Leveraging Smart Meter Data through Advanced Analytics:

Applications to Building Energy Efficiency

by

Saurabh Jalori

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2013 by the
Graduate Supervisory Committee:

T. Agami Reddy, Chair
Harvey Bryan
George Runger

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

The poor energy efficiency of buildings is a major barrier to alleviating the energy dilemma. Historically, monthly utility billing data was widely available and analytical methods for identifying building energy efficiency improvements, performing building Monitoring and Verification (M&V) and continuous commissioning (CCx) were based on them. Although robust, these methods were not sensitive enough to detect a number of common causes for increased energy use. In recent years, prevalence of short-term building energy consumption data, also known as Energy Interval Data (EID), made available through the Smart Meters, along with data mining techniques presents the potential of knowledge discovery inherent in this data. This allows more sophisticated analytical tools to be developed resulting in greater sensitivities due to higher prediction accuracies; leading to deep energy savings and highly efficient building system operations.

The research explores enhancements to Inverse Statistical Modeling techniques due to the availability of EID. Inverse statistical modeling is the process of identification of prediction model structure and estimates of model parameters. The methodology is based on several common statistical and data mining techniques: cluster analysis for day typing, outlier detection and removal, and generation of building scheduling. Inverse methods are simpler to develop and require fewer inputs for model identification. They can model changes in energy consumption based on changes in climatic variables and up to a certain extent, occupancy. This makes them easy-to-use and appealing to building managers for evaluating any general retrofits, building condition monitoring, continuous commissioning and short-term load forecasting (STLF).

After evaluating several model structures, an elegant model form was derived which can be used to model daily energy consumption; which can be extended to model energy consumption for any specific hour by adding corrective terms. Additionally, adding AR terms to this model makes it usable for STLF. Two different buildings, one synthetic (ASHRAE medium-office prototype) building and another, an actual office building, were modeled using these techniques. The methodologies proposed have several novel features compared to the manner in which these models have been described earlier. Finally, this thesis investigates characteristic fault signature identification from detailed simulation models and subsequent inverse analysis.

DEDICATION

Wish you'll were here, Baba / Dada…..

ACKNOWLEDGEMENTS

It has been a long journey that brings me to the culmination of this long sought dream. It all started many years back, when as a 22-year old Architect, I set foot in the real world trying to discover myself, wishing to change the world. The amazing set of people I met and the experiences that life paved my way with have absolutely changed my view of the world. In retrospect, I feel blessed to have made it this far and looking ahead, I feel the real voyage is about to begin. I don't know where life will take me from here, and where will I be a few weeks, months or years from now, but I stand ready for what the seas will bring. For my time at ASU, I have learnt not to fear the unknown, but to seek it out, to learn, to grow, to constantly challenge and conquer that what I do not know yet.

I would like to thank each one of the individuals who have directly and indirectly played an important role in shaping my life and actions. To begin with, I would like to thank my beautiful wife and best friend, Namitha Jalori, for standing by me and my decisions during the best and worst times of our life; helped me understand the big picture, showed me the way when in doubt and kept me grounded when required. Vaibhav Rathi, my best friend, confidant, brother and critic always served as a role model, lent an ear when I needed one and kept encouraging me. Vipul Singh, my friend and guide, who has seen all this unfold over the last few years and still has the patience to occasionally knock some sense into me. Ranojoy Dutta, my friend, critic, mentor, colleague and sounding board, taught me how and where to look for answers to the weirdest of my queries. Working through the nights on projects, long walks by the lake and up the A-mountain were the ideal settings for the most stimulating conversations I

have ever had. To my friends Apoorva Bhattad, Anant Bhattad, Raji Sunderkrishnan and Shervil Bhatia, for all those light and serious moments that are my memories leading up to this day. Anupam Bansal, my friend and guide, also my ex-boss, for believing in me. He gave me the freedom to explore, to learn, constantly ideate and shape my ideas. It was only under him that I realized the power of my ideas.

I had always wanted the perfect teacher, someone who could understand my curiosities and help me see this world far beyond what I would have been able to on my own. It is here at ASU that this wish got fulfilled and I found not one, but many such teachers. Prof. T. Agami Reddy, my Thesis chair and mentor taught me many things, most importantly, to believe in myself; he constantly challenged me, never doubted my abilities and helped me expand my thinking far beyond what I could have imagined. Prof. Harvey Bryan, who took a chance on a student two years back, is the reason why I am writing this document; I would be enjoying the Midwest winters if it was not for his support. I learned from him that simplicity usually leads to success, and that one step at a time would usually get you to where you would like to be and beyond. I would like to thank Prof. George Runger for his support and guidance which helped me shape my thesis. Prof. Marlin Addison for the time and dedication with which he replied to all my queries; the anecdotes that I looked forward to in all the extra classes, made learning so much fun. Prof. Muthukumar Ramalingam, who taught me to develop a clear understanding of my ideas, break-up complex problems into smaller pieces and then solve them. All my teachers taught me many things but the most important of them all, the virtues of patience and humility. I am lucky to have shared a personal bond with all

my teachers at ASU, one that reaches far beyond academia and I will always cherish this time and look up to them when in doubt, when not in doubt.

Finally, I would not be who I am today or where I am today, without the unending love and support of my family. My grandfathers and father are the greatest men I will ever know; their curiosities, ambition, vision, perseverance and hard work are the scales I measure myself with and I hope that one day I will come close to their achievements. My mother has been an energetic, strong and a supporting pillar; she pushed me constantly to achieve my dreams and taught me the virtues of a positive and go-getter attitude. Finally, to my brothers Sameer, Kshitij and Saaransh Jalori – Never has life been so meaningful without a real meaning.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ACRONYMS

AOB	Actual Office Building
AR	Auto regressive
ASHRAE	American Society of Heating, Refrigeration and Air Conditioning Engineers
AZ	Arizona
CCx	Continuous Commissioning
CMAC	Cerebellar Model Articulation Controller
CO	Colorado
CP	Change-point
CV-RMSE	Coefficient of Variation – Root Mean Square Error
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DBT	Dry-bulb Temperature
DM	Date Mining
DSM	Demand Side Management
ECM	Energy Conservation Measures
EID	Energy Interval Data
EIS	Energy Information Systems
EMCS	Energy Management and Control System
FDD	Fault Detection and Diagnosis
FEMP	Federal Energy Management Program
HVAC	Heating, Ventilation and Air-conditioning
IMDS	Information Monitoring and Diagnostic Systems
IPMVP	International Performance Measurement and Verification Protocol
KDD	Knowledge Discovery in Databases

LEED	Leadership in Energy and Environmental Design
LTLF	Long-term Load Forecasting
M&V	Monitoring and Verification
MBE	Mean Bias Error
MCP	Multiple Change Point
MTLF	Medium-term Load Forecasting
NC	New Construction
NEMVP	North American Energy Measurement and Verification Protocol
O&M	Operations and Management
OLS	Ordinary Least Squares
PRISM	Princeton Scorekeeping Method
RMSE	Root Mean Square Error
SAC	Sample Auto Correlation Function
SD	Standard Deviation
SMLR	Stepwise Multivariate Linear Regression
SOB	Synthetic Office Building
SPAC	Sample Partial Auto Correlation Function
SSE	Sum of Square Error
STLF	Short-term Load Forecasting
SVM	Support Vector Machine
USDOE	U.S. Department of Energy
WBE	Whole Building Electric
WSN	Wireless Sensor Network

CHAPTER 1 : INTRODUCTION

1.1 Overview

The United States, home to 4% of the World's population consumes roughly 20% of the total energy produced by the World (www.eia.gov), which is an indicator of the energy demands of the developed nations today. Buildings in the United States themselves account for roughly 41% of this total energy consumed. They are large consumers of electricity, water, alternative fuels and many other natural resources and have a sizeable environmental footprint; thereby requiring considerable attention on the path to Energy Independence and Security.

Buildings, or the built environment, are major contributors to the world's total energy consumption (Hunn, 1996). There are two important facets to this: from the front-end; it is a design problem, a lot of effort has to be put into the entire process, accounting for various variables such as the building form, architectural characteristics, system parameters and occupant behavior, and; from the back-end; it is a performance and operational problem; one needs to ensure that the buildings are performing as intended. Additionally, as pointed by many studies, (for example, Diamond, 2001) and U.S. Department of Energy 2010 Buildings Energy Data Book, a vast majority of the North American building stock is significantly old. Systems tend to deteriorate over time and become more inefficient, leading to higher energy consumption. Also, building managers need to adapt to the rapidly changing economics of power generation, and should be able to change building operations for purposes of peak shaving or peak shifting (Demand Side Management or DSM measures), to avoid any exorbitant energy charges.

Many of the initial building energy calculation tools were developed for use during the design phase, i.e. they catered to the need of comparing and evaluating different design alternatives for buildings that were to be built, i.e. New Construction (NC). However, the current building stock, which will account for a significantly large percentage of the total building stock a decade from now, presents a huge opportunity for energy savings. It is now widely accepted that it costs much less to retrofit buildings than to completely re-build them. This presents an enormous potential to develop strategies and techniques to evaluate these buildings and measure the impact of Operation and Maintenance (O&Ms) and Energy Conservation Measures (ECM's) undertaken.

Although the Green Building movement has been around for more than a few decades, technological advances in the recent decade have created a huge opportunity for people to get an insight into their energy consumption behavior and make informed decisions that help them lower their energy bills. New technologies such as Smart Metering and Smart Grids have made it easier for the utilities to collect data regarding the energy consumption patterns of their customers, which has been publicly made available through programs such as the Green Button Initiative for analysis purposes. Energy Interval Data is a record of energy consumption levels, with readings made at regular intervals throughout the day, every day, over an extended period of time collected by the utilities.

The main focus of this research is to develop and assess methods to evaluate the performance of existing buildings utilizing hourly energy consumption data, for purposes such as Monitoring and Verification (M&V), Building Condition Monitoring and Short-term Load forecasting for better electric Demand Response Management (DRM)

measures. Additional work has been carried out to develop methods that can be used for automated Fault Detection and Diagnosis purposes by analyzing the building energy data collected from Smart Meter.

1.2 Research Outline

Energy is a big question that most countries face today and one that will ultimately determine the future of any nation. Understanding the energy related problems and addressing these is the key agenda and the center of many policies framed by all the nations today. In general, the factors influencing the total energy consumption of buildings can be grouped into seven categories (Yu et al., 2011):

(i) Climate (e.g., outdoor-air temperature, solar radiation, humidity etc.),

(ii) Building-related characteristics (e.g., type, area, orientation, etc.),

(iii) Occupancy Schedules,

(iv) Type of building services systems and operation (e.g., space cooling / heating, hot water supplying, etc.),

(v) Building occupants' behavior and activities,

(vi) Social and economic factors (e.g., degree of education, energy cost, etc.), and

(vii) Indoor environmental quality desired.

Each of the above listed factors play an important in role in the total building energy consumption and must be clearly understood. This research specifically deals with point (i) and (iv) and tries to address the issues of improving building energy performance by analyzing data related to these two sources. Specifically, the types of issues that this research aims to address are as follows:

(a) How can we develop reliable building energy prediction and forecasting models that are easily interpretable and can be used with ease by people without advanced mathematical skills?

(b) How can we quantify building occupants' behavior and identify its effect on total building energy consumption?

(c) How can newer building energy data streams (such as Smart Meter Data) be utilized for building energy prediction purposes? How reliable are these data streams and what kind of inherent and reliable knowledge do they present to the building owner / manager?

(d) How can this building energy data be used for one-time Monitoring and Verification (M&V) purposes as well as for ongoing Building Condition Monitoring purposes?

(e) How can these energy data streams be used for advanced Fault Detection and Diagnosis purposes?

This research looks in details at all of the above identified issues and tries to propose simple, easy to implement methods, that can be used to address them.

1.3 Proposed Data Analysis Techniques

This research combines the more conventional Inverse Statistical data analysis techniques with the more recent Data Mining (DM) methods for purposes of building energy prediction purposes.

More specifically, regression modeling techniques are applied to Smart Meter Data after clustering this data using Data Mining algorithms for removal of outliers and

generation of day types so as to have a generalized model that can be used to predict energy consumption for multiple day types. Additionally, time series analysis techniques, such as Autoregressive (AR) modeling are applied to model the error structures for improving the prediction accuracies of these regression models.

Finally, for fault detection purposes, various commonly observed faults with office building types are identified and these are simulated using the hourly time-step energy simulation programs (e.g., eQuest) and the resulting energy data streams are sorted to generate residual patterns, which are characteristic of those particular faults.

1.4 Thesis Organization

This chapter provides an overview of the research topic. It outlines the problem statements or research objectives and gives a general introduction to the proposed data analysis techniques. The scope and limitations of the work carried out have also been clearly stated.

Chapter 2 will present a historical overview of the various studies and data analysis techniques that have been proposed in the past for building energy performance analysis purposes. As such, it outlines the historical and the more recent trends observed in this domain and present a clear picture of the direction in which they are headed.

Chapter 3 reviews existing literature in the building energy performance domain. We look at studies that have identified various techniques as well as addressed related issues. A short literature review of the existing studies is presented along with the use of data mining techniques for analyzing building energy.

Chapter 4 introduces the proposed data analysis framework. Various inter-connected and related processes, such as data pre-processing, clustering, day typing, outlier detection and removal, inverse statistical modeling and autoregressive modeling are described. It presents a detailed, step-by-step methodology which has been applied to two office buildings.

Chapter 5 presents and compares the results of the pre-processing and clustering portions of the research as applied to the two office building types described earlier.

Chapter 6 reports the results of identification of inverse statistical (daily and hourly energy prediction) models for M&V, short-term load forecasting and building condition monitoring purposes.

Chapter 7 presents the FDD work carried out as part of this study. Common equipment faults found in office buildings are simulated and resulting patterns are presented and discussed.

Chapter 8 presents a summary of findings and outlines potential future research directions.

1.5 Scope and Limitations

This report provides a theoretical insight into building performance and evaluation techniques related to the whole-building electric energy use only. A master framework has been proposed in Chapter 4 which tries to draw connections between the two broad strategies of Inverse Statistical model identification and Calibrated Simulation model development related to enhancing the energy performance of existing buildings.

Various techniques of knowledge extraction from Smart Meter Data have been proposed that can feed into both these strategies. However, the work in this research is limited to the inverse Statistical Model identification only. A detailed flowchart proposes a framework of how knowledge can be extracted from the energy interval data in developing more accurate calibrated simulation models. However, calibration simulation modeling itself is outside the purview of this research.

The analysis methods have been evaluated and refined using year-long energy consumption data from two office buildings, a smaller, 54,000 sq.ft. synthetic DOE office prototype building of three floors in Phoenix, Arizona and the other, much larger, 185,000 sq.ft actual offices building of 7 floors in Denver, Colorado. The future scope and proposed areas of study have been described in Chapter 8.

CHAPTER 2 : HISTORICAL OVERVIEW AND BACKGROUND

2.1 Historical Overview of Building Energy Performance Evaluation

Measurement and Verification (M&V) activities were heightened during the Middle East oil crisis. Prior to that, these activities were limited to simple, unadjusted comparisons of monthly utility bills. One of the earliest efforts was reported by Socolow (1978), wherein two identical townhouses were studied. Energy baselines were developed and two controlled experiments were carried out, studying changes due to retrofits and occupant awareness. Various reports and studies such as Fels (1986), Ruch et al. (1991), Claridge et al. (1992), Kissock et al. (1992), Kissock (1993), Ruch et al. (1993), Fels et al. (1995), Haberl (1996), Haberl et al. (1998), Saman et al. (1998), and Yazdani et al. (2000) describing the procedures, methodologies and findings of various energy performance evaluation programs began to appear during the 1980s and the 1990s. Measurement procedures, software and modeling toolkits were developed to aid the process of performance evaluation of buildings and HVAC components. Some of the prominent ones include the ASHARE RP-1050 for calculating linear inverse building energy analysis models (Kissock et al. 2001, 2003), and RP-1093 for compilation of diversity factors and schedules for energy and cooling load calculations (Abushakra et al. 2002).

The U.S. Department of Energy's (USDOE) North American M&V Protocol (NEMVP), published in 1996, and was the culmination of efforts in several states in United States for measuring the energy and demand savings in existing buildings. This was accompanied by USDOE's 1996 Federal Energy Management Program (FEMP)

Guidelines. Finally, in 1997 the NEMVP was updated and republished as the International Performance Measurement and Verification Protocols (IPMVP). In 2001, the IPMVP was expanded into two volumes: Volume I, covering Energy and Water Savings and Volume II, covering Environmental Quality. In 2003, Volume III of the IPMVP was published which covered the protocols for New Construction. In 2002, ASHRAE also released Guideline 14-2002: Measurement of Energy and Demand Savings, which intended to serve as a technical document for the IPMVP. Much of the foundation of the ASHRAE and IPMVP energy modeling procedures was provided by the Texas A&M LoanStar (Loan to Save Taxes and Resources) project, initiated in 1988 by the Governor's Energy Office of Texas.

Some of the more recently undertaken research includes the work by Deru & Torcellini (2005), wherein they conducted extensive research to establish a standard methodology for measuring and characterizing the energy performance of commercial buildings. The performance metrics determined therein may be compared with benchmarks to evaluate and verify performance. Torcellini et al. (2004) studied the performance of six high-performance buildings around the United States. All the buildings performed better than typical buildings; however, none of them performed as well as was initially predicted. Haves et al. (2008) worked on the development of a model specification for performance monitoring systems for commercial buildings and focused on four key aspects of performance monitoring: (a) performance metrics; (b) measurement system requirements; (c) data acquisition and archiving, and (d) data visualization and reporting. In ASHRAE RP-1286, Glazer (2006) proposed guidance regarding base lining of building energy use. Turner and Frankel (2008) undertook a

study of 121 North American LEED New Construction (NC) buildings and concluded that on an average, LEED buildings were delivering anticipated results. Eventually, awarding of LEED credits for advanced commissioning and M&V were recommended.

2.2 Background and Motivation

Each of the above-mentioned studies was either based on analysis of monthly utility bills or on-site monitored consumption data for building energy performance evaluation. On one hand, utility bill analysis can show significant variation between the predicted and measured energy use of the building. This can be attributed to factors such as changing occupancy and equipment schedule; which are unavoidable in all practical circumstances, but are not captured in the monthly bills. On the other hand, on-site end-use energy consumption monitoring for modeling purposes can prove to be significantly expensive and time-consuming. This is where the whole building electric (Smart Meter or Energy Interval) data comes into the picture. Energy Interval Data packs in all the required energy consumption information and is easily available from the electric utilities. It proves to be a good resource for building energy performance evaluation purposes and forms the basis of this study.

Looking at the historical overview and the more recent developments, the field of building energy performance evaluation has evolved over the years and the trends can be broadly classified as:

(i) **Self-help Methods and Tools**: the first wave in this industry was the development of building energy analysis methods wherein, experts in the field used monthly utility bills, along with some on-site measured data and

developed techniques that were used for evaluating buildings and informing energy-saving decisions. Projects such as the Princeton Scorekeeping Method (PRISM), Texas LoanStar described earlier fall within this period of evolving research in the building energy analysis domain.

(ii) **Customized Tools + Services:** the next trend was the commercialization and customization of the methods and techniques developed earlier and were used by large Energy Service Companies (ESCOs) to offer energy efficiency services to their clients. However, only large industries or commercial enterprises could afford these services and were also targeted by the service companies as better sources of revenue generation.

(iii) **Big Data + Cloud-based Building Management:** the most current trend in the industry leverages the recent evolution in various fields such as, availability of Smart Meter data, database storage and management systems, data analysis techniques involving advancement in statistics, machine learning etc. and the information technologies that help make evaluation methods and techniques available to any person at the ease of their desktop screens. These are the pre-packaged solutions that combine years of research and make these tools available to the mass market, increasing building energy efficiencies across time zones, borders, industries and sectors.

This study aims to develop tools and methods in alignment with the most recent trend in building energy analytics that leverage the advancement in multiple technologies and support rapid deployment of energy efficiency initiatives.

CHAPTER 3 : LITERATURE REVIEW

A vast number of published studies and extensive literature is available that describes various procedures relevant to performance measurement and evaluation of operational buildings. This vast pool of knowledge is often used by energy engineers and commissioning agents for detection and diagnosis of operational problems and commissioning errors in these buildings. Knowledge of these techniques can help architects, designers and engineers evaluate how design concepts actually work once applied and can help them make informed decisions.

This chapter presents a detailed literature review of the various techniques and related issues that underline various analytical methods which form the basis of this study; i.e. Inverse Statistical Modeling and Data Mining techniques pertinent to analyzing building energy performance data.

3.1 Smart Meter Data

Availability of smart meter electric consumption data about the buildings marks a paradigm shift in the evaluation methods of building energy performance. The data became available with the inception of electricity deregulation and market-driven pricing around the world which forced the utilities to match consumption with generation, eventually leading to detailed electric consumption data being collected by the utilities. This data is rich in information, recording the energy consumption of buildings with high-resolution and transferring it to the utilities in real-time (Buchmann et al., 2012). In the building energy performance domain, the energy interval data when combined with sophisticated analysis and visualization software can present streams of interval data for

analyzing usage patterns, identifying faulty equipment, validating monthly utility bills, evaluating rate options, verification of operational and control modifications (O&M's) or energy efficiency measures (ECM's) (Younger, 2007).

A number of studies were found utilizing this data for evaluating a building's performance. Price et al. (2002) used interval data from recently installed meters for identifying faulty equipment operation in buildings. They call the technique Bulls-eye commissioning, i.e. rapid commissioning without the wait and expense of full commissioning services and applied it to multiple buildings. Claridge et al. (1994) and Claridge et al. (1996) used hourly energy consumption data for "Continuous Commissioning" purposes, identifying the broad strategies of scheduling changes, efficient temperature settings, efficient system operational settings and summarize their findings. Piette et al. (1998) present the tests of an Information Monitoring and Diagnostic System (IMDS), targeted towards on-site building operators and engineers, and is based on a top-down approach; i.e., from a whole building analysis to system and component diagnostics. A study by Brown et al. (2010) revealed building energy-failure modes and various other anomalies in building operation based on the smart meter data. The main failures identified were heating (or cooling) not relevant to a season, building heating during unoccupied hours, high base load consumption and excessive energy consumption.

3.2 Building Energy Use and Performance Modeling

As described in Reddy (2011), a system is an object under study, simple or complex, and can be an ordered, inter-related set of things, and their attributes. A model is a

construct that allows anyone to represent a real-life system, which can further be used to predict the future behavior of the system under multiple "if-then" scenarios. A model helps gain insight about influential drivers and system dynamics, or predicting system behavior, determining optimal control conditions, operation management, and deciding on policy measures and planning.

Building energy use modeling is of three fundamental types (Hunn, 1996): (a) steady-state, (b) quasi-steady-state, and (c) dynamic. The steady-state models assume that there is no net energy storage in the mass during the entire time period and temperature condition under consideration. Also, it is assumed that all the system parameters (such as internal temperature, outdoor temperature, U-value, system efficiencies, glass shading coefficient etc.) assume the same value during this entire period. On the other hand, quasi-steady-state methods attempt to treat dynamic or transient behavior of the building by assuming parameter constancy for the calculation time period, say one hour, and the system parameters are re-calculated after that calculation period. Dynamic models may approximately represent the time-dependent operation of any system or equipment and variation in its capacity; fully dynamic models are based on sub-hourly time steps and they represent the continuous time variation of the building and its systems.

During the 1980's and early 1990's, many procedures and methodologies to baseline energy use in commercial buildings began to appear. A number of modeling methodologies have been proposed which are useful in developing performance metrics for buildings, as well as HVAC system components. Some of the studies are: the PRISM (Fels, 1986), development of a computer model for evaluating the economics of cool storage systems (Baughman, Jones, & Jacob, 1993), a hybrid monitoring and modeling

approach for analyzing the performance of large central chilling plants (Troncoso, 1997), a baseline energy modeling technique for facility level energy use (Reddy et al., 1997a), and a baseline energy modeling technique for utility bill analysis using both weather and non-weather related variables (Sonderegger, 1998).

The building energy modeling techniques can be broadly classified into the following two categories:

(i) Inverse statistical modeling techniques, and

(ii) Calibration simulation modeling techniques.

These will be discussed in detail in Section 3.3 - 3.4 along with a discussion of studies on related concepts.

3.3 Inverse Statistical Modeling

Inverse statistical modeling is the process of identifying a predictive model structure and estimates of the model parameters from measured system data. It helps one achieve a better understanding of the system dynamics by combining the basic physics of the system with statistical methods. As described earlier, inverse models could be steady-state or dynamic models. Steady state inverse models would be insensitive to dynamic effects (such as the thermal mass effects of the building) and may not perform well for buildings that exhibit such behavior or for short-time steps (such as 15 min or 1 hour periods).

Within the building energy performance domain, a baselining methodology is crucial to verify savings from energy conservation programs. The idea is to develop baseline energy consumption prediction models using inverse methods from pre-retrofit

consumption data and then, to use these models to predict the energy consumption during post-retrofit. The difference between the measured and predicted post-retrofit data thus gives an estimate of the savings achieved. The baseline models are developed using the monthly, daily or hourly energy consumption as the response variable and other parameters of interest as the regressor variables. There are many issues related to the development of a formal baselining methodology at the whole-building level. These issues along with the related studies are presented in the following sections.

3.3.1 Simple Linear Regression Modeling

Simple regression modeling methods and techniques are the mainstay for energy analysts and researchers around the world. The techniques have evolved over the years, overcoming and resolving many issues through multiple iterations and as such, a significant body of work can be found with applications in the domain of building energy performance evaluation.

Claridge (1998) provides a historical perspective on energy analysis of commercial buildings. The paper summarizes and discusses the capabilities and uncertainties inherent in the regression methods used for M&V purposes along with a discussion on the use of artificial neural networks, Fourier series and spectral analysis methods. The paper finally presents the need for graphical indices.

Reddy et al. (1997a and b), discuss various issues such as, normalizing annual energy use for changes in the conditioned area and the number of occupants, as well as correcting for increase in the connected load. The paper, however, identifies that correction of occupancy as presenting a huge challenge as it is not well documented, and

that conditioned area and occupancy are correlated to building operating schedules, and that correcting for all of the above factors can lead to over-correction. The current research identifies this issue and proposes a methodology to generate a comprehensive hourly occupancy that can be used to develop hourly energy prediction models. Additionally, use of outdoor-air temperature as the sole independent regressor variable is presented, also covered in Kissock et al. (1998). Finally, the development of prediction uncertainty bands for the regression model is discussed for reaching statistically significant conclusions about energy performance; an in-detail discussion of this aspect can also be found in Reddy et al. (2000).

Other studies on use of regression techniques for building energy performance analysis are reviewed by Farouz et al. (2001), especially pertinent to the basic procedures developed and used for monitoring and verification (M&V) purposes as part of the Texas Loan Star and Rebuild America programs. This program served as a foundation for several other state and federal M&V programs.

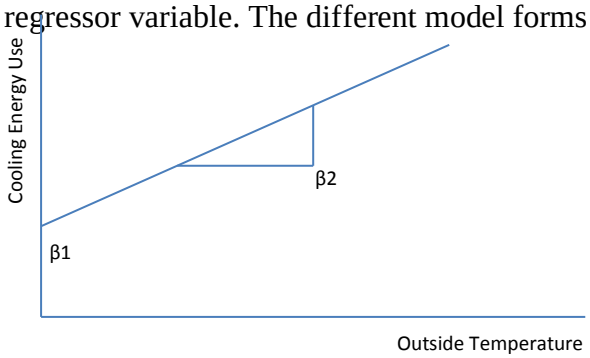
3.3.2 Model structure and change-point modeling (CP)

The simplest steady-state inverse models are the linear regression models to predict average behavior, for example, the monthly average electricity consumption of a building based on the average exterior temperatures. These models can be made more sophisticated by additional techniques such as multivariate regression modeling, wherein, instead of a single regressor or predictor variable, multiple variables could be used to predict the response variable. Another technique for modeling energy consumption of weather-load dominated buildings is change point (CP) linear regression modeling,

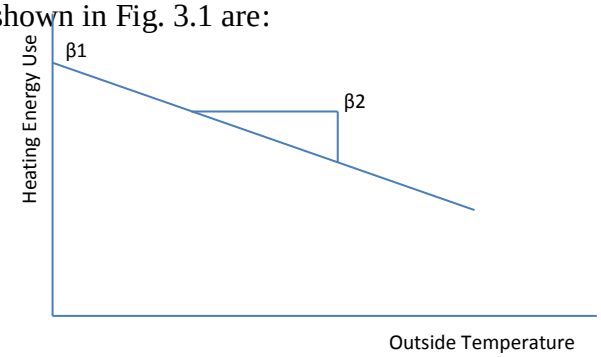
wherein, indicator variables are included to indicate temperature points, beyond which the electricity consumption alters drastically. Certain buildings might even require multiple change-points (MCP) modeling.

Katipamula et al. (1998), review the literature on multivariate linear regression (MLR) models of building energy use, highlighting their usefulness as baseline models and in detecting deviations in consumption owing to any operational changes. Katipamula et al. (1994) used multiple linear regression models with internal gain, solar radiation and humidity ratios as additional variables, in addition to temperature for modeling energy consumption.

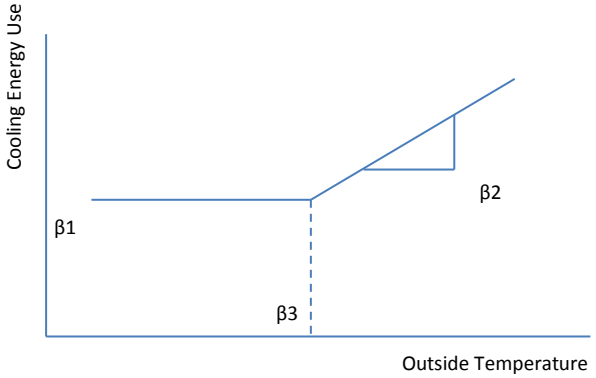
Reddy et al. (1997a) and Kissock et al. (1998) discuss the various functional forms assumed by the regression models with outdoor dry-bulb temperature as the sole regressor variable. The different model forms shown in Fig. 3.1 are:



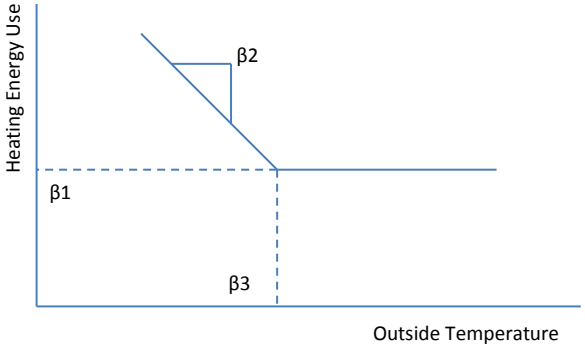
(a) 2P cooling energy model



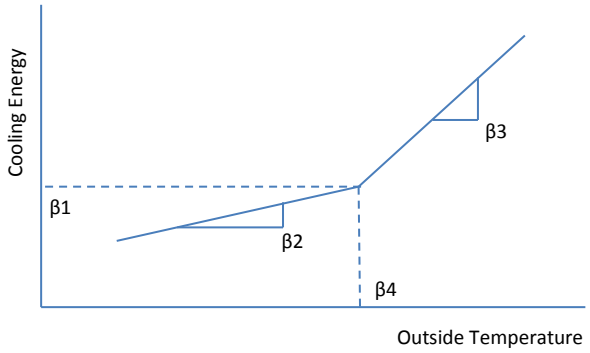
(b) 2P heating energy model



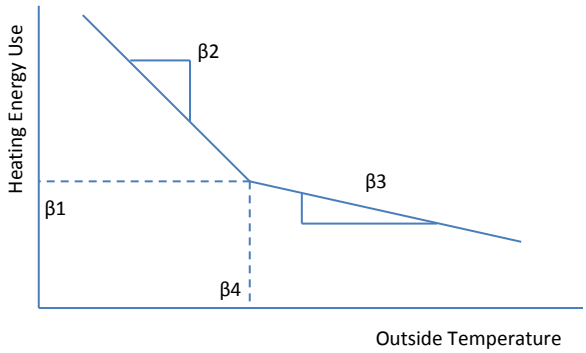
(c) 3P Cooling Energy Model



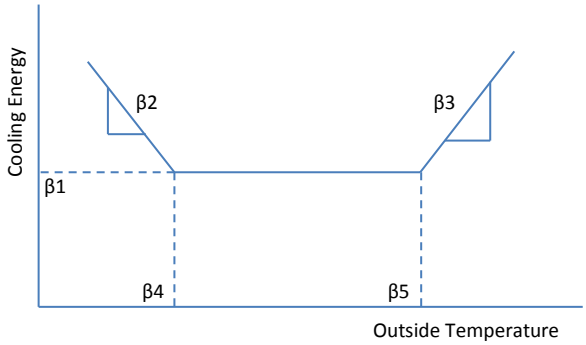
(d) 3P Heating Energy Model



(e) 4P Cooling Energy Model



(f) 4P Heating Energy Model



(g) 5P Energy Models

Figure 3.1. Change Point Models

(i) Mean or one-parameter (1P) models;

(ii) Two-parameter (2P) models for weather-dependent energy use;

$$\check{E} = \beta_1 + \beta_2 . T_{db} \qquad \text{Eq. 3-1}$$

(iii) Three-parameter (3P) change-point models for weather-dependent energy use;

$$\check{E}_c = \beta_1 + \beta_2 . (T_{db} - \beta_3)^+ \qquad \text{Eq. 3-2}$$

$$\check{E}_h = \beta_1 + \beta_2 . (T_{db} - \beta_3)^- \qquad \text{Eq. 3-3}$$

(iv) Four-parameter (4P) change-point models for heating and cooling energy use; and,

$$\check{E} = \beta_1 + \beta_2.(T_{db} - \beta_4)^- + \beta_3.(T_{db} - \beta_4)^+ \quad \text{Eq. 3-4}$$

(v) Five-parameter (5P) change-point models for energy use.

$$\check{E} = \beta_1 + \beta_2.(T_{db} - \beta_4)^- + \beta_3.(T_{db} - \beta_5)^+ \quad \text{Eq. 3-5}$$

Ruch et al. (1992) developed a four-parameter change point model for energy consumption as a function of dry-bulb temperature while Nassif (2012) developed single and multivariate regression models with and without change-points for estimating energy consumption in school buildings, and suggests that including occupancy along with temperature variable greatly improves the consumption predictions.

Abushakra (1999) proposed a stepwise multiple linear regression (SMLR) model that depends on one-time site measurements of certain end-uses and takes into account the daily and hourly variability in energy consumption. However, the process of model development involves on-site measurements of energy consumption data for a few weeks.

Multiple regression models may be developed based on different model structures, i.e., including different regressors or combination of regressors. After comparing their results, the best model can be selected using appropriate statistical or model performance criteria. The most commonly adopted statistical indices to measure model performance are the model coefficient of determination (R^2) and the coefficient of variation of the root mean square error (CV-RMSE). These statistical indices will be presented in Section 3.4.4.

3.3.3 Modeling occupancy

The occupancy factor proves to be an important one in estimating building energy consumption at higher resolution time scales, such as, hourly time scales. Although important, very little published literature has dealt with this predictor of building energy use and performance. One of the simplest methods is to separate the data into "occupied" and "unoccupied" groups and build models on these two respectively. Another solution to build a single model on the entire dataset is to include a dummy regressor variable, which assumes values of "0" for unoccupied and "1" for occupied hours. Although simple, this is not the best way to model occupancy, and thus helps make very little improvements to the model prediction accuracy.

The earliest known study addressing this issue was by Keith et al. (1999). They proposed a methodology for developing a simplified prediction tool to estimate peak occupancy rate from readily available information, specifically average occupancy rate and number of rooms within an office building and as such, required extensive monitoring of occupancy for a 12-month period. This study was a result of evaluating the economics of energy saving potential of occupancy sensors. The occupancy rate is equal to the number of occupied records divided by the sum total of occupied and unoccupied records. The average hourly occupancy is the monthly average of occupancy rate for that particular hour of all the workdays. Finally, a multiple linear regression model of peak occupancy rate was proposed as a function of average occupancy rate, number of rooms, and other variables that are combinations of these two variables.

Camden (1999) accounted for changes in occupancy for calculating the energy savings from retrofits, and proposed to recalculate the energy consumption baselines of

buildings experiencing change in occupancy. Linear and logarithmic correlation models between the whole building electricity consumption and demand, and the occupancy density (no. of people / 1000ft^2) were established.

Another study by Abushakra et al. (2001), derived surrogate occupancy variables by investigating the lighting and equipment load schedules (diversity factors), determined in an earlier study by Bronson (1992). Five different options were used to obtain fractions between 0 and 1 for the occupancy variable:

(i) based on a walk-through survey of the building,

(ii) occupancy derived from lighting and electrical load profiles,

(iii) occupancy derived from the lighting and electrical loads by dividing all values by the absolute maximum value of lighting and electrical consumption,

(iv) a value of 1 for weekdays occupied hours; 0 for unoccupied hours; 0.33 for weekend occupied hours, 0 for weekend unoccupied hours; and finally

(v) 1 for weekdays and 0 for weekends. Regression equations are proposed for deriving the occupancy based on lighting and electrical factors and the results are discussed.

3.3.4 Model Goodness-of-fit Criteria

There are a number of general indices to gauge the goodness-of-fit of various regression models. As described in Reddy (2011) and Dielman (2004), we will use the following model goodness-of-fit criterion in this study:

(i) Coefficient of determination: This is the most widely used goodness-of-fit criteria, where $0 \leq R^2 \leq 1$:

$$R^2 = \frac{explained\ variation\ of\ y}{total\ variation\ of\ y} = \frac{SSR}{SST} \quad \text{Eq. 3-6}$$

$R^2 = 1$ indicates a perfect model fit, whereas $R^2=0$ indicates that no model relationship exists.

where, SSE is the error sum of squares and is given as:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{Eq. 3-7}$$

SSR is the regression sum of squares and is given as:

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad \text{Eq. 3-8}$$

SST is the total sum of squares and is given as:

$$SST = SSE + SSR = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{Eq. 3-9}$$

(ii) Root Mean Square Error: Also known as the "standard error of the estimate", the root mean square error is an absolute value and is defined as follows:

$$RMSE = \left(\frac{SSE}{n-k}\right)^{1/2} \quad \text{Eq. 3-10}$$

where, SSE is the error sum of squares and is given as:

$$SSE = \sum(y_i - \hat{y}_i)^2 \quad \text{Eq. 3-11}$$

Here,

$\boldsymbol{y_i}$ = the actual value of the response variable,

$\hat{y}_i$ = the predicted value of the response variable,

$\bar{y}$ = the mean value of the response variable of the actual data set,

n = the number of data points in the actual dataset, and

k = the total number of regression parameters in the model.

(iii) Coefficient of Variation of the Root Mean Square Error: This is a normalized measure and implies the percentage of the unexplained variation in the response variable when compared to the mean value of the actual response variable. It is defined as follows:

$$\text{CV-RMSE} = \frac{RMSE}{\bar{y}} \qquad \text{Eq. 3-12}$$

With respect to inverse modeling, although, R^2 is a proper metric to use when the primary objective is to evaluate the model fit; the CV-RMSE becomes more relevant when the objective changes to evaluating the actual energy savings (Reddy et al., 2000).

3.4 Calibrated Simulation Modeling

Calibrated simulation entails reconciling the actual energy usage data of an existing building (such as monthly utility bills) with the modeled energy consumption of the same building. The process involves generating an energy model (using energy modeling software such as eQuest, Energy Plus), of the building based on prior knowledge of the various input parameters; such as architectural features of the building, HVAC system parameters, applicable energy rates etc., and generating an initial energy consumption output, yearly electricity / gas costs. Once this has been done, the various input parameters are adjusted to get the simulated energy data to match the actual energy consumption as closely as possible. As mentioned in Reddy et al. (2006), calibration has

been considered an art form that inevitably relies on user knowledge, past experience, statistical expertise, engineering judgment, and an abundance of trial and error. It can be a powerful tool for estimating energy savings and M&V purposes. Calibrated simulations are not within the scope of this research.

3.5 Knowledge Discovery and Data Mining

Rapid advances in sensor technology, data collection and data storage technologies have enabled the collection and storage of various types and amounts of data pertaining to building systems. Although sometimes collected for a specific purpose, this data is abundant in information and can provide great insights into the ways the buildings are being operated and their energy consumption patterns. Historically, the analyst was responsible to analyze this data and come up with important knowledge regarding building energy performance. However, such massive data sets can prove to be a challenge to analyze using traditional statistical techniques. Moreover, it might be impossible to discover patterns that were previously unknown.

A simple high level definition of Knowledge Discovery in Databases (KDD) as mentioned in Fayyad (1996) is: "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." On the whole, the KDD process entails, retrieving data from a database, selecting the appropriate subset of data and deciding on the sampling strategy, cleansing data and handling missing fields, applying appropriate transformation, reducing dimensionality, fitting models to extract information or patterns, evaluating extracted knowledge to check for useful information, visualization and finally consolidating with

existing knowledge. KDD is an inter-disciplinary field (Fayyad, 1996) including statistics, machine learning, artificial intelligence and reasoning with uncertainty, databases, knowledge acquisition, pattern recognition, information retrieval, visualization, intelligent agents for distributed and multimedia environments, digital libraries, and management information systems.

Data Mining is an integral part of knowledge discovery in databases (KDD) (Tan, Steinbach, and Kumar, 2005). It is the process of discovering useful information in large datasets. Matheus et al. (1993) and Bissantz et al. (2008) provide an interesting overview of the KDD and certain data mining tasks.

3.5.1 Data Mining Techniques

Data mining goals can be broadly classified into the following two categories:

(i) **Descriptive Data Mining:** Similar to descriptive statistics, here the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships inherent in the data.

(ii) **Predictive Data Mining:** The main objective here is to predict the value of an attribute (response variable) based on the values of certain other attributes (independent or regressor variables).

The above mentioned goals of data mining can be achieved by using a variety of data mining techniques Fayyad et al. (1996):

(i) **Classification:** Classification is the simple process of assigning different data objects into specific classes that are pre-defined. For this reason, this is also

known as supervised learning technique as the algorithm is trained on classified data before it can be used for prediction.

(ii) **Regression:** Similar to classical statistics, the aim of this technique is to develop a prediction model with the least error.

(iii) **Clustering:** Clustering is the task of discovering groups or structures in data containing data objects that are very similar in nature. This is a common descriptive technique, and the data objects may be earlier classified or not. For this reason, this qualifies as an unsupervised learning technique as there are no pre-set classes to begin with.

(iv) **Summarization:** This technique involves methods for finding compact descriptions for a subset of data.

(v) **Dependency Modeling:** Also known as association rule mining, this technique aims to find relationships or dependencies between variables.

(vi) **Change and Deviation detection:** Also known as anomaly detection, this technique helps to discover unusual data records that are different from previously measured or normative values.

Various clustering techniques have been used in this research which will be discussed further in Chapter 4.

3.5.2 Data Mining in Building Energy Performance

Application of data mining techniques in building energy performance domain is relatively sparse. Although this study uses only clustering techniques, we will present a

short literature review of some of the studies involving data mining techniques used for building performance evaluation purposes.

Piette et al. (2002) present a discussion on the applicability of data mining in automated fault detection using diagnostic tools and EMCS data. They present a broad list of the most common faults found across the office spaces that can be identified and evaluated using EMCS data and diagnostic tools along with a discussion of some of the techniques.

Morbitzer et al. (2004) describe the applicability of various data mining techniques in analyzing building performance data and present a comparative table with other techniques. Association rule mining, classification, outlier analysis, cluster analysis and evolution analysis techniques are presented along with examples of applications in the building energy performance domain.

Liang et al. (2007) proposed a combination of model-based fault detection and diagnosis (FDD) and the Support Vector Machine (SVM) methods for investigating the characteristics of three major faults generated by computer simulation: stuck recirculation damper, cooling coil fouling/blockage and supply fan speed decrease. The faults can be detected efficiently using the residual analysis method based on the variation of the system states under normal and faulty conditions of different degrees.

Wu et al. (2007) discuss the very noisy data, exhibiting temporal and spatial correlation, collected by the wireless sensor networks (WSNs). The paper presents clustering techniques to identify and remove outliers by analyzing and understanding the patterns in the data for purposes of optimizing the indoor air quality in office spaces.

Ahmed et al. (2011) present predictive data mining techniques that can integrate any thermal comfort standards and indoor daylight procedures, observe correlations between weather conditions, building characteristics and low-energy comfortable rooms, and finally build models that can optimize occupant's comfort and energy consumption. In another companion paper, Ahmed et al. (2011) couple data mining techniques with daylight analytical tools for assessing building performance by setting the daylight design criteria.

Clustering methodology is described for purposes of classifying the typical load profiles / patterns of a building (Jota et al., 2011) for assisting the facility manager for better load management; specifically, peak shaving or peak shifting purposes.

Schumann et al. (2011) address the challenges and discuss the contributions of artificial intelligence techniques such as transfer learning, ontologies, knowledge representation or diagnosis for developing easily adaptable, self-learning in-depth diagnostic approaches.

Yu (2012) presents different data mining methodologies for extracting hidden knowledge from building-related data. Specifically, classification analysis, cluster analysis and association rule mining techniques have been described. Classification is used for developing building energy demand predictive models. Clustering is employed for studying occupant behavior on building energy consumption. Association rule mining is used for examining all the correlations and associations between building operational data, discovering useful knowledge about energy conservation.

3.6 Time-Series Modeling and Short-term Electric Load Forecasting

A time-series is a chronological set of observations on a particular variable. The components of a time series (Bowerman et al., 2005) are as follows:

(i) **Trend:** reflects the long-run growth or decline in a time-series

(ii) **Cycle:** refers to the recurring up and down movements around the trend levels.

(iii) **Seasonal variations:** refer to the periodic patterns that complete over a given calendar year and repeat every year thereafter.

(iv) **Irregular fluctuations:** correspond to erratic movements or white noise in a time series that follow no recognizable pattern.

The stochastic time series modeling (Reddy, 2011) is an approach which explicitly treats the model residual errors after removal of the long-term trend and cycle using OLS models by adding a layer of sophistication. In essence, the stochastic time series modeling treats this systematic stochastic component by modeling the error structure, leading to higher prediction accuracies.

Current value at time (t) = [deterministic component] + [stochastic component] = [constant + long term trend + cyclic (or seasonal) trend] + [systematic stochastic component + white nose] Eq. 3-13

Load forecasting is an active area of research and there are a number of published studies which deal with numerous methods (Alfares et al., 2002), such as multiple regression, exponential smoothing, adaptive load forecasting, stochastic time series, ARMAX modeling and neural networks and their applications to load forecasting.

Load forecasting can be classified in terms of the planning horizon's duration (Hahn et al., 2009):

(i) Short-term load forecasting (STLF): up to 1 day.

(ii) Medium-term load forecasting (MTLF): 1 day up to 1 year

(iii) Long-term load forecasting (LTLF): from 1 – 10 years or more.

These are done for different purposes. However, our study addresses short-term load forecasting. STLF is useful in day to day operations of buildings as it would help building managers optimize building system operations for better implementation of demand response management strategies.

Seem et al. (1991) proposed an adaptive method for real-time forecasting of building electrical demand. They used a CMAC (Cerebellar Model Articulation Controller) model for modeling the deterministic trend and autoregressive (AR) models for stochastic time series models. However, the study does not make use of the Sample Autocorrelation (SAC) and the Sample Partial Autocorrelation (SPAC) functions to determine the order of the model. They instead propose building multiple order models and evaluating them by plotting the standard deviation (SD) of the residuals, ultimately selecting the order with the lowest SD. Also, the methodology used is an adaptive one, i.e. it recursively estimates the AR parameters. This study makes use of the SAC and SPAC functions to determine the AR model order and does not include changing the AR parameters recursively; instead it tries to model the systematic error structure for higher prediction accuracies.

3.7 Energy Information Visualization and Analytics

Building analytics has two important aspects; on one hand, it uses descriptive and predictive models to gain valuable knowledge from data – the data analysis side; on the other hand, using the analytical insight, it aims to recommend actions or guide the decision-making process – the communication side.

Energy information visualization or the visual representation of data, in the form of graphs and charts can greatly help understand the energy consumption trends and patterns, inherent in the data and help building managers quickly summarize and assess the overall performance of the building and conduct long-term planning. Over the years, a large number of researchers have proposed various methods to present the energy consumption information in forms that are palatable to energy managers and are informative, visually efficient and easy to comprehend.

An efficient graph is one that aids the decoding process of vast quantities of quantitative information encoded within. The decoding process by the viewer is known as graphical perception. Cleveland (1994), Tufte (1990), Tufte (2001), and Tukey (1977) are seminal works which provide a detailed understanding of the graphical perception problems associated with various kinds of graphs and also suggest a comprehensive set of principles to help enhance a graph's ability to visually represent data structure.

Some of the common graphs used in building energy domain are line graphs, bar charts, scatter plots, pie diagrams, time-series plots, 3-D surface plots and color density plots. The graphical features (Capehart, 2004) that form a part of many energy information systems (EIS) are:

(i) **Summary:** summarizes energy data by day, week, month or other selected time period.

(ii) **Energy-use breakdown**: represents energy use for individual or multiple buildings either by fuel type (electricity, gas, oil etc.) or by end-use type (lighting, space cooling, space heating, plug loads etc.)

(iii) **Load duration curve**: represents the percentage of time a particular load persisted.

(iv) **X-Y scatter plot:** represents the correlation between two measured quantities like energy consumption vs. outdoor ambient temperature etc.

(v) **Time series:** represents the time-dependent energy consumption trends. These could be further broken down into: daily profiles (energy consumption profiles displayed with time), day overlay (comparing various days energy profiles on a time scale), point overlay (multiple variable time series to understand the correlations, for example, 24 hour outdoor temperature and energy consumption profiles), calendar profile (viewing daily consumption profiles on a monthly basis) etc.

Haberl et al. (1998 a and b) present interesting graphical indices to represent detailed building energy consumption information for analytical purposes. Some of the graphical indices discussed are as follows:

(i) **3-D surface Plots:** These present the qualitative aspects of the energy consumption such as variations across the year, as well as diurnal variation across the days, changes between days of the week, periods of low

consumption as against high consumption peaks, missing data points as represented in Figure 3.2.

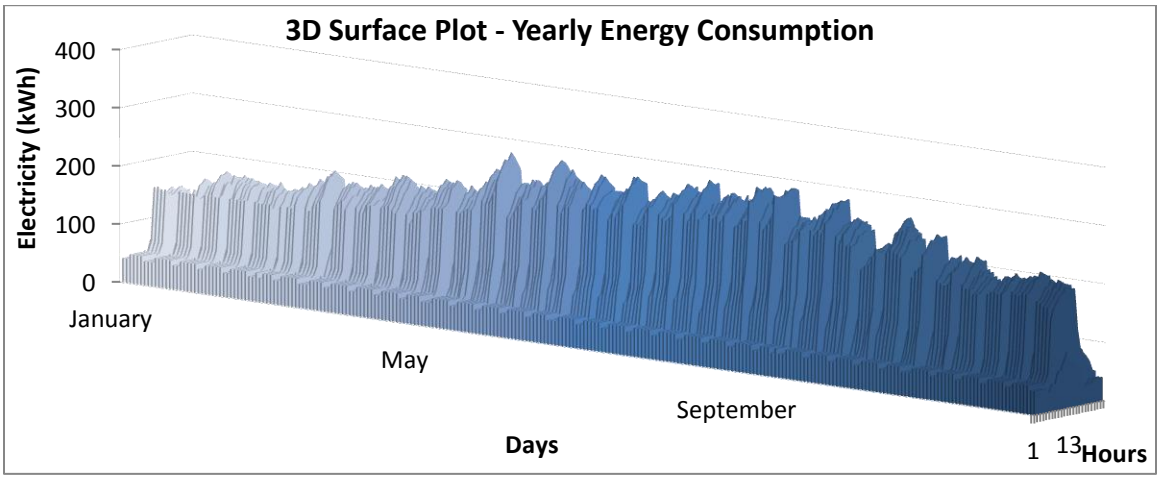


Figure 3.2. 3D Surface Plots

(ii) **Box-Whisker-Mean Plots:** These emphasize more on the quantitative aspect of the energy consumption information and present the relevant statistics. They can be used in two ways: (a) weekly plot, wherein, the 52 box-and-whisker plots show the statistics for all the weeks of the year. These can easily show the shifts in energy consumption across the various seasons throughout the year, as well as the peak and low consumption information during each of the weeks. Additionally, the means can be connected in each of these box-and-whisker plots and the relevant position of the mean and the median for the week can explain the occupancy as shown in Figure 3.3; and (b) hourly day type plot, as in Figure 3-4, where the 24 box-and-whisker plots show the statistics of energy consumption across all the hours of the day and as suggested earlier, these plots could be drawn for each of the day types. They

present information such as peak and low energy consumption across all the days, base load energy consumption and occupancy levels.

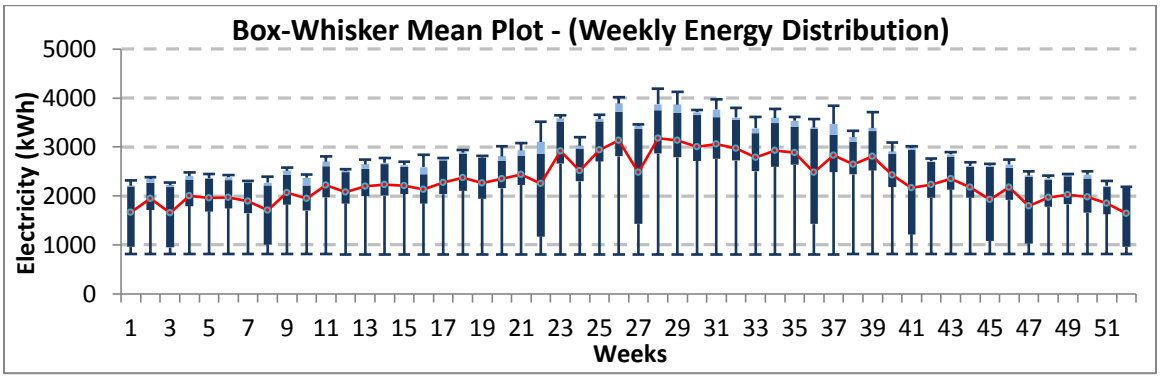


Figure 3.3. Box-Whisker-Mean Plot (Weekly Energy Distribution)

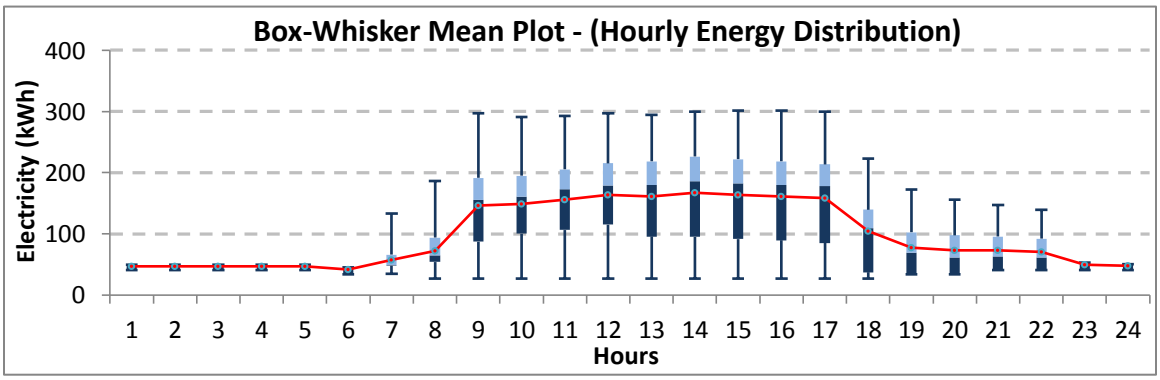


Figure 3.4. Box-Whisker-Mean Plot (Hourly Energy Distribution)

(iii) Scatter Plot and binned Box-Whisker-Mean Plots: Using juxtapositioning[1] and juxtapaging[2] techniques, the scatter plots of outdoor temperature vs. energy consumption are placed along with binned box-whisker-mean plots that show the energy consumption as a function of outdoor temperature bins divided into 5 °F or 10 °F segments as shown in Figures 3.5 – 3.6.

[1] Juxtapositioning is the vertical and/or horizontal axes alignment in the graphs.
[2] Juxtapaging is plotting the same data in different graphs in similar locations on succeeding pages.

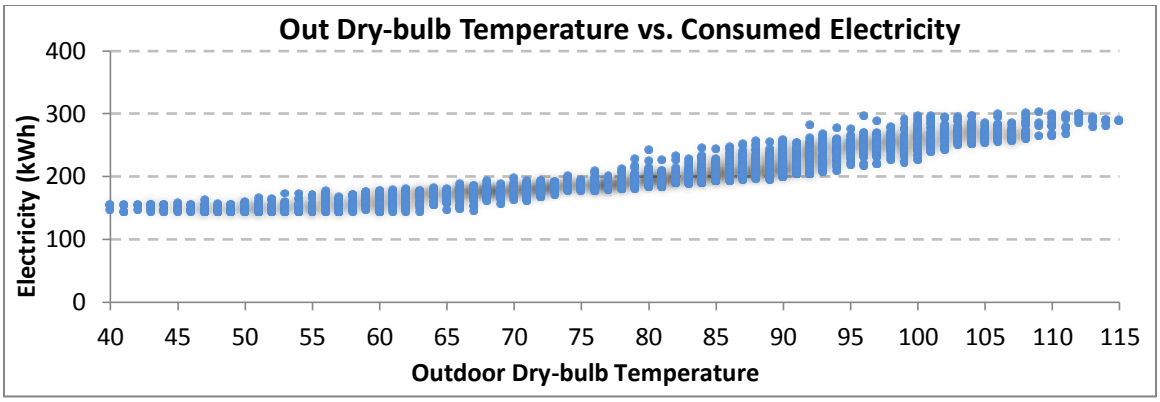


Figure 3.5. Scatter Plot – Outdoor Dry-bulb Temperature vs. Consumed Electricity

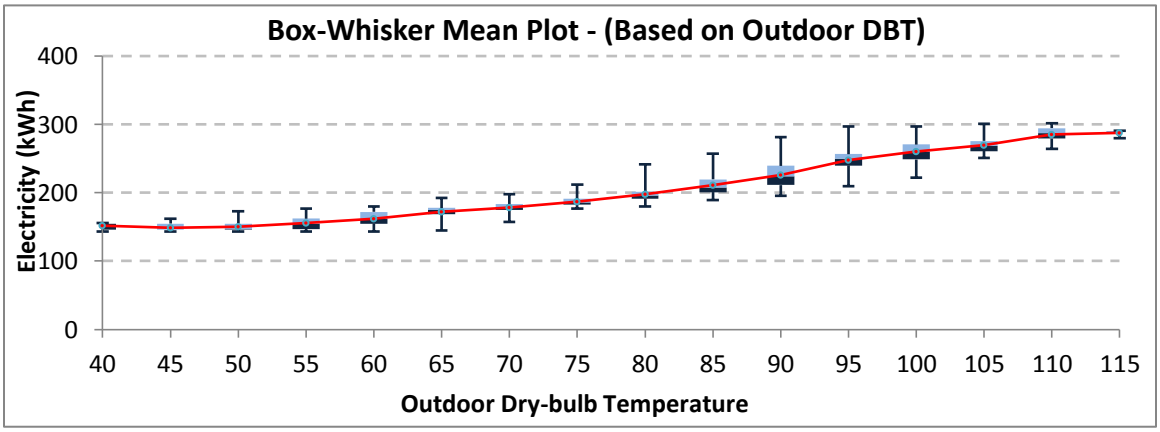


Figure 3.6. Temperature-binned Box-Whisker-Mean Plots

As described in (Reddy and Maor 2006), these plots eliminate data overlap and allow for a statistical characterization of the dense cloud of hourly points. This combined information provides a detailed as well as a statistical view of the data. The inter-quartile range explains the variation in energy consumption over a certain temperature range; this can help explain its dependence on any additional variables, such as humidity or occupancy.

Finally, the energy consumption data should be normalized before plotting which can help comparisons across multiple sites. Also, pre- and post-retrofit charts can be drawn

for energy consumption to measure the improved energy performance of a building after an energy retrofit.

Haberl et al. (1996) explored advanced data presentation techniques by enhancing the display of data with animation (or time sequencing). The specific tools presented are: (a) time sequenced contour plots, and (b) superimposed time-series and moving segment x-y plot.

CHAPTER 4 : METHODOLOGY - INVERSE STATISTICAL MODELING

The research proposes enhancement to building energy performance and operation analysis procedures due to the availability of Smart Meter data. A new methodology for inverse modeling of Smart Meter data is described in this chapter. Statistical methods have been adopted, assisted by Data Mining techniques for predictive modeling. The different application areas for this modeling approach are as follows:

(i) Building Retrofit Monitoring and Verification (M&V).

(ii) Building Condition Monitoring.

(iii) Short-term load forecasting for better demand response management.

The methodology, with additional work, can also be used in automated Fault Detection and Diagnosis (FDD), as well as for pre-processing data useful in developing better calibrated energy simulation models.

Figure 4.1 is a flowchart presenting a broad framework underlining the topic of research in this study and also the possible connections between the broad techniques of inverse statistical and calibrated simulation modeling. Further, flowcharts A, B, C and D in Figures 4.4, 4.7, 4.8 and 4.12 respectively provide greater detail of the methodologies and the various steps that were undertaken under each of the main data processing tasks.

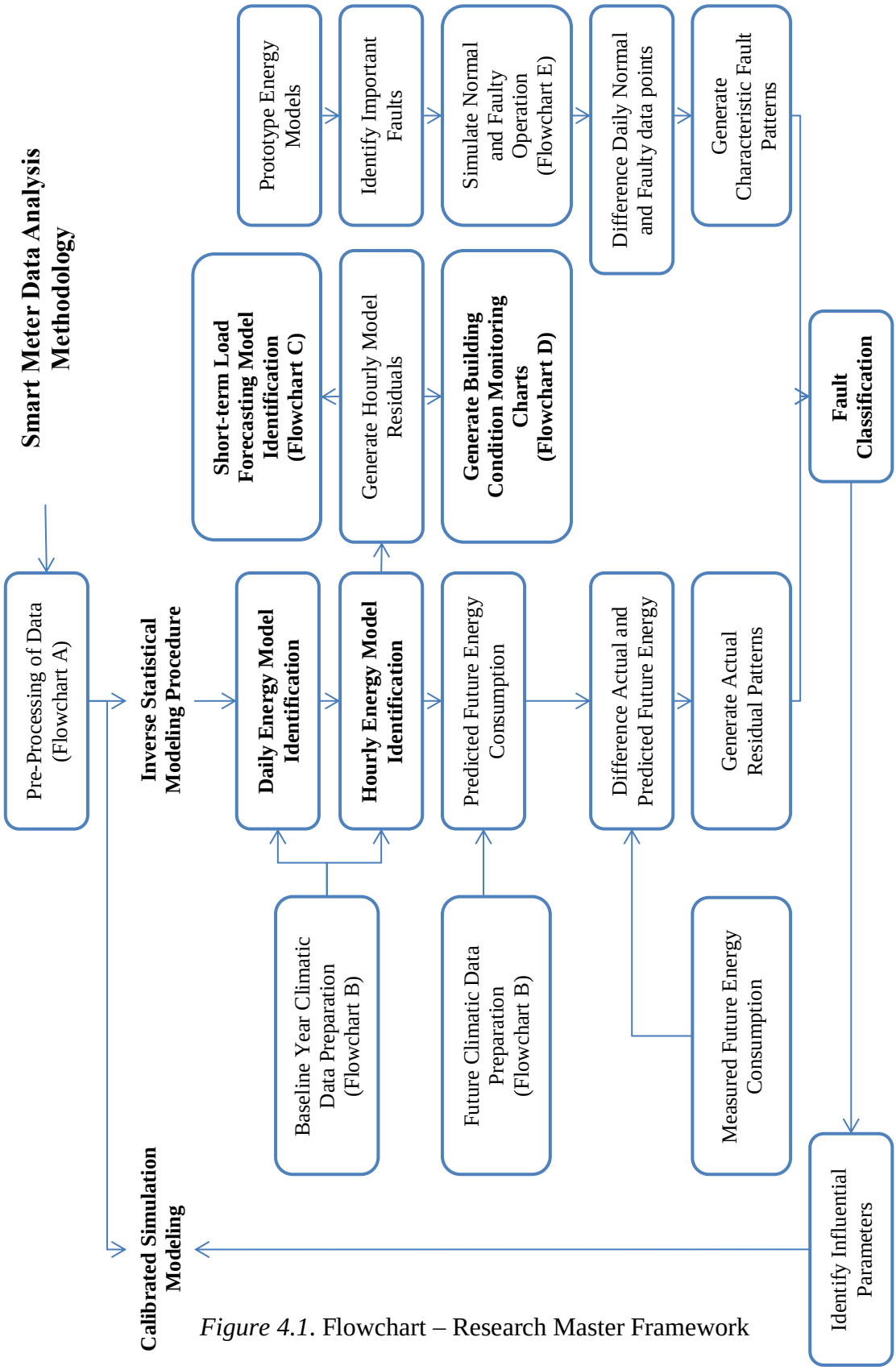


Figure 4.1. Flowchart – Research Master Framework

4.1 Smart Meter Data (Energy Interval Data)

The graphs in Figures 4.2-4.3 help illustrate typical energy consumption patterns of a building as it varies from one day to another during the week and also from one hour to another during the day. Figure 4.2 shows the weekly variations in energy consumption. Weekly energy consumptions are highest during the weekly operations of the building. The consumption drops during Saturdays due to reduced occupancy and operational hours and is lowest during Sundays, and Holidays which is indicative of the minimum base loads of the building, comprising of minimum lighting, equipment etc.

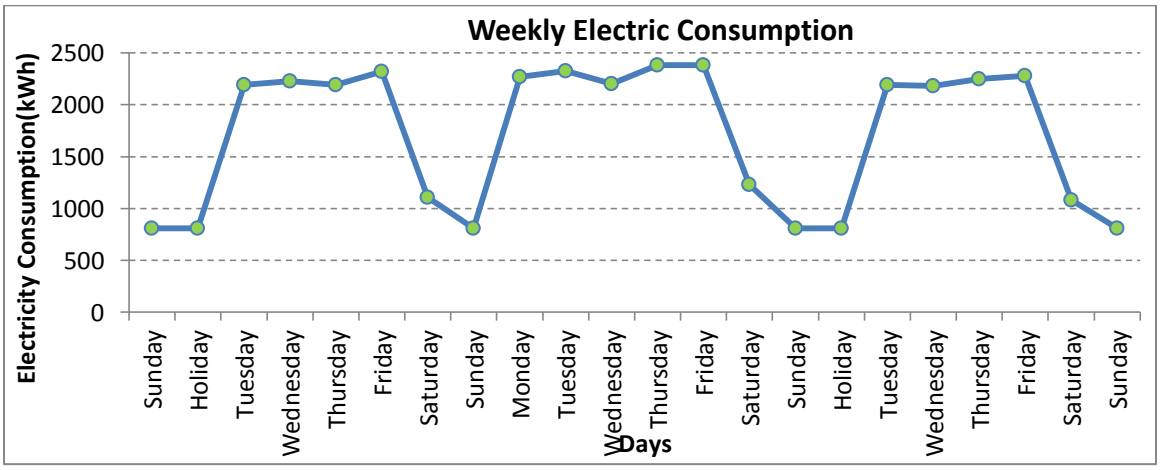


Figure 4.2. Weekly Energy Interval Data

Figure 4.3 exhibits the daily energy consumption variation of a building. The night-time hours have the minimum base load due to lighting and equipment. The energy consumption begins rising around 7:00 A.M. which indicates certain systems coming ON after which there is a steep rise in energy consumption as the building starts to get occupied and the normal functioning begins at around 9:00 A.M. This goes on till the evening hours when the energy consumption begins to decline around 7:00 P.M. when the operations start coming down for the day.

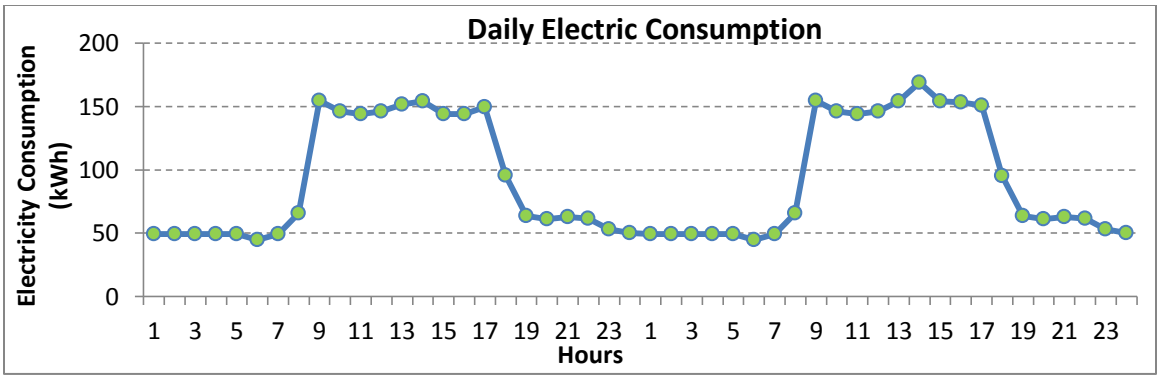


Figure 4.3. Hourly Energy Interval Data

4.2 Data Pre-processing and Preparation

Before any data can be used for modeling purposes, it has to pass through a series of processing and preparation steps that make it suitable for modeling purposes. Figure 4.4 presents the step-by-step flowchart of data processing and preparation needed to be done:

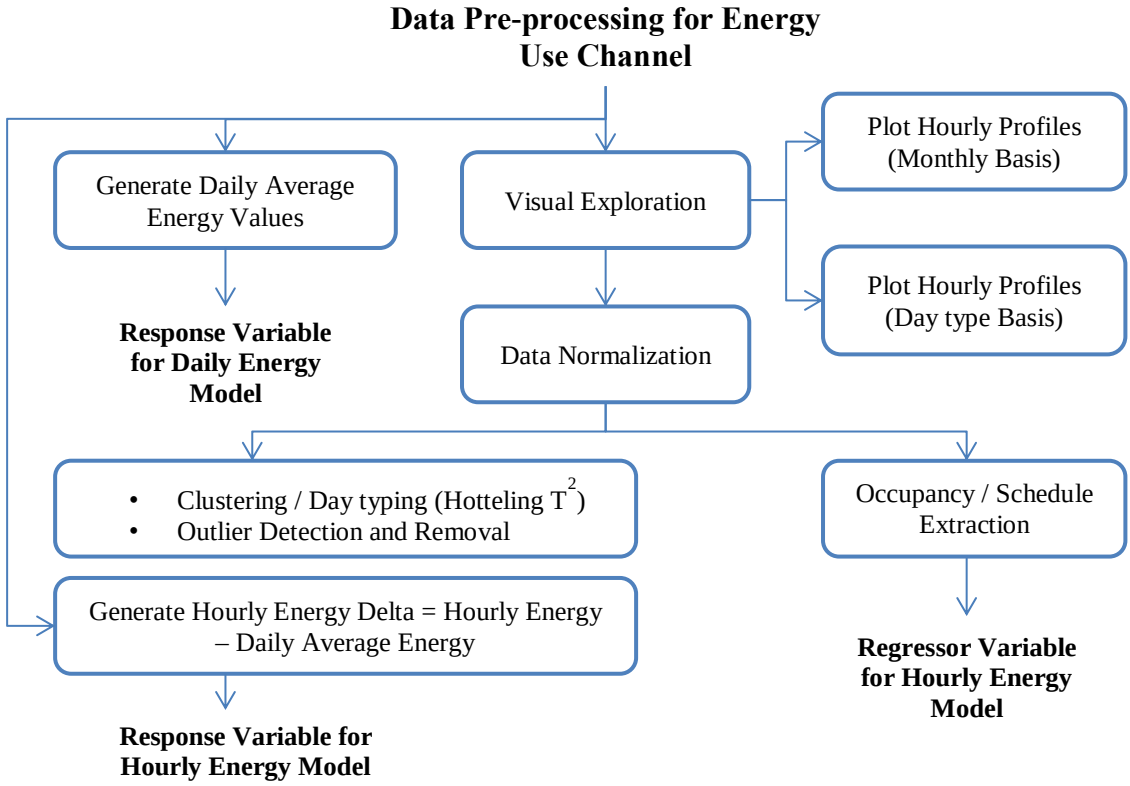


Figure 4.4. Flowchart A – Data Pre-processing for Energy use Channel

4.2.1 Visual Exploration of Data

The first step in data processing is the visual exploration of energy consumption data. Past studies, for example, Capehart (2004), have proposed multiple ways of visually exploring this data such as time series and scatter plots to develop an understanding of the underlying patterns. Time series plots of the energy consumption data were generated first on the monthly basis and finally, based on day types. The monthly plots show the variation in building energy consumption as it shifts from one month to another across the entire year, clearly marking the seasonal shifts in consumption. Once plotted based on day types, these profiles indicate the variation of energy consumption across similar day types, for example, Weekdays, Saturdays etc., as the consumption changes across the entire year. These graphs also show the seasonal shifts in consumption, and help visually identify any outliers, i.e. day profiles with irregular energy consumption patterns. Scatter plots of the energy consumption with the outdoor dry-bulb temperature help visually explore the relationship between the two dependent variables. It also allows for a quick identification of any change-points required for modeling purposes.

In our study, we limit the number of day types to Weekdays, Saturdays and Sundays / Holidays. However, we take into account any specific characteristics within these three broad day types, i.e., there can be two different weekday types or two different Saturday types, based upon some physical significance. This will require clustering of all the 24-dimensional daily energy consumption vectors of all three day types.

4.2.2 Data Normalization

As discussed in Reddy et al. (1997), total energy use in a building, or even a group of buildings is affected by changes in at least the following five sets of parameters:

(a) climatic variables,

(b) conditioned building floor area,

(c) population, (i.e. no. of occupants),

(d) connected loads and operating schedules and,

(e) energy-efficiency and O&M measures.

To be able to cluster energy profiles for all the day types, we would require them to be normalized against all the above parameters, which might lead to over-correction. However, since our clustering strategy uses the 24-dimensional daily energy vectors, occupancy was the most important parameter and we normalized the profiles for this parameter. The following formulae were used to normalize the 24-hourly profiles.

For a Weekday *i*, and hour *j*, the hourly energy use $E'_{i,j}$ is normalized as:

$$E'_{i,j} = \frac{1}{2} \cdot \frac{\mathrm{E}_{i,j}}{\check{\mathrm{E}}_i} \qquad \text{Eq. 4.1}$$

For a Saturday, Sunday or a Holiday*i*, and hour *j*, the hourly energy use $E'_{i,j}$ is normalized as:

$$E'_{i,j} = \frac{1}{2} \cdot \frac{\mathrm{E}_{i,j}}{\check{\mathrm{E}}_{i\ preceding\ Friday}} \qquad \text{Eq. 4.2}$$

where,

$\mathrm{E}_{i,j}$ = actual energy consumption for a given day *i* at a given hour *j*.

$\check{E}_i$ =average energy consumption for the given day

From Figure 4.3, the energy consumption of any hour can lie on either side of the average energy consumption for the given day. Dividing by the average energy consumption for a given day results in certain hours assuming normalized values greater than 1. To correct for this and to restrict the normalized data points to fractional values between 0 and 1, we divide the normalized data point by 2 as shown in Eqs. 4.1 - 4.2.

4.2.3 Clustering / Day typing and Outlier Detection and Removal

As described in Tan, Steinbach, & Kumar (2005), the goal of cluster analysis is to group data or objects such that, the objects within a group be similar (be related) to one another and different from (or unrelated to) the objects in other groups. A related issue is the identification and elimination of outliers, i.e. faulty data. A number of studies have been proposed in the past, such as Seem (2005 & 2007), that deal with the issue of outlier detection and day type clustering. However, there are a number of limitations with these studies. Firstly, the analysis techniques do not go down to the level of examining each of the 24-dimensions of the daily energy vector, but depend upon extracted features such as daily peak and average electricity consumption. Secondly, the user decides how many outliers they would like to remove before one can proceed with clustering. Finally, the clustering technique, agglomerative hierarchical clustering, requires the user to specify the 'stopping rule' which determines when to stop combining the nearest clusters.

After trying various clustering algorithms, DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm was selected because of its following properties and advantages:

Table 4.1

Properties and Advantages of DBSCAN algorithm

• Is a density-based partitional clustering algorithm that assigns each object to a single cluster or group.
• Can handle clusters of various shapes and sizes and is not strongly affected by noise or outliers.
• Makes no assumptions about the distribution of the data.
• Merges clusters that overlap and identifies and separates noise points.
• Automatically determines the number of clusters based on two parameters: *Eps* and *MinPoints*.

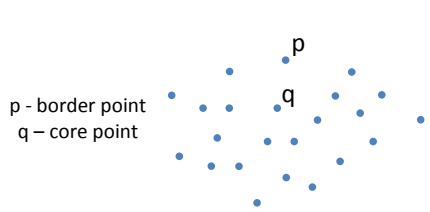
Ester et al. (1996) describe the DBSCAN algorithm. The key idea behind the algorithm is that for each point within a cluster, there will be at least a minimum number of points within a specified radius, i.e., the density in the neighborhood have to exceed some threshold. The following will explain the parameters and the workings of the DBSCAN algorithm (also illustrated in Figure 4.5):

(i) ***Eps*** – neighborhood of a point: The Eps-neighborhood of a point p specifies the density radius within which certain number of points exists.

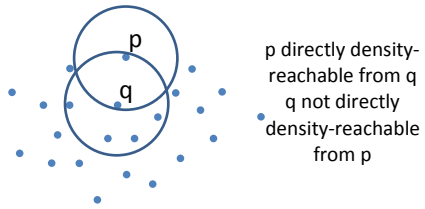
(ii) ***MinPoints*** – minimum number of points in a cluster: There are two kinds of points in any given cluster, the points within a cluster (core points) and the points on the border of the cluster (border points). This parameter specifies the density threshold below which no clusters will be formed. In general, the *Eps*-neighborhood of a border point will contain less number of points than that of

a core point. Hence, the *MinPoints* parameter should be set to a relatively low value in order to include all the points that belong to the same cluster.

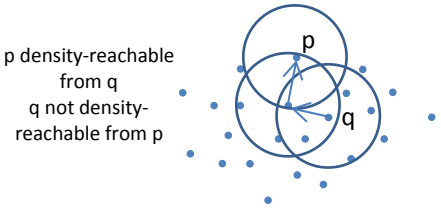
(iii) **Directly Density-Reachable**: The first method of cluster formation, any point p is directly density-reachable from any point q if: (1) point p falls within the *Eps*-neighborhood of point q, and (2) number of points within the *Eps*-neighborhood of point q crosses the *MinPoints* threshold (core point condition). Generally, directly-density reachable is symmetric for pairs of core points. However, it is not symmetric if a core point and a border point are involved.



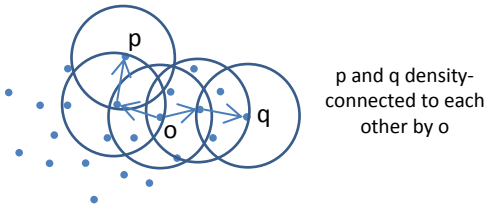
(a) Core points and border points;



(b) Direct-density reachability



(c) Density- reachability;



(d) Density - connectivity

Figure 4.5 (a-d). Working of DBSCAN Algorithm

(iv) **Density – Reachable**: A point p is density-reachable from a point q if there is a chain of points $p_{1,}$ …….. $p_{n,}$ p_1 = q, p_n = p, such that p_{i+1} is directly-density reachable from p_i. Two border points may not be density-reachable from each

other due to the non-fulfillment of the core point condition. However there will be a core point with which both the border points are density-reachable.

(v) **Density-Connected**: A point p is density-connected to another point q if there exists a point o, such that both p and q are density-reachable from o with respect to *MinPoints* and *Eps*-neighborhood conditions.

(vi) **Cluster**: Any cluster C in a database of points D is a non-empty subset of D satisfying the above conditions.

(vii) **Noise**: The points within the database D which do not satisfy any of the above conditions, and do not fall within any cluster C are classified as noise points.

Using the DBSCAN algorithm, clusters are generated and noise points are eliminated. As mentioned earlier in Section 4.2.1, there can be two Weekday clusters and two Saturday clusters; the next step is to verify whether these different clusters within the three broad day types are different enough to warrant inclusion of an indicator variable.

4.2.4 Hoteling T^2 Analysis of different clusters

As described in Reddy (2011), Hotteling T^2 is the extension of the univariate statistical tests applied to evaluate two or more samples to determine whether they originate from populations with: (i) different means, and (ii) different variance / covariance. Let us assume two separate samples of sizes n_1 and n_2. We wish to compare differences between p random variables among the two samples. Let X_1 and X_2 be the mean vectors of the two samples. A pooled estimate of the covariance matrix is:

$$C = \{(n_1 - 1)\, C_1 + (n_2 - 1)\, C_2\} / (n_1 + n_2 - 2) \qquad \text{Eq. 4.3}$$

where, C1 and C2 are the covariance vectors given by:

$$C1 = \begin{bmatrix} c11 & c12 & \cdots & c1p \\ c21 & c22 & \cdots & c2p \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ cp1 & cp2 & \cdots & cpp \end{bmatrix} \quad \text{Eq. 4.4}$$

where, c_{ii} is the variance for parameter i and c_{ik} the covariance for parameters i and k.

Then, the Hotelling T^2 statistic is defined as:

$$T^2 = \frac{n1\ n2\ (\dot{X}1 - \dot{X}2)' C^{-1} (\dot{X}1 - \dot{X}2)}{(n1 + n2)} \quad \text{Eq. 4.5}$$

A large numerical value of this statistic suggests that the two population mean vectors are different. Statistical tests are available for such hypothesis testing.

4.2.5 Occupancy / Schedule extraction

As discussed in Section 3.4.3, occupancy is an important regressor when predicting building energy consumption at higher resolution time scales, such as at the hourly level. The hourly energy consumption profiles are to a large extent an indicator of the occupancy, i.e. number of people inside the building and the connected loads and their operating schedules, as it is the variations in these parameters that changes the profiles to begin with. The hourly occupancy fractions from the Smart Meter Data can be extracted using the Equations 4.1 – 4.2 described earlier, and will serve as an important regressor while building the hourly energy prediction model. The Occupancy Fractions thus generated are a combined effect of Human Occupancy, Lighting Schedules, Equipment Schedules, HVAC Operation Schedules etc.

4.2.6 Energy Data Preparation

The model development methodology proposed is to develop a single energy prediction model with two separate inter-related parts which can be used for daily as well as hourly energy prediction purposes. As is clear from Figure 4.6, the daily prediction portion of the statistical model predicts the average daily energy consumption of the building. Once complete, the hourly portion of the model will predict the delta (or difference) between the average daily and the hourly energy consumption. Upon adding, this will produce the hourly energy consumption of the building. To facilitate this, the energy consumption data is prepared as follows for modeling purposes:

(i) Generate Daily Average Energy data stream – (For Daily Model)

(ii) Generate Hourly Energy Delta data stream – (For Hourly Model)

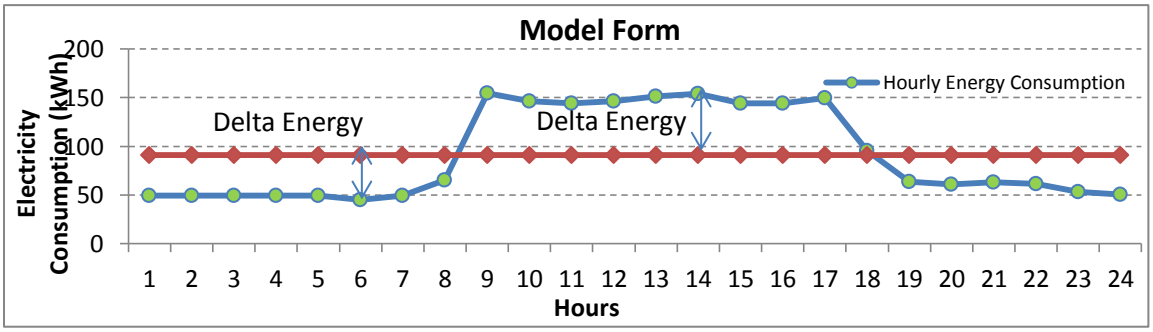


Figure 4.6. Model Form – Hourly Energy Prediction

All of the above steps conclude the energy data pre-processing and preparation methodology as suggested in the flowchart in Figure 4.4.

4.3 Baseline / Future Climatic Data Preparation

Climatic variables, i.e. outdoor dry-bulb temperature, humidity ratio differential, W^{+} = $(W\text{-}0.0008)^{+}$, as mentioned in Reddy (2011)), and total solar horizontal radiation are

used as regressors in daily and hourly regression modeling. As discussed in Section 4.2.6, climate data also needs to be prepared for daily as well as hourly modeling purposes. Figure 4.7 shows the flowchart indicating various steps needed to pre-process the baseline climate data prior to regression modeling.

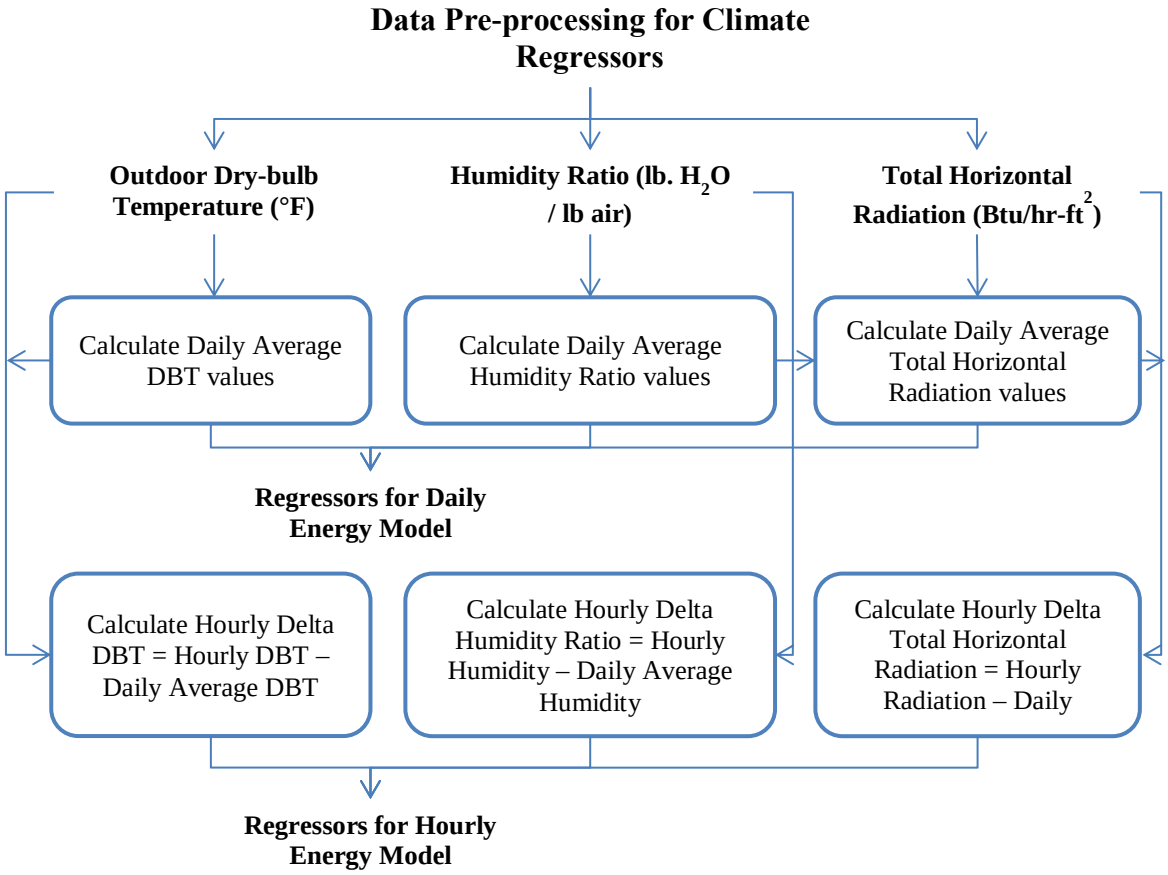


Figure 4.7. Flowchart B – Data Pre-processing for Climate Regressors

Once the modeling is identified, the future climatic data also needs to be processed along the same lines if it is to be used for predictive purposes.

4.4 Daily / Hourly Baseline Statistical Modeling

Based on the discussions of modeling techniques and the related issues in Section 3.4 and its sub-sections and the pre-processing of energy data (response variable) and

climatic data (regressor variables) as per Sections 4.2 and 4.3, regression analysis is carried out to identify daily and hourly energy consumption prediction models. Multiple regression models are built based on different parameters and combination of parameters, multiple change points are accounted for as described in Section 3.4.2 and these models are assessed based on model R^2 and CV-RMSE indices as described earlier in Section 3.4.4. The identified models assume the following functional forms:

$$\check{E}_{i,j} = \{\text{Daily Average Energy Prediction}\}_i + \{\text{Correction for Hour of the day}\}_j + AutoRegressive\ Component \quad \text{Eq. 4.6}$$

where,

i = 1 to 365, index for the day of the year, and

j = 1 to 24, index for the hour of the day.

The elegant model form above can be used for multiple purposes. It can be used for predicting daily average energy consumption, and can then be extended, by adding corrective terms, to model energy consumption for a specific hour. Additionally, by adding AR terms to this model, it can be used for short-term load forecasting.

4.5 Short-term Load Forecasting (STLF) – Modeling the Error Terms

Once the hourly regression modeling is complete, the next step in the analysis is stochastic time series modeling. As discussed in Section 3.7 earlier, we try and model the systematic stochastic component in the residuals generated after building the hourly prediction model, thereby increasing the model prediction accuracies. The steps of the

process are shown in the flowchart in Figure 4.8 below along with a brief discussion of each of the steps.

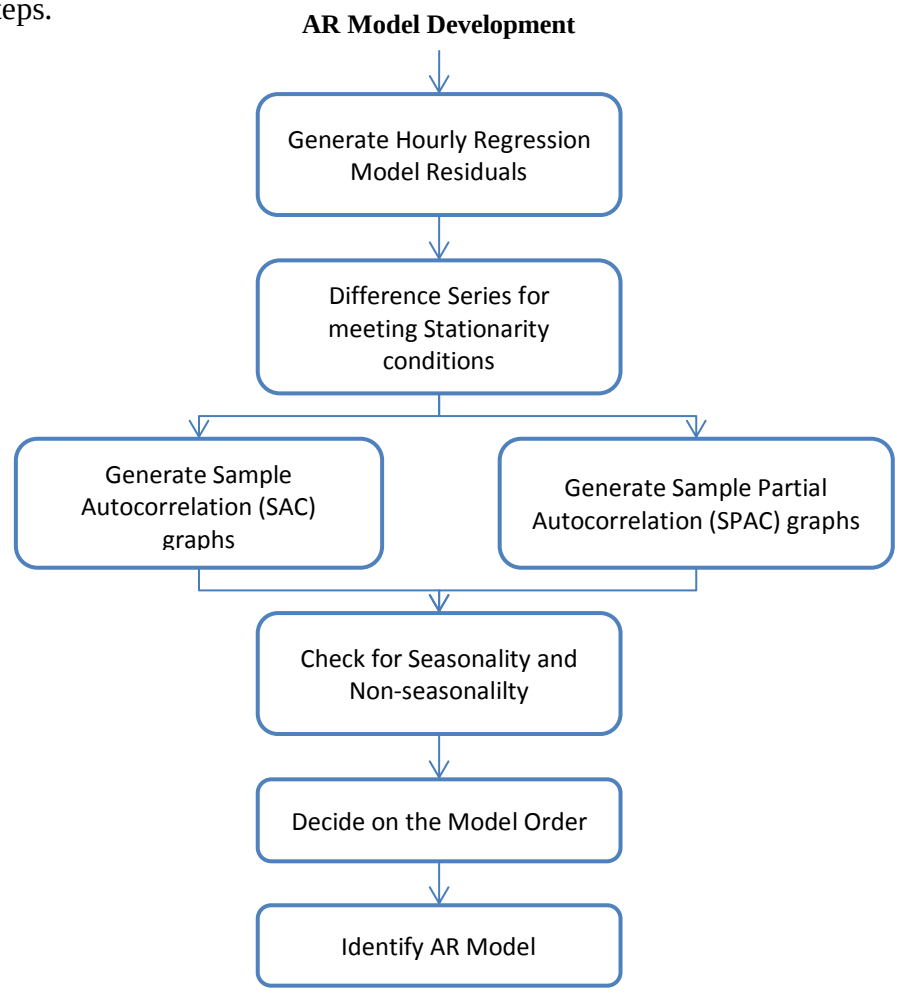


Figure 4.8. Flowchart C – AR Model Development

4.5.1 Differencing: checking for series stationarity

The basis for any time series analysis is a stationary time series (Bisgaard, 2011). It is essentially for stationary time series that we can develop models and forecasts. Thus, it is extremely important that we first determine that the time series we wish to forecast is stationary. As described in Bowerman et al. (2005), a time series is stationary if the statistical properties (for example, the mean and the variance) of the time series are

essentially constant through the series; a plot of data points against time can help us determine if the time series is stationary. If the *n* values seem to fluctuate around a constant mean with a constant variation, it would be reasonable to believe that the time series is stationary and vice versa. When non-stationary, we can convert the time series into a stationary time series by differencing. Bisgaard (2011) proposes multiple ways by which differencing can be done. In this study, we have used the following differencing schemes:

$$\text{(i)} \qquad \mathbf{z_t} = y_t - y_{t-1}; \text{ and, (ii) } \mathbf{z_t} = y_t - y_{t-24}$$

i.e., 1-hour lag and 24-hour lag are used for differencing. The above methods produce a new time series z_t which can be further checked for stationarity. Additionally, the Sample Autocorrelation function can also be used to evaluate the stationarity of a time series.

4.5.2 Sample Autocorrelation (SAC) and Partial Autocorrelation (SPAC)

Box-Jenkins forecasting models can be identified by evaluating the behavior of the Sample Autocorrelation (SAC) and the Partial Autocorrelation (SPAC) functions for the values of a stationary series, z_t, z_{t+1},……., z_n. As defined in Bowerman et al. (2005):

(i) **Sample Autocorrelation (SAC)** measures the linear relationship between time-sequenced observations separated by a lag of k units. It assumes values between 1 and -1. A value close to 1 indicates that observations separated by lag k are linearly correlated with a positive slope, and a value close to -1 indicates that observations separated by the lag k tend to vary linearly with a negative slope. It is given by the formula:

$$r_k = \frac{\sum_{t=b}^{n-k}(z_t - \check{z})(z_{t+k} - \check{z})}{\sum_{t=b}^{n}(z_t - \check{z})^2} \quad \text{Eq. 4.7}$$

where,

$$\check{z} = \frac{\sum_{t=b}^{n} z_t}{(n - b + 1)} \quad \text{Eq. 4.8}$$

SAC can be used to check stationarity of a time series. In general, it can be shown that for non-seasonal data:

- If the SAC of the time series either cuts of fairly quickly, or dies down fairly quickly, as in Figure 4.9, then the time series can be considered stationary.

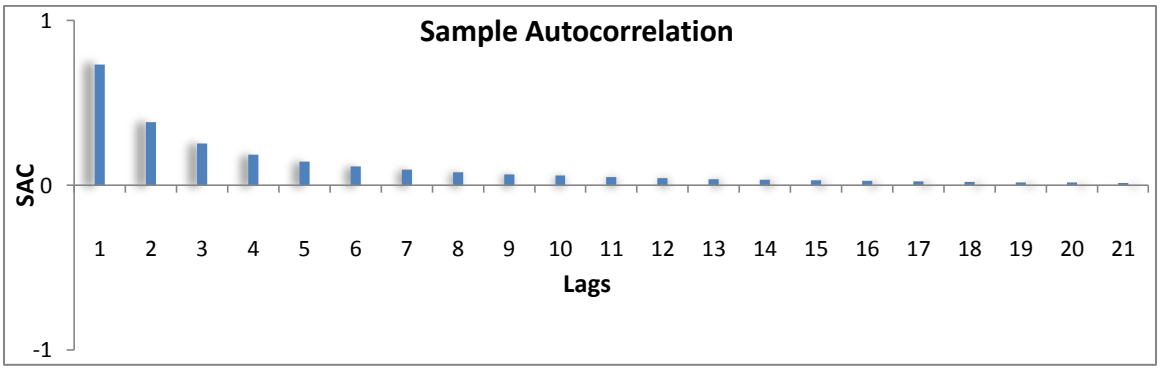


Figure 4.9. SAC – Stationary Time Series

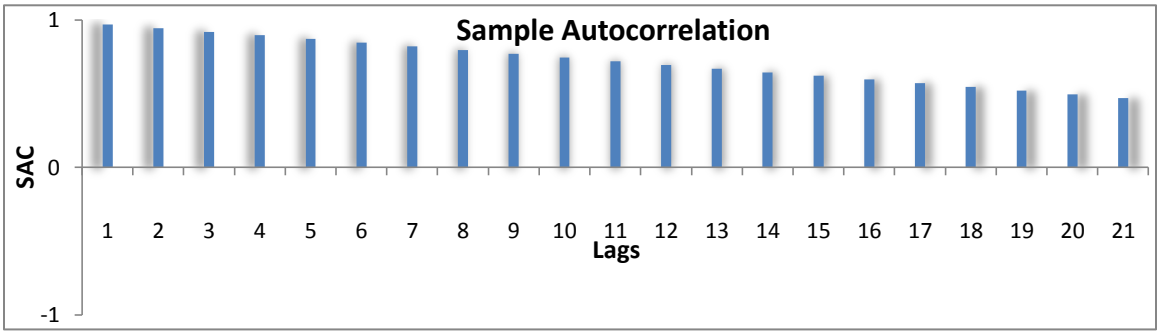


Figure 4.10. SAC – Non-stationary Time Series

- If the SAC of the time series dies down extremely slowly, as in Figure 4.10, then the time series can be considered non-stationary.

(ii) **Sample Partial Autocorrelation (SPAC)** can be thought of as the sample autocorrelation of time series observations separated by lag k units with the effects of the intervening observations eliminated. It is given by the formula:

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j}\, r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j}\, r_j}, \qquad \text{if } k = 2,3,\ldots\ldots \qquad \text{Eq. 4.9}$$

where,

$$r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j} \qquad \text{for } j = 1,2,\ldots\ldots\ k\text{-}1 \qquad \text{Eq. 4.10}$$

SPAC, like the SAC, can exhibit a variety of behaviors such as cutting off abruptly or dying down fairly quickly or extremely slowly; such behaviors are useful for model order identification purposes.

4.5.3 Seasonality and Non-seasonality in Time Series Models

As pointed out in Section 3.7, time series often represent both seasonal and non-seasonal behavior. In the building energy domain, the energy consumption of the building at any given hour during the day is directly related to the preceding hours, but also bears a relation to the same hour on a 24-hour daily cycle. SAC exhibits and help understand this dependence.

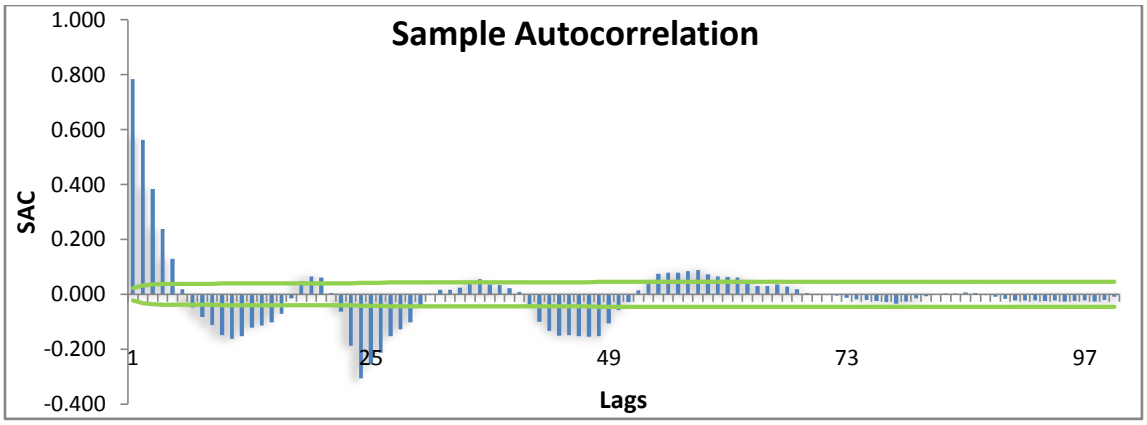


Figure 4.11. SAC - Seasonality and Non-seasonality Check

As seen in Figure 4.11, the series exhibits correlations with observations at lags 1, 2, 3 ….. and so on, which are at the 'non-seasonal' level; it also exhibits correlations with observations at lags 24, 48, 72 ….. and so on, which can be viewed as at the 'seasonal' level.

Finally, once we have finished plotting the SAC and the SPAC, we decide on the model order and model the error structure. Multiple models formed by incorporating different error terms or combination of error terms are evaluated based on the Model CV-RMSE index defined in Section 3.4.4.

4.6 Building Condition Monitoring

Finally, in this section we demonstrate the use of these modeling techniques for generating building condition monitoring charts that can be used for Building Commissioning purposes. Flowchart D in Figure 4.12 describes the various steps to be undertaken to generate these charts.

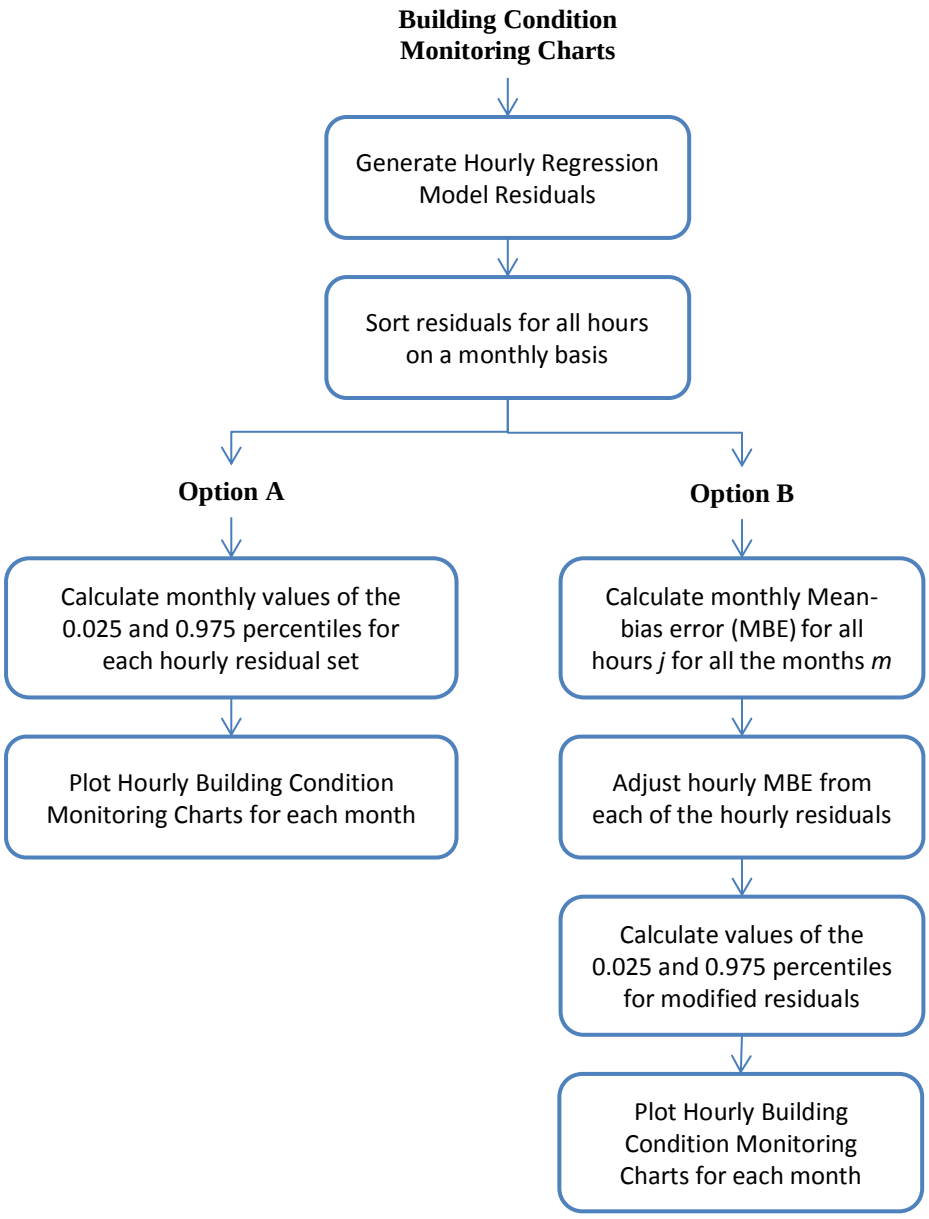


Figure 4.12. Flowchart D – Building Condition Monitoring Charts Generation

In Option A, hourly regression model residuals can be directly used to identify the residual ranges for the condition monitoring charts. However, these would suffer from the fact that the monthly value of Mean-Bias error is used which may introduce biases.

Hence, we adjust the biases from the hourly prediction model by adjusting Eq. 4.6 as follows:

$$\check{E}_{i,j} = \{\text{Daily Average Energy Prediction}\}_i + \{\text{Correction for Hour of the day}\}_j - MBE_{m,j} \quad \text{Eq. 4.11}$$

where,

$MBE_{m,j}$ = Mean bias error for month m at hour j.

The residuals from this corrected model would now need to be plotted on a MBE-corrected condition monitoring chart which forms the second parallel chart generation case described in the flowchart above. These corrected charts would exhibit much tighter residual variances for identifying consumption deviations and generating alarms. This procedure is shown as Option B in Figure 4.12.

CHAPTER 5 : RESULTS – DATA PRE-PROCESSING

The methodology described in Chapter 4 has been applied to two buildings, one synthetic office building in Phoenix, Arizona and another, an actual office building in Denver, Colorado. A complete years' worth of data, i.e. 8760 hourly energy consumption data points were available for these two buildings. In this chapter, we present the results of the data pre-processing steps explained in the flowchart described in Figure 4.4 earlier.

5.1 Buildings' Summary

Table 6.1 below summarizes the key features of the two office buildings. Detailed building descriptions of the buildings can be found in Appendix A.

Table 5.1

Office Buildings Summary

S. No.	Feature	Description	
1	Building Type	Synthetic Office Building (SOB)	Actual Office Building (AOB)
2	Location	Phoenix, AZ	Denver, CO
3	Area (sq.ft.)	53,600 sq.ft.	185,220
4	No. of Floors	3	6 + Lower level
5	Response Data Channel	Whole Building Electric (WBE)	
6	Regressor Data Channels	Dry-bulb Temperature (°F), Humidity Ratio (lb H_2O/lb air), Total Horizontal Solar Radiation (Btu/hr-ft^2)	

As mentioned earlier, 8760 hourly values of whole building electric (WBE) signal are used for analysis purposes and the main regressors used are the climatic variables, i.e. outdoor dry-bulb temperature (°F), humidity ratio (lb H_2O/lb air) and total horizontal

solar radiation (Btu/hr-ft^2). Additionally, an hourly occupancy regressor is identified from the electric Interval Data itself for hourly energy prediction purposes.

The model R^2 and CV-RMSE are used for evaluating various models in terms of their predictive accuracies and form the basis of this evaluation. Models are built and evaluated in two parts: in the first one, models are built on 100% data points and important regressors and their combinations are evaluated; in the second part, the data is divided into 60% training and 40% testing data sets. Models are built on the 60% data set and their predictive accuracies are evaluated on the testing data set kept aside. Following sections will summarize the analysis results for each of the detailed steps described earlier.

5.2 Energy Data Visualization

The first step of the analysis is the visual exploration of the energy data for the given building. We plot the 24-hourly energy profiles on the monthly and day type basis as shown in Figures 5.1 - 5.4 (a-d).

Synthetic Office Building (SOB): Judging from the monthly graphs in Figure 5.1 (a-d), it is clear that the energy consumption of the building cycles throughout the year, with consumption levels being the lowest during the winter month of January, transitioning to the higher levels of July through April and again lowering through the month of November before hitting the lowest again during the winters. The different profile types in the monthly graphs are a result of the different day types within the month, i.e. Weekdays, Saturday and Sundays. These become clearer as we segregate the profiles based on day types in the next set of graphs shown below. Being a synthetic office

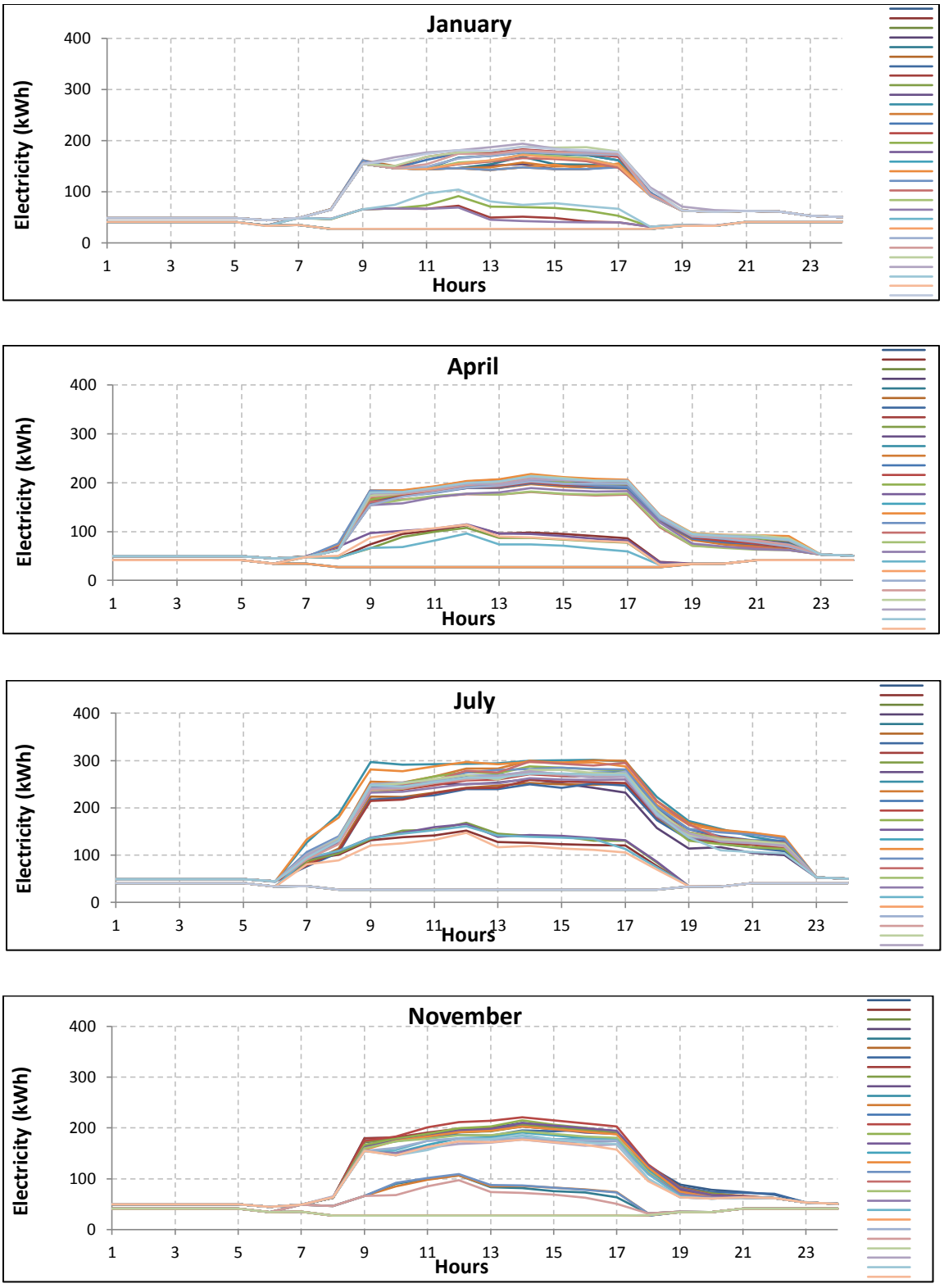


Figure 5.1(a-d). SOB Energy Interval Data (Monthly Basis)

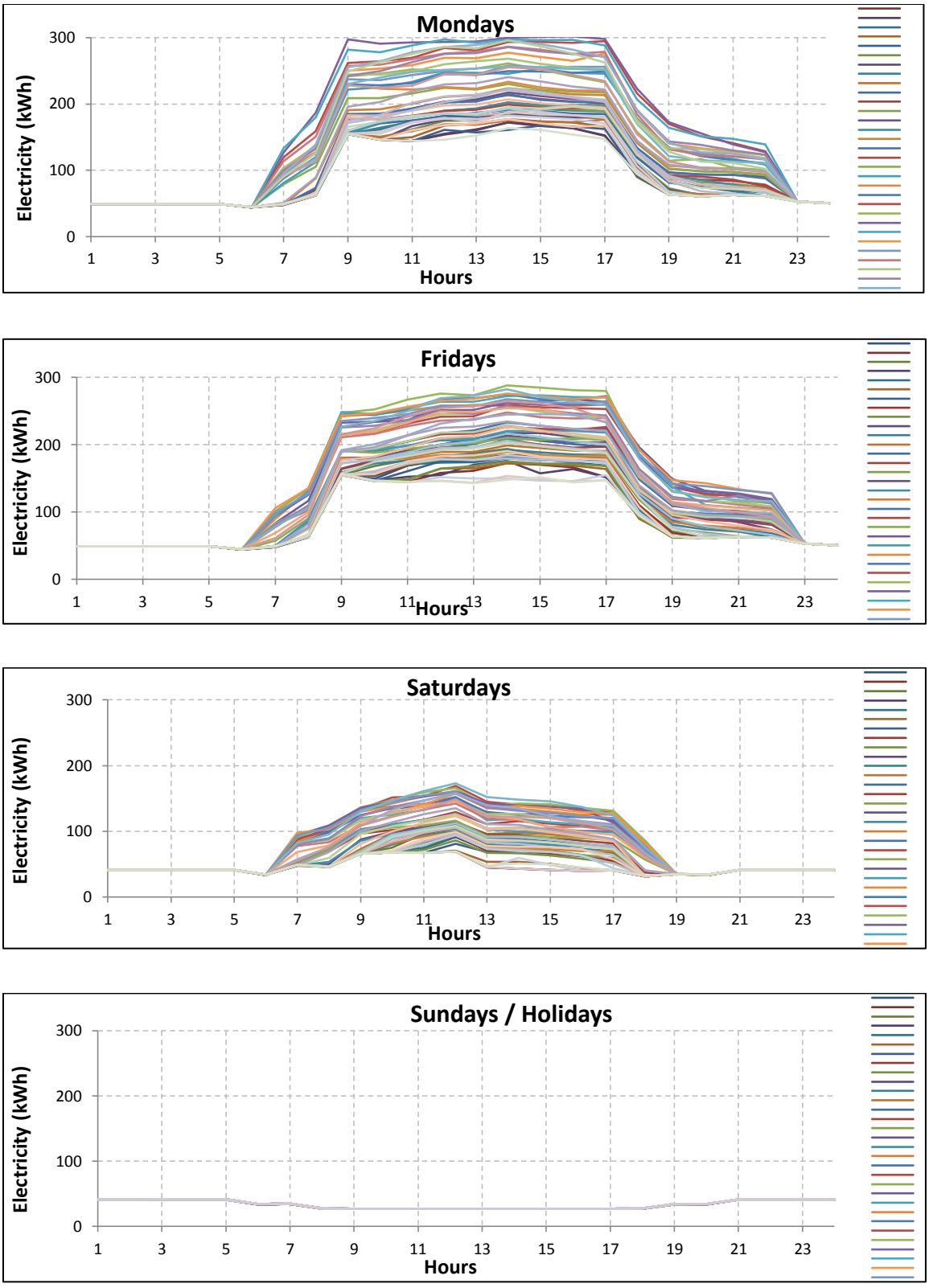
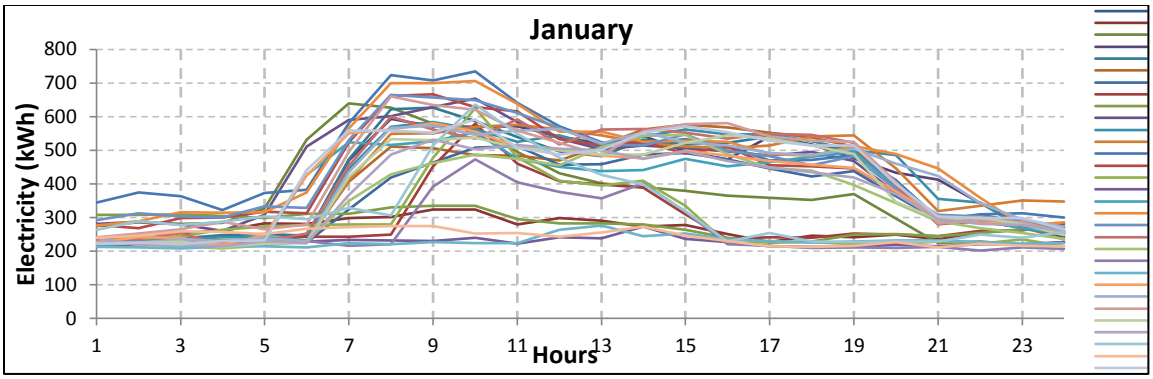


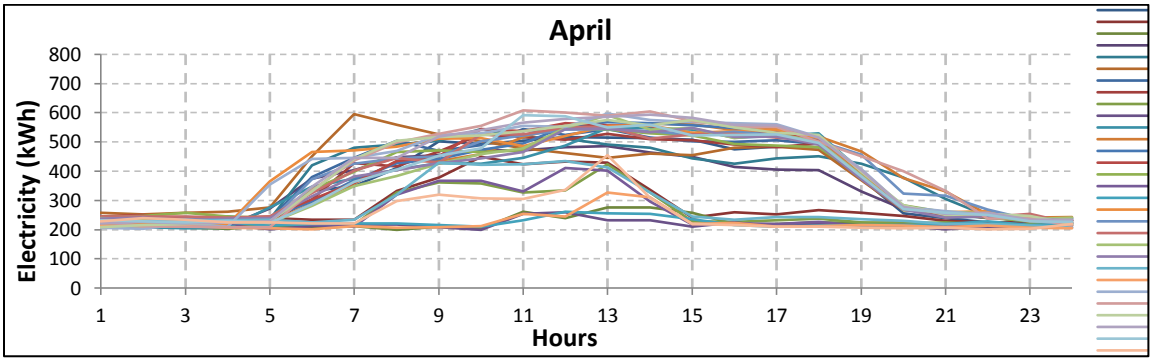
Figure 5.2(a-d). SOB Energy Interval Data (Day type Basis)

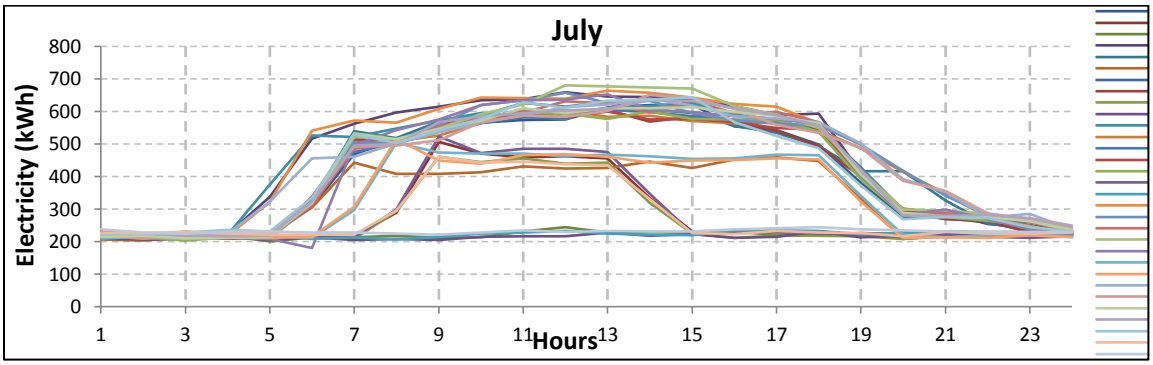
building, the energy profiles are very well-behaved unlike the actual building profiles shown in Figures 5.2 (a-d), that exhibit a lot of variations in energy consumptions. This is because of the synthetic nature of the building, which was developed in eQuest and simulated, and the resulting energy consumption points are extracted and presented here.

Figures 5.2 (a-d) shows the variations in energy consumption profiles based on different days of the week. The graphs represent 52 Monday, Friday, Saturday and Sunday / Holiday profiles in the entire year; the variations in the profiles are a result of the different seasons that the building sees throughout the year. As such, segregating the profiles on day types helps identify any outliers which should be eliminated before we begin building the daily energy prediction model.

Actual Office Building (AOB): Looking at the monthly graphs in Figures 5.3 (a-d) and comparing it to the synthetic building monthly graphs, it is clear that the energy consumption of the building stays constant throughout the year and does not cycle based on the different seasons that the building sees. Also, this building has much higher base load consumption as compared to the synthetic building, which is evident from the amount of electricity the building consumes during the non-operational hours. Hence, it would be safe to assume that this building is a weather-independent, internal-load dependent building. Again, the different profile types in the monthly graphs are a result of the different day types within the month, i.e. Weekdays, Saturday and Sundays which become clearer as we segregate the profiles based on day types in the next set of graphs shown in Figures 5.4 (a-d).







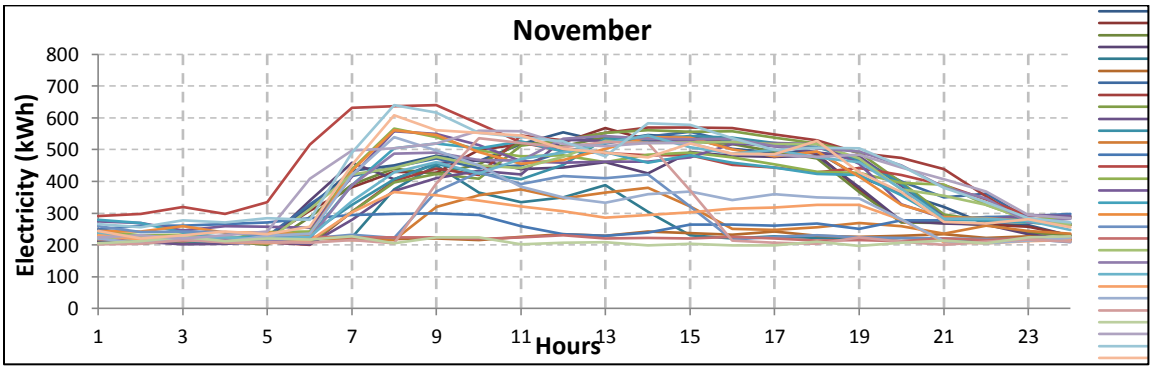
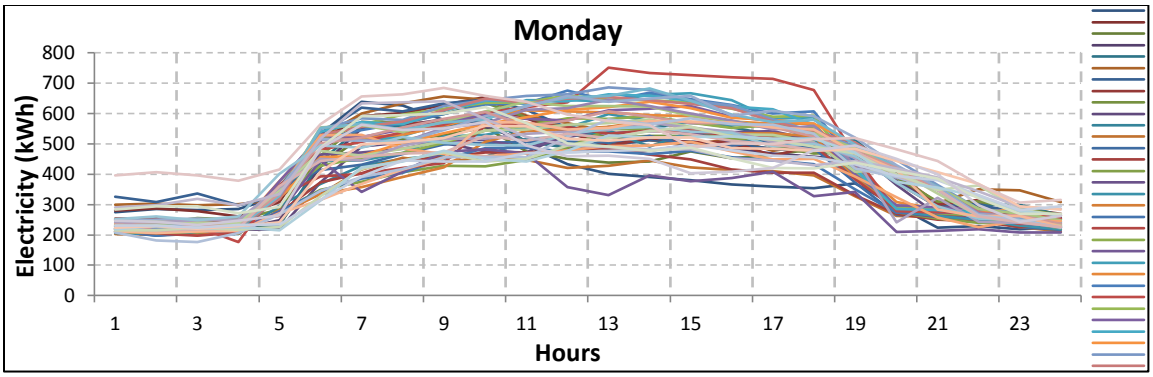
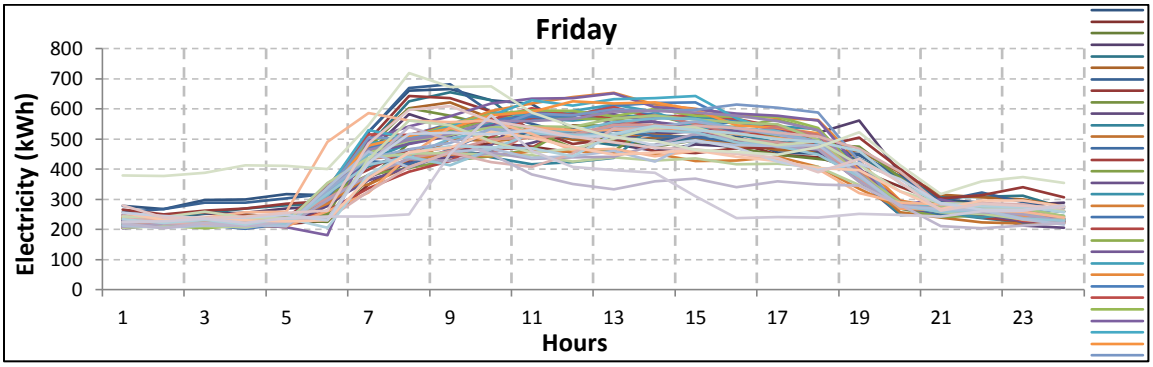
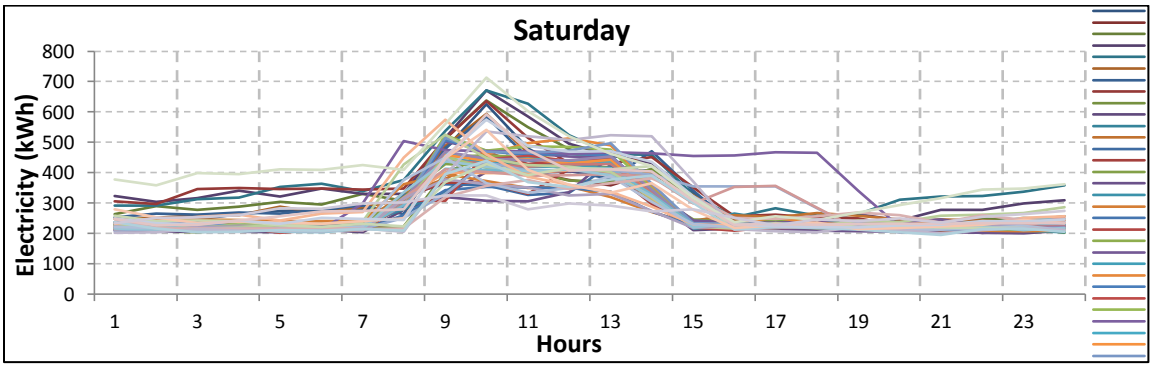


Figure 5.3(a-d). AOB Energy Interval Data (Monthly Basis)







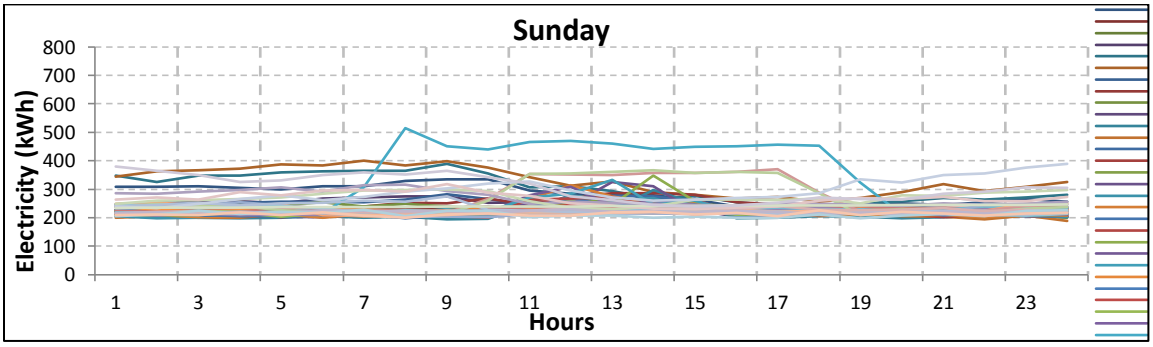


Figure 5.4(a-d). AOB Energy Interval Data (Day type Basis)

Comparing the monthly and daily energy consumption profiles of the two buildings, we arrive at two important conclusions:

(i) Actual buildings would tend to exhibit far more variations in their energy consumption profiles as compared to the profiles of synthetic buildings.

(ii) Sorting the energy profiles based on different day types very clearly indicates the outliers which are almost non-existent in a synthetic building. These profiles are of interest as these represent deviations from the general energy consumption behaviors which could be attributed to an operational change or the result of a fault occurrence in one of the many building systems.

5.3 Data Normalization and Clustering

The daily energy profiles are normalized as mentioned in Section 4.2.2 and the normalized profiles are used for clustering. To be able to identify a generalized energy prediction model, it is important to identify the different day types within the available data. Clustering helps identify the optimum number of day types and separates the noise or outlier points. We utilize the DBSCAN algorithm, described in Section 4.2.3, for clustering purposes. Clustering with DBSCAN requires two parameters to be determined, *MinPoints* (minimum number of points required for a cluster to be formed) and *Epsilon* (density radius). Different values for both of these parameters were evaluated, and the resulting confusion matrices were studied to arrive at the optimal clustering keeping in mind the following objectives:

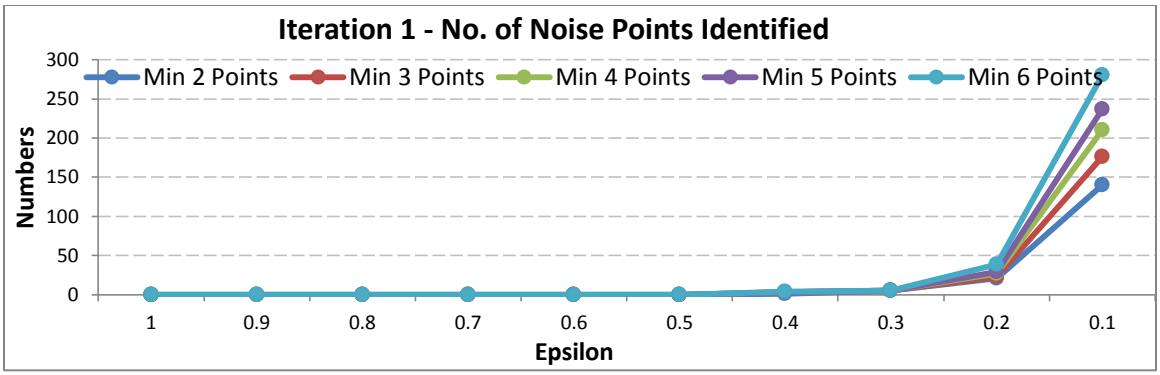
(i) Fewer numbers of Outliers is preferable as we did not want to remove a lot of data points.

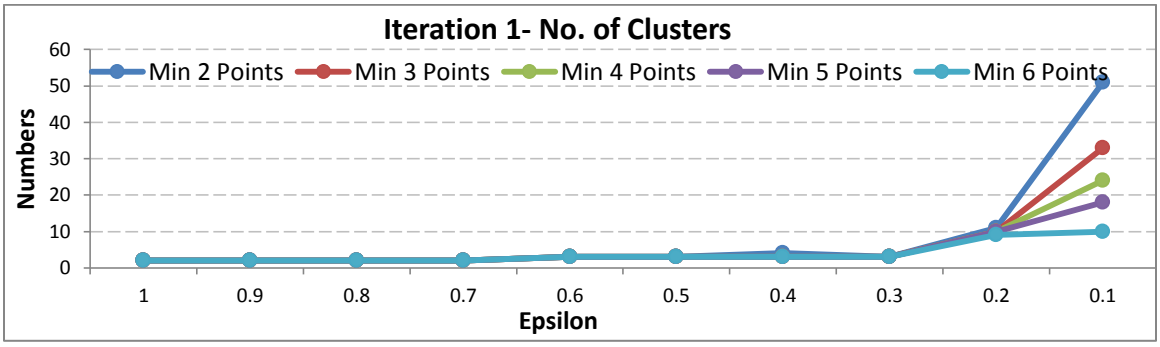
(ii) Fewer numbers of Clusters is preferable so that we could arrive at a more generalized model that can predict energy consumption for most of the days in a given year.

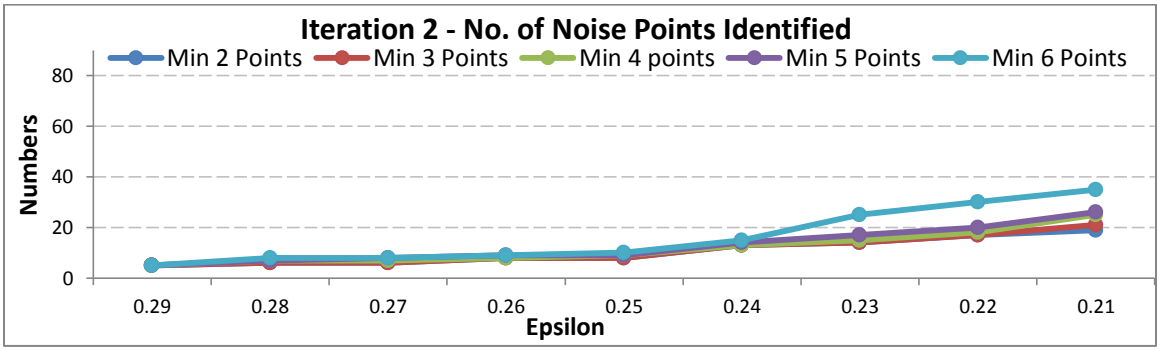
It is important to understand the importance of the above mentioned points as these directly affect the energy prediction process. A large number of outliers represents that the building is not properly operated and sees a lot of variation in energy consumption patterns. Hence, it might be difficult to identify a generalized energy prediction model for such a building. On the other hand, too many clusters also represent a irregularly operated building with large variations in energy consumption. Additionally, this may also affect the identified model accuracy as many different day types have been accounted for. Smaller number of clusters is preferable as this indicates regular building operation with lesser variations and also helps identify a more robust model based on lesser number of day types.

Two separate iterations were carried out for the *Eps* parameter. Iteration 1 involved taking large increments of *Eps* to check where the clusters actually started forming. Eventually, Iteration 2 is carried out with *Eps* sub-step values for the identified range in Iteration 1 that helps decide on the optimal value of this parameter. We pair these values of *Eps* with different values of the parameter *MinPoints* ranging from 1 to 6 and study the resulting confusion matrices.

Figures 5.5 - 5.6 illustrate the changes in the objectives mentioned above as we decide on the final values of the two clustering parameters.







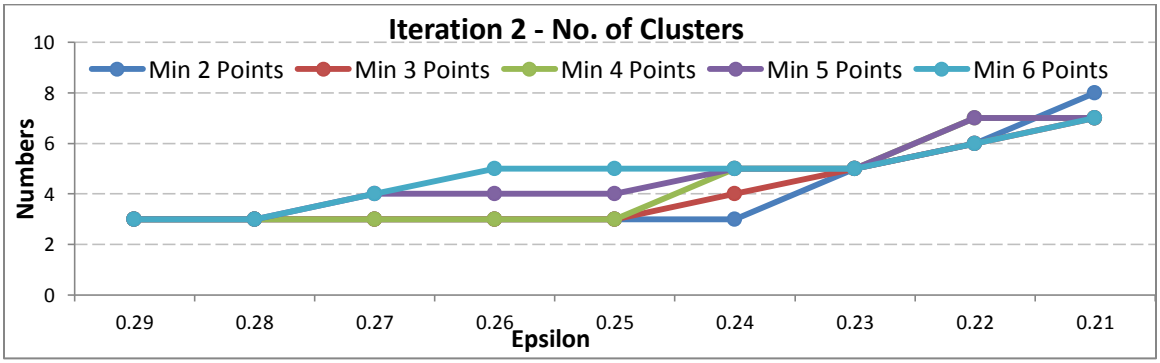


Figure 5.5(a-d). SOB Clustering Iterations

Synthetic Office Building (SOB): Looking at Iteration 1 from the Figures 5.5 (a-d), we notice that the clustering within each of the three day types only happens as we change the value of *Eps* from 0.3 to 0.2 also shown in the confusion matrices in Table 5.2. For *Eps* value 0.3 in Iteration 1, the clustering returns a perfect result, with each of the profiles being clearly identified and a nil misclassification error. Finally, when *Eps* is changed to 0.2, 10 different clusters are formed and many noise points are identified. Additionally, we combine these with values of *MinPoints* ranging from 1 to 6. The noticeable trend is that as the value of this parameter increases from 1 to 6, the number of clusters reduces from 51 to 10, but the number of noise points increases from 140 to 280, and vice versa. This is obvious as a larger value of this parameter results in points being unable to form clusters, and are then identified as noise points.

In Iteration 2, values of *Eps* were changed from 0.29 to 0.21 and different combinations with *MinPoints* assuming values from 1 to 6 were carefully examined using the resulting confusion matrices for each case. Looking at the confusion matrices in Table 5.3, at *Eps* value of 0.25, the clustering still returns a perfect result, only when its value is changed to 0.24, an additional cluster is formed with22 Saturdays during the hot summer months. The *MinPoints* parameter exhibits the same trend described above. In Iteration 2, as the value of this parameter is changed from 1 to 6, the number of clusters reduces from 8 to 7, and the number of noise points increases from 19 to 35. We finally select the case Eps-0.24, MinPoints-3, because the number of clusters is 4 (all Saturdays during the summer being the additional cluster), and the number of noise points is much smaller. Other clusters also resulted in number of clusters being 4 with lesser noise points, but they were not as meaningful. Additionally, it is found that the NOISE points identified by

Table 5.2

SOB Confusion Matrices (Iteration 1)

CONFUSION MATRIX (Iteration 1; *MinPoints* – 3; *Epsilon* – 0.3)											
Unclustered Instances:			5		Class Attribute:			Day type 2			
Classes to Cluster:											
0	1	2	←	Assigned to cluster							
63	0	0	|	Sunday							
0	245	0	|	Weekday							
0	0	52	|	Saturday							
Cluster 0	←	Sunday									
Cluster 1	←	Weekday									
Cluster 2	←	Saturday									
Incorrectly Clustered Instances:				0							
CONFUSION MATRIX (Iteration 1; *MinPoints* – 3; *Epsilon* – 0.2)											
Unclustered Instances:			23		Class Attribute:			Day Type 2			
Classes to Cluster:											
0	1	2	3	4	5	6	7	8	9	←	Assigned to
6	0	40	0	0	0	0	1	0	0	|	Sunday
0	2	0	0	0	0	0	0	0	0	|	Weekday
0	0	0	1	11	3	10	0	7	5	|	Saturday
Cluster 0	←	No Class	Cluster 4	←	Saturday	Cluster 8	←	No Class			
Cluster 1	←	Weekday	Cluster 5	←	No Class	Cluster 9	←	No Class			
Cluster 2	←	Sunday	Cluster 6	←	No Class						
Cluster 3	←	No Class	Cluster 7	←	No Class						
Incorrectly Clustered:			58								

Table 5.3

SOB Confusion Matrices (Iteration 2)

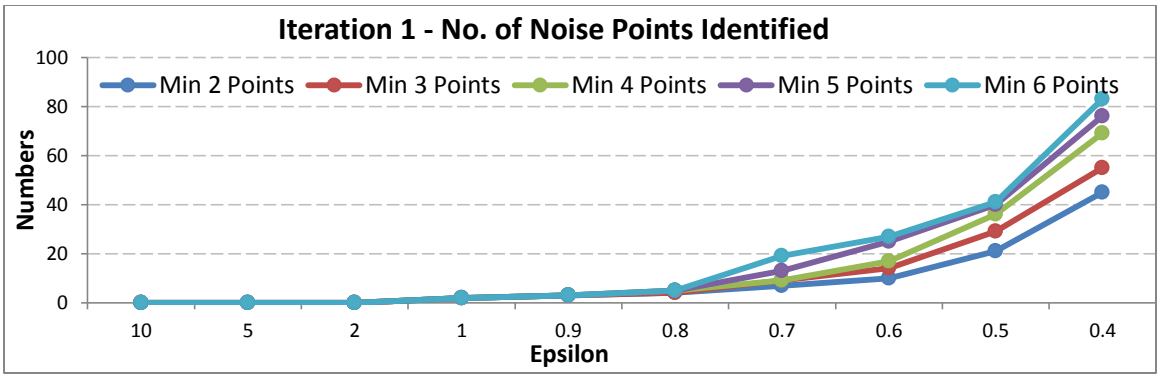
CONFUSION MATRIX (Iteration 2; *MinPoints* – 3; *Epsilon* – 0.25)					
Unclustered Instances:			8	Class Attribute:	Day Type 2
Classes to Cluster:					
0	1	2	←	Assigned to cluster	
63	0	0	|	Sunday	
0	242	0	|	Weekday	
0	0	52	|	Saturday	
Cluster 0	←	Sunday			
Cluster 1	←	Weekday			
Cluster 2	←	Saturday			
Incorrectly Clustered Instances:			0		

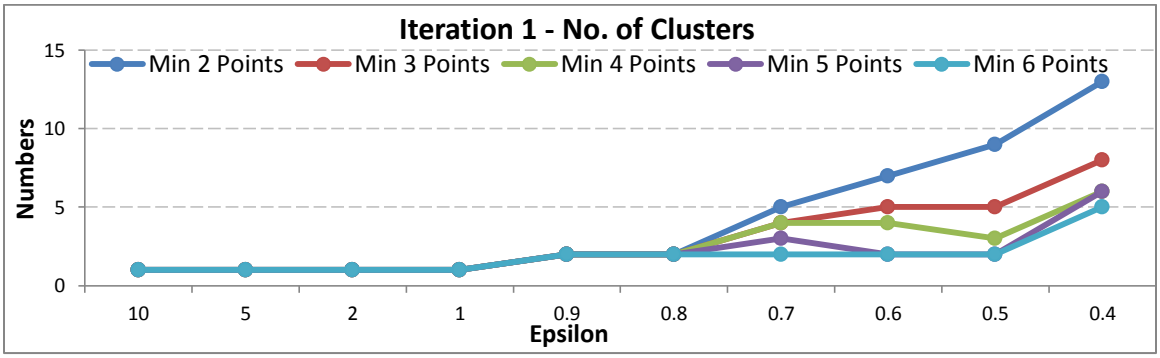
CONFUSION MATRIX (Iteration 2; *MinPoints* – 3; *Epsilon* – 0.24)					
Unclustered Instances:			15	Class	Day Type 2
Classes to Cluster:					
0	1	2	3	←	Assigned to cluster
63	0	0	0	|	Sunday
0	235	0	0	|	Weekday
0	0	30	22	|	Saturday
Cluster 0	←	Sunday			
Cluster 1	←	Weekday			
Cluster 2	←	Saturday			
Cluster 3	←	No Class			
Incorrectly Clustered Instances:			22		

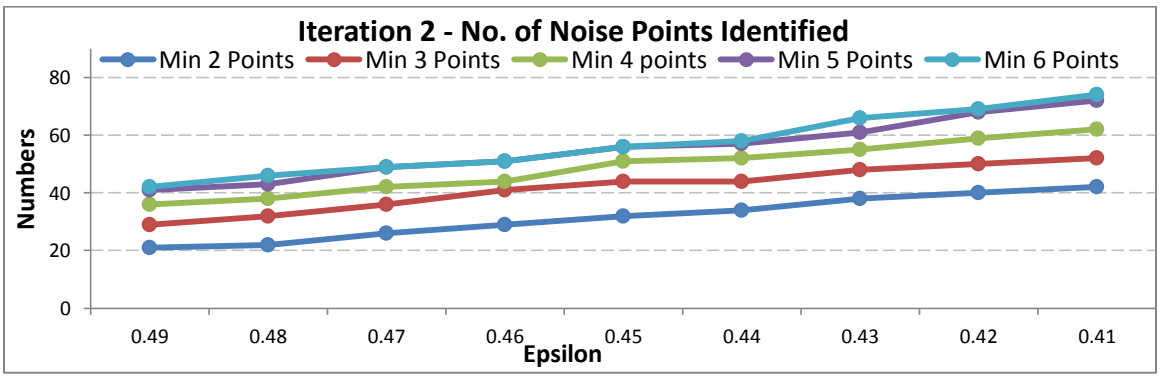
the clustering algorithm were mostly Mondays and Tuesdays (when Mondays were holidays) during the summer months of May, June, July and August when there was a sudden spike of energy consumption in the mornings when the HVAC system came ON. This is attributable to the fact that the HVAC system was scheduled to stay OFF during the weekend and hence the building would retain the heat which had to be removed on Monday mornings to make the building ready for occupancy.

Actual Office Building (AOB): Similar to the synthetic building, two clustering iterations were carried out for the actual office building as well. Looking at Iteration 1 graphs from Figures 5.6 (a-d) and comparing with the confusion matrices from Table 5.4, it is clear that Saturday profiles are clustered separately from Sundays as we change the Eps value from 0.5 to 0.4 Hence, this identifies the range of this parameter for Iteration 2. For the *MinPoints* parameter, as the value of this parameter increases from 1 to 6, the number of clusters reduces from 13 to 5, but the number of noise points increases from 45 to 83, and vice versa.

In Iteration 2, values of *Eps* were changed from 0.49 to 0.41 and different confusion matrices with *MinPoints* assuming values from 1 to 6 were carefully examined. Although Saturdays are separated from Sundays at *Eps*-0.49, we select the parameter values *Eps*-0.43 and *MinPoints*-3. This is because at Eps-0.43 value, a separate cluster is found that identifies 10 consecutive summer Mondays wherein the building started operations early. The *MinPoints* parameter exhibits similar trends with number of clusters reducing from 12 to 5 and noise points increasing from 42 to 74 as *MinPoints* value changes from 1 to 6. We trade a slightly more number of noise points against a smaller number of clusters by selecting *MinPoints*-3.







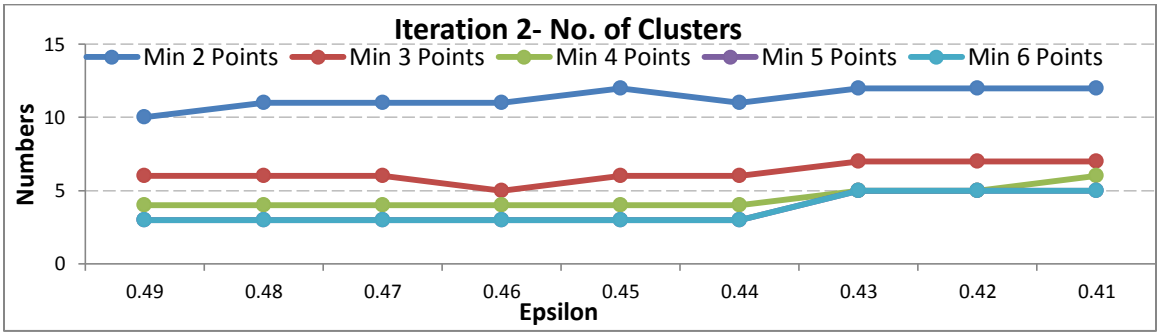


Figure 5.6(a-d). AOB Clustering Iterations

Table 5.4

AOB Confusion Matrices (Iteration 1)

CONFUSION MATRIX (Iteration 1; *MinPoints* – 3; *Epsilon* – 0.5)						
Unclustered Instances:			29	Class Attribute:		Day Type 2
Classes to Cluster:						
0	1	2	3	4	←	Assigned to cluster
44	1	4	0	0	|	Sunday
0	244	0	0	3	|	Weekday
38	0	0	3	0	|	Saturday
Cluster 0	←	Sunday	Cluster 4	←	No Class	
Cluster 1	←	Weekday				
Cluster 2	←	No Class				
Cluster 3	←	Saturday				
Incorrectly Clustered:		46				

CONFUSION MATRIX (Iteration 1; *MinPoints* – 3; *Epsilon* – 0.4)									
Unclustered Instances:			55	Class Attribute:			Day Type 2		
Classes to Cluster:									
0	1	2	3	4	5	6	7	←	Assigned to cluster
41	1	0	0	0	4	3	0	|	Sunday
0	212	0	0	10	0	0	3	|	Weekday
0	0	28	9	0	0	0	0	|	Saturday
Cluster 0	←	Sunday	Cluster 3	←	No Class	Cluster 6	←	No Class	
Cluster 1	←	Weekday	Cluster 4	←	No Class	Cluster 7	←	No Class	
Cluster 2	←	Saturday	Cluster 5	←	No Class				
Incorrectly Clustered: 30									

Table 5.5

AOB Confusion Matrices (Iteration 2)

CONFUSION MATRIX (Iteration 2; *MinPoints* – 3; *Epsilon* – 0.44)							
Unclustered Instances:			44	Class Attribute:			Day Type 2
Classes to Cluster:							
0	1	2	3	4	5	←	Assigned to cluster
41	1	0	4	3	0	|	Sunday
0	233	0	0	0	3	|	Weekday
0	0	37	0	0	0	|	Saturday
Cluster 0	←	Sunday		Cluster 3	←	No Class	
Cluster 1	←	Weekday		Cluster 4	←	No Class	
Cluster 2	←	Saturday		Cluster 5	←	No Class	
Incorrectly Clustered:		30					

CONFUSION MATRIX (Iteration 2; *MinPoints* – 3; *Epsilon* – 0.43)								
Unclustered Instances:			48	Class Attribute:			Day Type 2	
Classes to Cluster:								
0	1	2	3	4	5	6	←	Assigned to cluster
41	1	0	0	4	3	0	|	Sunday
0	219	0	10	0	0	3	|	Weekday
0	0	37	0	0	0	0	|	Saturday
Cluster 0	←	Sunday		Cluster 4	←	No Class		
Cluster 1	←	Weekday		Cluster 5	←	No Class		
Cluster 2	←	Saturday		Cluster 6	←	No Class		
Cluster 3	←	No Class		Incorrectly Clustered:		21		

Comparing the two clustering iteration results of the two buildings from Figures 5.5 and 5.6 (a-d), we conclude the following:

(i) Actual building has higher number of clusters owing to variations in the way actual buildings are operated.

(ii) Actual buildings also have higher number of outliers due to the deviations from normal operations observed. These could be a result of a planned operational change or the occurrence of a fault in any of the building systems.

(iii) A smaller value of *Eps* parameter can result in optimal clustering for a synthetic building due to well-behaved energy profiles, whereas, in an actual building, this value needs to be higher to accommodate minor variations between regularly operated days to be clustered together.

(iv) Finally, looking at the graphs and resulting confusion matrices in Tables 5.2 – 5.5, it is clear that multiple outcomes are possible in clustering and it is best left to user interpretation.

5.4 Daily and Hourly Energy Consumption Distribution

Once clustering of the energy profiles is complete, we study the distribution of the daily and hourly energy consumption profiles for all the days within the various identified clusters.

Synthetic Office Building (SOB): First, we plot the daily energy consumption distribution of the days within different clusters:

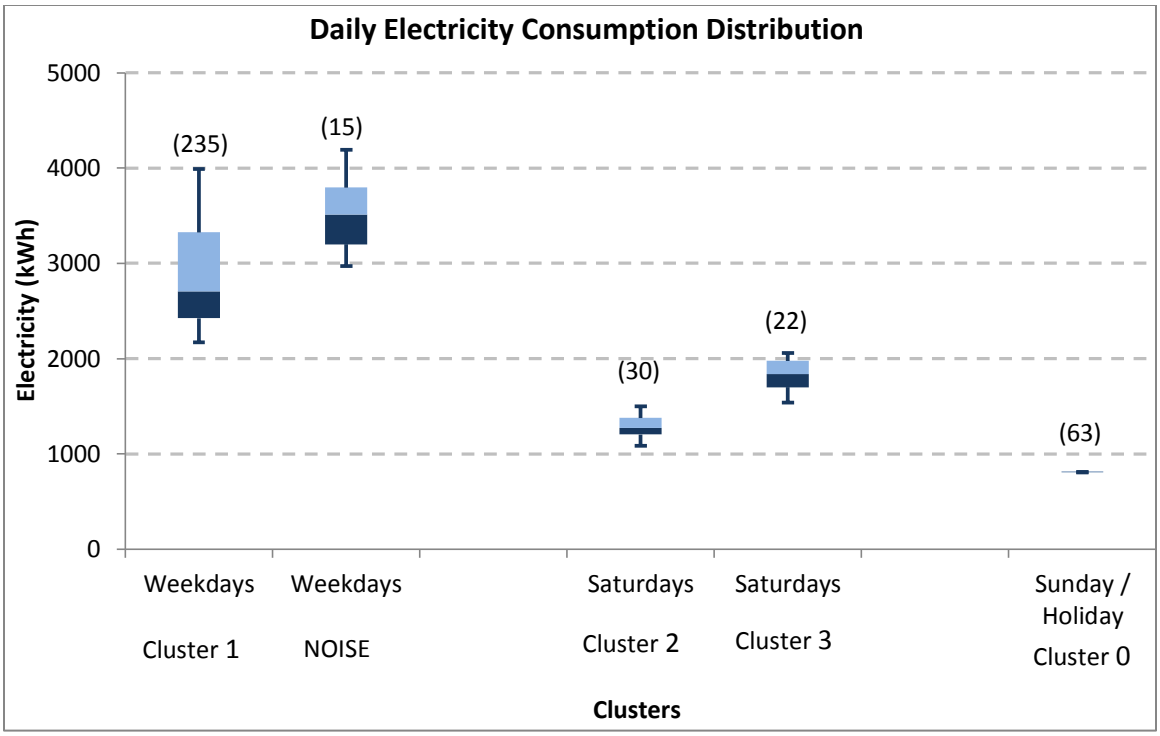


Figure 5.7. SOB - Daily Energy Consumption Distribution of various Clusters

It is important to note that Figure 5.7 represents the actual daily energy consumption of the days within the cluster and the wide distributions during the day types are a result of the temperature effects. As such, the current CV values of these clusters are given in Table 5.6:

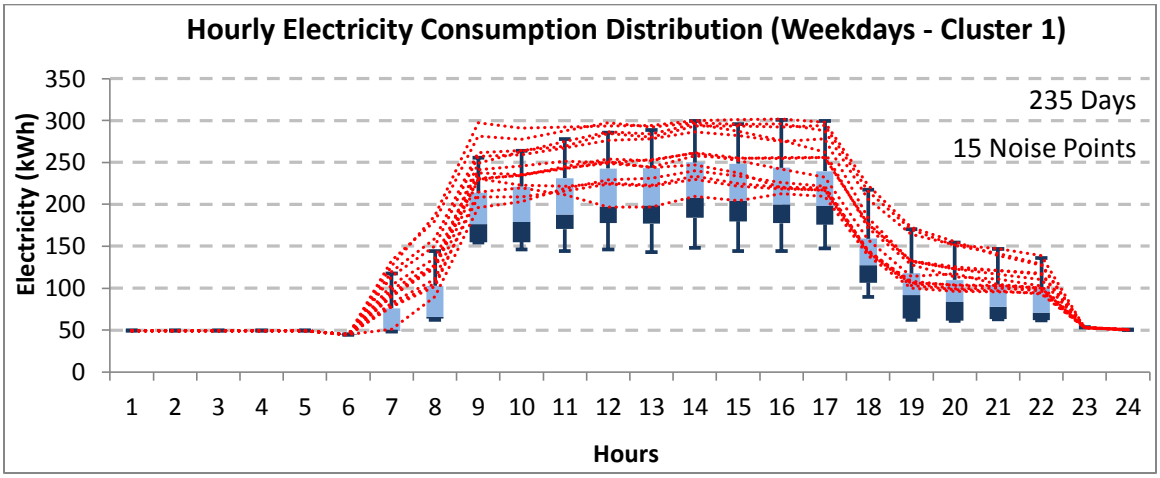
Table 5.6

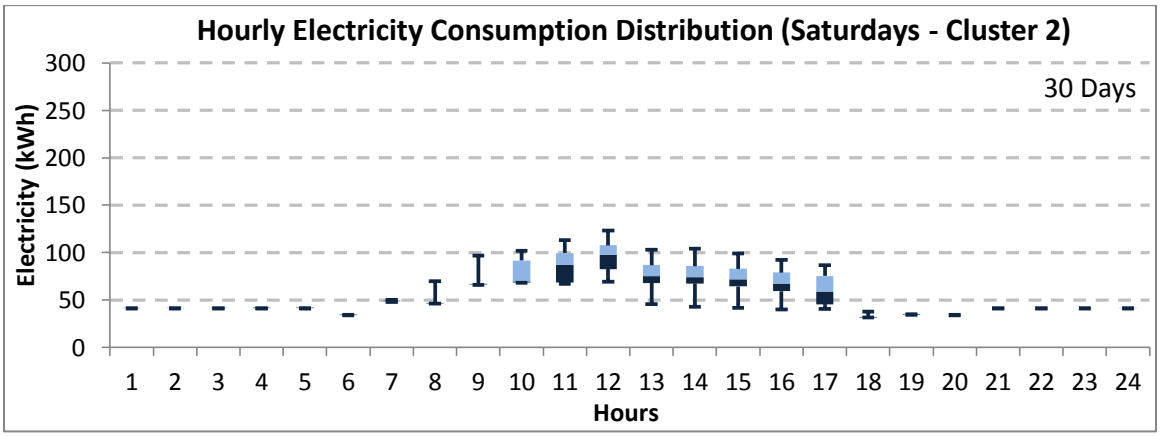
SOB – Clustering Statistics

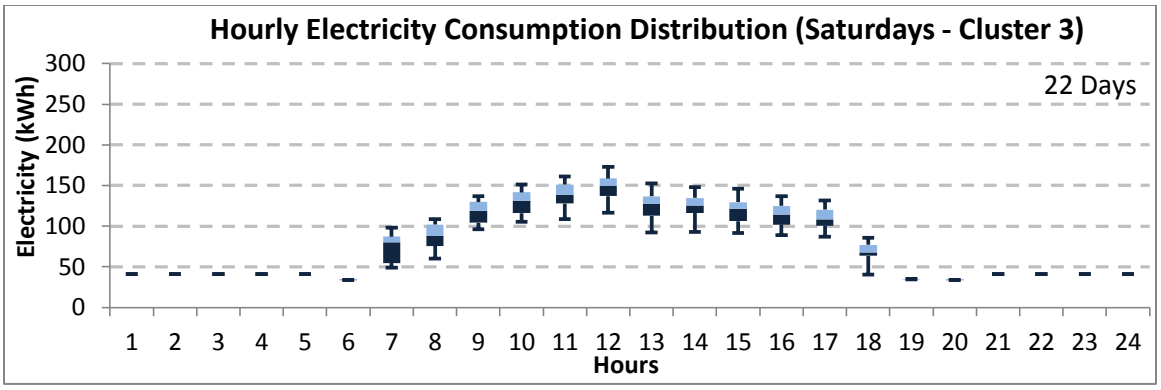
Day Type	Standard Deviation	Mean	CV (%)
Weekdays (Cluster 1)	510.15	2844.03	17.94%
Saturdays (Cluster 2)	126.65	1276.52	9.92%
Saturdays (Cluster 3)	167.82	1828.67	9.18%
Sundays / Holidays (Cluster 0)	1.16	808.15	0.14%

Once we build the daily energy prediction model, we will compare the resulting CV values which can help us evaluate the robustness of the clustering thus achieved.

Next, we plot the hourly energy consumption distribution of the clusters thus formed shown in Fig. 5.8. The resulting NOISE profiles are also plotted to show the distinction between them and the regular days within that cluster. It is important to note that there are no NOISE points for Saturdays (Clusters 2 and 3) and Sundays. The NOISE points identified in the weekday's clusters are Mondays during the summer months. These are a result of excess energy consumption on Monday mornings when the systems turn ON after being shut down over the weekends to make the building fit for occupancy.







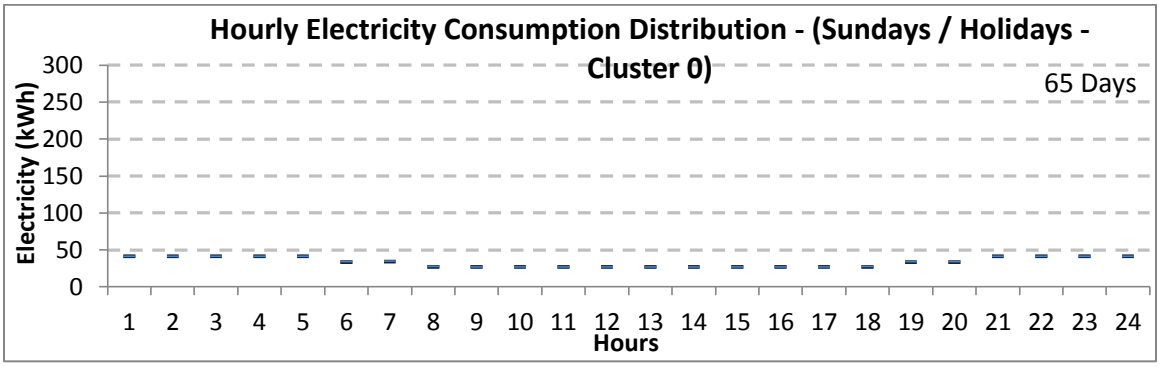


Figure 5.8(a-d). SOB - Hourly Energy Consumption Distribution of various clusters

Actual Office Building (AOB): We plot the daily energy consumption distribution of the days within different clusters:

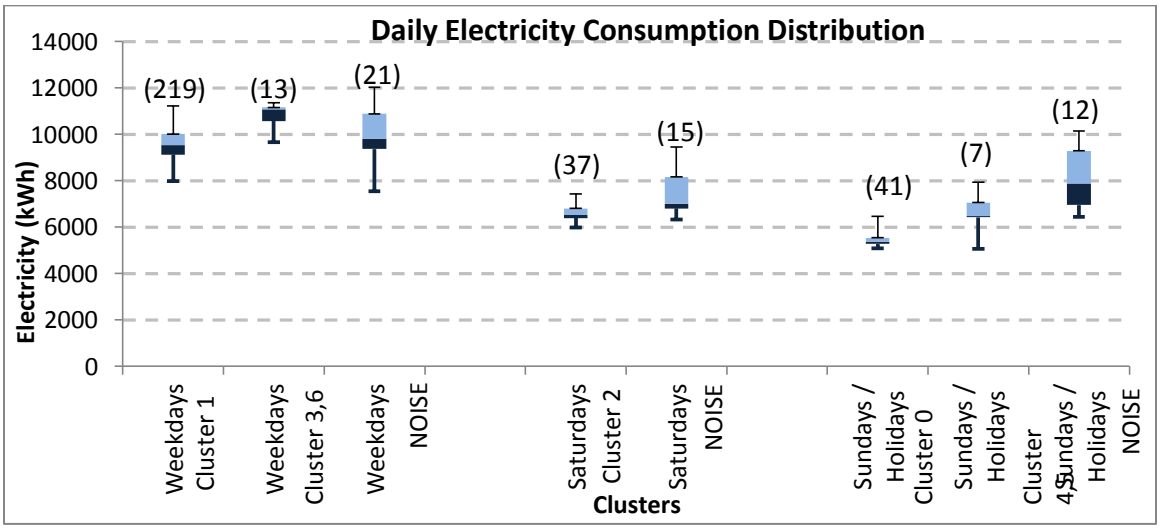


Figure 5.9. AOB - Daily Energy Consumption Distribution of various Clusters

Figure 5.9 represents the actual daily energy consumption of the days within the cluster and the wide distributions during the day types are a result of the temperature effects. As such, the current CV values of these clusters are assembled in Table 5.7:

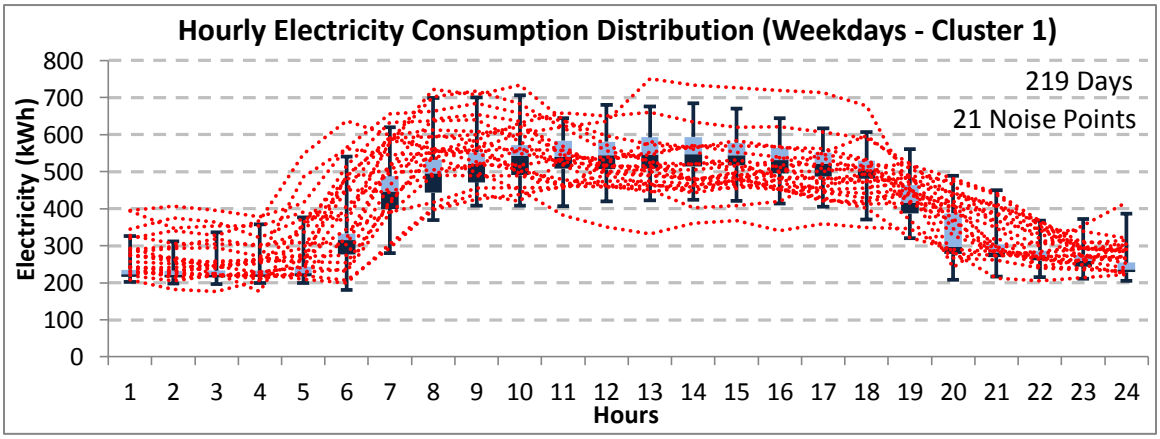
Table 5.7

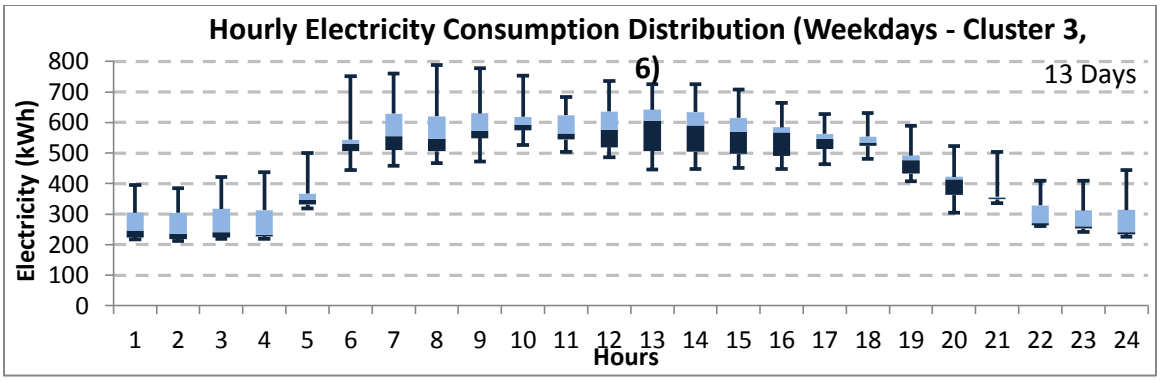
AOB – Clustering Statistics

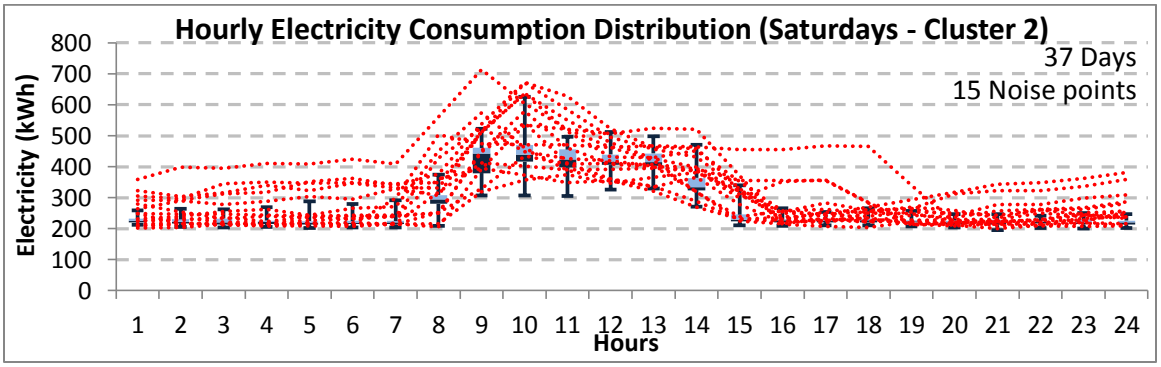
Day Type	Standard Deviation	Mean	CV (%)
Weekdays (Cluster 1)	631.64	9579.70	6.59 %
Weekdays (Cluster 3,6)	540.87	10820	5.00 %
Saturdays (Cluster 2)	366.83	6596.80	5.56 %
Sundays / Holidays (Cluster 0)	259.74	5434.20	4.78 %
Sundays / Holidays (Cluster 4,5)	889.15	6634.20	13.40 %

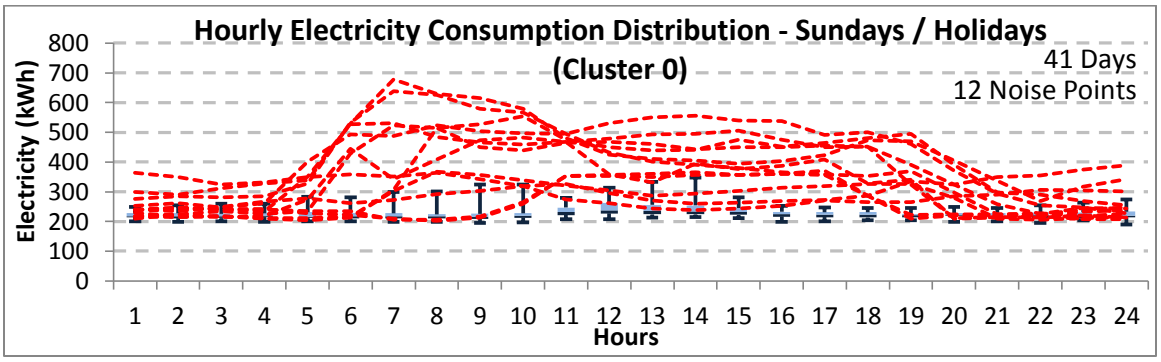
Once we build the daily energy prediction model, we will compare the resulting CV's, which can help us evaluate the robustness of the clustering thus achieved.

Next, we plot the hourly energy consumption distribution of the clusters thus formed shown in Fig. 5.8. The resulting NOISE profiles are also plotted to show the distinction between them and the regular days within that cluster.









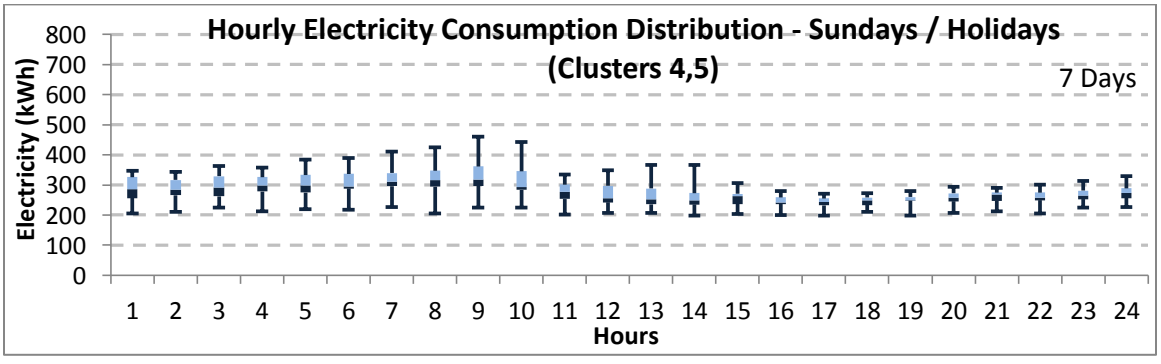


Figure 5.10(a-e). AOB-Hourly Energy Consumption Distribution of various clusters

Finally, comparing the clustering between the two office buildings reinforces our earlier conclusions stated towards the end of Section 5.2 that:

(i) Actual buildings need far more clusters owing to the wide variations in the energy consumption profiles.

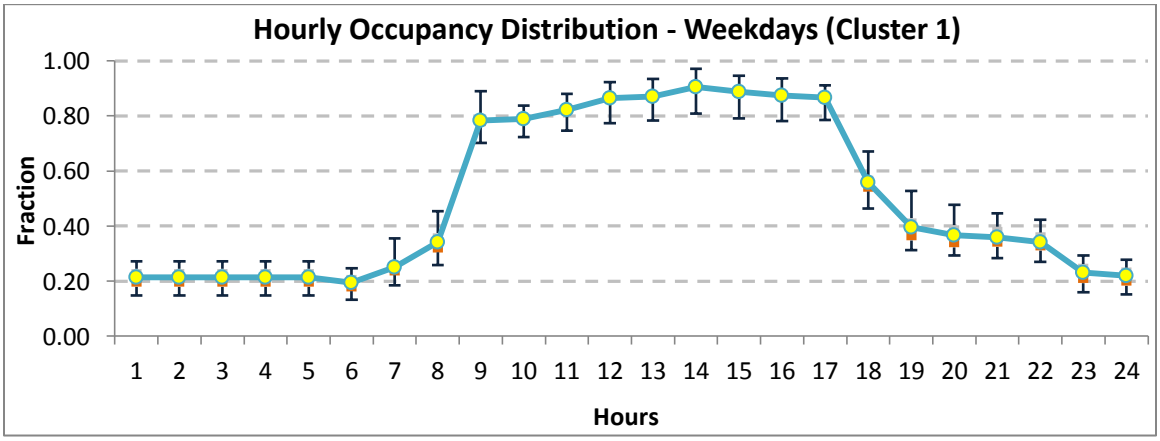
(ii) Actual buildings have more identified NOISE points which are a direct result of the deviations that occur in the day-to-day functioning of the real workd buildings.

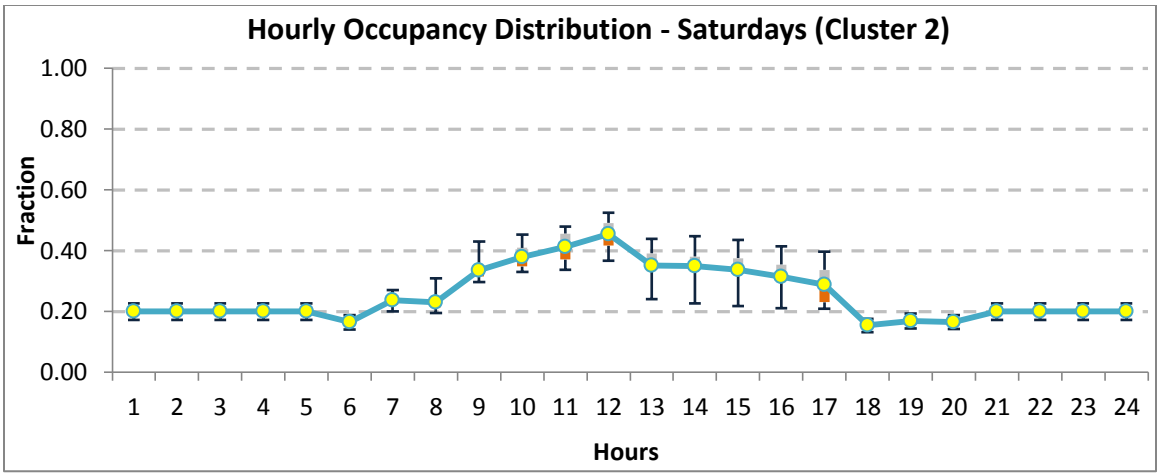
Hence, we would conclude that the clustering algorithm is quite robust in identifying and presenting this information to the analyst.

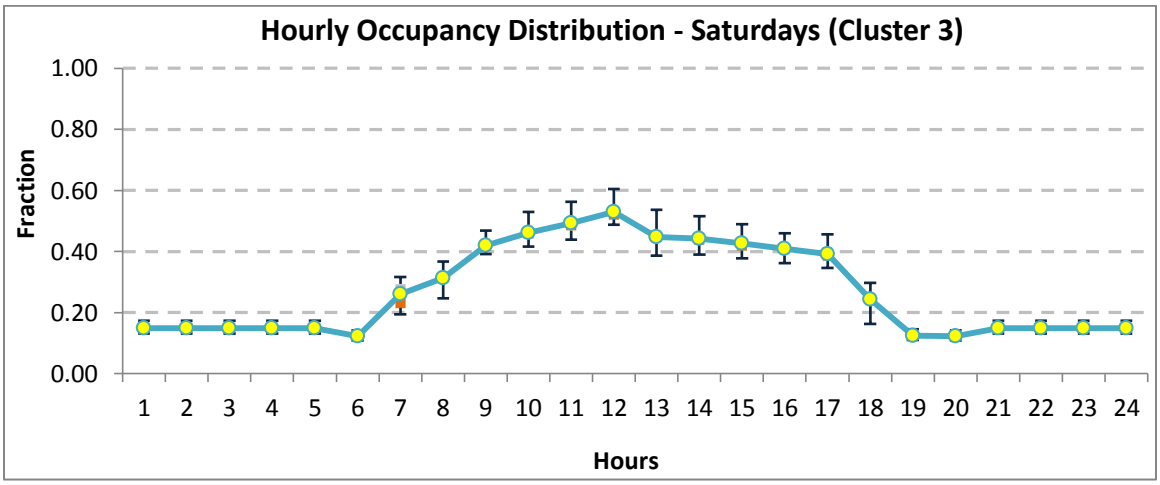
5.5 Occupancy

As described in Section 4.2.5, the occupancy fractions for each day of the year are generated. Finally, we plot the hourly occupancy distribution for all days of the different clusters and choose the median as the occupancy fraction for that particular hour for that specific cluster.

Synthetic Office Building (SOB):







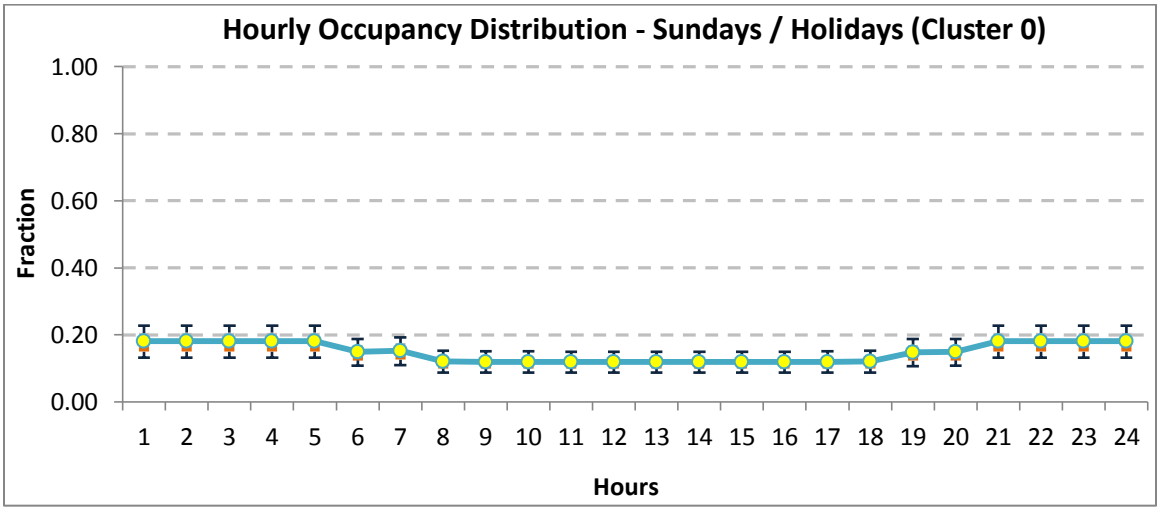


Figure 5.11(a-d). SOB – Occupancy Generation

The occupancy fractions thus derived for all the identified clusters are assembled in Table 5.8:

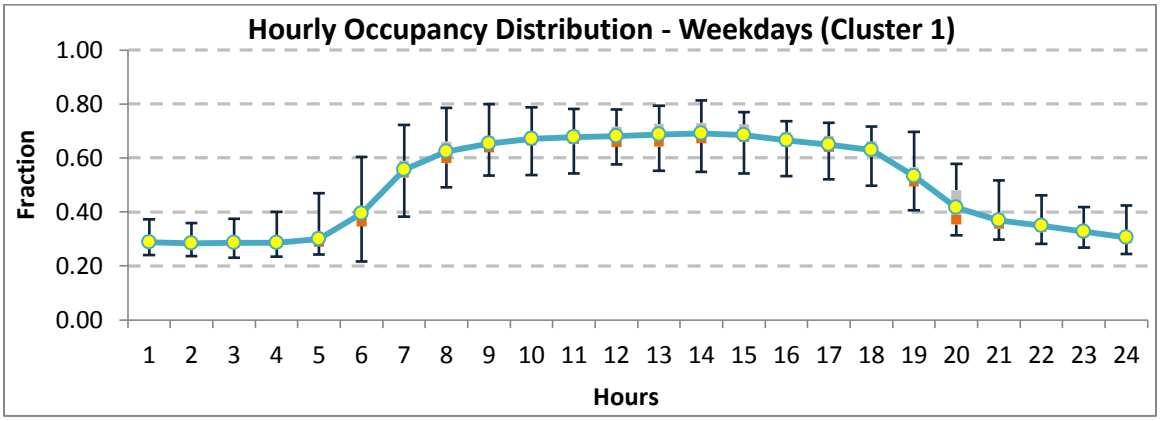
Table 5.8

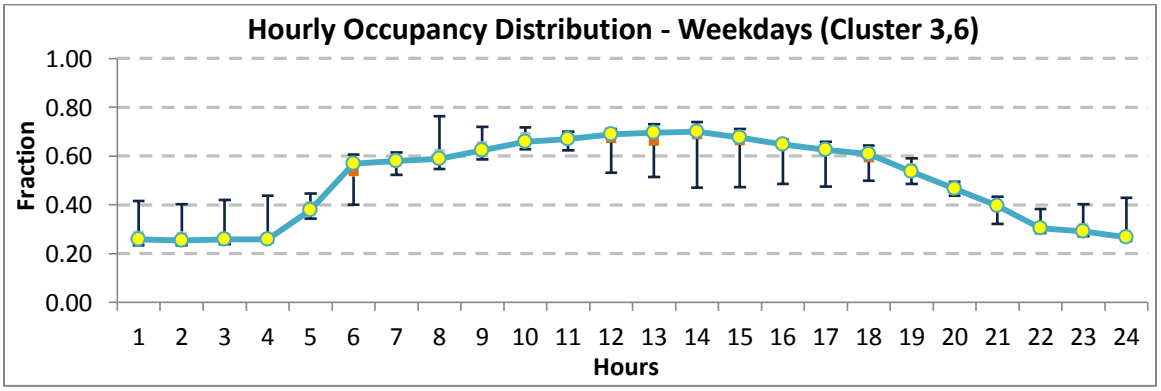
SOB – Occupancy Fractions

Hour	**Weekdays (Cluster 1)**	**Saturdays (Cluster 2)**	**Saturdays (Cluster 3)**	**Sundays (Cluster 0)**
1	0.22	0.20	0.15	0.18
2	0.22	0.20	0.15	0.18
3	0.22	0.20	0.15	0.18
4	0.22	0.20	0.15	0.18
5	0.22	0.20	0.15	0.18
6	0.20	0.17	0.12	0.15
7	0.25	0.24	0.28	0.15
8	0.34	0.23	0.31	0.12
9	0.78	0.33	0.42	0.12
10	0.79	0.37	0.45	0.12
11	0.82	0.42	0.49	0.12
12	0.87	0.47	0.53	0.12
13	0.87	0.36	0.44	0.12
14	0.91	0.36	0.44	0.12
15	0.89	0.35	0.43	0.12
16	0.88	0.32	0.41	0.12
17	0.87	0.29	0.39	0.12

18	0.56	0.16	0.25	0.12
19	0.4	0.17	0.12	0.15
20	0.37	0.17	0.12	0.15
21	0.35	0.2	0.15	0.18
22	0.34	0.2	0.15	0.18
23	0.24	0.2	0.15	0.18
24	0.22	0.2	0.15	0.18

Actual Office Building (AOB): We repeat the same procedure for the actual building as shown in Figures 5.12 (a-e). We note that the variability around the diurnal profiles are much greater than the synthetic building:





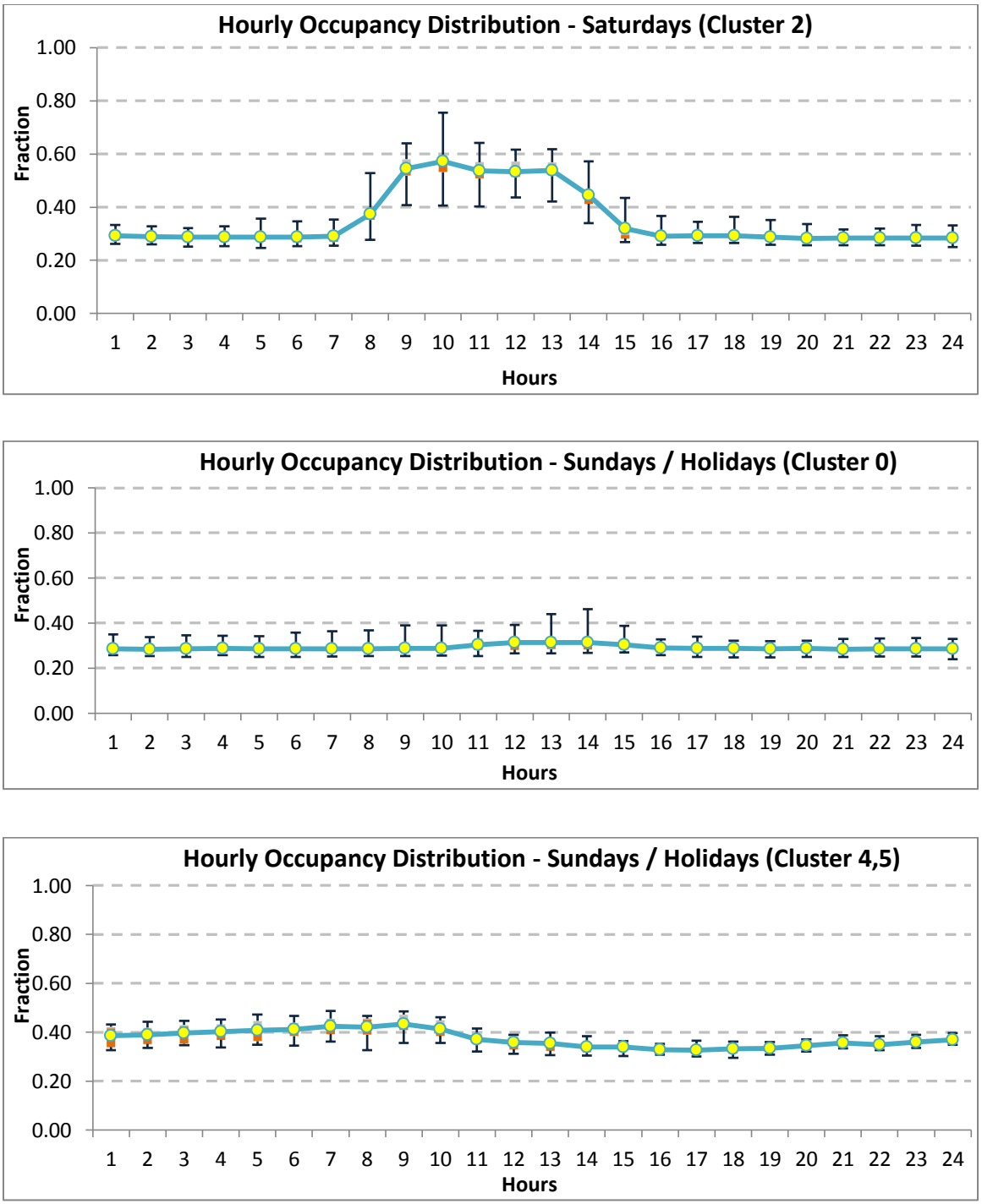


Figure 5.12(a-e). AOB – Occupancy generation

The occupancy fractions thus derived for all the identified clusters are as given in Table 5.9:

Table 5.9

AOB – Occupancy Fractions

Hour	Weekdays (Cluster 1)	Weekdays (Cluster 3,6)	Saturdays (Cluster 2)	Sundays (Cluster 0)	Sundays (Cluster 4,5)
1	0.28	0.27	0.29	0.28	0.42
2	0.28	0.26	0.29	0.28	0.40
3	0.28	0.27	0.28	0.28	0.42
4	0.29	0.26	0.29	0.28	0.42
5	0.29	0.38	0.28	0.28	0.41
6	0.39	0.57	0.28	0.28	0.42
7	0.56	0.58	0.28	0.28	0.43
8	0.61	0.59	0.37	0.28	0.45
9	0.65	0.63	0.55	0.28	0.43
10	0.68	0.67	0.56	0.28	0.41
11	0.68	0.67	0.55	0.30	0.36
12	0.69	0.68	0.54	0.31	0.37
13	0.70	0.69	0.54	0.30	0.36
14	0.71	0.69	0.43	0.30	0.33
15	0.70	0.66	0.30	0.29	0.34
16	0.68	0.64	0.28	0.29	0.33
17	0.66	0.62	0.29	0.29	0.32
18	0.64	0.59	0.29	0.29	0.34
19	0.52	0.54	0.29	0.29	0.33
20	0.39	0.46	0.28	0.28	0.35

21	0.36	0.39	0.28	0.28	0.36
22	0.34	0.31	0.28	0.29	0.35
23	0.32	0.30	0.28	0.29	0.36
24	0.30	0.27	0.28	0.28	0.36

This concludes the results of the Data Pre-processing steps described in the flowchart in Figure 4.4. Additionally, climate regressor data is pre-processed as described in the flowchart in Figure 4.7 and then, we move to the next step of energy prediction model identification. In the next chapter, we present the results of the Inverse Statistical modeling.

CHAPTER 6 : RESULTS – INVERSE STATISTICAL MODELING

The results obtained in Chapter 5 are used to identify the energy prediction models for the two office buildings. Such models can be used for enhancement of energy performance and operations analysis of buildings with the availability of energy interval data. Various areas of applications include building M&V, CCx, condition monitoring and FDD. In this Chapter, we assemble and discuss the results for the base-lining portion of the flowchart described in Figure 4.1.

6.1 Change Point Identification

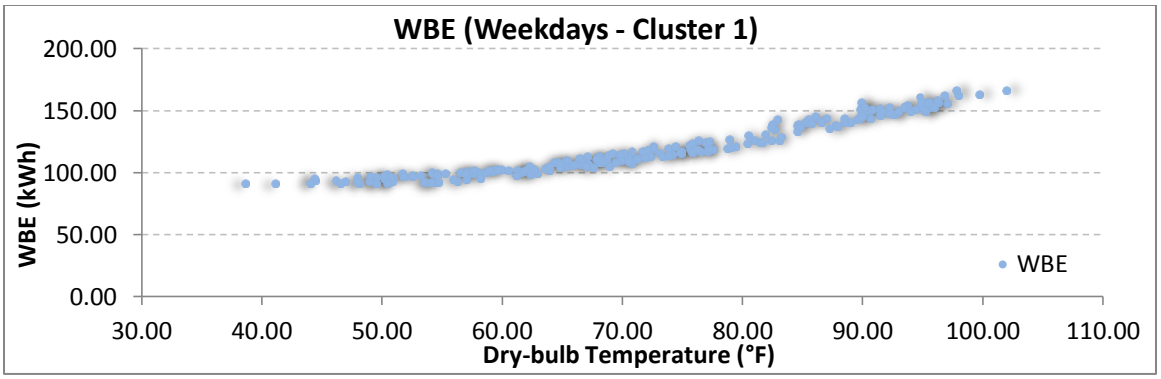
We create scatter plots for the daily energy consumption vs. the outdoor dry-bulb temperature for the two buildings to visually look for any change-points that should be accounted for in the daily energy prediction models.

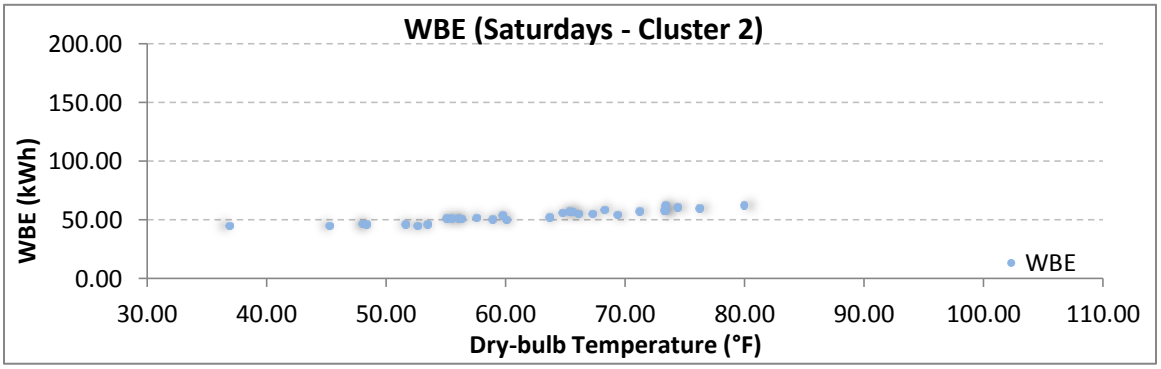
Synthetic Office Building (SOB): From Figures 6.1 (a-d), energy use during weekdays exhibit a data scatter that would require multiple change-points (MCP) to be identified and included in model identification. This is akin to the 5P regression model described earlier in Section 3.3.2. Apart from these, most of the other clusters identified exhibit linear trend. The change points identified during regression for the daily models of different clusters in this building are shown in Table 6.1:

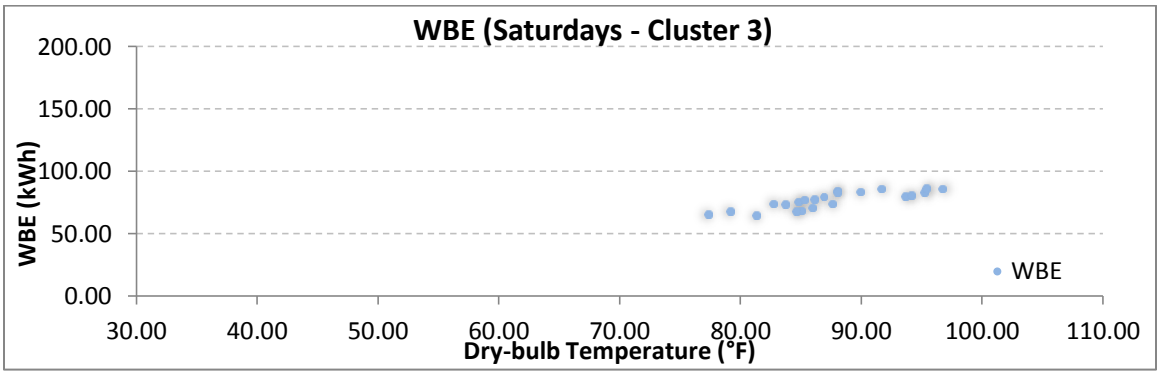
Table 6.1

SOB Daily Energy Model Change Points

Day Type	Cluster Number	Change – Point (°F)
Weekdays	1	60.8 and 77.6







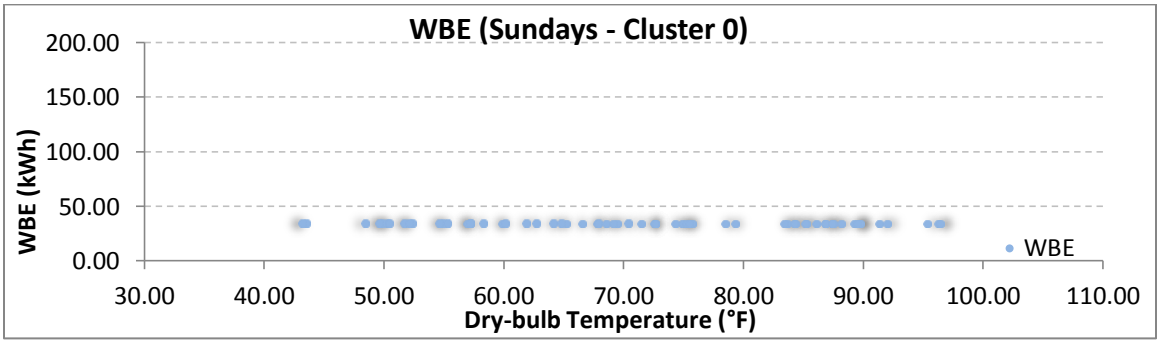


Figure 6.1(a-d). SOB Scatter Plots – Outdoor DBT vs. Daily Energy Consumption

Actual Office Building (AOB): Looking at Figures 6.2(a-c), one of the weekdays cluster (Cluster 1) in this office building exhibits a pattern that is akin to a 5P model form described earlier in Section 3.3.2 and would require multiple change-point modeling. Also, one of the Sunday clusters (Cluster 0), exhibits a pattern similar to a 2P model and requires a single change-point modeling.

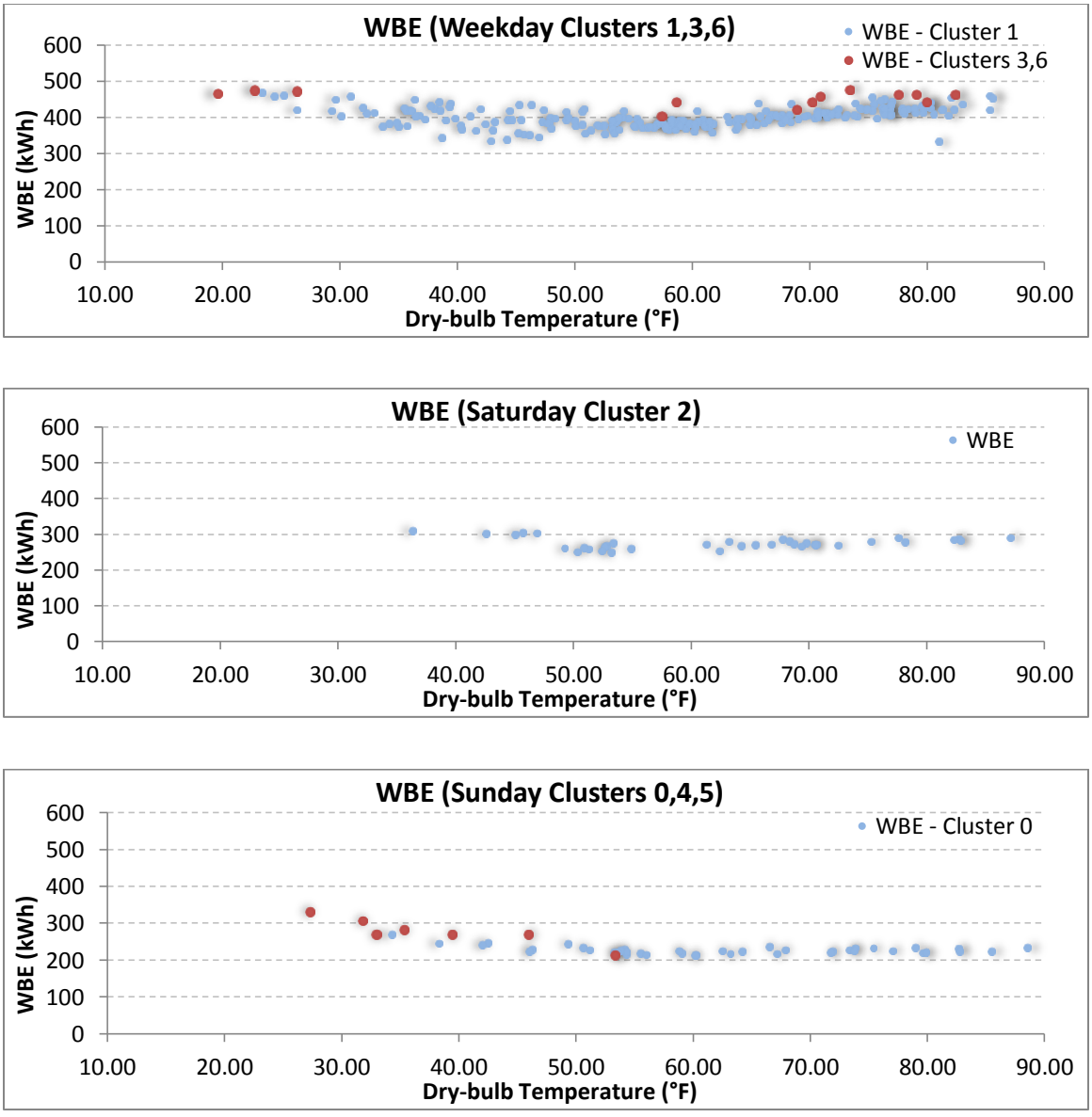


Figure 6.2 (a-c). AOB Scatter Plots – Outdoor DBT vs. Daily Energy Consumption

The change points identified for the daily models for the different clusters in this building are given in Table 6.2:

Table 6.2

AOB Daily Energy Model Change Points

Day Type	Cluster Number	Change – Point (°F)
Weekdays	1	42.9
Weekdays	1	59.5
Sunday	0	56

6.2 Daily Energy Consumption Model Identification

As described earlier, the entire dataset was divided into two parts – training and testing datasets. The Daily Average energy consumption model form identified from the training set is given by Equation 6.1:

$$\check{\mathrm{E}}_i = a + \sum_{k=1}^{K}\{b_k.\check{\mathrm{T}}_{db,i} + c_k.\left(\check{\mathrm{T}}_{db,i} - X_{1,k}\right)^{+} + d_k.\left(\check{\mathrm{T}}_{db,i} - X_{2,k}\right)^{+} + e_k.\left(\hat{\mathrm{W}}_i - 0.008\right)^{+} + f_k.\bar{\mathrm{Q}}_{sol} + g_k.\mathrm{D}_k\} \quad \text{Eq. 6.1}$$

Where,

$i = 1$ to 365, index for day of the year; $j = 1$ to 24, index for hour of the day,

$m = 1$ to 12, index for month of the year; and $k = 1$ to K, index for day type,

$\check{\mathrm{E}}_i$ = Daily average energy consumption,

$\check{\mathrm{T}}_{db,i}$ = Daily average dry-bulb temperature,

$X_{1,k}$ = Change-point 1 for any given day type k,

$X_{2,k}$ = Change-point 2 for any given day type k,

$\hat{W}_i$ = Daily average humidity ratio potential,

$\bar{Q}_{sol}$ = Daily average total horizontal radiation,

D_k = Any given day type *k.*

The model coefficients for both the office buildings are shown in Table 6.3:

Table 6.3

Daily Average Energy Model Coefficients – SOB and AOB

Daily Average Energy Prediction Model Coefficients							
Synthetic Office Building (SOB)				**Actual Office Building (AOB)**			
Code	**Value**	**Code**	**Value**	**Code**	**Value**	**Code**	**Value**
a	56.4825	e_1	1233.1454	a	541.6698	e_1	0
b_1	0.7354	e_2	0	b_1	- 3.5970	e_2	0
b_2	- 0.4794	e_3	0	b_2	0	e_3	0
b_3	- 0.5703	e_4	-623.9243	b_3	- 1.0786	e_4	0
b_4	- 1.2858			b_4	0	e_5	0
		f_1	0	b_5	- 3.2295	f_1	0
c_1	0.4359	f_2	0	c_1	2.9248	f_2	0.2107
c_2	0	f_3	- 0.1446	c_2	0	f_3	0
c_3	0	f_4	0	c_3	0	f_4	0
c_4	0			c_4	- 2.0541	f_5	0

Code	Value	Code	Value	Code	Value	Code	Value
		g_1	0	c_5	0	g_1	0
		g_2	-20.9314			g_2	0
d_1	0.5669	g_3	0	d_1	3.2819	g_3	- 56.3569
d_2	0	g_4	7.9057	d_2	0	g_4	-154.1647
d_3	0	$k=1$	Weekdays (C 1)	d_3	0	$k=1$	Weekdays (C 1)
d_4	0	$k=2$	Saturdays (C 2)	d_4	0	$k=2$	Weekdays (C 3,6)
		$k=3$	Saturdays (C 3)	d_5	0	$k=3$	Saturdays (C2)
$x_{1,1}$	60.8	$k=4$	Sundays (C 0)	$x_{1,1}$	42.9	$k=4$	Sundays (C 0)
$x_{2,1}$	77.6			$x_{2,1}$	59.5	$k=5$	Sundays (C 4,5)
				$x_{1,4}$	56		

The above model predicts daily average energy consumption for each of the two buildings. The actual daily energy consumption of the buildings is then given by:

$$\mathrm{E}_i = 24 * \check{\mathrm{E}}_i \qquad \text{Eq. 6.2}$$

The model statistics are assembled in the Table 6.4:

Table 6.4

SOB and AOB – Daily Model Statistics

DAILY MODEL STATISTICS					
Training dataset (60% points)			**Testing Dataset (40% points)**		
	SOB	**AOB**		**SOB**	**AOB**
Model – R^2	0.994	0.958			
RMSE (kWh)	69.50	374.32	**RMSE (kWh)**	73.7	480.0
CV – RMSE	3.16%	4.35%	**CV – RMSE**	3.08%	5.44%
Durbin - Watson	1.7	1.91			

Looking at the model statistics assembled in Table 6.4 along with the model predictions and the residual plots below, we conclude that the models have a high model R^2 and a fairly low RMSE and CV-RMSE. Also, the model CV's calculated now are much lower than the CV's for energy consumptions of the different clusters identified in Tables 5.6 and 5.7 which strengthens the fact that those wide distributions were a result of the temperature variations and that the clustering achieved is robust. Finally, the residual plots indicate that the residuals are fairly evenly distributed and there is no observable structure to the residuals.

The model predictions are plotted against the actual whole building electric (WBE) values as show in Figures 6.3 and 6.5 and the corresponding residual plots are shown in Figures 6.4 and 6.6:

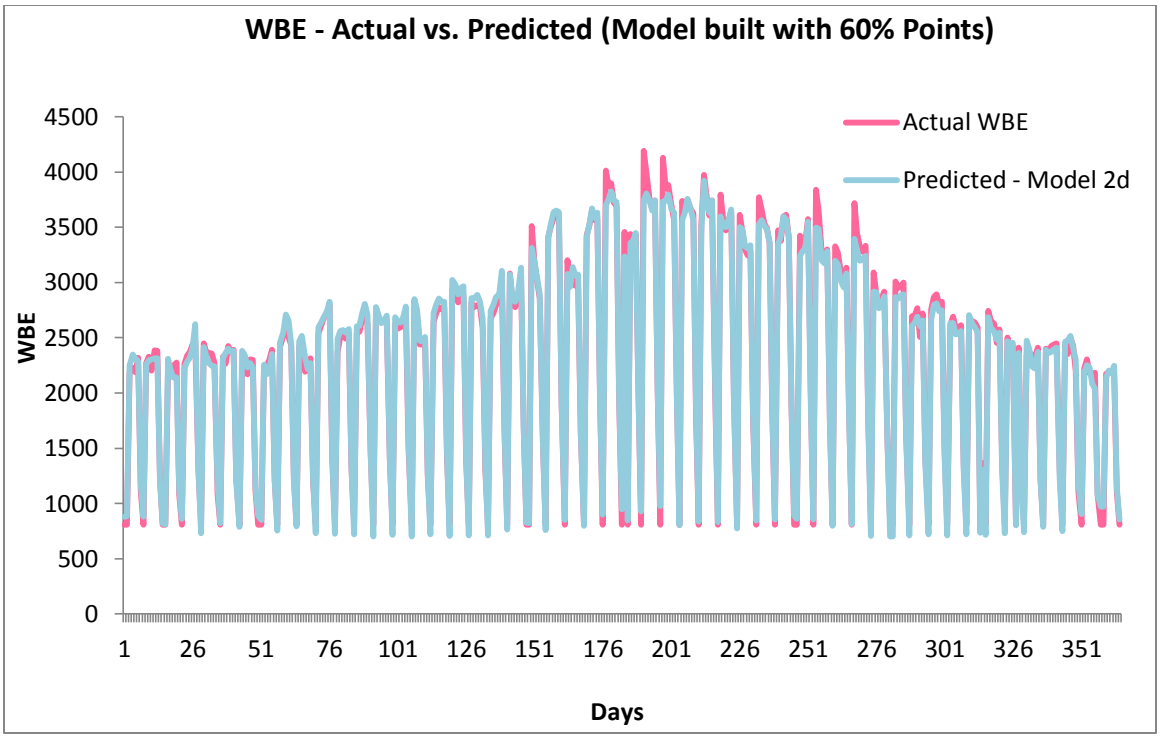


Figure 6.3. SOB – Daily model predicted vs. Actual WBE

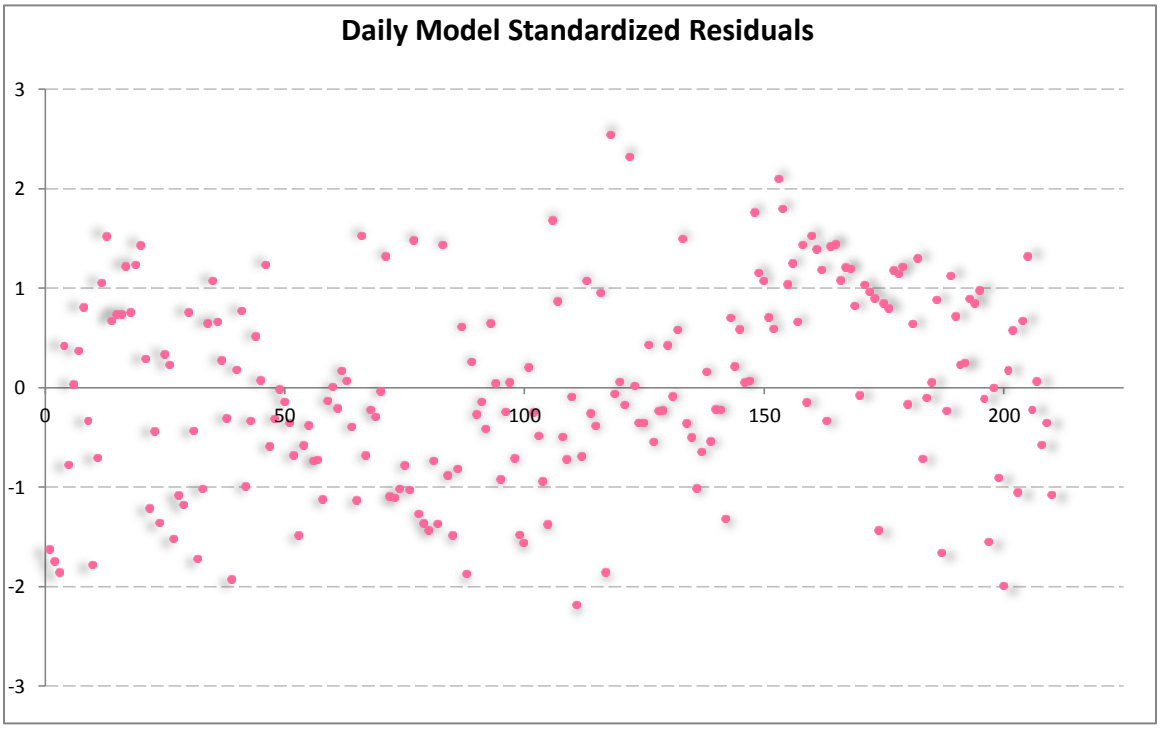


Figure 6.4. SOB – Daily Model Standardized Residuals

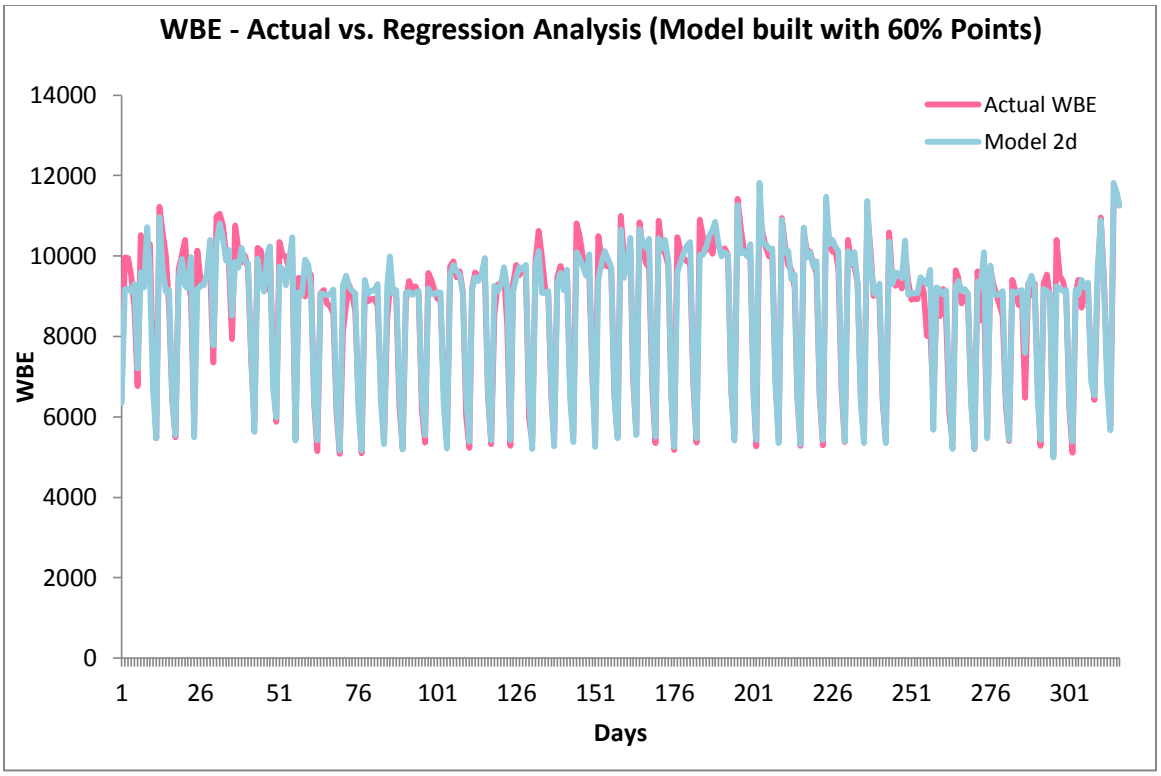


Figure 6.5. AOB – Daily model predicted vs. Actual WBE

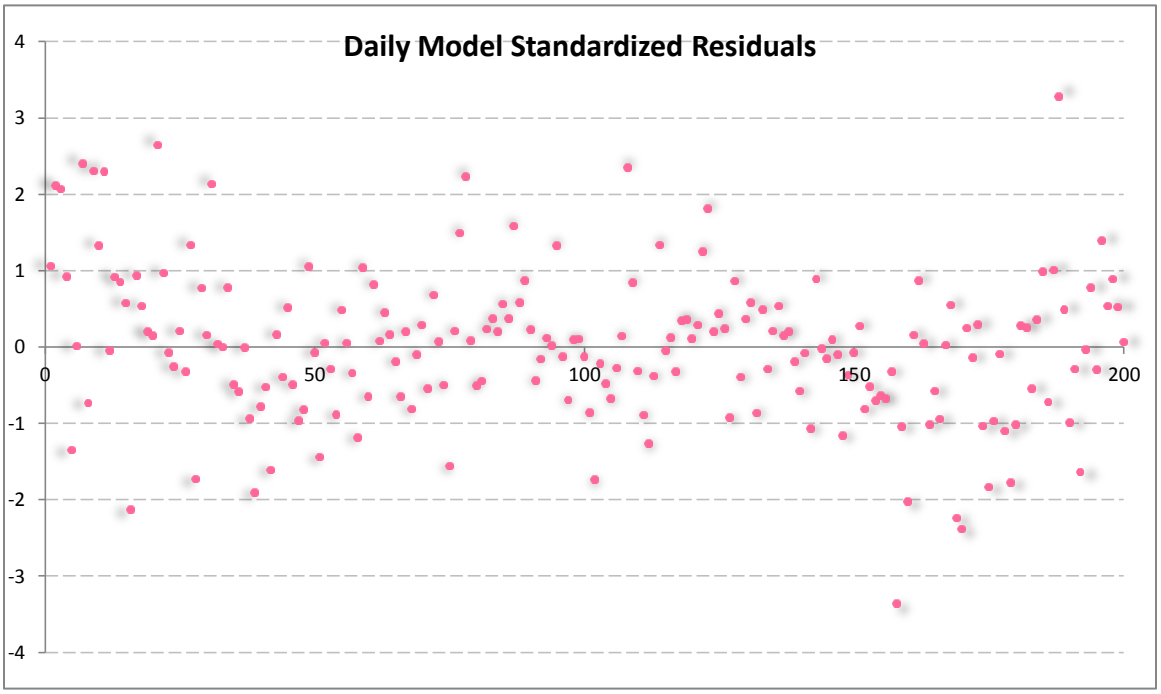


Figure 6.6. AOB – Daily Model Standardized Residuals

6.3 Hourly Energy Consumption Model Identification

At this point, we have already identified the average daily energy prediction models. As described in Section 4.2.6, we develop the hourly energy prediction models as a deviation from the average energy consumption based on the deviation of the hourly climate regressors. Additionally, we include the hourly occupancy fractions which have already been calculated for each of the clusters and described in Tables 5.8 and 5.9. These are important regressors in the hourly model identification process. Once again, the entire dataset was divided into training and testing dataset and modeling was done. The Hourly energy consumption models thus developed from the training set are as follows:

$$\check{\mathrm{E}}_{i,j} = \check{\mathrm{E}}_i + [\alpha + \sum_{k=1}^{K}\{\beta_k.\Delta \mathrm{T}_{db,i,j} + \gamma_k.\Delta \mathrm{W}_{i,j} + \delta_{k.}\Delta \mathrm{Q}_{sol,i,j} + \varepsilon_{k.}\mathrm{D}_k\} + \tau.O_{k,j}] \quad \text{Eq. 6.3}$$

Where,

$i = 1$ to 365, index for day of the year; $j = 1$ to 24, index for hour of the day,

$m = 1$ to 12, index for month of the year; and $k = 1$ to K, index for day type,

$\check{\mathrm{E}}_i$ = Daily average energy consumption,

$\check{\mathrm{E}}_{i,j}$ = Hourly energy consumption prediction,

$\Delta \mathrm{T}_{db,i,j}$ = Hourly DBT – Daily average DBT,

$\Delta \mathrm{W}_{i,j}$ = Hourly humidity – Daily average humidity,

$\Delta \mathrm{Q}_{sol,i,j}$ = Hourly Radiation – Daily Average Radiation,

$O_{k,j}$ = Hourly occupancy fraction for any given day type *k* and any given hour *j*.

D_k = Any given day type *k*.

The model coefficients for both the office buildings are given in Table 6.5:

Table 6.5

Hourly Energy Model Coefficients – SOB and AOB

Hourly Energy Prediction Model Coefficients							
Synthetic Office Building (SOB)				**Actual Office Building (AOB)**			
Code	**Value**	**Code**	**Value**	**Code**	**Value**	**Code**	**Value**
α	- 88.520288	δ_4	- 0.154717	α	-371.549214	δ_4	0
β_1	0.722481			β_1	0.620489	δ_5	0
β_2	- 0.732464	ε_1	0	β_2	0	ε_1	0
β_3	- 1.016259	ε_2	42.485407	β_3	- 0.628518	ε_2	-11.16882
β_4	- 0.768401	ε_3	88.817794	β_4	- 1.078778	ε_3	111.9695
		ε_4	62.78503	β_5	0	ε_4	158.5118
Y_1	0			Y_1	-800.017793	ε_5	76.89253
Y_2	0	τ	176.20271	Y_2	0	τ	741.7109
Y_3	0	***k=1***	Weekdays (C 1)	Y_3	0	***k=1***	Weekdays (C 1)
Y_4	0	***k=2***	Saturdays (C 2)	Y_4	0	***k=2***	Weekdays (C 3,6)
		k=3	Saturdays (C 3)	Y_5	0	***k=3***	Saturdays (C2)
δ_1	0.145539	***k=4***	Sundays (C 0)	δ_1	0.014917	***k=4***	Sundays (C 0)
δ_2	- 0.110585			δ_2	0.027938	***k=5***	Sundays (C 4,5)

$\boldsymbol{\delta_3}$	0.203358		$\boldsymbol{\delta_3}$	0		

Finally, the hourly model assumes the form:

$$\check{E}_{i,j} = \{Daily\ Average\ Energy\ Prediction\}_i + \{Correction\ for\ Hour\ of\ the\ day\}_j$$

Eq.6.4

The model statistics are assembled in the Table 6.6:

Table 6.6

SOB and AOB – Hourly Model Statistics

HOURLY MODEL STATISTICS					
Training dataset (60% points)			**Testing Dataset (40% points)**		
	SOB	**AOB**		**SOB**	**AOB**
Model – R^2	0.948	0.893			
RMSE (kWh)	13.193	40.46	**RMSE (kWh)**	13.5	40.0
CV - RMSE	13.93%	11.19%	**CV - RMSE**	14.26%	10.98%
Durbin - Watson	0.635	0.895			

Synthetic Office Building (SOB): Looking at the model statistics in Table 6.6, we can conclude that the model has a high R^2; the CV values are quite high as compared to the daily values. This is to be expected given the random variations and heat flows which assume relative importance at this finer time scale. We plot one week's actual and

predicted energy consumption during a winter month (January) and a summer month (July) to understand this:

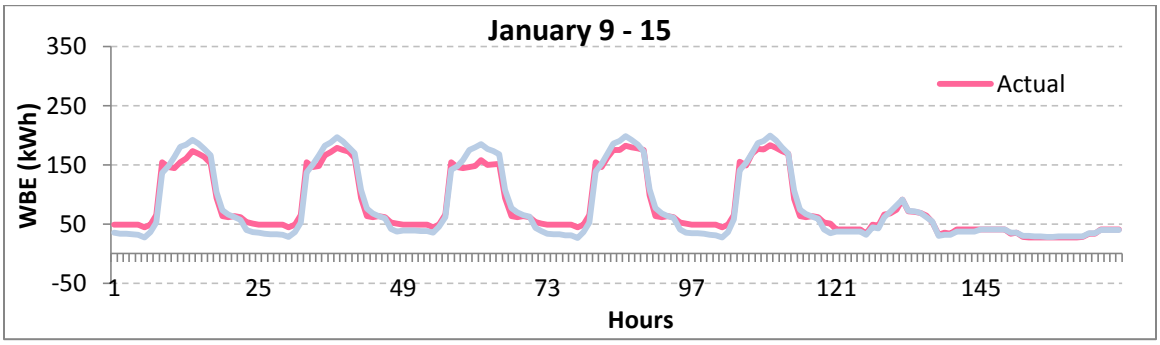


Figure 6.7. SOB – Actual vs. Predicted Energy Consumption (January)

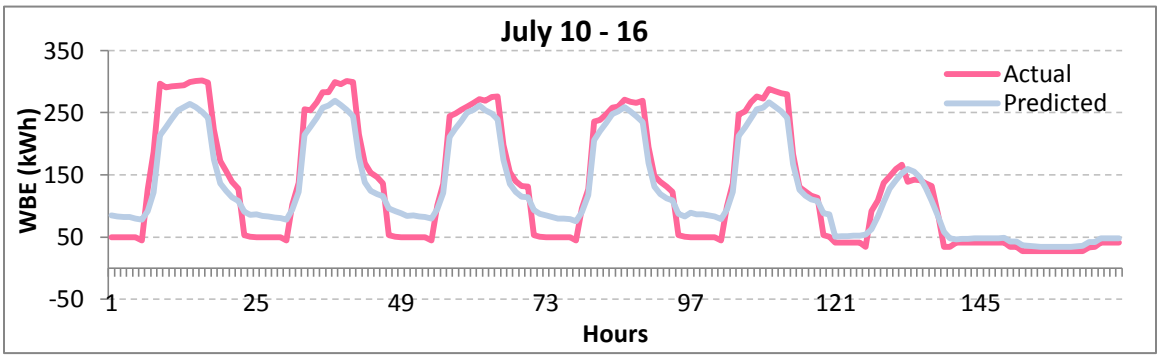


Figure 6.8. SOB – Actual vs. Predicted Energy Consumption (July)

Observing Figures 6.7 and 6.8, we can clearly see that the model is predicting higher energy consumption during the winter months, while it is predicting lower energy consumption during the summer months as compared to the actual energy consumption. This can be attributed to the thermal mass effect of the building envelope. During the winter months, the lower outdoor temperatures cool the building envelope thus reducing the energy required for cooling purposes of certain building zones. In the summer months, the same higher outdoor temperatures heat the building envelope and these result

in extra cooling requirement. Ultimately, these variations between the hourly values lead to a higher model CV, which is unable to capture the thermal mass effect.

Actual Office Building (AOB): The hourly model has a high R^2 value but again, has a higher CV. Observing the actual vs. predicted energy graph in Figure 6.11 for this building, we find that the model is under-predicting during the winter months at the beginning of the year, while being accurate in the winter months towards the end of the year. This can be attributed to unsystematic changes in the building operations that the model is not trained to capture. Also, during the summer months, the model is systematically under-predicting which can be attributed to the thermal lag effect described above.

The model predictions are plotted against the actual whole building electric (WBE) values as shown in Figures 6.9 and 6.11 and the corresponding residual plots are shown in Figures 6.10 and 6.12:

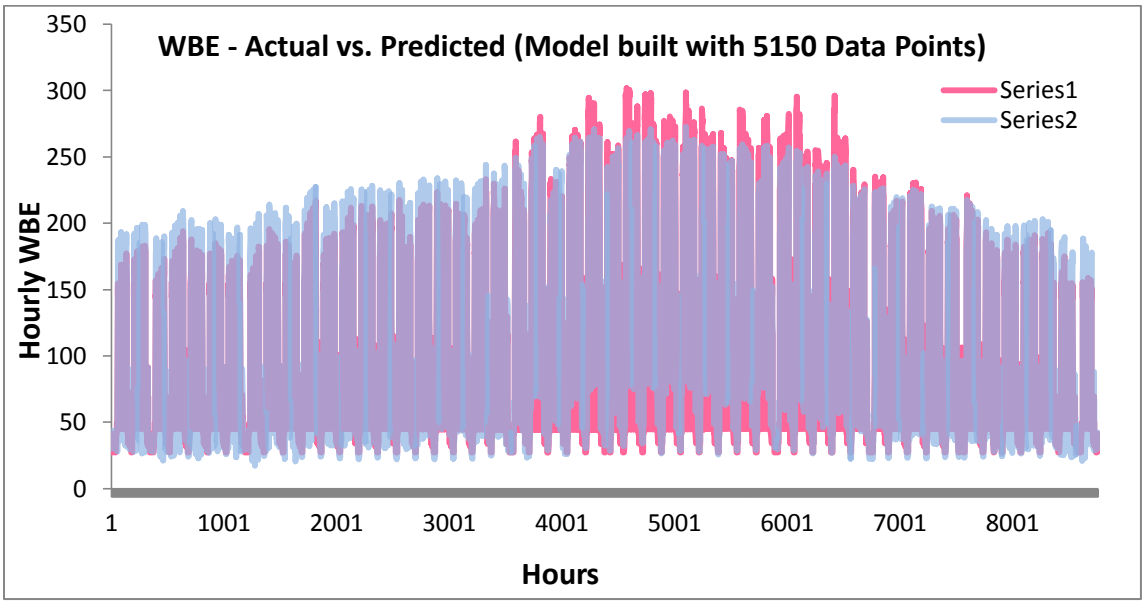


Figure 6.9. SOB – Hourly model predicted vs. Actual WBE

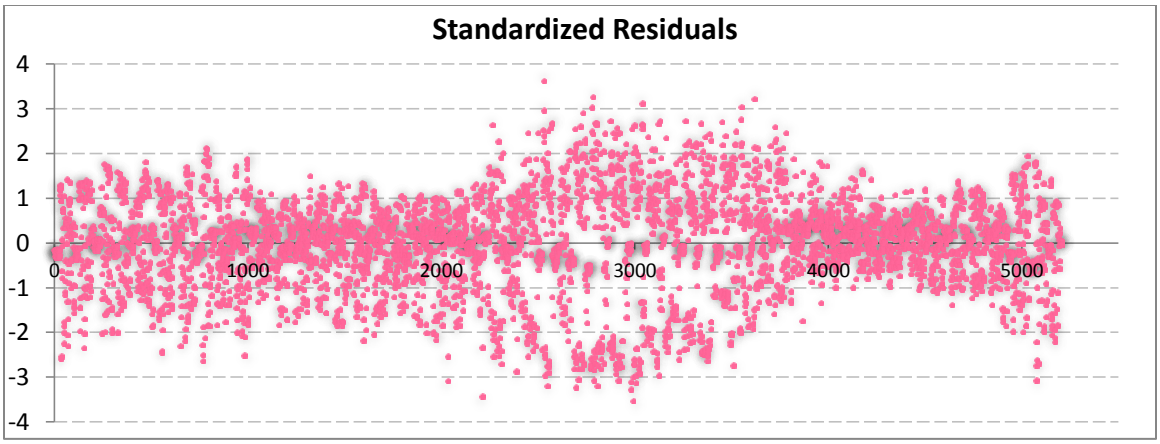


Figure 6.10: SOB – Hourly model standardized residuals

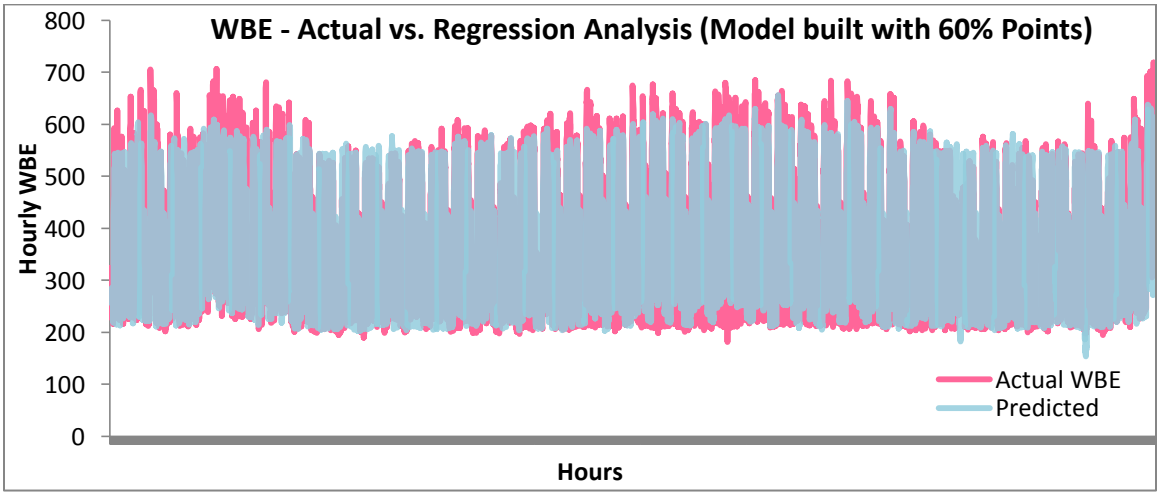


Figure 6.11. AOB – Hourly model predicted vs. Actual WBE

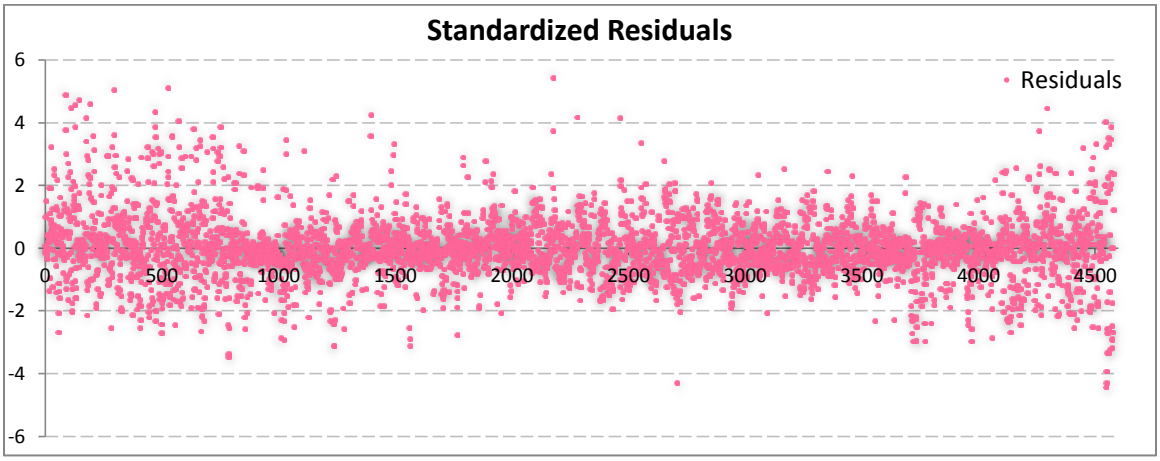


Figure 6.12. AOB – Hourly model standardized residuals

The Durbin – Watson statistic for both the models in Table 6.6 (SOB – 0.635 and AOB – 0.895) indicates that serial autocorrelation is present between the model residuals and the residual plots exhibit patterns indicative of a thermal lag effect. In the next section, we will address this issue.

6.4 Short-term Load Forecasting (STLF) – Modeling the error terms

As discussed in Section 3.7 and described in Section 4.5, we will model the systematic stochastic component of the hourly model residuals, eventually leading to a model with higher prediction accuracies. In essence, we treat the residual patterns as an individual series and try to establish the correlations between consecutive observations on a seasonal and non-seasonal level as discussed in Section 4.5.3. As such, this relationship helps us to predict the next residual observation, which when adjusted in the hourly energy prediction increases the overall accuracy and this model can be used for STLF.

6.4.1 Detecting Non-stationarity and Differencing

Plotting the residuals of the hourly model can help us visually decide on the stationarity of the time series. As described earlier in Section in 4.5.1, if the values seem to fluctuate around a constant mean with a constant variance, it would be reasonable to believe that the series is stationary. If not, then we will transform the series by differencing and re-evaluate this transformed series. Additionally, we will plot the SAC and the SPAC functions to check for stationarity as described in Section 4.5.2

Synthetic Office Building (SOB): We begin by evaluating the actual hourly model residuals of the synthetic office building. The actual time series plots along with the SAC and SPAC plots are given in Figures 6.13 (a-c):

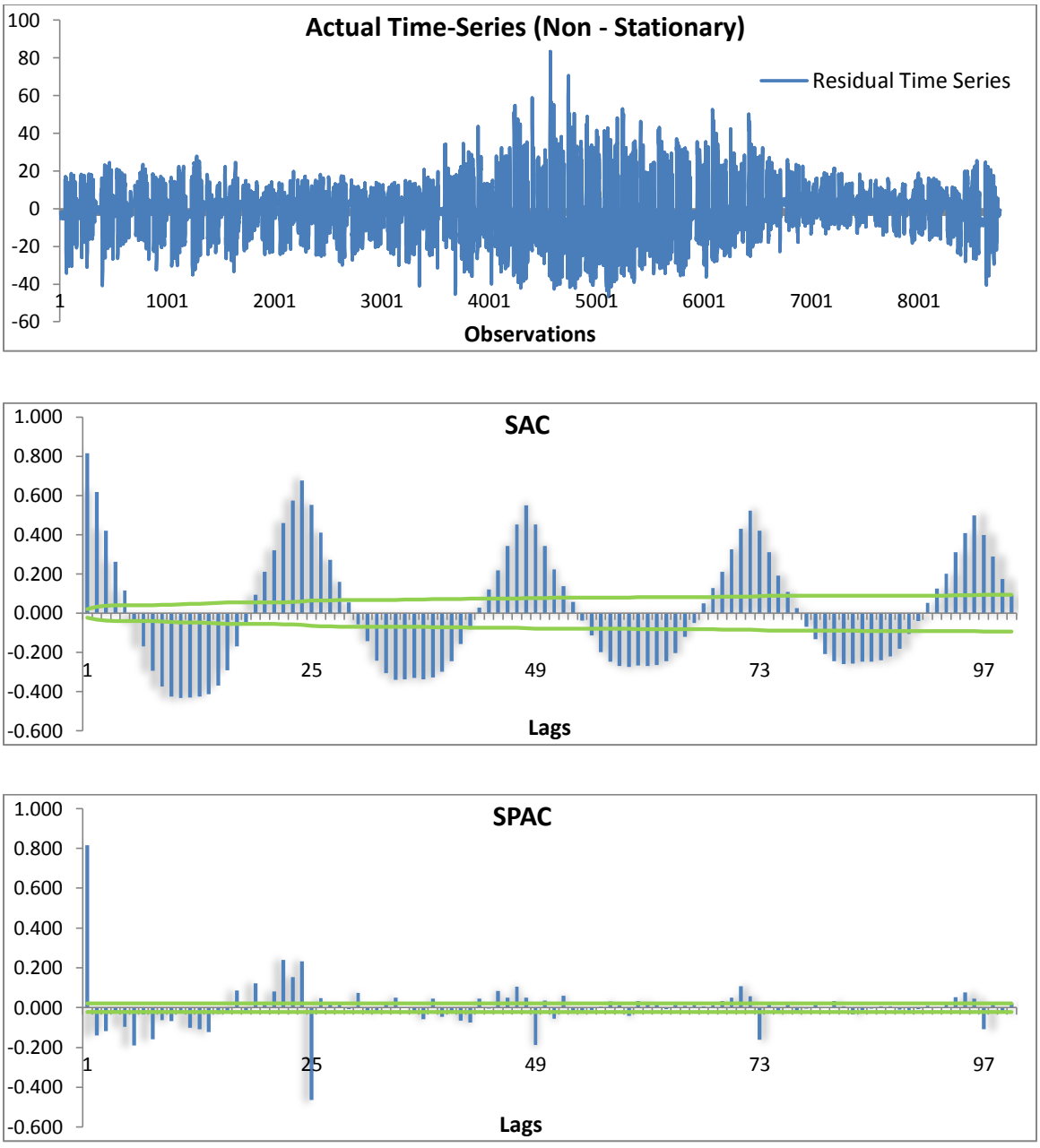
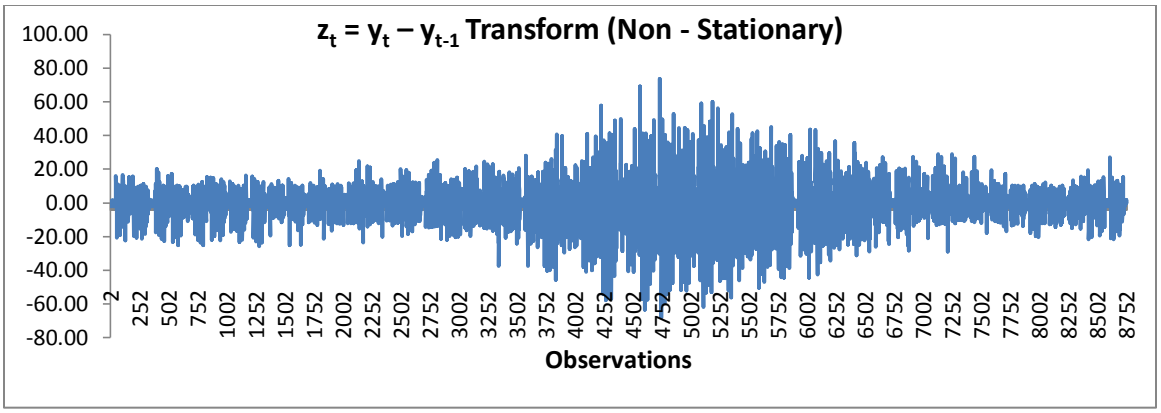
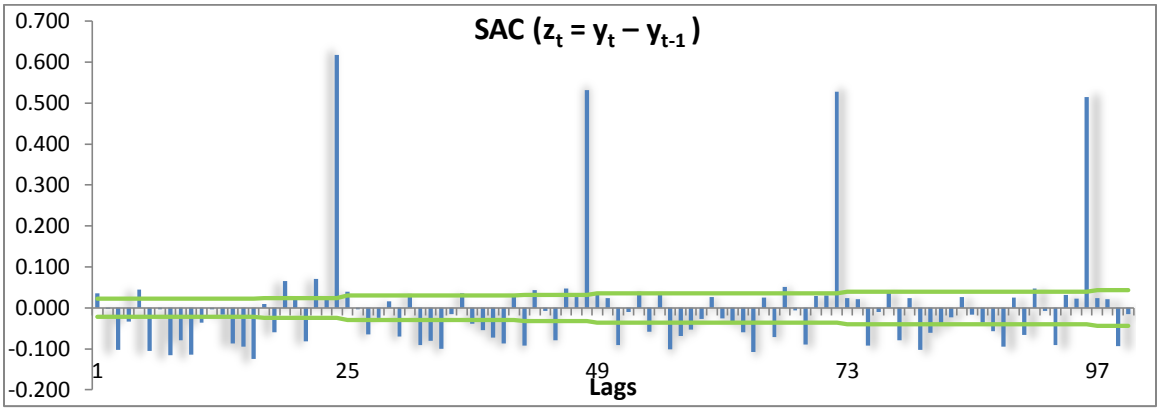


Figure 6.13 (a-c). SOB Actual Time Series – Stationarity Check

In Figure 6.13(a-c), the actual residual time series plots do not exhibit constant variation around the mean and the SAC and SPAC functions do not cut off fairly quickly. Hence, we conclude that this series is non-stationary. We will transform the series as described in Section 4.5.1 and re-plot the functions.





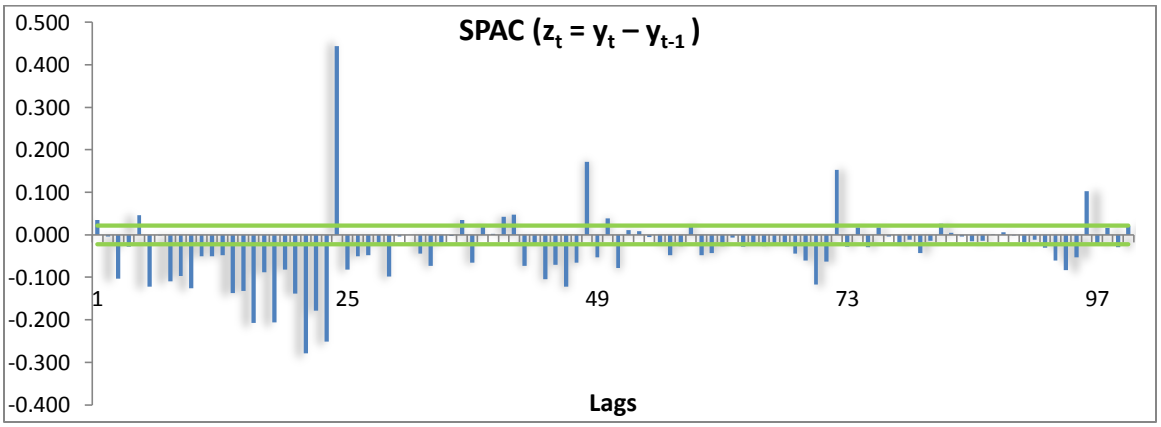


Figure 6.14 (a-c). SOB Transformed Series – Stationarity Check

The 1 hour lag transformed series in Figure 6.14 (a-c) still do not show constant variance and the SAC and SPAC functions don't seem to be cutting off fairly quickly at the seasonal and non-seasonal level. Hence, we try the next transformation and re-plot the functions.

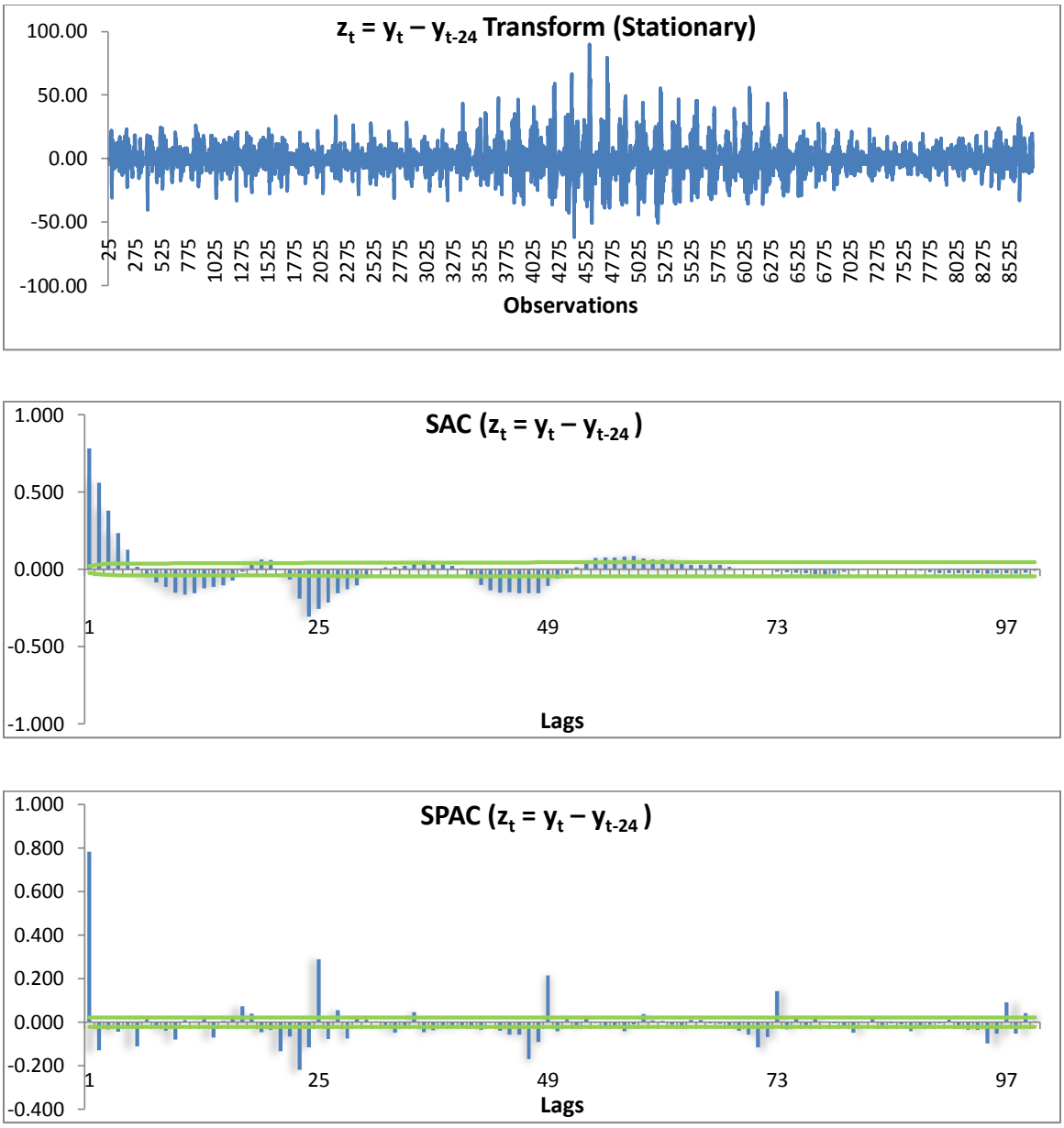
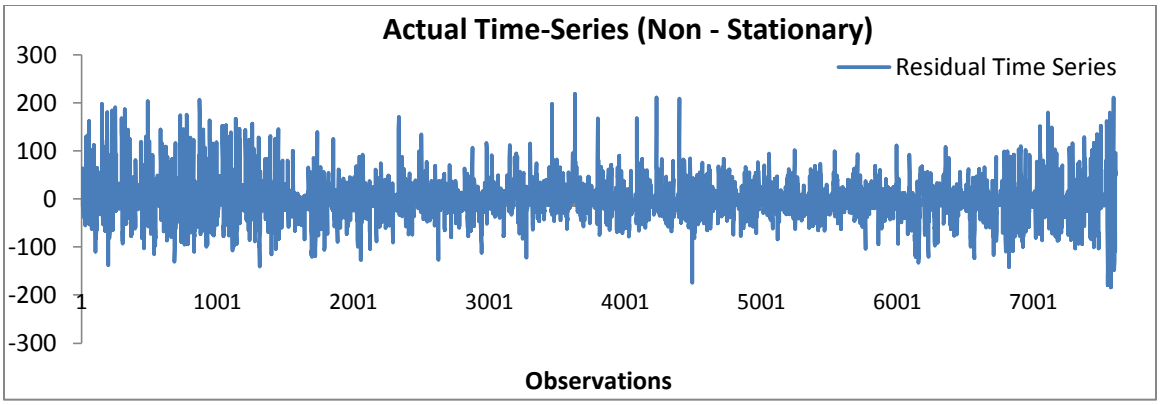
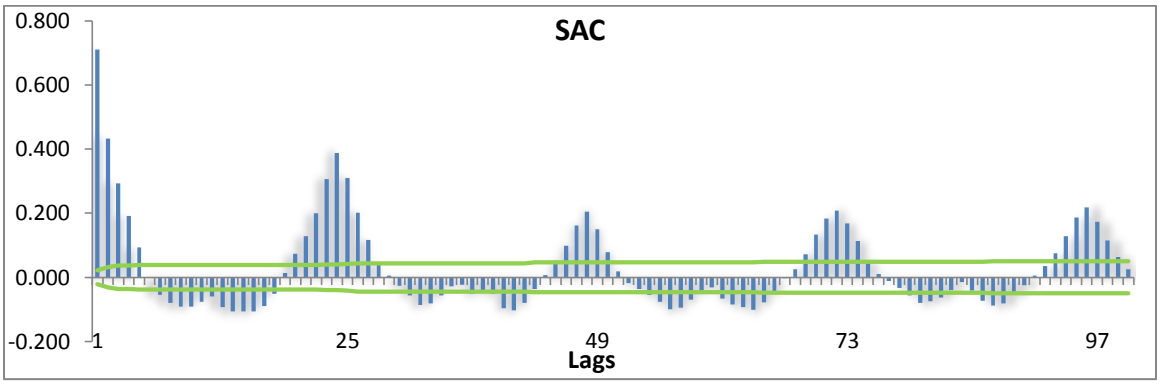


Figure 6.15 (a-c). SOB Transformed Series – Stationarity Check

We next try the 24 hour lag transformation. Figures 6.15 (a-c) indicate proper residual behavior, i.e. constant variation around the mean. Also, the SAC and SPAC die down fairly quickly both at the seasonal (lag 1, 2, 3, 4, and so on) and the non-seasonal (lag 24, 48, 72, and so on) level. Hence, we will adopt this transformation for our model development. Next, we evaluate the actual office building in a similar manner:





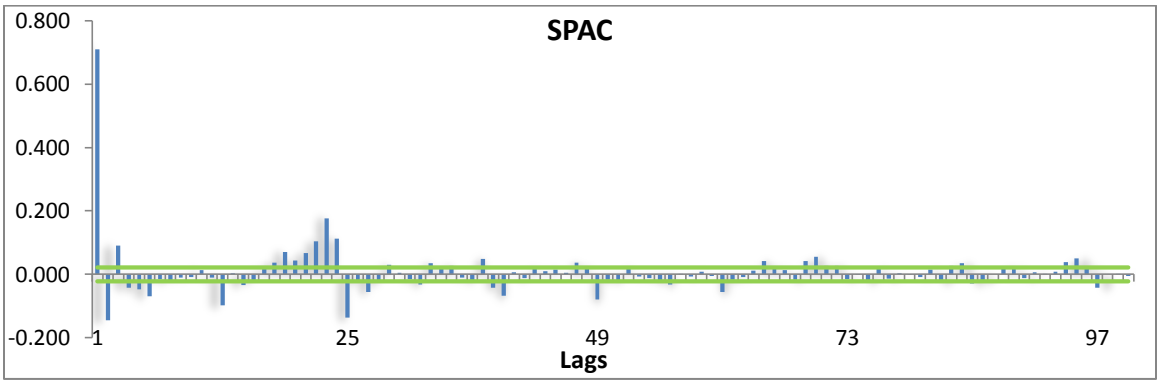
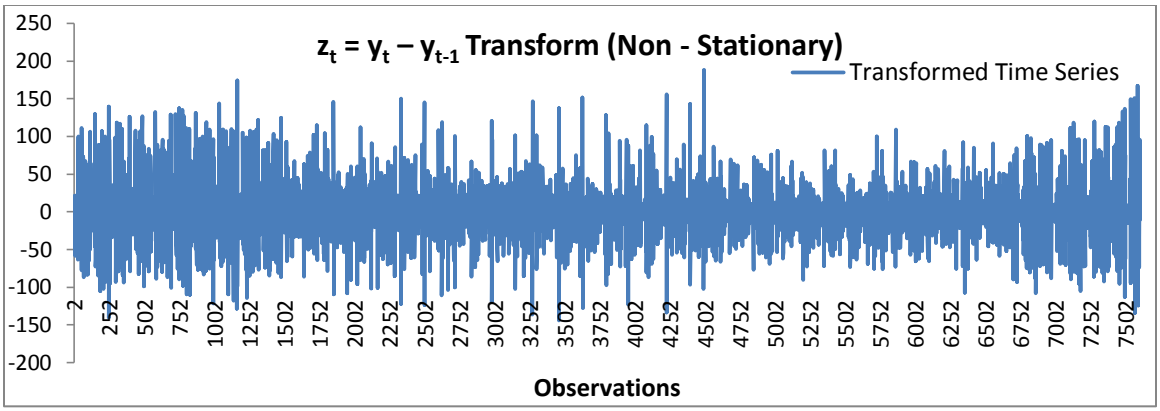
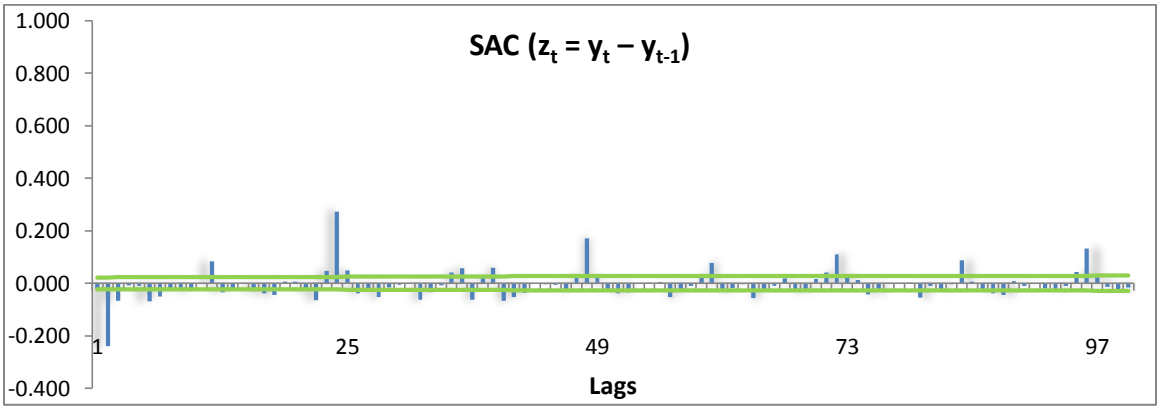


Figure 6.16 (a-c). AOB Actual Time Series – Stationarity Check

The actual residual time series does not exhibit constant variation around the mean and the SAC function do not cut off fairly quickly. Hence, we conclude that this series is non-stationary. We will transform the series as described in Section 4.5.1 and re-plot the functions.





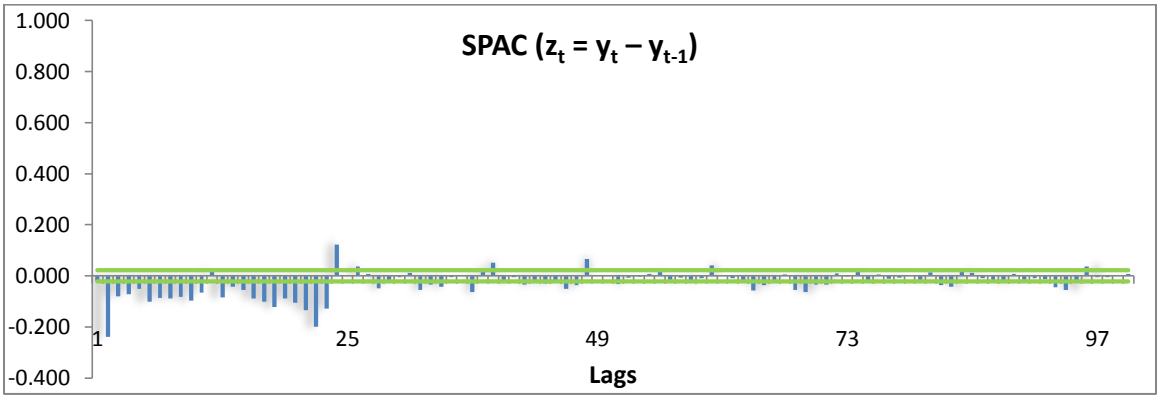


Figure 6.17(a-c). AOB Transformed Series – Stationarity Check

The transformed series in Figures 6.17 (a-c) still do not show constant variance and the SAC and SPAC functions do not seem to be cutting off fairly quickly at the seasonal level (lags 24, 48, 72, and so on). Hence, we try the next transformation and re-plot the functions.

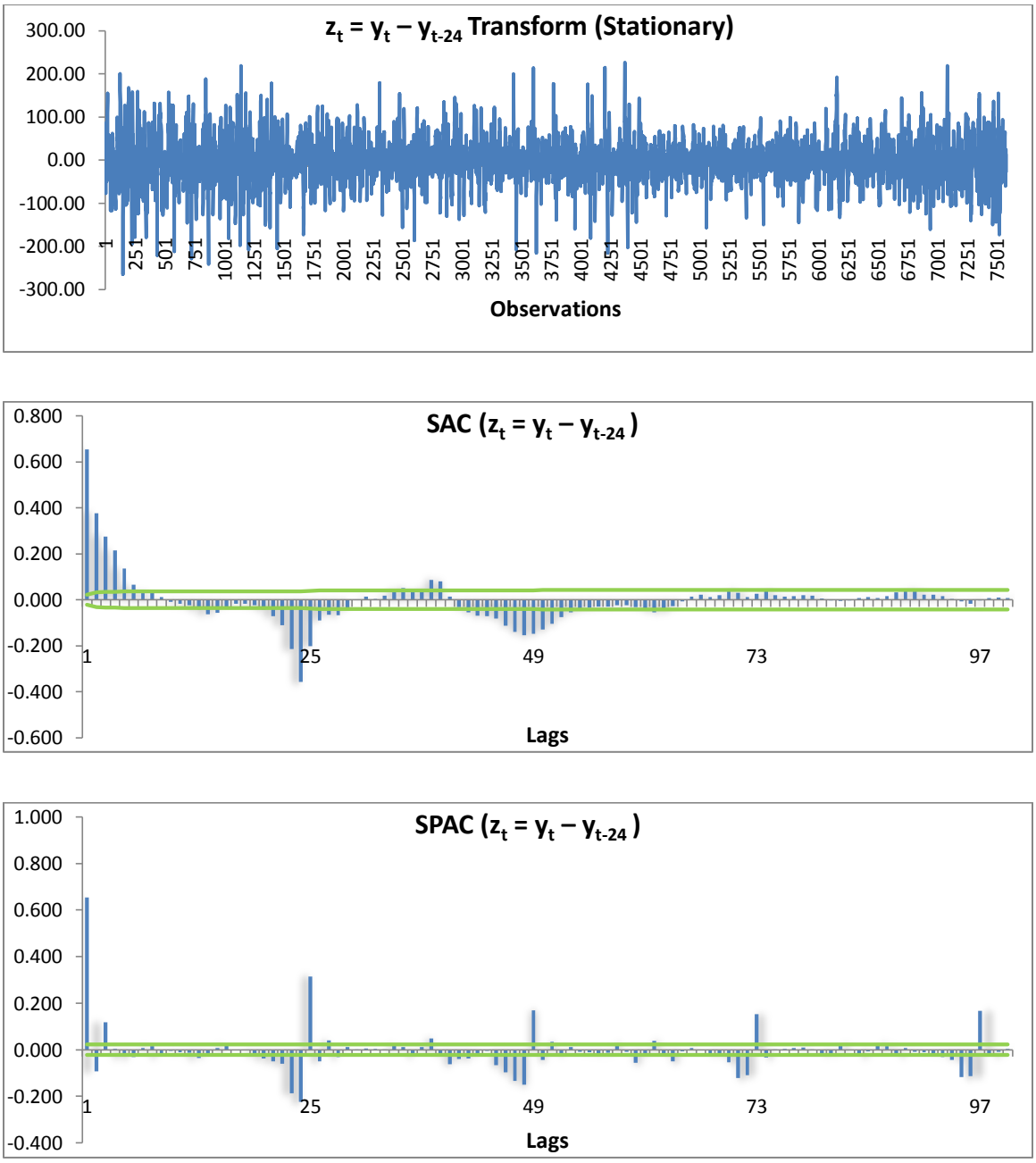


Figure 6.18(a-c). AOB Transformed Series – Stationarity Check

The transformed series in the Figures 6.18 (a-c) above seem to exhibit proper residual behavior. Also, the SAC and SPAC die down fairly quickly both at the seasonal (lags24, 48, 72, and so on) and the non-seasonal (lags1, 2, 3, 4, and so on) level. Hence, we will go ahead with this transformation.

6.4.2 Error Modeling

The SAC and SPAC of the transformed error series, $z_t = y_t - y_{t-24}$, exhibits seasonality at lags 24, 48, 72 and so on, and non-seasonality at lags 1,2,3 and so on. Based on this, we will try three Autoregressive (AR) models:

(i) Model based on Lag 1 error term,

(ii) Model based on Lag 24 error term. And,

(iii) Model based on Lags 1 and 24 error term.

Once complete, we plot the SAC and SPAC again to check if for any remaining autocorrelations. The final error series prediction models are as follows:

$$Z_{i,j+1} = \Phi_1 . (Z_{i,j} - Z_{i,j-1}) + \text{white noise} \quad \text{Eq. 6.5}$$

$$Z_{i,j+1} = \Phi_{24} . (Z_{i,j} - Z_{i,j-24}) + \text{white noise} \quad \text{Eq. 6.6}$$

$$Z_{i,j+1} = \Phi_1 . \left(Z_{i,j} - Z_{i,j-1} \right) + \Phi_{24} . (Z_{i,j} - Z_{i,j-24}) + \text{white noise} \quad \text{Eq. 6.7}$$

The coefficients for the above models are assembled in Table 6.7 and the model statistics in Table 6.8:

Table 6.7

Hourly Forecasting Model Coefficients – SOB and AOB

Hourly Energy Forecast Model Coefficients			
Synthetic Building (SOB)		**Actual Building (AOB)**	
Code	**Value**	**Code**	**Value**
AR(1) Model		AR(1) Model	
Φ_1	0.8158	Φ_1	0.7162

AR(24) Model		AR(24) Model	
Φ_{24}	0.6762	Φ_{24}	0.4166
AR(1) + AR(24) Model		AR(1) + AR(24) Model	
Φ_1	0.7520	Φ_1	0.6053
Φ_{24}	- 0.1655	Φ_{24}	- 0.2275

Table 6.8

SOB and AOB – Hourly Forecasting Model Statistics

Model Statistics				
	Synthetic Office Building		**Actual Office Building**	
Model No.	**RMSE (kWh)**	**CV – RMSE (%)**	**RMSE (kWh)**	**CV – RMSE (%)**
WBE_Hourly + AR(1)	7.112	7.32	33.36	9.21%
WBE_Hourly + AR(24)	10.912	11.21	41.13	11.36%
WBE_Hourly + AR(1) + AR (24)	6.884	7.07	31.84	8.79%

Table 6.8 makes it clear that models with both AR(1) and AR(24) terms have the lowest CV values. After AR modeling is complete, we evaluate the residuals to check for any significant remaining correlations. We plot the SAC and the SPAC functions for the residuals. As is clear from the Figures 6.19 and 6.20, both the SAC and SPAC do not exhibit any significant correlations both at the seasonal and non-seasonal level. Hence, we conclude that the AR(1) + AR(24) model accounts for the systematic stochastic component in the residuals generated after the hourly energy prediction model.

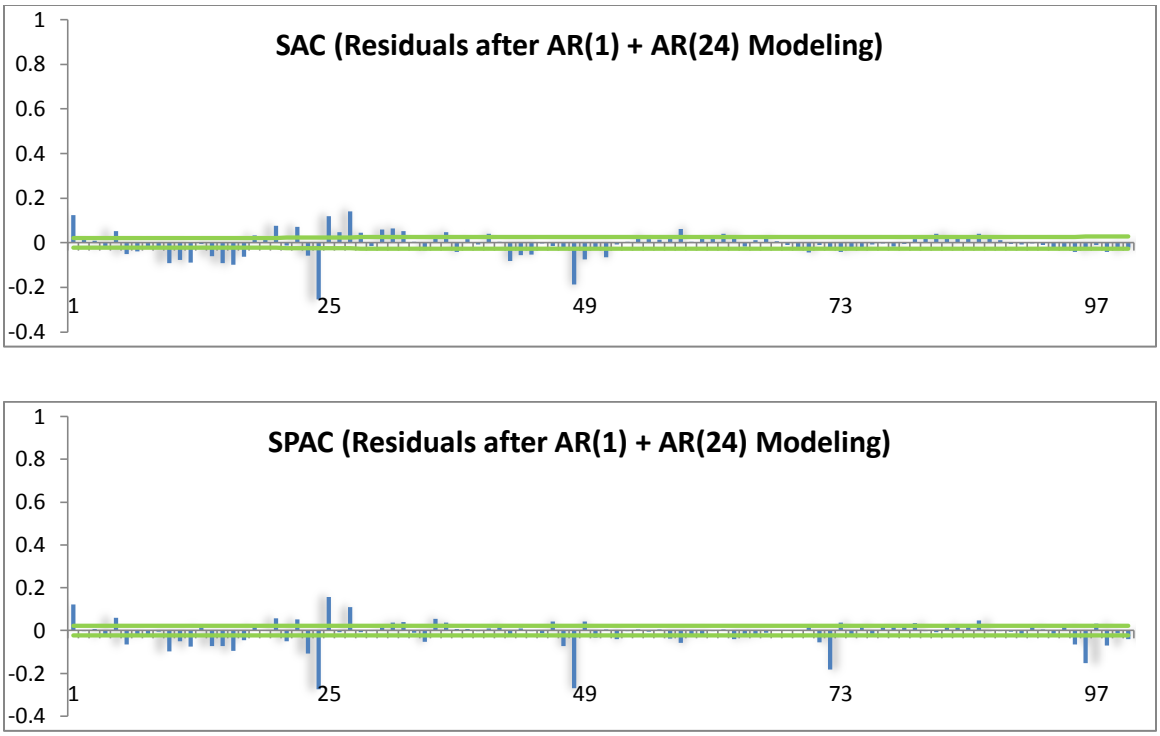


Figure 6.19 (a-b). SOB – SAC and SPAC (Residuals after AR Modeling)

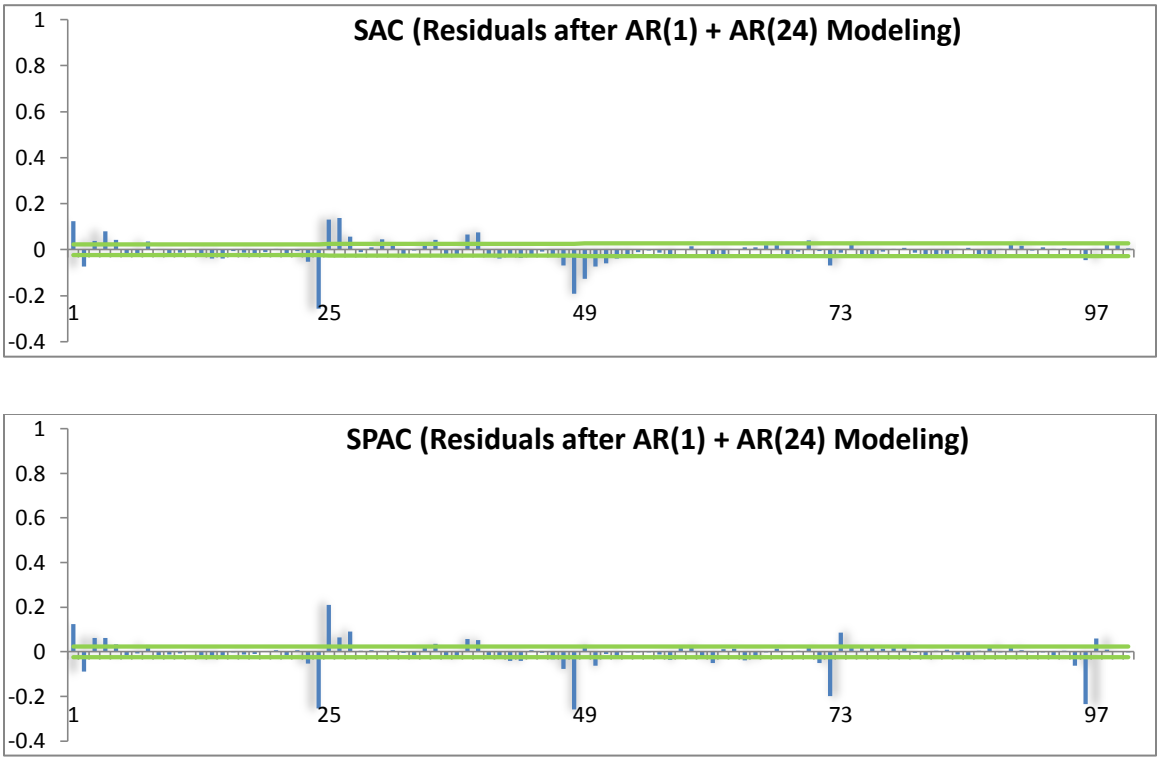


Figure 6.20 (a-b). AOB – SAC and SPAC (Residuals after AR Modeling)

6.5 Conclusions

The complete models summary is given in Table 6.9:

Table 6.9

SOB and AOB – Final Model Summary

MODELS SUMMARY (Synthetic Office Building)					
S.No.	**Model**	**Model Form**	**Model – R^2**	**RMSE (kWh)**	**CV – RMSE (%)**
1	WBE_Daily	$24 * \check{E}_i$	0.994	73.68	**3.08%**
2	WBE_Hourly	$\check{E}_i + \Delta\check{E}_{i.j}$	0.948	13.52	**14.26%**
3	WBE_Hourly + AR(1)	$\check{E}_i + \Delta\check{E}_{i.j} + \check{Z}_{t1}$		7.11	**7.32%**
4	WBE_Hourly + AR(24)	$\check{E}_i + \Delta\check{E}_{i.j} + \check{Z}_{t24}$		10.91	**11.21%**
5	WBE_Hourly + AR(1)+AR(24)	$\check{E}_i + \Delta\check{E}_{i.j} + \check{Z}_{t(1+24)}$		6.88	**7.07%**
MODELS SUMMARY (Actual Office Building)					
S.No.	**Model**	**Model Form**	**Model – R^2**	**RMSE (kWh)**	**CV – RMSE (%)**
1	WBE_Daily	$24 * \check{E}_i$	0.958	479.86	**5.44%**
2	WBE_Hourly	$\check{E}_i + \Delta\check{E}_{i.j}$	0.893	39.82	**10.98%**
3	WBE_Hourly + AR(1)	$\check{E}_i + \Delta\check{E}_{i.j} + \check{Z}_{t1}$		33.36	**9.21%**
4	WBE_Hourly + AR(24)	$\check{E}_i + \Delta\check{E}_{i.j} + \check{Z}_{t24}$		41.13	**11.36%**

5	WBE_Hourly + AR(1) + AR(24)	$\check{E}_{\boldsymbol{i}} + \Delta\check{E}_{\boldsymbol{i,j}} + \check{Z}_{\boldsymbol{t(1+24)}}$		31.84	**8.79%**

From Table 6.9, we can conclude that the daily energy prediction models are very robust and have high prediction accuracies as is indicated by the low CV values (about 3.1% for SOB and 5.4% for AOB). As explained in Section 6.3, the high CV values for the hourly prediction models (about 14.3% for SOB and 11% for AOB) are a result of systematic operational changes and the thermal mass effects of the buildings that the models are unable to capture. These high CV values at the hourly time scale can be reduced to some extent by including the AR terms. Finally, we account for the systematic stochastic component in the hourly model residuals by integrating these AR terms for the non-seasonal (lag 1 error term) and seasonal (lag 24 error terms) effects. These models ultimately result in the lowest CV values (about 7.1% for SOB and 8.8% for AOB) and have much higher prediction accuracies than the hourly model.

6.6 Building Condition Monitoring

As described in Section 4.5, there exists a mean-bias error at the monthly level since model identification was done using year-long data. Looking at Figure 6.21, months from July up till November have a positive bias; months from December up till June exhibit a negative bias. These model predictions can be corrected for this bias leading to reduced error variances thereby increasing deviation detection sensitivities.

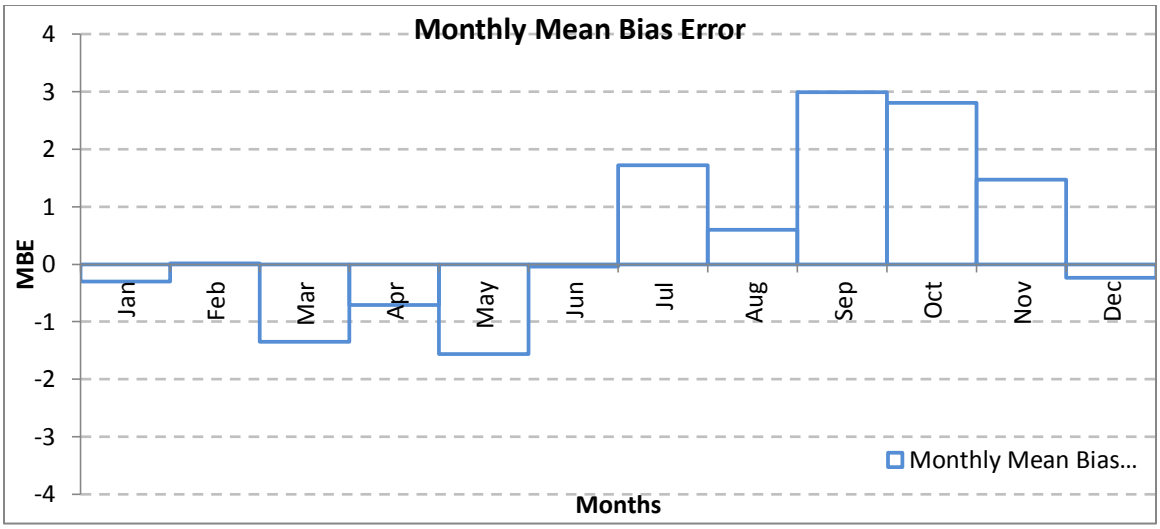


Figure 6.21. Monthly Mean-Bias Error

The mean-bias errors are calculated at the monthly scale for ease in explaining the concept.

Synthetic Office Building (SOB): In order to implement a Building Condition Monitoring scheme at the hourly level for all the months, we calculate mean-bias errors for each of the hours for different months. Table 6.10 assembles the calculated mean-bias errors at the hourly level for three of the 12 months of the year:

Table 6.10

SOB - Hourly Mean-Bias errors for different months

Hour	January (kWh)	April (kWh)	July (kWh)
1	13.5	7.4	-36.9
2	14.2	9.4	-35.1
3	15	10.8	-34.5
4	15.8	12.4	-33.3

5	16.3	13.8	-32.5
6	14.3	9.2	-37
7	10.7	-1.9	17.9
8	9.7	-12.8	24.7
9	16.3	2.2	47.6
10	-2.8	-4.2	31.6
11	-14.4	-13.1	25.3
12	-21.6	-19.7	20.2
13	-24.8	-21.9	8.3
14	-25.4	-19.4	11.8
15	-21.8	-15.1	16.6
16	-16	-7.5	28
17	-8.5	5.5	41.7
18	-11.6	-3.5	23.7
19	-11.5	-1.8	10.1
20	-7.2	3.1	13.4
21	0.4	5.8	16.4
22	1.7	3.7	12.9
23	12	3.5	-38.3
24	14	6.4	-36.3

These mean-bias errors are adjusted in the hourly predictions. Finally, regular and bias-adjusted condition monitoring charts are plotted as shown in Figures 6.22 – 6.24 (a-b):

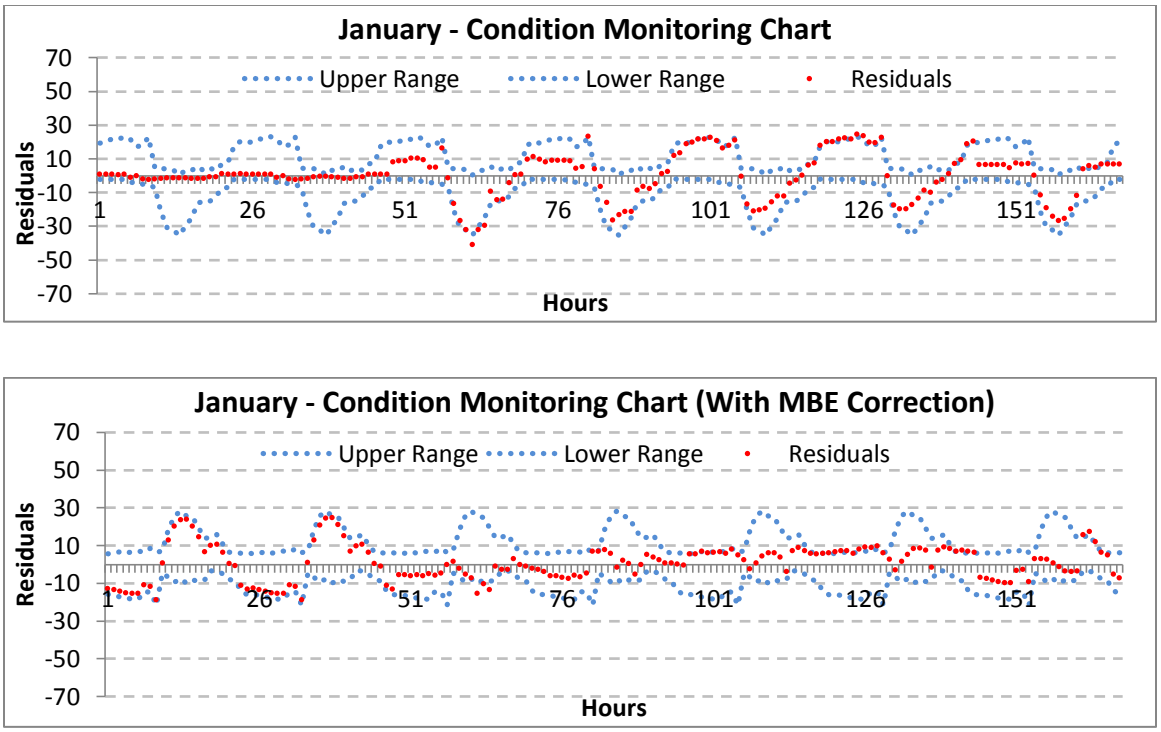


Figure 6.22 (a-b). SOB - Condition Monitoring Charts (January)

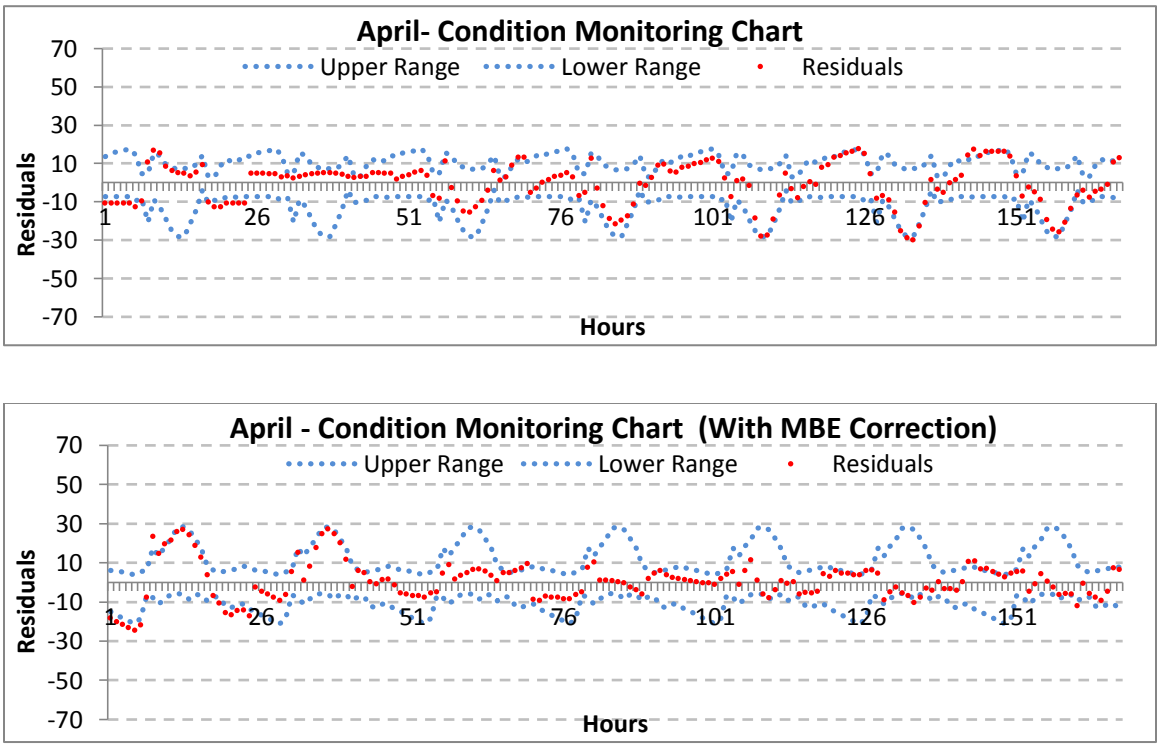
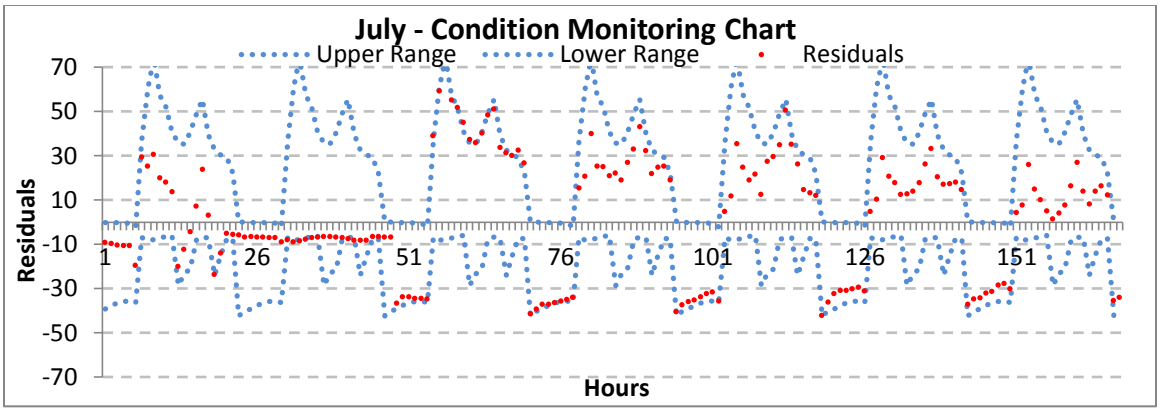


Figure 6.23 (a-b). SOB - Condition Monitoring Charts (April)



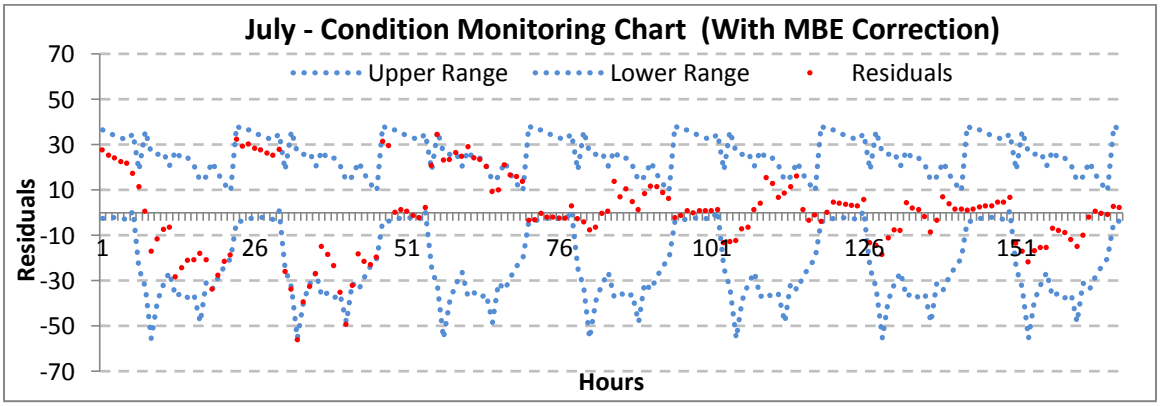


Figure 6.24 (a-b). SOB - Condition Monitoring Charts (July)

Looking at these charts above, and correlating them with Figure 6.21 above, we notice that the month of April has the tightest condition monitoring range due to it being a weather transition month; the months of January and July exhibit much broader ranges due to them being the extreme winter and summer months respectively. This shows the models inability to capture the high energy consumption owing to thermal lag effects, leading to much larger values of residuals in these months. Finally, we calculate the error variances before and after MBE correction assembled here in Table 6.11. There is a 35% reduction in error variances during the months of January and July and 25% reduction in the month of April due to MBE correction. This reinforces the weather-dependence of the building's energy consumption and the systematic over-prediction during winter months

and under-prediction during summer months discussed in Section 6.3 above. Ultimately, this reduction in error variances increases the sensitivity of detection.

Table 6.11

SOB – Residual Variances (Regular and MBE-Corrected)

Residual Variance Reduction by MBE correction					
January		**April**		**July**	
Residual Variance	Residual Variance (MBE Corrected)	Residual Variance	Residual Variance (MBE Corrected)	Residual Variance	Residual Variance (MBE Corrected)
139	90	98.5	76	524.5	341

Actual Office Building: We repeat the above procedure to generate building condition monitoring charts for the actual office building. Table 6.12 assembles the calculated mean-bias errors at the hourly level for three of the 12 months of the year:

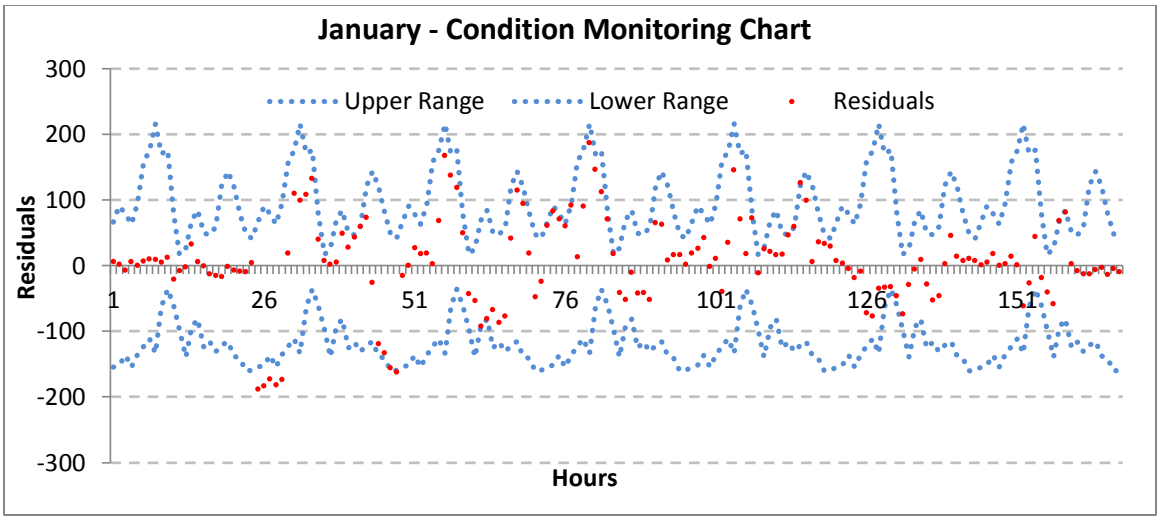
Table 6.12

AOB - Hourly Mean-Bias errors for different months

Hour	**January (kWh)**	**April (kWh)**	**July (kWh)**
1	-18.7	14.5	-67.3
2	-9.7	13.3	-64.6
3	-3.6	13.5	-65.4
4	-17.3	3.7	-68.7
5	1.7	31.5	-49.6
6	-17.2	59.8	-10

7	49	-11	38.3
8	109.4	-10.4	16.5
9	94.5	-44.5	15.6
10	87.1	-45.5	12.2
11	19.5	-23.5	32.5
12	-40	-0.6	23.3
13	-63.6	14.8	24.3
14	-36.5	-9.4	12.6
15	-20.2	-17.6	23.3
16	-37.4	-11.9	19.4
17	-38	4.5	20.8
18	-33.3	5.6	15.9
19	55.9	-6.2	-30.8
20	70.6	-11.7	-47.6
21	-0.4	-16.5	-43.3
22	1.	-19.2	-56.1
23	-7.7	-20.4	-57
24	-12	-6.4	-67.5

Finally, we plot the regular and MBE corrected condition monitoring charts as shown in Figures 6.25 – 6.27 (a-b) and calculate the residual variances in Table 6.13:



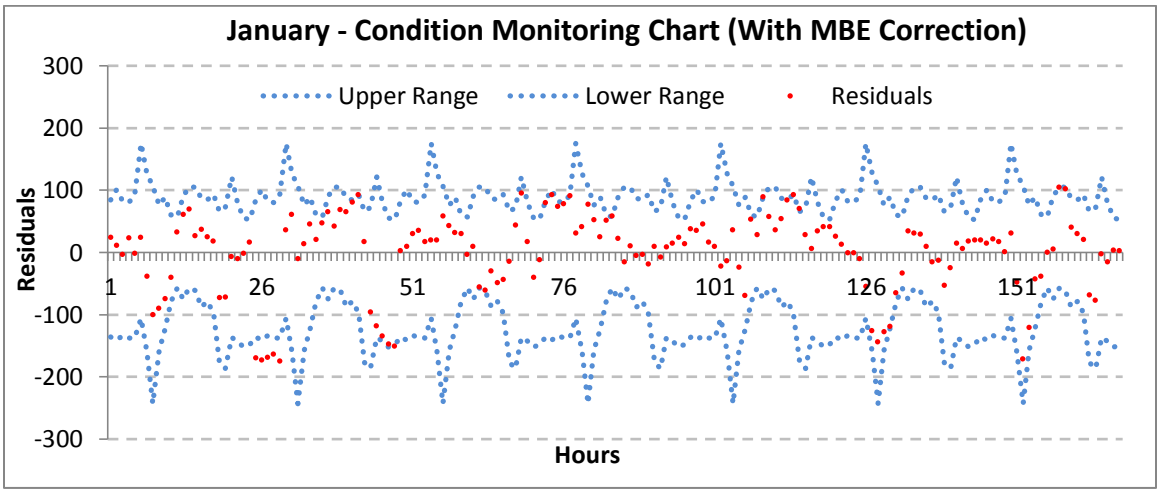
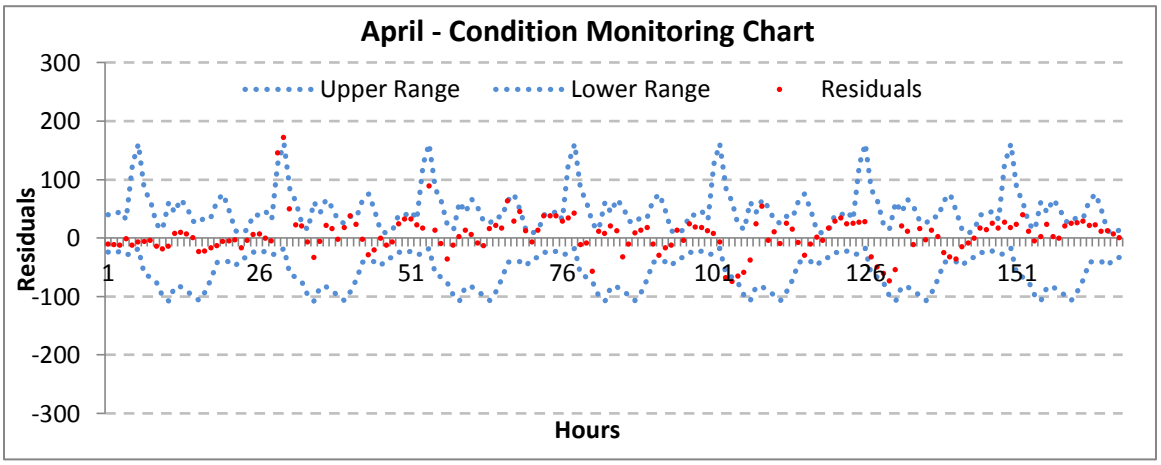


Figure 6.25 (a-b). AOB - Condition Monitoring Charts (January)



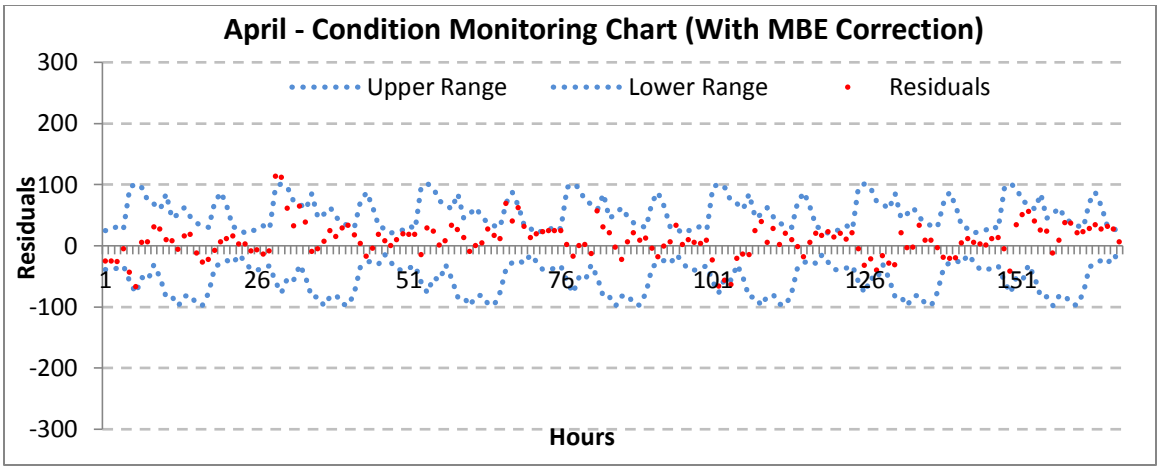
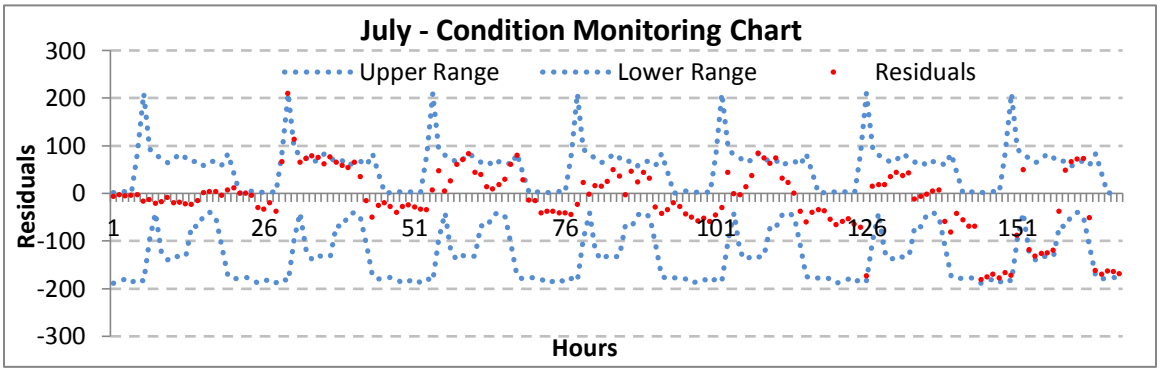


Figure 6.26 (a-b). AOB - Condition Monitoring Charts (April)



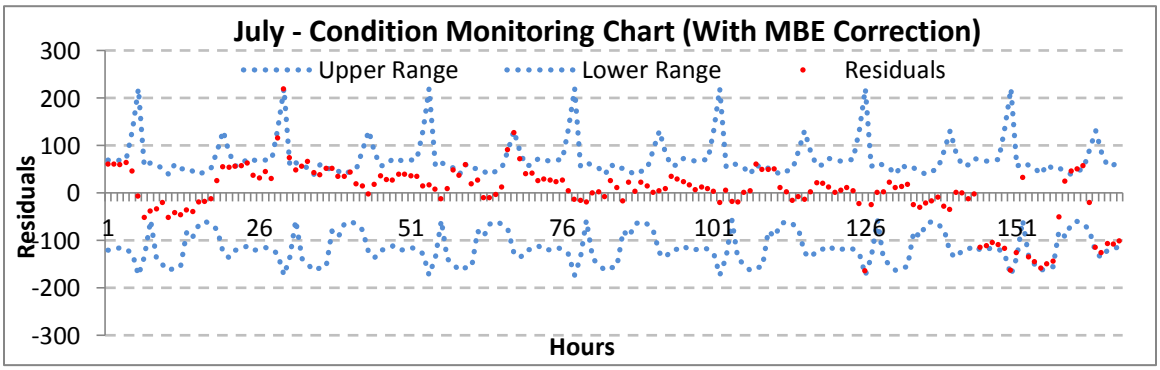


Figure 6.27 (a-b). AOB - Condition Monitoring Charts (July)

Looking at the plots in Figure 6.25 – 6.27 (a-b), we see that the month of April has the tightest condition monitoring ranges followed by the months of January and July,

which exhibit similar ranges. These can be explained by the operational changes of the building during the winter month of January and the thermal lag effects during the summer month of July as explained in Section 6.3. Finally, we calculate the reduction in error variances after MBE correction. There is an 8.5% reduction during the month of April and 11.5 % reduction during the months of January and July. Comparing these figures with the synthetic building above, we can conclude that actual buildings have much broader condition monitoring ranges owing to larger variations in these buildings.

Table 6.13

AOB – Residual Variances (Regular and MBE-Corrected)

Residual Variance Reduction by MBE correction					
January		April		July	
Residual Variance	Residual Variance (MBE Corrected)	Residual Variance	Residual Variance (MBE Corrected)	Residual Variance	Residual Variance (MBE Corrected)
4333	3850	1160	1062	2881.5	2536.7

CHAPTER 7 : METHODOLOGY AND PRELIMINARY RESULTS - FDD

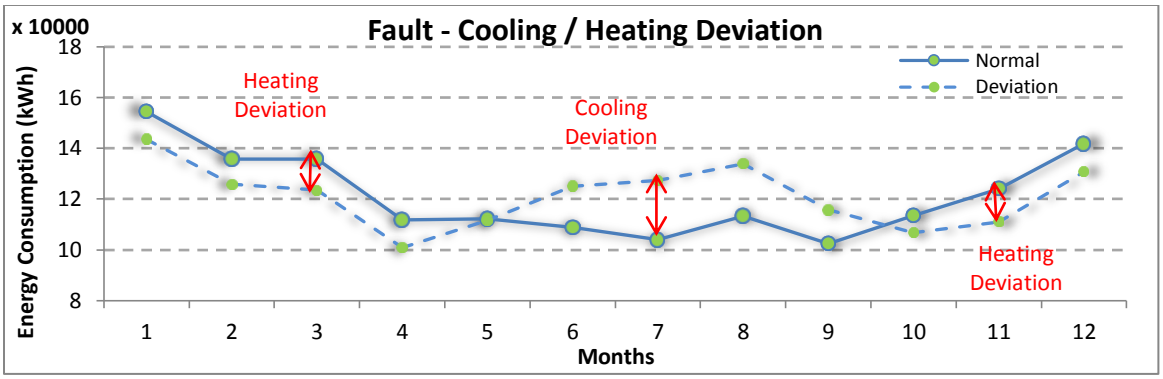
7.1 Introduction

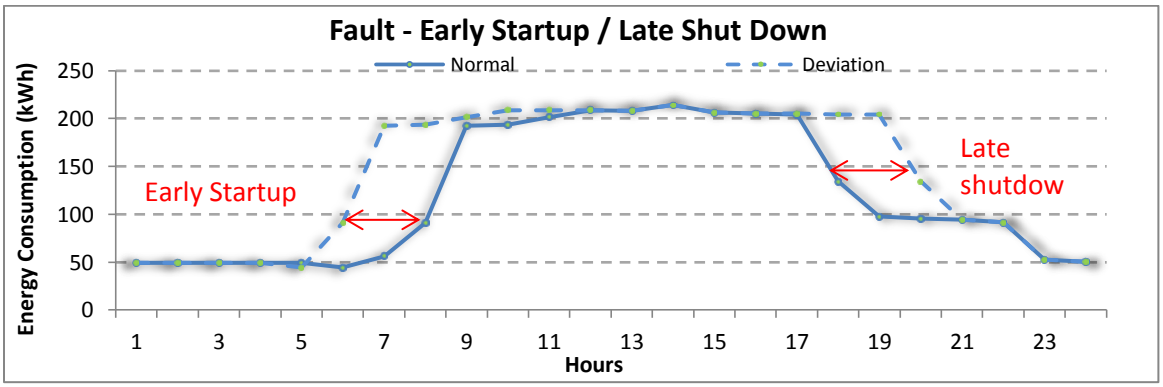
Fault Detection and Diagnosis (FDD) is a key component of many operations management automation systems. A 'fault' in the context of building energy may refer to any systems malfunctioning, and may either be a 'root cause' fault, i.e., a fault that may not be directly observable that could potentially lead to other problems, or 'directly observable' fault. In fault detection, a symptom is an observed event or a variable value needed to detect and isolate faults. Faults may be detected by a variety of quantitative or qualitative means. In this study, we present a discussion of the use of Energy Interval data stream for common fault detection purposes.

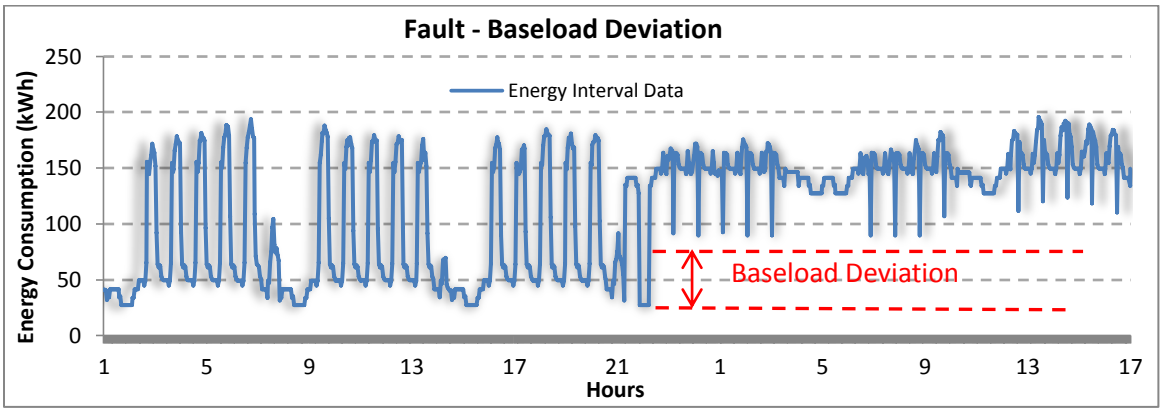
There are a number of past studies, such as Wei et al. (1998), that have proposed analyzing certain key energy consumption signatures and linking those to specific types of equipment related faults. The above study proposes the use 'calibration' and 'characteristic' signatures presenting their application for energy model calibrations. Characteristic signatures are energy signatures generated by inducing faults into simulation software and generating the energy consumption patterns. The difference between the normally operated and fault-induced energy consumption data will produce a residual pattern that would be characteristic to the fault induced. Calibration signatures, on the other hand, are the energy signatures generated by taking a difference between the actual consumption and the simulated consumption during energy model calibration. This residual pattern can be matched to different characteristic signatures to assess the variable changes required for calibrating the energy model. In the study above, the author

generated characteristic signatures for CHW and HW energy consumption due to different HVAC related faults.

Waltz (2000) provides an excellent overview of profile errors in his chapter of critiquing simulation output. It contains a good heuristic discussion of how to perform diagnosis from profile errors. Some of the examples presented are given in Figs 7.1 (a-d):







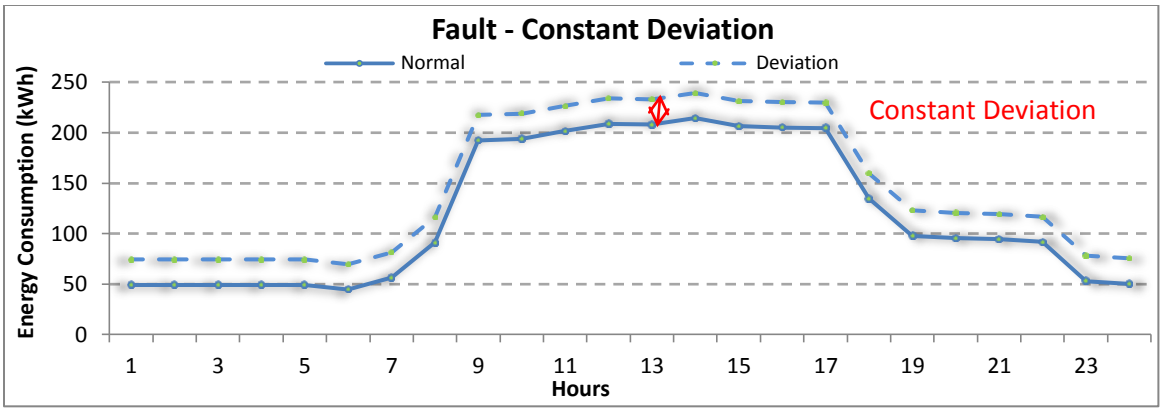


Figure 7.1(a-d). Various energy profile errors

7.2 Methodology

We take the idea of 'characteristic fault' generation and use prototype building energy models[3] to induce some of the common faults found in office building types. In this regard, the first step for us was to identify some of the fault types found most commonly in office buildings. Table 7.1 below assembles a list of various operational and component-related faults found most commonly in office buildings.

Table 7.1

Common Office Building Faults

COMMON OFFICE BUILDING FAULTS		
Fault Area	**S.No.**	**Fault Type**
Building Envelope	1	Excessive Air Infiltration
	2	Wall / Ceiling / Roof insulation is inadequate or has been damaged.

[3] http://www.energycodes.gov/development/commercial/90.1_models

	3	Poor Windows
Scheduling Related Faults	4	Early Building Startup
	5	Late Building Shutdown
	6	Improper Unoccupied Operation
	7	Unexpected Seasonal Variability
	8	Unexpected spikes in the evening
Operation	9	Improper Thermostat Set points
	10	Lights / Equipment left on at night
	11	HVAC ON during Unoccupied hours
AHU Related Faults	12	Excessive Ventilation Airflow
	13	Excessive Supply Air
	14	Improper Cold Coil Set point
	15	Clogged Filters
	16	COP reduction of Cooling Unit
	17	Improper AC refrigerant charge
	18	Improper Air Condenser Fan Operation
	19	Air Supply / Return Duct Leakage
Chiller Related Faults	20	Loss in efficiency of Units
	21	Improper Circulating Pump Operation
	22	Fouled Condenser Tubes
Boiler Related Faults	23	Drop in Combustion efficiency
	24	High / low air supply

Once complete, we decided to start by inducing two common faults in the Synthetic Office Building energy model, described earlier in Section 5.1. The two faults along with their variations are given in Table 7.2.

Table 7.2

List of simulated faults

FINAL SET OF SIMULATED FAULTS
(A) Thermostat Set point Deviation
Case I – Cooling Set point shifted from 75°F to 72°F. HVAC system turns OFF at night
Case II – No change in Set point. HVAC system stays ON at night
Case II - Cooling Set point shifted from 75°F to 72°F. HVAC system turns ON at night
Case IV – Heating Set point shifted from 70°F to 73°F. HVAC system turns OFF at night
Case VI - Heating Set point shifted from 70°F to 73°F. HVAC system turns ON at night
(B) Rooftop Chiller EER Degradation
Case I – 10% reduction in cooling system EER.
Case I – 20% reduction in cooling system EER.

Each of these faults was induced in the SOB energy model and it was simulated for three different climate types – Phoenix, Arizona, Denver, Colorado and Atlanta, Georgia. Once complete, the difference between the normal and fault-induced daily energy consumption points was taken and daily residuals were generated and plotted against outdoor dry-bulb temperature to generate fault patterns.

7.3 Preliminary Fault Generation Results

This section assembles the preliminary results of the daily residuals as a result of the faults discussed earlier in Table 7.2:

7.3.1 Cooling Set point Deviation

We first plot the outdoor dry-bulb temperature plots for the three climate types which would help us in analyzing these residual patterns:

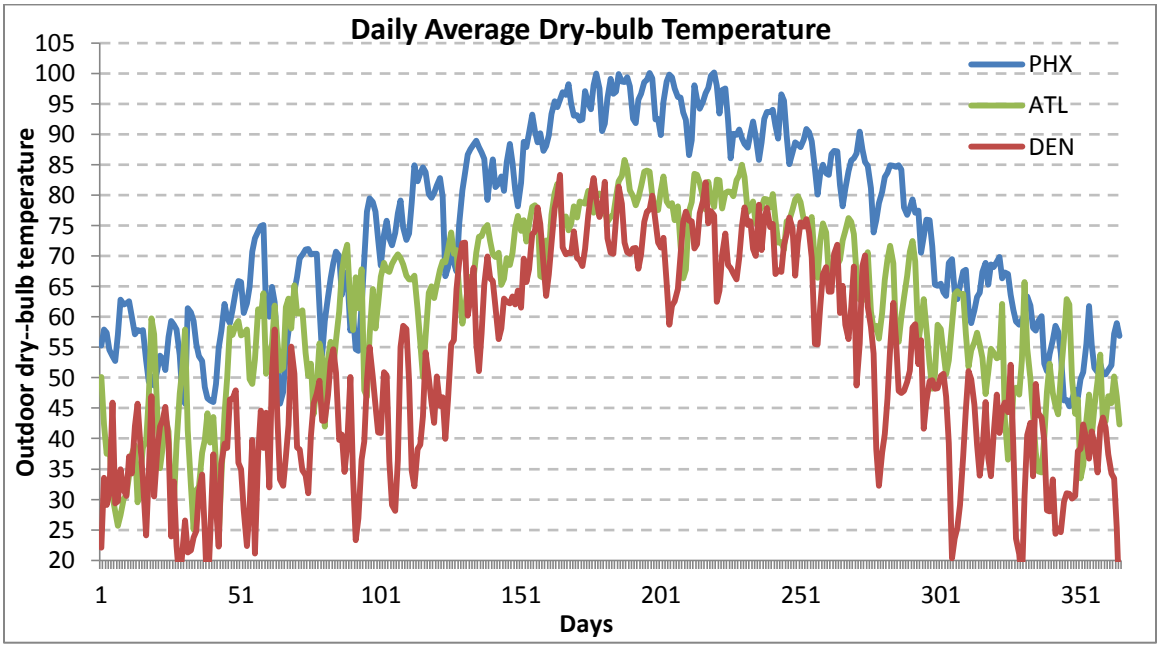
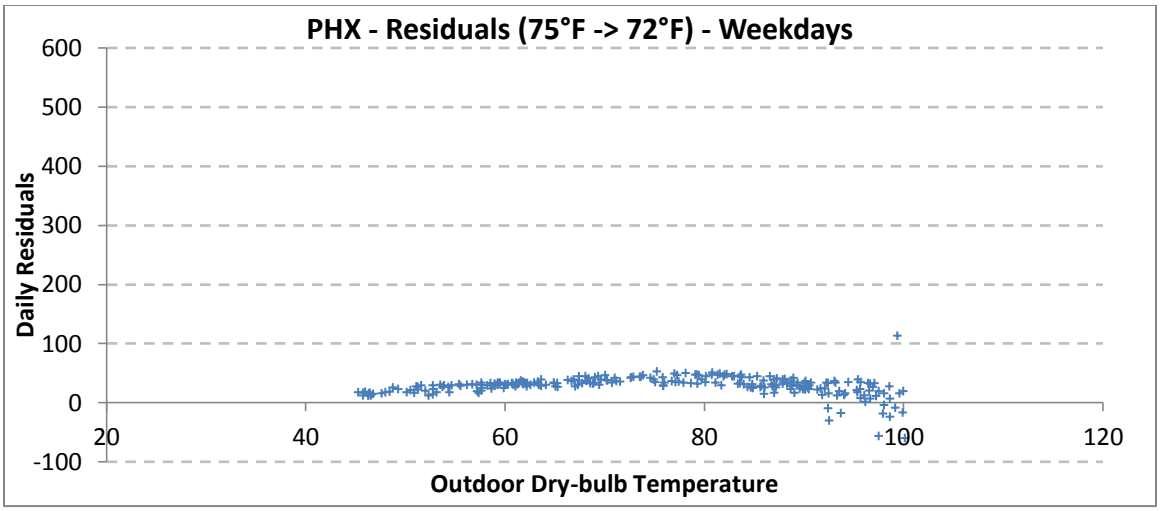
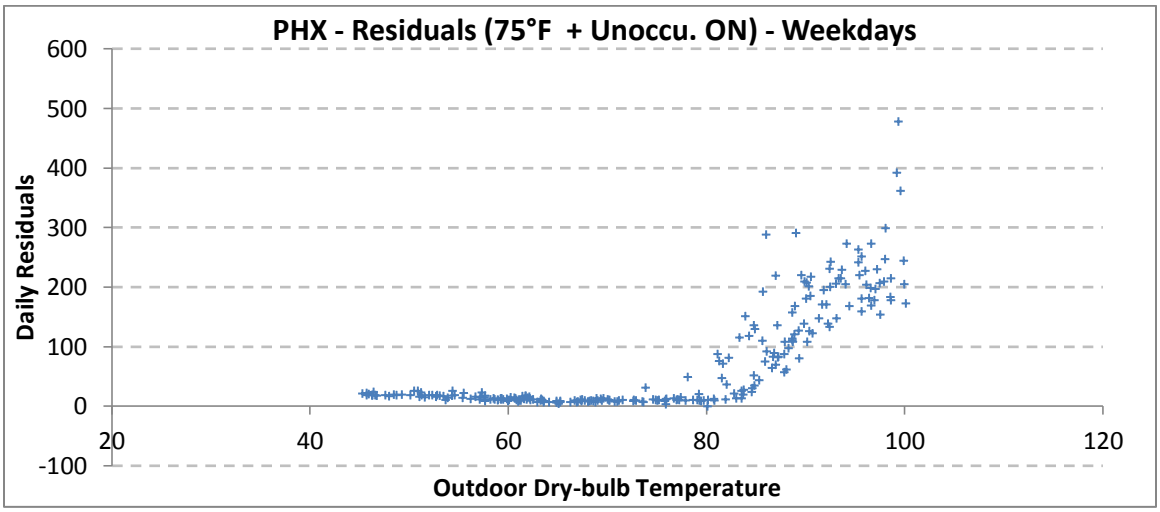


Figure 7.2. Outdoor dry-bulb temperature plots

Phoenix is the hottest climate type with almost 6 months above daily average temperatures of 75°F touching up to 100°F, followed by Atlanta with about 3 months but mostly within 80°F while Denver is the coolest with most of the time the temperatures are less than 75°F. Next, we plot the individual fault residual patterns for Cases I, II and III described earlier for all the three locations:





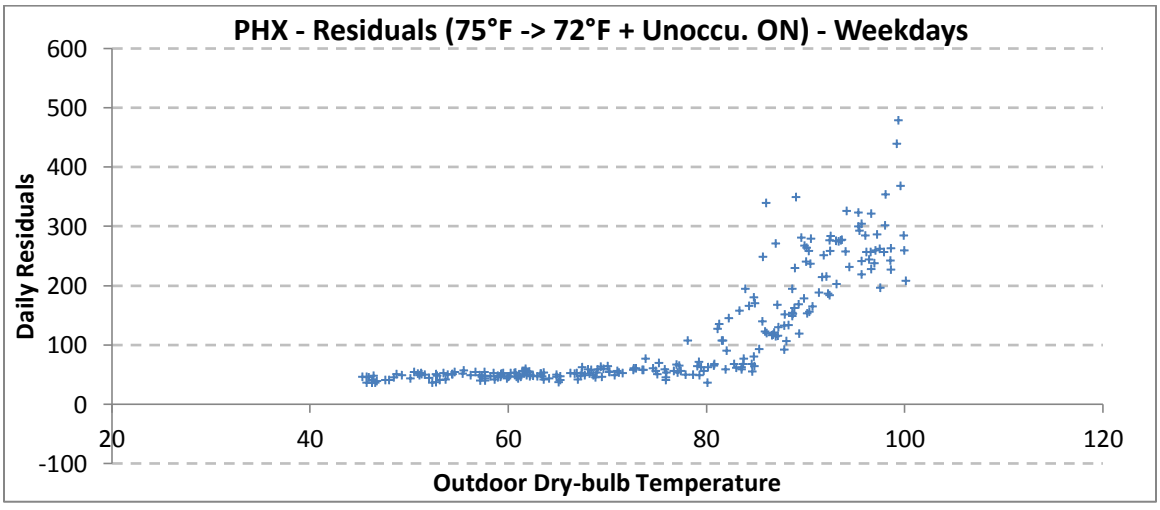
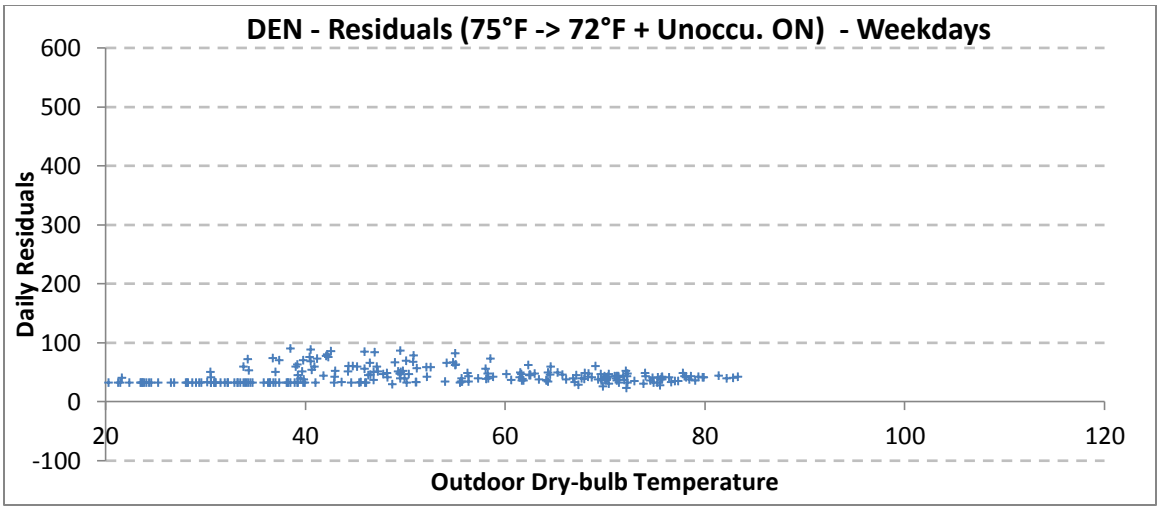
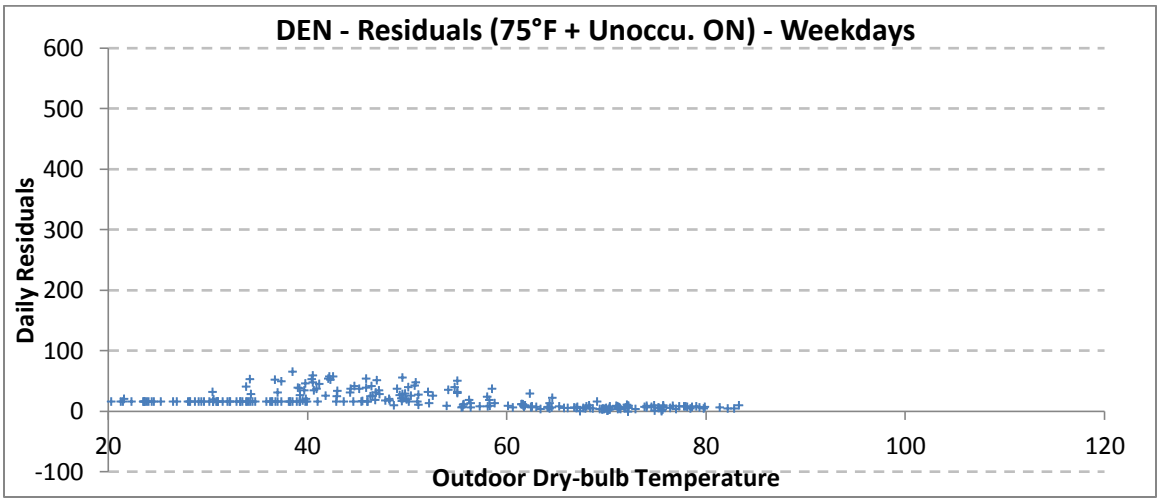


Figure 7.3(a-c). PHX - Cool Setpoint Deviation | HVAC ON during Unoccu. Hours





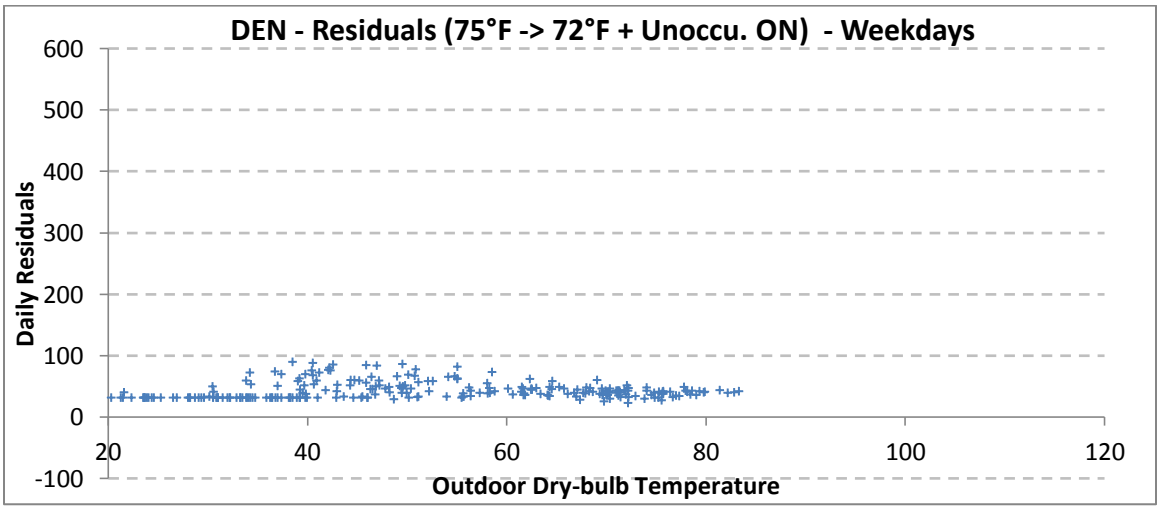
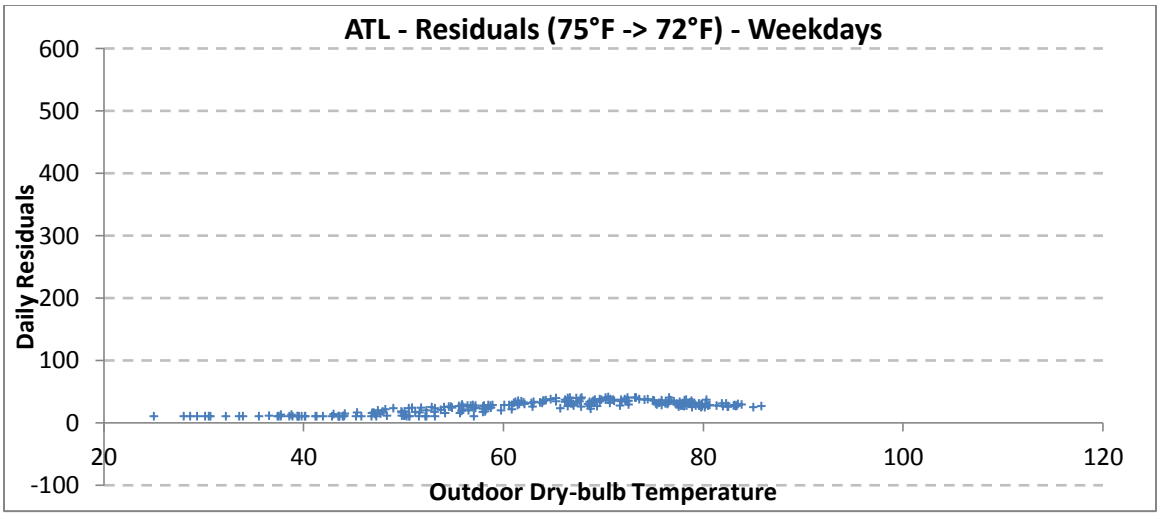
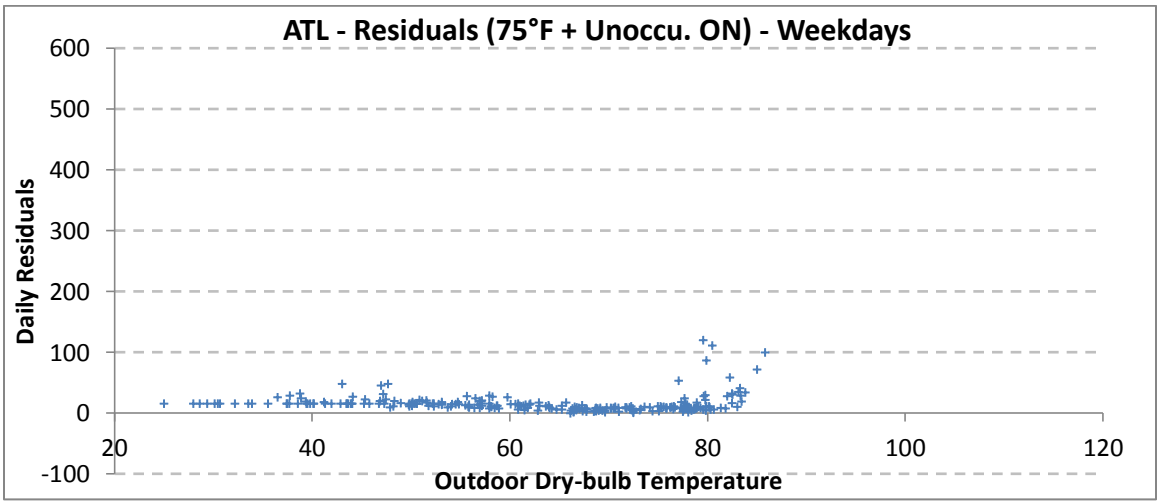


Figure 7.4(a-c). DEN - Cool Setpoint Deviation | HVAC On during Unoccu. Hours





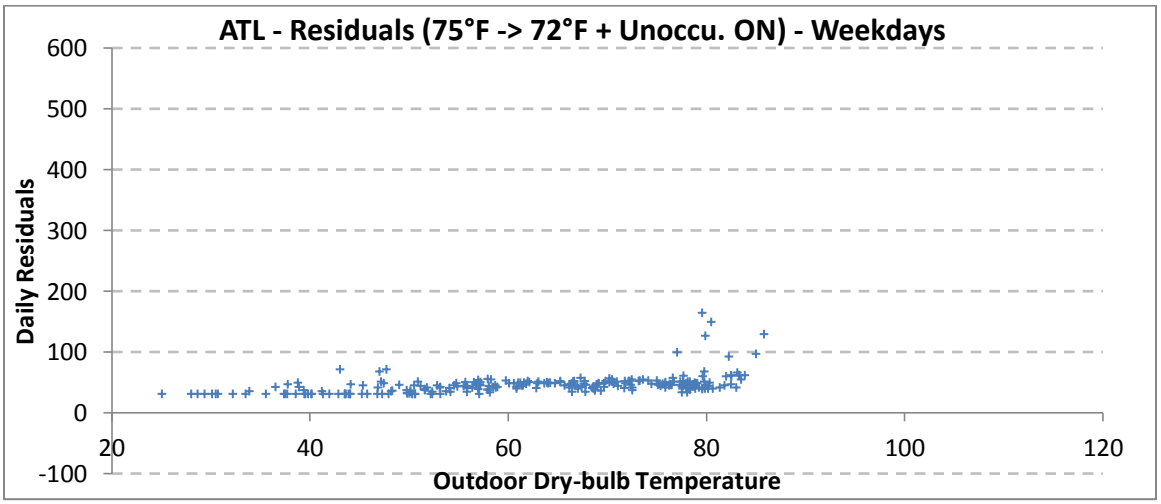
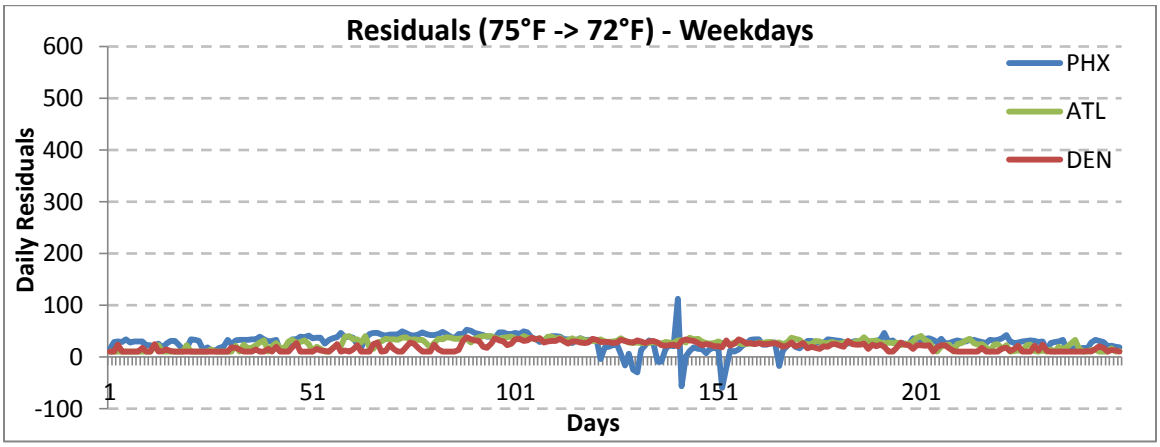
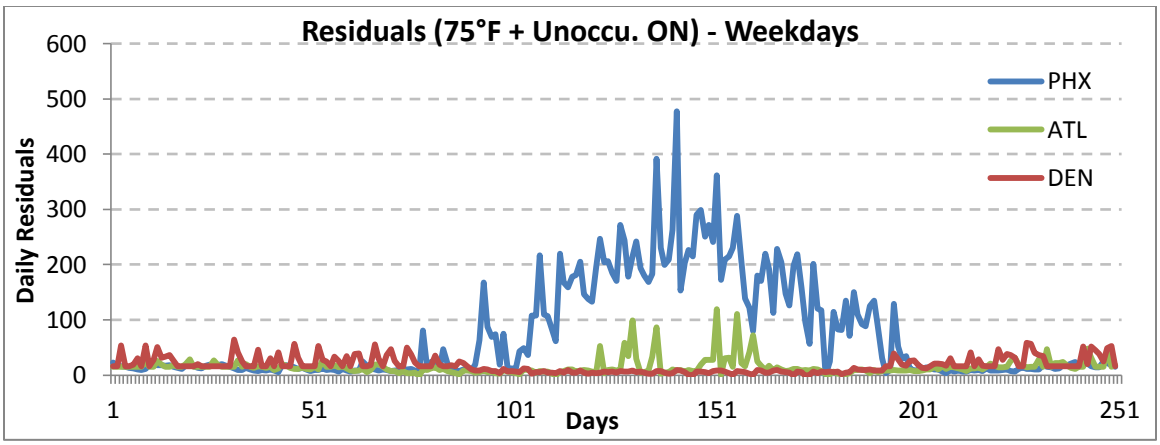


Figure 7.5(a-c). ATL - Cool Setpoint Deviation | HVAC On during Unoccu. Hours

Additionally, to understand the residual patterns above, we will plot the daily residuals against days for all the climate types together:





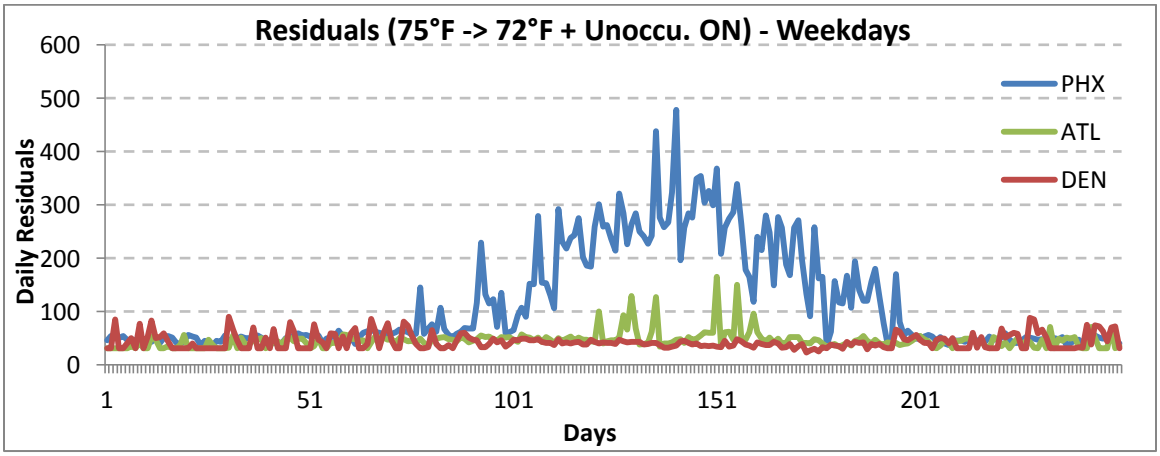


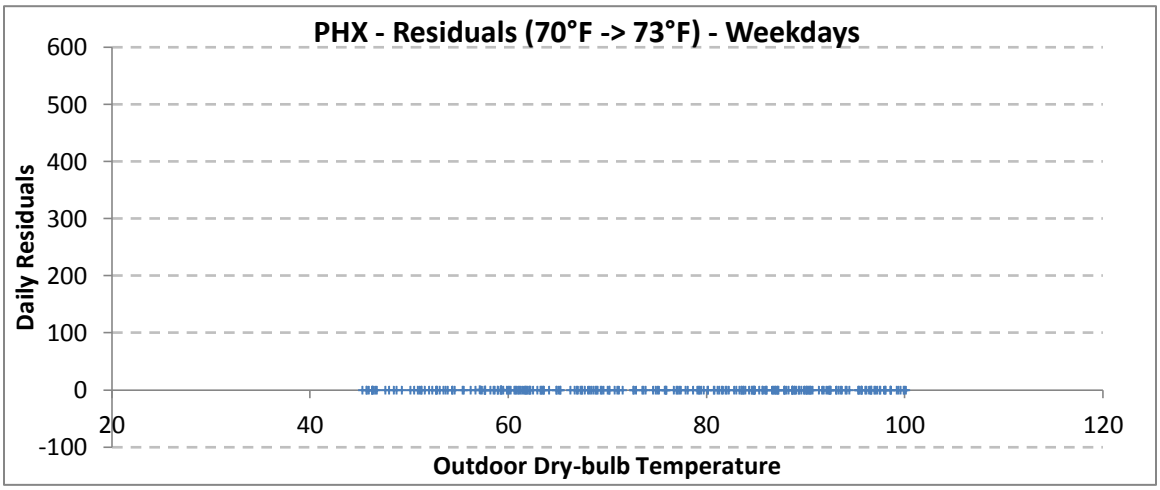
Figure 7.6(a-c). Comparative Residual Pattern Plots (Cooling Set point Deviation)

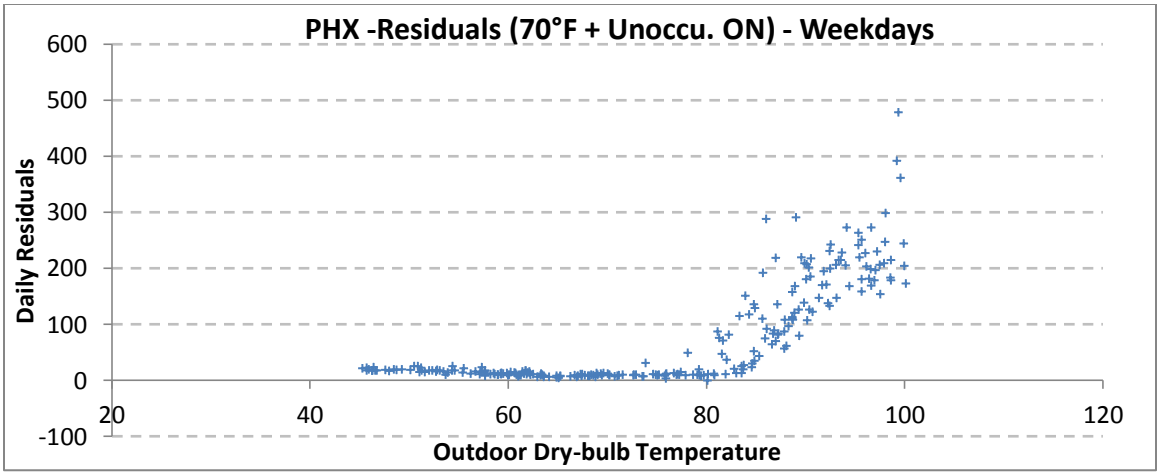
Looking at Figures 7.6 (a), and correlating these to Figures 7.3a, 7.4a and 7.5a, cooling set point deviation fault only increases energy consumption during the summer months and magnitude wise, Phoenix sees the highest increase followed by Atlanta and Denver, due to their individual temperature profiles as described in Figure 7.2. Additionally, for a small period during the peak summer, Phoenix shows a dip in the residuals which is also exhibited in the scatter moving downwards after 80°F in Fig 7.3a. Since the building starts in the morning and cools down significantly due to the lower set point, the building thermal mass helps retain some of the cooling energy and hence slightly lower energy is consumed during the early afternoon hours for further cooling. By late afternoon, the system picks up but quickly reduces as the outdoor temperature begins to drop leading to lower energy consumption during early evening hours again.

Leaving the HVAC system ON during the unoccupied hours results in a huge penalty on energy consumption and is the largest for Phoenix followed by Atlanta. Denver exhibits a slight increase in energy consumption of low magnitudes between 30°F-60°F. Correlating figure 7.4b with 7.6b, we can see that Atlanta begins to show a penalty after 77-78°F, but the rise in the pattern is curbed because the weather does not go much beyond the 80°F described earlier. For Phoenix, we can correlate the residual patterns in Figure 7.6b to 7.3b and can safely say that the resulting excess in energy consumption is due to high-night time temperatures in Phoenix.

Finally, combining the above two faults, it is clear that the resulting residual patterns are dominated by the unoccupied hours operation and not so much so by the cooling set point deviation. On close inspection, we can say that residual patterns in Figure 7.3c-7.5c can be obtained by adding up the patterns in 7.3a-7.5a and 7.3b-7.5b. .

7.3.2 Heating Set point Deviation





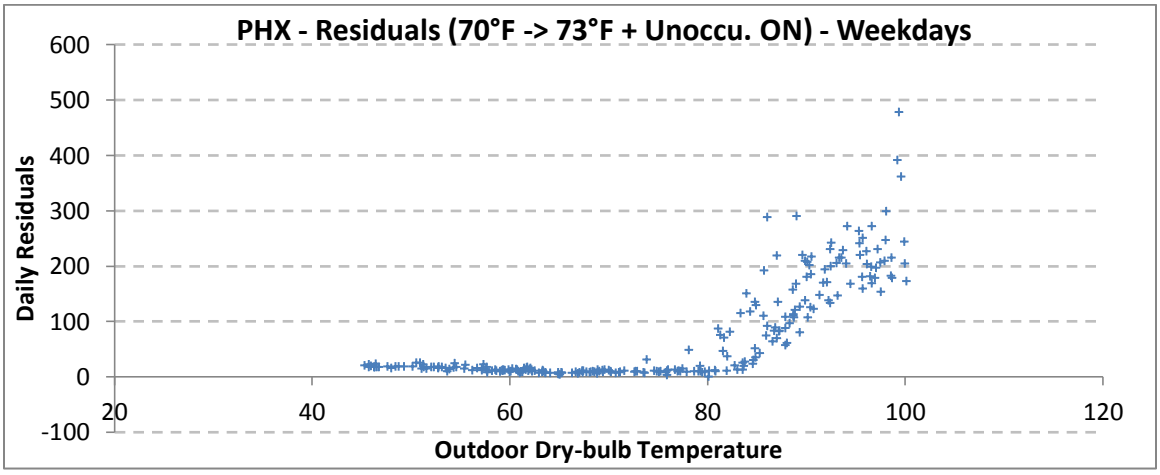
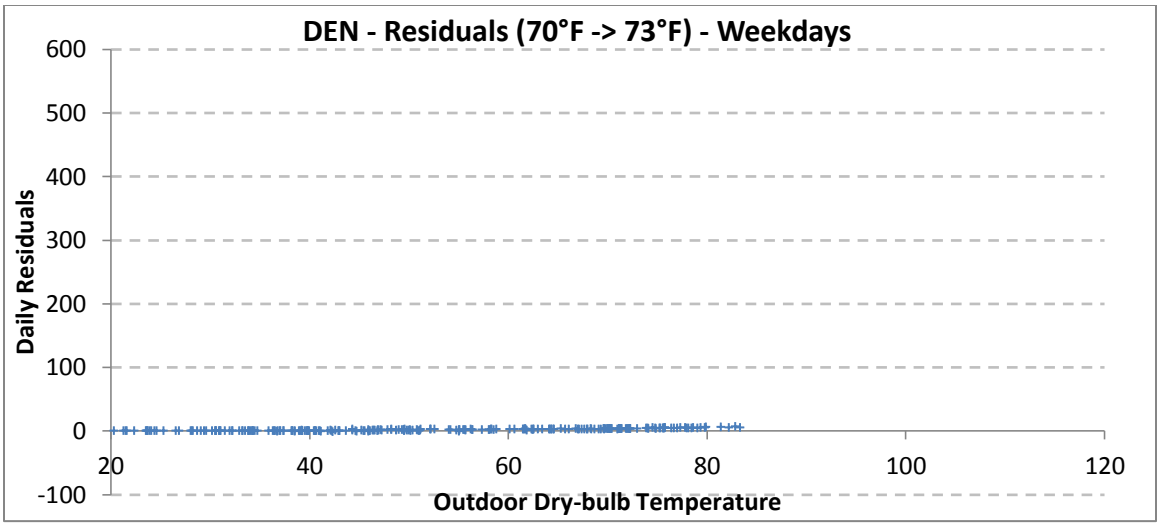
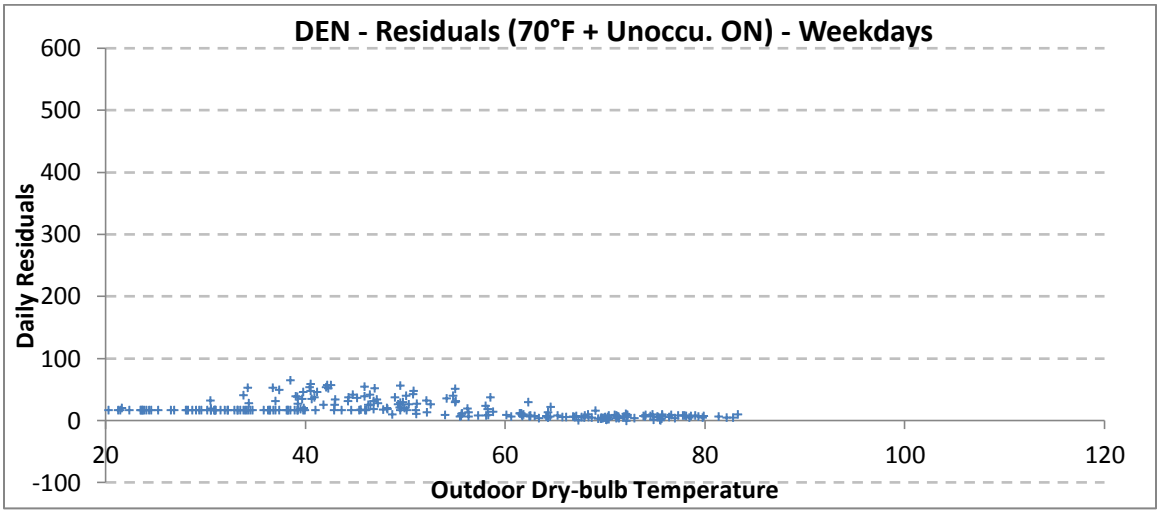


Figure 7.7(a-c). PHX-Heating Setpoint Deviation | HVAC On during Unoccu. Hours





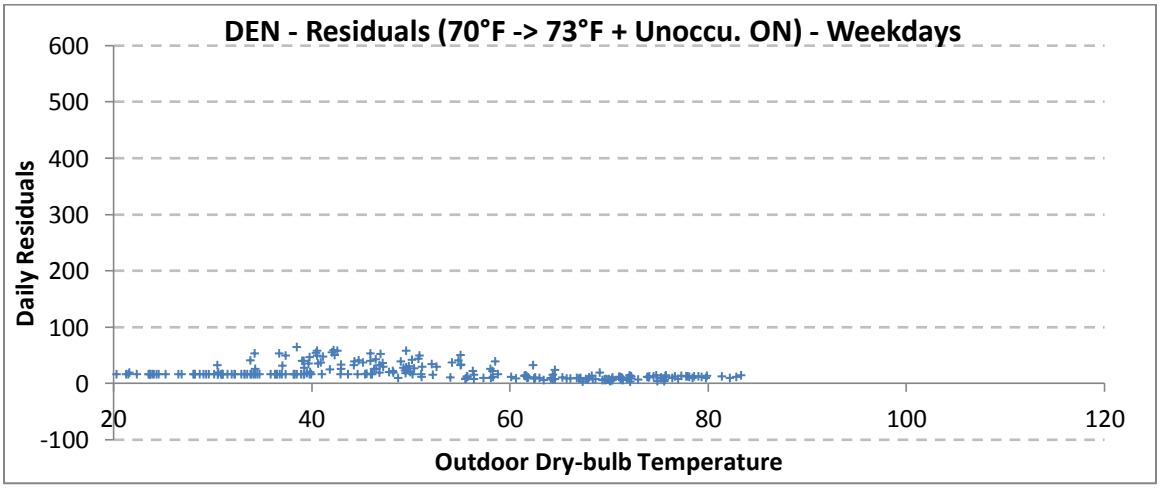
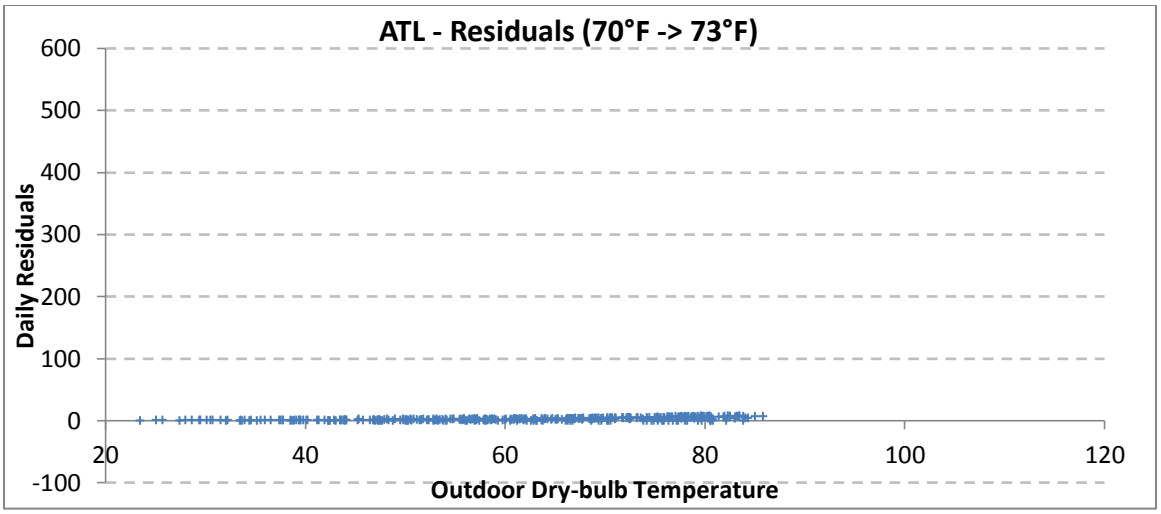
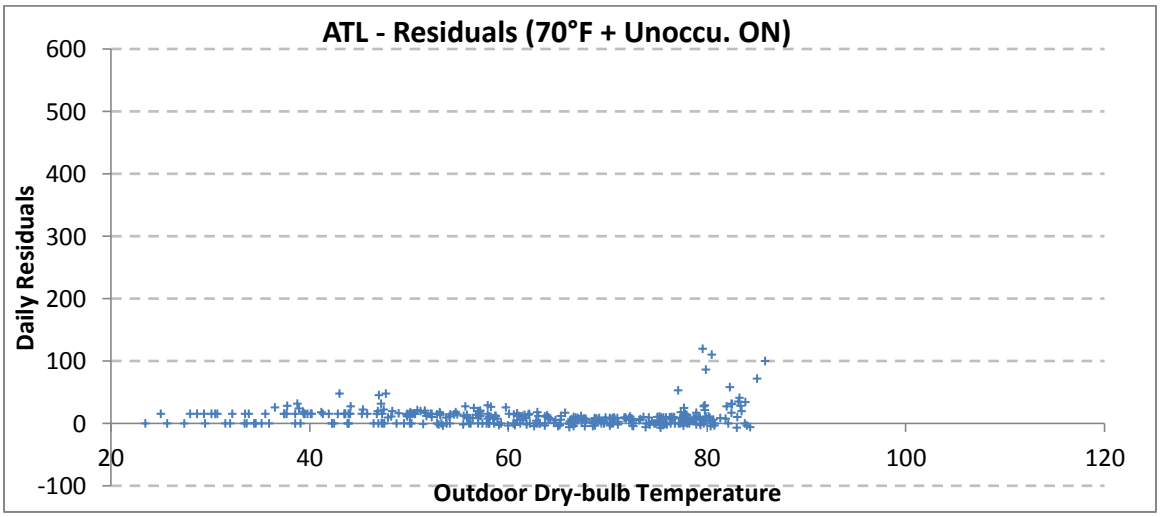


Figure 7.8(a-c). DEN-Heating Setpoint Deviation | HVAC On during Unoccu. Hours





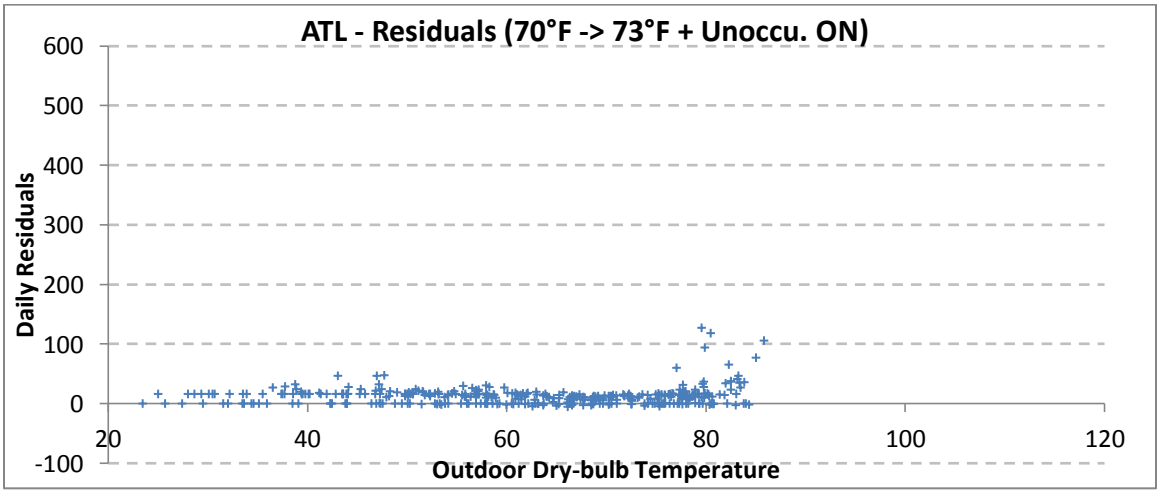


Figure 7.9(a-c). ATL-Heating Setpoint Deviation | HVAC On during Unoccu. Hours

Repeating the exercise, this time with heating set point deviations, we notice similar behaviors regarding the operations during unoccupied hours. However, the changes in set points have little to no effects in all the three climate types.

7.3.3 EER Degradation

Next, we generate the residual patterns due to EER degradation of the HVAC systems for the three different climate types. We evaluate two levels of EER degradation:

(i) 10% EER degradation from the baseline of 9.8 EER

(ii) 20% EER degradation from the baseline of 9.8 EER

The resulting patterns are as follows:

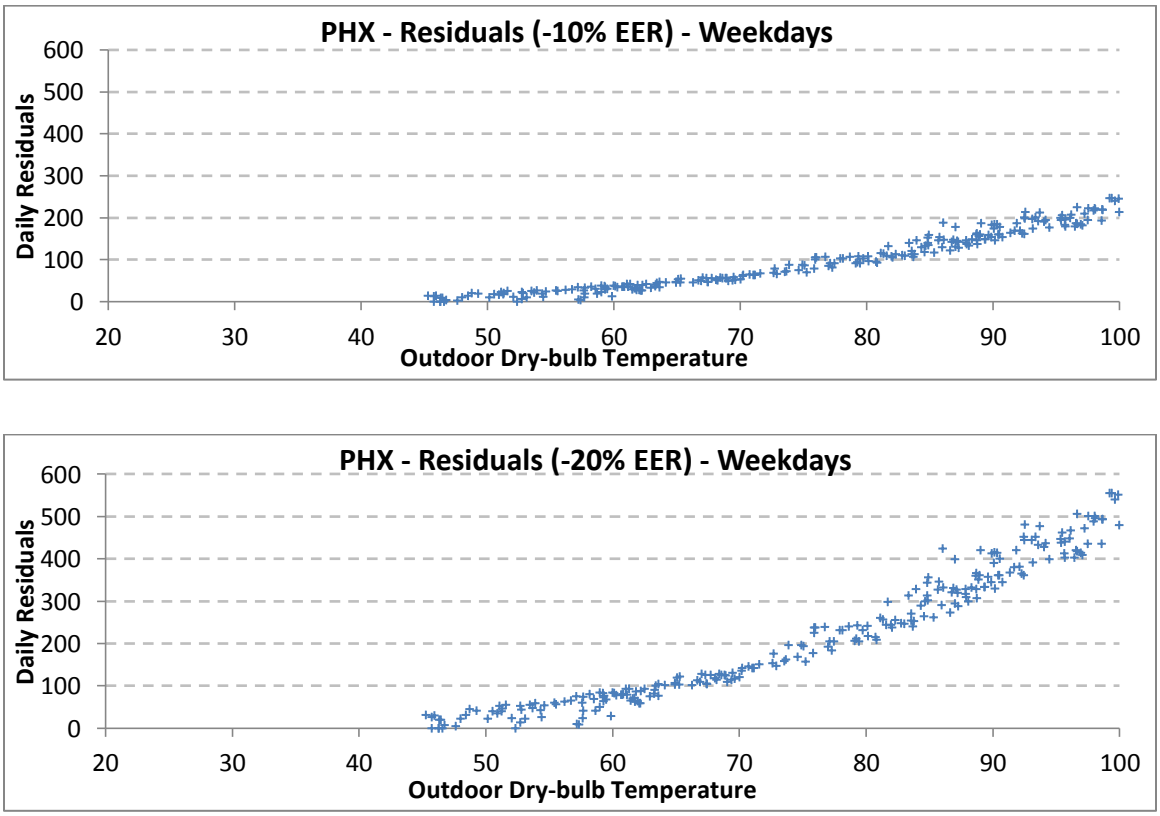
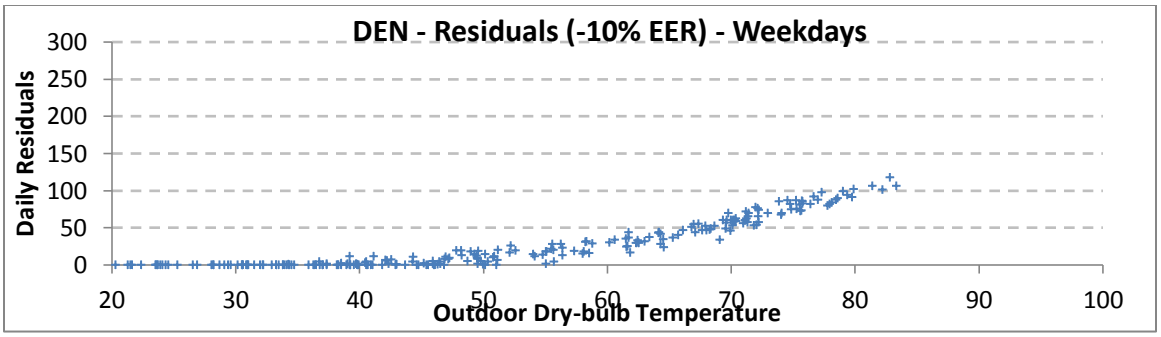


Figure 7.10(a-b). PHX – EER Degradation (-10% and -20%)



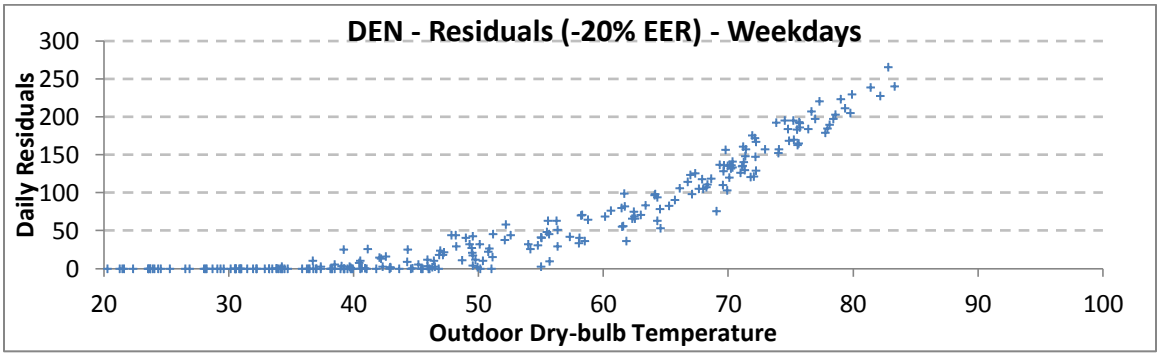
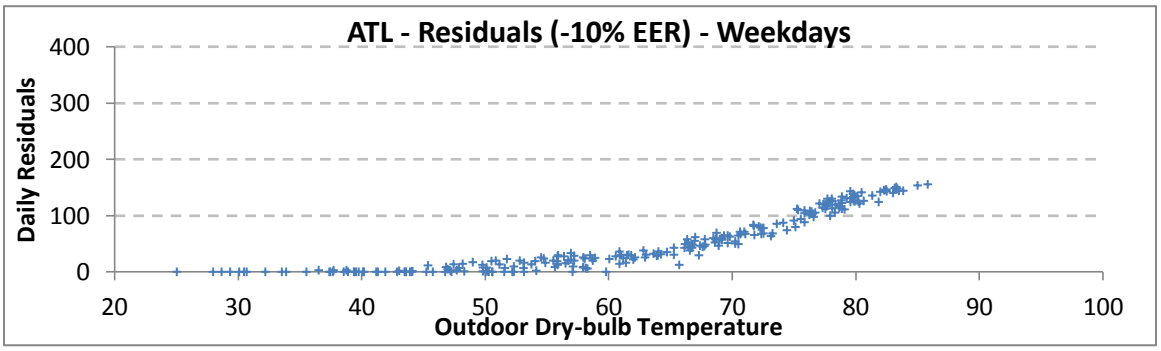


Figure 7.11(a-b). DEN – EER Degradation (-10% and -20%)



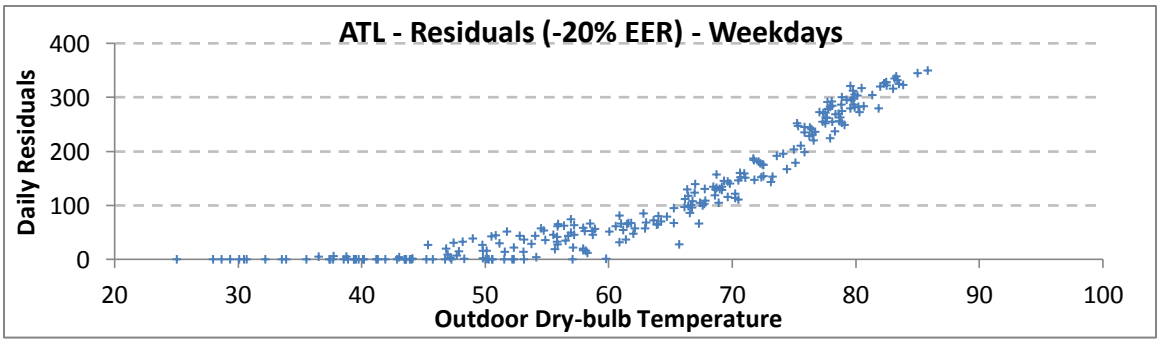


Figure 7.12(a-b). ATL – EER Degradation (-10% and -20%)

Looking at Figure 7.10, 7.11 and 7.12 (a-b), it is evident that the EER degradation increases energy consumption in all the climate types and that this increase is very strongly correlated to the increase in outdoor dry-bulb temperature. Increased degradation leads to higher energy consumption. We plot comparative residual plots to understand the differences in residual patterns for the different climates types due to the same fault.

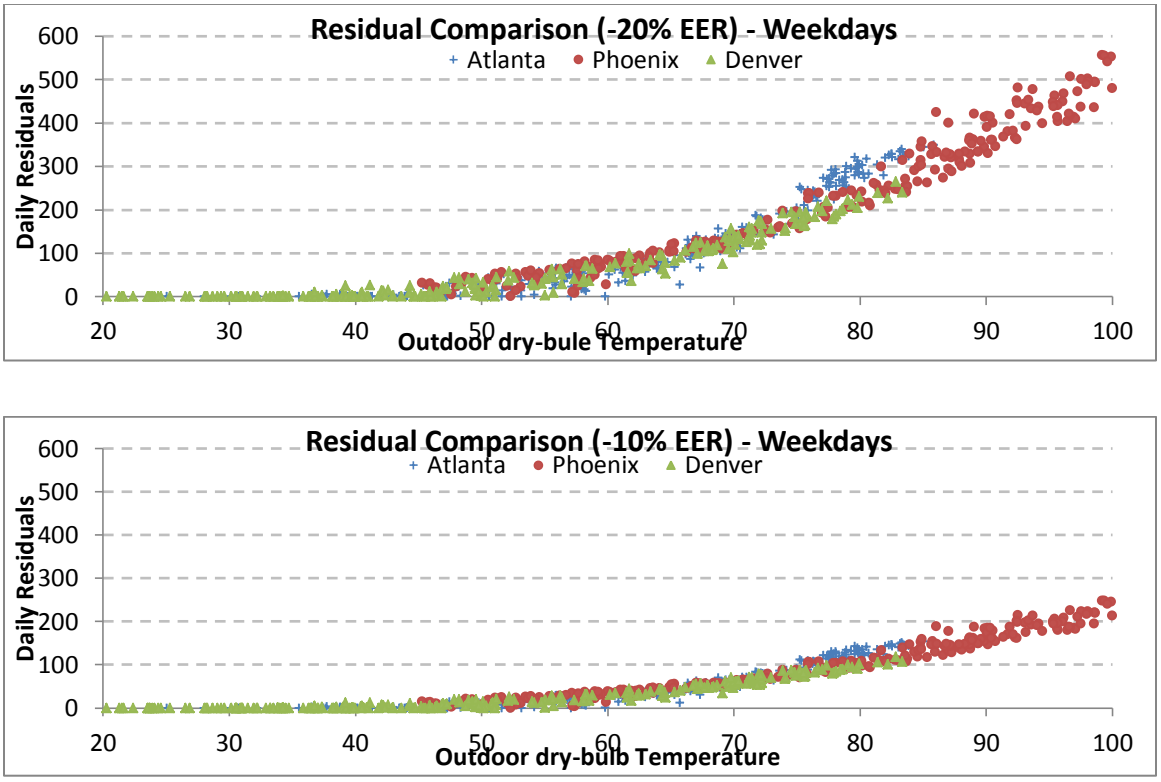


Figure 7.13(a-b). Comparative Residual Pattern Plots (EER Degradation)

Figure 7.13 (a-b) makes it clear that for the same building in different climate types, the changes in energy consumption due to this kind of fault exhibit similar patterns, i.e., the energy penalty magnitude seems to be consistent irrespective of the weather. Additionally, increases in degradation leads to similar increases in energy penalty (20% EER degradation leads to twice the energy penalty when compared to 10% EER degradation).

The analysis above can help us conclude the following:

(i) Some faults result in energy use residual patterns which are independent of climate types, whereas some others are climate-specific.

(ii) Some faults may not exhibit a deviation in the residuals for some climates.

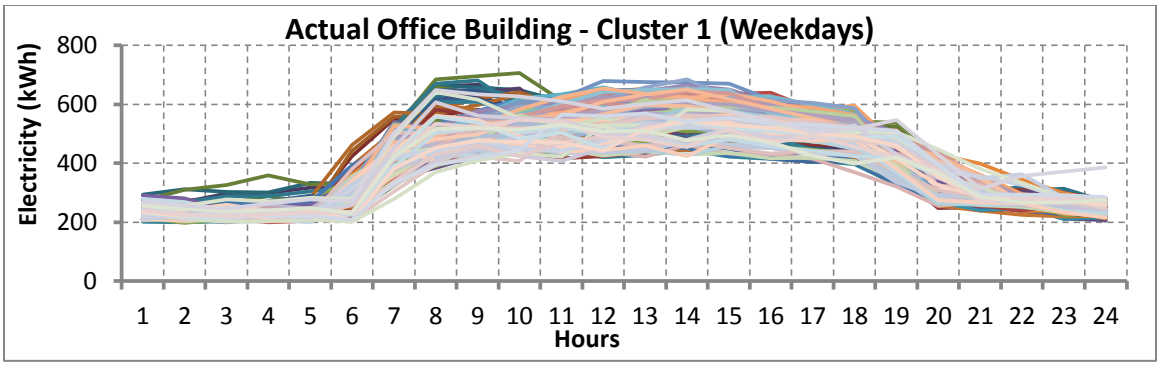
(iii) Incorrect scheduling of equipment in buildings can lead to much higher energy consumption compared to some of other faults and as such warrant the first place on the list of corrective measures.

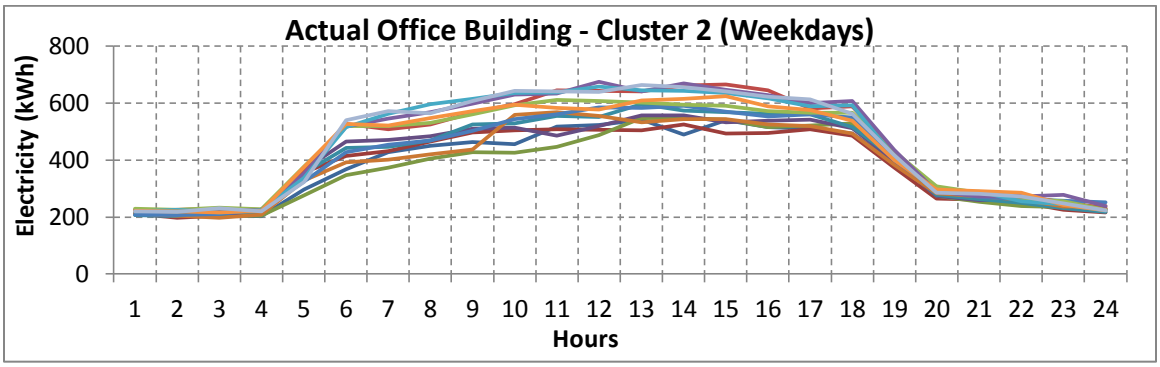
7.4 Clustering for FDD

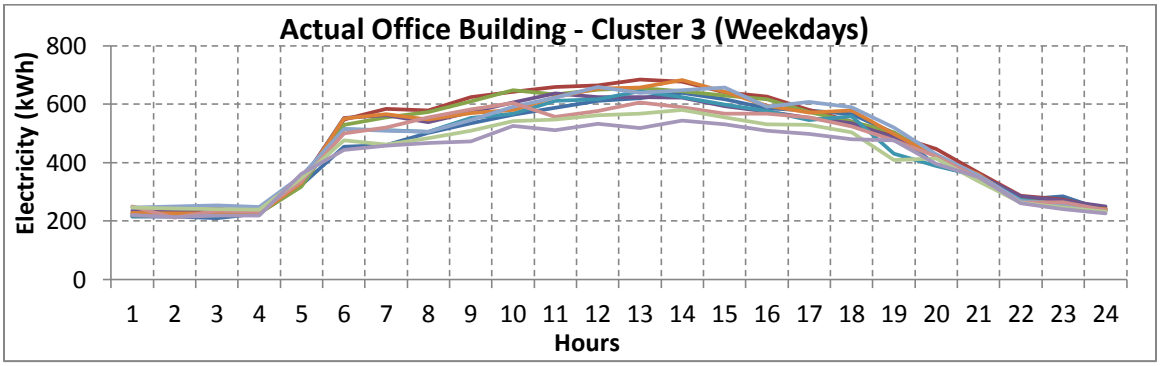
From point (iii) mentioned above and discussions about the use of clustering for FDD purposes, we go ahead and demonstrate its use for identifying changes in the hourly profiles that could be a result of improper or incorrect scheduling of a building. We applied this to the Actual Office Building described in Section 5.1.

As described in Section 5.3, multiple outcomes are possible in clustering depending upon the selected values of the two parameters *Eps* and *MinPoints*. In the following analysis, we assumed *MinPoints*-3 and tightened the *Eps*-0.39 from 0.43, resulting in more number of meaningful clusters as presented below in Figure 7.14(a-d):

(i) **Cluster 1** below contains the maximum number of day profiles from the yearly data and forms the baseline profile. This shows the building starting to operate at 6:00 AM, energy-use climbing up until 8:00 AM, going down at 6:00 PM, and finally reaching the base load during the night time operation from 8:00 PM onwards.







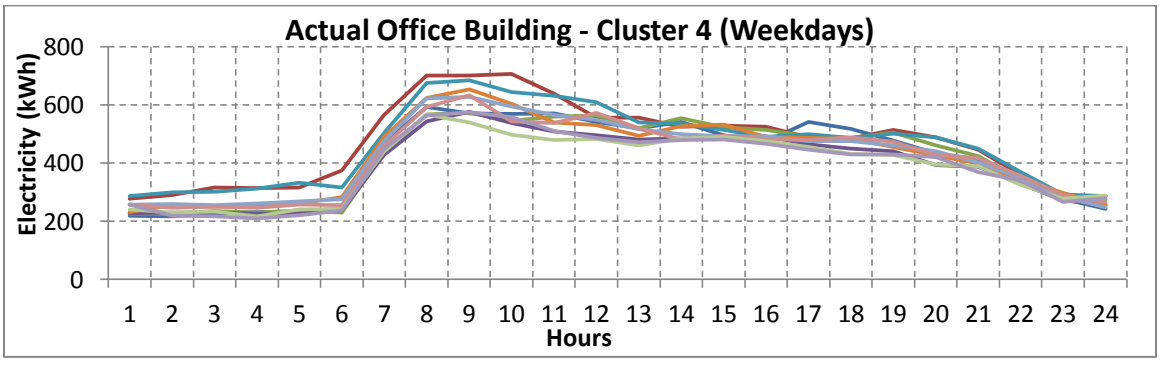


Figure 7.14(a-d). AOB - Clustering for FDD

(ii) **Cluster 2** shows the daily profiles of 13 Mondays from March up till the month of July, wherein, the building starts early at 4:00 AM and gradually climbs until 8:00AM when the building starts operations and starts declining like other baseline day profiles. This is due to the fact that since the HVAC systems are turned OFF during the weekends, the building is heated up and to account for this, it is started up early on Mondays to remove that heat and be ready for occupancy.

(iii) **Cluster 3** shows the profiles for 10 Mondays from July up till the end of September, wherein the buildings are started early as explained in point (ii) above, but stay ON until 10:00 PM in the nights before reducing to night-time base loads. This is a result of an operational change wherein, maybe certain work was planned for those Monday evenings in the building.

(iv) **Cluster 4** indicates the profiles of 10 Tuesday profiles from December uptil the end of February, wherein the building starts at regular hours of 6:00 AM but continues operating until 11:00 PM; again a result of an operational change.

Looking at the results of clustering for profile deviation purposes, we can study these differences in profiles, creating a library of rules, which can be used to train algorithms that can finally be used for profile classification purposes.

CHAPTER 8 : SUMMARY, CONCLUSIONS AND FUTURE WORK

8.1 Summary

The focus of this research was to propose and evaluate various methods by which hourly whole building electric energy consumption data, recorded by the deployment of Smart Meters, could be utilized for analyzing improvements in building energy performance due to energy conservation (ECMs) and operational and maintenance (O&Ms) measures. Additionally, data mining techniques were applied along with the more conventional inverse statistical modeling techniques. The proposed framework consists of the following parts:

(i) Visual exploration of energy interval data,

(ii) Data normalization,

(iii) Clustering, outlier detection and removal,

(iv) Occupancy generation or schedule extraction,

(v) Climate regressor data pre-processing,

(vi) Daily energy consumption prediction model,

(vii) Hourly energy consumption prediction model,

(viii) Short-term load forecasting (AR model), and

(ix) Generating condition monitoring charts.

Simple modeling techniques and model forms have been proposed. The applicability of the proposed processes within the framework has been demonstrated through their applications to two office buildings (one DOE synthetic office and another, an actual office building). Various application areas have been described earlier that include, but

are not limited to, building monitoring and verification (M&V), building commissioning, building condition monitoring and short-term load forecasting. Additional work proposed suggests the use of these techniques for fault detection and diagnosis purposes.

The model form proposed is as follows:

$$\check{E}_{i,j} = \{Daily\ Average\ Energy\ Prediction\}_i + \{Correction\ for\ Hour\ of\ the\ day\}_j + AutoRegressive\ Component$$

where, $i = 1$ to 365, index for day of the year, $j = 1$ to 24, index for hour of the day.

As mentioned earlier, by carefully using different portions of this model, one can predict daily energy consumption, hourly energy consumption and make short-term forecasts for the two building types.

8.2 Conclusions – Data Pre-processing and Clustering

Data pre-processing included preparing the data for modeling purposes. It began with visually exploring the energy profiles on monthly as well as day type basis. The monthly basis graphs help understand the variations in buildings' energy consumption as it goes through the yearly seasonal cycle. At the day type level, it is easy to visually identify profiles that deviate from the regular energy consumption profiles and need to be removed for energy prediction model identification purposes. The day type graphs also exhibit the yearly seasonal variations.

Energy profiles were normalized for clustering purposes. The clustering algorithm selected, DBSCAN, has numerous benefits such as, no prior assumptions about the data distribution; can easily cluster complex shapes, automatic separation of data points into

suitable number of clusters, and identification and presentation of noise, or outlier points. It was very clear that multiple outcomes were possible from clustering depending upon the selection of the values of the clustering algorithm criteria, i.e. *Eps* and *MinPoints*, and that optimal value of these parameters can be arrived by carefully evaluating the confusion matrices for various cases. Clustering was presented and discussed with the help of resulting confusion matrices. Once applied to both the synthetic and the actual buildings, it helped understand issues such as:

(i) Actual buildings have more number of clusters owing to variations in energy consumption behaviors, and

(ii) Actual buildings having more number of noise points which are also a result of the deviations in consumption behaviors.

The synthetic building in this study had a total of 4 clusters or day types, 2 for weekdays and 1 each for Saturdays and Sundays whereas; the actual building had 5 clusters or day types, 2 for weekdays, 1 for Saturdays and 2 for Sundays. Apart from this, synthetic building had lesser noise points (15 points) as compared to the actual office building (48 points). The noise points identified during clustering are discarded before we move to the next step of inverse statistical model identification as we would like to develop a generalized model for energy prediction purposes and these noise points would lead to biased coefficients. But, from the FDD perspective, these noise points are of great interest as they represent the profiles of days wherein there was a significant deviation in energy consumption due to operational changes or system degradation.

Finally, building scheduling or occupancy was extracted from the daily energy profiles. These were generated for all the days within each of the clusters. The fractional distributions for each of the hours for all days within a cluster were plotted and finally, the median fractional value was selected to avoid any bias due to any extreme points.

8.3 Conclusions – Inverse Statistical Modeling

Once the clustering was complete, inverse statistical models for energy prediction purposes were identified. The model form proposed has multiple applications. By carefully using different terms in the model, we can use it to predict daily and hourly energy consumption and further use it for short-term load forecasting purposes. Additionally, the identified models account for all the different day types or clusters found during clustering. The model performance was tested by observing the statistical indices of R^2 and CV-RMSE.

Daily average energy prediction models were based on the daily average values of the climatic variables such as dry-bulb temperature, humidity potential and total horizontal solar radiation. Models for both the synthetic and actual office buildings had a high model R^2 (SOB – 0.994 and AOB – 0.958) and a low RMSE and the CV-RMSE (SOB – 3.08% and AOB – 5.44%) which indicates the high prediction accuracies of the models. Additionally, these were much lesser than the CV's of the daily energy consumption distributions calculated during the clustering process, clearly reinforcing that the clustering algorithm was, in fact very robust and the clusters generated were very accurate. We can conclude that these models were very accurately capturing the effects of the multiple climatic variables including the deviations due to the change-points.

Hourly energy prediction model included an hourly correction term added to the daily average energy model. The hourly model form was based on the hourly deviations from daily average energy consumption due to the hourly deviations in climatic variables from their daily average values. Additionally, occupancy fractions generated during the data pre-processing step were used as regressors for this hourly model. Models for both the synthetic and the actual buildings had a high model R^2 (SOB – 0.948 and AOB – 0.893); the CV values were very high (SOB – 14.26% and AOB – 10.98%) as compared to the daily model. This was expected given the random variations and heat flows which assume relative importance at this finer time scale. For the synthetic building, the hourly model was over-predicting during the winter months and under-predicting during the summer months, which can be attributed to its inability to capture the thermal lag effects of the building envelope. For the actual building, the model was under-predicting during the winter months at the beginning of the year, but the predictions were accurate during the winter months at the end of the year. This suggested that there were operational changes during the beginning of the year that the model was not trained for. Also, adding to the high CV for the actual office building was its inability to capture the thermal lag effects during the summer months of the year.

Finally, AR models were proposed to model the systematic stochastic component of the residual series generated from the hourly energy prediction model. The residual series was treated as an individual time series and the correlations between subsequent observations was evaluated both at the seasonal (lags 1, 2, 3, 4, and so on) level as well as the non-seasonal (lags 24, 48, 72, 96, and so on) level by carefully observing the SAC

and the SPAC plots. The model form proposed predicted the next observation of the residual series based on the previous observations. Three model forms were evaluated:

(i) Model based on Lag 1 error term,

(ii) Model based on Lag 24 error term, and

(iii) Model based on Lag 1 and Lag 24 error term.

Out of the these three model forms, the one based on both Lag 1 and Lag 24 error terms to predict the next observation of the residual series resulted in the lowest model CV's (SOB – 7.70% and AOB – 8.79%). As such, the AR models identified were able to capture the random variations up to a certain extent, and these AR term models can be used for short-term load forecasting purposes for better demand response management.

Finally, residuals from the hourly energy prediction models were used for developing the building condition monitoring charts. As discussed earlier, there exists a mean-bias error at the monthly level since model identification was done using year-long data. The MBE correction was done at the hourly level for each of the months. The MBE correction leads to reduced error variances (25% – 35% for the SOB and 8.5% - 11.5% for the AOB depending upon the season of the year), which ultimately results in higher sensitivity of detection. These charts were plotted for both the regular and MBE corrected residuals. The charts exhibited that the residual ranges were broader during the extreme winter and summer months of January and July respectively and narrower during the weather transition months of April. This reinforces the fact that the model cannot capture the finer time scale variations leading to larger residuals.

8.4 Conclusions – Fault Detection and Diagnosis

The methodology proposed for automated FDD in this research was based upon carefully evaluating the residual patterns and attributing them to a prior known characteristic fault signature. For this purpose, a list of commonly known system faults in medium-scale office buildings was proposed and two of them (Thermostat Set point Deviations and Chiller EER Degradation) were simulated. The difference between the normally operated, simulated building energy stream and fault-induced, simulated building energy stream would produce a characteristic residual pattern representative of a specific fault. We presented the preliminary results of simulating and generating these fault signatures. After evaluating these resulting patterns, we conclude that:

(i) Different faults are relevant in different climate types, i.e. some faults may not exhibit a deviation in the residuals for some climates.

(ii) Some faults result in energy use residual patterns which are independent of climate types, whereas some others are climate-specific.

(iii) Operational changes tend to have a higher energy penalty as compared to changes in operating system parameters.

Based on our conclusion of point (iii), we finally used clustering for fault detection purposes and found that the clustering algorithm DBSCAN was able to identify and isolate different profiles by changing its two parameters *Eps* and *MinPoints*. Different clusters were identified including days with normal operation, days indicating early building startup, days indicating early building startup and late shutdown. This further reinforced the fact that multiple outcomes were possible depending upon the user intention and interpretation.

8.5 Advancements to current work

This research proposes several new techniques and methodologies based on the analysis of short-term energy interval data for purposes of enhancing building energy performance and operation. As such, there are a number of advancements possible to this work as well as parallel work to finally arrive at techniques robust enough for mass market-implementation. Some of the current advancements to this work are as follows:

(i) To begin with, the inverse statistical modeling technique should be applied to a wider sample of actual buildings, preferably different building types and from different geographical regions to generalize this approach.

(ii) This study proposes a clustering technique (DBSCAN) for day typing purposes. However, the parameters *Eps* and *MinPoints* are manually selected based on certain defined objectives such as fewer numbers of outliers and clusters. Ways to automate this would be an improvement to the current method.

(iii) In this study, the occupancy regressors for hourly energy prediction model represent combined occupancy for occupants, equipment, lighting and HVAC schedules. The data stream analyzed was the whole-building electric (WBE) available through the use of Smart Meters. Disaggregated data streams, i.e., separate energy consumption data for occupants, equipment and lighting, when available, might help improve the hourly level energy predictions of the models proposed, eventually improving the short-term load forecasting models that will result in better energy-use planning.

8.6 Future work

Referring to the master research framework drawn in Figure 4.1, there are a number of possible connections that could be explored in future studies. To begin with:

(i) In this study, we presented the preliminary results of the automated FDD methodology. This area requires a much wider and deeper attention in future studies. We presented the preliminary results of simulating two faults and the resulting residual patterns at the daily time scale. A much larger library of fault residual patterns can be prepared for training algorithms for fault classification purposes. Additionally, these patterns can be analyzed at much finer time scales, such as at the hourly level.

(ii) Once trained, energy interval data from actual buildings can be analyzed and run through the trained algorithms for fault identification and classification.

(iii) With the availability of energy consumption data at the hourly level, energy model calibration can be carried out at the daily or hourly level to see if the process deserves any merit over monthly utility bill calibration.

(iv) The schedules extraction procedure discussed in this study could be used for energy model calibration purposes to evaluate its usefulness.

(v) Once faults are classified as suggested in point (ii) above, these can be correlated to irregularly functioning system parameters and to specific times of the year when the deviations occurred. This information can further be used to improve energy model calibration.

REFERENCES

Abushakra, B., & Claridge, D. E. (2001). Accounting for the occupancy variable in inverse building energy baselining models. In Proceedings of the International Conference for Enhanced Building Operations (ICEBO). Retrieved from http://txspace.di.tamu.edu/handle/1969.1/5161

Abushakra, B., Sreshthaputra, A., Haberl, J. S., & Claridge, D. E. (2002). Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations, (ASHRAE Research Project 1093-RP, Final Report) (Energy Systems Laboratory Technical Report, ESL-TR-01/04-01,). Department of Mechanical Engineering, Texas A&M University.

Abushakra, Bass. (1999). An inverse model to predict and evaluate the energy performance of large commercial and institutional buildings. Quebec: Concordia University. In: Proceedings of the 1994 ACEEE Summer Study on Energy Efficiency in Buildings, 8, 8–49.

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. doi:10.1016/j.aei.2010.10.002

Ahmed, A., Otreba, M., Korres, N. E., Elhadi, H., & Menzel, K. (2011). Assessing the performance of naturally day-lit buildings using data mining. Advanced Engineering Informatics, 25(2), 364–379.

Alfares, H. K., & Nazeeruddin, M. (2002). Electric load forecasting: Literature survey and classification of methods. International Journal of Systems Science, 33(1), 23–34.

Baughman, M. L., Jones, J. W., & Jacob, A. (1993). A model for evaluating the economics of cool storage systems. Power Systems, IEEE Transactions on, 8(2), 716–722.

Bisgaard, S. (2011). Time series analysis and forecasting by example. Hoboken, N.J.: Wiley.

Bissantz, N., & Hagedorn, J. (2008). Data Mining. Business & Information Systems Engineering, 1(1), 118–122.

Bowerman, B. L., Koehler, A. B., & O'Connell, R. T. (2005). Forecasting, time series, and regression: an applied approach [...] [...]. Belmont, Calif. [u.a.: Thomson Brooks/Cole.

Brown, N., Wright, A. J., Shukla, A., & Stuart, G. (2010). Longitudinal analysis of energy metering data from non-domestic buildings. Building Research & Information, 38(1), 80–91.

Buchmann, E., Böhm, K., Burghardt, T., & Kessler, S. (2012). Re-identification of Smart Meter data. Personal and Ubiquitous Computing, 17(4), 653–662.

Capehart, B. L. (2004). Information Technology for Energy Managers. The Fairmont Press, Inc.

Claridge, D. E. (1998). A Perspective on Methods for Analysis of Measured Energy Data from Commercial Buildings. Journal of Solar Energy Engineering, 120(3), 150–155.

Claridge, D., Liu, M., Zhu, Y., Abbas, M., Athar, A., & Haberl, J. S. (1996). Implementation of Continuous Commissioning in the Texas LoanSTAR Program: Can you Achieve 150% of Estimated Retrofit Savings: Revisited. Proceedings of the 1996 ACEEE Summery Study, August. Retrieved from http://cgec.ucdavis.edu/ACEEE/1994-96/1996/VOL04/059.PDF

Claridge, D. E., Haberl, J. S., Sparks, R. J., López, R.E., Kissock, J. K., (1992). "Monitored Commercial Building Energy Data: Reporting the Results,"ASHRAE Transactions: Symposia, 1992, pp. 881 –889.

Cleveland, W. S. (1994). The Elements of Graphing Data (2nd ed.). Hobart Press.

Deru, M. P., & Torcellini, P. A. (2005). Procedure for Measuring and Reporting Commercial Building Energy Performance. National Renewable Energy Laboratory. Retrieved from http://www.nrel.gov/docs/fy06osti/38600.pdf

Diamond, R. C. (2001). AN OVERVIEW OF THE US BUILDING STOCK1. Retrieved from http://energy.lbl.gov/IED/pdf/LBNL-43640.pdf

Dielman, T. E. (2004). Applied Regression Analysis: A Second Course in Business and Economic Statistics (with CD-ROM and InfoTrac) (4th ed.). Brooks/Cole.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Retrieved from http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Farouz, S., Baltazar-Cervantes, J. C., Haberl, J. S., & Claridge, D. E. (2001). Monitoring and Verification Procedures Used in the Texas LoanSTAR and Rebuild America Programs.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. IEEE expert, 11(5), 20–25.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Fels, M. F. (1986). PRISM: An introduction. Energy and Buildings, 9(1-2), 5–18.

Fels, M.F., Kissock, K. Marean, M. and Reynolds C.,(1995). "PRISM (Advanced Version 1.0) Users' Guide", Center for Energy and Environmental Studies, Princeton University, Princeton, NJ, January.

Glazer, J. (2006). Evaluation of Building Energy Performance Rating Protocols. Retrieved from http://energyiq.lbl.gov/EnergyIQ/SupportPages/pdf/ASHRAE1286-FinalReport-draft11.pdf

Haberl, J. S., & Abbas, M. (1998a). Development of Graphical Indices for Viewing Building Energy Data: Part I. Journal of Solar Energy Engineering, 120(3), 156–161.

Haberl, J. S., & Abbas, M. (1998b). Development of Graphical Indices for Viewing Building Energy Data: Part II. Journal of Solar Energy Engineering, 120(3), 162–167.

Haberl, J., Sparks, R., & Culp, C. (1996). Exploring new techniques for displaying complex building energy consumption data. Energy and buildings, 24(1), 27–38.

Haberl, J., Thamilseran, S. (1996). "Predicting Hourly Building Energy Use: The Great Energy Predictor Shootout II: Measuring Retrofit Savings -- Overview and Discussion of Results", ASHRAE Transactions, Vol. 102, Pt. 2, (June).

Haberl, J., Thamilseran, S., Reddy, A., Claridge, D., O'Neal, D., Turner, D. (1998). "Baseline Calculations for Measuring and Verification of Energy and Demand Savings in a Revolving Loan Program in Texas", ASHRAE Transactions, Vol. 104, Pt. 2, (June).

Hahn, H., Meyer-Nieberg, S., & Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. European Journal of Operational Research, 199(3), 902–907.

Haves, P., Hitchcock, R. J., Gillespie, K. L., Brook, M., Shockman, C., Deringer, J. J., & Kinney, K. L. (2008). Development of a Model Specification for Performance Monitoring Systems for Commercial Buildings.

Hunn, B. D. (1996). Fundamentals of building energy dynamics (Vol. 4). MIT Press.

Jota, P. R. S., Silva, V. R. B., & Jota, F. G. (2011). Building load management using cluster and statistical analyses. International Journal of Electrical Power & Energy Systems, 33(8), 1498–1505.

Katipamula, S., Reddy, T. A., & Claridge, D. E. (1998). Multivariate Regression Modeling. Journal of Solar Energy Engineering, 120(3), 177–184.

Keith, D. M., & Krarti, M. (1999). Simplified prediction tool for peak occupancy rate in office buildings. with discussion, 28(1), 43–56.

Kissock, J.K., Claridge, D.E., Haberl, J.S. and Reddy, T.A., (1992). "Measuring Retrofit Savings for the Texas LoanSTAR Program: Preliminary Methodology and Results", Proceedings of the ASME/JSES/KSES International Solar Energy Conference, pp.299-308, Hawaii, April.

Kissock, J. K., Reddy, T. A., & Claridge, D. E. (1998). Ambient-Temperature Regression Analysis for Estimating Retrofit Savings in Commercial Buildings. Journal of Solar Energy Engineering, 120(3), 168–176.

Kissock, J. K. (1993). A methodology to measure retrofit energy savings in commercial buildings (Ph.D.). Texas A&M University, United States -- Texas. Retrieved from http://search.proquest.com.ezproxy1.lib.asu.edu/docview/304091754/abstract?accountid=4485

Liang, J., & Du, R. (2007). Model-based Fault Detection and Diagnosis of HVAC systems using Support Vector Machine method. International Journal of Refrigeration, 30(6), 1104–1114.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. Knowledge and Data Engineering, IEEE Transactions on, 5(6), 903–913.

Morbitzer, C., Strachan, P., & Simpson, C. (2004). Data mining analysis of building simulation performance data. Building Services Engineering Research and Technology, 25(3), 253–267.

Nassif, N. (2012). Regression Models for Estimating Monthly Energy Consumptions in Schools in Hot and Humid Climates. ASHRAE Transactions, 118(1), 225–232.

Piette, M. A., & Friedman, H. (2002). Data mining using HVAC diagnostic tools and EMCS data. Heating/Piping/Air Conditioning Engineering : HPAC, 30.

Piette, M. A., Gartland, L., Khalsa, S., Rumsey, P., Lock, L. E., Sebald, A., & Shockman, C. (1998). Development and testing of an information monitoring and diagnostic system for large commercial buildings. Retrieved from http://escholarship.org/uc/item/6g22x685.pdf

Price, W., Hart, R., & Water, E. (2002). Bulls-Eye Commissioning: Using Interval Data as a Diagnostic Tool. In Proceedings of the 2002 ACEEE Summer Study on

Energy Efficiency in Buildings. Retrieved from http://cgec.ucdavis.edu/ACEEE/2002/pdfs/panel03/23_295.pdf

Reddy, T. A. (2011a). Applied data analysis and modeling for energy engineers and scientists. New York: Springer.

Reddy, T. A. (2011b). Applied Data Analysis and Modeling for Energy Engineers and Scientists (2011th ed.). Springer.

Reddy, T. A., & Claridge, D. (2000). Uncertainty of "Measured" Energy Savings from Statistical Baseline Models. HVAC&R Research, 6(1), 3–20.

Reddy, T. A., & Maor, I. (2006). Procedures for Reconciling Computer-Calculated Results With Measured Energy Data.

Reddy, T. A., Saman, N. F., Claridge, D. E., Haberl, J. S., Turner, W. D., & Chalifoux, A. T. (1997a). Baselining methodology for facility-level monthly energy use-part 1: Theoretical aspects. TRANSACTIONS-AMERICAN SOCIETY OF HEATING REFRIGERATING AND AIR CONDITIONING ENGINEERS, 103, 336–347.

Reddy, T. A., Saman, N. F., Claridge, D. E., Haberl, J. S., Turner, W. D., & Chalifoux, A. T. (1997b). Baselining Methodology for Facility-Level Monthly Energy Use-Part 2: Application to Eight Army Installations. TRANSACTIONS-AMERICAN SOCIETY OF HEATING REFRIGERATING AND AIR CONDITIONING ENGINEERS, 103, 348–364.

Ruch, D., & Claridge, D. E. (1992). A Four-Parameter Change-Point Model for Predicting Energy Consumption in Commercial Buildings. Journal of Solar Energy Engineering, 114(2), 77–83.

Ruch, D.K. and Claridge, D.E., (1993). "A Development and Comparison of NAC Estimates for Linear and Change-Point Energy Models for Commercial Buildings", Energy and Buildings, Vol. 20, pp.87-95.

Saman, N., Haberl, J., Turner, D. (1998). "Overview of the Rebuild America Program in Texas", Proceedings of the Eleventh Symposium on Improving Building Systems in Hot and Humid Climates, Texas Building Energy Institute, Ft. Worth, Texas, pp.185-193, (June).

Schumann, A., Hayes, J., Pompey, P., & Verscheure, O. (2011). Adaptable Fault Identification for Smart Buildings. In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/download/3943/4216&embedded=true

Seem, J. E. (2005). Pattern recognition algorithm for determining days of the week with similar energy consumption profiles. Energy and Buildings, 37(2), 127–139.

Seem, J. E. (2007). Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy and Buildings, 39(1), 52–58.

Sonderegger, R. C. (1998). A baseline model for utility bill analysis using both weather and non-weather-related variables. Transactions-American Society of Heating Refrigerating and Air Conditioning Engineers, 104, 859–870.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining (1st ed.). Addison-Wesley.

Troncoso, R. (1997). A hybrid monitoring-modeling procedure for analyzing the performance of large central chilling plants. In Proceedings of Building Simulation (Vol. 97, pp. 421–428). Retrieved from http://www.ibpsa.org/%5Cproceedings%5CBS1997%5CBS97_P029.pdf

Tufte, E. R. (1990). Envisioning Information. Graphics Pr.

Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.). Graphics Pr.

Tukey, J. W. (1977). Exploratory Data Analysis (1st ed.). Pearson.

Wu, S., & Clements-Croome, D. (2007). Understanding the indoor environment through mining sensory data—A case study. Energy and Buildings, 39(11), 1183–1191.

Yazdani, B., Turner, D., Haberl, J., Myers, M. (2000). "The Brazos Valley Energy Conservation Coalition, Part of the Rebuild America Program in Texas: Program Update", Proceedings of the Twelfth Symposium on Improving Building Systems in Hot and Humid Climates, Texas Building Energy Institute, San Antonio, Texas, (May), pp. 207-215.

Younger, W. J. (2007). Using interval meter data for improved facility management. Energy engineering, 104(5), 21–33.

Yu, Z. (2012). Mining Hidden Knowledge from Measured Data for Improving Building Energy Performance. Concordia University. Retrieved from http://spectrum.library.concordia.ca/973713/

Yu, Z., Fung, B. C. M., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy and Buildings, 43(6), 1409–1417.