ABSTRACT

ANJIE, JIANG. Building Load Analysis and Forecasting -- A Case Study of the Building Load of the North Carolina State University Centennial Campus. (Under the direction of Dr. Ning Lu).

This research focuses on developing methodologies for building load analysis and forecasting. A set of one-year, 15-min energy consumption data, collected at a substation feeder supplying the Centennial Campus load at the NC State University, is used in the study. Detailed load analysis is first conducted to identify signatures of building energy consumption. The correlations of building load consumption with respect to the outdoor temperature, type of the day, time of the day, and humidity are studied. Key parameters that influence building energy consumptions are identified and their correlations are formulated. Then, four two-hour-ahead forecast methods are developed. The performance of each method in terms of accuracy, robustness, and efficiency is compared. Simulation results show that the similar day algorithm using the mean value of historical similar days as the forecast value produces the best forecasting results. This algorithm is also very stable and robust and its forecasting parameters can be dynamically updated using the 15-min data. The contributions of the research are 1) the building energy signatures allow building owners to identify abnormal operation scenarios and quantify the status of building operation states; 2) the load profile analysis helps building owners to decide how to optimize the building operation by shifting consumptions; and 3) the load forecasting methods predicting future operation conditions help building owner to take precaution on abnormal operation conditions to avoid economic losses and improve operational reliability.
Building Load Analysis and Forecasting -- A Case Study of the Building Load of the North Carolina State University Centennial Campus

by
Anjie Jiang

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Electrical Engineering

Raleigh, North Carolina
2014

APPROVED BY:

Dr. Ning Lu
Committee Chair

Dr. David Lubkeman

Dr. Aranya Chakrabortty
DEDICATION

To my parents and my future family.
To teachers who nourished me with knowledge.
To all the friends who have given me support and courage.
BIOGRAPHY

Anjie Jiang was born in Hefei, China. She received her Bachelor of Engineering degree in Measuring Testing Technologies and Instruments from Hefei University of Technology in June 2012. She started to pursue a Master of Science degree in North Carolina State University in August, 2012. Her research interests include load forecast and smart meter data analysis.
# TABLE OF CONTENTS

LIST OF TABLES .................................................................................................................................. vi
LIST OF FIGURES .................................................................................................................................. vii

## CHAPTER 1 Introduction ...................................................................................................................... 1

1.1 Background .................................................................................................................................. 1
1.2 Data Introduction ......................................................................................................................... 2
1.3 Modeling Methodologies ............................................................................................................ 4

## CHAPTER 2 Energy Use Signatures ...................................................................................................... 6

2.1 Identification of Energy Signatures Considering Time Factors ....................................................... 6
  2.1.1 Daily, Weekly, and Monthly Load Analysis ............................................................................. 7
  2.1.2 Yearly Load Analysis ............................................................................................................ 19

2.2 Identification of Energy Signature Considering Weather Factors .................................................. 26

## CHAPTER 3 Building Load Forecast .................................................................................................. 32

3.1 Load Forecast Algorithms .............................................................................................................. 32
  3.1.1 Algorithm 1: Direct Curve Fitting (Polynomial Regression) .................................................. 34
  3.1.2 Algorithm 2: Similar Day Approach ....................................................................................... 35
  3.1.3 Algorithm 3: Multiple Linear Regression (MLR) .................................................................. 36
  3.1.4 Algorithm 4: Multiple Linear Regression with Interactions between Predictors .............. 36
  3.2 Implementation and Performance Evaluation ............................................................................. 36
    3.2.1 Algorithm 1: Direct Curve Fitting ...................................................................................... 37
    3.2.2 Algorithm 2: Similar Day Approach .................................................................................... 45
    3.2.3 Algorithm 3: Multiple Linear Regression (MLR) .............................................................. 49
    3.2.4 Algorithm 4: Multiple Linear Regression with Interactions between Predictors ............ 51
  3.3 Curve Fitting Quality and Forecast Accuracy .............................................................................. 54
  3.4 Conclusions ............................................................................................................................... 55

## CHAPTER 4 Conclusion and Summary ............................................................................................... 57
References...................................................................................................................58
Appendix.....................................................................................................................62
LIST OF TABLES

Table 1-1 Data Description ......................................................................................................... 4
Table 3-1 Comparison of Four Algorithms ..............................................................................33
Table 3-2 MAPEs of Algorithm 1 .................................................................................................44
Table 3-3 MAPEs of Algorithm 2 .................................................................................................47
Table 3-4 Counts of Each Date (Forecast without Type Filter) ...................................................48
Table 3-5 Counts of Each Date (Forecast with Type Filter) ......................................................48
Table 3-6 MAPEs of Algorithm 3 .................................................................................................51
Table 3-7 MAPEs of Algorithm 4 .................................................................................................54
Table 3-8 R-squared and MAPE comparison of Algorithm 1 .....................................................55
LIST OF FIGURES

Figure 1-1 Map of NCSU Centennial Campus ................................................................. 3
Figure 2-1 Plots of Maximum/Minimum/Average Loads of April 2013............................... 8
Figure 2-2 Load Profile of Weekdays/Holidays of April 2013 ........................................... 9
Figure 2-3 Load in Different Day of Week ........................................................................ 10
Figure 2-4 Plot of Daily Consumption .............................................................................. 11
Figure 2-5 Boxplot of Load Profile of April 2013 ............................................................... 13
Figure 2-6 Boxplot of Load Profile of March 2013 ........................................................... 14
Figure 2-7 Load Ramp Rates of April 2013 ..................................................................... 15
Figure 2-8 FFT signatures for each month ...................................................................... 16
Figure 2-9 Load Profile of Year (From May 2012 to April 2013) ....................................... 19
Figure 2-10 Load in Different Seasons .............................................................................. 20
Figure 2-11 Daily Maximum/Minimum/Average loads (From May 2012 to April 2013) ....... 21
Figure 2-12 Daily energy consumptions of Weekends/holidays .......................................... 22
Figure 2-13 Monthly load duration curves ....................................................................... 23
Figure 2-14 Yearly load duration curve ............................................................................. 24
Figure 2-15 Peak load hours categorized by month of occurrence ..................................... 24
Figure 2-16 Distribution of peak load hours (From May 2012 to April 2013) ..................... 26
Figure 2-17 Temperature and Humidity of the Year .......................................................... 27
Figure 2-18 Correlation of Temperature and Load ............................................................. 28
Figure 2-19 Average Load of Each Temperature Range ...................................................... 29
Figure 2-20 Hour and Day Distribution When Temperature Is Very High or Low ............. 30
Figure 2-21 Load Envelop (Typical Curves of Different Temperature) .............................. 31
Figure 2-22 Correlation between Humidity and Load ...................................................... 31
Figure 3-1 Algorithm 1 Direct Fitting Demonstration ......................................................... 35
Figure 3-2 Comparison of Algorithm 1 using 5 Direct Fitting Methods (March 12\textsuperscript{th} 2013, 10AM) .................................................................................................................. 39
Figure 3-3 MAPEs of Algorithm 1 with 4 Direct Fitting Methods on March 2013 (Black:
Figure 3-4 Comparison of Algorithm 1 with Different Data Points on March 12th 2013 10AM .................................................................40
Figure 3-5 MAPEs of Algorithm 1 with Different Data Points on March 2013 (Black: MAPE Pink: Max APE) ..................................................................................42
Figure 3-6 General MAPEs of Algorithm 1 Linear Regression and Quadratic Regression (Black: MAPE Pink: Max APE) ........................................................................43
Figure 3-7 Comparison of MAPE of Algorithm 2 Whether to Add Type Filter on March 12th 2013 10AM ..................................................................................................44
Figure 3-8 MAPEs of Algorithm 2 Whether to Add Type Filter on March 2013 ................45
Figure 3-9 General MAPEs of Algorithm 2 Without/With Type Filter (Black: MAPE Pink: Max APE) .......................................................................................................46
Figure 3-10 Dates Which has a Forecast Error over 15% with Algorithm 2 ....................47
Figure 3-11 MAPEs of Algorithm 3 With/Without Type or Hour Factor March 2013 (Black: MAPE Pink: Max APE) ..........................................................................................50
Figure 3-12 General MAPEs of Algorithm 3 (Black: MAPE Pink: Max APE) .............51
Figure 3-13 MAPEs of Algorithm 4 With/Without Type or Hour Factor March 2013 (Black: MAPE Pink: Max APE) ..........................................................................................52
Figure 3-14 MAPEs of Algorithm 4 without/with Type Filter (Black: MAPE Pink: Max APE) ..........................................................................................53
Figure 3-15 General MAPEs of Algorithm 4 (Black: MAPE Pink: Max APE) .............54
Figure 3-16 MAPE of Each Algorithm ........................................................................56
CHAPTER 1  Introduction

1.1  Background

Building energy consumption data usually is collected by building energy management systems (BEMS) ([1][9]). Traditionally, energy meters record building electricity usage and are read at the end of each month to record the monthly energy consumption [2]. With the installation of smart meters and sensors, data of all kinds (humidity, temperature, and occupants schedules) with far better resolution (e.g. minute-by-minute energy consumptions) can be recorded and stalled. The BEMS dataset can be used in many ways:

1) Diagnosing building energy use for fault detection [1]. Typical energy use patterns and amounts can be identified as energy signatures for buildings. When real-time measurements deviate significantly from the signatures, building owners and operators can investigate to make sure it is not due to system faults or incorrect settings. Thus, the reliability of the building energy system will be enhanced.

2) Participating demand response programs [3]. Demand Response (DR) is defined as: ‘Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized.’[1]. An analysis of historical building energy use data can reveal load composition and the potential of the building to participate demand response programs. Real-time BEMS data can identify the states each end use are at and help the building owners to make decisions on which end use is suitable for responding to demand response signals.

3) Improving energy efficiency ([4][5]). Basic characteristics of the energy consumption can be identified using data mining or artificial intelligence ([27][30]). Then, hourly-ahead, day-ahead, or even longer term load forecast of building energy consumptions can be obtained based on customer’s requirements [6]. Accurate load forecast allows optimal energy purchasing schedules to be made to optimize energy payment and energy efficiency [7].

4) Integrating renewable resources [35]. Rooftop photovoltaic (PV) systems and small wind turbines can be installed to supply building loads. The BEMS data can help the building owners to determine the size of the renewable generation resources and the operation needs
for auxiliary devices such as whether or not additional energy storage system is needed.

At the beginning of the project, we have contacted building managers at the North Carolina State University. The building loads at the Centennial campus are increasing steadily every year. A distribution system upgrade plan is needed to expand the circuit capacity and a co-gen system or distributed renewable generation system (primarily PVs) may be needed to supply the growing load. The building managers are also considering demand response programs for lowering energy bills. Therefore, they have extremely high interests in identifying building energy use patterns and short-term and mid-term load forecasting for predicting the load increases. As a first step, we were requested to focus on the following four tasks:

- Identify the building energy consumption patterns
- Develop a systematic process to identify building energy signatures
- Derive the parameters that have strong correlations with building energy consumptions
- Develop short term load forecasting methods to improve building operation

A set of one-year, 15-min energy consumption data, collected at a substation feeder supplying the Centennial Campus load at the NC State University, is provided by the building managers. Weather data such as temperature and humidity are also provided.

1.2 Data Introduction

Centennial Campus is a research park and educational campus owned and operated by North Carolina State University (NCSU) in Raleigh, North Carolina, USA. Composed of two locations, the 1,334 acres (5.4 km²) property provides office and lab space for corporate, governmental and not-for-profit entities, in addition to providing space for 75 university research centers, institutes, laboratories and departmental units. Currently, 2,700,000 sq ft (250,000 m²) of constructed space has been built. Upon completion, Centennial Campus is anticipated to have 9,000,000 sq ft (840,000 m²) of constructed space (http://en.wikipedia.org/wiki/Centennial_Campus_of_North_Carolina_State_University).

The electric demand of Centennial campus includes heating, ventilation, air conditioning (HVAC), lighting, office equipment (e.g. laptops, printers, desktops, etc.), lab equipment (different motor loads), computer servers, etc. Among those demand sources, HVAC units are important contributors to total energy consumption in summer because in Raleigh area,
both the temperature and humidity can be very high for days in summer time. The electric demands are varied by seasons, weathers, days of the week, and special days such as national holidays, breaks, and long vacations. Since the data used in this thesis were obtained within Centennial Campus of NCSU, it is greatly related to the behaviors and schedules of the building occupants, faculties and students.

The data set we use for this thesis comes from a feeder which supplies five building loads. The data was recorded every 15 minutes from May 2012 – April 2013. As highlighted in Figure 1-1, the five buildings are EB1, EB2, EB3, James. B. Hunt Library, and the Monteith Engineering Research Center. Also, we have the hourly data of the temperature and humidity. The data details are shown in Table 1-1.

Figure 1-1 Map of NCSU Centennial Campus
### Table 1-1 Data Description

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Resolution</th>
<th>Time Range</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load(kW)</td>
<td>15 mins</td>
<td>May 2012 – April 2013</td>
<td>14865.12</td>
<td>3499.20</td>
</tr>
<tr>
<td>Temperature(°C)</td>
<td>1 hour</td>
<td>40.6</td>
<td>40.6</td>
<td>-7.2</td>
</tr>
<tr>
<td>Humidity(%)</td>
<td>1 hour</td>
<td>100</td>
<td>100</td>
<td>16</td>
</tr>
</tbody>
</table>

1.3 Modeling Methodologies

Statistical analysis and data mining techniques are used to identify and summarize the key features of building energy consumption. Data mining ([3][16][27]) is the process of analyzing data from different views and summarizing it into useful information. In this thesis, 15-minute load data of a year (May 2012-April 2013) was obtained and analyzed. From the yearly perspective, the electric demand is varying by seasons. Since NCSU is located at Raleigh in North Carolina, the southeast part of United States, it has a relatively warm and mild climate, which gives rise to the circumstance that the electric demand of air conditioning in summer is much higher than in winter. From the monthly perspective, the demand is not only driven by seasons, but also by the academic calendar schedule [2]. Commonly the load is much lower in consecutive holidays, such as winter vacation, summer vacation and spring break, than in regular days, and higher when it is in exam week. And from the daily perspective, the demand differs from weekday and weekend, and influenced by students’ and faculty’s working behavior. Generally, the loads are categorized as laboratory load, library load and classroom load. Signatures of load are obtained and analyzed in Chapter 2.

Load forecast ([2][20][24]) can be classified into 3 categories depending on the forecasting duration: long-term forecast which is longer than a year, mid-term forecast which is usually from one week to one year, and short-term forecast which is usually from one hour to one week. The forecasts for different time horizons can meet different needs in the electrical energy industry. For example, mid-term and long-term forecast are normally used to

---

1 This temperature indicates the Dry Bulb Temperature (DBT), which means the temperature of air measured by a thermometer freely exposed to the air but shielded from radiation and moisture.
schedule fuel storage and determine the placement of new generators. Short-term forecast ([11][13]) is mainly used in real-time power system operation for making better energy transactions plans, security analysis, automatic generation control, hydro-thermal coordination, economic dispatch, managing load and making outage plans.

The forecasting techniques are categorized as statistical methods or artificial intelligence algorithms, such as regression, neural networks, and fuzzy logic, etc. Among those methods, End-use and Econometric approach is widely used in Mid-term and long-term forecasting. And similar day approach, neural networks, statistical learning algorithms, fuzzy logic and various regression models have been developed for short-term forecasting [8]. The load forecasting that the thesis focuses on is short-term load forecasting. Among those methods mentioned above, 3 kinds of regression and similar day approach will be discussed and compared in Chapter 3.
CHAPTER 2 Energy Use Signatures

This Chapter presents the process and methodologies used for deriving energy use signatures.

Building energy consumptions are determined by the following main factors ([1][14]): time, weather, occupants’ schedules and behaviors, electricity prices, and special events. Building energy consumptions have seasonal, weekly, holiday and daily patterns. In general, seasonal effect reflects in the number of daylight hours in one season. The schedules for building lighting systems (e.g. work hours and off-work hours) and the modes of the HVAC systems (heating/cooling) are changed seasonally. In addition, loads in weekdays usually will be higher than that of the weekends and holidays. In a day, the loading levels of off-work hours will be less than that of the office hours. Among weather related factors, temperature and humidity have greater influences on load patterns. The temperature change will affect the amount of power needed for heating in winter and air conditioning in summer. Other weather factors, such as humidity, precipitation, thunderstorms, and the wind and light intensity of the day, will influence the building energy consumptions by increasing the lighting and HVAC system loading. Random disturbances by special events such as exam days, public activities, blackouts, etc. will also result in abnormal energy consumption patterns.

In this research, we focus on identifying the correlation of energy consumptions of buildings on university campus with time and weather factors. The main methodology we used is statistical analysis and data mining. The goal is to extract information from the data set and derive energy use signatures and patterns to be used in fault diagnosis [29] and building load forecast [28]. Compared with other existing building energy use studies, our data sets are complete and have high resolution and quality. The 15-min data collected has no missing data and is aligned with a whole set of weather data such as solar radiation, temperature, wind, and humidity. North Carolina State university calendar can be used for identify different types of days. Our main focus has been placed on identifying energy signatures that have not been fully studied in the past: hourly, daily, and weekly energy consumption and power variation signatures considering different types of day and weather conditions.

2.1 Identification of Energy Signatures Considering Time Factors

North Carolina State University is an educational institution that has clearly outlined class schedules and holiday arrangements. Therefore, the occupant schedules play a vital role in
load patterns ([12][13]). In this section, different kinds of load patterns and energy consumption signatures are discussed. The analysis is categorized two parts: daily, weekly, monthly (short-term) analysis and yearly (long-term) analysis. This is because short-term analysis reveals more detailed information about daily load patterns and weekday/weekend differences will clearly be identified; while yearly analysis yields seasonal load changes and the maximum/minimum load of the year will guide the capacity selection for different equipment in a distribution system.

2.1.1 Daily, Weekly, and Monthly Load Analysis

The goal of daily-weekly-monthly analysis is to identify the maximum and minimum loads of the year and the durations of each loading levels. Making circuit expansion plans needs to consider different options that may result in either expanding circuit capacity or reducing peak loads. Therefore, it is critical for the analysis to reveal information such as to what extent, the peak load can be reduced for how long so that adding new transformers or co-gens are not needed or what the capacity of the co-gen is to supply the increased peak load.

In general, monthly building load analysis shows building owners the general energy consumption features in a period that weather conditions are similar. The loads changes periodically following a daily pattern and a weekly pattern. The data of April 2013 are given as an example to interpret the monthly analysis results we obtained. Seven Energy Signatures are derived and results are summarized as follows.

**Energy Signature 1: Maximum, minimum, and average loads of a day**

The maximum, minimum, and average loads, \( P_{\text{max}} \), \( P_{\text{min}} \) and \( P_{\text{ave}} \), of a day is calculated by

\[
P_{\text{max}} = \max(P_i) \quad P_{\text{min}} = \min(P_i) \quad P_{\text{ave}} = \frac{\sum_{j=1}^{24 \times 4} P_i(j)}{24 \times 4}
\]

(2-1)

where \( i \) is the \( i^{th} \) month and \( j \) is the \( j^{th} \) data point in a month. The black dashed line in Figure 2-1 is the original load and the red, blue, and cyan lines in Figure 2-1 indicate the maximum,
average, and minimum load of each day. As shown in Figure 2-1, the building load profiles variations follow daily and weekly patterns. This analysis yields an upper and lower envelope for energy consumptions, which if violated, show abnormal energy consumptions.

![Graph showing load profile variations](image)

**Energy Signature 2: Shapes of Daily Load Profile**

Figure 2-2 show daily load profiles by type of days in April 2013. The reason we picked the month of April is because there are days that the air conditioning (a/c) loads are off or on [33]. By observation, there are four types of load profiles: weekday with a/c loads, weekday without a/c loads, weekend with a/c loads and weekend without a/c loads. Because in holidays, students, staffs, and faculties have special working schedules and different activities, the building energy consumptions can be very different. Therefore, we divide the days into two types: weekdays and holidays. Figure 2-2 clearly shows that there are two kinds of load curves, which are separately represented in Figure 2-2(b) and (c). The figure of weekdays (Figure 2-2 (b)) has a “hat” shape, showing that the load increases rapidly at around 8am and decreases when classes end at around 6pm. During holidays (see Figure 2-2(c)), loads do not have a significant peak but a low load period is observed between 6am
to 9am due to the lighting schedule changes.

Weekday with/without air conditioning loads: It can be inferred from the figure that the load changes follow a similar trend each day. The load valley period is between 0-7 am. A sharp load increase is observed between 7-11 a.m. The load peak is flat and duration is approximately six hours (11 a.m. -17 p.m.). Then, the load decreases sharply between 17
p.m. to 19 p.m. and the load continues to decrease slowly after 19 p.m. until midnight. As we observed in Figure 2-4, for weekday loads, there can be 3-4 MW differences between two load curves for with and without a/c loads. Note that the maximum load happens around 3pm. The pink dots in the figure show the lowest consumption of weekdays and can be used as an approximate base line of weekday load profile.

*Weekend with/without air conditioning loads:* the load and stable load curve in the other time. Since the figure demonstrates very different curve between the highest and lowest day, a problem of why there is such a difference is raised. The pink dots in the figure show the lowest consumption of weekends or holidays and can be used as an approximate base line of weekend load profile.

Figure 2-3 shows the average load of different day of the week. It shows significant difference between weekday and weekends. Also, there are slight difference between weekdays/weekends, which can reveal the behavior of faculty and students. For weekends, Saturdays have a slightly higher load than Sundays, but the load of Sunday evening starts to increase since the following day will be a weekday. For weekdays, Mondays have lowest morning load but highest evening load, while Tuesdays and Wednesdays have the highest load. In NCSU, most classes are scheduled on Monday through Thursday, and professors
have preference to set the exams and quiz in the middle of the week. Apparently, load of Friday evening is relatively low for it is near weekend.

**Energy Signature 3: Daily Energy Consumption**

The daily energy consumptions are also very important signatures to quantify the customer’s electricity usage. Because 15-minute power consumption $P_i(j)$ is available, the daily energy consumption $E_{\text{daily}}^i$ of the $i^{th}$ day can be calculated by

$$E_{\text{daily}}^i = \frac{1}{4} \sum_{j=1}^{4*24} P_i(j)$$  \hspace{1cm} (2-2)

Figure 2-4(a) of daily consumption data demonstrates that the energy consumption changes following a weekly cycle. The daily consumption data are divided into two groups: weekdays and weekends/holidays, as shown in Figure 2-4(b).
The results show that:

- The weekday consumptions are significantly higher than weekends and holidays. The differences of the energy consumptions between weekdays and weekends show the additional energy consumed when classes are held in teaching facilities.
- The weekday energy consumptions show changes while the holiday consumptions are fairly steady. This is because during weekends, there are little activities in buildings and the a/c has a set-back schedule so it is less sensitive to outdoor temperature changes.

**Energy Signature 4: 15-minute Energy Consumption**

The 15-minute energy consumptions are signatures that reveal peak load conditions. Because 15-minute load \( P_i(j) \) is available, the 15-minute energy consumption can be calculated by

\[
E_{15-min}^i = \frac{1}{4} P_i(j)
\]

As in Figure 2-5(a), where 96 small boxplots are shown, the power consumptions at the same time period but different days have different distributions. There are no outliers in the figure, which means there are no extraordinary days when the power consumptions are abnormal in April. Figure 2-5(b), which is boxplot of the weekday and holiday load, is very useful to identify outliers for any specific hour within a month. Therefore, it can be used as a power signature for building operation.
Using March 2013 data as an example, compared to Figure 2-5, the boxplot of load of March 2013 (Figure 2-6) has many outliers. Those outliers are mainly above average and mostly happening in the off-work hours. Furthermore, when they are divided into weekday and holiday boxplots, more detailed information are revealed.

The weekday boxplot shows both above-average outliers and below-average outliers during working hours. In addition, the interquartile ranges (IQR)s\(^2\) [18] of weekday boxplot are small, which means the daily consumptions are similar if the outliers are excluded. When looking into the data sets, we find that all the outliers above the average came from one specific period when the outdoor temperatures are exceptionally high and a/c units need to be run at cooling modes, resulting in an energy consumption jump.

In the holiday boxplot, the outliers mostly happen during off-work hours. In this case, a conclusion can be drawn as follows: there are one or two particular days that the energy consumption is obviously higher than the other days, and these days have a high possibility that they are weekdays. Actually, by looking into the data the particular day can be found out to be March 12th, which was the day just after the spring break (See Appendix A). The

\(^2\) IQR(Interquartile Range) is defined as: IQR = Q3 – Q1, where Q1 is 25th percentile data, which means 25% of observations are below it, and Q3 is 75th percentile data, which means 75% of observations are below it. Thus the IQR is the range of the 50% observation in the middle.
reason might be that students and faculty have returned to school to catch up with studies and works after the spring break.

![Boxplot of Consumption of March 2013](image)

![Boxplot of Weekday Consumption of March 2013](image)

![Boxplot of Weekend/Holiday Consumption of March 2013](image)

**Figure 2-6 Boxplot of Load Profile of March 2013**

**Energy Signature 5: Power Ramps**

Power ramps, $R$, are defined as the power changes between the $j^{th}$ point and the $(j-1)^{th}$ point.

$$ R = P(j) - P(j-1) $$ (2-4)

The power ramp shows how fast the energy consumption changes in a period of 15-min. Fast change of power can result in power outages if not anticipated. Therefore, the power ramps are considered as a signature for energy changes. The plots of ramp rates (Figure 2-7) express that most of the ramp rates are among -1 to 1. And the average ramp rate of weekdays is higher than the average rate of holidays, which leads to a conclusion of weekday power consumption changes more than holiday power consumption. Besides, based on the ramp rate range, the holiday power ramp rates have more outliers.
Energy Signature 6: Frequency Domain Characteristics

As building load follows certain cycles, a frequency domain analysis (See Figure 2-8) can yield a unique signature that reflects the cyclic loads [4]. Because natural loads have diversified frequency domain signatures, it is hard to reproduce a natural load profile that matches them. This signature therefore is very important to identify man inserted building load profiles. As pointed out in [4], this signature can be used to identify energy thefts or any abnormal periodically use of electricity. An observation of the FFT signatures of the building loads in Centennial Campus show that there are loads cycling at 6, 12, and 24 hours. The limitation of the FFT analysis is the data resolution. Loads cycling smaller than 30 minutes cannot be observed.
Figure 2-8 FFT signatures for each month
2.1.2 Yearly Load Analysis

A yearly load analysis reveals seasonal load variations because of different occupant schedules and weather changes. The energy signatures derived from yearly load analysis are going to be used for planning of circuit expansion, assessment of overall capacity limits and needs, and evaluation of different energy efficiency plans. Traditionally building load analysis is based on yearly load analysis using monthly energy consumptions [12]. The methodologies are well defined. In this section, we will discuss what our findings are and suggest a few new discoveries that can serve as energy signatures.

As shown in Figure 2-9, a few facts can be quickly observed from yearly data profile:

- Outages. There are a few significant outliers in September which clearly indicate loss of supplies for several short periods of time.
- Major holidays. There is also an extremely low low-load period at the end of December corresponding to the Christmas break and winter holiday.
- Annual peak load. Summer load is much higher than the other seasons because in Raleigh area, which is at the southeast part of USA, summers are hot and humid and winters are rather mild. In addition, a lot of spacing heating units in the area use gas as fuel.
Figure 2-10 Load in Different Seasons shows the load in different seasons. Obviously, the summer has the highest load, which is almost twice as the load of spring and fall, while winter has the lowest load. The load of different seasons is positively related to the temperature.

More detailed analysis on yearly data shows that if we combine the energy signatures derived in daily-weekly-monthly analysis, yearly data can let us identify a few more very useful energy signatures.

**Energy Signature 7: Load Composition**

In Figure 2-11, the maximum, average, minimum daily load curves are plotted for the whole year. As shown in the figure, the building load, $P_{Baseline} = 6.8$ MW, is at the lowest during the Christmas break. This reflects the lowest possible energy consumption because in this time period, all buildings on the Centennials Campus were running at their minimum energy consumption levels. Thus, using this baseline load, one can easily derive the baseline load
for the weekend non-mission critical loads (0.5 MW: including lighting, dominant computer energy consumption, non-mission critical lighting, etc.), $P_{\text{Wkend NonCritical operation}}$, occupant activity loads (0.5 MW: including active computer loads, lab equipment, lighting loads, etc.), $P_{\text{Wkend Activity}}$, weekend a/c loads (up to 5 MW), $P_{\text{Wkend Air Conditioning}}$. Note that by simply observe the differences between the base load profile and other weekend load profiles (see Figure 2-12), one can easily desegregate the building loads and estimate the composition of the a/c loads. This is very important for controlling slow voltage recovery that is caused by the stalling of a/c motors [21].

![Figure 2-11 Daily Maximum/Minimum/Average loads (From May 2012 to April 2013)]
Figure 2-12 Daily energy consumptions of Weekends/holidays
Energy Signature 8: Peak Load and Peak Load Durations

Monthly sorted consumption curves are used to compare energy consumptions for each month. As shown in Figure 2-13, 12 sorted consumption curves clearly mark the maximum and minimum load levels of this feeder and provide a straightforward view of peak load durations. It is clear that the campus would have more electricity demand in summer time, especially in July. November, December and January have least consumption. The sorted yearly load profile (Figure 2-14) shows the peak load magnitudes and durations. Figure 2-15 further shows in which months, those peak loads occurred and for how long. The signature can help building operators to determine the peak shaving programs. For example, if the peak load needs to be lower than 14 MW, we can observe from Figure 2-15 that 233 data points are higher than 14 MW, which is about 58 hours and occurred in Months June-Sept.
Figure 2-14 Yearly load duration curve

Figure 2-15 Peak load hours categorized by month of occurrence
**Energy Signature 9: Distribution of Peak Load Hours**

The hour that daily peak power occurs at is another key parameter that reflects the characteristics of building energy consumption. As demonstrated in the distribution histogram (Figure 2-16 (a)), the max load often happens around 3 p.m. in the afternoon, mostly happen between 10AM to 8PM. However, for rare occasions, peak load can occur at the midnight. Looking into the data, those outliers are Sundays with high outdoor temperatures. Compared with regular weekends, Sunday nights are the nights that students return to school to finish homework assignments and get ready for the coming week.

As shown in Figure 2-16, 2-4 p.m. are the most likely peak-load hours in a day. Therefore, any peak shaving and load shifting programs should target those hours to be the most effective. The histograms of weekdays and holidays (Figure 2-16 (b)) give more detailed information. It can be seen in the figure that the weekdays has a relatively regular schedule and peak time is almost concentrated around 3 p.m. But the time interval of peak load is more scattered in holidays, and has more outliers occurring at 12AM, mainly during Sunday nights. It can be inferred that weekdays have relatively regular demand patterns, while holidays have electricity usage patterns that are more difficult to detect. Part of the reason is normally after weekend break, students are working till late night on Sundays to catch up with their assignments. This figure can be used to detect whether there’s an unusual situation, like extreme weather or specific public events, if the peak load happens at an unexpected time point.
2.2 Identification of Energy Signature Considering Weather Factors

Among different weather factors, such as solar radiations, wind speeds, humidity, and temperatures, building energy consumptions are strongly correlated with temperature. This is mainly because of the a/c load is a significant portion of building loads. Humidity has a secondary impact on building loads because a/c units are derated in high humidity conditions [15]. Because Raleigh is located on the southeast part of USA, it has long humid and hot summer, we will focus our studies on the temperature and humidity in this study. In Figure 2-17, the temperature and humidity of the study year is plotted.
In statistics [18][34], correlations are to evaluate whether or not and how strongly pairs of variables are related. To show the temperature and load correlations (see Figure 2-18), Pearson's product moment coefficient is calculated by

\[ \rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \]

\[ \mu_X = E(X) \quad \sigma_X^2 = E(X^2) - E^2(X) \]

\[ \rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \]

where \( E \) is the expected value operator, \( \text{cov} \) means covariance, and \( \text{corr} \) a widely used alternative notation for the correlation coefficient. In the case of NCSU centennial campus, \( \rho_{X,Y} = 0.8654 \).
Regression [34] methods are used to fit a linear model. R-squared is a parameter indicating how well data points fit a statistical model. It is defined by

\[ R^2 = 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2} \]  \hspace{1cm} (2-6)

Where \( y_i \) is the observed values, \( f_i \) is the modelled values, \( \bar{y} \) is the mean value of \( y_i \).

In this case, R-squared equals to 0.749, which is close to one. This result indicates that the temperature and load are strongly linear related with each other.

In Figure 2-19, the average load of each temperature range is plotted. This figure indicated that the load and temperature are roughly positively related, which means as the temperature increases, the load increases. Intuitively, when the temperature is very high or very low, the air conditioner or heating system would consume much more energy. But it can be seen in the figure that there is no significant rise when the temperature is below 0 °C or over 35 °C. On the contrary, the average load decreases when the temperature is more than 35°C. Thus,
to discover the reason of this circumstance, the hours and the days having temperatures ranging below 0 °C or exceeding 35 °C are plotted in Figure 2-20.

From Figure 2-20 (a), the hour distribution shows that the cold weather happens at night and early morning, most of which are in off-work hours. In Figure 2-20(b), the day distribution shows that the hot weather happens mostly on holidays. It gives rise to the little increase in load when the temperature reaches low or high, for the extreme temperatures happened to occur at off-work hours.

Figure 2-19 Average Load of Each Temperature Range
Energy Signature 10: Typical Load Profiles for Signature Temperature Ranges

This energy signature shows an expected load profile for weekend and weekdays for a range of temperatures at 0°C, 10°C, 20°C, and 30°C. This gives the building operator a typical building load profile for estimating energy consumptions without much information needed in a corresponding temperature range, as shown in Figure 2-21. The load difference between 10 and 20 is significantly more than the difference between 0 and 10. This is because the cooling mode are triggered normally at 17°C and when air conditioning is running in cooling mode, the building energy consumption will jump up 4 MW.

Note that humidity does not show strong correlations with load unless we categorize the data points into groups in different temperature ranges (Figure 2-22). Because a day with high humidity usually is a cloudy, rainy, or snowy day, the temperature of the day will usually be mild. Extremely cold or hot days are usually with low humidity. In a high humidity day, the a/c compressor will be derated so that more energy will be consumed. Therefore, if outdoor temperature is the same, then in a humid day, the building energy consumption is normally higher, as shown in Figure 2-22. If the temperature is lower than the thermostat setting for cooling in a building, then the energy consumption tends to have no obvious correlation with humidity. Therefore, a refined Energy signature will include the humidity to adjust the typical load profiles between 20°C, and 30°C.
Figure 2-21 Load Envelop (Typical Curves of Different Temperature)

Figure 2-22 Correlation between Humidity and Load (On different Temp)
CHAPTER 3 Building Load Forecast

Building Load Forecast ([19][20]) is an area that has been well understood in terms of mid- and long-term load forecast. Therefore, this research focuses only on very short-term building load forecasts that only made possible when high resolution building energy consumption data (15-min or less) is available. The goals of very short-term building load forecast are [10]:

- identify adverse operation conditions (overloading, high load ramps, etc.)
- detect failures and faults
- optimize of energy payments
- identify load characteristics

Although the load forecast methodologies are developed for two-hour-ahead load forecast, those methods can easily be extended to longer time periods. Four load forecasting algorithms and models are developed and the performances of each algorithm are evaluated based on the forecast accuracy. A load forecast engine is built based on the four methods as a product for the building operators to use in their operation.

3.1 Load Forecast Algorithms

In this chapter, four load forecast algorithms are introduced: direct curve fitting (Algorithm 1), similar day approach (Algorithm 2), multiple linear regression (Algorithm 3) and multiple linear regression with second order interactions (Algorithm 4) ([25][26]). A brief comparison of the four algorithms is shown in Table 3-1.

All the regression models we used are formulated based on Least Square Method ([18][23]). For a data set consists of \( n \) points (data pairs) \( (x_i, y_i) \), \( i = 1, ..., n \), where \( x_i \) is an independent variable and \( y_i \) is a dependent variable. The model can be denoted as \( f_i \). The least square method can minimize the sum, \( S \), of the square of residuals, \( r_i \). The problem is formulated as

\[
\min \quad S = \sum_{i=1}^{n} r_i^2 \quad \text{where} \quad r_i = y_i - f_i
\]  

(3-1)
Table 3-1 Comparison of Four Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Algorithm 1</th>
<th>Algorithm 2</th>
<th>Algorithm 3</th>
<th>Algorithm 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>Load of previous n points. ( y = (y_1, y_2, \ldots, y_n) )</td>
<td>Load of previous 2 months, same hours as the forecast hours. (same day type for Algorithm 2). ( y = \begin{pmatrix} y_{i1} \ \vdots \ y_{in} \ y_{i8} \ \vdots \ y_{n8} \end{pmatrix} = (y_{ij}) )</td>
<td>( i ) is the day number denotation</td>
<td>( j ) is the number of point, ( j = 1, 2, \ldots, 8 ).</td>
</tr>
<tr>
<td>Predictors</td>
<td>Number of points ( n ). ( x = (1, 2, 3, \ldots, n) )</td>
<td>Historical Load ( y )</td>
<td>Month ( x_1 ), Date ( x_2 ), Day of Week ( x_3 ).</td>
<td>Temperature ( x_4 ), Humidity ( x_5 ), Hour ( x_6 ).</td>
</tr>
<tr>
<td>Regression Model</td>
<td>( \hat{y} = a_0 + a_1 x )</td>
<td>( \hat{y} = a_0 + a_1 x + a_2 x^2 )</td>
<td>( \hat{y} = \sum_{k=0}^{6} a_k x_k )</td>
<td>( \hat{y} = \sum_{k=0}^{6} a_k x_k + \sum_{l=1}^{6} b_{lj} x_l x_j ) where ( x_0 ) is the constant, ( a_k ) is the linear regression coefficients</td>
</tr>
<tr>
<td>Or Forecast Value</td>
<td>( \hat{y} = mean(y) )</td>
<td>where ( x_0 ) is the constant, ( a_k ) is the first-order coefficients, ( b_{lj} ) is the second-order coefficients</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.1.1 Algorithm 1: Direct Curve Fitting (Polynomial Regression)

Direct Fitting (DF) considers only data points immediately ahead of the to-be-forecasted data points. This method does not consider any other elements such as data from previous weeks, current temperature and humidity, etc. Polynomial Regression [23] is used to calculate the curve fitting parameters which can be used to calculate the data points to be forecasted.

Figure 3-1 illustrates how Algorithm 1 works. Assume that the forecast starts at 11 a.m.. First, n past data points before 11 a.m. are used to calculate the parameters of a second order polynomial curve. In our case, we have run a Monte Carlo simulation and n=10 yields the best results. So we select 10 points before the current time to calculate the curve fitting parameters. Then, the fitted curve is used to calculate the next 8 data points for eight intervals. Since the data has a 15-min resolution, eight intervals mean two hours. Note that first data point of the forecast curve needs to be shifted to align with the actual measurement obtained for the current time interval (see shaded areas Figure 3-1 at 11 a.m.). This step is to eliminate the bias caused by forecasting errors. We will use the first data point as a pivot to align the forecast curve with the actual curve.

The forecast values are calculated as:

For Simple Linear Regression (SLR): \( \hat{P} = a_0 + a_1 t + bias \)  \( \text{(3-2)} \)

For k\textsuperscript{th}-order Polynomial Regression (PR): \( \hat{P} = a_0 + a_1 t + a_2 t^2 + \ldots + a_k t^k + bias \)

where \( a_0, a_1, \ldots, a_k \) are the regression coefficients, \( t \) is the time and \( \hat{P} \) is the forecasted 15-min power values.

In this study, we set \( k=1 \) or 2. The reason we choose the parameter \( k \) will be mentioned later in 3.2.1.
3.1.2 Algorithm 2: Similar Day Approach

The similar day approach ([17],[20]-[22]) has been widely used in load forecast. This method takes an artificial intelligent approach by searching historical data that have similar characteristics (e.g. weather, day of the week, and the date). As discussed in the previous chapters, days with different types or weather patterns would have very different load profiles. Thus, defining the characteristics for similar days is essential to this approach.

In this research, the training data are from the past two months. There are two day types: weekday and weekend/holiday. Similar forecasting intervals refers to the same 2-hour. For example, if we are forecasting load consumptions on Tuesday between 10 a.m. to 12 a.m., the training data will be picked between 10 a.m. to 12 a.m. from Tuesdays in the past two months. The means of the historical data are used as the forecast values. Similar to Method 1, the forecast curve is shifted to eliminate the bias using the first point as a pivot. The mathematical expression of Algorithm 2 is:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} y_i + bias$$  \hspace{1cm} (3-3)

Where $n$ is the number of points selected and $y_i$ represents the measurements.
3.1.3 Algorithm 3: Multiple Linear Regression (MLR)

Regression is the one of most widely used statistical methods to predict future values. In load forecasting applications ([20][31][32]), regression techniques are usually used to model the relationship between load consumption and factors such as day type, day of the week, and weather factors. In this research, a set of variables including month, date, day of week, temperature, humidity, and hour are used as predictors to fit the measurements into the multiple linear model[23]. We will discuss in the later sections why we discarded the type of the day as an input. Similar to Algorithm 2, all the training data sets are from the past two months and have the same data types and within the same time intervals. Algorithm 3 is formulated as

$$\hat{y} = \sum_{k=0}^{6} a_k x_k + bias$$  \hspace{1cm} (3-4)

where $x_0$ is the constant, $a_k$ is the partial regression coefficients of MLR.

3.1.4 Algorithm 4: Multiple Linear Regression with Interactions between Predictors

Algorithm 4 is implemented based on the past two-month historical data at the same time intervals. Similar to Algorithm 2 and 3, training load data in Algorithm 4 are also from the past two months with the same data type and the same time interval. The predictors included in Algorithm 4 are month, date, day of week, temperature, humidity, and hour. The difference between Algorithm 4 and 3 is that second order interactions between predictors are added to the regression to make the curve more sensitive to each factor. Algorithm 4 is formulated as:

$$\hat{y} = \sum_{k=0}^{6} a_k x_k + \sum_{j \geq i}^{6} b_{ij} x_i x_j + bias$$ \hspace{1cm} (3-5)

where $x_0$ is the constant, $a_k$ is the coefficient of first-order terms (‘main effects’), and $b_{ij}$ is the coefficient of second-order interaction terms.

3.2 Implementation and Performance Evaluation

Because Algorithms 2-4 require 2 months of historical data as the training data, we select July 2012 to April 2013 data as the testing data sets for performance comparison of the four

---

3 An interaction term between the two predictors would allow the effect of one predictor consumed to depend on the other one, and vice versa.
algorithms. In addition, one month data is used to test the accuracy of each algorithm. Parameters of the algorithms, such as the number of data chosen and the order of terms, will be varied to show how an optimal set of parameters have been selected for each algorithm.

The criterion used to evaluate the accuracy of forecast is the absolute percentage error (APE). The mean absolute percentage error (MAPE)[14] is the average value of APE. APE and MAPE are calculated by:

\[
APE = \frac{|\text{Forecast Load} - \text{Actual Load}|}{\text{Actual Load}} \times 100\% \tag{3-1}
\]

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} APE \tag{3-2}
\]

APE are calculated for each forecasting point. The current load is set as a reference such that the closer the point to the reference the more accurate it is. The MAPE of each forecasting point is obtained and plotted (as a shape of horn). The parameter used to evaluate the general accuracy of each algorithm is the MAPE of the eighth forecasting point. Also, the max APE will be calculated to evaluate whether or not the performance of each algorithm is stable. In all of the MAPE plot, the black line means the MAPE, and the pink line means the max APE of that forecasting point.

3.2.1 Algorithm 1: Direct Curve Fitting

a. Fitting method selection:

For direct fitting, one can use linear fitting, second-order and third-order fitting, or other higher order methods. Sometimes, the exponential and reciprocal fitting can also be used. The mathematical models for the fitting methods are listed as follows:

First-Order (Linear) Fitting: \[ \hat{y} = a_0 + a_1 x \]

Quadratic Fitting: \[ \hat{y} = a_0 + a_1 x + a_2 x^2 \]

Cubic Fitting: \[ \hat{y} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \]

n-Order Fitting: \[ \hat{y} = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \]

Mixed Fitting: \[ \hat{y} = a_0 + a_1 x + \cdots + a_n x^n + b e^x + c \frac{1}{x} \]
To identify the best fitting function, we compared with different fitting methods. In this case, the number of points is set to be 10. Using March 12th 2013, 10 a.m. data as example, we modeled different curve fitting methods and the results are shown in Figure 3-2. In the analysis of single time point, the mean value of APEs of the eight points are calculated as the evaluation criterion, because for single trial the eighth point may not be the point which has the largest APE. Figure 3-2 (a) is the first-order (linear) fitting, (b) is the second-order (quadratic) fitting, (c) is the third-order (cubic) fitting, (d) is the second-order plus exponential, (e) is the second-order plus reciprocal. Because (d) has an MAPE that is 100 times of the value of itself, it is immediately excluded from our options. The other 4 methods have similar MAPE at this point.
Figure 3-2 Comparison of Algorithm 1 using 5 Direct Fitting Methods (March 12th 2013, 10AM)
Then, we compared the performance of the methods using data of a whole month (in this case March 2013 data). As shown in Figure 3-3, where (a) is the first-order (linear) fitting, (b) is the second-order fitting, (c) is the third-order fitting, (d) is the second-order plus reciprocal, the latter two method (c) and (d) have a relatively large error which cannot be accepted. So the linear and quadratic curve fitting are selected.

Figure 3-3 MAPEs of Algorithm 1 with 4 Direct Fitting Methods on March 2013
(Black: MAPE Pink: Max APE)
b. **Determining the Number of Previous Data**

In Figure 3-4, the left column is the first-order fitting, while the right column is the second-order fitting. By comparison, we find that the second-order fitting has a better fit to the training data, but the forecasting is not accurate. A comparison is made for a month and the result is shown in Figure 3-5.

Figure 3-5 shows that the accuracy of the first-order fitting is very stable (around 4%) no matter how the number of points changes; while for the second-order fitting, the forecast performance is improving as the number of data points increases.
Figure 3-4 Comparison of Algorithm 1 with Different Data Points on March 12\textsuperscript{th} 2013 10AM
Figure 3-5 MAPEs of Algorithm 1 with Different Data Points on March 2013
(Black: MAPE  Pink: Max APE)
c. Accuracy of Algorithm 1

The MAPEs are calculated for the data from July 2012 to April 2013 and the results are shown in Figure 3-6 and Table 3-2. The results demonstrate that the linear fitting has a better performance than the second-order fitting.

![Figure 3-6 General MAPEs of Algorithm 1 Linear Regression and Quadratic Regression](image)

*Figure 3-6 General MAPEs of Algorithm 1 Linear Regression and Quadratic Regression*

*Black: MAPE  Pink: Max APE*

<table>
<thead>
<tr>
<th>Algorithm 1</th>
<th>Point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Fitting</td>
<td>Max APE(%)</td>
<td>21.80</td>
<td>26.90</td>
<td>27.99</td>
<td>28.65</td>
<td>31.71</td>
<td>32.05</td>
<td>32.52</td>
<td>33.30</td>
</tr>
<tr>
<td></td>
<td>MAPE(%)</td>
<td>0.77</td>
<td>1.25</td>
<td>1.72</td>
<td>2.19</td>
<td>2.68</td>
<td>3.18</td>
<td>3.69</td>
<td>4.21</td>
</tr>
<tr>
<td>Second Order Fitting</td>
<td>Max APE(%)</td>
<td>21.47</td>
<td>27.41</td>
<td>34.02</td>
<td>41.54</td>
<td>48.75</td>
<td>53.39</td>
<td>58.42</td>
<td>63.09</td>
</tr>
<tr>
<td></td>
<td>MAPE(%)</td>
<td>0.82</td>
<td>1.40</td>
<td>2.00</td>
<td>2.63</td>
<td>3.32</td>
<td>4.05</td>
<td>4.84</td>
<td>5.67</td>
</tr>
</tbody>
</table>
3.2.2 Algorithm 2: Similar Day Approach

a. Training Data Set Selection

To determine what inputs we will use as the training data set, we conducted a number of simulations. Using March 12th 2013 10a.m. as an example for illustration purpose, the comparison between with and without type of day is shown in Figure 3-7. Figure 3-7(a) is the plot of data without the type of day filter and (b) is the plot with type filter. There is no significant difference between these two methods. Both methods have small errors which are far more accurate than Algorithm 1. Then, March data are used to obtain a monthly performance evaluation.

It can be inferred from Figure 3-8 that the training data with type of day as an input will have a slightly more accurate performance than without day type as an input. For the Similar Day Approach, we will choose the data with type filter to see its accuracy.
b. Accuracy of Algorithm 2

Figure 3-8 MAPEs of Algorithm 2 Whether to Add Type Filter on March 2013

(a) Without Type Filter

(b) With Type Filter

Figure 3-9 General MAPEs of Algorithm 2 Without/With Type Filter

(Black: MAPE  Pink: Max APE)
Algorithm 2, with or without type of day filter for the training data, shows better performance than Method 1. The MAPE is below 3%, which is better than Algorithm 1 (5%). But the maximum APE for the farthest point is around 30%, which does not improve compared to Algorithm 1. Further analysis is then made to find out the points with large errors.

<table>
<thead>
<tr>
<th>Algorithm 2</th>
<th>Point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max APE(%)</td>
<td>21.83</td>
<td>27.41</td>
<td>28.64</td>
<td>29.72</td>
<td>29.98</td>
<td>30.35</td>
<td>30.49</td>
<td>31.20</td>
</tr>
<tr>
<td></td>
<td>MAPE(%)</td>
<td>0.68</td>
<td>1.02</td>
<td>1.32</td>
<td>1.59</td>
<td>1.84</td>
<td>2.09</td>
<td>2.33</td>
<td>2.56</td>
</tr>
<tr>
<td>Without Type Filter</td>
<td>Max APE(%)</td>
<td>21.62</td>
<td>26.89</td>
<td>27.84</td>
<td>28.59</td>
<td>29.09</td>
<td>29.42</td>
<td>29.15</td>
<td>29.21</td>
</tr>
<tr>
<td></td>
<td>MAPE(%)</td>
<td>0.67</td>
<td>0.97</td>
<td>1.23</td>
<td>1.44</td>
<td>1.64</td>
<td>1.82</td>
<td>2.00</td>
<td>2.17</td>
</tr>
<tr>
<td>With Type Filter</td>
<td>Max APE(%)</td>
<td>21.62</td>
<td>26.89</td>
<td>27.84</td>
<td>28.59</td>
<td>29.09</td>
<td>29.42</td>
<td>29.15</td>
<td>29.21</td>
</tr>
<tr>
<td></td>
<td>MAPE(%)</td>
<td>0.67</td>
<td>0.97</td>
<td>1.23</td>
<td>1.44</td>
<td>1.64</td>
<td>1.82</td>
<td>2.00</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Figure 3-10 Dates Which has a Forecast Error over 15% with Algorithm 2

Figure 3-10 shows the dates distribution when forecast errors over 15% are observed. The date distribution is more scattered without the day-type as a filter and both methods have seen several days in December that have large errors.
### Table 3-4 Counts of Each Date (Forecast without Type Filter)

<table>
<thead>
<tr>
<th>Month</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>1</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td></td>
<td>26</td>
<td>9</td>
<td>15</td>
<td>17</td>
<td>24</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Count</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 3-5 Counts of Each Date (Forecast with Type Filter)

<table>
<thead>
<tr>
<th>Month</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>12</th>
<th>1</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td></td>
<td>26</td>
<td>15</td>
<td>17</td>
<td>24</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>Count</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>


Looking into those days with large forecast errors (Table 3-4 & Table 3-5), we found that in Table 3-4 and Table 3-5, several dates in common are Jan. 15th, April 19th, June 26th, July 15th, July 17th, July 24th, September 14th, December 24th, and December 25th. Most of those days are Tuesdays and Fridays, which are the max power day of the week and the pre-weekend day. December 24th and 25th is the Christmas, during which time the school is shut down. Thus, they are special days that are hard to use historical data to produce accurate load forecast.

3.2.3 Algorithm 3: Multiple Linear Regression (MLR)

a. Predictors Selection

As mentioned above in Chapter 2, variables such as month, date, day of week, temperature, humidity, day type and hour of day may have influence on load patterns. These variables can be categorized into 3 groups [24]:

- Quantitative variables: temperature, humidity
- Dummy variables: hour of the day, month, date, day of week
- Categorical variables: type of day

As experimented, the variables of month, date, day of week, temperature, humidity will greatly contribute to the prediction accuracy when they are served as predictors. But the effect of the categorical variable type of day (will serve as an indicator variable t, t = 1 if weekday, t = 2 if weekend/holiday) and hour (h) of day is not so obvious. To determine which variables should be used as inputs to the regression model, we used the March data to compare the accuracy of the forecast for a combination of type-of-day and hour-of-the-day.
In Figure 3-11, we can see that if the type-of-day factor is included, although the MAPE is low, there may be many unexpected outliers because the maximum APE is over 50%. This suggests that if the type-of-day is used as an input, some forecast values can have very large forecasting errors, which are not desirable.

Therefore, we decided to use the hour-of-the-day as an input and discard the type-of-day as an input for Algorithm 3.

b. Accuracy of Algorithm 3

It has a MAPE no more than 3% and maximum APE around 30%, which is a relatively good
performance, similar with the performance of Algorithm 2.

![Figure 3-12 General MAPEs of Algorithm 3](image)

Figure 3-12 General MAPEs of Algorithm 3
(Black: MAPE   Pink: Max APE)

<table>
<thead>
<tr>
<th>Point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max APE(%)</td>
<td>20.81</td>
<td>24.92</td>
<td>26.68</td>
<td>27.88</td>
<td>28.86</td>
<td>29.73</td>
<td>30.35</td>
<td>30.96</td>
</tr>
<tr>
<td>MAPE(%)</td>
<td>0.76</td>
<td>1.18</td>
<td>1.54</td>
<td>1.84</td>
<td>2.10</td>
<td>2.35</td>
<td>2.58</td>
<td>2.82</td>
</tr>
</tbody>
</table>

### 3.2.4 Algorithm 4: Multiple Linear Regression with Interactions between Predictors

#### a. Predictors Selection

Similar to Algorithm 3, a comparison is first made to determine the input factors of the
regression model. As shown in Figure 3-14, without type-of-the-day filters, we will achieve much better performance. This is mainly because in short-term load forecast, the data points immediately before the to-be-forecasted data contains the most relevant information that contributes to the accuracy of forecasting. Thus, we exclude the type-of-the-day as an input factor in the regression models.

Figure 3-13 MAPEs of Algorithm 4 With/Without Type or Hour Factor March 2013
(Black: MAPE  Pink: Max APE)
The method with h but without t shows the best performance. Although the performances are really close to each other, we choose the same predictors as Algorithm 3 for better comparison. In addition, the interactions of each of the two variables have been added into the regression model.

b. Training Data Set Selection

Similarly with Algorithm 3, the forecasting performance of training data with/without type filter is compared.

In Figure 3-14, (a) is the MAPE demonstration without the type filter, while (b) is with the type filter. It is obvious that the forecast without type filter has a far better performance. Thus, we choose the forecast method without type filter.
c. Accuracy of Algorithm 4

Figure 3-15 General MAPEs of Algorithm 4
(Black: MAPE  Pink: Max APE)

Table 3-7 MAPEs of Algorithm 4

<table>
<thead>
<tr>
<th>Point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max APE(%)</td>
<td>24.00</td>
<td>45.84</td>
<td>63.80</td>
<td>77.26</td>
<td>88.95</td>
<td>101.46</td>
<td>114.65</td>
<td>125.15</td>
</tr>
<tr>
<td>MAPE(%)</td>
<td>0.94</td>
<td>1.54</td>
<td>2.05</td>
<td>2.49</td>
<td>2.89</td>
<td>3.26</td>
<td>3.61</td>
<td>3.96</td>
</tr>
</tbody>
</table>

Figure 3-15 and Table 3-7 shows Algorithm 4 has a MAPE of 3.9% but its maximum MAPE exceeds 100%. Thus, the performance of Algorithm 4 is not very robust and stable.

3.3 Curve Fitting Quality and Forecast Accuracy

In this thesis, we applied 3 kinds of different regression methods to develop short-term
building load forecast models which warrant a discussion on the observation (in this case: load) and the terms (predictors, interactions). In statistics, as mention above, R-squared is a parameter indicating how well data points fit a statistical model [34]. In the analysis of Algorithm 1, we compared 5 regression models: the linear, quadratic, cubic, with exponential and with reciprocal models. When comparing the fitting curves, it is obviously that the higher order regression has a better fitting (close to original curve) for the historical data sets as shown in Table 3-8, in which R-squared and MAPE are calculated for each model.

The results show that the model with better fitting (R-squared closer to 1) can have worse performance on forecasting future values. Similar phenomena have been observed when using Algorithm 3 and 4. Therefore, when used for curve fitting, Algorithm 4 achieved a better fit than Algorithm 3 because it contains more terms and considers more conditions. However, when forecasting future values, Algorithm 4 is not as good as Algorithm 3. This phenomenon is caused by parameter sensitivities. The better the model fits a curve using historical data, the more sensitive it will be to data point changes.

Table 3-8 R-squared and MAPE comparison of Algorithm1

<table>
<thead>
<tr>
<th>Algorithm 1</th>
<th>Linear</th>
<th>Quadratic</th>
<th>Cubic</th>
<th>With Exponential</th>
<th>With Reciprocal</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.9042</td>
<td>0.9577</td>
<td>0.9624</td>
<td>0.9644</td>
<td>0.9577</td>
</tr>
<tr>
<td>MAPE(%)</td>
<td>3.2066</td>
<td>3.3783</td>
<td>5.2546</td>
<td>1430.1421</td>
<td>3.4035</td>
</tr>
</tbody>
</table>

3.4 Conclusions

A comparison of MAPEs of the four algorithms is shown in Figure 3-16. From the above modeling results, we have drawn the following conclusions:

- Algorithm 1, the direct fitting method, is the easiest way to forecast. It forecasts the load of following two hours only based on the load of the former hours, regardless of all the other elements that may have influenced the load. Algorithm 1 has a better
accuracy with linear fitting, which means the future load is strongly correlated to previous data points.

- Algorithm 2, the similar day approach, is the forecast algorithm that has the best overall performance. It uses the mean value of the historical similar days as the forecast value and yields very accurate forecasts without a lot of calculations. This algorithm is also very stable and robust.

- Algorithm 3, the linear regression, is also a very accurate algorithm. It is implemented with more input factors, such as month, day, day-of-the-week, temperature and humidity, which makes it a more adaptive forecasting method.

- Algorithm 4, the second-order regression, is less accurate than the linear regression. This method fits the historical curve well, but it can overfit or underfit future values. This is because the forecast values are very sensitive to certain input variables.

---

**Figure 3-16 MAPE of Each Algorithm**

![Graph showing the Mean Percentage Error (MAPE) of each algorithm over time. The x-axis represents time after the current point in minutes, ranging from 0 to 120. The y-axis represents mean percentage error (%), ranging from 0 to 8. The graph compares the performance of different algorithms, showing how they vary with time. The algorithms are Algorithm 1 Linear, Algorithm 1 Second-order, Algorithm 2, Algorithm 3, and Algorithm 4. Each line on the graph represents a different algorithm, with distinct markers for easy identification.](image-url)
CHAPTER 4  Conclusion and Summary

In this thesis, a thorough statistical analysis is conducted to identify the energy signatures of a university building load using the 15-min building energy consumption data collected at a substation feeder that supplies the North Carolina State University Centennial Campus. Then, a two-hour short-term load forecaster is developed. Four load forecast methods are developed and their performance compared.

The contributions of the thesis are summarized as follows:

- Developed a set of energy signatures that can be conveniently used in building energy management systems for building energy analysis, building load desegregation, energy use anomaly detection, and building load modeling.
- Revealed the unique characteristics of university building energy consumption in terms of the correlation of building energy use with respect to school calendars, faculty and students behaviors, and weather factors.
- Developed an efficient building load forecaster that equipped with four load forecasting methods. This forecaster allows building managers and operators to make smart energy purchasing decisions and dispatch co-gen more efficiently. Potentially save the university millions dollars a year.

Our future research direction is to obtain multi-year data and integrate our energy signatures and building load forecaster into the university building energy management system.
References


[5] Yang Gao; Emmanuel Tumwesige; Brian Cahill; Karsten Menzel; “Using data mining in optimisation of building energy consumption and thermal comfort management,” 2nd International Conference on Software Engineering and Data Mining (SEDM), 2010

[6] Du Xin-hui; Tian Feng; Tan Shao-qiong; “Study of Power System Short-term Load Forecast Based on Artificial Neural Network and Genetic Algorithm,” International Conference on Computational Aspects of Social Networks (CASoN), page(s): 725-728, 2010


[10] Koo, Bon-Gil; Kim, Min-Seok; Kim, Kyu-Han; Lee, Hee-Tae; Park, June-Ho; Kim, Cheol-Hong; “Short-Term Electric Load Forecasting Using Data Mining Technique,” 7th International Conference on Intelligent Systems and Control (ISCO), page(s): 153-157, 2013


[12] Xudong Ma; Ran Cui; Yu Sun; Changhai Peng; Zhishen Wu; “Supervisory and Energy Management System of Large Public Buildings,” International Conference on Mechatronics and Automation (ICMA), 2010


Appendix
Appendix A

Academic Calendar of 2011-2013

**Academic Calendar**

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fall Semester</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Day of Classes</td>
<td>Wed, Aug 17</td>
<td>Th, Aug 16</td>
</tr>
<tr>
<td>No Classes; Labor Day</td>
<td>Mon, Sept 5</td>
<td>Mon, Sept 3</td>
</tr>
<tr>
<td>No Classes; Fall Break</td>
<td>Th-Fri, Oct 6-7</td>
<td>Th-Fri, Oct 4-5</td>
</tr>
<tr>
<td>No Classes; Thanksgiving</td>
<td>Wed-Fri, Nov 23-25</td>
<td>Wed-Fri, Nov 21-23</td>
</tr>
<tr>
<td>Last Day of Classes</td>
<td>Fri, Dec 2</td>
<td>Fri, Nov 30</td>
</tr>
<tr>
<td>Reading Days</td>
<td>Mon-Tu, Dec 5-6</td>
<td>Mon-Tu, Dec 3-4</td>
</tr>
<tr>
<td>Final Examinations</td>
<td>Wed-Th, Dec 7-15</td>
<td>Wed-Th, Dec 5-13</td>
</tr>
<tr>
<td>Graduation</td>
<td>Sat, Dec 17</td>
<td>Sat, Dec 15</td>
</tr>
<tr>
<td><strong>Spring Semester</strong></td>
<td>2012</td>
<td>2013</td>
</tr>
<tr>
<td>First Day of Classes</td>
<td>Mon, Jan 9</td>
<td>Mon, Jan 7</td>
</tr>
<tr>
<td>No Classes; MLK Jr Day</td>
<td>Mon, Jan 16</td>
<td>Mon, Jan 21</td>
</tr>
<tr>
<td>No Classes; Spring Break</td>
<td>Mon-Fri, Mar 5-9</td>
<td>Mon-Fri, Mar 4-8</td>
</tr>
<tr>
<td>No Classes; Spring Holiday</td>
<td>Th-Fri, Apr 5-6</td>
<td>Th-Fri, March 28-29</td>
</tr>
<tr>
<td>Last Day of Classes</td>
<td>Fri, Apr 27</td>
<td>Fri, Apr 26</td>
</tr>
<tr>
<td>Reading Days</td>
<td>Mon-Tu, Apr 30-May 1</td>
<td>Mon-Tu, Apr 29-30</td>
</tr>
<tr>
<td>Final Examinations</td>
<td>Wed-Th, May 2-10</td>
<td>Wed-Th, May 1-9</td>
</tr>
<tr>
<td>Commencement</td>
<td>Sat, May 12</td>
<td>Sat, May 11</td>
</tr>
<tr>
<td><strong>First Summer Session</strong></td>
<td>2012</td>
<td>2013</td>
</tr>
<tr>
<td>First Day of Classes</td>
<td>Mon, May 21</td>
<td>Mon, May 20</td>
</tr>
<tr>
<td>No Classes; Mem. Day</td>
<td>Mon, May 28</td>
<td>Mon, May 27</td>
</tr>
<tr>
<td>Last Day of Classes</td>
<td>Fri, June 22</td>
<td>Fri, June 21</td>
</tr>
<tr>
<td>Final Examinations</td>
<td>Mon-Tu, June 25-26</td>
<td>Mon-Tu, June 24-25</td>
</tr>
<tr>
<td><strong>Second Summer Session</strong></td>
<td>2012</td>
<td>2013</td>
</tr>
<tr>
<td>First Day of Classes</td>
<td>Th, June 28</td>
<td>Mon, July 1</td>
</tr>
<tr>
<td>No Classes; Independence Day</td>
<td>Wed, July 4</td>
<td>Th, July 4</td>
</tr>
<tr>
<td>Last Day of Classes</td>
<td>Wed, Aug 1</td>
<td>Fri, Aug 2</td>
</tr>
<tr>
<td>Final Examinations</td>
<td>Th-Fri, Aug 2-3</td>
<td>Mon-Tu, Aug 5-6</td>
</tr>
</tbody>
</table>