It helps in notifying the building manager, energy management system, and the owner about the anomalies that helps them in determining the appropriate energy saving strategies. The proposed method aids as a notification to the building management system to know that there are anomalies in certain buildings, thus alerting them about sudden or unexpected surge or blackout. The detection will lead to the development of energy control systems that could be used to optimize energy design and operation. With the optimized EEMD-Isolation forest-based method, anomalies can be detected reliably for all the kind of buildings.

4.8 Future work

The research was limited to seven buildings due to data quality and lack of reliable data. The integration of smart meters and EEMD-Isolation framework will help in eliminating the chances of data loss or quality. In addition, the algorithm will be improved to improve the accuracy of anomaly detection. The findings also indicate the need for data preprocessing to serve the energy industry with highly accurate results.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of Results and Contributions

The dissertation provides the energy analytic frameworks through various main body chapters. The objective of this research to develop frameworks that contributes to the total energy analytic framework. The objectives are met through three main body chapters, which are published or prepared for submitting to peer-reviewed journals. All three main body chapters implement first, a data preprocessing technique using mathematical and computational models, and second, a data processing technique using data mining, machine learning, and deep learning algorithms to analyze supply-demand characteristics, predict electricity consumption, and to detect anomalies from the database collected from Energy Information System (EIS) at ASU.

The analysis presented in Chapter 2 contributes to the overall objective by analyzing the supply-demand characteristics to elevate the energy loss reduction potential through k-means clustering and semi-supervised energy (deep learning) framework. Prior studies indicated the need for an automated model to evaluate energy losses between supply and demand and provide solutions to improve the loss reduction potential. After necessary data cleaning process, K-means clustering technique was applied to the electricity consumption data of buildings that are connected to different substations at ASU Tempe Campus. Thus, the buildings are grouped based on their consumption patterns to their respective substations. The second and final step is to implement Semi-Supervised Learning (SSL) algorithms that train the machine using both labeled and unlabeled data to improve the accuracy. The findings of this study indicate a potential of 15% energy loss reduction between supply (substation) and demand (building consumption) and the process is automated to visualize the loss reduction potential through graphical representation.

Chapter 3 presents a novel hybrid Wavelet-Deep learning framework to pre-process and predict electricity consumption of ten different ASU buildings located at various campuses and different cities in Arizona. After data screening, three of the buildings are omitted to avoid biased data in the model for predicting electricity. Prior studies provided extensive information on various energy forecasting models using mathematics, statistics, and computer-based models. The motivation of this study relied on the utilizing the most accurate (close to human brain accuracy) deep learning framework (CRBM) to predict the electricity consumption using wavelet treated data. First, the data is treated using an extensively implemented mathematical technique called Wavelet decomposition. After obtaining decomposed data, Conditional Restricted Boltzman's Machine learning algorithm is implemented to predict the electricity consumption of seven ASU building. The results of this WT-DL framework showed an accuracy of more than 90% on all seven buildings selected for the research. Thus this chapter contributes to overall objective through a framework development that can preprocess and predict electricity consumption of institutional buildings.

Chapter 4 details on utilizing a mathematical model and a data mining technique to detect anomalies from institutional buildings. Educational and institutional buildings have a broad range of abnormalities due to their building type, utilization, occupancy variations and other interdependent factors. The preprocessing technique used in this chapter is Ensemble Empirical Mode Decomposition (EEMD) that decomposes the data by removing noises and biased values and provide more comprehensive data for anomaly detection model. It, in turn, increases the accuracy of the model since all random noises from the data are neglected through EEMD. The decomposed data from EEMD is utilized to perform anomaly detection using iForest algorithm. iForest technique has been employed in various fields, and this chapter proposed a hybrid EEMD-Isolation forest framework to detect anomalies of institutional buildings. The findings of the study included results of seven buildings' anomaly points, their path length and anomaly score using three different algorithms under Isolation forest framework. All seven buildings had a considerable amount of anomalies indicating the regular system maintenance requirement and improving consumption strategies using the better building energy management system.

5.2 Limitations of the Research

The dissertation addressed the overall process of energy analytics (without cost savings) through supply-demand characteristics, energy predictions and anomaly detection. While each chapter (2,3, and 4) had their related works, objectives, research methods and results, every chapter had their limitations which have been discussed in the following section. One of the obvious limitation for this research is the quality of data collected from EIS at ASU. EIS at ASU was established in 2011 and made constant effort to collect data and integrate all old and new buildings into the building management system. Hence, the data collected from ASU sometime possess biased meter readings. It is one of the motivation to implement a preprocessing technique for different energy analytic purposes.

Chapter 2 addressed the clustering and semi-supervised learning technique to evaluate the loss reduction potential between supply-demand data. Although the reliability and accuracy of this model have been demonstrated to be acceptable, more works are still needed. The effort is necessary to overcome the complexity embedded in both models and data, and extensive knowledge of machine learning and cybernetic concepts, and power generations are also required. Training the machine with algorithms is the highly complex exercise after which the machine learns the pattern and automate the model for greater accuracy.

In Chapter 3, the research study was limited to only ten buildings because of data quality issues from all different campuses at ASU. Also, the methodology used in this chapter involves deep learning algorithms, which is highly complex and requires high-level computer science knowledge. These algorithms are utilized by very few industry pioneers to optimize their energy demand in their data centers. Hence, this serves as one of the major limitation since energy managers or researchers can not easily utilize it without extensive computer background.

In chapter 4, the research study was again limited to seven buildings due to lack of reliability of data. Also, the reliability of the model is still questionable because of the irrelevant results on a couple of buildings. The effort is necessary to reduce complexity and to elevate the accuracy of anomaly detection algorithms through a comprehensive database that can train the machine for better results.

5.3 Future Research

The main body chapters (Chapter 2,3 and 4) contributes to the overall objective of the research through analyzing and visualizing supply-demand characteristics, predict electricity consumption and detecting anomalies in the electricity consumption data of the institutional buildings. Each chapter developed their data mining, machine learning or deep learning framework to perform energy analytics. One of the main future work on this research would be on integrating all the framework into one comprehensive framework under the same platform (Example: Hadoop or Matlab) that can ease the researchers to identify the knowledge sharing platform. In addition, the research analyses data only from institutional buildings because of its availability. It is essential to develop the comprehensive framework into a data model that can be utilized for all kind of infrastructure. Also, the future work will include attributes and variables that contribute to the total energy consumption of any infrastructure. It can make this framework a generic tool to reduce losses, predict consumption and production and to detect anomalies from the comprehensive dataset.

5.3.1 Applications

All the frameworks developed in this dissertation are data based. The advantages of these frameworks are each framework has their data quality testing, or data transformation model integrated thus aids in enhancing the data before performing analytics. The research study has been extensively implemented in institutional because of the availability of data from the sustainably-elevated institution at Arizona. The application of this research includes

1. Transportation-based data analytics such as traffic signal optimization, real-time accident prediction, and self-driving vehicle's accuracy prediction.

2. Bridge engineering analytics where deflection of beams and girders can be detected and predicted using the real-time and historical database.

3. The complete framework can also be applied to projects that include database from renewable energy grids.

REFERENCES

Abe, N., Zadrozny, B., & Langford, J. (2006). Outlier detection by active learning. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’06, 504. http://doi.org/10.1145/1150402.1150459

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. ACM SIGMOD Record, 30(2), 37–46. http://doi.org/10.1145/376284.375668

Ahmed, A., Korres, N. E., Ploennigs, J., Elhadi, H., & Menzel, K. (2011). Mining building performance data for energy-efficient operation. Advanced Engineering Informatics, 25(2), 341–354. http://doi.org/10.1016/j.aei.2010.10.002

Al-Hamadi, H. M., & Soliman, S. A. (2004). Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. Electric Power Systems Research, 68(1), 47–59. http://doi.org/10.1016/S0378-7796(03)00150-0

Ambrosone, G., Catalanotti, S., Matarazzo, M., & Vicari, L. (1933). A Dynamic Model for the Thermal Energy Management of Buildings. Applied Energy, 15, 285–297. http://doi.org/10.1016/0306-2619(83)90058-2

Amjady, N. (2001). Short-term hourly load forecasting using time-series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 16(4), 798–805. http://doi.org/10.1109/59.962429

Amjady, N., & Keynia, F. (2009). Short-term load forecasting of power systems by combination of wavelet transform and neuro-evolutionary algorithm. Energy, 34(1), 46–57. http://doi.org/10.1016/j.energy.2008.09.020

ASHRAE. (2017). Index | ASHRAE 201 Facility Smart Grid Information Model. Retrieved June 24, 2017, from http://spc201.ashraepcs.org/

Azadeh, M. A., & Sohrabkhani, S. (2006). Annual electricity consumption forecasting with Neural Network in high energy consuming industrial sectors of Iran. Proceedings of the IEEE International Conference on Industrial Technology, 49, 2166–2171. http://doi.org/10.1109/ICIT.2006.372572

Azimi, R., Ghofrani, M., & Ghayekhloo, M. (2016). A hybrid wind power forecasting model based on data mining and wavelets analysis. Energy Conversion and Management, 127, 208–225. http://doi.org/10.1016/j.enconman.2016.09.002

Bahrami, S., Hooshmand, R.-A., & Parastegari, M. (2014). Short term electric load forecasting by wavelet transform and grey model improved by PSO (particle swarm optimization) algorithm. Energy, 72(2014), 434–442. http://doi.org/10.1016/j.energy.2014.05.065

Bandaragoda, T. R., Ming Ting, K., Albrecht, D., Tony Liu, F., & Wells, J. R. (2014). Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble. IEEE Data Mining.

Begovich, O., & Valdovinos-Villalobos, G. (2010). DSP application of a water-leak detection and isolation algorithm. Program and Abstract Book - 2010 7th International Conference on Electrical Engineering, Computing Science and Automatic Control, CCE 2010, (Cce), 93–98. http://doi.org/10.1109/ICEEE.2010.5608570

Benaouda, D., & Murtagh, F. (2006). Electricity Load Forecast using Neural Network Trained from Wavelet-Transformed Data. 2006 IEEE International Conference on Engineering of Intelligent Systems, (1), 6–11. http://doi.org/10.1109/ICEIS.2006.1703163

Benaouda, D., Murtagh, F., Starck, J. L., & Renaud, O. (2006). Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting. Neurocomputing, 70(1–3), 139–154. http://doi.org/10.1016/j.neucom.2006.04.005

Beniger, J. R., Barnett, V., & Lewis, T. (1980). Outliers in Statistical Data. Contemporary Sociology, 9(4), 560. http://doi.org/10.2307/2066277

Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network Anomaly Detection: Methods, Systems and Tools. IEEE Communications Surveys & Tutorials, 16(1), 303–336. http://doi.org/10.1109/SURV.2013.052213.00046

Biswas, M. A. R., Robinson, M. D., & Fumo, N. (2016). Prediction of residential building energy consumption: A neural network approach. Energy, 117, 84–92. http://doi.org/10.1016/j.energy.2016.10.066

Cappers, P., Goldman, C., & Kathan, D. (2010). Demand response in U.S. electricity markets: Empirical evidence. Energy, 35(4), 1526–1535. http://doi.org/10.1016/j.energy.2009.06.029

Carrasquilla, U. (2010). Benchmarking Algorithms for Detecting Anomalies in Large Datasets. Review Literature And Arts Of The Americas, 1–16.

Catalão, J. P. S., Pousinho, H. M. I., & Mendes, V. M. F. (2009). Neural networks and wavelet transform for short-term electricity prices forecasting. 2009 15th International Conference on Intelligent System Applications to Power Systems, ISAP '09. http://doi.org/10.1109/ISAP.2009.5352834

Catalão, J. P. S., Pousinho, H. M. I., & Mendes, V. M. F. (2011). Hybrid wavelet-PSO-

ANFIS approach for short-term electricity prices forecasting. IEEE Transactions on Power Systems, 26(1), 137–144. http://doi.org/10.1109/TPWRS.2010.2049385

Charytoniuk, W., Chen, M. S., & Van Olinda, P. (1998). Nonparametric regression based short-term load forecasting. IEEE Transactions on Power Systems, 13(3), 725–730. http://doi.org/10.1109/59.708572

Chen, Y., Luh, P. B., Guan, C., Zhao, Y., Michel, L. D., Coolbeth, M. A., … Member, S. (2010). Short-Term Load Forecasting : Similar Day-Based Wavelet Neural Networks, 25(1), 322–330.

Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. IEEE Transactions on Power Systems, 21(2), 933–940. http://doi.org/10.1109/TPWRS.2006.873122

Chitsaz, H., Shaker, H., Zareipour, H., Wood, D., & Amjady, N. (2015). Short-term electricity load forecasting of buildings in microgrids. Energy and Buildings, 99, 50–60. http://doi.org/10.1016/j.enbuild.2015.04.011

Chou, J., Telaga, A. S., Chong, W. K., & Jr, G. E. G. (2017). Early-warning application for real-time detection of energy consumption anomalies in buildings. Journal of Cleaner Production, 149(March), 711–722. http://doi.org/10.1016/j.jclepro.2017.02.028

Citroen, N., & Ouassaid, M. (2015). Moroccan Long Term Electricity Demand Forecasting Using Wavelet Neural Networks.

Conejo, A. J., Plazas, M. A., Espínola, R., Member, S., & Molina, A. B. (2005). Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models. IEEE Transactions On Power Systems, 20, 1035–1042. http://doi.org/10.1109/TPWRS.2005.846054

Crawley, D. B., Lawrie, L. K., Winkelmann, F. C., Buhl, W. F., Huang, Y. J., Pedersen, C. O., … Glazer, J. (2001). EnergyPlus: Creating a new-generation building energy simulation program. Energy and Buildings, 33(4), 319–331. http://doi.org/10.1016/S0378-7788(00)00114-6

Cugliari, J., Goude, Y., & Poggi, J. M. (2016). Disaggregated electricity forecasting using wavelet-based clustering of individual consumers. 2016 IEEE International Energy Conference, Energycon 2016. http://doi.org/10.1109/ENERGYCON.2016.7514087

de Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, 41, 40–49. http://doi.org/10.1016/j.autcon.2014.02.009

De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, 41, 40–49. http://doi.org/10.1016/j.autcon.2014.02.009

Deb, C., Eang, L. S., Yang, J., & Santamouris, M. (2016). Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. Energy and Buildings, 121, 284–297. http://doi.org/10.1016/j.enbuild.2015.12.050

Deguen, S., & Zmirou-Navier, D. (2010). Social inequalities resulting from health risks related to ambient air quality--A European review. European Journal of Public Health, 20(1), 27–35. http://doi.org/10.1093/eurpub/ckp220

Deo, R. C., Wen, X., & Qi, F. (2016). A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Applied Energy, 168, 568–593. http://doi.org/10.1016/j.apenergy.2016.01.130

Dincer, I., & Dost, S. (1997). Energy and GDP. International Journal of Energy Research, 21(2), 153–167. http://doi.org/10.1002/(SICI)1099-114X(199702)21:2<153::AID-ER227>3.0.CO;2-Z

Djuric, N., & Novakovic, V. (2012). Identifying important variables of energy use in low energy office building by using multivariate analysis. Energy and Buildings, 45, 91–98. http://doi.org/10.1016/j.enbuild.2011.10.031

Dodoo, A., Gustavsson, L., & Sathre, R. (2011). Building energy-efficiency standards in a life cycle primary energy perspective. Energy and Buildings, 43(7), 1589–1597. http://doi.org/10.1016/j.enbuild.2011.03.002

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5), 545–553. http://doi.org/10.1016/j.enbuild.2004.09.009

Dong, B., Oneill, Z., Luo, D., & Bailey, T. (2014). Development and calibration of an online energy model for campus buildings. Energy and Buildings, 76, 316–327. http://doi.org/10.1016/j.enbuild.2014.02.064

Dudek, G. (2016). Pattern-based local linear regression models for short-term load forecasting. Electric Power Systems Research, 130, 139–147. http://doi.org/10.1016/j.epsr.2015.09.001

Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings, 49, 591–603. http://doi.org/10.1016/j.enbuild.2012.03.010

EIA. (2016). Arizona State Energy Profile. Retrieved June 17, 2017, from https://www.eia.gov/state/print.php?sid=AZ

EIA. (2017). What is U.S. electricity generation by energy source? - FAQ - U.S. Energy

Information Administration (EIA). Retrieved May 2, 2017, from https://www.eia.gov/tools/faqs/faq.php?id=87&t=1

Energy Information Administration (EIA)- About the Commercial Buildings Energy Consumption Survey (CBECS). (2013). Retrieved May 15, 2015, from http://www.eia.gov/consumption/commercial/about.cfm

Espinoza, M., Joye, C., Belmans, R., & De Moor, B. (2005). Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. IEEE Transactions on Power Systems, 20(3), 1622–1630. http://doi.org/10.1109/TPWRS.2005.852123

Eynard, J., Grieu, S., & Polit, M. (2011). Wavelet-based multi-resolution analysis and artificial neural networks for forecasting temperature and thermal power consumption. Engineering Applications of Artificial Intelligence, 24, 501–516. http://doi.org/10.1016/j.engappai.2010.09.003

Fan, S., & Hyndman, R. J. (2012). Short-Term Load Forecasting Based on a Semi-Parametric Additive Model. Ieee Transactions on Power Systems, 27(1), 134–141. http://doi.org/Doi 10.1109/Tpwrs.2011.2162082

Fawkes, S. (1987). Soft-systems model of energy management and checklists for energy managers. Applied Energy, 27(3), 229–241. http://doi.org/10.1016/0306-2619(87)90028-6

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005a). An electric energy consumer characterization framework based on data mining techniques. IEEE Transactions on Power Systems, 20(2), 596–602. http://doi.org/10.1109/TPWRS.2005.846234

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005b). An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. IEEE Transactions on Power Systems. http://doi.org/10.1109/TPWRS.2005.846234

Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews, 23, 272–288. http://doi.org/10.1016/j.rser.2013.03.004

Frimpong, E. A., & Okyere, P. Y. (2010). Monthly Energy Consumption Forecasting Using Wavelet Analysis and Radial Basis Funtion Neural Network, 30(2), 157–164.

Fumo, N., Mago, P., & Luck, R. (2010). Methodology to estimate building energy consumption using EnergyPlus Benchmark Models. Energy and Buildings, 42(12), 2331–2337. http://doi.org/10.1016/j.enbuild.2010.07.027

Gee, G. C., & Payne-Sturges, D. C. (2004). Environmental health disparities: A framework integrating psychosocial and environmental concepts. Environmental Health Perspectives. http://doi.org/10.1289/ehp.7074

Georgescu, M., & Mezić, I. (2015). Building energy modeling: A systematic approach to zoning and model reduction using Koopman Mode Analysis. Energy and Buildings, 86, 794–802. http://doi.org/10.1016/j.enbuild.2014.10.046

Ghelardoni, L., Ghio, A., & Anguita, D. (2013). Energy load forecasting using empirical mode decomposition and support vector regression. IEEE Transactions on Smart Grid, 4(1), 549–556. http://doi.org/10.1109/TSG.2012.2235089

Ghiassi, M., Zimbra, D. K., & Saidane, H. (2006). Medium term system load forecasting with a dynamic artificial neural network model. Electric Power Systems Research, 76(5), 302–316. http://doi.org/10.1016/j.epsr.2005.06.010

Ghiaus, C. (2006). Experimental estimation of building energy performance by robust regression. Energy and Buildings, 38(6), 582–587. http://doi.org/10.1016/j.enbuild.2005.08.014

Ghosh, D., & Vogt, A. (2012). Outliers: An Evaluation of Methodologies. Joint Statistical Metings, 3455–3460.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human Semi-Supervised Learning. Topics in Cognitive Science, 5(1), 132–172. http://doi.org/10.1111/tops.12010

Gogoi, P., Bhattacharyya, D. K., Borah, B., & Kalita, J. K. (2011). A survey of outlier detection methods in network anomaly identification. Computer Journal, 54(4), 570–588. http://doi.org/10.1093/comjnl/bxr026

Goldberg, A. B., Zhu, X., Furger, A., & Xu, J. (2011). OASIS : Online Active SemI-Supervised Learning. In Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI) (pp. 362–367).

Gross, G., & Galiana, F. D. (1987). Short-term load forecasting. Proceedings of the IEEE, 75(12), 1558–1573. http://doi.org/10.1109/PROC.1987.13927

Hagan, M. T., & Behr, S. M. (1987). The Time Series Approach to Short Term Load Forecasting. IEEE Transactions on Power Systems, 2(3), 785–791. http://doi.org/10.1109/TPWRS.1987.4335210

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. Annals of Physics, 54, 770. http://doi.org/10.5860/CHOICE.49-3305

Harish, V. S. K. V, & Kumar, A. (2016). A review on modeling and simulation of building energy systems. Renewable and Sustainable Energy Reviews, 56, 1272–1292. http://doi.org/10.1016/j.rser.2015.12.040

He, Y., Liu, R., Li, H., Wang, S., & Lu, X. (2017). Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. Applied Energy, 185, 254–266. http://doi.org/10.1016/j.apenergy.2016.10.079

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review. http://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hong, T., Wilson, J., & Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. IEEE Transactions on Smart Grid, 5(1), 456–462. http://doi.org/10.1109/TSG.2013.2274373

Hrasko, R., Pacheco, A. G. C., & Krohling, R. A. (2015). Time Series Prediction Using Restricted Boltzmann Machines and Backpropagation. Procedia Computer Science, 55, 990–999. http://doi.org/10.1016/j.procs.2015.07.104

Hu, R., Wen, S., Zeng, Z., & Huang, T. (2017). A short-term power load forecasting model based on the generalized regression neural network with decreasing step fruit fly optimization algorithm. Neurocomputing, 221(September 2016), 24–31. http://doi.org/10.1016/j.neucom.2016.09.027

Ibrahim, E. S. (2000). Management of loss reduction projects for power distribution systems. Electric Power Systems Research, 55(1), 49–56. http://doi.org/10.1016/S0378-7796(99)00073-5

Iec. (2007). Efficient electrical energy transmission and distribution, 24. Retrieved from http://www.iec.ch/about/brochures/pdf/technology/transmission.pdf

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys. http://doi.org/10.1145/331499.331504

Jebaraj, S., & Iniyan, S. (2006). A review of energy models. Renewable and Sustainable Energy Reviews, 10(4), 281–311. http://doi.org/10.1016/j.rser.2004.09.004

Kafaie, S., Kashefi, O., & Sharifi, M. (2011). Energy Transformed: Sustainable Energy solutions for climate change mitigation (Vol. 3).

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, a., Stevanovic, Z., & Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. Building and Environment, 45(7), 1683–1697. http://doi.org/10.1016/j.buildenv.2010.01.021

Kelso, J. (2012). Energy Efficiency and Renewable Energy databook. Silver Spring: Department of Energy. Retrieved from buildingsdatabook.eere.energy.gov

Khoa, T. Q. D., Phuong, L. M., Binh, P. T. T., & Lien, N. T. H. (2004). Application of wavelet and neural network to long-term load forecasting. Power System

Technology, 2004. PowerCon 2004. 2004 International Conference on, 1(November), 840–844 Vol.1. http://doi.org/10.1109/ICPST.2004.1460110

Kim, C., Yu, I., & Song, Y. (2002). Kohonen neural network and wavelet transform based approach to short-term load forecasting. Electric Power Systems Research, 63(3), 169–176. http://doi.org/10.1016/S0378-7796(02)00097-4

Koopmans, C. C., & te Velde, D. W. (2001). Bridging the energy efficiency gap: using bottom-up information in a top-down energy demand model. Energy Economics, 23(1), 57–75. http://doi.org/10.1016/S0140-9883(00)00054-2

Kumar, U., & Jain, V. K. (2010). Time series models (Grey-Markov, Grey Model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India. Energy, 35(4), 1709–1716. http://doi.org/10.1016/j.energy.2009.12.021

Lai, Y. (1998). Analytic signals and the transition to chaos in deterministic flows. Physical Review E, 58(6), 6911–6914. http://doi.org/10.1103/PhysRevE.58.R6911

Landsberg, P. T. (1976). A simple model for solar energy economics in the U.K. Energy, 2, 149–159.

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 161(2), 121–160. http://doi.org/10.1111/1467-985X.00094

Lanzisera, S., Dawson-Haggerty, S., Cheung, H. Y. I., Taneja, J., Culler, D., & Brown, R. (2013). Methods for detailed energy data collection of miscellaneous and electronic loads in a commercial office building. Building and Environment, 65, 170–177. http://doi.org/10.1016/j.buildenv.2013.03.025

Le Cam, M., Daoud, A., & Zmeureanu, R. (2016). Forecasting electric demand of supply fan using data mining techniques. Energy, 101, 541–557. http://doi.org/10.1016/j.energy.2016.02.061

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. http://doi.org/10.1038/nature14539

Lee, D. H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. ICML Workshop on Challenges in Representation Learning. Retrieved from https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf%5Cnpapers3://publication/uuid/FAC81169-4273-4668-90FA-2B849AAA67C1

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 86(10), 2249–2256. http://doi.org/10.1016/j.apenergy.2008.11.035

Li, X., Tan, H., & Rackes, A. (2015). Carbon footprint analysis of student behavior for a sustainable university campus in China. Journal of Cleaner Production, 106, 97–108. http://doi.org/10.1016/j.jclepro.2014.11.084

Li, X., Zhang, Y., & Cai, L. (2008). Electrical Load Forecasting Based on Fuzzy Wavelet Neural Networks. 2008 International Seminar on Future BioMedical Information Engineering, 122–125. http://doi.org/10.1109/FBIE.2008.67

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2010). Can Isolation-Based Anomaly Detectors Handle Arbitrary Multi-modal Patterns in Data?

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings - IEEE International Conference on Data Mining, ICDM (pp. 413–422). http://doi.org/10.1109/ICDM.2008.17

Liu, W., & Hwang, I. (2011). Robust estimation and fault detection and isolation algorithms for stochastic linear hybrid systems with unknown fault input. IET Control Theory &amp; Applications, 5(12), 1353–1368. http://doi.org/10.1049/iet-cta.2010.0287

Majcen, D., Itard, L. C. M., & Visscher, H. (2013). Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications. Energy Policy, 54, 125–136. http://doi.org/10.1016/j.enpol.2012.11.008

Mann, G. S., & McCallum, A. (2007). Simple, Robust, Scalable Semi-Supervised Learning via Expectation Regularization. Proceedings of the 24th International Conference on Machine Learning - ICML ’07, 593–600. http://doi.org/10.1145/1273496.1273571

Mocanu, E., Nguyen, P. H., Gibescu, M., & Kling, W. L. (2016a). Deep learning for estimating building energy consumption. Sustainable Energy, Grids and Networks, 6, 91–99. http://doi.org/10.1016/j.segan.2016.02.005

Mocanu, E., Nguyen, P. H., Gibescu, M., & Kling, W. L. (2016b). Deep learning for estimating building energy consumption. Sustainable Energy, Grids and Networks, 6, 91–99. http://doi.org/10.1016/j.segan.2016.02.005

Mocanu, E., Nguyen, P. H., Kling, W. L., & Gibescu, M. (2016). Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning. Energy and Buildings, 116, 646–655. http://doi.org/10.1016/j.enbuild.2016.01.030

Monteiro, C., Ramirez-Rosado, I., Fernandez-Jimenez, L., & Conde, P. (2016). Short-

Term Price Forecasting Models Based on Artificial Neural Networks for Intraday Sessions in the Iberian Electricity Market. Energies, 9(9), 721. http://doi.org/10.3390/en9090721

Moreno-Chaparro, C., Salcedo-Lagos, J., Rivas, E., & Canon, A. O. (2012). State of the art of electricity demand forecasting based on wavelet analisys and a nonlinear autoregressive model NAR, 1–6. http://doi.org/10.1109/WEA.2012.6220078

Mourad, M., Bouzid, B., & Mohamed, B. (2012). A hybrid wavelet transform and ANFIS model for short term electric load prediction. 2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), (1), 292–295. http://doi.org/10.1109/ICTEA.2012.6462886

N. E. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, … H. Liu. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 454(1971), 903–995. http://doi.org/10.1098/rspa.1998.0193

Naganathan, H., Chong, W. O., & Chen, X. (2016). Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. Automation in Construction. http://doi.org/10.1016/j.autcon.2016.08.002

Naganathan, H., Seshasayee, S. P., Kim, J., Chong, W. K., & Chou, J.-S. (2016). Wildfire Predictions: Determining Reliable Models using Fused Dataset. Global Journal of Computer Science and Technology: C Software & Data Engineering, 16(4), 4–11. Retrieved from https://globaljournals.org/GJCST_Volume16/6-Wildfire-Predictions-Determining.pdf

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohamad, M. (2010). Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Transactions on Power Delivery, 25(2), 1162–1171. http://doi.org/10.1109/TPWRD.2009.2030890

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1. http://doi.org/10.1186/s40537-014-0007-7

Newsham, G. R. (2009). Post-occupancy evaluation of energy and indoor environment quality in green buildings : a review. 3rd International Conference on Smart and Sustainable Built Environments, 1–7. Retrieved from http://www.sasbe2009.com/proceedings/documents/SASBE2009_paper_post-occupancy_evaluation_of_energy_and_indoor_environment_quality_in_green_buildings_-_a_review.pdf

Nguyen, H. T., & Nabney, I. T. (2010). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. Energy, 35(9), 3674–3685. http://doi.org/10.1016/j.energy.2010.05.013

Nizar, A. H., Dong, Z. Y., & Zhao, J. H. (2006). Load profiling and data mining techniques in electricity deregulated market. 2006 IEEE Power Engineering Society General Meeting. http://doi.org/10.1109/PES.2006.1709335

Norford, L. K., & Leeb, S. B. (1996). Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms. Energy and Buildings, 24(1), 51–64. http://doi.org/10.1016/0378-7788(95)00958-2

Nunes, J. C., Bouaoune, Y., Delechelle, E., Niang, O., & Bunel, P. (2003). Image analysis by bidimensional empirical mode decomposition. Image and Vision Computing, 21(12), 1019–1026. http://doi.org/10.1016/S0262-8856(03)00094-5

O'Neill, Z., Eisenhower, B., Yuan, S., Bailey, T., Narayanan, S., & Fonoberov, V. (2011). Modeling and calibration of energy models for a DoD building. ASHRAE Transactions, 117(860), 358–365.

Osório, G. J., Matias, J. C. O., & Catalão, J. P. S. (2015). Short-term wind power forecasting using adaptive neuro-fuzzy inference system combined with evolutionary particle swarm optimization, wavelet transform and mutual information. Renewable Energy, 75, 301–307. http://doi.org/10.1016/j.renene.2014.09.058

Pandey, A. S., Singh, D., & Sinha, S. K. (2010). Intelligent hybrid wavelet models for short-term load forecasting. IEEE Transactions on Power Systems, 25(3), 1266–1273. http://doi.org/10.1109/TPWRS.2010.2042471

Pao, H. T. (2009). Forecasting energy consumption in Taiwan using hybrid nonlinear models. Energy, 34(10), 1438–1446. http://doi.org/10.1016/j.energy.2009.04.026

Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. IEEE Transactions on Power Systems, 5(4), 1535–1547. http://doi.org/10.1109/59.99410

Pérez-Lombard, L., Ortiz, J., Coronel, J. F., & Maestre, I. R. (2011). A review of HVAC systems requirements in building energy regulations. Energy and Buildings, 43(2–3), 255–268. http://doi.org/10.1016/j.enbuild.2010.10.025

Perez-Lombard, L., Ortiz, J., & Maestre, I. R. (2011). The map of energy flow in HVAC systems. Applied Energy, 88(12), 5020–5031. http://doi.org/10.1016/j.apenergy.2011.07.003

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. Energy and Buildings, 40(3), 394–398.

http://doi.org/10.1016/j.enbuild.2007.03.007

Petersen, S., & Svendsen, S. (2011). Method for simulating predictive control of building systems operation in the early stages of building design. Applied Energy, 88(12), 4597–4606. http://doi.org/10.1016/j.apenergy.2011.05.053

Pitt, B. D., & Kitschen, D. S. (1999). Application of data mining techniques to load profiling. Proceedings of the 21st International Conference on Power Industry Computer Applications Connecting Utilities PICA 99 To the Millennium and Beyond Cat No99CH36351, 131–136. http://doi.org/10.1109/PICA.1999.779395

Platon, R., Dehkordi, V. R., & Martel, J. (2015). Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. Energy and Buildings, 92, 10–18. http://doi.org/10.1016/j.enbuild.2015.01.047

Province, S. (2011). A Fault Detection and Isolation Algorithm for Distribution Systems Containing Distributed Generations, 1(2), 1753–1756.

Raftery, P., Keane, M., & Costa, A. (2011). Calibrating whole building energy models: Detailed case study using hourly measured data. Energy and Buildings, 43(12), 3666–3679. http://doi.org/10.1016/j.enbuild.2011.09.039

Raftery, P., Keane, M., & O'Donnell, J. (2011). Calibrating whole building energy models: An evidence-based methodology. Energy and Buildings, 43(9), 2356–2364. http://doi.org/10.1016/j.enbuild.2011.05.020

Rana, M., & Koprinska, I. (2016). Forecasting electricity load with advanced wavelet neural networks. Neurocomputing, 182, 118–132. http://doi.org/10.1016/j.neucom.2015.12.004

Ren, H., Wang, Y.-L., Huang, M.-Y., Chang, Y.-L., & Kao, H.-M. (2014). Ensemble Empirical Mode Decomposition Parameters Optimization for Spectral Distance Measurement in Hyperspectral Remote Sensing Data. Remote Sensing, 6(3), 2069–2083. http://doi.org/10.3390/rs6032069

Rodrigues, F., Duarte, J., Figueiredo, V., Vale, Z., & Cordeiro, M. (2003). A Comparative Analysis of Clustering Algorithms Applied to Load Profiling. In Machine Learning and Data Mining in Pattern Recognition (pp. 73–85). http://doi.org/10.1007/3-540-45065-3_7

Ryan, E. M., & Sanquist, T. F. (2012). Validation of building energy modeling tools under idealized and realistic conditions. Energy and Buildings, 47, 375–382. http://doi.org/10.1016/j.enbuild.2011.12.020

Ryghaug, M., & Sørensen, K. H. (2009). How energy efficiency fails in the building industry. Energy Policy, 37(3), 984–991. http://doi.org/10.1016/j.enpol.2008.11.001

Sartori, I., Napolitano, A., & Voss, K. (2012). Net zero energy buildings: A consistent definition framework. Energy and Buildings, 48, 220–232. http://doi.org/10.1016/j.enbuild.2012.01.032

Schlueter, A., & Thesseling, F. (2009). Building information model based energy/exergy performance assessment in early design stages. Automation in Construction, 18(2), 153–163. http://doi.org/10.1016/j.autcon.2008.07.003

Scofield, J. H. (2009). Do LEED-certified buildings save energy? Not really... Energy and Buildings, 41(12), 1386–1390. http://doi.org/10.1016/j.enbuild.2009.08.006

Silva, D. De, & Yu, X. (2011). A data mining framework for electricity consumption analysis from meter data. IEEE Transactions on Industrial Informatics, 7(3), 399–407. http://doi.org/10.1109/TII.2011.2158844

Sinha, N., Lai, L. L., Ghosh, P. K., & Ma, Y. (2007). Wavelet-GA-ANN based hybrid model for accurate prediction of short-term load forecast. 2007 International Conference on Intelligent Systems Applications to Power Systems, ISAP. http://doi.org/10.1109/ISAP.2007.4441661

Song, K.-B., Baek, Y.-S., Hong, D. H., & Jang, G. (2005). Short-Term Load Forecasting for the Holidays Using Fuzzy Linear Regression Method. IEEE Transactions on Power Systems, 20(1), 96–101. http://doi.org/10.1109/TPWRS.2004.835632

Sudheer, G., & Suseelatha, a. (2015). Short term load forecasting using wavelet transform combined with Holt–Winters and weighted nearest neighbor models. International Journal of Electrical Power & Energy Systems, 64, 340–346. http://doi.org/10.1016/j.ijepes.2014.07.043

Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting - A review. Renewable and Sustainable Energy Reviews. http://doi.org/10.1016/j.rser.2011.08.014

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews. http://doi.org/10.1016/j.rser.2008.09.033

Tan, Z., Zhang, J., Wang, J., & Xu, J. (2010). Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. Applied Energy, 87, 3606–3610. http://doi.org/10.1016/j.apenergy.2010.05.012

Tang, J.-P., Chiou, D.-J., Chen, C.-W., Chiang, W.-L., Hsu, W.-K., Chen, C.-Y., & Liu, T.-Y. (2010). A case study of damage detection in benchmark buildings using a Hilbert-Huang Transform-based method. Journal of Vibration and Control, 17(4), 623–636. http://doi.org/10.1177/1077546309360053

Tang, L., Yu, L., Wang, S., Li, J., & Wang, S. (2012). A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. Applied Energy, 93, 432–443. http://doi.org/10.1016/j.apenergy.2011.12.030

Tarca, A. L., Carey, V. J., Chen, X., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. PLoS Computational Biology. http://doi.org/10.1371/journal.pcbi.0030116

Toftum, J., Andersen, R. V., & Jensen, K. L. (2009). Occupant performance and building energy consumption with different philosophies of determining acceptable thermal conditions. Building and Environment, 44(10), 2009–2016. http://doi.org/10.1016/j.buildenv.2009.02.007

Torcellini, P., Pless, S., & Deru, M. (2006). Zero Energy Buildings : A Critical Look at the Definition. In ACEE Summer study. http://doi.org/10.1016/S1471-0846(02)80045-2

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32, 1761–1768. http://doi.org/10.1016/j.energy.2006.11.010

Vasan, A., & Sivasubramaniam, A. (2015). Energy disaggregation analysis of a supermarket chain using a. Energy & Buildings, 97, 65–76. http://doi.org/10.1016/j.enbuild.2015.03.053

Vengertsev, D., & Thakkar, H. (2005). Anomaly Detection in Graph : Unsupervised Learning , Graph-based Features and Deep Architecture.

Vu, D. H. (2014). Combinatorial approach using wavelet analysis and artificial neural network for short-term load forecasting, (Aupec), 1–6. http://doi.org/10.1109/AUPEC.2014.6966607

Wang, J., Wang, J., Li, Y., Zhu, S., & Zhao, J. (2014). Techniques of applying wavelet de-noising into a combined model for short-term load forecasting. International Journal of Electrical Power and Energy Systems, 62, 816–824. http://doi.org/10.1016/j.ijepes.2014.05.038

Wang, J., & Zhu, Q. (2015). Short-term Electricity Load Forecast Performance Comparison Based on Four Neural Network Models. The 27th Chinese Control and Decision Conference (2015 CCDC), 2928–2932.

Wang, L., Mathew, P., & Pang, X. (2012). Uncertainties in energy consumption introduced by building operations and weather for a medium-size office building. Energy and Buildings, 53, 152–158. http://doi.org/10.1016/j.enbuild.2012.06.017

Wang, Z., Yang, F., Ho, D. W. C., Swift, S., Tucker, a, & Liu, X. (2008). Stochastic dynamic modeling of short gene expression time-series data. IEEE Transactions on Nanobioscience, 7(1), 44–55. http://doi.org/10.1109/TNB.2008.2000149

Weston, J., Ratle, F., Mobahi, H., & Collobert, R. (2012). Deep learning via semi-supervised embedding. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7700 LECTU, 639–655. http://doi.org/10.1007/978-3-642-35289-8-34

Wood, G., & Newborough, M. (2003). Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design. Energy and Buildings, 35(8), 821–841. http://doi.org/10.1016/S0378-7788(02)00241-4

Wu, Zhaohua and Huang, N. E. (2009). Ensemble empirical mode decomposition for high frequency ECG noise reduction. Advances in Adaptive Data Analysis, 55(4), 193–201. http://doi.org/10.1515/BMT.2010.030

Wu, J., Wang, J., Lu, H., Dong, Y., & Lu, X. (2013). Short term load forecasting technique based on the seasonal exponential adjustment method and the regression model. Energy Conversion and Management, 70, 1–9. http://doi.org/10.1016/j.enconman.2013.02.010

Wu, Z., & Huang, N. E. (2009). Ensemble Empirical Mode Decomposition : A Noise Assisted Data Analysis Method. Advances in Adaptive Data Analysis, 1(1), 1–41. http://doi.org/10.1142/S1793536909000047

Xia, C., Wang, J., & McMenemy, K. (2010). Short, medium and long term load forecasting model and virtual load forecaster based on radial basis function neural networks. International Journal of Electrical Power and Energy Systems, 32(7), 743–750. http://doi.org/10.1016/j.ijepes.2010.01.009

Xu, H., & Niimura, T. (2004). Short-term electricity price modeling and forecasting using wavelets and multivariate time series. IEEE PES Power Systems Conference and Exposition, 2004., 858–862. http://doi.org/10.1109/PSCE.2004.1397570

Yalcinoz, T., & Eminoglu, U. (2005). Short term and medium term power distribution load forecasting by neural networks. Energy Conversion and Management. http://doi.org/10.1016/j.enconman.2004.07.005

Yalçınkaya, T., & Lai, Y.-C. (1997). Phase Characterization of Chaos. Physical Review Letters, 79(20), 3885–3888. http://doi.org/10.1103/PhysRevLett.79.3885

Yao, S. ., Song, Y. ., Zhang, L. ., & Cheng, X. . (2000). Wavelet transform and neural networks for short-term electrical load forecasting. Energy Conversion and Management, 41(18), 1975–1988. http://doi.org/10.1016/S0196-8904(00)00035-2

Yokoyama, R., Wakui, T., & Satake, R. (2009). Prediction of energy demands using neural network with model identification by global optimization. Energy Conversion and Management, 50(2), 319–327. http://doi.org/10.1016/j.enconman.2008.09.017

Zhang, B.-L., & Dong, Z.-Y. (2001). An adaptive neural-wavelet model for short term load forecasting. Electric Power Systems Research, 59, 121–129. http://doi.org/10.1016/S0378-7796(01)00138-9

Zhang, J., & Tan, Z. (2013). Day-ahead electricity price forecasting using WT, CLSSVM and EGARCH model. International Journal of Electrical Power and Energy Systems, 45, 362–368. http://doi.org/10.1016/j.ijepes.2012.09.007

Zhang, P., & Wang, H. (2012). Fuzzy Wavelet Neural Networks for City Electric Energy Consumption Forecasting. Energy Procedia. http://doi.org/10.1016/j.egypro.2012.02.248

Zhao, H., Liu, R., Zhao, Z., & Fan, C. (2011). Analysis of Energy Consumption Prediction Model Based on Genetic Algorithm and Wavelet Neural Network. 2011 3rd International Workshop on Intelligent Systems and Applications, 1–4. http://doi.org/10.1109/ISA.2011.5873468

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. Machine Learning-International Workshop Then Conference-, 20(2), 912. http://doi.org/10.1.1.5.68

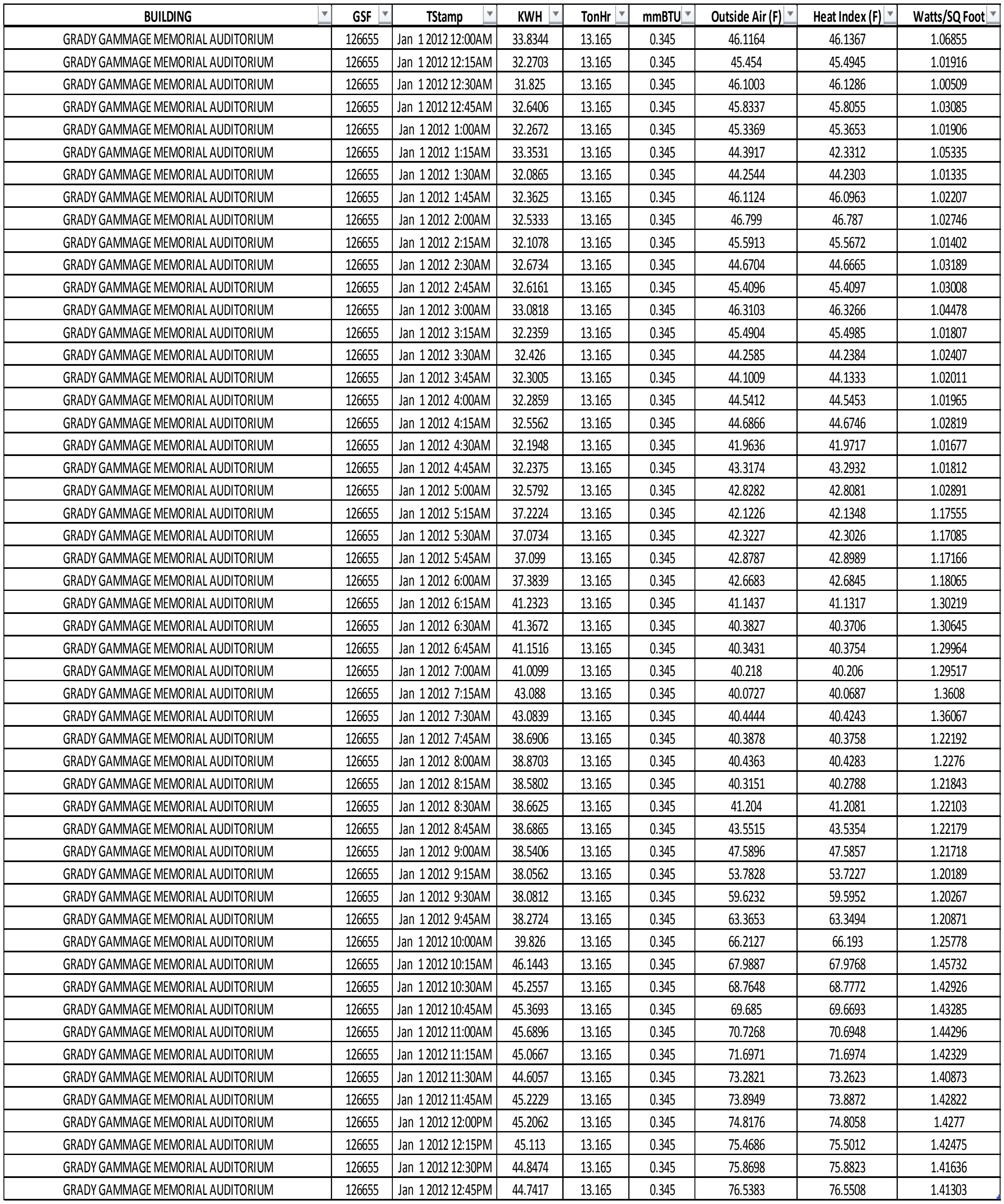
Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. http://doi.org/10.2200/S00196ED1V01Y200906AIM006

Zhu, Y. (2006). Applying computer-based simulation to energy auditing: A case study. Energy and Buildings, 38(5), 421–428. http://doi.org/10.1016/j.enbuild.2005.07.007

APPENDIX A

CHAPTER 2 SAMPLE DATA COLLECTED FROM EIS

A.1 sample data (15 minutes interval)



BUILDING	GSF	TStamp	KWH	TonHr	mmBTU	Outside Air (F)	Heat Index (F)	Watts/SQ Foot
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:00AM	33.8344	13.165	0.345	46.1164	46.1367	1.06855
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:15AM	32.2703	13.165	0.345	45.454	45.4945	1.01916
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:30AM	31.825	13.165	0.345	46.1003	46.1286	1.00509
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:45AM	32.6406	13.165	0.345	45.8337	45.8055	1.03085
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 1:00AM	32.2672	13.165	0.345	45.3369	45.3653	1.01906
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 1:15AM	33.3531	13.165	0.345	44.3917	42.3312	1.05335
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 1:30AM	32.0865	13.165	0.345	44.2544	44.2303	1.01335
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 1:45AM	32.3625	13.165	0.345	46.1124	46.0963	1.02207
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 2:00AM	32.5333	13.165	0.345	46.799	46.787	1.02746
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 2:15AM	32.1078	13.165	0.345	45.5913	45.5672	1.01402
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 2:30AM	32.6734	13.165	0.345	44.6704	44.6665	1.03189
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 2:45AM	32.6161	13.165	0.345	45.4096	45.4097	1.03008
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 3:00AM	33.0818	13.165	0.345	46.3103	46.3266	1.04478
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 3:15AM	32.2359	13.165	0.345	45.4904	45.4985	1.01807
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 3:30AM	32.426	13.165	0.345	44.2585	44.2384	1.02407
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 3:45AM	32.3005	13.165	0.345	44.1009	44.1333	1.02011
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 4:00AM	32.2859	13.165	0.345	44.5412	44.5453	1.01965
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 4:15AM	32.5562	13.165	0.345	44.6866	44.6746	1.02819
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 4:30AM	32.1948	13.165	0.345	41.9636	41.9717	1.01677
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 4:45AM	32.2375	13.165	0.345	43.3174	43.2932	1.01812
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 5:00AM	32.5792	13.165	0.345	42.8282	42.8081	1.02891
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 5:15AM	37.2224	13.165	0.345	42.1226	42.1348	1.17555
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 5:30AM	37.0734	13.165	0.345	42.3227	42.3026	1.17085
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 5:45AM	37.099	13.165	0.345	42.8787	42.8989	1.17166
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 6:00AM	37.3839	13.165	0.345	42.6683	42.6845	1.18065
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 6:15AM	41.2323	13.165	0.345	41.1437	41.1317	1.30219
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 6:30AM	41.3672	13.165	0.345	40.3827	40.3706	1.30645
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 6:45AM	41.1516	13.165	0.345	40.3431	40.3754	1.29964
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 7:00AM	41.0099	13.165	0.345	40.218	40.206	1.29517
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 7:15AM	43.088	13.165	0.345	40.0727	40.0687	1.3608
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 7:30AM	43.0839	13.165	0.345	40.4444	40.4243	1.36067
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 7:45AM	38.6906	13.165	0.345	40.3878	40.3758	1.22192
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 8:00AM	38.8703	13.165	0.345	40.4363	40.4283	1.2276
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 8:15AM	38.5802	13.165	0.345	40.3151	40.2788	1.21843
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 8:30AM	38.6625	13.165	0.345	41.204	41.2081	1.22103
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 8:45AM	38.6865	13.165	0.345	43.5515	43.5354	1.22179
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 9:00AM	38.5406	13.165	0.345	47.5896	47.5857	1.21718
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 9:15AM	38.0562	13.165	0.345	53.7828	53.7227	1.20189
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 9:30AM	38.0812	13.165	0.345	59.6232	59.5952	1.20267
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 9:45AM	38.2724	13.165	0.345	63.3653	63.3494	1.20871
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 10:00AM	39.826	13.165	0.345	66.2127	66.193	1.25778
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 10:15AM	46.1443	13.165	0.345	67.9887	67.9768	1.45732
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 10:30AM	45.2557	13.165	0.345	68.7648	68.7772	1.42926
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 10:45AM	45.3693	13.165	0.345	69.685	69.6693	1.43285
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 11:00AM	45.6896	13.165	0.345	70.7268	70.6948	1.44296
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 11:15AM	45.0667	13.165	0.345	71.6971	71.6974	1.42329
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 11:30AM	44.6057	13.165	0.345	73.2821	73.2623	1.40873
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 11:45AM	45.2229	13.165	0.345	73.8949	73.8872	1.42822
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:00PM	45.2062	13.165	0.345	74.8176	74.8058	1.4277
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:15PM	45.113	13.165	0.345	75.4686	75.5012	1.42475
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:30PM	44.8474	13.165	0.345	75.8698	75.8823	1.41636
GRADY GAMMAGE MEMORIAL AUDITORIUM	126655	Jan 1 2012 12:45PM	44.7417	13.165	0.345	76.5383	76.5508	1.41303

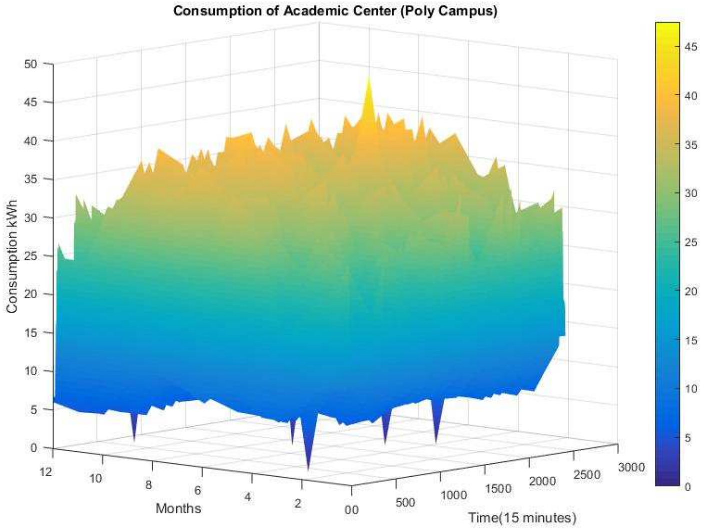
A.2 Cluster 1 Sample data

Time stamp	Buildings																		Supply kW	Demand	Supply kWh	Loss kWh
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18				
Jan 1 2013 12:00AM	3.4786	83.176	66.0401	69.7684	11.0688	49.6595	66.7984	159.463	19.129	26.4315	39.4892	24.9495	2.31927	27.057	17.6745	0.266433	33.4021	13.9917	8146.29	714.163003	2036.5725	1322.409497
Jan 1 2013 12:15AM	3.49423	83.2433	66.6594	67.9738	10.748	48.8634	64.8583	158.319	18.8869	26.5541	40.4561	24.1802	2.42135	27.3367	17.5682	0.28362	33.5427	14.1349	8516.57	709.5242	2129.1425	1419.6183
Jan 1 2013 12:30AM	3.51195	81.6773	65.3057	68.7427	10.9772	49.4316	65.0802	159.445	18.9728	26.3077	40.0062	24.3062	2.36042	27.0297	17.7354	0.259922	32.2391	14.513	8516.57	707.902092	2129.1425	1421.240408
Jan 1 2013 12:45AM	3.5307	82.3779	65.4589	70.0535	11.7772	49.2452	65.2089	162.23	23.6448	26.4002	40.673	24.2786	2.43385	27.0961	17.5917	0.279715	33.4875	13.9672	8516.57	719.734965	2129.1425	1409.407535
Jan 1 2013 1:00AM	3.47967	82.1785	64.325	70.2509	11.0105	48.8318	67.0698	161.718	40.1807	26.7688	40.8232	24.2005	2.30469	27.0141	17.4828	0.256276	33.5474	15.1589	8640	736.601536	2160	1423.398464
Jan 1 2013 1:15AM	3.54113	81.6125	65.374	68.9972	10.7064	49.3572	65.3495	159.031	57.4594	27.4222	40.3733	24.2698	2.26354	27.1187	17.6547	0.241943	32.5703	14.0969	8640	749.938153	2160	1410.061847
Jan 1 2013 1:30AM	3.49531	82.0798	65.074	69.4715	10.8064	49.3448	65.9328	159.963	72.8614	27.5732	40.5568	24.2521	2.38594	27.05	17.5479	0.222476	33.8479	14.2839	8886.86	769.616416	2221.715	1452.098584
Jan 1 2013 1:45AM	3.46198	82.0138	66.099	68.8044	11.5814	49.3989	65.0292	161.725	76.5807	27.7298	40.407	24.2385	2.41484	27.1117	17.6854	0.222866	33.2333	14.6177	8763.43	775.000796	2190.8575	1415.856704
Jan 1 2013 2:00AM	3.96331	82.4811	66.3109	69.4148	11.419	49.6862	67.5307	163.072	74.2688	27.5292	40.4238	24.8484	2.3237	27.1617	17.6328	0.226578	33.0896	15.2167	8886.86	776.80658	2221.715	1444.90842
Jan 1 2013 2:15AM	3.53908	82.1817	65.6557	69.3436	10.6773	49.2127	66.3443	163.008	73.7069	27.5944	40.1406	24.2266	2.47839	27.2414	17.7464	0.212319	33.0938	14.4885	8516.57	770.891689	2129.1425	1358.250811
Jan 1 2013 2:30AM	3.54742	82.7823	64.9224	69.2811	11.0273	49.2096	64.3805	160.953	70.8637	27.9152	39.7908	24.2068	2.32474	27.9992	17.7391	0.224234	33.387	14.9625	7529.14	765.516894	1882.285	1116.768106
Jan 1 2013 2:45AM	3.48284	82.2496	65.6437	68.6877	10.8524	49.1059	67.0812	159.391	67.5517	27.8328	40.5076	24.1828	2.22708	27.1875	17.8297	0.211538	33.2214	15.137	7529.14	762.383458	1882.285	1119.901542
Jan 1 2013 3:00AM	3.55889	82.1503	65.9427	69.7825	11.5315	50.3596	68.0115	159.613	64.8647	27.6955	40.9411	24.2104	2.37318	27.3617	18.0688	0.24624	33.737	15.1318	7529.14	765.58041	1882.285	1116.70459
Jan 1 2013 3:15AM	3.60056	82.6842	66.1964	69.6625	11.0024	49.0396	66.3714	161.548	64.3652	27.8972	41.2079	24.174	2.28359	27.0695	18.0625	0.236473	34.8438	14.5521	7776	764.797323	1944	1179.202677
Jan 1 2013 3:30AM	3.51619	83.9515	65.1135	70.0062	10.7774	49.2731	67.6583	162.299	63.5844	27.7408	41.2581	24.1854	2.28802	27.0047	17.9984	0.269945	38.1688	15.2911	8763.43	770.384855	2190.8575	1420.472645
Jan 1 2013 3:45AM	3.50058	83.4522	65.7599	68.7406	12.109	49.5601	66.6427	163.443	62.3036	27.7517	43.2416	24.8234	2.40859	27.2867	17.8432	2.294	38.663	14.587	9010.29	774.41087	2252.5725	1478.16163
Jan 1 2013 4:00AM	3.52246	82.4528	65.4771	69.226	13.587	48.5378	68.7344	163.113	63.4916	27.7068	42.9917	24.2094	2.39531	27.5078	17.9911	2.26804	39.0042	15.3901	8886.86	777.60661	2221.715	1444.10839
Jan 1 2013 4:15AM	3.49538	83.2868	75.6521	69.7246	14.5186	50.7855	68.3563	163.006	64.2108	27.1152	49.0252	24.1927	2.3875	27.0227	17.8901	8.67449	38.7844	15.1781	8886.86	803.30647	2221.715	1418.40853
Jan 1 2013 4:30AM	3.57768	83.4874	80.9536	69.1602	15.4287	50.6967	67.1385	161.845	67.9926	27.3604	50.5588	24.251	2.30104	27.7219	18.2635	10.0473	39.7661	15.0917	9010.29	815.64212	2252.5725	1436.93038
Jan 1 2013 4:45AM	3.54956	83.4547	80.0542	68.3066	15.3937	51.3502	69.3167	162.873	68.4306	27.1263	49.9256	23.7234	2.39792	27.2141	17.9469	10.233	52.4385	16.175	9010.29	829.90998	2252.5725	1422.66252
Jan 1 2013 5:00AM	3.53603	83.2887	78.3172	69.7818	15.7663	50.3929	66.7948	162.78	67.3686	27.0214	50.0425	23.4917	2.40937	27.4453	18.2495	10.2563	57.9734	14.9885	9010.29	829.9043	2252.5725	1422.6682
Jan 1 2013 5:15AM	3.48812	84.1226	79.1271	68.119	15.9979	59.0455	68.1922	162.156	66.9941	27.1049	50.226	24.1318	2.42813	27.3133	17.7167	10.0983	58.4589	15.0911	9133.71	839.81165	2283.4275	1443.61585
Jan 1 2013 5:30AM	3.48708	84.7233	79.4813	68.9619	15.7146	59.5573	67.4818	162.619	67.7134	26.7717	50.2262	23.6031	2.29115	27.0617	17.8026	9.81091	58.6172	16.1724	9133.71	842.09664	2283.4275	1441.33086
Jan 1 2013 5:45AM	3.53292	83.9239	80.7682	68.6977	19.3398	59.0117	69.7062	162.473	65.5264	27.0667	48.1764	23.5094	2.4099	27.2414	17.9922	8.81534	58.8349	15.8271	9133.71	842.85316	2283.4275	1440.57434
Jan 1 2013 6:00AM	3.50272	83.6579	80.5156	68.7169	20.8304	59.6681	68.651	163.227	67.4332	27.6661	48.2099	23.4344	2.35521	27.1641	17.8484	8.50068	58.913	15.5667	9133.71	845.86131	2283.4275	1437.56619
Jan 1 2013 6:15AM	3.50168	83.9586	80.2443	69.1656	21.7555	59.2683	67.7802	162.575	69.1524	28.3912	48.1767	23.4979	2.49193	27.2297	18.0859	7.76268	59.0432	16.2807	9257.14	848.36149	2314.285	1465.92351
Jan 1 2013 6:30AM	3.53919	84.5592	79.4146	69.0603	20.7639	59.4394	69.9865	163.016	67.3405	28.5597	47.3936	23.4698	2.2987	27.2758	17.9312	7.65234	58.5708	15.9552	9257.14	846.22673	2314.285	1468.05827
Jan 1 2013 6:45AM	3.57774	85.1598	79.6599	68.9499	20.4307	59.7084	68.3359	161.337	65.5285	28.5765	48.3604	23.551	2.38646	26.9797	17.8453	7.48984	58.162	15.4578	9380.57	841.49684	2345.1425	1503.64566
Jan 1 2013 7:00AM	3.534	85.2605	80.1078	68.7082	20.6807	59.6616	68.2802	162.678	64.4665	27.6599	48.8106	23.7792	2.36094	27.3117	17.8771	7.90075	58.3667	16.4156	9257.14	843.85999	2314.285	1470.42501
Jan 1 2013 7:15AM	3.48296	87.2278	78.8714	69.8282	20.7808	59.3091	69.1354	162.71	68.5607	27.6132	48.4775	23.4427	4.61953	27.3008	17.8964	8.12831	58.8396	15.4901	9133.71	851.7145	2283.4275	1431.713
Jan 1 2013 7:30AM	3.53089	86.1285	80.3411	68.8115	21.0226	59.1468	69.7156	163.007	66.9987	27.7075	48.011	22.9849	5.58906	27.2813	17.8521	7.51356	58.8438	16.888	9133.71	851.37391	2283.4275	1432.05359
Jan 1 2013 7:45AM	3.54756	85.6291	79.9849	68.9778	20.4893	58.9831	69.7495	164.158	66.6867	28.0235	47.7279	22.9802	5.57734	27.318	17.7755	7.27996	57.237	16.3797	8269.71	848.50506	2067.4275	1218.92244
Jan 1 2013 8:00AM	3.52361	84.7965	79.0927	69.1684	20.7727	58.0267	69.2557	162.884	66.7497	27.6653	48.2614	23.0031	5.16406	27.1625	17.7177	7.89097	56.7182	16.2615	7652.57	844.11474	1913.1425	1069.02776
Jan 1 2013 8:15AM	3.46216	84.9638	79.0938	68.9182	20.7145	58.2951	69.6424	161.706	65.719	27.6012	44.5782	22.8865	5.14479	27.1781	17.7297	4.54241	55.9286	16.199	7652.57	834.30346	1913.1425	1078.83904
Jan 1 2013 8:30AM	3.4882	85.6644	78.0385	69.3883	21.1312	57.676	69.0203	162.939	62.657	27.6155	47.7784	23.0375	5.07578	27.2359	17.7641	7.32853	56.1203	14.3141	7529.14	836.27301	1882.285	1046.01199
Jan 1 2013 8:45AM	3.45488	85.6651	78.624	69.6084	20.8396	58.4089	70.6417	161.341	59.0476	27.5597	47.0952	23.9984	5.10651	27.2398	17.4953	7.50993	56.1109	14.0672	7405.71	833.81412	1851.4275	1017.61338
Jan 1 2013 9:00AM	3.49447	85.2991	80.0026	69.3499	20.398	58.1118	68.0865	159.259	58.9071	27.1498	46.1454	23.3302	5.11719	27.075	17.6927	5.86134	55.8411	13.613	7035.43	824.7342	1758.8575	934.1233
Jan 1 2013 9:15AM	3.83018	85.8997	80.4042	69.0502	20.6564	58.3751	69.9073	160.414	57.8136	27.8058	45.8456	23.638	5.18776	27.3844	17.6734	5.70726	56.3677	12.7047	7035.43	828.6653	1758.8575	930.1922
Jan 1 2013 9:30AM	3.49865	84.8337	79.5839	69.5168	20.6815	58.0357	68.0208	159.361	55.2825	27.1267	46.7124	23.6594	5.06667	25.4812	17.7766	6.84069	56.3422	13.0547	6665.14	820.87511	1666.285	845.40989
Jan 1 2013 9:45AM	3.49032	84.901	78.5344	69.4116	21.2899	58.3901	66.4724	158.537	55.2827	27.5369	46.7459	23.5906	5.10573	23.9328	17.774	6.08662	57.3172	12.024	6541.71	816.42317	1635.4275	819.00433
Jan 1 2013 10:00AM	3.4997	84.1017	79.9599	68.6121	21.0984	58.2174	67.0365	159.209	53.9861	27.5061	44.8961	23.4177	5.16406	27.0008	17.8693	4.78671	56.549	11.6135	6418.29	814.52407	1604.5725	790.04843
Jan 1 2013 10:15AM	3.49867	84.5023	80.0792	69.1971	20.2484	58.9822	66.5297	159.876	50.9706	27.3979	45.4463	23.3516	5.12604	26.832	17.9266	5.34305	56.3641	11.3396	6171.43	813.01136	1542.8575	729.84614
Jan 1 2013 10:30AM	3.4393	84.6696	82.5943	68.704	20.5402	58.8762	65.8328	156.378	52.2833	27.8467	45.9465	23.6526	5.10807	27.1031	17.8318	5.56402	55.6776	11.2031	6541.71	813.25119	1635.4275	822.17631
Jan 1 2013 10:45AM	3.48097	85.2703	80.7526	69.4154	20.8903	58.2493	64.3094	156.095	50.1117	27.1516	44.9133	23.5938	5.1888	27.1336	17.9521	3.87199	56.7484	10.1964	6665.14	805.32496	1666.285	860.96004
Jan 1 2013 11:00AM	3.46848	86.2709	80.3781	69.5455	21.7653	58.6206	66.1333	156.621	50.69	27.7065	47.3301	23.2349	5.11641	26.8938	17.7969	6.17451	55.0151	10.1495	6048	812.9109	1512	699.0891
Jan 1 2013 11:15AM	3.56224	86.0716	79.8021	69.9518	21.2488	58.289	65.2797	158.501	48.7683	27.359	45.497	23.3031	5.25911	27.2266	17.9328	3.88892	56.1615	10.2833	6294.86	808.38587	1573.715	765.32913
Jan 1 2013 11:30AM	3.50495	86.9722	81.8828	69.1308	20.7155	57.7163	63.4615	157.207	46.5497	27.5943	44.9805	23.2823	5.21953	27.4281	17.8083	3.783	55.0562	9.27292	6171.43	801.5659	1542.8575	741.2916
Jan 1 2013 11:45AM	3.49558	88.3729	82.5224	69.4903	20.6656	58.1957	65.0396	156.049	47.3624	27.336	46.9973	23.3974	4.83724	27.2828	17.9823	5.54555	54.7167	9.28958	6171.43	808.57835	1542.8575	734.27915
Jan 1 2013 12:00PM	3.51226	87.3069	80.7859	69.4074	20.9657	58.5045	64.8617	156.488	46.6126	27.6713	45.2975	23.4969	5.18932	27.3523	18.025	3.36566	54.6234	9.60521	6048	803.07155	1512	708.92845
Jan 1 2013 12:15PM	3.52476	87.7409	81.2807	68.969	21.7657	59.3401	62.4594	155.289	47.4878	27.5946	45.6977	23.3839	5.12656	27.1234	17.7573	4.31043	54.1375	9.57292	6048	802.56167	1512	709.43833
Jan 1 2013 12:30PM	3.47685	88.3083	82.8625	69.5528	21.4742	59.127	62.9617	156.202	46.8004	27.2131	46.6479	23.5203	5.21042	27.6203	18.7443	4.99965	61.4854	9.30156	6048	815.50868	1512	696.49132
Jan 1 2013 12:45PM	3.47998	88.2756	83.0297	69.3405	20.8576	58.5864	61.6198	154.391	46.1287	27.8055	43.1314	24.1901	5.1125	27.9617	19.001	2.08068	64.5203	9.65573	6171.43	809.16819	1542.8575	733.68931
Jan 1 2013 1:00PM	3.50707	88.2429	83.7073	69.9251	20.8493	58.1931	64.6161	154.642	44.957	27.4467	45.8482	23.5005	5.15573	28.2625	17.6833	3.62546	64.5635	8.74375	6171.43	813.46951	1542.8575	729.38799
Jan 1 2013 1:15PM	3.58	88.577	80.9594	69.4665	20.9911	58.579	63.8813	154.847	46.3635	27.8796	47.2984	23.6448	5.21771	27.6828	17.724	4.59146	64.613	8.72552	6171.43	814.62209	1542.8575	728.23541
Jan 1 2013 1:30PM	3.48625	89.1443	81.5849	69.1769	21.4495	58.6879	62.4839	154.202	48.4574	27.8636	45.3486	23.4854	5.35703	27.8477	17.6573	2.44157	64.3422	8.80052	6171.43	811.81697	1542.8575	731.04053
Jan 1 2013 1:45PM	3.49459	88.345	82.224	69.6449	21.3246	57.8896	60.8661	154.432	46.4107	27.5692	45.2654	23.5281	5.2237	27.4156	17.5849	2.98125	64.7021	9.46354	6171.43	808.36528	1542.8575	734.49222
Jan 1 2013 2:00PM	3.51543	89.2457	83.8943	68.5272	20.6913	59.3333	62.0641	156.487	44.1921	27.7716	46.4156	23.4312	5.28776	27.1773	17.5599	4.42256	64.6688	8.78385	6294.86	813.469	1573.715	760.246
Jan 1 2013 2:15PM	3.52482	89.3463	83.7203	69.2062	20.4497	61.1695	61.2	155.309	44.0048	27.6473	44.1491	23.5083	5.28073	27.1578	17.9167	1.93429	64.2792	9.18073	6294.86	808.98477	1573.715	764.73023

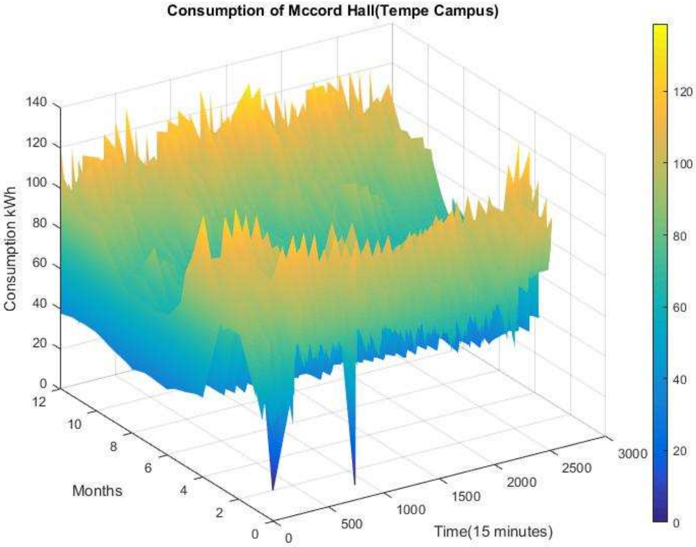
APPENDIX B

CHAPTER 3 RAW DATA SURFACE GRAPHS

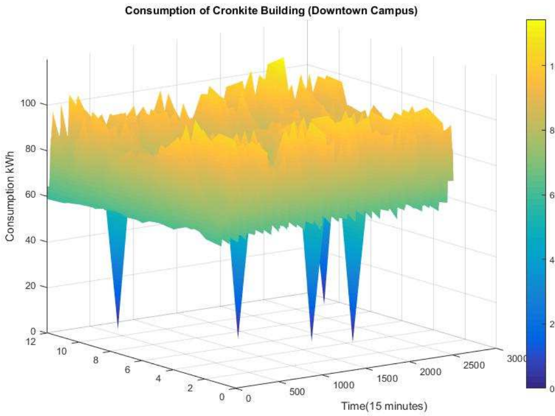
B.1. Raw Data Graph of Academic Center (Polytech Campus)



B.2. Raw Data Graph of McCord Hall (Tempe Campus)



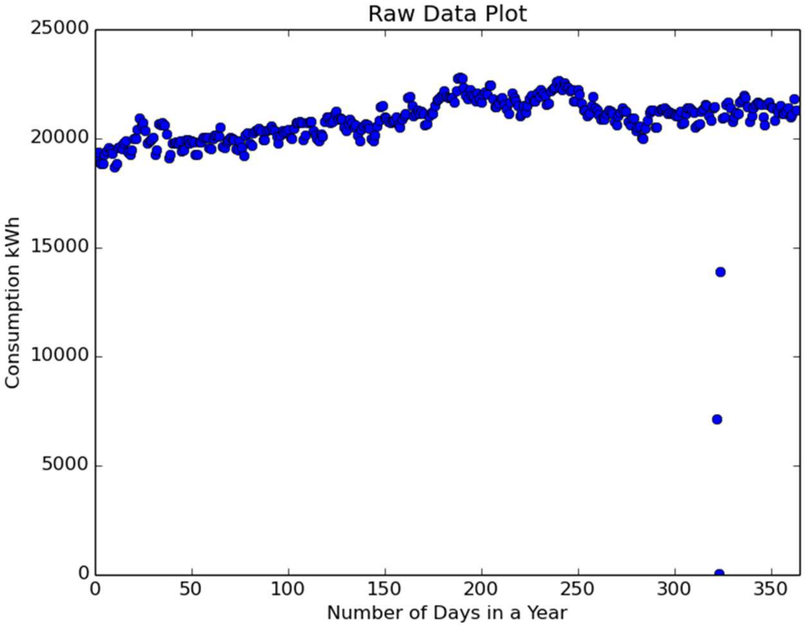
B.3. Raw Data Graph of Cronkite Building (Downtown Campus)



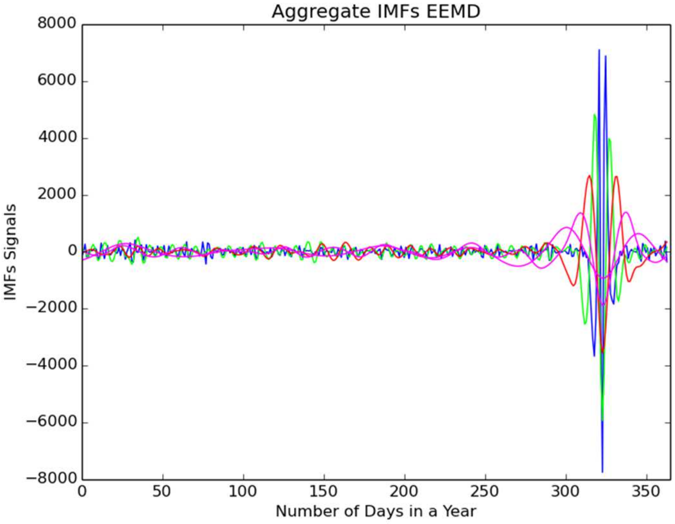
APPENDIX C

CHAPTER 4 AGGREGATED IMFs OF THE BUILDINGS

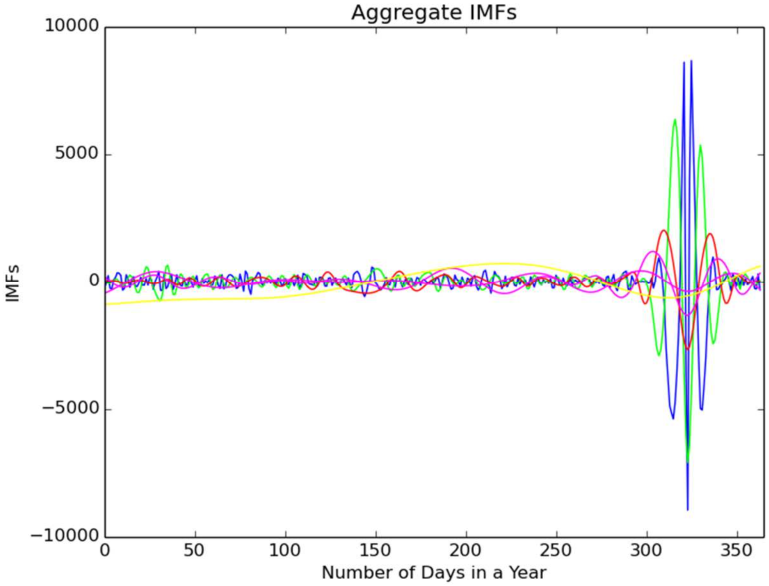
C.1 Raw Data Graph of Building C



C.2 Aggregated IMF Graph of Building D



C.3 Aggregated IMF Graph of Building E


Aggregate IMFs
IMFs
Number of Days in a Year