A LOAD-BASED TEMPERATURE PREDICTION MODEL FOR ANOMALY DETECTION

by

Masoud Sobhani

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2017

Approved by:

_____
Dr. Tao Hong

_____
Dr. Gang Chen

_____
Dr. Pu Wang

ABSTRACT

MASOUD SOBHANI. A load-based temperature prediction model for anomaly detection. (Under the direction of DR. TAO HONG)

Electric load forecasting, as a basic requirement for the decision-making in power utilities, has been improved in various aspects in the past decades. Many factors may affect the accuracy of the load forecasts, such as data quality, goodness of the underlying model and load composition. Due to the strong correlation between the input variables (e.g., weather and calendar variables) and the load, the quality of input data plays a vital role in forecasting practices. Even if the forecasting model were able to capture most of the salient features of the load, a low quality input data may result in inaccurate forecasts. Most of the data cleansing efforts in the load forecasting literature have been devoted to the load data. Few studies focused on weather data cleansing for load forecasting. This research proposes an anomaly detection method for the temperature data. The method consists of two components: a load-based temperature prediction model and a detection technique. The effectiveness of the proposed method is demonstrated through two case studies: one based on the data from the Global Energy Forecasting Competition 2014, and the other based on the data published by ISO New England. The results show that by removing the detected observations from the original input data, the final load forecast accuracy is enhanced.

DEDICATION

To My Wife: Shaghayegh Javan

ACKNOWLEDGMENTS

When I look back to my past two years, I see two important landmarks in my life as a student: a great chance and a right choice. I was lucky that I started my Master's program with taking a course instructed by Dr. Tao Hong. He opened the door of a new world to me, and made me motivated to continue my education in the field of energy analytics. The right choice that I made was joining Dr. Hong's research group, which ended up to this thesis. I would like to express my sincere gratitude to my advisor Dr. Tao Hong for guiding me in learning not only in the scientific area, but also on a personal level.

I also want to thank my committee members. Dr. Gang Chen and Dr. Pu Wang, for generously offering their time and guidance through my preparation of this thesis.

Last but not the least; I want to express my deepest appreciation and love to my family: my wife, Shaghayegh Javan, my parents, Mostafa Sobhani and Azar Fadakhah, who have provided me emotional support in my life and encouraged me all the way along.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Electric load forecasting is one of the basic requirements for power system planning and operations. Since the inception of the power industry, people have recognized the importance of accurate load forecasts to the business. Virtually every sector of the industry, from generation, transmission to distribution and retail are using load forecasts for their decision-making processes [1]. Due to the restructure of the market, deployment of smart grid technologies and the push for increasing renewable penetration, the demand for accurate load forecasts is rising to an even higher level than before.

We can classify the load forecasting problems based on the forecast horizon into four categories: very short-term load forecasting (VSTLF), short-term load forecasting (STLF), medium-term load forecasting (MTLF) and long-term load forecasting (LTLF) with the cut-off horizons of one day, two weeks, and three years, respectively [2]. Each of them has specific applications in various departments of a utility company, such as planning, operations and energy trading. The research in this paper is generic to load forecasting across all horizons.

Figure 1.1 shows a schematic diagram of a load forecasting process. The process consists of three main components: input data (weather history, load history and calendar variables), the forecasting model, and the output (load forecast). There are various factors than can affect the performance of a load forecasting process. The forecasting model is the principal component that processes the input data and makes the forecasts. Load

forecasting models are typically grouped into statistical and artificial intelligence (AI) models. Numerous amount of forecasting models have been implemented in the power industry to predict the future demand of electricity. Both statistical and AI models provide accurate outcomes and help the decision-makers in various stages.

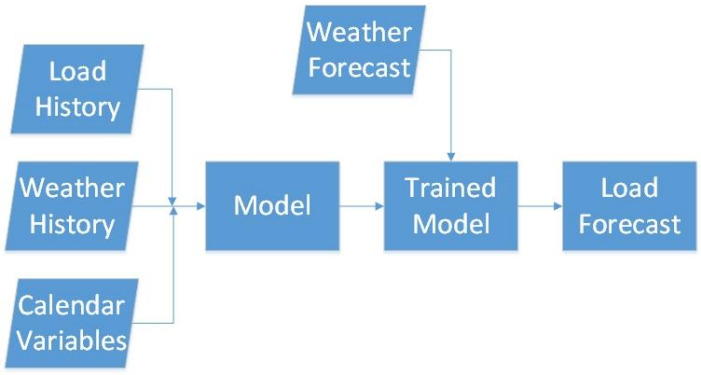

Figure 1.1: Schematic of a load forecasting process

The data requirements of a load forecasting process are provided from two general sources: the load history from the previous measurements, and weather information such as temperature, relative humidity or wind speed. In some cases, depending on the forecasting horizon, the economy information or geographical variables are also being used in load forecasting practices. Therefore, a load forecasting process uses a combination of various variables such as calendar variables, weather and load to train the forecasting model.

A low quality input, even when the forecasting model is strong, can produce unsatisfying outcomes. In other words, garbage input results in garbage output. In a real world, there are various reasons leading to inconsistencies and anomalies in the data. Aging instruments, human error and reporting failures are some of the reasons that cause

abnormal conditions. Before feeding the raw data to the forecasting model, we need to detect the anomalies, and then replace them by better values to make the data cleaner.

Most of the researches in the load forecasting literature have been focused on developing the cleansing methods for the load data with minimal attention to the weather data cleansing. The basic assumption of these studies is based on a wrong belief that the weather data is perfect and reliable. The cleansing methods have been focused/concentrated on the load data, whereas in a load forecasting process, the weather data is as important as the load data.

The aim of this research is proposing a method to detect the anomalies of the input data of temperature to tackle the weather data issues in the load forecasting literature. The proposed method consists of a temperature prediction model and an anomaly detection technique. To study the performance of the method, it is evaluated in two case studies: one based on the data from Global Energy Forecasting Competition 2014, and the other based on the data from New England ISO.

The rest of this thesis is organized as follows: Chapter 2 presents the literature review; Chapter 3 introduces the background and theoretical concepts of the study; Chapter 4 develops a load-based temperature prediction, which later combined with a technique to build an anomaly detection method; Chapter 5 presents the anomaly detection technique and the performance of the proposed technique is studied by simulating the anomalies in different levels; Chapter 6 presents a case study, in which the proposed method is

implemented on NE-ISO dataset; Chapter 7 concludes the thesis and discusses the possible future works.

CHAPTER 2: LITERATURE REVIEW

The subject of this study covers a wide range of research topics. Many papers are published each year in the field of electric load forecasting or weather prediction. To obtain a sufficient overview, this chapter reviews the representative papers in the following three subjects: (1) Electric load forecasting; (2) Temperature prediction methods; (3) Outlier detection and data cleansing in load forecasting practices.

2.1 Electric Load Forecasting

Based on forecasting horizon, load forecasting can be broadly categorized in two main groups: Short-term load forecasting (STLF) and long-term load forecasting (LTLF). The number of studies in the literature on STLF is larger than on LTLF [1]. Load forecasting techniques are generally divided in two classes: Statistical techniques and artificial intelligence techniques. In this section, we will present some recently published review papers, and then we will bring the representative papers regarding to the load forecasting techniques.

Gross and Galiana [3] presented a comprehensive and tutorial survey about the short term load forecasting. The authors define the short-term as the interval of one hour to one week. The paper begins with describing the importance of short-term load forecasting in the power industry and then the basic concepts of load forecasting are explained. The paper presents a broad literature survey in different subjects such as the type of load models, the data needs of the model, the computational requirements of the forecasting algorithm, and the availability of experimental results. After reviewing the techniques and other features, the authors discuss about the applications of different models, the methods for model selection and the evaluation measurements.

Alfares and Nazeeruddin [4] reviewed the load forecasting techniques in the literature. They classified the forecasting methods into nine groups: multiple regression, exponential smoothing, iterative reweighted least-squares, adaptive load forecasting, stochastic time series, autoregressive moving average models with exogenous inputs (ARMAX) based on genetic algorithms, fuzzy logic, ANN, and expert systems. In each

section, they review the representative researches in a chronological order. In addition, some other notable studies are also reviewed that compare different forecasting model.

Hahn et al. [5] provided a brief review of the load forecasting models. The authors introduced some literature survey papers and then discussed about the forecasting methods. Similar to many other researches, they classified the load forecasting models into two main groups: statistical methods and artificial intelligence methods. In each group, some representative papers and models are listed. Among the statistical methods, the paper reviewed regression models and time series approaches (univariate and multivariate). The AI models reviewed include Neural Network, Support Vector Machine and hybrid approaches.

Hong [1] reviewed 50 years of short-term load forecasting papers from three aspects: the developed techniques, the variables deployed and the representative works done by major research groups. The research's aim was not covering any aspects of the field, but focusing on the major developments. He reviewed conceptual and experimental literature reviews on short-term load forecasting studies. The representative publications that implemented different types of the statistical and AI models are presented. The review also highlighted that the benchmarking and the standard reporting format in the field are the two issues that have not received much attention.

Hyndman and Fan [6] proposed a methodology to forecast the density of long-term peak electricity demand. The methodology is split into two models: a semi-parametric additive model for half-hourly demand, which is in the regression framework; and the other

model for annual effect to capture the correlation with the economic and demographic factors. The proposed models were implemented in a case study to forecast the long-term peak demand for South Australia.

Hong [1] proposed a regression based model for short-term load forecasting. The model includes the interaction (cross effects) of the weather variables with the calendar variables to capture the seasonal patterns of load profile. This model has been used as a benchmark model in many studies and competitions. The Vanilla Benchmark Model (VBM) was used as the benchmark for the electric load track in the Global Energy Forecasting Competition 2012 [7]. Wang et al. studied the relationship of the load demand and the temperature of preceding hours by inserting recency effects to the VBM [8].

ANNSTLF (artificial neural network short-term load forecaster) is one of the well-known ANN load forecasting models, which has been widely used in the power utilities [9]. The third generation of this model consists of two ANN load forecasters and an adaptive combiner. One ANN predicts the base load and the other is used to capture the change from the previous day's load.

Hong and Wang [10] proposed a fuzzy interaction regression approach to short-term load forecasting. The aim of the paper was to improve the underlying linear models for fuzzy regression. The proposed model outperforms the other two fuzzy regression models and one MLR model significantly.

The papers reviewed in this section are only some representative papers and notable studies among thousands papers that are published each year in the field of load forecasting.

There are other aspects of electric load forecasting that has not been covered in this research, such as variable selection, model selection methods, evaluation measurements, probabilistic load forecasting, etc.

2.2 Temperature Prediction Methods

More than hundred years ago, Abbe and Bjerknes [11] proposed the very first weather prediction model by using partial differential equations to mathematically model the laws of physics. Improvement of the technology over the past 40 years enables the scientists to implement complicated numerical weather prediction models with huge datasets. Accurate forecasts create considerable financial revenue, save lives, support emergency management and mitigation of impacts and prevent economic losses from high-impact weather [12].

The weather prediction models are generally classified into two groups [13]: numerical or dynamic models and empirical or statistical models. The numerical weather predictions (NWP) use numerical methods to model a set of equations that describes the flow of fluids. The statistical models are based on the past information and pattern recognition skills that are mainly in four categories: linear regression; principal component analysis (PCA); canonical correlation analysis; and neural network (NN) models [14].

The first operational NWP could model only the temporal variation of the vertical structure of the atmosphere, because it consists only one layer [15]. Due the improvement of the technology and computers, now the NWP models can resolve the variations of temperature, wind and humidity.

The first implementation of artificial neural network (ANN) in weather prediction was done by Hu [16]. Hsieh and Tang [14] reviewed the applications of NN models in meteorology and oceanography. They believed that the difficulties in adopting NN to

weather predictions are nonlinear instability with short data records and its interpretation in large spatial data fields.

Most of numerical weather prediction models rely on physics of the weather flows and physical properties of climate. The operational statistical models, also just mainly use the previous measurements of the weather parameters and they use the correlation between weather related variables to predict the future values. On the other hand, different weather conditions can affect the performance of many systems, such as air traffics or energy consumption. These correlations can be used to model the performance of the mentioned systems and in a reverse approach, we can use the correlations to study the weather behaviors as well.

2.3 Anomaly Detection in Load Forecasting Practices

Accurate prediction is the principle achievement of a forecasting practice. Quality of the data is an important factor in obtaining accurate outcomes. In the real world, the raw data includes significant amount of anomalies and inconsistencies that makes preprocessing of input data the first step in load forecasting practice. Therefore, outlier detection followed by a data cleansing is important in any load forecasting. This section reviews the common approaches of the data cleansing in load forecasting practices.

[17] is a review paper that introduces some representative outlier detection methods. The authors believe that there are three fundamental approaches to the problem of outlier detection: Type 1 determines the outliers with no prior knowledge of the data (unsupervised clustering); Type 2 models the cases where there are both normal and abnormal observations (supervised classification); Type 3 models only normality (semi-supervised detection). The paper reviews different type of techniques used for outlier detection, including statistical, neural network, Machine learning and hybrid algorithms.

In the power industry, a naïve or seasonal naïve method is often used for anomaly detection. The mean and standard deviation of the all history observations or all observations at the same hour of the day are used to define a threshold for the load value. If the load is out of the threshold interval, it is marked as a detected anomaly. Xie [18] utilized similar idea for data cleansing component of her submission in GEFComp2014 [19]. The load values are estimated in her approach by implementing a multiple linear regression (MLR) model to calculate the absolute percentage error (APE) for each

observation. The observations with APE values of greater than 50% are treated as outliers and replaced by the predicted values from the regression model.

Akouemo and Povinelli [20] proposed an algorithm for anomaly detection in natural gas time series data. The detection algorithm is a two-stage method. In the first stage, a linear regression is fitted to the natural gas time series and then the geometric probability distribution of the residuals is calculated to detect the anomalies. In the second stage, based on the identified anomalies from the first stage, a Bayesian maximum likelihood classifier is trained to find the type of the anomaly. The detected anomalies are replaced by using a naïve imputation model.

Luo et al. [21] investigated the effect of data integrity attacks on the accuracy of the load forecasting. The robustness of four representative load forecasting models are studied including multiple linear regression, support vector regression (SVR), artificial neural networks, and fuzzy interaction regression. They simulated the data integrity attacks by injecting errors randomly to the data. The errors are some multipliers, which have uniform or normal distribution. Among the four models, SVR was the most robust with respect to the point forecast accuracy, while the fuzzy interaction regression was the least robust. In addition, while the level of the integrity attacks is large, all four models fail to provide accurate forecasts.

Gaun et al. [22] proposed a pre-filtering method to detect and correct spikes in load data. In this research, Spikes are classified as “micro spikes” and “macro spikes” based on their widths. The main idea of filtering is using a smoothed load data by applying a zero

phase filter. The method detects the spikes when the difference between smoothed load and the original load exceeds a threshold. The smoothed load replaces the detected spikes.

Fidalgo and Lopes [23] present a NN-based procedure to avoid the performance failures under anomalous conditions such as historical data bugs, anomalous behavior (like holidays or atypical days) and shape changes following switching operations. The pre-processing approaches used in this study include simple filtering methods like if-then rules; low pass filtering and ANN based self-filtering. A similar approach was utilized in [24]. The zero value replaces the spikes and missing observations and then a splining algorithm or an interpolation from similar day actual load data (when f, the length of zero gaps, is big) is used to fix these values.

All papers reviewed in this section focused on anomaly detection and data cleansing of the load data. There is a gap in the load forecasting literature regarding to the weather data cleansing. The aim of this research is to propose a method for detecting the anomalies in the temperature data.

CHAPTER 3: BACKGROUND

Multiple Linear Regression (MLR) is the main technique used in this research. The proposed load-based temperature prediction model is developed based on linear regression analysis. One of the application of the proposed model is outlier detection and data cleansing in load forecasting practices. The performance of the proposed anomaly detection method is evaluated by measuring the load forecasting accuracy for before and after cleansing. Tao's Vanilla Benchmark model, which is a MLR based model, is implemented for load forecasting. The accuracy of the temperature predictions is measured in Mean Absolute Error (MAE) and the accuracy of the load forecasting is calculated in Mean Absolute Percentage Error (MAPE). Therefore, in this chapter the background of the following techniques and methods are presented; (1) Multiple linear Regression (2); Tao's Vanilla Benchmark Model; (3) Evaluation measures including MAE and MAPE.

3.1 Multiple Linear Regression

MLR has been widely used in load forecasting practices. The proposed temperature prediction model (described in Chapter 4) is a regression based model, as well as, a Vanilla Benchmark Model, which has been for load forecasting. In temperature prediction, the temperature is the dependent variable and the other variables are independent ones. In load forecasting, load is dependent. Each MLR model includes quantitative and class variables; and the correlations between these variables are considered through main effects and cross effects. In this section, the theoretical background of MLR is explained.

The general linear regression model, in terms of *X* variables, is defined [25]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (3.1)$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters

$X_{i1}, X_{i2}, \ldots, X_{i,p-1}$ are known constants

ε_i is the independent normally distributed random variable $N(0,\sigma^2)$

$i = 1, 2, \ldots, n$

Since $E\{\varepsilon_i\} = 0$, the response function for regression model is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \quad (3.2)$$

The general linear regression equation indicates that the observations Y_i are independent normal variables, with mean $E\{Y_i\}$ as given by (3.1) and with constant variance σ^2.

3.2 Tao's Vanilla Benchmark Model

Vanilla Benchmark Model (VBM) is a MLR based model which was proposed first in [1]. The model is widely used in the literature. It was used as a benchmark in the GEFCom2012 for the hierarchical load forecasting track [26]. In this study, we also used VBM as a benchmark for load forecasting. The performance of the proposed anomaly detection method is evaluated by implementing VBM on raw and cleaned data. The data is cleansed by removing the detected anomalies.

Components of the Vanilla Benchmark Model are [27]:

a) Data: The model requires two sets of data: hourly load and hourly temperature for at least two years of history data.
b) Linear Trend: In the utilities with stable service territory and local economics, the locally increasing (or decreasing) trend of load demand can be captured by a quantitative variable. The *Trend* variable is defined by assigning a natural number to each hour in ascending order.
c) Temperature (*TMP*): Temperature is a driving factor electric load demand and the correlation between the load and the temperature is significantly considerable in load forecasting practices. Among all suggested functions between load and temperature based on previous studies, VBM uses 3rd ordered polynomials of the temperature to predict the load.
d) Calendar Variables: The load demand behavior is different at different times of a day, in different days of a week and in different months of a year. These three seasonal blocks are used in the VBM to input calendar variables as predictors. The

qualitative variables (Hour, Weekday, and Month) with 24, 7, and 12 classes are used.

e) Cross effects: There are obvious correlations between temperature and the hour of the day or the month of the year. The relationship between the day of the week and temperature is not evident. To capture these correlations, the interaction effects are included in the model.

Therefore, the components of the benchmarking model are as follows:

1) Quantitative variables: *Trend, TMP*;
2) Class variables: *Hour, Weekday, Month*;
3) Main effects: *Trend, Month*;
4) Cross effects: *Day*×*Hour*, *Month*×*TMP*, *Month*×*TMP*^{2}, *Month*×*TMP*^{3}, *Hour*×*TMP*, *Hour*×*TMP*^{2}, *Hour*×*TMP*^{3};

The equation of the regression model is as follows:

$$Load = \beta_0 + \beta_1 \times Trend + \beta_2 \times Day \times Hour + \beta_3 \times Month + \beta_4 \times Month \times TMP + \beta_5 \times Month \times TMP^2 + \beta_6 \times Month \times TMP^3 + \beta_7 \times Hour \times TMP + \beta_8 \times Hour \times TMP^2 + \beta_9 \times Hour \times TMP^3, \quad (3)$$

3.3 Evaluation Measurement

To evaluate the performance of a prediction we need measurement scales. In point load forecasting, there are various evaluation measures such as absolute error (AE), percentage error (PE), mean absolute error (MAE), root square mean error (RSME) and mean absolute percentage error (MAPE). In this study, for temperature prediction we use MAE and for load forecasting we use MAPE. In the remaining of this section, these measures are explained in details.

MAE is scale dependent and measures the size of error in units. Equation (3.4) shows how to calculate MAE. Since this measurement is scale dependent, it should not be used across different data sets.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Actual_t - Predict_t| \tag{3.4}$$

Mean Absolute Percentage Error (MAPE) is a scale-free measurement. The equation is shown in (3.5).

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Actual_t - Predict_t}{Actual_t}\right| \tag{3.5}$$

The percentage error measures result unreliable accuracies when the variation of the data is high, because for small amounts of data, the percentages are very big. In the case study, since we have small and negative amount of temperature data, we use MAE. On the other hand, since the load data is mainly aggregated, we use MAPE for load forecasting.

3.4 Cross validation

Cross validation is a well-known and popular method for model selection. The main idea behind cross validation is to split data, once or several times, for estimating the risk of each algorithm [28]. For model selection, part of data is used for training the model and the remaining is used for validating the predictions. There are two different approaches for cross validation in load forecasting practices; X-fold cross validation and sliding simulation.

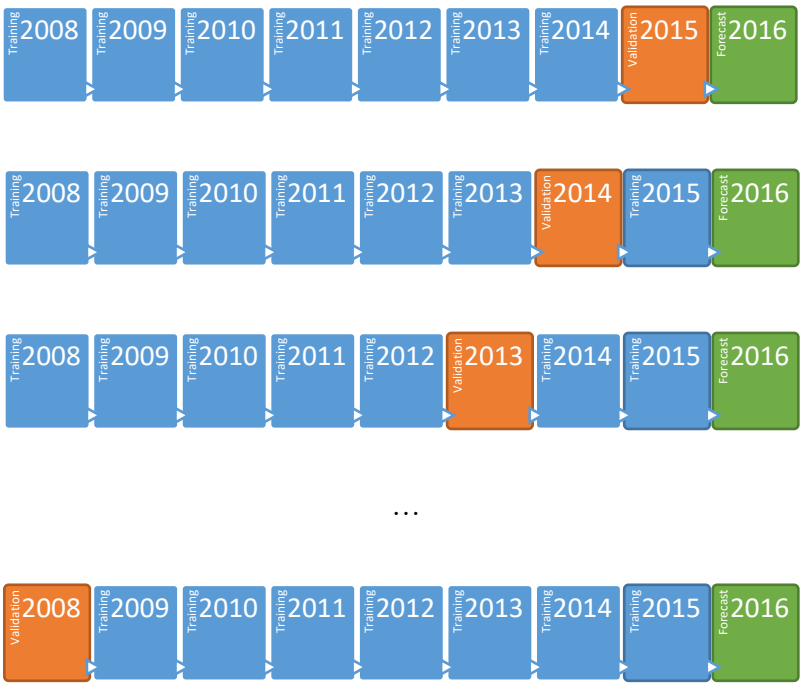

Figure 3.1: X-fold cross validation

Assume that we have 8 years of history data and we are supposed to forecast the next year. For x-fold cross validation, in each step, 7 years of the history data is used as the training window and the remaining year is used as validation year. We repeat the forecasting for all eight possible attempts and based on the forecasts we can validate the

performance of the model. Figure 3.1 shows a schematic graph of an x-fold cross validation.

In sliding simulation, in each attempt the training windows moves forward one-step to forecast the next year. The length of history can be fixed or changing. This method can be used for variable selection, variable combination, selecting the length of history etc. Figure 2 shows schematic steps of a sliding simulation.

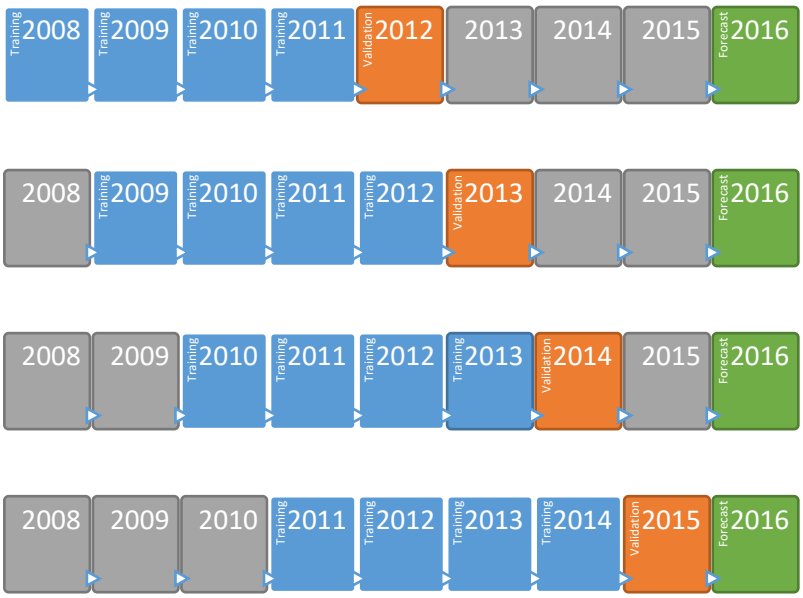

Figure 3.2: Sliding simulation

CHAPTER 4: A LOAD-BASED TEMPERATURE PREDICTION MODEL

Temperature is a driving factor in electric load demand. The correlation between load and temperature has been utilized in many load forecasting models. The aim of this chapter is to use the correlation between temperature and load to develop a temperature prediction model. The proposed model is a MLR based one where load and calendar variables are used as the independent variables to predict the temperature. In this section, we study the correlation between load and temperature and we use the extracted features to develop the temperature prediction model.

The dataset used in this chapter comes from load forecasting track of Global Energy Forecasting Competition 2014 (GEFCom 2014). The data includes 69 months of hourly load data (from January 2005 to September 2010) and 117 months of hourly weather data (from January 2001 to September 2010) [26].

4.1 Temperature vs. Load

Figure 4.1 shows the boxplot of the temperature from 2008 to 2010. Temperature follows a seasonal pattern each year with peak period in the summer and valley period in the winter. The increasing tendency of the temperature cannot be captured in figure 4.1, but this trend should be considered in modeling the temperature behavior. We define a quantitative variable (*Trend*) to consider this increase into the model.

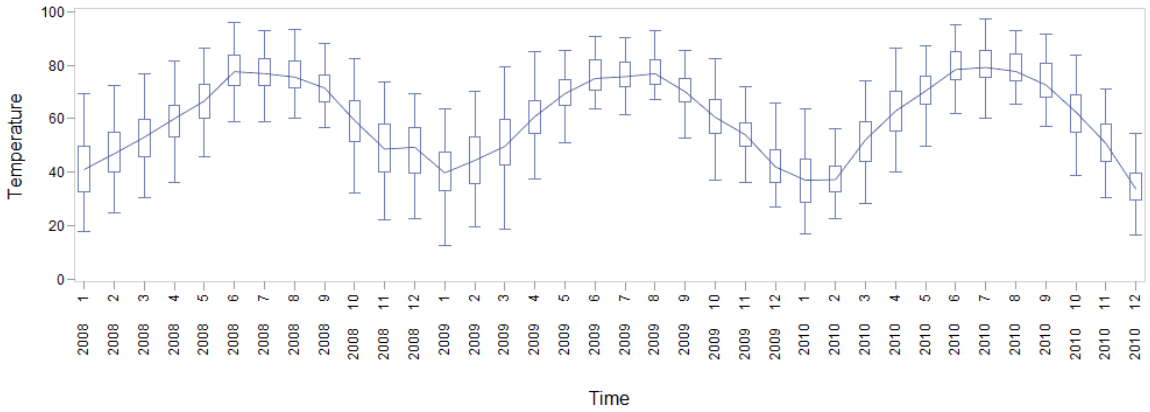

Figure 4.1: Three years temperature boxplot (2008-2010)

Correlation between temperature and load has been studied in previous researches. For example, piecewise linear function was implemented in [29]; piecewise quadratic function was used in [30]; and 3rd ordered polynomials were considered in [31]. Figure 4.2 shows a scatter plot of temperature-load for a utility in the US. The plot shows overall nonsymmetrical shape. Therefore, a piecewise linear or piecewise cubic function are proper options to model the relationship between temperature and load. Two separate multiple linear regression functions are fitted to upper and lower zones. The cut-off point is allocated to the comfort temperature, which is generally between 57 F to 63 F. For this dataset, the

cut-off point is 60 F. The third ordered polynomials of the load are used to capture the nonlinear relationship between temperature and load in each piece. Therefore, the square of load ($load^2$) and the cube of load ($load^3$) are considered as the independent variables in the MLR model.

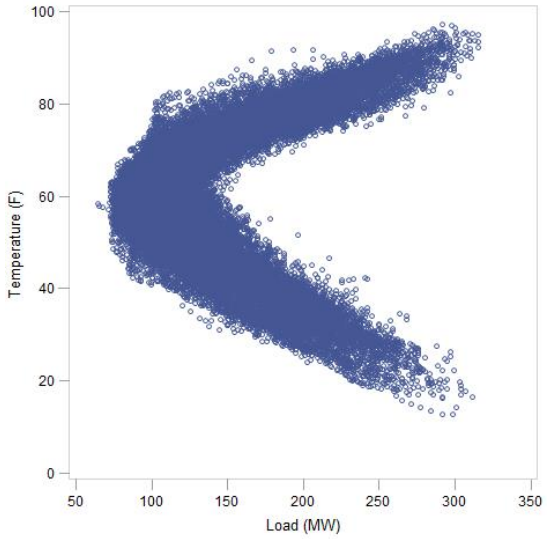

Figure 4.2: Temperature-load scatter plot (2008-2010)

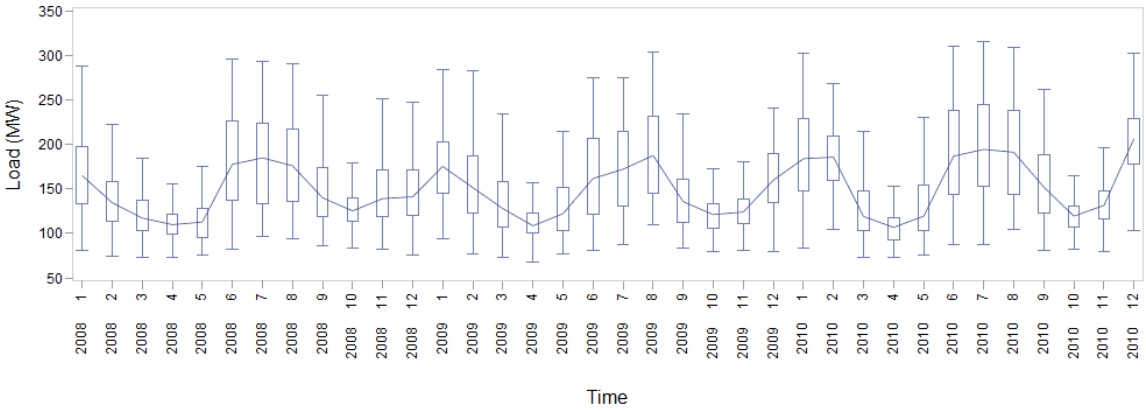

Figure 4.3: Load boxplot (2008-2010)

4.2 Seasonality

Typically, there are two seasonal blocks in temperature series: year and day. On the other hand, there are three seasonal blocks in load series: year, week and day. Since the temperature prediction model is load-based, the seasonal behaviors in both temperature and load should be considered. In this section, we discuss each seasonal block of load and temperature.

A. Month of the Year

As we saw in figure 4.1, there is a correlation between the seasons of a year and the temperature. Figure 4.3 shows the load boxplot for three years (2008 to 2010). The seasonal pattern of the load demand through a year is the result of the seasonal behavior of the temperature. Typically, there is one peak load in winter and another one in summer. Therefore, seasonality of the temperature leads to the seasonality of the energy consumption [1]. To study the seasonal correlation between the temperature and load, a scatter plot for each month is shown in Figure 4.4. It is obvious that the scatter plot in each month is different and we can construct the plot for the whole year (figure 4.2) by putting all the 12 plots together. To sum up, we define variable *Month*, which is a class variable. The interaction between *Month* and *Load* captures the seasonality of the load through a year.

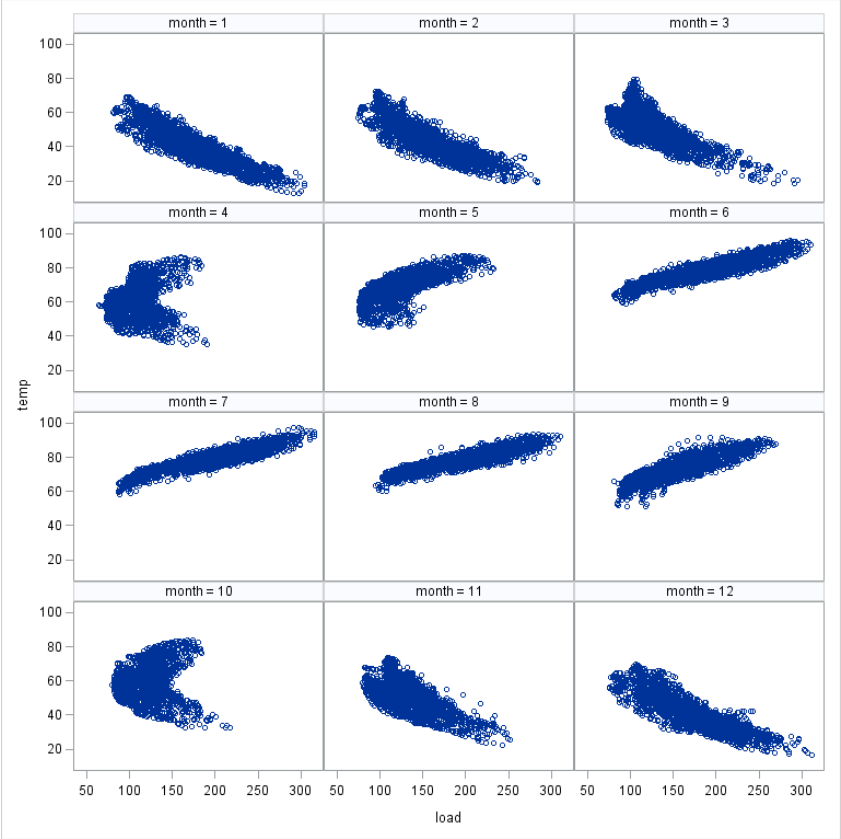

Figure 4.4: Temperature-load scatter plot (12 plots, one for each month)

B. Hour of the day

Sunrise and sunset cause the temperature changes through a day. This correlation between the temperature and daytime seems to be seasonal. Since the load and the temperature are correlated, the seasonality of the temperature should be observed in the daily load profiles as well. Figure 4.5 shows the boxplot of hourly load for week. Figure 4.6 shows 24 temperature vs. load scatter plots by each hour of a day. The load levels at the same temperature in different hours changes, which shows the interaction effects between the load and the hour. On the other hand, the load profile during the summer days

are different from the load profile during winter days. A typical load profile in a winter day has two peak values; one in the morning and one in the afternoon, whereas in a summer day, the load profile generally has one peak at noon when the temperature is at the highest amount. To capture the daily seasonality of the load, variable *Hour*, representing hour of the day, is assumed a class predictor and the interactions between the hour and the load polynomials are considered in the model. The interaction between the *Hour* and *Month* is considered to model different load profiles in winter and summer days.

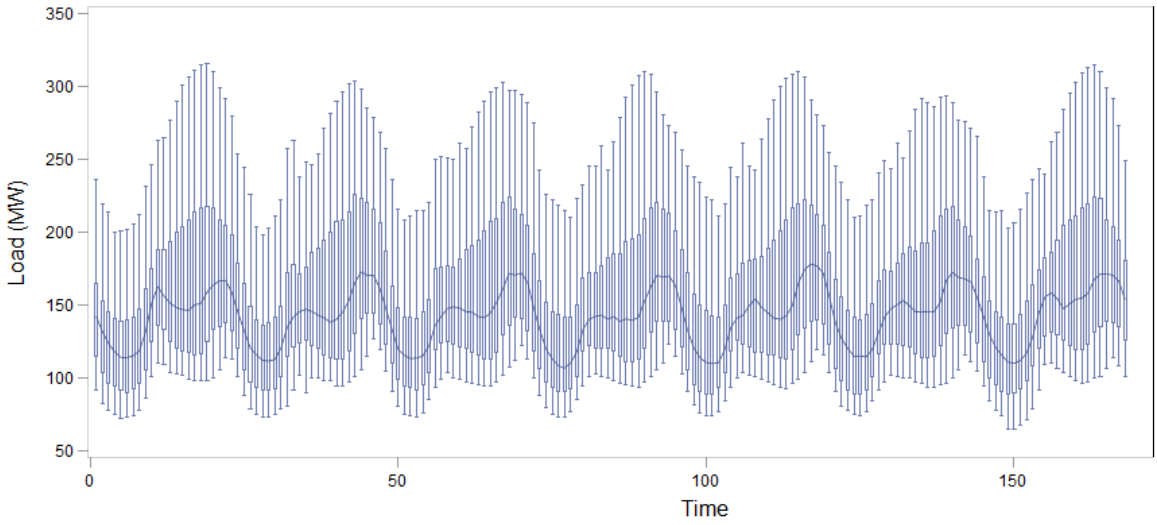

Figure 4.5: Boxplot of hourly load through a week

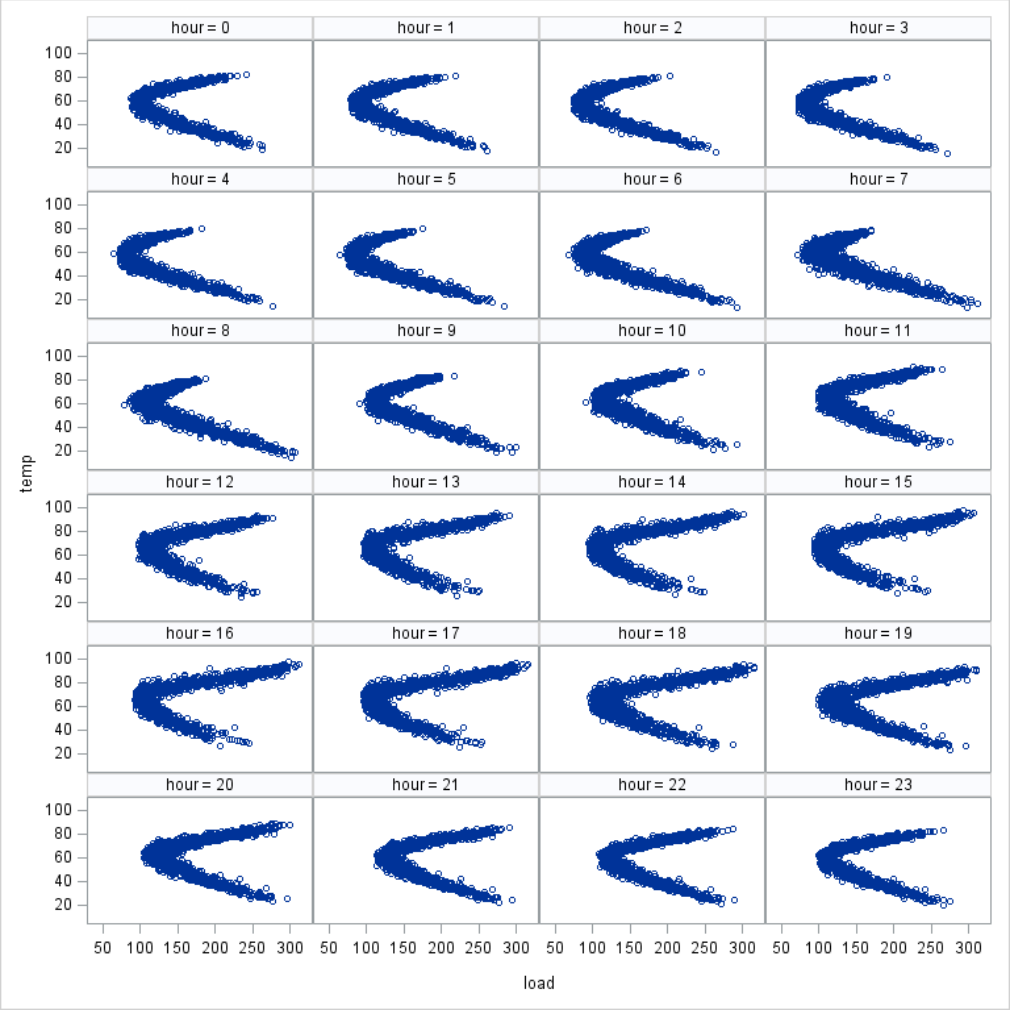

Figure 4.6: Temperature-load scatter plot for all 24 hours of a day

C. Day of the Week

In weekly block, the human activities are not the same on different days. For example, people may sleep late on Friday night, wake up late on Saturday and Sunday mornings, most offices are closed and so forth. This makes the load profiles different on weekends. On the other hand, the weather conditions are not significantly different due to

weekends. The temperature is not affected by human activities during days of a week, but since the model is a load-based model and the load is correlated to the day of the week, we should consider this correlation to our model. There are various grouping methods in the literature. The days of a week were grouped into four groups in [32]: Monday; Tuesday-Thursday; Friday; Saturday, Sunday. In [33] the dataset is divided to five day-type categories: Monday; Tuesday-Thursday; Friday; Saturday; Sunday. Another way to group the days of a week is to use dummy variable for each day. In the other words , we can categorize the days into 7 classes [1][34]. In the proposed model, the *Weekday* is a class variable and the interaction effect between *Hour* and *Weekday* is included in the model.

4.3 Model Selection Method

As mentioned in section 4.1, two separate MLR equations are fitted to upper and lower zones of temperature-load scatter plot. The cut-off point is the average temperature of all observation in the history. We name the equation for the lower zone and upper zone, equation 1 and equation 2 respectively. To predict the temperature for a new observation, we need to select the proper equation. For this reason, we grouped the data based on month and hour. Therefore, we have 12×24 categories. For each category, we take a simple average among all temperatures of that category. If the average is less than the cut-off point, the corresponding equation is number one. Otherwise, it is number two. Table 4.1 shows how the data is categorized based on the month and hour number and how each equation is allocated to a specific group.

Table 4.1: Categorizing data based on the month and hour (cut-off point is 61 ^{o}F)

Month	Hour	Mean Temperature	Equation
4	8	53.79	1
4	9	57.09	1
4	10	60.43	1
4	11	63.12	2
4	12	65.27	2
4	13	66.87	2
4	14	68.06	2
4	15	68.79	2
4	16	68.98	2
4	17	68.63	2
4	18	67.70	2
4	19	65.88	2
4	20	62.74	2
4	21	60.27	1
4	22	58.77	1
4	23	57.62	1

4.4 Benchmark model

In previous sections of this chapter, various features of temperature, data and correlative factors of load, temperature and calendar variables were discussed. In this section, we develop a benchmark model for temperature prediction based on introduced features of the data and then the performance of the proposed model will be analyzed. To improve the benchmark model, more considerations will be applied to the model in the following sections.

Based on the features of load and temperature, their correlations and seasonal blocks, we assumed nine different predictor candidates. Using these predictor candidates, eight different models (M1-M8) are developed. Table 4.2 shows the predictor candidates and the models.

Table 4.2: Candidate benchmark models

	Predictor candidates								
	Trend	*Load* $Load^2$ $Load^3$	*Hour*	*Month*	*Weekday*	*Month×Load* $Month \times Load^2$ $Month \times Load^3$	*Hour×Load* $Hour \times Load^2$ $Hour \times Load^3$	*Hour×Weekday*	*Month×Hour*
M1	×	×							
M2	×	×	×						
M3	×	×	×	×					
M4	×	×	×	×	×				
M5	×		×	×	×	×			
M6	×		×	×	×	×	×		
M7	×			×		×	×	×	
M8	×					×	×	×	×

The predictor variables are divided into two types: quantitative, and class variables. Quantitative variables: *Trend, Load,* $Load^2$, $Load^3$

Class variables: *Month, Hour, Weekday*

The effect of the predictors are either main or cross effect. The interaction (cross) effects are shown by "×". When a class variable has cross effect with another predictive variable, it is not required that the quantitative variable has a main effect in the model. In addition, when two class variables have cross effects, we can remove the individual class variables from the model.

For model selection, we did an ex-post forecasting. Two years of hourly load and temperature data is used as history to forecast the next hourly temperature. To evaluate the forecasts, MAE (Mean Absolute Error) is used as the measurement. The equation selection method proposed in the section 4.3 to switch between the piecewise equations for the new observations in the forecasting window. Table 4.3 shows the forecasts results for each model.

Table 4.3 Ex-post forecasting MAEs for the candidate models

Model	MAE (^{o}F)
M1	4.30
M2	3.14
M3	2.97
M4	2.96
M5	2.93
M6	2.85
M7	2.79
M8	2.74

Table 4.3 indicates that the model M8 gives better prediction among all the candidate models. As discussed in the section 4.1, a piecewise regression model is used to capture the correlation between the temperature and the load in two MLR equation; one for upper zone of the temp-load scatter plot and the other for the lower zone (Figure 4.2). A

comparison is shown in the table 4.4 to confirm that a piecewise regression works better than one regression on the whole points. We can see that the piecewise regression wins the single regression in any candidate model implementation.

Table 4.4: Comparison between piecewise regression v. single regression

	MAE (^{o}F)	
Model	Piecewise regression	Single regression
M1	4.30	15.19
M2	3.14	14.97
M3	2.97	5.92
M4	2.96	5.97
M5	2.93	3.85
M6	2.85	3.72
M7	2.79	3.71
M8	2.74	3.00

As a conclusion, among all candidate models (M1 to M8), piecewise version of model M8 in combination with the model selection technique gives the best results. Therefore, the final benchmark model is as follows:

$$Temperature = \beta_0 + \beta_1 \times Trend + \beta_2 \times Weekday \times Hour + \beta_3 \times Month \times Hour + \beta_4 \times Month \times Load + \beta_5 \times Month \times Load^2 + \beta_6 \times Month \times Load^3 + \beta_7 \times Hour \times Load + \beta_8 \times Hour \times Load^2 + \beta_9 \times Hour \times Load^3 \quad (4.1)$$

where *Month*, *Hour* and *Weekday* are class variables.

4.5 Recency Effects

Recency effect in psychology refers to a principle that it is most likely for human beings to remember the most recent experiences. Same analogy can be adopted to the power grid where the demand tends to be affected by the recent temperatures [8]. Lagged temperature terms were used in [1] for the first time in a load forecasting methodology. To study the Recency effect on the temperature prediction benchmark model, we use lagged load terms in the model. The proceeding load values were added to the model including their polynomials and their interactions with *Month* and *Hour*. In addition, we mentioned that the demand tends to be affected by the recent temperature. Therefore, in a load-based temperature prediction model, we can use the following load values to capture the Recency effect of temperature on the load demand. In the customized model, we will add lead-lagged load values to the model including their polynomials and their interactions with *Month* and *Hour*.

Table 4.5: Improving the benchmark model (M8) by adding lagged and lead-lagged variables

ID	**Model**	**MAE (^{o}F)**
M8	Benchmark	2.74
R1	Benchmark + $Load_{(t-1)}$	2.54
R2	Benchmark + $Load_{(t-1)}$ + $Load_{(t-2)}$	2.51
R3	Benchmark + $Load_{(t-1)}$ + $Load_{(t-2)}$+ $Load_{(t-3)}$	2.50
R4	R3 + $Load_{(t+1)}$	2.42
R5	R3 + $Load_{(t+1)}$ + $Load_{(t+2)}$	2.38
R6	R3 + $Load_{(t+1)}$ + $Load_{(t+2)}$+ $Load_{(t+3)}$	2.37

Table 4.5 shows the results of adding lagged and lead-lagged variables to the benchmark model. In each step one level of lagged variable, its polynomials, and the interactions with

hour and month are added to the benchmark model. Adding lagged and lead-lagged values of load variable for three preceding and three ahead hours (R6) improved the benchmark model 13%.

4.6 Holiday Effects

Generally, the load profile in holidays are not similar to the one in a regular day. There are different types of holidays with respect to which day of week they are falling on. Some holidays are in fixed weekdays. For instance, Memorial Day and Labor Day are on Monday. There are some holidays such as New Year's Day and Christmas Day, which are fixed date holidays and may fall into any day of the week. Depending on the situation of the holiday and features of location, the load profile of the holiday would be different. In addition, the surrounding days of the holidays are affected by the human behavior through holidays. To capture the holidays effects on the correlation of the temperature and load, a new variable, *Holiday*, is defined which is a class variable. Among all holidays in the United States, New Year's Day, Memorial Day, Independence Day, Labor Day, Thanksgiving Day and Christmas Day are called "big six" holidays which almost all companies observe them and they are closed.

The holiday effect in load forecasting models has been studied in the literature. For instance, the holidays were grouped together as a dummy variable along with weekday and weekend in [34]. The weekday holidays are grouped into Monday holidays, winter holidays and summer holidays in [35]. In some researches, the holidays are treated as weekends, for example Saturday [36].

In this section, we take only "big six" holidays to study the holiday effects on the improved version of the benchmark model (R6). The holidays are grouped together as a class variable. Each holiday is assigned to a number (from 1 to 6) and the other days of the year are zero. The interaction between the predictor variables of *Holiday*, *Month* and *Hour*

are added to the model to see whether accuracy of the model will be improved. Table 4.6 shows the results of holiday effects on the improved model. We can see that the holiday effects improved the accuracies where MAE of the holidays is improved by 5%.

Table 4.6: Holiday effect on the benchmark model (R6) for whole year and for the holidays

Model	MAE (Whole year)	MAE (Holidays)
R6	2.37	2.56
R6 + Holiday Effect	2.36	2.41

CHAPTER 5: ANOMALY DETECTION

The quality of the data has significant impact on the accuracy of the load forecasting outcomes. Having a high quality weather data and load history as the inputs to a load forecasting model, provides good and reliable predictions. When the input data is not clean and consistent, even if the model is strong enough, the output of the forecasting is not good. In the real world, as a result of various factors such as old instruments and meters, human errors and so forth, the data often includes errors and inconsistencies. Therefore, the preprocessing of the data is one of the first steps to prepare the data for the forecasting. Depending on the quality of the data, in many practices, the data is required to be cleansed. In this chapter, we discuss about the application of the proposed model in anomaly detection of the weather data.

Weather data as a major driver of load demand plays an essential role in load forecasting. The aged equipment of weather stations, human errors and failures in reporting procedures lead to the quality issues in the weather data such as outliers and missing values. To have a high quality data, we need to cleanse the data out of the anomalies, but first we should detect the outliers of the data. In this chapter, we propose a method for anomaly detection, which is one of the application of the proposed temperature prediction model.

5.1 Methodology

As discussed before, temperature follows a seasonal pattern during a day and through a year. The weather reports generally includes errors and inconsistencies, which can be detected visually. In this section, we introduce a brand new method for anomaly detection of temperature data.

The load-based temperature prediction model (LBTP) proposed in the chapter 4 estimates the value of the temperature using the load history data and calendar variables. These estimations can be used for detection of the outliers. The method is described as follows:

1- Implementing the LBTP model on the history data (In-sample fit) to have estimations for all the temperature observations
2- Calculating the absolute error (AE) using the reported temperature and the estimated value at each observation
3- Calculating the average (μ) and variance (σ) of the MAE values through the all observations
4- At each observation, if the absolute error of the predicted temperature is greater than ($\mu + \sigma$), the point is marked as an outlier.

The method is shown as a flow chart in Figure 5.1.

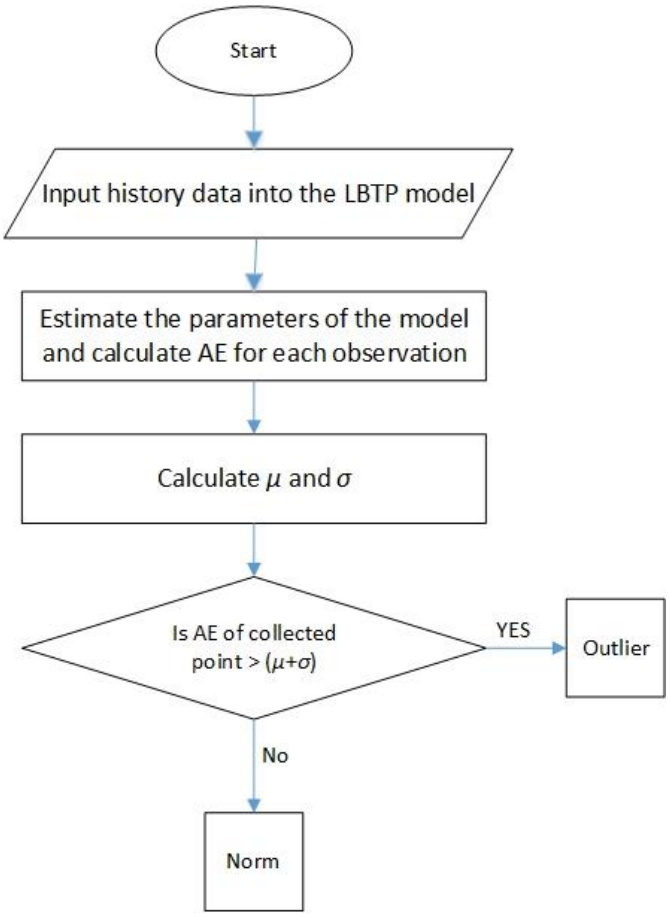

Figure 5.1: Flow chart of the proposed anomaly detection method

5.2 Performance Analysis

In this section, we study the performance of the proposed anomaly detection method. An anomaly simulation method is used to measure the detecting power of the proposed method. In addition, the quality of the data for before and after anomaly detection is measured by inputting them into a load forecasting practice.

The dataset similar to Chapter 4 is used in this section. The hourly load and the hourly temperature data of years 2008 and 2009 are used in training period and the year 2010 is the forecast window. For anomaly simulation, n% of the history (training) data is selected randomly with their temperature altered by multiplying with 1+k% to make these selected data points anomalies. Different levels of corruption were injected to the history data and in each level, the performance of anomaly detection method is studied. Two measures are used to study the performance of the method [37][38]:

- False Negative Rate (FNR): is defined as the ratio of the number of undetected anomalies to the number of all anomalies.
- False Positive Rate (FPR): is defined as the ratio of the number of normal points, which is detected as anomalies to the number of normal points.

The smaller amounts of FNR or FPR shows the effectiveness of the proposed method.

Different stages of corruption are injected to the data including 4 levels of sampling (n=10% to 40% of history) and in each sampling 8 levels of corruption are injected to the sample (k=5% to 40%). In each stage, the corrupted data is used as a history to forecast the

hourly loads of the next year and after the anomaly detection, the detected values are removed from the data and the cleansed data is used as the history for the load forecasting. The MAPEs for before and after cleansing is compared to evaluate the quality of the data.

Table 5.1: Error analysis of the anomaly detection method

				MAPE (%)	
n	k	FNR (%)	FPR (%)	Corrupted data	Cleansed data
10	5	0.80	0.06	6.07	6.06
	10	0.41	0.05	6.43	6.18
	15	0.23	0.05	6.72	6.18
	20	0.14	0.04	6.80	6.13
	25	0.08	0.03	6.95	6.11
	30	0.06	0.03	7.18	6.15
	35	0.05	0.02	7.31	6.11
	40	0.04	0.02	7.49	6.12
20	5	0.81	0.05	6.24	6.14
	10	0.46	0.04	6.89	6.24
	15	0.29	0.04	7.32	6.34
	20	0.21	0.03	7.63	6.22
	25	0.18	0.02	7.95	6.32
	30	0.14	0.02	8.21	6.23
	35	0.12	0.01	8.36	6.34
	40	0.11	0.01	8.52	6.29
30	5	0.81	0.05	6.41	6.24
	10	0.51	0.04	7.35	6.39
	15	0.35	0.03	7.94	6.56
	20	0.29	0.02	8.41	6.70
	25	0.25	0.02	8.87	6.90
	30	0.23	0.01	9.09	6.87
	35	0.20	0.01	9.34	6.85
	40	0.19	0.00	9.68	7.13
40	5	0.83	0.04	6.69	6.49
	10	0.57	0.03	7.88	7.00
	15	0.41	0.02	8.70	7.06
	20	0.36	0.02	9.18	7.10
	25	0.33	0.01	9.71	7.54
	30	0.30	0.01	10.20	7.98
	35	0.29	0.00	10.23	7.93
	40	0.28	0.00	10.63	8.24

Table 5.1 shows the error analysis of the anomaly detection method. We can see that in a given *n*, as *k* increases, FNR and FPR decreases which means the anomaly detection method works better in more corrupted datasets. In the other words, the methods can detect the bigger outliers better. In the most corrupted data (n=40% and k=40%), the detection method was able to detect 72% of the simulated anomalies. In addition, if we compare the MAPEs for before and after removing the detected anomalies form the corrupted data, we can see that the cleansed data produced more accurate forecasts in all levels of simulation (32 out of 32).

CHAPTER 6: ISO NEW ENGLAND CASE STUDY

In this chapter, we present a case study using another dataset to investigate the performance of the proposed model and method and to confirm that the model is not specific to only one dataset. For this reason, we use ISO-NE data, which is publicly available. The anomaly detection method will be applied to the dataset. To evaluate the quality of the data before and after removing the detected anomalies, the accuracy of the load forecasts is compared. In the following sections, first, we describe the ISO-NE dataset, then the results of the model implementation are presented, and finally we will discuss about the results.

6.1 Data

Two different datasets are used in the thesis: the first one is from GEFComp2014 which has been used in chapter 4 to develop the model and the second one is from ISO-NE which has been used in a case study as described later in this chapter. These two datasets cover load data of seven states in the United States. The datasets from GEFCom2014 and ISO-NE are publicly available and have been used in previous studies. The portion of the information that we use in the case study includes the hourly load demand in different zones, as well as the hourly temperature data obtained from various weather stations.

The dataset includes system load and weather data for the ISO New England Control Area (ISO NE CA) and its eight wholesale load zones. The load zones are shown in Figure 6.1. There are weather data from seven different weather stations in the ISO-NE territory. The names of weather stations are listed in Table 6.1 and their locations are shown in Figure 6.2. The data contains 14 years (2003 to 2016) of hourly load and hourly temperature data.

Table 6.1: ISO-NE weather station names and codes

Weather Station	Station Code	State
Boston	BOS	MA
Burlington	BTV	VT
Concord	CON	NH
Portland	PWM	ME
Providence	PVD	RI
Windsor Locks	BDL	CT
Worcester	ORH	MA

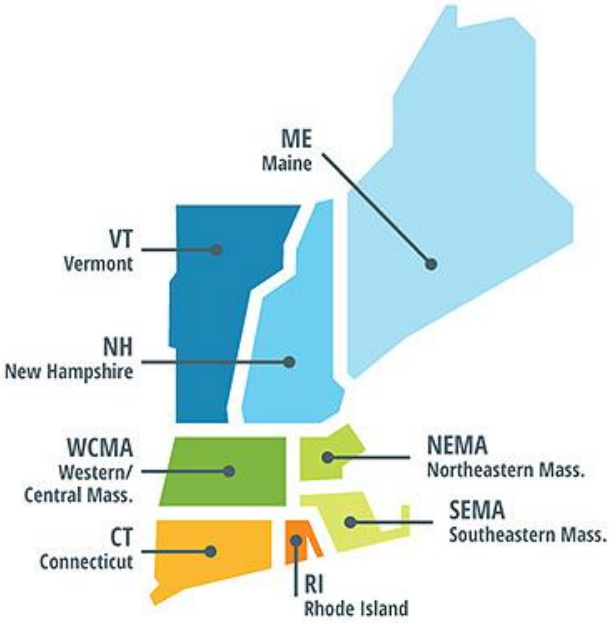

Figure 6.1: Wholesale load zones in New England [39]

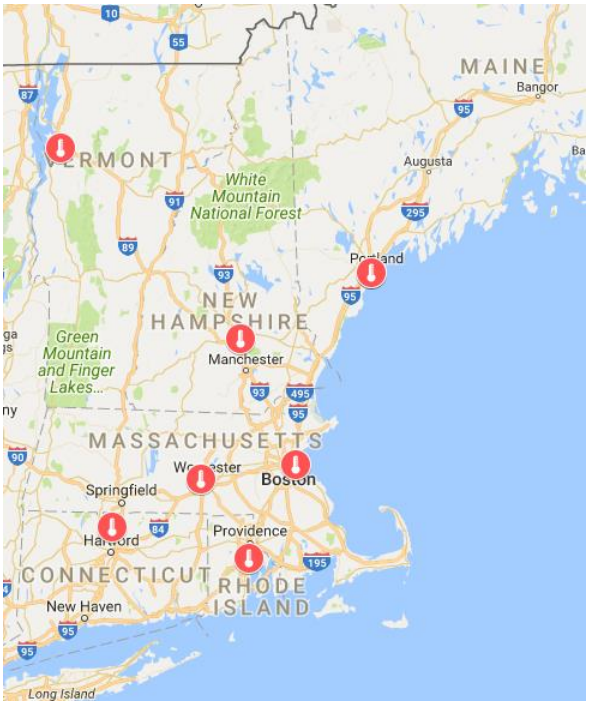

Figure 6.2: Weather station location in ISO-NE dataset

6.2 Temperature Prediction

In the proposed anomaly detection method, which was introduced in the Chapter 5, the first step is estimating the temperature by using the load-based model. Similar to the previous chapter, in this section we implement the load-based temperature prediction model on ISO-NE dataset. The dataset includes eight load zones, as well as seven weather stations. We take all pairs of load zone and weather station to make 56 different cases. Therefore, we will have 56 different temperature predictions. The history period is the hourly data of the years 2014 and 2015, and the forecasting window is the year 2016. Mean Absolute Error (MAE) is used as the measurement to evaluate the predictions.

Table 6.2: The MAE value of the temperature predictions for all 56 load zone/weather station pairs in ISO-NE.

Zone / WS	PWM	CON	BTV	BDL	PVD	ORH	BOS
ME	5.54	6.32	7.09	6.01	5.55	6.08	5.46
NH	4.54	4.97	5.88	4.54	4.6	4.74	
VT	5.57	5.9	5.71	5.26	5.24	5.46	5.5
CT	5.23	5.56	6.14	4.21	4.14	4.85	4.7
RI	5.1	5.8	6.6	4.6	4.29	4.97	4.92
SEMASS	4.87	5.56	6.3	4.62	4.02	4.86	4.41
WCMASS	5.13	5.58	6.08	4.49	4.72	4.89	4.95
NEMASSBOST	4.71	5.46	6.07	4.63	4.23	4.77	4.15

Table 6.2 shows the MAE values for each pair of load zone and weather station. The numbers are displayed as a heat map: the lower amounts, which are more accurate predictions, are marked green and the higher amounts are shown in red.

The load zones and weather stations are distributed in different locations of ISO-NE territory (Figure 6.1 and Figure 6.2). The heat map of Table 6.2 shows that there is a

best load data (from a specific load zone) for each weather station. For instance, the temperature data from weather station “BOS” can be estimated more accurately if we use the load data from “MEMASSBOST”. The distance of the weather station from the load zone, and the area of the load zone can affect the quality of the temperature predictions. For instance, all predictions of weather station BVT is not good. If we look at the weather station location in the Figure 6.2, we can see that this weather station is very far from all load zones except from VT and obviously the best prediction is for the pair of VT/BVT.

On the other hand, the load zone “ME” is not a good option to be paired with any of the weather station for the temperature prediction. The reason might be the large area of the load zone and wider distribution of the electricity consumers. The weather stations measure the temperature for a specific location. For the load zones with large areas, the correlation between the load and the temperature of a specific weather station is weaker than for the smaller zones. This weak correlation leads to less accurate predictions.

6.3 Anomaly Detection

In this section, the proposed anomaly detection method is implemented in ISO-NE dataset to study the performance of the method in another dataset. We used the data from years 2014 and 2015 as the history data of a load forecasting practice. The forecasting window is hourly load demand of the year 2016. ISO-NE dataset includes 8 load zones and 7 weather stations. We tried each pair of load zone and weather station. Therefore, we have 56 pairs in total. For each pair, the anomaly detection method is implemented. After detection, the anomalies are removed to make the data cleansed. The cleansed data is used for load forecasting. The accuracy of the load forecasts (in MAPE) is compared for before and after anomaly detection.

Table 6.3 shows the percentage of detected anomalies in the history data by the proposed method. Table 6.4 shows the MAPE numbers of load forecasts in all 56 pairs of load zone and weather station. The Figure 6.3 shows samples of detected anomalies in two different months of the history. The anomalies are determined by small blue circles.

Table 6.3: The percentage of the detected anomalies for all pairs of load zone vs. weather station of ISO-NE

Zone / WS	PWM	CON	BTV	BDL	PVD	ORH	BOS
ME	7.41	7.49	8.11	7.71	8.26	7.21	8.06
NH	5.50	5.91	7.69	6.66	6.72	5.99	3.40
VT	7.63	7.20	6.84	7.19	7.65	7.09	7.56
CT	7.32	6.73	8.19	5.64	6.79	6.27	7.48
RI	7.36	6.06	8.21	6.35	6.07	6.17	6.57
SEMASS	7.13	5.66	8.41	6.58	5.98	5.84	7.12
WCMASS	7.16	6.10	8.14	6.25	6.52	6.03	6.87
NEMASSBOST	7.14	5.89	8.45	6.31	6.32	5.70	6.56

Table 6.4: MAPEs of load forecasts for each pair of the ISO-NE. In each cell, the upper number is for raw data and the lower one is for the data, which the detected anomalies are removed.

Zone / Weather Station		PWM	CON	BTV	BDL	PVD	ORH	BOS
ME	Raw	3.749	3.779	4.513	4.196	3.951	3.871	3.989
	Removed	3.744	3.744	4.456	4.291	3.943	3.831	3.990
NH	Raw	3.918	3.455	4.583	3.606	3.852	3.357	4.023
	Removed	3.937	3.484	4.587	3.683	3.788	3.444	3.979
VT	Raw	5.142	5.003	4.854	4.774	5.160	4.751	5.188
	Removed	5.167	4.980	4.789	4.757	5.109	4.713	5.174
CT	Raw	6.361	5.505	6.858	4.881	4.881	4.931	6.189
	Removed	6.329	5.455	6.840	4.911	4.814	4.888	5.984
RI	Raw	5.334	4.858	6.165	4.247	3.877	4.162	4.885
	Removed	5.354	4.865	6.171	4.271	3.852	4.240	4.839
SEMASS	Raw	5.769	5.386	6.894	5.076	4.456	4.685	5.423
	Removed	5.810	5.350	6.830	5.085	4.418	4.740	5.372
WCMASS	Raw	5.380	4.726	5.818	4.310	4.756	4.490	5.179
	Removed	5.341	4.708	5.740	4.326	4.677	4.493	5.143
NEMASSBOST	Raw	4.831	4.502	6.055	4.624	4.044	3.794	4.325
	Removed	4.831	4.526	6.046	4.622	3.968	3.832	4.277

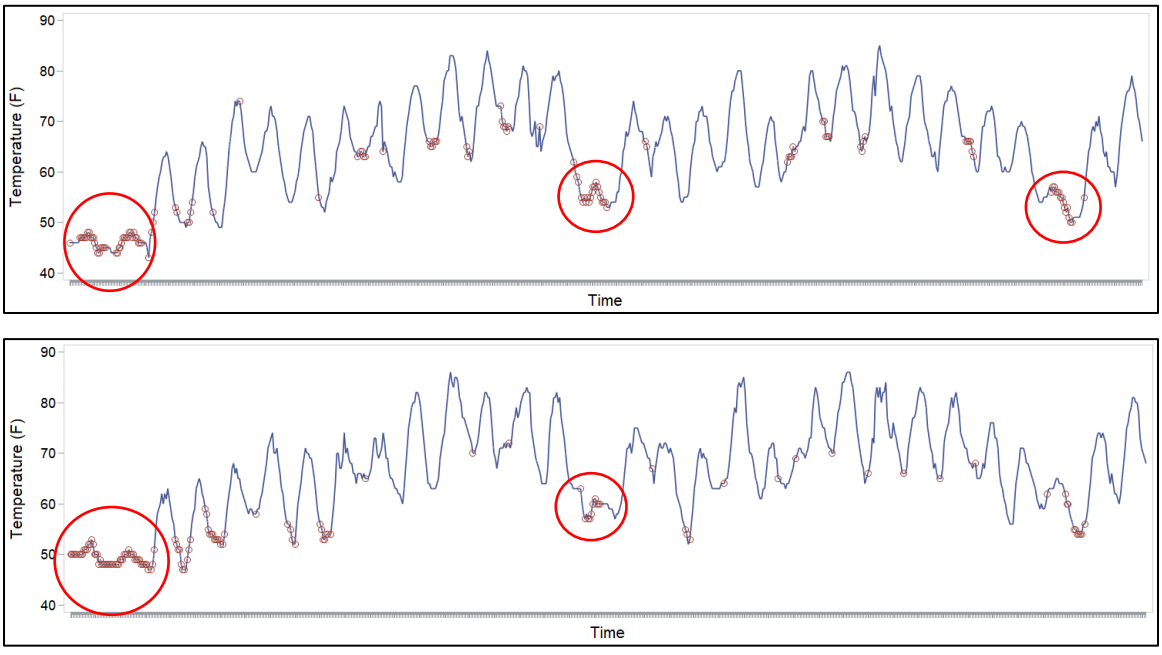

Figure 6.3: Samples of the detected anomalies

The following observations can be made from Table 6.3, Table 6.4 and Figure 6.3:

- The percentages of the detected anomalies show that the quality of the original dataset is reasonably high and we should not expect a big difference after removing the anomalies. Based on the results shown in chapter 5 (Table 5.1), the effectiveness of the method depends on how high the anomaly ratio is.
- The ratio of the detected anomalies in all pairs of weather station “BTV” with any load zone is high. Respectively in the corresponding attempts for load forecasting, the accuracy of the outcomes gets better after removing the detected anomalies in most cases (6 out of 8). This shows us that the quality of the data in weather station “BTV” is not good and the detection method works effectively and improves the quality of the data.
- Figure 6.3 shows two sample months of history data, where some of the anomalies are obvious. For instance, in some time intervals, the temperatures goes flat for a couple of hours in a row, which is not possible in real world. It might be because of an incorrect report, or a wrong procedure for replacing the missing values. The graph shows that the proposed method can detect these types of unusual behaviors.
- In overall, when we removed the detected anomalies from the raw data, the MAPE is decreased in 32 out of 56 cases. It shows that even with a high quality data, the detection method works effectively and finds the possible anomalies.

CHAPTER 7: CONCLUSION

Temperature is a driving factor in electric consumption and it has been utilized as a predictor variable in many load forecasting models. The quality of the input data plays a vital role in the performance of the load forecasting process. Since the real world data includes various types of anomalies and inconsistencies, the raw data requires to be prepared for using in a forecasting process. The anomalies are not specified only to the load data. Due to many reasons like human errors or aging instruments, the weather data includes anomalies such as missing data, outliers and inconsistencies.

This research proposes an anomaly detection method for the temperature data. The method consists of two components: a load-based temperature prediction model; and a detection technique. The prediction model is a regression-based model using the calendar variables, lagged and lead load values, and the interactions among them to estimate the temperature. The estimated temperature profile is used as a baseline or reference to detect the anomalies.

We developed a load-based temperature prediction model to build a foundation for anomaly detection. The temperature prediction is a regression-based model. We used calendar variables, load, lagged and lead values of the load and interaction among these variable to create a MLR model. The model is developed based on the data from Global Energy Forecasting Competition 2014 (GEFCom 2014). By analyzing different features on the basic model such as recency and holiday effect, the best model turned to be our proposed model.

The proposed anomaly detection method was evaluated through two case studies. The first one is based on the data from GEFCom 2014. In this case study, an anomaly simulation was used to study the performance of the proposed method. We created 32 scenarios to inject different levels of corruptions into the data. In each attempt, the detected anomalies were removed and both corrupted and cleansed data were used separately for load forecasting. The proposed method improved the accuracy of the load forecast for all 32 scenarios.

We conducted another case study based on the publicly available data from NE-ISO. In this case study, the method was implemented on the raw data. Based on the available data, 56 combinations of load zone and weather station were created. The proposed method improved the accuracy of the load forecast for 32 out of 56 pairs. The MAPE values were improved by about 1% on average.

The significance of this work lies in its interdisciplinary nature by building a bridge between meteorology and energy. During the past several decades, meteorologists have been analyzing weather data and developing numerical weather prediction models without taking electricity demand as an input. On the other hand, load forecasters have been taking weather history with the full faith without devoting much effort into data cleansing. The research demonstrates the benefits of marrying the two domains by using load data to validating temperature data.

The research conducted in this thesis opens a door to data cleansing for other weather variables used in load forecasting models, such as humidity and wind speed.

Temperature is not the only weather variable used in the load forecasting models. Other variables such as humidity and wind speed have some levels of correlation with the load. The similar approach can be implemented on the other weather variables to improve the quality of predictors and improving the forecast accuracies.

The introduced technique is used for the detection part of the cleansing process. The next step of the preprocessing is fixing these detected anomalies. An extension to this thesis would be developing a method to estimate appropriate temperature values for replacing the anomalies.

We assumed the proposed load-based temperature prediction model as a benchmark model. Improving the temperature prediction model leads to better estimations for the temperature and consequently enhances the power of anomaly detection. Another future work would be improving the temperature prediction model.

REFERENCES

[1] T. Hong, “Short Term Electric Load Forecasting.,” Sep. 2010.

[2] T. Hong and S. Fan, “Probabilistic electric load forecasting: A tutorial review,” *Int. J. Forecast.*, vol. 32, pp. 914–938, 2016.

[3] G. Gross and F. D. Galiana, “Short-term load forecasting,” *Proc. IEEE*, vol. 75, no. 12, pp. 1558–1573, 1987.

[4] H. K. Alfares and M. Nazeeruddin, “Electric load forecasting: Literature survey and classification of methods,” *Int. J. Syst. Sci.*, vol. 33, no. 1, pp. 23–34, Jan. 2002.

[5] H. Hahn, S. Meyer-Nieberg, and Pickl Stefan, “Electric load forecasting methods: Tools for decision making,” *Eur. J. Oper. Res.*, vol. 199, no. 3, pp. 902–907, Dec. 2009.

[6] R. J. Hyndman and S. Fan, “Density Forecasting for Long-Term Peak Electricity Demand,” *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 1142–1153, May 2010.

[7] T. Hong, P. Pinson, and S. Fan, “Global Energy Forecasting Competition 2012,” *Int. J. Forecast.*, vol. 30, no. 2, pp. 357–363, Apr. 2014.

[8] P. Wang, B. Liu, and T. Hong, “Electric load forecasting with recency effect: A big data approach,” *Int. J. Forecast.*, vol. 32, no. 3, pp. 585–597, Jul. 2016.

[9] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam, “ANNSTLF-Artificial Neural Network Short-Term Load Forecaster- generation three,” *IEEE Trans. Power Syst.*, vol. 13, no. 4, pp. 1413–1422, 1998.

[10] T. Hong and P. Wang, “Fuzzy interaction regression for short term load forecasting,” *Fuzzy Optim. Decis. Mak.*, vol. 13, no. 1, pp. 91–103, Mar. 2014.

[11] P. C. Abbe and P. C. Abbe, “THE PHYSICAL BASIS OF LONG-RANGE WEATHER FORECASTS1,” *http://dx.doi.org/10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2*, Dec. 1901.

[12] P. Bauer, A. Thorpe, and G. Brunet, “The quiet revolution of numerical weather prediction,” *Nature*, vol. 525, no. 7567, pp. 47–55, Sep. 2015.

[13] G. Shrivastava, S. Karmakar, M. K. Kowar, P. Guhathakurta, and S. Nagar, “Application of Artificial Neural Networks in Weather Forecasting: A Comprehensive Literature Review,” *Int. J. Comput. Appl.*, vol. 51, no. 18, pp. 975–8887, 2012.

[14] W. W. Hsieh, B. Tang, W. W. Hsieh, and B. Tang, “Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography,” *Bull. Am. Meteorol. Soc.*, vol. 79, no. 9, pp. 1855–1870, Sep. 1998.

[15] A. C. Lorenc, “Analysis methods for numerical weather prediction,” *Q. J. R. Meteorol. Soc.*, vol. 112, no. 474, pp. 1177–1194, Oct. 1986.

[16] M. J. C. Hu, H. E. Root, M. J. C. Hu, and H. E. Root, “An Adaptive Data Processing System for Weather Forecasting,” *J. Appl. Meteorol.*, vol. 3, no. 5, pp. 513–523, Oct. 1964.

[17] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[18] J. Xie and T. Hong, “GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation,” *Int. J. Forecast.*, vol. 32, pp. 1012–1016, 2016.

[19] T. Hong and S. Fan, “Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond,” *Int. J. Forecast.*, vol. 32, no. 3, pp. 896–913, Jul. 2016.

[20] H. N. Akouemo and R. J. Povinelli, “Probabilistic anomaly detection in natural gas time series data,” *Int. J. Forecast.*, vol. 32, no. 3, pp. 948–956, Jul. 2016.

[21] J. Luo, T. Hong, and S.-C. Fang, “Benchmarking robustness of load forecasting models under data integrity attacks,” *Int. J. Forecast.*, vol. 34, no. 1, pp. 89–104, Jan. 2018.

[22] C. Guan, P. B. Luh, L. D. Michel, Y. Wang, and P. B. Friedland, “Very Short-Term Load Forecasting: Wavelet Neural Networks With Data Pre-Filtering,” *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 30–41, Feb. 2013.

[23] J. N. Fidalgo and J. A. Lopes, “Load Forecasting Performance Enhancement When Facing Anomalous Events,” *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 408–415, Feb. 2005.

[24] P. Shamsollahi, K. W. Cheung, Quan Chen, and E. H. Germain, “A neural network based very short term load forecaster for the interim ISO New England electricity market system,” in *pica 2001. Innovative Computing for Power - Electric Energy Meets the Market. 22nd IEEE Power Engineering Society. International Conference on Power Industry Computer Applications (Cat. No.01CH37195)*, pp. 217–222.

[25] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th editio. 2005.

[26] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, “Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond,” *Int. J. Forecast.*, vol. 32, no. 3, pp. 896–913, Jul. 2016.

[27] T. Hong, P. Wang, and H. L. Willis, “A Naïve multiple linear regression benchmark for short term load forecasting,” in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–6.

[28] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection *,” *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[29] Shu Fan, K. Methaprayoon, and Wei-Jen Lee, “Multiregion Load Forecasting for System With Large Geographical Area,” *IEEE Trans. Ind. Appl.*, vol. 45, no. 4, pp. 1452–1459, Jul. 2009.

[30] Tao Hong, Min Gui, M. E. Baran, and H. L. Willis, “Modeling and forecasting hourly electric load by multiple linear regression with interactions,” in *IEEE PES General Meeting*, 2010, pp. 1–8.

[31] M. T. Hagan and S. M. Behr, “The Time Series Approach to Short Term Load Forecasting,” *IEEE Trans. Power Syst.*, vol. 2, no. 3, pp. 785–791, 1987.

[32] K. Methaprayoon, W.-J. Lee, S. Rasmiddatta, J. R. Liao, and R. J. Ross, “Multistage Artificial Neural Network Short-Term Load Forecasting Engine With Front-End Weather Forecast,” *IEEE Trans. Ind. Appl.*, vol. 43, no. 6, pp. 1410–1416, 2007.

[33] T. M. Peng, N. F. Hubele, and G. G. Karady, “Advancement in the application of neural networks for short-term load forecasting,” *IEEE Trans. Power Syst.*, vol. 7, no. 1, pp. 250–

257, 1992.

[34] B.-J. Chen, M.-W. Chang, and C.-J. Lin, “Load Forecasting Using Support Vector Machines: A Study on EUNITE Competition 2001.”

[35] A. Khotanzad, Rey-Chue Hwang, A. Abaye, and D. Maratukulam, “An adaptive modular artificial neural network hourly load forecaster and its implementation at electric utilities,” *IEEE Trans. Power Syst.*, vol. 10, no. 3, pp. 1716–1722, 1995.

[36] S. Rahman, “Formulation and analysis of a rule-based short-term load forecasting algorithm,” *Proc. IEEE*, vol. 78, no. 5, pp. 805–816, May 1990.

[37] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers,” 2003.

[38] A. Ashok, M. Govindarasu, and V. Ajjarapu, “Online Detection of Stealthy False Data Injection Attacks in Power System State Estimation,” *IEEE Trans. Smart Grid*, pp. 1–1, 2016.

[39] “About the Different Types of Costs and Reports.” [Online]. Available: https://www.iso-ne.com/markets-operations/market-performance/load-costs.