2013

Zonal And Regional Load Forecasting In The New England Wholesale Electricity Market: A Semiparametric Regression Approach

Jonathan Farland
University of Massachusetts Amherst

ZONAL AND REGIONAL LOAD FORECASTING IN THE NEW ENGLAND WHOLESALE ELECTRICITY MARKET: A SEMIPARAMETRIC REGRESSION APPROACH

A Thesis Presented

By

JONATHAN T. FARLAND

Submitted to the Graduate School of the University of Massachusetts Amherst
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

September 2013

Department of Resource Economics

ZONAL AND REGIONAL LOAD FORECASTING IN THE NEW ENGLAND WHOLESALE ELECTRICITY MARKET: A SEMIPARAMETRIC REGRESSION APPROACH.

A Thesis Presented

By

JONATHAN T. FARLAND

Approved as to style and content by:

_____

Bernard J. Morzuch, Chair

_____

Erin Baker, Member

_____

Daniel Lass, Member

_____

Daniel Lass, Department Chair

Resource Economics

DEDICATION

In my lifetime, I have yet to tackle a more challenging or exhausting task as completing this thesis. I would like to dedicate this work to my mother, who has never once shown any signs of giving up, no matter how hard life got for us.

ACKNOWLEDGMENTS

I was fortunate enough to have been surrounded by incredibly hard working, intelligent, and positive individuals while composing this thesis. The persistent encouragement, revision of my work and guidance provided by my advisor Bernie Morzuch particularly stands out. Bernie is a brilliant and tenacious man that became like a second father for me while working and studying under him.

I must thank Dan Lass and Erin Baker for the encouragement and education they provided. I know how difficult it was to review the first manifestations of this thesis, and I am glad they didn't give up on me right then and there.

Jonathan Black, David Ehrlich, Dan O'Connor, Wayne Coste, Gail Adams, as well as many others at ISO New England also contributed and made this research possible. However, the constant flow of questions that I have for them may not end with the completion of this thesis.

Most importantly, I have to thank my friends, colleagues and fellow graduate students who kept my head above water during this entire process. In particular, Avelino Amado, Karl Allen, Francesca Colantuoni, Chris Burns, Rodolfo Magno, Andrew Goggins, Bob Barron, Amritanshu Pandey, Ariel Newman, Ted Sullivan, Jess Rogean, Larissa Vangel, Jon Holt, Sarah Tuttle and Mike Griffin.

My dog Bailey also played an integral role in maintaining my sanity.

ABSTRACT

ZONAL AND REGIONAL LOAD FORECASTING IN THE NEW ENGLAND WHOLESALE ELECTRICITY MARKET: A SEMIPARAMETRIC REGRESSION APPROACH.

SEPTEMBER 2013

JONATHAN T. FARLAND

M.S., University of Massachusetts Amherst

Directed by: Professor Bernard Morzuch

Power system planning, reliability analysis and economically efficient capacity scheduling all rely heavily on electricity demand forecasting models. In the context of a deregulated wholesale electricity market, using scheduling a region's bulk electricity generation is inherently linked to future values of demand. Predictive models are used by municipalities and suppliers to bid into the day-ahead market and by utilities in order to arrange contractual interchanges among neighboring utilities. These numerical predictions are therefore pervasive in the energy industry.

This research seeks to develop a regression-based forecasting model. Specifically, electricity demand is modeled as a function of calendar effects, lagged demand effects, weather effects, and a stochastic disturbance. Variables such as temperature, wind speed, cloud cover and humidity are known to be among the strongest predictors of electricity

demand and as such are used as model inputs. It is well known, however, that the relationship between demand and weather can be highly nonlinear. Rather than assuming a linear functional form, the structural change in these relationships is explored. Those variables that indicate a nonlinear relationship with demand are accommodated with penalized splines in a semiparametric regression framework. The equivalence between penalized splines and the special case of a mixed model formulation allows for model estimation with currently available statistical packages such as R, STATA and SAS.

Historical data are available for the entire New England region as well as for the smaller zones that collectively make up the regional grid. As such, a secondary research objective of this thesis is to explore whether or not an aggregation of zonal forecasts might perform better than those produced from a single regional model. Prior to this research, neither the applicability of a semiparametric regression-based approach towards load forecasting nor the potential improvement in forecasting performance resulting from zonal load forecasting has been investigated for the New England wholesale electricity market.

Keywords: Semiparametric Regression, Load Forecasting, Penalized Splines, Mixed Models

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

In a simple electrical circuit, at least one source of power is connected to one or more sources of resistance (Soliman and Al-Kandari, 2010, Benjamin, 2013). The *electrical load* of the resistors on the circuit is defined as the amount of power, measured in watts, being drawn from that circuit or system at a given point in time. This term is synonymous with demand and is often used interchangeably. For example, a small LED light bulb connected to a battery represents a simple circuit. A windmill providing electricity to a remote household not connected to the local transmission infrastructure represents a larger-scale example. In fact, if that household were connected to a regional or local transmission grid, it would simply become a resistor itself in a much larger electrical system.

The United States electricity grid is an extremely complicated system involving interactions of economic agents, regulatory oversight, and physical and engineering constraints. Generating plants are the system's source of power and consumers represent the load(s). The infrastructure was originally designed and constructed entirely by individual monopolistic utility companies. Under this scheme, each vertically integrated utility was separately responsible for generating, transmitting and distributing electricity to its respective residential, commercial and industrial customers. Similar to the effect that sometimes results from having multiple authors compose a single paper, each with their own point of view, the U.S. electricity grid was composed of many different types of electricity networks.

The utility-specific segmentation of the electricity grid resulted in a system where fuel, resources and, most importantly, power were difficult to share among utilities. The systemic vulnerability of this design was made fully apparent during the Great Northeast Blackout of 1965. Since then, utilities have formed "power pools" to ensure regional system reliability. In the Northeast, the New England Power Pool (NEPOOL) was established in 1971 to facilitate collaboration among the utilities in Massachusetts, New Hampshire, Vermont, Rhode Island, Connecticut, and Maine. After three decades of development and operation, NEPOOL finally produced an electricity grid with its own system operator and sufficient generations to ensure that the New England region of the United States never again experiences a full system failure.

In addition to system reliability, economic factors prompted regulators to consider a competitive market design for the industry in view of the monopolistic alternative. Events such as the Oil Embargo by Saudi Arabia in 1973 and Three Mile Island in 1979 as well a general increase in inflation created a substantial surge in electricity prices. The formerly-adopted vertically integrated business model of supplying electricity provided little incentive to reduce prices, and as such Congress passed legislation to create an industry where suppliers must compete for customers.

In conjunction with governmental policy, the Federal Energy Regulatory Commission (FERC) was called upon to oversee the national electricity industry, and it began by restructuring the wholesale side of the industry. It encouraged states to mandate their utilities to sell generational capacity as a means to abolish regulator-enforced rates that had once been required in a monopolistic market and ideally to replace it with a purely

market driven price. FERC established several regional markets, each with its own portfolio of generation and equal access to transmission. In these areas of the country, the electricity grid is directly overseen and operated by an Independent System Operator (ISO).

ISO New England (ISONE) is an independent, not-for-profit corporation created in 1997 to accomplish three primary objectives: (1) manage the daily operation of the regional power grid, (2) develop and oversee a market for wholesale electricity generation, and (3) ensure a reliable source of electricity to the New England region through system and market planning[2]. It serves as a special type of ISO referred to as a Regional Transmission Operator (RTO). In this role, it is the system operator for the six-state region of New England. In the subsequent two years of its founding, ISONE completed the design and implementation of the region's first wholesale electricity market. Since its inception, renewed investments have allowed more than 1.3 GW of new generational capacity to be installed in the region, as well as a 2% increase in generator reliability. This market design allows generators to respond to economic incentives when demand is highest , e.g., during hot summer days. A centralized approach to managing power flow allows for scheduling required plant maintenance without concern of insufficient peak period generation. To improve system reliability and to mitigate price-volatility, ISONE administers Day-Ahead (DA) and Forward Capacity markets (FCM) as a means to efficiently schedule future supply and demand requirements. On the supply side, ISONE can call upon generating units that range from small 'peak-load' units, to medium-sized units used in the presence of a quick increase in demand, to large base-load units that are online nearly continuously. In total,

[2] http://www.iso-ne.com/aboutiso/co_profile/overview/index.html

ISONE has jurisdiction over 300 electricity generators within the region as well as ties among neighboring regional grids in New York and Canada.

The complexity of the regional grid has grown into a large, sophisticated and perhaps overly complicated electrical system. With the onset of competition in the market for electricity, generators and utilities are driven to decrease their costs and to streamline operational efficiencies. The reliable daily operation of the grid, as well as the financial and economic performance of suppliers (e.g., power plants) and demanders (e.g., utility companies) depend on understanding what future demand will be. As such, forecasts of electricity load are used by suppliers, municipalities, utilities and others within the electricity industry. Where short and medium term forecasts allow for the scheduling of sufficient generation, long term forecasts are used to predict demand in the face of changes in the economic, demographic or political landscape of the region. Everything from scheduling maintenance, making investment decisions, and establishing contractual fuel purchasing obligations is critically dependent on accurate demand forecasts. As such, load forecasting holds a central role in the operation of a regional power grid and is a well-researched topic in electrical engineering, mathematics and economics.

CHAPTER 2 LOAD FORECASTING

The act of generating predictions for future demand or energy usage of an electrical grid is referred to as *load forecasting* (Bunn and Farmer, 1985). Electric power system operators make use of these forecasts for the daily operation of a system, structural planning for the system (e.g., construction of new power plants), or to meet long term trends of demand requirements. Subsequently, load forecasting can be categorized by three different horizons:

1) Long-Term Load Forecasts – These predictions are made with lead times of a year or more. Their primary purpose is to accommodate the changes in economic and demographic environments that occur over long periods of time.
2) Medium-Term Load Forecasts – Power plant maintenance and fuel supply requires estimation of future load and energy usage between a week and a year out.
3) Short-Term Load Forecasts – Electrical system operations are critically dependent on predicting what the load on the system will be over the course of the next week with an immediate precedence for the next operating day.

We focus solely on the last of these: Short-Term Load Forecasting (STLF). The rest of this chapter describes the role of forecasting load in the modern energy industry. In sections 2.1 and 2.2, we discuss the role that short-term forecasts have in the operational and economic contexts of operating a regional power system. A review of load forecasting

models is given in section 2.3 and the specific procedures used in making day-ahead load forecasts at ISONE are discussed in section 2.4.

2.1 Resource Scheduling, Economic Dispatch, and System Security

As presented in Chapter 1, one of the principal responsibilities of ISONE is scheduling the mix of generators used to meet short-term demand. The term *Unit Commitment* often refers to the process of scheduling available capacity in advance to meet total system load at every moment of the day. This can become extremely complicated as start-up times, fuel availability, and operational and staffing constraints vary from power plant to power plant. For example, a large nuclear power plant takes a significant amount of time to begin generating electricity, while a small natural gas-fired turbine can start providing electricity almost immediately. For most plants, there are also additional fixed or "no-load" costs just to remain available should its capacity be required. Predictions of system load are required in order to (1) meet variable demand and (2) satisfy the operational constraints associated with scheduling a particular generating unit.

There are additional economic costs associated with scheduling power plants. The mix of generator type in a power system determines the economic sensitivity to prediction error in load forecasts. For systems primarily dependent on fossil-fueled thermal generation, expensive gas turbines would be required to satiate peak demand if the forecasted value of load ended up being far less than the actual value. Therefore, the economic costs associated with quick-firing peak units and the fixed costs of large steam or coal-firing units are directly proportional to prediction error. However, the presence of renewable, hydroelectric or pumped storage resources can alleviate the economic loss

resulting from errors in load forecasts. In particular, pumped storage is a technology that allows for extra capacity to be called on at times of peak demand rather than expensive gas turbines. Specifically, by pumping water into a reservoir during off peak hours when the price of electricity is low, this technology actually demands electricity from the grid before supplying it. The reservoir is typically on top of a mountain or hill where the water can be stored and then released through openings in the bottom of the reservoir. The water can fall unaided to pass through hydroelectric turbines in order to generate power in times of need. In the context of a mixed thermal and hydro system, the economic dispatch is determined by the costs associated the last thermal generator to be dispatched (Bunn and Farmer, 1985).

Load dispatch and *economic dispatch* are terms used to denote the process of minimizing the total cost of meeting the demand while maintaining the security of the system. These terms can be used interchangeably. While scheduling the operation of power plants can be done for the next day and up to a week out, dispatch is done on a minute-to-minute basis in order to satisfy demand. This near real-time activity is often referred to as "online" while longer term scheduling can be done "offline". Subsequently, economic dispatch requires comparable "online" forecasting methods. These have typically been limited to time-series or adaptive methods that use the most recent observations of demand in order to track the variation in demand. While the errors in offline plant scheduling often result in significant economic losses, dispatch errors typically only disrupt control of system frequencies and can cause excessive stress on system infrastructure from the rapid changes in which generators are dispatched to serve load.

Short term forecasts of load are also required to supply reliable electrical power by ensuring system security and infrastructure maintenance. In a regional electrical grid, overloading transmission lines with more electricity than they are rated for can cause system imbalances and eventually system failure. Forecasting load at the geographical supply points, where electricity is physically generated, allows for scheduling plants so as not to overload the local high-powered transmission lines.

In addition to the physical limitations of an electrical system, allocation of a *reserve capacity* is also necessary to guarantee reliable electricity generation and distribution. Supply interruptions, or *loss-of-load events*, can cause serious harm to the infrastructure of a large scale electricity system; the reserve capacity provides for a continual buffer against such events. However, determining the appropriate amount of available generation to allocate for reserve capacity requires knowledge of what system load will be like in the short term. Continual overestimation of required reserve capacity can lead to economic losses that stem from scheduling unnecessary plant availability.

Clearly, accurate short term load forecasts are essential for maintaining the day-to-day operation of an electricity grid. They are required to mitigate potential financial and economic losses in a competitive wholesale electricity market as well as to ensure reliability.

2.2 Review of Short Term Load Forecasting Models

There have been many approaches toward developing highly accurate load forecasting models. Surveys and reviews of this expansive literature can be found in Matthewman and

Nicholson (1968), Bunn and Farmer (1985), Gross and Galiana (1987), Alfares and Nazeeruddin (2002) and more recently by Soliman and Al-Kandari (2010). Even when restricted to short-term applications, the variety of methodologies applied to forecasting electricity load is exhaustive.

For simplicity, we classify these approaches into two major categories: conventional and artificial intelligence (AI). Load forecasting is a required task for the operation of any electrical grid and has been for many years. However, there has been a recent surge of applications that use artificial neural networks and other AI-based methods applied to load forecasting. For AI-based and conventional methods, a list of major approaches is provided below, and each is described.

2.2.1 Conventional Methods

2.2.1.1 Multiple Regression

Regression is a statistical and econometric technique used to explain relationships between independent and dependent variables as well as make predictions of the latter. Its adoption to forecasting load is among the earliest of any method. Regression-based load forecasting models analyze the statistical relationship between total load and weather conditions as well as time-of-year effects (Alfares et al., 2002). As such, multiple regression models benefit from incorporating additional predictor variables as opposed to the univariate methods described in sections 2.2.1.2 and 2.2.1.3. Section 3.2 goes into detail regarding regression methodology; as such, its discussion is limited here.

Many studies have successfully used regression to produce forecasts of electricity demand. In addition to the current context of day-ahead forecasting, regression models have been used to predict peak-period electricity demand, probability density functions of load, and medium- to long-term forecasts of energy consumption. Adams et al., (1991) provide forecast distributions of weekly peak load using nonparametric simulation and three separate regression models. These regressions involve forecasting weekly peak load based on trend, socioeconomic and weather indicators. A daily peak load model is also used in conjunction with weekly models. Engle et al., (1986) use smoothing splines within a regression model to estimate the functional relationship between weather and load. This semiparametric approach revealed features and relationships that were not clear in other parametric approaches. Fan and Hyndman (2011) use a similar semiparametric approach but focus on providing forecasts rather than simply estimating the functional relationships. Other well-known applications include Bernard and Veal (1987), Heineman et al., (1966), Corpening (1973), and more recently, Hippert et al., (2001). A thorough set of references regarding regression-based load forecasting is contained in Soliman and Al-Kandari (2010).

2.2.1.2 Exponential Smoothing

While regression can be used to explain relationships between different variables, univariate methods focus on the relationship among the observations within the series itself. Time series can have four components: level, trend, seasonal and cyclical. An example of a widely used univariate approach to load forecasting is exponential smoothing. Using one or more smoothing equations, this method fits past observations of itself in

order to predict future values. As there may be several possible components in a time series (e.g., seasonal, trend, and cyclical) there are several corresponding exponential smoothing methods. For instance, simple exponential smoothing addresses only changes in level while Holt's method can be used to accommodate changes in level and trend. The Holt-Winters (HW) method accomplishes the same as the Holt method, but is also applicable when seasonal patterns are present as well. (Alfares et al., 2002; Gelper, et al., 2010). The HW method, also referred to as double-exponential smoothing, is a simple and recursive method originally introduced in Holt (1959) and Winters (1960). While this methodology is used in load forecasting, it has the disadvantage of neglecting the relationship between load and weather. Because patterns of load, weather and the load-weather relationship are constantly changing, the HW method is often characterized by poor forecasting performance at longer lead times. However, this approach can incorporate the most recent observations of load and therefore performs well for online forecasting but not for day-ahead forecasting (Soliman and Al-Kandari, 2010). Recently, Hyndman et al. (2005) have explored the statistical properties of the Holt-Winters methodology as applied to load forecasting while El-keib et al. (1995) and Infield et al. (1998) have successfully developed hybrid models which incorporate adaptive and time series methods in conjunction with exponential smoothing.

2.2.1.3 Stochastic Time Series

Another approach in univariate load forecasting treats the load series as a purely stochastic variable. Time series approaches are a broad category of forecasting methods where some internal structure, which may incorporate seasonality or trend, is assumed

and requires estimation. The Box and Jenkins approach (BJ) has been widely used and involves first identifying the internal structure and then the estimation of the structural components. For instance, the autoregressive (AR) model assumes that electricity load can be expressed solely as a linear combination of previous loads. The autoregressive moving-average (ARMA) model extends the AR model to include the disturbances from previous periods into the model as well. Both the AR and ARMA approaches have been successfully applied to load forecasting, but they require the condition of stationarity in the load series in order to assure validity of the forecasts. A stationary process is one where its first and second moments (e.g., mean and variance) and covariances (i.e., time displacements) remain constant over time. In the presence of nonstationarity, a transformation (e.g., first-differencing or logarithmic) is required. The transformation is referred to as "integrating" the series. In turn, the autoregressive integrated moving-average (ARIMA) model can be employed for forecasting.

Barakat et al. (1992) used an ARIMA model to identify the stochastic components of monthly peak demand and then, with additional deterministic components, produce forecasts. Jubieras et al. (1999) was able to demonstrate the union of ARIMA models and weather predictions to produce an online load forecasting model. Time series models such as this are referred to as autoregressive moving-average models with exogenous variable (ARMAX). Where other time series models are limited by the availability of past observations, ARMAX models have had success in load forecasting applications given the clear presence of load weather relationships. In other words, ARMAX represents a possible means of addressing the presence of load-weather relationships because it can include an

exogenous covariate such as weather. Other examples of load forecasting using time-series models include Liu (1996), Zhao et al. (1997), and Huang (1997).

2.2.1.4 Similar Day

Rather than a purely mathematical model, the similar day approach matches the currently forecasted conditions with observed historical data. This forecasting method is simple, intuitive and is often used for benchmarking and model forecast comparison (Chen et al., 2010). It also allows forecasters to make manual adjustments based on experience and intuition. ISONE currently employs a similar day approach as one of its short term load forecasting models. This is described in detail in Section 2.3.

Recently, Chen et al., (2010) have combined a similar day approach with wavelet decomposition and neural networks using data from ISONE. Where the similar day methodology by itself may be too simple to capture the complex load relationships, Chen et al. (2010) use similar day load as inputs for a hybrid forecasting model . This approach provided accurate forecasts across different forecasting periods and using a variety of weather inputs. Mu et al. (2010) demonstrate further forecasting improvement by weighting the most similar days more than others.

2.2.2 Artificial Intelligence

2.2.2.1 Neural Networks

Artificial neural networks (ANNs) have received a great deal of attention in load forecasting studies conducted over the last decade. These complicated models are designed to simulate the mechanics of the human brain and are characterized by their ability to learn

the load-weather relationships as well as the relationships within the load series itself. ANNs can be trained using historical data and can be employed even in the presence of nonlinearity. This characteristic has made ANNs a popular choice for short-term load forecasting.

The structure of neural networks can vary greatly and depends on architectural choices such as the number of neurons, the training approach and the function used for each neuron. For instance, Fan and Hyndman (2010) use a three-layer feed-forward network and use the Levenburg-Marquardt approach to train the network as a benchmark to their semiparametric regression approach. There have been a great deal of ANN-based approaches used in load forecasting with accurate predictions reported (Ferreira and Alves da Silva, 2007; Yun et al., 2008; Amjady, 2006). A typical neural network approach to load forecasting uses predicted weather, the most recent observations of load, and a weekday indicator variable to make day-ahead predictions of load (Chen et al., 2010). Hippert et al.(2001) note that, while ANNs may have many advantages to forecasting hourly load, the ANN models in the load forecasting literature are generally poorly validated. Hippert and Pedreira (2004) also state that ANNs have the potential to become heavily over-fitted and their performance is not universally accepted.

2.2.2.2 Fuzzy Logic

The term "fuzzy logic" is used to denote a logical system where conditional statements (such as IF-THEN clauses) are allowed to be approximate rather than exact. For example, rather than using binary indicator variables to indicate TRUE or FALSE, fuzzy logic variables can have a 'truth value' ranging between 0 and 1. In this way, the degree of 'truth'

is enumerated. It has been shown that this flexibility allows artificial intelligence based fuzzy logic systems to approximate any dynamic system of (statistical) relationships (Liu et al., 1996).

Given that the relationships between load and its predictors are changing over time, fuzzy logic has been applied to load forecasting by treating electricity demand as a dynamic system. This is not a statistical approach. Rather, it is widely used in solving robust linear programming or comparable optimization problems. However, fuzzy logic has been successfully used in conjunction with purely statistical (regression) models as well as neural networks in numerous load forecasting applications (Srinivasan et al., 1992; Dash et al., 1995; Chow et al., 1998). Notably, Srinivasan et al. (1999) created an autonomous approach toward short term load forecasting using a combination of fuzzy logic, neural networks and expert systems (Alfares et al., 2002).

2.2.2.3 Expert Systems

Knowledge-based expert systems (ES) are among the most recent applications of artificial intelligence to forecast electricity demand. The idea behind ES is to capture the task-specific expertise of a human and transfer it to a computer. As such, an "expert system" is a computer with the ability to learn, reason and give advice (Liao, 2004). An expert system is designed by a "knowledge engineer" who extracts knowledge from load forecasting experts to build a central knowledge component of the system. Similar to fuzzy logic, these are stored as IF-THEN clauses which can be used to establish relationships between the changes in load and changes in the factors that drive electricity demand.

Typical arguments in an ES load forecasting system include season, day of the week, temperature and change in temperature (Alfares et al., 2002).

For short-term load forecasting, Rahman and Hazim (1996) developed an expert system that performed extremely well and was not dependent upon location. Other ES load forecasting models include Brown et al. (1999) and Ho et al. (1990). Kim et al., (1995) were able to use a neural network to produce an initial set of hourly load predictions which were then adjusted by a second-stage fuzzy expert system as a means to accommodate the presence of holidays and changes in temperature. Other hybrid models have incorporated expert systems alongside other AI-based and statistical methods alike. (Rahman and Shreshta, 1991; Mohamad et al., 1996; Chandrashekara et al., 1999).

2.3 Day Ahead Forecasting at ISO New England

For every hour of the day, there is a forecaster assigned to the ISONE control room. The main responsibility of this position is to develop an hourly demand forecast for the next operational day as well as the next six days. Specifically, the day-ahead demand forecast must be completed and published before 10:00 am. These forecasts fall into the classification as "offline" as they are not continually updated for dispatch purposes but rather for day-ahead unit commitment. As ISONE is required to provide this forecast before 10:00 am, the day-ahead forecasting horizon is 36 hours ahead[3].

The forecaster begins each short term forecast by accumulating weather forecasts for the eight New England cities provided by three separate weather vendors[4]. These weather data are analyzed and validated against past performances and other sources such as the National Weather Service. Each vendor's provided weather forecast is visually inspected for five different variables: temperature, dew point, wind speed, cloud cover and precipitation intensity. Once validated, the weather forecasts for each vendor are aggregated to a single regional weather forecast.

ISONE is provided with both actual and forecasted weather for weather stations corresponding to each load zone. However, these stations are chosen based solely on their geographical proximity to the load zone and therefore do not represent the actual weather that occurs over the entire load zone. ISONE has not undergone any analysis to determine

[3] http://www.iso-ne.com/rules_proceds/operating/sysop/out_sched/sop_outsch_0040_0010.pdf
[4] http://www.iso-ne.com/rules_proceds/operating/sysop/rt_mkts/sop_rtmkts_0050_0030.pdf

the appropriateness of these weather station locations for each load zone. These data are not used in any other analysis done at ISONE, nor is there a forecast made for each load zone.[5]

These predictions of short-term weather conditions are then used as inputs for forecasting models available to ISONE. Expert knowledge and experience are blended with numerical predictions of demand to produce a short term load forecast for the New England region. Currently, the ISONE control room forecaster employs three separate modeling approaches for short term load forecasting. These consist of the following:

1) **SimDay**: The similar day approach allows the forecaster to review a range of historical daily load shapes and their corresponding weather conditions. The forecaster may choose criteria such as how many years back, which days of the year, and the percent deviation of actual weather from forecasted weather. These predetermined criteria limit the search, and five historical days are eventually selected. The forecaster then has the option to adjust for historical energy growth and even to manually enter specified hourly weights as a means to blend each hour into a single 'similar-day' load curve.
2) **MetrixND** (Metrix Next Day): This model uses only weather inputs to produce load forecasts. Specifically, it uses effective temperature (EFF) during the heating months of October through April and a temperature-humidity index (THI) during the cooling months of May through September. MetrixND is a product offered by

[5] Note: This is stated within the data dictionary of the historical data sets ISONE publicly provides.

ITRON, a leading technology company in the energy and water industries.[6] While the software has the capability of producing load forecasts by means of exponential smoothing, ARIMA, regression, and neural networks, ISONE utilizes the last of these.

3) **ANN (**Artificial Neural Network**):** Four different ANNs are used by ISONE to produce short term load forecasts. Two of these ANNs are "fast" learners which weight the most recent demand and weather more than earlier observations. The other two ANNs are "regular" learners and weigh all observations evenly. All four models are retrained on an annual basis. Similar to the MetrixND model, all four ANNs use EFF during heating months and THI during the summer months.

Both the MetrixND and ANN models are used to create day-ahead forecasts, the seven-day forecast, and to update the current day load forecast.[7] In the context of the day-ahead forecasts, the forecasts produced by both the ANN and MetrixND models are then combined by the forecaster with the Simday load shape to produce a single regional forecast for the next 36 hours. Regression models are not currently used for short-term load forecasting at ISONE.

[6] https://www.itron.com/na/productsAndServices/pages/MetrixND.aspx
[7] http://www.ferc.gov/industries/electric/indus-act/rto/metrics/iso-ne-rto-metrics.pdf

CHAPTER 3 MODELING APPROACH

3.1 Overview

Classical linear regression has many applications. However, it is based on assumptions that are usually very restrictive. One particular difficulty relates to situations where the functional relationship may be nonlinear. Semiparametric regression models offer a bridge between classical linear regression models and those that assume no specific functional form. The latter approach is generally referred to as nonparametric and is determined purely by sample data. While linear regression is fully capable of addressing nonlinearities, there are instances where linear regression may not provide a suitable fit or assumes an inappropriate functional form.

The proposed methodology here focuses on the complex relationships that exist between electricity demand and its driving forces. Particular attention is paid to the functional form of the relationship between weather and demand as it is often highly nonlinear. The nonparametric approach suggested here uses a regression framework but permits adaptations (referred to as nonparametric terms) to enter the model specification. The resulting fitted model retains the parsimony of a linear model while simultaneously relaxing some of the underlying assumptions of the classical linear model.

The theoretical research that underlies the mechanisms of semiparametric regression is not a recent development. For instance, smoothing techniques using splines

have long been used to accommodate apparent nonlinearities in observed data (Craven and Wahba 1979, De Boor 1978, Eubank 1988). However, the ability to incorporate these ideas into a single conceptual framework is a fairly recent development. In 2003, Ruppert, Wand and Carrol published the book *Semiparametric Regression*, the first to formally show the connection between penalized splines and mixed-effects analysis. Smoothing in a mixed-model setting allows for well-established estimation procedures such as maximum likelihood and best prediction. Mixed-model software such as SAS and R conveniently performs the estimation. This text is considered a comprehensive treatment of the semiparametric methodology (Ruppert, et al., 2009).

Since 2003, research into the use of semiparametric modeling has exploded; a 2009 review of progress in this field, again by Ruppert, et al., reported successful applications in on-line auctions, genomics, air pollution, agriculture and even cosmology. These studies have also shown that nonparametric components can be accurately modeled relatively simply using low-rank smoothing splines. In view of the numerous complicated and highly sophisticated models that have gained popularity in load forecasting applications, we seek a simpler approach.

This chapter begins by briefly reviewing classical linear regression and provides examples of when parametric and nonparametric methods may be more appropriate, respectively. Section 3.2 describes curve fitting with splines and focuses on penalized splines as an attractive method for scatter plot smoothing. Section 3.3 shows how penalized spline models are equivalent to a mixed model while section 3.4 concludes by

including additional parametric and nonparametric terms in an additive semiparametric model.

3.2 Classical Linear Regression

Given data points (x_i, y_i), where $i = 1, \dots, n$, consider the following bivariate statistical model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

where x_i and y_i are observed variables in period i, β_0 and β_1 are population parameters and ε_i is a disturbance that is $i.i.d.(0, \sigma^2)$. This is an example of a simple linear regression model. Equation (1) is simple because there is only one explanatory variable. Also, the model is linear in the parameters. This means that the exponent on each parameter is one and no parameter is multiplied by another parameter.

Note that the only other term used to capture variation in the response variable is the random or *stochastic* disturbance ε_i. As this term is assumed to have mean zero and constant variance σ^2, taking expectations of Equation (1) leads to the expected mean of y_i conditioned on the observed value of x_i. Specifically,

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i \quad (2)$$

Equation (2) asserts that there is a linear relationship between the observed response variable y_i and the observed predictor variable x_i. That linear relationship could be estimated with a (linear) model if values for the parameters in Equation (2) were estimated. This estimated model can be expressed as

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i \quad (3)$$

where the error at time period *i* is given by

$$y_i - \hat{y}_i = e_i \tag{4}$$

Estimates of the parameters β_0 and β_1 are obtained by minimizing the sum of squared errors defined as

$$\sum_{i=1}^{n} \left(y_i - [\widehat{\beta_0} + \widehat{\beta_1} x_i]\right)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

The parameter estimates that minimize Equation (5) are found by taking the first partial derivatives with respect to both β_0 and β_1, respectively. Setting these derivatives to zero and solving forms the necessary conditions for an optimum. Checking the second partial derivatives guarantees a minimum. Simultaneously solving yields estimates of the population parameters β_0 and β_1. If this effort is successful, then population parameter estimates have been found that make the sum of squared errors as small as possible. *Ordinary Least Squares* (OLS) or just *least squares* is a method that can provide these parameter estimates.

This method can be applied when simple linear regression is extended to include more predictor variables. The multivariate linear statistical model with m predictors is given by the following equation:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_m x_{m,i} + \varepsilon_i \tag{6}$$

where $x_{m,i}$ is an observation of explanatory variable *m* at time period *i* and β_m is a scalar valued regression coefficient associated with explanatory variable *m.* Equation (6) is referred to as the population regression function while its fitted counterpart is referred to

as the sample regression function. As more predictors are included in a regression model, matrix notation is often used to compactly express these equations.

We can express Equation (1) or Equation (6) with more general matrix notation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_i \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{bmatrix}, \boldsymbol{\beta} = [\beta_0 \ \ \beta_1 \ \dots \ \beta_m]^{\mathrm{T}}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{8}$$

Here, $\boldsymbol{X}$ is an $(n \times [m+1])$ *design matrix* where each row corresponds to an observation and each column after the first corresponds to predictor variable $j = 1, \dots, m$. While there are *m* predictors, the first column of $\boldsymbol{X}$ is an $(n \times 1)$ vector of ones corresponding to a single intercept parameter. Hence, the number of columns in $\boldsymbol{X}$ is $m + 1$. It is clear now that simple regression is a special case of multivariate regression where *m*=1 predictor. Both $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are $(n \times 1)$ column vectors denoting observations of the response variable and the disturbance, respectively. Finally, $\boldsymbol{\beta}$ is a $([m+1] \times 1)$ vector of regression coefficients where each entry is a coefficient corresponding to the *j*th predictor variable in $\boldsymbol{X}$. From here on, matrices and vectors are denoted with bold notation.

For the linear regression model in matrix notation, the sum of squared residuals is given by the following equation:

$$\boldsymbol{S} \equiv (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \tag{9}$$

Its expansion leads to

$$= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{10}$$

Ordinary least squares corresponds with taking this quantity and treating it in the context of a minimization. It minimizes by taking the partial derivative of Equation (10) with respect to $\widehat{\boldsymbol{\beta}}$ and setting it equal to zero to yield the first order necessary conditions (FONC) for an optimum as follows:

$$\text{FONC} \qquad \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}} = 0 \tag{11}$$

$$-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0 \tag{12}$$

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{13}$$

Pre-multiplying both sides of Equation (13) by $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ yields the normal equations:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{14}$$

These equations can now be solved for the vector of parameters that minimizes the residual sum of squares:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{15}$$

Equation (15) is the least squares estimator for the ($[m + 1] \times 1$) vector of regression coefficients. If the goal of the linear regression model is to explain marginal effects, then inference can made regarding the entries of the coefficient vector. If prediction is the goal, then the least squares coefficients along with observations of explanatory variables can be

used to make predictions of the dependent variable. The fitted values of the response variable can be determined with the following calculation:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{16}$$

where $\widehat{\boldsymbol{y}}$ is an $(n \times 1)$ vector of fitted or predicted values of the response variable. We note here that an interesting result occurs when Equation (15) is substituted into Equation (16) :

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \tag{17}$$

Here, $\boldsymbol{A}$ is referred to as the *hat* or *smoothing* matrix as it puts a 'hat' on the response variable. Equation (17) shows that the relationship between $\widehat{\boldsymbol{y}}$ and y is a straightforward linear transformation.

3.3 Assumptions of Linear Regression

Linear regression can be a useful tool for establishing functional relationships among variables and for predicting behavior. The model itself – referred to as the classical linear regression model (CLRM) – follows a very strong set of assumptions. The assumptions of the CLRM are as follows:

1) Linearity: The specified population regression function is the *true* data generating process and is *linear* in its parameters.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{ki} + \varepsilon_i \tag{18}$$

None of the parameters is raised to a power or multiplied by another parameter.

2) Strict Exogeneity of the Explanatory Variables:

$$E[\varepsilon_i|x_i, \dots, x_n] = 0 \quad (19)$$

Equation (19) states that the expected value of the residual at time period *i* is not a function of the explanatory variables at *any* time period. In essence, this assumption defines something called strict exogeneity and requires that the predictor variables are nonrandom and are uncorrelated with the disturbances at any time period. Notice that Equation (19) is a conditional expectation. Furthermore, this implies that the unconditional mean of the disturbance is zero as well; i.e., $E[\varepsilon_i] = 0$. In one statement, we have:

$$E\big[E[\varepsilon_i|x_i, \dots, x_n]\big] = E[\varepsilon_i] = 0 \quad (20)$$

$$cov(\varepsilon_i x_i) = 0, \quad \forall i = 1, \dots, n \quad (21)$$

3) No Multicollinearity: This requires that there is no perfect linear correlation among explanatory variables in the design matrix; i.e., the columns of X must be linearly independent of each other. Absence of *perfect* correlation results in full rank, where the rank of **X** is defined to the number of linearly independent columns m+1. If this assumption is violated, multicollinearity (or just collinearity) is the result. It leads to inflated standard errors on the regression coefficients. This has a negative effect on hypothesis tests and confidence interval precision.
4) Constant Variance (Homoscedasticity): The variance of the stochastic disturbances is a constant over all observations.

$$E[\varepsilon_i{}^2|x_i] = \sigma^2 \quad (22)$$

Nonconstant variance results in heteroscedasticity. Its negative consequence is to bias the standard errors on the regression coefficients.

5) No Serial Correlation among the Disturbances:

$$E\left[\varepsilon_i \varepsilon_j \middle| x_i\right] = 0, \quad \forall i \neq j \tag{23}$$

Correlation of the disturbances is known as autocorrelation. Its negative consequence is to bias the standard errors on the regression coefficients.

The method of least squares does not provide unbiased and consistent parameter estimates when its underlying assumptions are violated. When the assumptions of the CLRM do hold, however, the Gauss-Markov Theorem guarantees that the resulting estimators are best, linear, and unbiased (BLU).

3.4 Applications of Linear Regression

To illustrate a simple application of parametric regression, we use a data set containing 2,208 hourly observations of temperature and electricity demand in New England. These observations are limited to the months of June, July and August for 2011. Here, electricity demand (i.e., load) is measured in Megawatts (MW). Drybulb temperature in degrees Fahrenheit is a measurement of air temperature using a thermometer that is feely exposed to the air while being shielded from moisture. This method of measuring temperature is used as a proxy for the true air temperature.

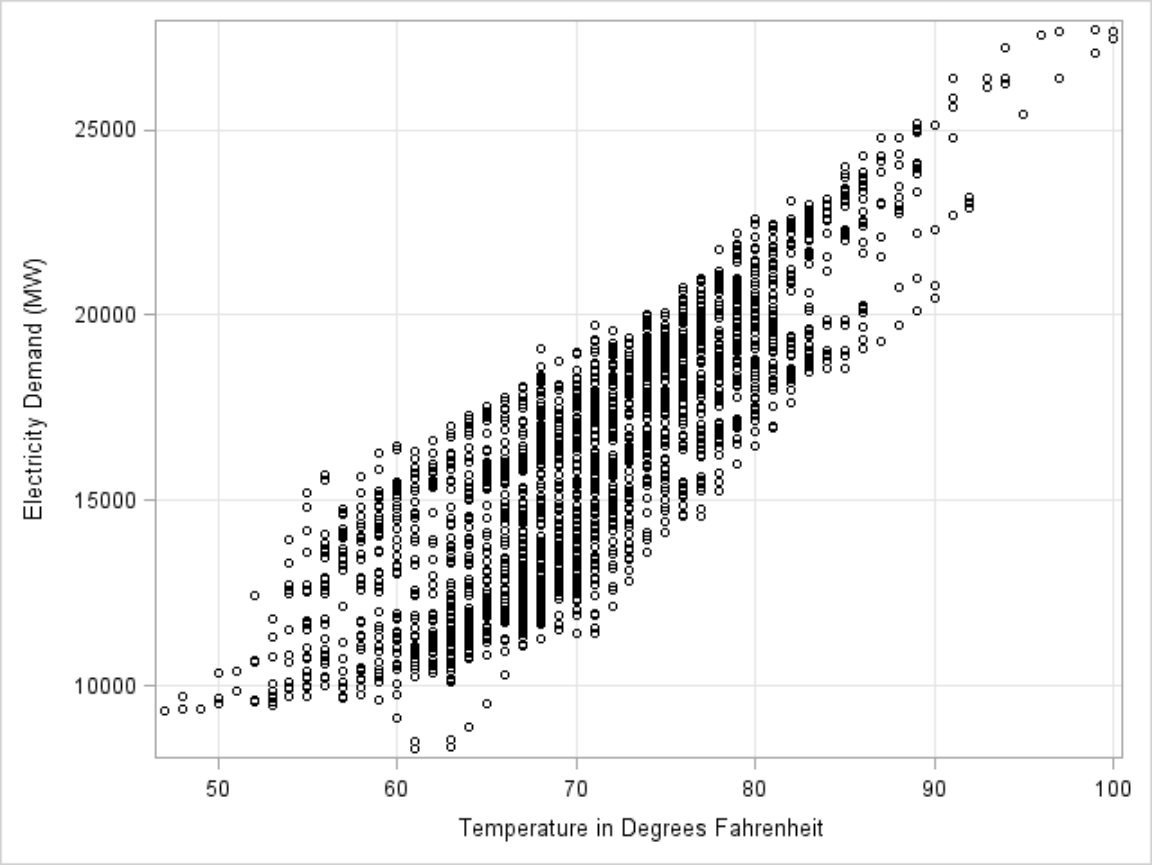
The Pearson correlation coefficient (r) between load and summer temperature is calculated and reported in Table 3-1. This statistic quantifies the degree of *linear* association between summer temperature and electricity demand.

Table 3-1: Correlation Coefficient for Electricity Load and Temperature for June, July and August, 2011

Correlation Coefficient	.85
Pr >| r |	<.0001

The degree of linear association is strong (r=.85) and highly statistically significant (p-value < .0001). Figure 1 displays a scatterplot for the n=2,208 observations.

Figure 1: Electricity Load versus Temperature for June, July, and August, 2011



Least squares is used to estimate a possible relationship between electricity demand (y_i) and temperature (x_i) over all time periods *i*. The fitted equation is:

$$\hat{y}_i = -10{,}047 + 368.94x_i \qquad (24)$$

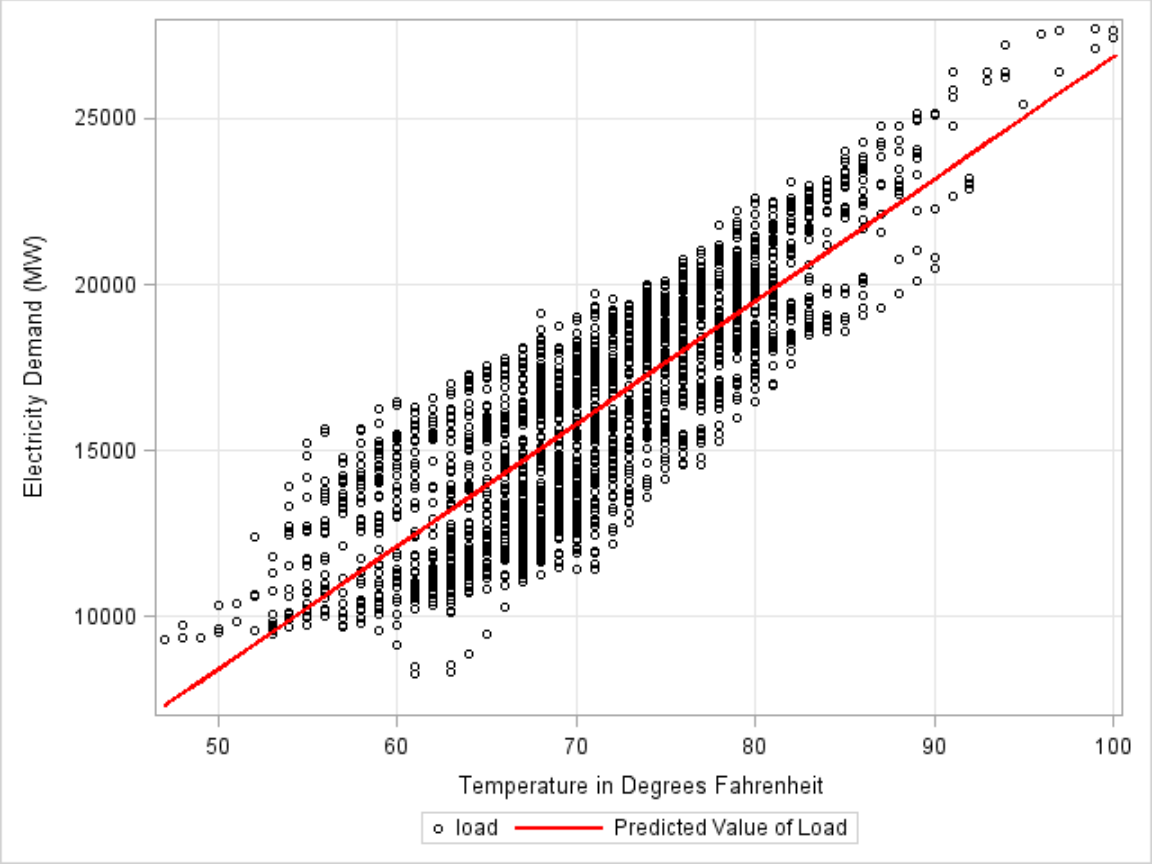
Table 3-2 summarizes the relevant estimation results for Equation (24) .

Table 3-2: Least Squares Estimates from Equation (24)

Parameter Estimates						
Independent Variable	*Estimate*	*Standard Error*	*t-Statistic*	*Pr >| t |*	*95% Lower Confidence Limit*	*95% Upper Confidence Limit*
Intercept	-10,047	356.01	-28.22	<.0001	-10,746.00	-9,349.30
Temperature	368.94	4.98	74.05	<.0001	359.17	378.71

The coefficient on temperature is positive and highly statistically significant (p-value <.0001). It suggests that a one degree increase in drybulb temperature is expected to increases average electricity demand by approximately 369 MW. This is entirely plausible as higher temperatures in the summer months lead to increased use of air conditioning and other cooling methods that rely on electricity. Figure 2 superimposes the fitted regression Equation (24) onto the observations presented in Figure 1.

Figure 2: Electricity Load and Fitted Values versus Temperature for June, July, and August 2011



Visually, the regression provides a good fit. The coefficient of determination is respectable (R^2=0.71) for this simple specification. It suggests that 71% of the variation in electricity demand is explained by temperature. Table 3-3 reports the analysis-of-variance (ANOVA) results.

Table 3-3: ANOVA table from Equation (24)

Analysis of Variance					
Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F - Statistic*	*Pr > F*
Model	1	19,581,176,979	19,581,176,979	5482.91	<.0001
Error	2206	7,878,308,844	3,571,310		
Total	2207	27,459,485,823			

The overall fit (F=5482.91) is very large and also highly statistically significant (p-value <.0001). This simple linear specification is a springboard for refinement.

The problems of autocorrelation and heteroscedasticity could very well be present but we do not address these issues here. A formal diagnostic analysis of any regression includes a Durbin-Watson[8] test for autocorrelation and hypothesis tests for heteroscedastic errors such as White's[9] or the Breusch-Pagan test[10]. In the event that these diagnostic measures indicate a violation of any one of the classical regression modeling assumptions, there are well-established methods to address each issue. We realize this simple model does not represent the true data generating process.

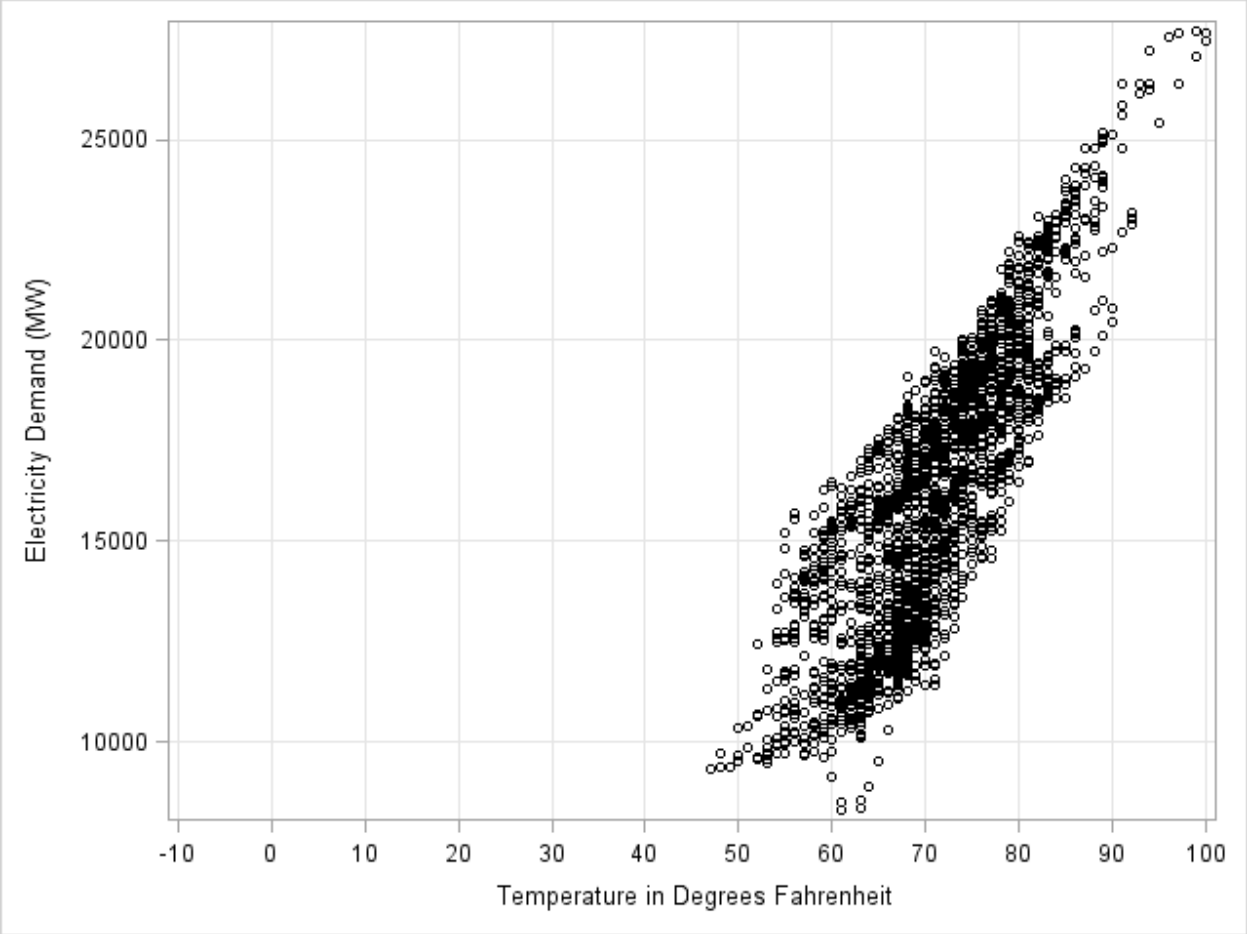
The relationship between temperature and summer electricity demand in New England is clearly evident, and linear regression is a viable modeling choice. There are instances where a model linear in the parameters is not an accurate functional form of the relationship between predictor and response. As an example, we plot the same data set of drybulb temperature and load with an expanded horizontal axis. Figure 3 is the same scatterplot as Figure 2 with the exception that temperature's units on the horizontal axis are extended to -10° Fahrenheit.

[8] Durbin, J. (1969), "Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals," *Biometrika*, 56, 1–15.

[9] White, H. (1980), "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica,* 48, 817-838

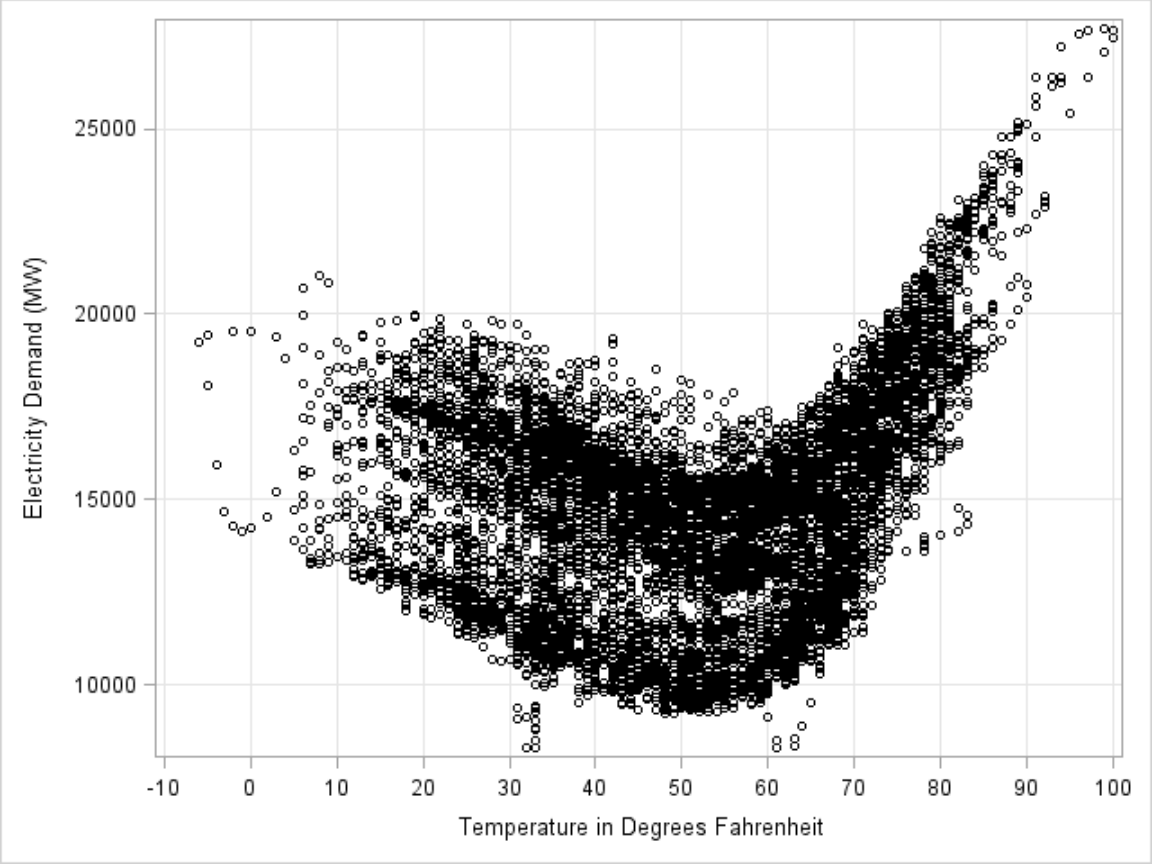
[10] Breusch, T. S. and Pagan, A. R. (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics," *The Review of Economic Studies,* 47:1, 239-253.

Figure 3: Electricity Load versus Temperature for June, July, and August, 2011



Rather than limiting our observations to the summer months only, we now consider a data set comprised of hourly observations of electricity demand and temperature for the entire year of 2011. Expanding the units on the horizontal axis of Figure 2 permits the inclusion of data points for other months of the year. Figure 3 displays all observations of temperature and demand in New England for all of 2011. The immediate impact of the inclusion of these additional observations is the need for an amended functional form.

Figure 4: Hourly Observations of Electricity Demand versus Drybulb Temperature: All Months in 2011



The linear form that was appropriate for the summer months only (Figure 3) would no longer be appropriate when considering data for the entire year. The correlation coefficient between summer temperature and load dropped to 0.24 when using the entire data set, suggesting a weak linear association. The p-value was less than .0001. It is not surprising that the correlation coefficient is significant given the thick band associated with the data in Figure 4.

The same bivariate linear regression model that was estimated for the summer was applied to the complete set of 2011 hourly observations. The resulting fitted equation for the annual data is

$$\hat{y}_i = 12{,}824 + 37.61x_i \quad (25)$$

Compared to the summer fitted regression, the annual fitted regression has an intercept that is drastically different in magnitude and sign as well as a slope coefficient that is significantly smaller. The R^2 for the annual regression is now a mere 0.06 compared to the summer regression of 0.71. Only 6% of the variation in electricity demand is explained by temperature. Table 3-4 reports the annual regression coefficient estimates.

Table 3-4: Least Squares Estimates from Equation (25)

Parameter Estimates						
Independent Variable	*Estimate*	*Standard Error*	*t-Statistic*	*Pr >| t |*	*95% Lower Confidence Limit*	*95% Upper Confidence Limit*
Intercept	12,824	87.19	147.08	<.0001	12,653.00	12,995.30
Temperature	37.61	1.61	23.42	<.0001	34.46	40.76

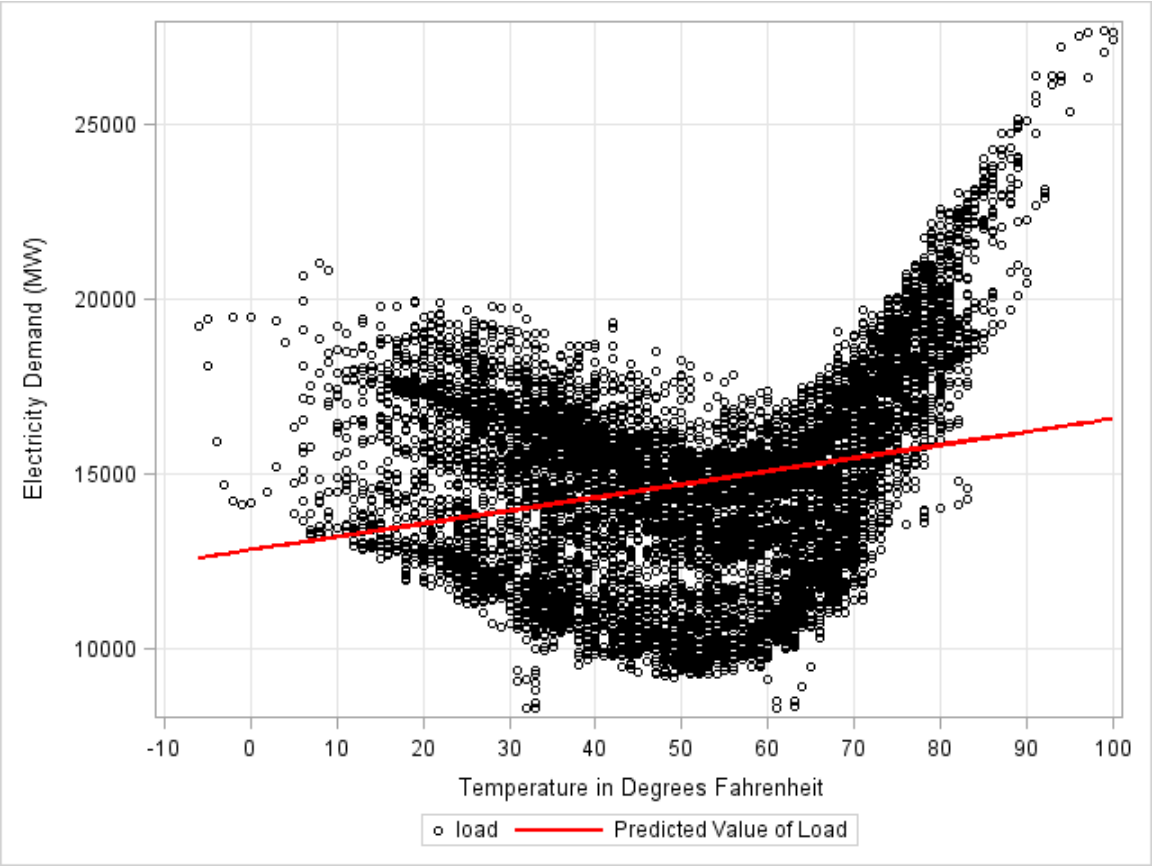
Temperature is still positive and a statistically significant predictor of average electricity demand in New England. However, the test statistics are smaller. Table 3-5 presents the amended ANOVA table.

Table 3-5: ANOVA table from Equation (25)

Analysis of Variance					
Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F - Statistic*	*Pr > F*
Model	1	4,222,886,758.00	4,222,886,758.00	548.59	<.0001
Error	8758	67,416,949,381.00	7,697,756.00		
Total	8759	71,639,836,139.00			

The F-Statistic for the annual regression is almost exactly one tenth the magnitude of the one resulting from the summer regression. It is, however, still statistically significant. Finally, Figure 5 superimposes fitted regression Equation (25) onto the data.

Figure 5: Electricity Load and Fitted Values versus Temperature for 2011



There is strong graphical evidence that the true relationship between temperature and electricity demand, over time, may in fact be highly nonlinear. In the presence of this nonlinear relationship, we amend Equation (1) to fit a quadratic model where temperature squared is included as an additional predictor. The reformulated model is presented as:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad (26)$$

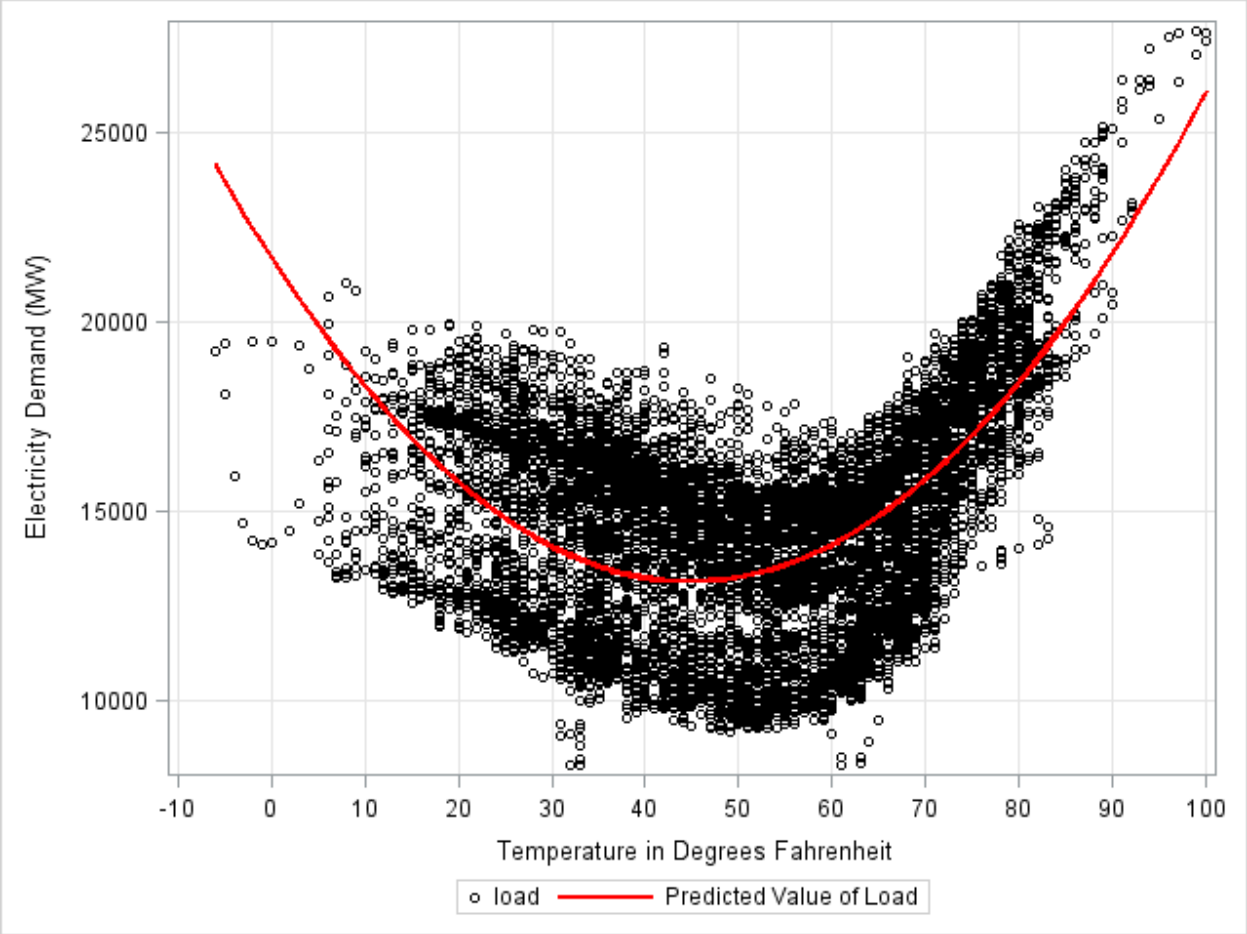
Equation (26) retains linearity in its coefficients while permitting the function itself to be nonlinear due to nonlinearity in the variables. Transforming the explanatory variable to a higher degree and then including the higher-degree terms in the model is a way to accommodate nonlinear relationships. The estimation results from fitting the alternative quadratic specification are presented in Table 3-6.

Table 3-6: Least Squares Estimates from Equation (26)

Parameter Estimates						
Independent Variable	*Estimate*	*Standard Error*	*t-Statistic*	*Pr >| t |*	*95% Lower Confidence Limit*	*95% Upper Confidence Limit*
Intercept	21,733	155.293	139.95	<.0001	21,428	22,037
Temperature	-383.17	6.63847	-57.72	<.0001	-396.19	-370.16
Temperature Squared	4.27	0.06596	64.68	<.0001	4.13713	4.39574

The fitted equation from Equation (26) is superimposed onto the full year observations in Figure 6.

Figure 6: Electricity Load and Fitted Values versus Temperature for 2011



The restriction of a linear association is relaxed by including the squared term for temperature. In this sense, Equation (26) is an unrestricted model while Equation (1) is a restricted model.

The ANOVA results associated with the Table 3-6 estimation results are presented in Table 3-7.

Table 3-7: ANOVA table from Equation (26)

Analysis of Variance					
Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F - Statistic*	*Pr > F*
Model	2	26,018,039,453.00	13,009,019,726	2497.05	<.0001
Error	8757	45,621,796,686.00	5,209,752		
Total	8759	71,639,836,139.00			

The overall fit of the unrestricted model, as reflected by F=2,497, is greater than that of the restricted model (F=548.59). We conduct a Chow test to formally test which model is better. Under the null hypothesis that the coefficient on the quadratic term in Equation (26) is zero, the appropriate F-statistic is given by:

$$F = \frac{(SSE_R - SSE_U)/j}{\frac{SSE_U}{DF_U}} \tag{27}$$

where $SSE_R$is the residual sum of squares for restricted model (1), SSE_U is the residual sum of squares and DF_U is the degrees of freedom for the unrestricted model (26), and j is the number of restrictions.

Using the ANOVA output from Table 3-5 and Table 3-7, the test statistic is calculated as:

$$F = \frac{(67{,}416{,}949{,}381.00 - 45{,}621{,}796{,}686)/1}{\frac{45{,}621{,}796{,}686}{8757}} = \frac{21{,}795{,}152{,}695}{5{,}209{,}751.82} = 4{,}183.53 \tag{28}$$

At the 5% level of significance, the critical value (F_c) with (1, 8757) degrees of freedom is 3.84. As $F = 4{,}183.53$ is greater than $F_c = 3.84$, the null hypothesis is resoundly rejected.

This is strong statistical evidence that a nonlinear relationship exists between temperature and electricity demand[11].

Nonlinear least squares is a competing approach for accommodating nonlinearities in a regression framework. While the appropriately specified linear regression model results in unbiased estimators, the nonlinear least squares counterpart may be biased owing to a lack of an explicit solution. In this case, iterative procedures are available using, for example, the Gauss-Newtown and Levenberg-Marquandt algorithms (Ruppert et al. 2003, pp. 49). Comparable to the first CLM assumption, nonlinear regression assumes only nonlinear relationships between predictor and response variables. This is also not an ideal modeling technique for short term load forecasting as there may be additional predictors whose effect can be accurately captured in a linear regression model.

3.5 Nonparametric Estimation

While nonlinear relationships can be accommodated in a linear regression model, an alternative approach that introduces flexibility regarding the specification of a functional form is *nonparametric regression*. The terms *curve fitting* and *scatterplot smoothing* have also been used interchangeably for nonparametric regression (albeit primarily in the univariate case). This methodology works toward a smooth functional relationship and is based solely on sample observations; i.e., it is *data-driven*. There are many methods available to approximate a smooth function (Silverman, 1985) Examples are kernel density

[11] The same result is obtained by doing a straightforward t-test on the coefficient for the temperature-squared variable in Table 3-6. In particular, notice that the F-statistic (=4,183.53) in Equation (28) can likewise be obtained by squaring the t-statistic for Temperature Squared reported in Table 3-6; i.e., $64.68^2 = 4{,}183.5$. This is no coincidence. $F = t^2$ when testing a single coefficient.

estimation, exponential smoothing and smoothing splines. For exposition, consider this simple nonparametric model:

$$y_i = f(x_i) + \varepsilon_i \,, \tag{29}$$

where $f(x_i)$ is a smooth function and all other notation is the same as before. This model is nonparametric in the sense that the function f is generalized and is not assigned or assumed to take on a specific form. In other words, there is no population parameter or set of parameters that define the relationship.

A simple method of estimating the function f is by piecing together several line segments or polynomials at different locations in the domain of f to form a 'grand' curve. These line segments are variable constructs called *splines,* and the locations where they are tied together are referred to as *knots*. Choosing the knot locations and the number of line segments allows flexibility in the estimation of $f(x_i)$.

The simplest spline model uses polynomial functions of degree 1. In this case, each piece-wise function is linear and can be mathematically expressed as follows:

$$(x_i - k_\kappa)_+ = \begin{cases} 0, & x_i - k_\kappa < 0 \\ x_i - k_\kappa, & x_i - k_\kappa \geq 0 \end{cases} \tag{30}$$

Here k_κ refers to a scalar-valued knot, indexed by $\kappa = 1, \ldots, K$. The relative value of k_κ within the sample range x_i defines the knot's *location*. Choosing the location of a knot can be arbitrary or based on sample information such as sample quartiles. Equation (30) is sometimes referred to as the "positive" portion of $(x_i - k_\kappa)$ as it takes a nonzero value only

when the value of $(x_i - k_\kappa)$ is nonzero and zero otherwise. In other words, the variable x_i is truncated at the value k_κ.

In general, these constructs can be referred to as *basis functions* or *spline basis functions* because they represent a change in direction or the structural characteristics of a particular relationship. A *basis* for modeling structural change is formed when more than one basis function is employed[12]. As an example, consider the same annual data set used to get fitted Equation (25) and where one knot is placed at 65 degrees. With k=1 knot, there is only one basis function. Specifically,

$$(x_i - 65)_+ \tag{31}$$

By including this spline basis function and the temperature variable itself as explanatory variables in a linear regression, we estimate the marginal effect that temperature has on electricity demand both below and above 65 degrees Fahrenheit. The resulting regression model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - 65)_+ + \varepsilon_i \tag{32}$$

For this example, least squares was used to estimate this model and yielded the following fitted equation:

$$\widehat{y_i} = 16{,}041 - 46.63 x_i + 485 (x_i - 65)_+ \tag{33}$$

As the basis function $(x_i - 65)_+$ is just a function of temperature, its evaluated value is inherently dependent on the value of temperature. Readers uncomfortable with this

[12] A *basis* in linear algebra is a set of elements within a vector space, a linear combination of which can be used to uniquely express any other element in that vector space. used to uniquely express any other element in that vector space.

notation should just consider that this new term simply represents the deviation of x_i from 65 degrees. However, if this deviation is negative (i.e., $x_i < 65$) then evaluation of $(x_i - 65)_+$ produces a value of zero. Therefore, the inclusion of $(x_i - 65)_+$ into model (33) conveniently allows us to estimate temperature's effect (i.e., slope) both below and above 65 degrees. In the context of fitted Equation (33), least squares has provided parameter estimates that seem to indicate a relatively small and negative marginal effect of temperature up until 65 degrees (-46.63) and a much larger, positive marginal effect afterwards (e.g., 485). Table 3-8 presents the complete estimation results.

Table 3-8: Least Squares Estimates from Equation (33)

Parameter Estimates						
Independent Variable	*Estimate*	*Standard Error*	*t-Statistic*	*Pr >| t |*	*95% Lower Confidence Limit*	*95% Upper Confidence Limit*
Intercept	16,041	80.46	199.36	<.0001	15,884.00	16,199.30
Temperature	-46.63	1.68	-27.70	<.0001	-49.96	-43.29
Basis Function	485.89	6.48	74.95	<.0001	474.68	497.09

Using the estimates provided in Table 3-8, fitted values of electricity demand can be produced with corresponding observations of temperature. The negative coefficient reported on Temperature (-46.63) in Table 3-8 is the marginal effect of each additional degree Fahrenheit on load. However, the basis function plays a unique role when calculating a fitted load value. It modifies the impact of each degree Fahrenheit depending on where the temperature falls relative to 65°. The positive coefficient on the basis function affects temperature's impact. Specifically, expanding Equation (33) and collecting terms confirms that there is in fact a completely different slope before and after 65 degrees. At

$x_i > 65$, the enumerated value of $(x_i - 65)_+$ is nonzero and Equation (33) can be expanded:

$$\begin{aligned}\widehat{y_i} &= 16{,}041 - 46.63x_i + 485x_i - 485 * 65 \\ &= 16{,}041 + 438.37x_i - 31{,}525 \\ &= -15{,}484 + 438.37x_i \end{aligned} \quad (34)$$

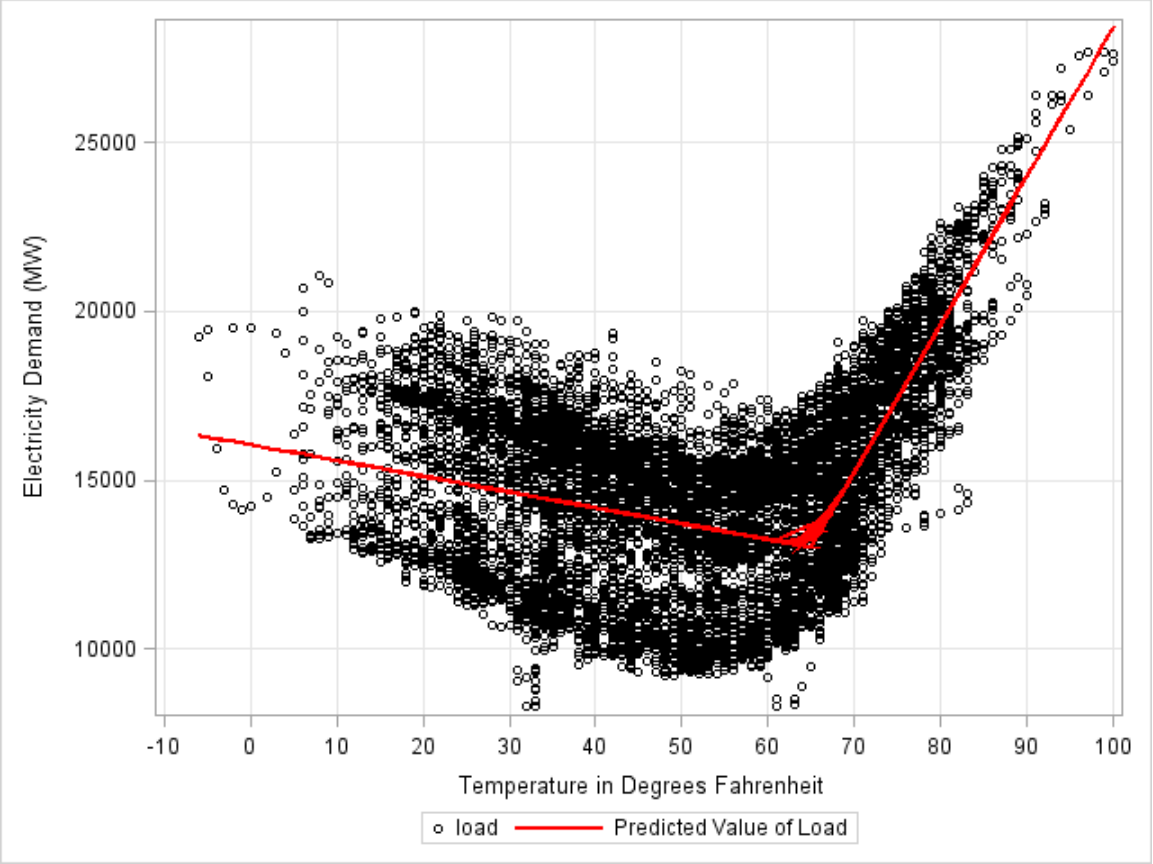
Equation (34) reveals a significantly positive slope at temperatures above 65 degrees. Effectively, there are two estimated equations that are based on this single knot and specific basis function used in estimation. Specifically,

$$\widehat{y_i} = 16{,}041 - 46.63x_i, \qquad \forall\, x_i < 65 \quad (35)$$

$$\widehat{y_i} = -15{,}484 + 438.37x_i, \qquad \forall\, x_i \geq 65 \quad (36)$$

Figure 7 superimposes the fitted Equation (33) onto the scatterplot of annual temperature and demand data.

Figure 7: Scatterplot of Annual Data and Fitted Values using a Single Knot at 65 Degrees



From this scatterplot, there appears to be a distinctive change in the relationship between demand and temperature at 65 degrees Fahrenheit. This single-knot representation is also referred to as the *broken-stick model.*

Choosing 65 degrees in this application as a knot location was not an arbitrary choice. This temperature is often used to indicate human comfort and to delineate between heating and cooling effects of temperature on electricity demand[13]. For instance, at temperatures above 65 degrees, there is said to be a cooling effect as electricity is used to bring the temperature back down to 65 degrees and vice versa. In our spline model, this knot is designed to represent change in curvature. If more knot locations are chosen, then

[13] http://www.srh.noaa.gov/key/?n=climate_heat_cool

each would represent a point where some change is expected to occur in the relationship of interest. While employing a single basis function at $k = 65$ appears to have captured an important structural change, the fit is still rough and the true functional might be better approximated with more knot locations.

Equation (32) could be amended to include additional knot locations that correspond to generating additional truncations of the variable x_i. These collectively form what is referred to as the *truncated line basis* or *truncated power functions* (TPF). The former is named from the resulting line segments in a regression model while the latter incorporates higher degree polynomial line segments that are also truncated at specified values. For instance, the TPF basis of degree *p* is given by

$$1,\ x_i,\ x_i^p,\ (x_i - k_1)_+^p,\ \dots,\ (x_i - k_\kappa)_+^p \qquad (37)$$

It can be clearly seen that the truncated line basis is a special case of the TPF basis where $p = 1$. Using a higher degree TPF basis in a regression might allow for more curvature and smoothness in the estimated function but does not necessarily result in a better fit. In addition, the ability to choose among alternative bases allows flexibility in spline models. In particular, the TPF basis can cause numerical instability in least squares estimation as variable truncation leads to design matrices with columns whose entries are mostly zero. With columns of the design matrix nearly identical to each other, this can lead to problems calculating $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ in the least squares estimator as $\boldsymbol{X}'\boldsymbol{X}$ becomes singular and, thus, its inverse does not exist.

In light of this drawback, bases with improved numerical properties, such as the radial and B-spline bases, are commonly used (Eilers and Marx, 1996). However, most statistical packages automatically treat the design matrix with an orthogonal

transformation in the presence of an ill-conditioned design matrix. Demmler- Reinsch Orthogonalization and QR decomposition are widely used (Calderon et al, 2009). Therefore the basis that is used in constructing a spline model is not necessarily the one used in estimating that model. Using more sophisticated spline bases has not been shown to necessarily improve overall fit, and linear splines have been shown to perform adequately in many instances (Ruppert et al., 2003).

For the current context, we focus only on the TPF basis of degree 1 where each spline basis function is linear.

$$1,\ x_i,\ \ (x_i - k_1)_+,\ \ \dots\ ,\ (x_i - k_\kappa)_+ \tag{38}$$

Using these basis functions as explanatory variables in a linear regression amounts to connecting line segments at the knots used to construct the respective function. By using the basis in Equation (38), the number of knots can be extended to yield the following linear spline:

$$f(x_i) \ = \ \beta_0 + \beta_1 x_i + \sum_{\kappa=1}^{K} u_\kappa (x_i - k_\kappa)_+ \tag{39}$$

where u_κ is a regression coefficient (i.e., parameter) associated with knot k_κ. Since there is a single coefficient to be estimated for each basis function (at each knot), it can be shown that using spline models in this way is a simple extension of classical linear regression.

Including more knot locations in a spline model can result in capturing more of the structural characteristics of the target relationship. Rewriting Equation (29) in the context of Equation (39), we have general formulation for nonparametric regression using linear splines and K knots as follows:

$$y_i = \ f(x_i) + \varepsilon_i \,, \tag{29}$$

$$y_i = \beta_0 + \beta_1 x_i + \sum_{\kappa=1}^{K} u_\kappa (x_i - k_\kappa)_+ + \varepsilon_i \tag{40}$$

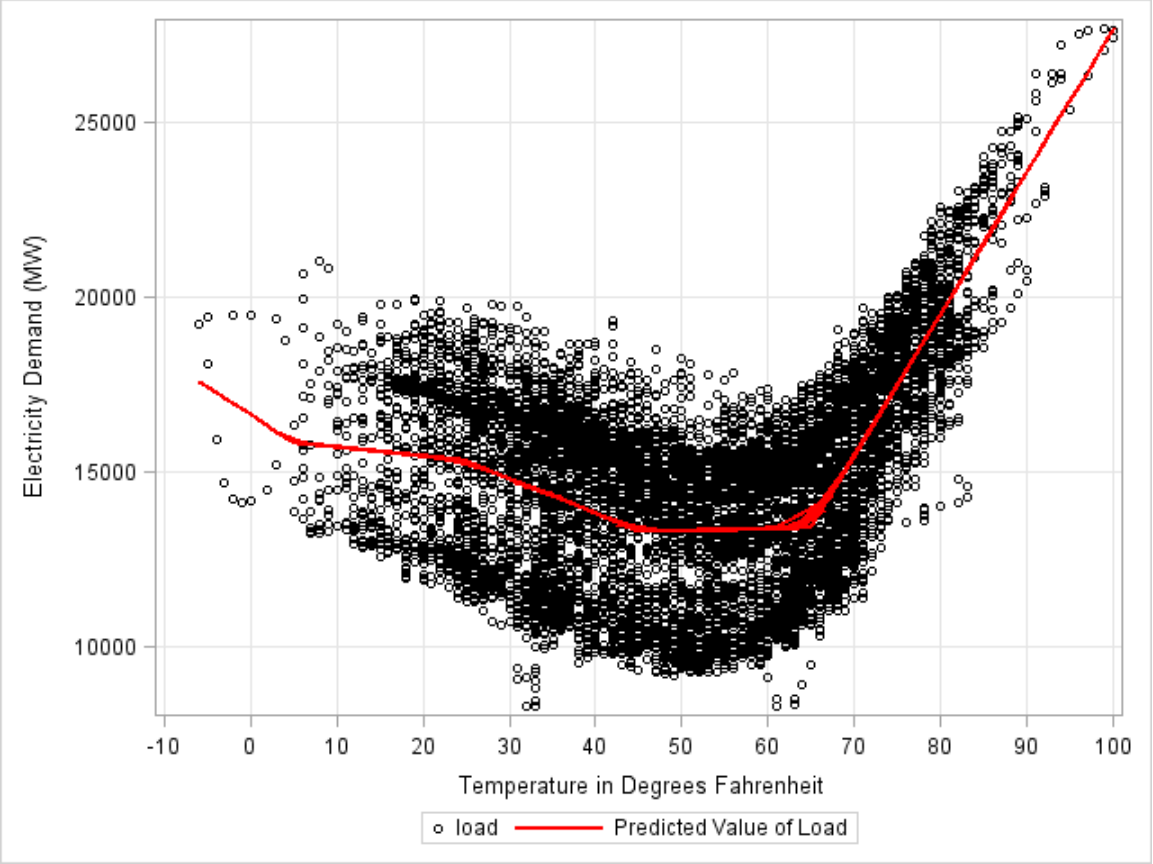
For illustrative purposes, we chose knots at 5, 25, 45, and 65 degrees and estimate Equation (40) using least squares. The resulting parameter estimates are presented in Table 3-9 below.

Table 3-9: Least Squares estimates for fitting Equation (40) with four knots.

Parameter Estimates						
Independent Variable	*Estimate*	*Standard Error*	*t-Statistic*	*Pr >| t |*	*95% Lower Confidence Limit*	*95% Upper Confidence Limit*
Intercept	16,649	420.96	39.55	<.0001	15,823	17,474
Temperature	-159.20	86.94	-1.83	0.06	-329.62	11.22
5-25	133.85	91.01	1.47	0.14	-44.54	312.24
25-45	-77.53	14.91	-5.2	<.0001	-106.76	-48.30
45-65	109.33	8.45	12.94	<.0001	92.76	125.90
>65	400.35	9.06	44.2	<.0001	382.59	418.10

Fitted values for the K=4 model are plotted in Figure 8 below.

Figure 8: Scatterplot of Annual Data and Fitted Values using Knots at 5, 25, 45 and 65 Degrees



There are several noticeable differences between the fitted curves in Figure 7 and Figure 8. In the latter, there appears to be no marginal effect between 45 and 65 degrees, as suggested by the near-horizontal fit of the equation between these two temperatures. Also, the changes in slope for earlier segments suggest additional features of the relationship at lower temperatures. Including more knots has resulted in a smoother curve than that provided by the simple broken-stick model.

This is meant only as a modeling example. Our intent is to model the underlying relationship itself. A linear spline fit with four knots might very well be an approximation of that relationship but it is not necessarily the case that the true relationship is linear

between each of the pre-selected knot locations. Our goal is to estimate an assumed smooth functional relationship.

3.5.1 Penalized Splines

It is clear that modeling the relationship between temperature and electricity demand, as well as the performance of splines in general, depend on choices such as number of knots, their location, and the degree of the basis functions used in their application. In addition, a smoother fit can be achieved by using higher-degree polynomials or by directly penalizing the roughness of the fit. For instance, smoothing splines are spline models that use every unique data point as a knot (Reinsch, 1967; Wahba, 1990). The "roughness" of a smoothing spline fit is penalized by constraining the integrated squared second derivative (See Equation (41) below). The purpose of this constraint is to permit a large number of knots while requiring the u_{κ} coefficients to be within a certain limit. This penalty was pioneered by Reinsch (1967).

The sum of squared residuals and the penalty for the general smoothing spline can be written as

$$S = \sum_{i=1}^{n} \{y_i - \ f(x_i)\}^2 - \lambda \int_a^b \{f''(x_i)\}^2 dx, \qquad (41)$$

where λ is referred to as the *smoothing parameter* and is used to control the degree of smoothing. Notice that the second term in Equation (41) is the integral over the squared second derivative of f. Penalizing the second derivative allows for direct control over the smoothness of the fit. However, this penalty is not limited to the second derivative, and any derivative can be used. The second derivative is a common choice as it represents a

compromise between the linear fit which results from penalizing the first derivative and the complex equations that result from penalizing higher degree derivatives (Eilers and Marx, 1996, p.91). Also, because every unique data point is used as a knot location, smoothing splines use *n* basis functions in their estimation (where, again, *n* is the number of observations). As such, this can cause difficulty in calculating model selection criterion when the sample size is large or when additional predictors are included.

Choosing the appropriate number of knots for a spline model has received much attention and, in general, "overfitting" is a consequence of too many knots while "underfitting" results from too few knots (Eilers and Marx, 1996). Overfitting a spline with too many knots is of particular concern as it can results in approximating both the structural features as well as the random fluctuations (noise) of the underlying process. O'Sullivan (1986, 1988) showed how choosing a relatively large number of knots, less than *n*, in conjunction with a penalty on the second derivative could provide comparable results to a smoothing spline with *n* knots.

An alternative roughness penalty was introduced by Eilers and Marx (1996) using splines with equally spaced knots. It penalizes the differences between adjacent spline coefficients rather than basing the penalty on a derivative. While the penalty of Eilers and Marx was applied to a particular spline basis, referred to as B-splines (de Boor, 1978), it can also be applied to any other corresponding basis. An advantage of this approach is the ability to choose the spline model and penalty separately. These penalized splines have since been generally referred to as P-splines and share a lot in common with the low-rank *pseudosplines*, which were proposed by Hastie (1996). Ruppert (2002) discusses choosing

the number of knots and the penalty simultaneously. The results of Monte Carlo simulations in his study suggest a default number of knots that is sufficiently large enough to capture the underlying process but has little impact on the fit when this default value is exceeded.

Ruppert, et al. (2003) show how penalized spline regression can be generalized and applied by (1) choosing the degree of the spline and the number and location of knots for the spline model and (2) choosing the way that the roughness of the fitted spline will be quantified and penalized. They go on to say that choosing the basis functions used in model construction and the basis functions used in actual estimation are secondary, yet still essential choices. Our approach follows that of Ruppert, et al (2003) very closely. The general definition of a penalized spline given in Ruppert, et al (2003) allows for a straightforward application using mixed models discussed in the following section. The P-spline approach used here is laid out below using the familiar example of annual electricity usage and temperature.

3.5.2 Penalized Spline Example

Let **X** be an n x (2+K) design matrix consisting of temperature observations and defined as

$$\mathbf{X} \equiv \begin{bmatrix} 1 & x_i & (x_i - k_1)_+ & \dots & (x_i - k_{\mathrm{K}})_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - k_1)_+ & \vdots & (x_n - k_{\mathrm{K}})_+ \end{bmatrix} \qquad (42)$$

In addition, let **y** be an (n x 1) vector of observations of electricity load, **β** be a (2+K) x 1 vector of regression coefficients, and **ε** be a n x1 vector of independently and identically distributed residuals. **β** can be written out fully as:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ u_1 \\ \vdots \\ u_K \end{bmatrix} \tag{43}$$

which is of dimension (2+*K*) x 1.

The impact that temperature has on electricity load can be expressed as a linear spline (where *p=1*) and reformulated in matrix notation in the classical linear model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{44}$$

Even if K is sufficiently large, least squares estimation of Equation (44) can still result in a rough fit (Ruppert, et al. 2003, p. 65). If there are many knots, then there are many line segments that make up the spline. Each of these segments has its own slope. Collectively, these can result in an overall rough or wiggly fit. This can be avoided by imposing a constraint. As previously discussed, smoothing splines 'shrink' the influence that each knot has on the fit by penalizing the integrated second derivative of the whole fit. This concept can also be applied to low-rank smoothers, where the number of basis functions is less than the sample size *n* (Hastie, 1996). Penalized splines are low-rank smoothers where the influence of each spline coefficient u_k is constrained by imposing the following restriction:

$$\sum \hat{u}_k^2 < \Upsilon \tag{45}$$

where Υ is some arbitrary constant. The purpose of this constraint is to permit a large number of knots while requiring the u_k coefficients to be within a certain limit. Each u_k is, by definition, a slope. Restricting the collective set of these slopes means promoting smoothness in the spline without compromising the number of knots.

There exist other constraints that would result in a smooth fit, but this particular constraint is attractive as it is easy to implement with least squares. To see how, consider the penalty matrix[14] defined by

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1_1 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1_K \end{bmatrix} \tag{46}$$

The matrix **D** (of dimension 2+K by 2+K) has nonzero entries on only those diagonal elements that correspond to knot coefficients (as opposed to the intercept or the coefficient for temperature). These nonzero entries correspond to the columns of **X** that are populated with the spline basis functions. The other columns, specifically the first and second, and their associated regression coefficients are not affected by this penalty because the first two diagonal entries in **D** are zero. While our choice of **D** coincides with a P-spline using the truncated line basis, other penalties are possible. In fact, Eilers and Marx (1996) originally constructed a matrix **D** by using second-differences in order to approximate the penalty on the second derivative which is commonly used in smoothing splines. Using **D** to target only the spline coefficient entries in the parameters vector **β,** the constraint in Equation (45) can be formulated in matrix notation as:

$$\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} \leq \Upsilon \tag{47}$$

Note that **β** is defined by Equation (43).

Minimizing the least squares objective function (i.e., the sum of squared residuals) in Equation (9) subject to Equation (45) amounts to a Lagrange multiplier constrained

[14] The penalty matrix can take numerous forms appropriate for the corresponding choice of basis. The only necessity is that the penalty matrix must be positive semi-definite.

minimization. This is sometimes referred to as penalized sum of squares or regularized regression. The criterion for estimating a P-spline model with least squares is given by

$$Minimize\ (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \quad + \lambda^{2p}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} \tag{48}$$

The term $\lambda^{2p}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}$ is known as a roughness penalty as it quantifies and restricts the roughness of the resulting fit. Ruppert, et al. (2003) justify the exponent $2p$ on the smoothing parameter, where again p is the degree of the spline, by noting that transforming x (e.g., with quadratic or cubic splines) warrants a comparable transformation of λ. Using the same calculations to find the least squares estimator, the solution to the minimization problem, for a given λ, can be shown to be[15]

$$\widehat{\boldsymbol{\beta}_{\lambda}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda^{2p}\mathbf{D}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{49}$$

Fitted values of the response variable are calculated by

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda^{2p}\mathbf{D}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}_{\lambda}} \tag{50}$$

Analogous yet separate to the classic smoothing spline, the degree of smoothing for a penalized spline is likewise controlled through a smoothing parameter $\lambda \geq 0$. When λ is set to zero, the roughness penalty on the spline coefficients becomes zero. As such, a smoothing parameter of zero in Equation (49) results in the constrained estimator becoming the least squares estimator. On the other hand, when $\lambda \to \infty$ the fit 'shrinks' towards the pth degree polynomial regression.

There exist several ways to determine the appropriate degree of smoothing. Both smoothing splines and penalized splines can have smoothness chosen by criteria such as Cross-Validation (*CV*), General Cross-Validation (*GCV*), Akaike's Corrected Information Criterion *(AICc*) and Mallows' criterion (*Cp*). However, using criteria such as these becomes

[15] Ruppert et al, 2003, p.66

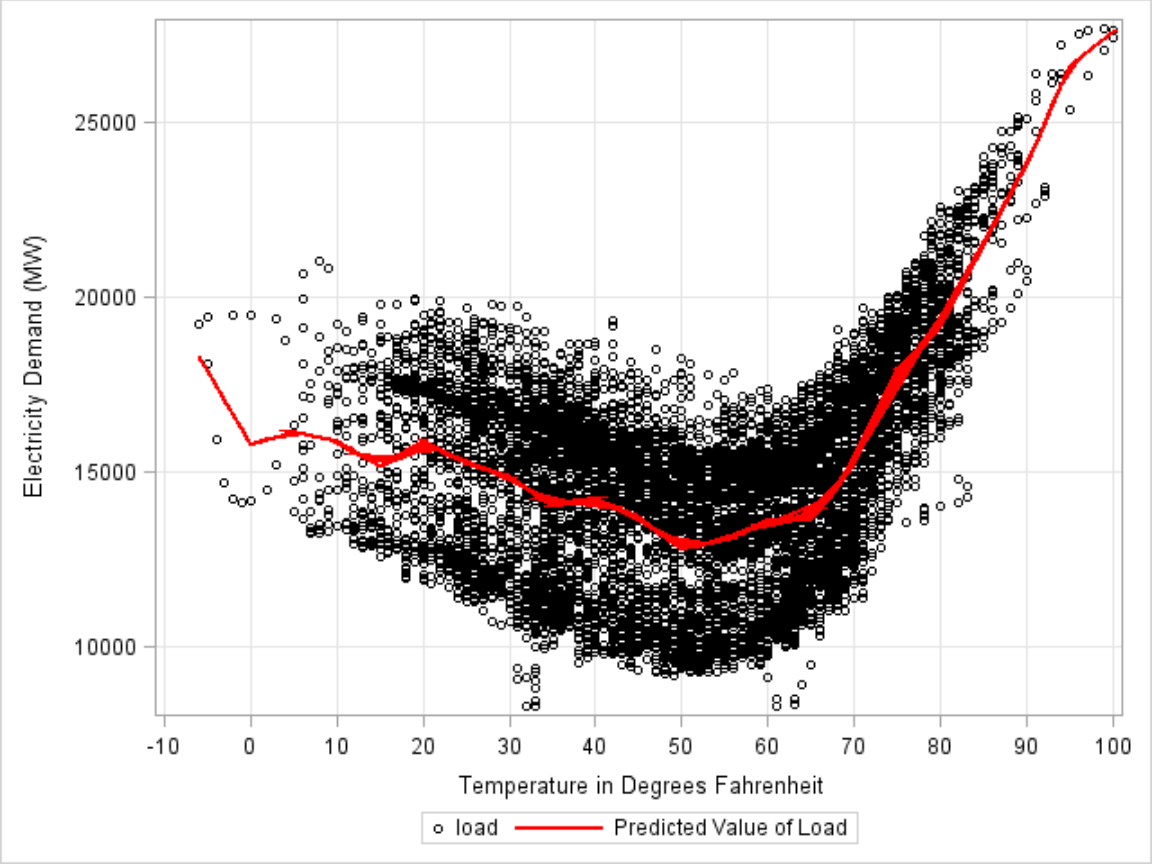
complicated when a large range of candidate values is used to estimate λ or if smoothing parameter estimates are simultaneously required for multiple nonparametric terms (Calderon et al., 2009).

Alternatively, the degree of smoothing can be obtained through maximum likelihood (ML) and restricted maximum likelihood (REML) estimation. The likelihood approach toward penalized spline smoothing has a strong connection with mixed model methodology, and current software packages can be used. This approach is described in section 3.6.

To illustrate how this new formulation results in a smoother fit, consider changing the number of knots from 4 (see Figure 8) to 20. Using K=20 knots, a linear spline model was fit to the annual data with least squares[16]. Figure 9 provides a pictorial representation. Notice that Figure 9 is still rough, even with more knots.

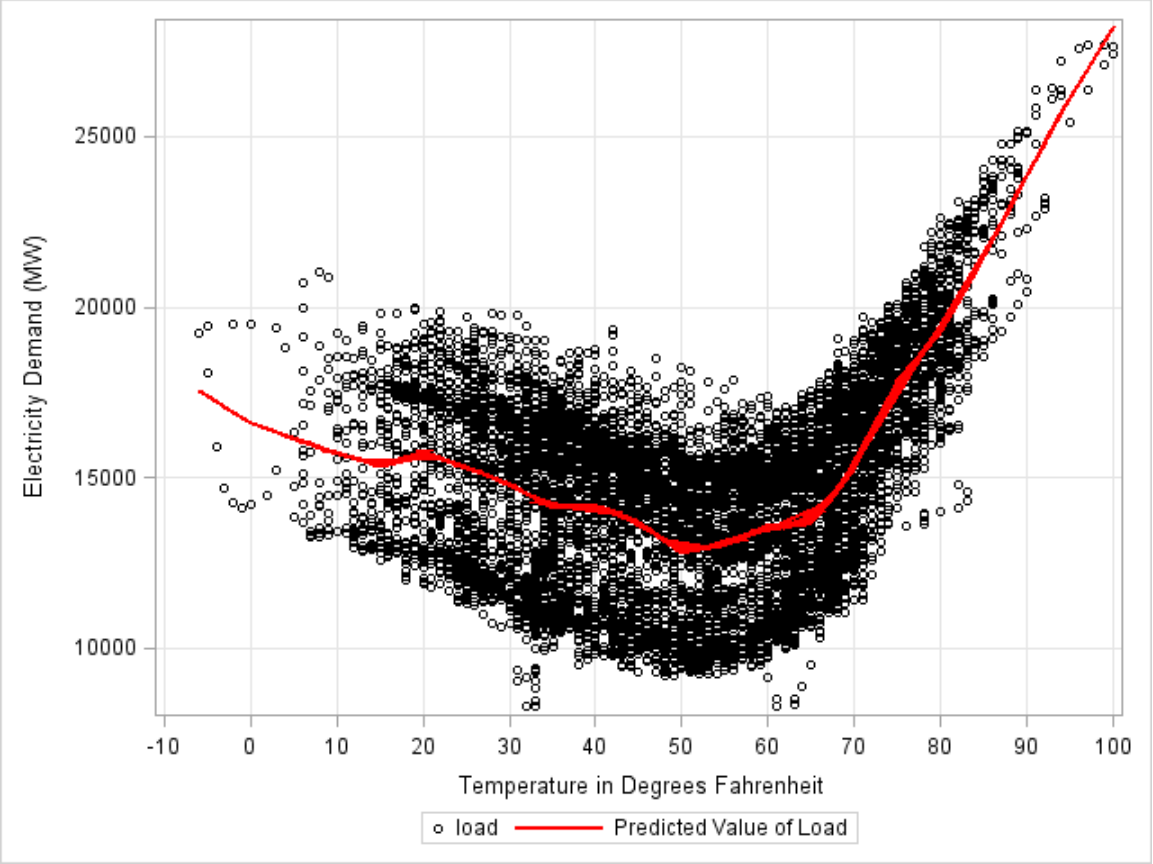
[16] Figure 9 was fit using PROC REG is SAS while Figure 10 was fit using PROC MIXED. The latter can easily be used for standard least squares, but for simplicity, PROC REG is used when smoothing is not required.

Figure 9: Fitted Values from a Linear Spline Model with K=20 Knots



Next, a P-spline (penalized) model was estimated using the same 20 knots. Figure 10 displays the fitted function.

Figure 10: Fitted Values from a Penalized Spline with K = 20 Knots



The degree of smoothing in Figure 10 was obtained with REML estimation. Both spline models represented by Figures 9 and 10 use the truncated line basis of degree one[17] with the same equally distanced knots. It is clear that shrinking the spline coefficients in the P-spline formulation has resulted in a smoother fit.

3.6 Penalized Splines as Mixed Models

Linear regression models that incorporate random effects are known as *mixed models*. Recall that the classical linear regression model treats all of the regression coefficient coefficients as fixed. A mixed model framework allows for some coefficients to be random.

[17] Note: Ruppert (2003) shows how the smoothing parameter is proportional to the degree of the spline basis on page 66.

There exists a surprisingly simple equivalence between the penalized spline proposed by Eilers and Marx (1996) and an analogous mixed-model representation. By treating the coefficients of the spline basis (i.e., the u_{k}) as random, the linear spline model in Equation (40) can be represented as a *random coefficient linear regression spline* (Brumback et al., 1999). The model is:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{\kappa=1}^{K} u_{\kappa}(x_{\ i} - k_{\kappa})_{+} + \varepsilon_i \tag{51}$$

The first thing to notice is that Equation (51) is a repeat of Equation (40). However, the u_{k} are no longer treated as (fixed) population parameters. Instead, each coefficient is assumed to be random, with $\mathrm{u_k} \sim \mathrm{N}(0, \sigma_u^2)$ and σ_u^2 denotes the variance of the random coefficients.

By separating the fixed and random effects, Equation (51) can be reformulated as a general linear mixed model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{52}$$

where the design matrices are now defined as

$$\mathbf{X} \equiv \begin{bmatrix} 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},\ \mathbf{Z} \equiv \begin{bmatrix} (x_i - k_1)_+ & \dots & (x_i - k_{\mathrm{K}})_+ \\ \vdots & \ddots & \vdots \\ (x_n - k_1)_+ & \vdots & (x_n - k_{\mathrm{K}})_+ \end{bmatrix}, \tag{53}$$

and $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are assumed to behave as follows:

$$E\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{54}$$

The covariance structure for this mixed-model representation is defined as

$$cov\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix} \tag{55}$$

where

$$\boldsymbol{G} = \sigma_u^2\mathbf{I}, \quad \boldsymbol{R} = \sigma_\varepsilon^2\mathbf{I} \tag{56}$$

By Equation (55) note that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent of each other.

It is clear that fitting a mixed model entails estimation of both fixed effects and covariance parameters, as well as the prediction of random effects. There is a significant distinction between the term *estimation* and *prediction;* a population parameter is a constant value that may be estimated while random coefficients are stochastic and must be predicted. Preferred estimation methods, such as Bayesian likelihood based approaches, do not apply to random effects. Ruppert et al. (2003, p.98-99) provides a treatment of this approach and begins by rewriting Equation (52) as follows:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}^*, \text{ where } \boldsymbol{\varepsilon}^* = \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{57}$$

Presenting the mixed-model in this way allows us to consider the stochastic term $\boldsymbol{\varepsilon}^*$ in Equation (57) to be viewed as a summation of two separate stochastic components, namely the random coefficients **u** and the stochastic disturbances **ε.** The covariance matrix (**V**) used in estimation is now generalized to incorporate the additional stochastic terms in Equation (57):

$$\boldsymbol{V} \equiv cov(\mathbf{y}) = cov(\boldsymbol{\varepsilon}^*) = \mathbf{ZGZ^T} + \mathbf{R} \tag{58}$$

where all notation has been defined in Equations (52) – (56).

In the presence of covariance matrix **V** in Equation (58), the appropriate estimator is the *generalized least squares (GLS) estimator* written as follows:

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X^T}\boldsymbol{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X^T}\boldsymbol{V}^{-1}\mathbf{y} \tag{59}$$

The GLS estimator of **β** is appropriate in the presence of **u** in Equation (52) or $\boldsymbol{\varepsilon}^*$ in Equation (57). Furthermore, Maximum Likelihood (MLE) or Restricted Maximum

Likelihood (REML) can be used to estimate the parameters $\boldsymbol{\beta}$ as well as the components of $\boldsymbol{V}$ in this context**.** When there is an unknown, generalized covariance matrix to be estimated, the consistent estimator is called the *estimated generalized least squares (EGLS)* estimator. However, when ML or REML is used to optimize first over **β,** for a fixed **V**, the likelihood function is maximized by Equation (59) (Ruppert, et al., 2003, p 100). This is illustrated below by substituting $\widetilde{\boldsymbol{\beta}}$ into the log likelihood function.

Fitting the mixed-model in Equation (52) requires predictions of the random effects **u**, which the GLS estimator does not provide. Robinson (1991) as well as Hayes and Haslett (1999) show that the notion of *Best Linear Unbiased Prediction* (BLUP) can produce predictions of the random coefficients once the parameters $\boldsymbol{\beta}$ and $\boldsymbol{V}$ have been estimated. The goal of BLUP is to minimize the prediction error[18], subject to a condition of unbiasedness, over all candidate linear random and fixed effects vectors $\widetilde{\mathbf{u}}$ and $\widetilde{\boldsymbol{\beta}}$. In this context, Robinson (1991, p.19) has shown that the solution for the random effects vector is given by

$$\widetilde{\mathbf{u}} = \boldsymbol{G}\mathbf{Z}^{\mathbf{T}}\boldsymbol{V}^{-\mathbf{1}}\left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right) \qquad (60)$$

Note that $\widetilde{\mathbf{u}}$ can be found only after $\widetilde{\boldsymbol{\beta}}$ has been determined. Thus the BLUP prediction of the random effects from Equation (52) can be solved once the appropriate estimators have been applied.

For a penalized spline model in mixed model representation, the appropriate degree of smoothing is determined by the estimated variance components σ_u^2 andσ_ε^2 and the two popular estimation techniques mentioned above, MLE and REML, can be applied. To illustrate this important and advantageous aspect of our approach, the log likelihood

[18] Ruppert, et al., 2003, p. 99, Robinson, 1991.

function used in estimating the variance components is derived. Here again we follow Ruppert, et al. (2003) very closely. Our justification for this rests in that the text itself is a comprehensive collection of approaches and applications.

The log likelihood function for estimating Equation (52) can be written as

$$\mathrm{L}(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2}[log|\boldsymbol{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + nlog(2\pi)] \qquad (61)$$

The process of MLE first optimizes Equation (61) over all $\boldsymbol{\beta}$**.** The result is the GLS estimator $\widetilde{\boldsymbol{\beta}}$ in Equation (59)[19]. Substituting the resulting GLS estimator $\widetilde{\boldsymbol{\beta}}$ into the log likelihood function results in the *profile log likelihood* function and can be written as

$$\mathrm{L}(\mathbf{V}) = -\frac{1}{2}\left[log|\boldsymbol{V}| + \mathbf{y}^{\mathrm{T}}\boldsymbol{V}^{-1}\left\{\boldsymbol{I} - \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{V}^{-\mathbf{1}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\right\}\boldsymbol{y}\right] - \frac{n}{2}log(2\pi) \qquad (62)$$

Thus, using MLE and optimizing Equation (62) over $\mathbf{V}$ results in parameter estimates for both $\boldsymbol{\beta}$ and $\mathbf{V}$.[20]

Implementing REML is more complicated than ML and is asymptotically equivalent. However, REML has the advantage of accounting for the degrees of freedom that are attributable to the fixed effects components in a mixed model. The restricted log likelihood function is given by

$$\mathrm{L}_R(\mathbf{V}) = \mathrm{L}(\mathbf{V}) - \frac{1}{2}log\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right| \qquad (63)$$

[19] PROC MIXED documentation in SAS version 9.3 confirms this. It also indicates that the Newton-Raphson algorithm is implemented as default and Lindstrom and Bates (1988) give details on why it is preferable.

[20] Much of the material on pages 48-53 here are a summary of Ruppert, et al (2003) pages 98 – 102.

In the case of a penalized spline with random spline coefficients, the variance components, σ_ε^2 and σ_u^2, are assumed known and therefore so is $\mathbf{V}$**,** Specifically, that $\mathbf{V} = \sigma_u^2\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \sigma_\varepsilon^2\mathbf{I}$. In this context, the solutions for $\widetilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ in Equations (59) and (60) can therefore be rewritten as

$$\tilde{\mathbf{u}} = \sigma_u^2\mathbf{Z}^{\mathbf{T}}\left(\sigma_u^2\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}\sigma_\varepsilon^2\right)^{-1}\left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right) \tag{64}$$

and

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\left(\sigma_u^2\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}\sigma_\varepsilon^2\right)^{-1}\mathbf{X}\right)^{-\mathbf{1}}\ \mathbf{X}^{\mathbf{T}}\left(\sigma_u^2\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}\sigma_\varepsilon^2\right)^{-1}\mathbf{y} \tag{65}$$

Equations (64) and (65) are a special case of the mixed model BLUP results but with the disturbance behavior and covariance structure specified in Equations (54), (55) and (56).

The connection between mixed models and penalized splines now becomes palatable. Brumback et al. (1999) showed that the solution of the penalized spline model in Equation (49) is exactly equivalent to the mixed model where the spline basis coefficients are treated as random effects[21]. Assuming that $\mathbf{y}|\mathbf{u} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}, \boldsymbol{R})$ and $\mathbf{u} \sim \mathrm{N}(0, \boldsymbol{G})$, i.e. as is the case of a random effects coefficient linear spline, then maximizing the log likelihood of $(\mathbf{y}, \mathbf{u})$ over the unknown $\boldsymbol{\beta}$ and $\mathbf{u}$ leads to the following criterion:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^{\mathrm{T}}\boldsymbol{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \quad + \mathbf{u}^{\mathrm{T}}\boldsymbol{G}^{-1}\mathbf{u} \tag{66}$$

Ruppert, et al. (2003) points out that Equation (66) is a just a generalized least squares criterion that is subject to the penalty $\mathbf{u}^{\mathrm{T}}\boldsymbol{G}^{-1}\mathbf{u}$**.** As such, the criterion involved with fitting a mixed-model also, in some manner, involves penalization.

[21] See also Ruppert and Carrol, *Spatially-Ddapted Penalties For Spline Fitting,* 1999.

Returning to the corresponding P-spine criterion in Equation (48) and noting that the spline coefficients in a mixed model representation are limited to the vector **u,** we divide the P-spline criterion by σ_ε^2 to form the following expression:

$$\frac{1}{\sigma_\varepsilon^2}\,(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})^{\mathrm{T}}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu}) \;\; +\frac{\lambda^{2p}}{\sigma_\varepsilon^2}\mathbf{u}^{\mathrm{T}}\mathbf{D}_{\mathrm{K}}\mathbf{u} \qquad (67)$$

where $\mathbf{D}_{\mathrm{K}}$ is defined as a $(K \times K)$ identity matrix. Compared to **D** in Equation (46) where only the spline coefficients are to be penalized, the penalization term in Equation (67) requires that every coefficient in **u** be penalized. As such, the identity matrix is used for the penalty matrix $\mathbf{D}_{\mathrm{K}}$.

If $cov(\mathbf{u}) = \sigma_u^2\mathbf{I} = \mathbf{G}$, (i.e., as is the case of a penalized spline) then its inverse $\mathbf{G}^{-1}$ in Equation (66) is just $\frac{1}{\sigma_u^2}\mathbf{I} = \frac{1}{\sigma_u^2}\mathbf{D}_{\mathrm{K}}$ **.** Analogously, if $cov(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2\mathbf{I} = \mathbf{R}$, then its inverse, $\mathbf{R}^{-1}$ is just $\frac{1}{\sigma_\varepsilon^2}\mathbf{I}$**.** Specifically, these relationships can be defined as

$$\mathbf{R}^{-1} = \begin{bmatrix} 1/\sigma_\varepsilon^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sigma_\varepsilon^2 \end{bmatrix} = \frac{1}{\sigma_\varepsilon^2}\mathbf{I}, \qquad \mathbf{G}^{-1} = \begin{bmatrix} 1/\sigma_u^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sigma_u^2 \end{bmatrix} = \frac{1}{\sigma_u^2}\mathbf{D}_{\mathrm{K}} \qquad (68)$$

It can now be seen that by defining the matrices **G** and **R** as is done in Equation (56) and substituting the specific form of these variance components into Equation (66) yields mixed model criterion identical to the penalized spline criterion. In this context, λ can be expressed explicitly in terms of the variance components:

$$\frac{\lambda^{2p}}{\sigma_\varepsilon^2} = \frac{1}{\sigma_u^2} \qquad (69)$$

Multiplying both sides by σ_ε^2 allows us to solve for lambda:

$$\lambda^{2p} = \frac{\sigma_\varepsilon^2}{\sigma_u^2} \tag{70}$$

Thus, the degree of smoothing for a penalized spline model can be chosen through likelihood-based estimation of σ_ε^2 and σ_u^2. To further appreciate the relationship between mixed models and P-splines, the fitted values that result from both can also be shown to be equivalent. For given solutions $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$, the BLUP of the response variable is given by

$$\hat{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}} \tag{71}$$

By denoting the combined design matrix of both fixed and random effects as $\mathbf{C} = [\mathbf{X}\ \mathbf{Z}]$ the fitted values of the response variable are shown to be

$$\hat{\mathbf{y}} = \mathbf{C}\left(\mathbf{C}^{\mathbf{T}}\mathbf{C} + \lambda\mathbf{D}\right)^{-1}\mathbf{C}^{\mathbf{T}}\mathbf{y} \tag{72}$$

where $\mathbf{D}$ is already defined. Brumback et al. (1999) note that this is exactly equivalent to the fitted values of the penalized spline smoother of Eilers and Marx (1996) found in Equation (48)

PROC MIXED in SAS can be used in smoothing applications with mixed models. The relationship between penalized splines and mixed-model theory allows for both REML and ML to be employed in parameter estimation as well as choosing the degree of smoothing. Our approach exploits this relationship and extends the case of a single smooth function to that of multiple smooth functional relationships.

3.7 Semiparametric Additive Models

The last two sections have shown how penalized splines can be used to estimate a smooth functional relationship and how they can be formulated as a mixed model. There

are many instances, however, when multiple predictors should be modeled in such a flexible fashion. In the previous section, we discussed the extension of simple regression to include multiple predictors. Whether a parametric, nonparametric or semiparametric model specification is employed, it is easy to see that each is a special case of an *additive model,* where the response variable is a simple summation of each predictor's effect and the residual. Ezekiel (1924) provided much of the early work on additive models while Friedman and Stuetzle (1981) continued. The simple assumption of additivity was explored by Stone (1985) and allowed for several nonparametric relationships to be investigated simultaneously. Additional flexibility was introduced with the class of Generalized Linear Models (Nelder and Wedderburn, 1972), which relaxes the distributional assumptions of the linear model[22]. However, it wasn't until 1990 that the seminal monograph *Generalized Additive Models* (GAM) by Hastie and Tibshirani took additive models and generalized them to other families of distributions. Since its publication, GAMs have been widely used and implemented in statistical packages such as S-PLUS and SAS.

An example of an additive model with three predictor variables $x_{1,i}, x_{2,i}$, and $x_{3,i}$ is given by

$$y_i = f(x_{1,i}, x_{2,i}, x_{3,i}) + \varepsilon_i = f_1(x_{1,i}) + f_2(x_{2,i}) + f_3(x_{3,i}) + \varepsilon_i \,, \qquad (73)$$

An additive model is a multivariate regression model where the functional form of each term is generalized. Specifically, the functions f_1, f_2, and $f_3$can *each* be estimated as smooth functions or be required to be linear, simultaneously. This flexibility allows the estimation of classical linear, nonparametric and semiparametric specifications.

[22] See section on Mixed-Model Theory in the PROC MIXED documentation in SAS version 9.3,

Our approach focuses on the inclusion of both linear ("parametric") and nonparametric terms in a regression context. For example, if the predictor variable $x_{1,i}$ in Equation (74) is modeled linearly while allowing smoothing functions for $x_{2,i}$ and $x_{3,i}$ then the subsequent specification would be

$$y_i = \beta_0 + \beta_1 x_{1,i} + f_2(x_{2,i}) + f_3(x_{3,i}) + \varepsilon_i \qquad (74)$$

When multiple smoothing functions require estimation, the classic approach and the one used by SAS and S-PLUS is referred to as the *backfitting algorithm.* This algorithm can be unattractive with large data sets as well as with many predictors. It is an iterative process that ultimately fits smoothing splines for each nonparametric component until the best overall fit is achieved. In this case, the smoothing parameter for each is chosen by a criterion such as Cross Validation (CV), Generalized Cross Validation (GCV), Akaike's Information Criterion (AIC) or its adjusted value (AICc). Each of these can require many potentially unnecessary calculations as indicated in section 3.6.

Alternatively, the smoothing functions that appear in Equation (74) can be fit using penalized splines. A semiparametric additive model using penalized splines results in the following equation:

$$y_i = \beta_0 + \beta_1 x_{1,i} + x_{2,i} \sum_{\kappa_2=1}^{K_2} u_{\kappa_2}^2 (x_{2,i} - k_{\kappa_2}^2)_+ + \sum_{\kappa_3=1}^{K_3} u_{\kappa_3}^3 (x_{3,i} - k_{\kappa_3}^3)_+ + \varepsilon_i \qquad (75)$$

where for a general predictor ω, the term $k_{\kappa_\omega}^{\omega}$ is a knot location associated with ω, indexed by $\kappa_\omega = 1, \ldots, K_\omega$, and $u_{\kappa_\omega}^{\omega}$ is the corresponding knot coefficient. (Note that each superscript in Equation (75) is *not* intended to indicate a polynomial power).

As we have shown, this approach can be accommodated in a mixed-model. The appropriate design matrices are

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,i} & x_{2,i} & x_{3,i} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} \end{bmatrix},$$

$$\mathbf{Z_2} = \begin{bmatrix} (x_{2,i} - k_1^2)_+ & \dots & (x_{2,i} - k_{K_2}^2)_+ \\ \vdots & \ddots & \vdots \\ (x_{2,n} - k_1^2)_+ & \vdots & (x_{2,n} - k_{K_2}^2)_+ \end{bmatrix}, \tag{76}$$

$$\mathbf{Z}_3 = \begin{bmatrix} (x_{3,i} - k_1^3)_+ & \dots & (x_{3,i} - k_{K_3}^3)_+ \\ \vdots & \ddots & \vdots \\ (x_{3,n} - k_1^3)_+ & \vdots & (x_{3,n} - k_{K_3}^3)_+ \end{bmatrix},$$

$$\mathbf{Z} = [\mathbf{Z}_2 \quad \mathbf{Z}_3]$$

where the dimensions of $\mathbf{X}$, $\mathbf{Z_2}$, $\mathbf{Z}_3$, and $\mathbf{Z}$ are (n x 4), (n x $k_{K_2}^2$), (n x $k_{K_3}^3$) and (n x ($k_{K_2}^2 + k_{K_3}^3$)), respectively. The corresponding mixed model representation and covariance structure is given by

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{77}$$

$$\boldsymbol{G} = \boldsymbol{cov}[\mathbf{u}] = \begin{bmatrix} \sigma_{x_2}^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_{x_3}^2 \mathbf{I} \end{bmatrix}, \qquad \boldsymbol{R} = \boldsymbol{cov}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2 \mathbf{I} \tag{78}$$

There are several advantages for fitting an additive model with a mixed model formulation. Firstly, the degree of smoothing for each nonparametric term can be determined through ML and REML. Analogous to the case of a single predictor, the degree of smoothing for a general predictor ω in this context is given by

$$\lambda_\omega = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\omega^2}} \tag{79}$$

By estimating the within-knot variance of each ω, an appropriate degree of smoothing proportional to the residual variance is determined. A second advantage is the reduced dimensionality that results from using low-rank smoothers. Stone (1985) defines dimensionality as "…the variance in estimation, 'the curse of dimensionality' being that the amount of data required to avoid an unacceptably large variance increases rapidly with increasing dimensionality".[23] Where smoothing splines require a basis function at each unique data point, they are referred to as being full rank. In contrast is the penalized spline where knot location and number can be chosen appropriately. Lastly, and as already discussed, the connection between penalized splines and mixed models allows for current mixed model software to be used in fitting a semiparametric additive model.

[23] Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, 689–705

CHAPTER 4 NEW ENGLAND SHORT – TERM LOAD FORECASTING MODEL

The previous section has developed a framework in the family of semiparametric additive models for making predictions. This chapter outlines and describes how this framework is applied to regional and zonal load forecasting in the New England region of the United States. While there exist numerous New England utilities and municipalities that use short term load forecasting, we focus on the prediction of regional load primarily done at ISO New England (ISONE). Modeling choices are discussed and the functional form for each candidate predictor is discussed.

4.1 Overview

The modeling approach here embodies that of a similar study by Fan and Hyndman (2010). Where sophisticated neural networks have become a popular modeling choice, Fan and Hyndman (2010) have applied semiparametric additive regression as an alternative forecasting framework for electricity demand in the short run. While significant differences exist between the modeling choices of Fan and Hyndman (2010) and those adopted here, both use semiparametric regression to forecast short term electricity load.

We start with the following simple relationship:

$$y_t = h(time) + \alpha(recent\ demand) + f(weather) + \varepsilon_t \qquad (80)$$

where electricity demand, y_t , in megawatts at hour t responds to calendar effects, dynamic effects, and prevailing weather effects represented in Equation (80) as $h(\cdot)$, $\alpha(\cdot)$, and $f(\cdot)$, respectively. While other studies have typically decomposed load into base and weather

sensitive components, this decomposition of load in Equation (80) follows that used in Fan and Hyndman (2010).

A primary objective of this modeling application is to avoid making unnecessary assumptions about the structural relationship between electricity demand and its short term driving forces. For each of the general effects listed above, functional forms for $h(\cdot)$, $\alpha(\cdot)$, and $f(\cdot)$ are explored and modeled first individually and then collectively in the family of additive models.

4.2 The Functional Form of $h(time)$

Calendar effects refer to how the time of year impacts electricity usage. The form of $h(time)$ is expressed as

$$h(time) = \delta_0 + \sum_{j=1}^{4} \delta_j D_{jt} + v N_t + \sum_{m=1}^{11} \gamma_m M_{mt} \qquad (81)$$

where

D_{jt} = an indicator, i.e. dummy, variable for each of four working days of the week, $j = 1\ to\ 4$[24], where $D_{jt} = 1$ for working day j and zero otherwise.

N_t = an indicator variable for nonworking days: Saturdays, Sundays or holidays, where $N_t = 1$ for a Saturday, Sunday or holiday and zero otherwise.

M_{mt} = an indicator variable for each of 11 months of the year, $m = 1\ to\ 11$, where M_{mt} = 1 for month m and zero otherwise.[25]

δ_0 = an intercept parameter.

δ_j = a parameter for each working day *j.*

v = a parameter for nonworking days.

γ_m = a parameter for each month *m.*

Federal holidays as well as weekend days are designated nonworking days. These are listed in the following table.

[24] Note: The effect of the fifth working day is captured in the intercept.
[25] Note: The effect of the twelve month is captured in the intercept.

Table 4-1: Dates considered to be Nonworking.

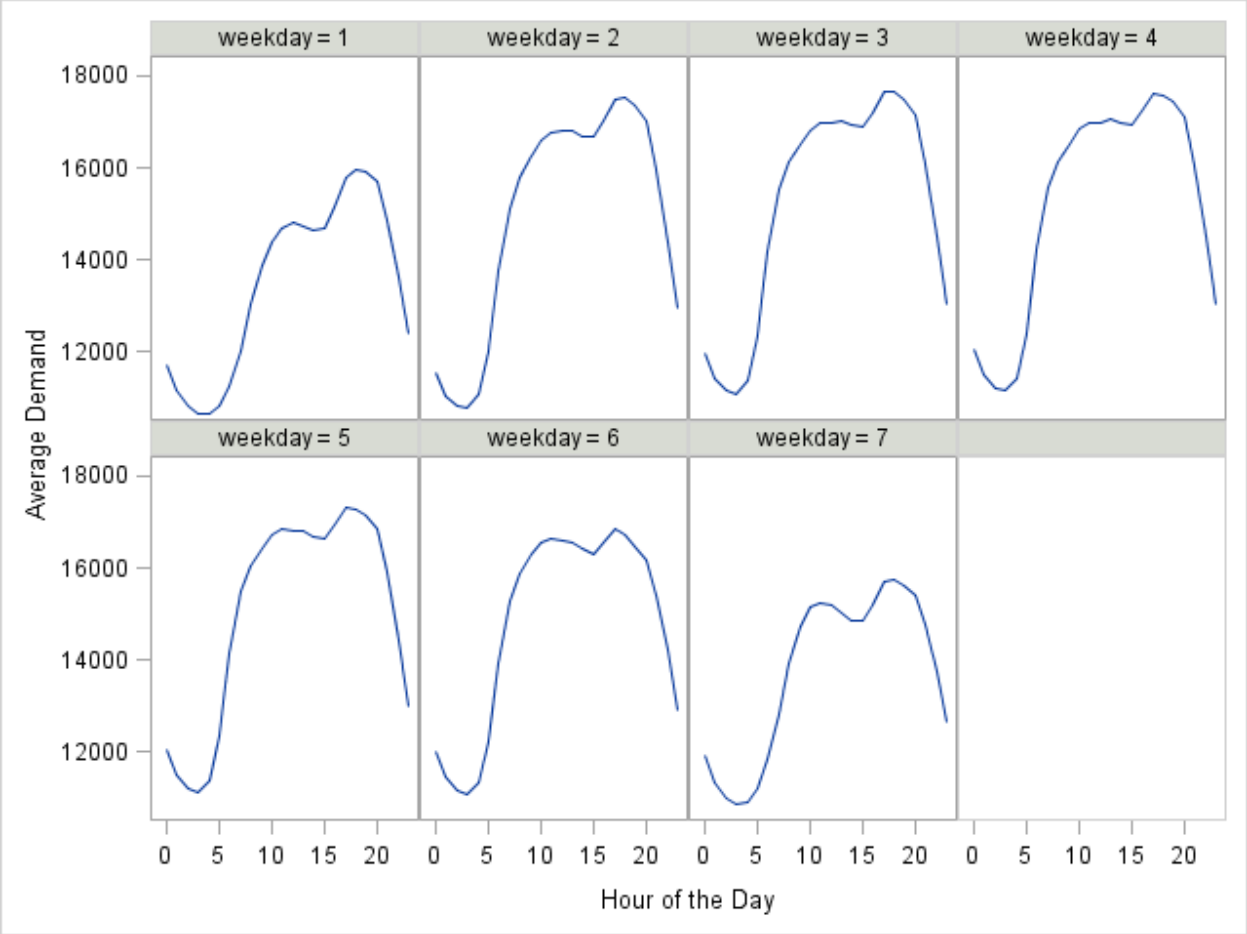
	Nonworking Days
(1)	Saturday
(2)	Sunday
(3)	New Year's Eve
(4)	New Year's Day
(5)	Martin Luther King Jr. Day
(6)	President's Day (Washington's Birthday)
(7)	Memorial Day
(8)	Independence Day
(9)	Labor Day
(10)	Thanksgiving Eve
(11)	Thanksgiving Day
(12)	Christmas Day

If the expression for $h(time)$ in Equation (81) is explicitly substituted into Equation (80), we have

$$y_t = \delta_0 + \sum_{j=1}^{4} \delta_j D_{jt} + v N_t + \sum_{m=1}^{11} \gamma_m M_{mt} + \alpha(recent\ demand) + f(weather) + \varepsilon_t \quad (82)$$

In modeling calendar terms, we make the assumption that day of the week, month of the year, and nonworking days have the effect of shifting electricity usage in a parallel fashion. Each parameter associated with the day-of-week and month dummy variables simply shifts the intercept of the model. Figure 11 plots average daily load curves by each weekday. This graphic was created using hourly observations of regional demand from January 1, 2009 to May 31, 2012. Hour 1 on the horizontal axis of each graph is equivalent to 1:00 a.m., hour 3 to 2:00 a.m., and so no. Each plot in Figure 11 retains the same general shape, yet each is of a slightly different size. A simple intercept shifter seems an adequate way to control for the differences among weekdays.

Figure 11: Plots of Average Hourly Demand in MW by Weekday

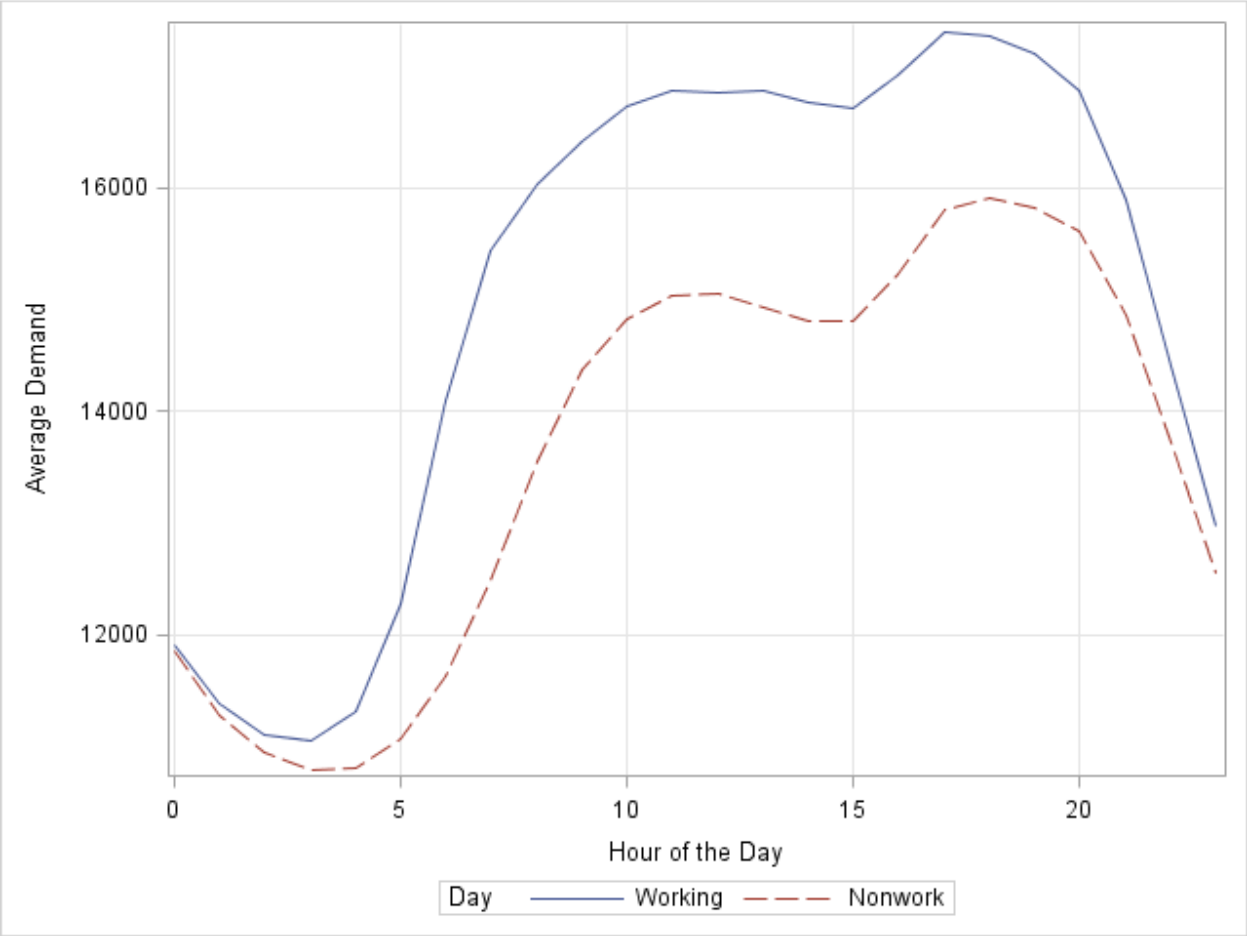


Note: Weekdays 1 and 7 correspond to the nonworking days of Saturday and Sunday. Their average load shapes are visibly less (or shorter) than that of weekdays 2 through 5 which are working days. This suggests that there is a diurnal effect influencing electricity load. Because industrial and commercial electricity usage across the New England region is reduced on holidays and weekends, the average load shape of nonworking days is different than that of the generic working day.

Figure 12 plots the average load for each hour from 2009 through 2011. Again, hour 1 on each horizontal axis signifies 1:00 a.m., hour 2 is 2:00 a.m., and so on. The delineation is

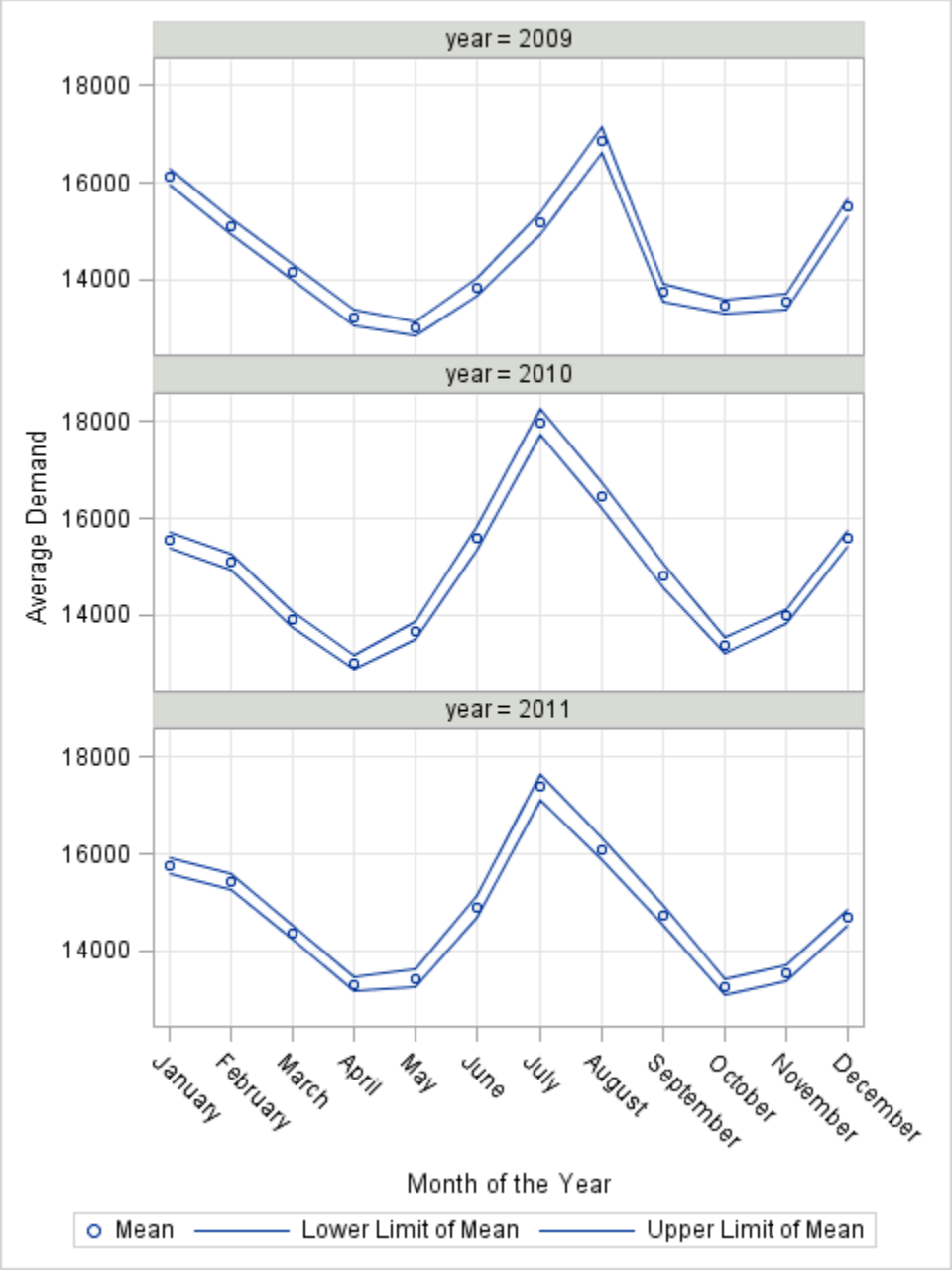
made between nonworking days and working days. The reduction of average hourly usage during a nonworking day is clearly visible. The visible impact of a nonworking day in Figure 12 is captured by the parameter v in Equation (82) with the expectation of a statistically significant negative shift in average demand.

Figure 12: Average Hourly Electricity Demand of Nonworking Days vs. Working Days.



There are also monthly effects that correspond to the changing seasons during a calendar year. Figure 13 displays the average load as well as the corresponding 95% upper and lower confidence limits for each month in 2009, 2010 and 2011, respectively.

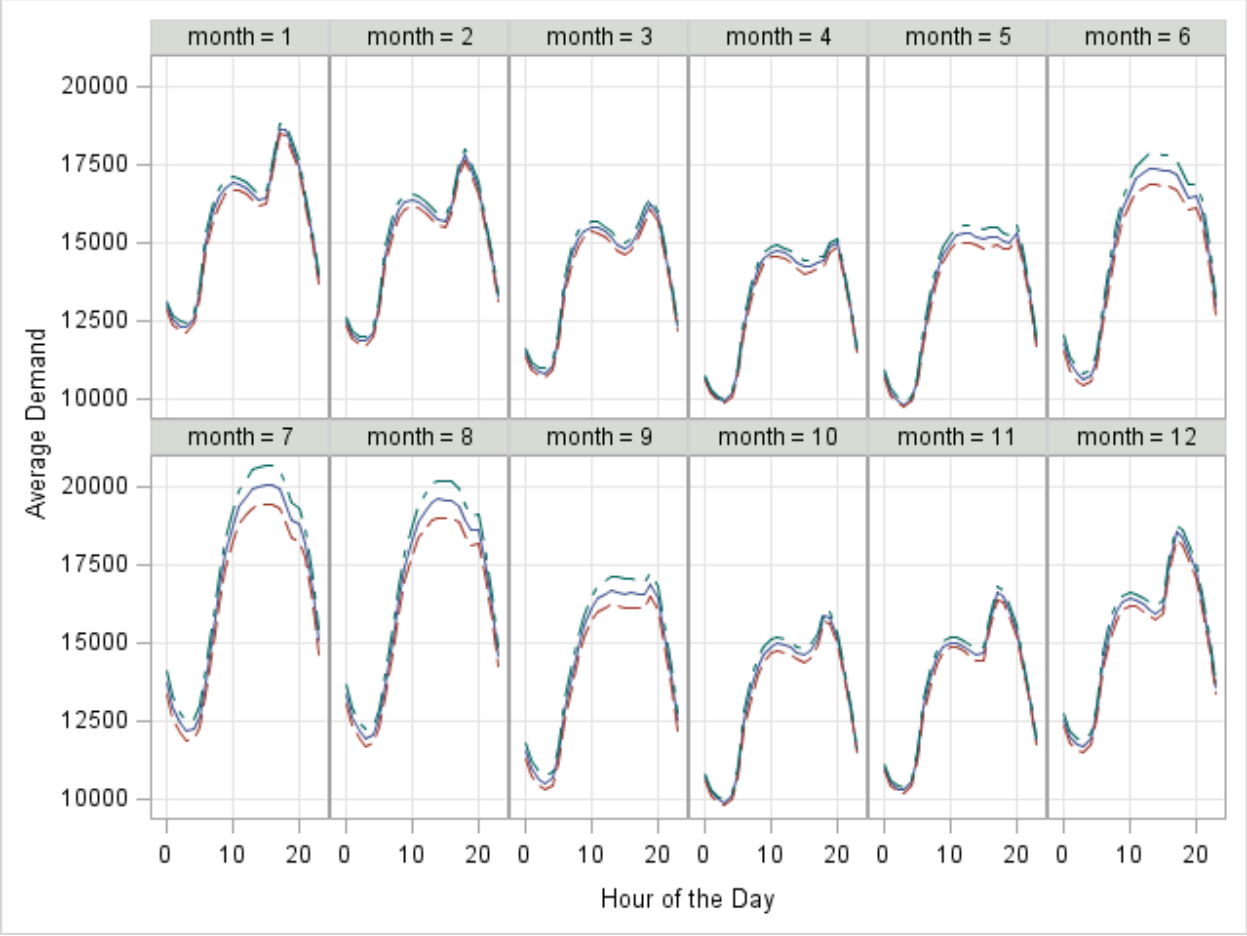
Figure 13: Average Monthly Load in New England for 2009, 2010 and 2011.



There is a significant change from each monthly average load to the next. New England seasonality is among the strongest in the world and has a direct impact on average monthly usage (Zielinski, 2003). The impact a particular calendar month m has on usage is treated as fixed and captured by the shift parameter γ_m in Equation (82).

It is important to note that treating each month as a fixed effect for each observation of load does not take into account the variety in daily load shapes from one season to another. In other words, there is evidence that each month has a fixed effect on average hourly electricity usage, but not that that effect is constant across each hour of the day. Figure 14 displays the average daily load shapes as well as the 95% upper and lower confidence limits for each month during 2009, 2010, and 2011. While the increase in average usage during summer and winter is visible and corroborates a fixed monthly effect, the average daily load shape takes a variety of forms.

Figure 14: Average Hourly Load Shapes by Month



The approach taken here with regard to calendar terms is identical to that of Fan and Hyndman (2010), as well as ANN and MetrixND models at ISONE. Fan and Hyndman (2010) address the variability in average daily load shape from month to month through the estimation of a separate model for each half-hour. We estimate a separate model for each hour of the day. This is discussed in section 4.5.

4.3 The Functional Form of $\alpha(\boldsymbol{recent\ demand})$

Regardless of the time of year, there is also a decaying effect of recent electricity usage when used specifically for cooling or heating. For buildings and structures, this includes the ability to retain heat and the subsequent speed of heat loss. Structure-specific characteristics such as these produce patterns of dynamic effects on future values of load. Take a typical winter weekday for example: the load which is used for electric heating at hour 10 (e.g., 10:00 a.m.) depends on how much heating related load was drawn from the grid at hour 8 (e.g., 8:00 a.m.). In other words, the amount of heating provided at 8:00 a.m. and how much of the heat remains insulated, has an effect on how much heating and therefore load will be needed at 10 a.m. Therefore, the future value of load depends directly on values of recent demand. ISO New England (ISONE) does not consider incorporate these structure-specific effects in its forecasts. Incorporating terms that capture these lagged effects can provide more accurate short term forecasts.

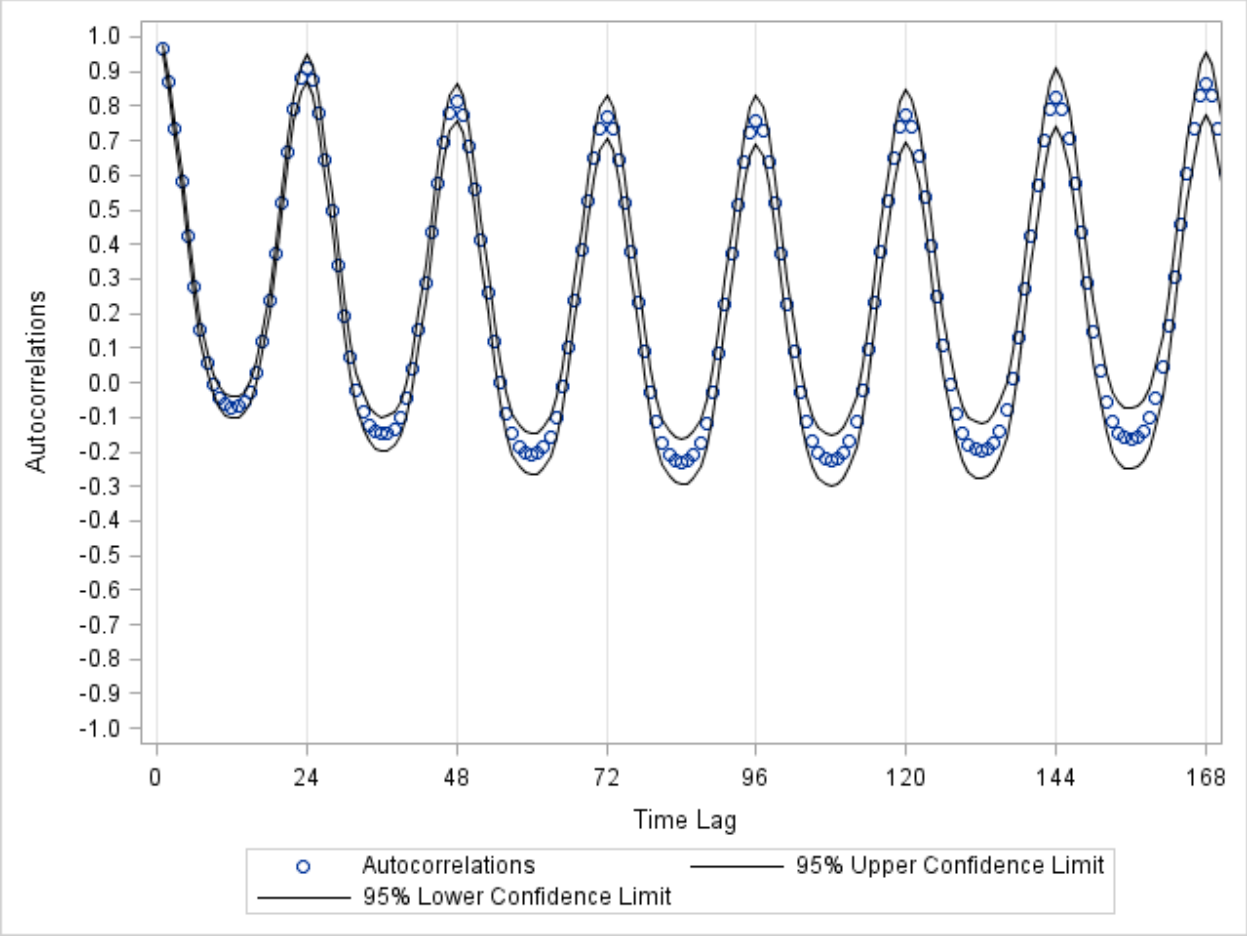
Our approach toward modeling dynamic effects of load follows Fan and Hyndman (2010). The term $\alpha(\cdot)$ in Equation (80) is designed to capture how recent values of electricity load can help predict future values in the short term. One can get a flavor for the

importance of lagged values by viewing the autocorrelation function (ACF) for electricity load.

An autocorrelation function is a construct that shows the relationship between a variable at time *t* and its lagged values (e.g., t-1, t-2, t-3, etc.). It is also flexible enough to show the potential relationship at patterned (i.e., "seasonal") intervals (e.g., every 24 hours – t-24, t-48, t-72, etc.). For example, it is reasonable that electricity usage 24 hours earlier, i.e., at t-24, would provide accurate predictions of the electricity usage at hour *t*. These *lagged effects* are due to the high degree of diurnal activity that exists in electricity usage. In other words, consumers of electricity typically exhibit consistent daily patterns of usage. Fan and Hyndman (2010) take this a step further and also use lagged values around the same time period from the previous two days.

This relationship is supported by observing the autocorrelation function for hourly electrical load. Analysis using the ACF as well as a generalized ACF modified to identify nonlinear association was done by Darbellay and Slama (2000). For our data, electricity usage at t-1 and at each 24 hour displacement provides the highest correlations. Figure 15 shows the correlation between load at any time t and each lag t-1, t-2, …, t-168. Irrespective of when *t* occurs, Figure 15 illustates that the highest correlation with t takes place with itself lagged 24 hours. This lagged effect is demonstrated with the repeated pattern out to 168 hours. Figure 15 plots the value of each correlation in addition to its 95% upper and lower confidence limits. 24 hour displacements remain the most statistically significant even at t-168.

Figure 15: Autocorrelation Function of Regional Electricity Demand



While lagged values may be useful predictors, these observations are not necessarily available when a prediction is to be made. In order to support the Day-Ahead market, ISONE is required to issue forecasted hourly values of regional demand before 10 am on the previous day. As such, there is a maximum of 38 and a minimum of 14 hours between when ISONE makes it forecasts (10 am) and when it must forecast load for the Day Ahead market (midnight the following day). Therefore, the most recent observation of load available at the time of forecasting is at *t-38*. However, the forecasts generated and described in Chapter 5 are limited to an ex post analysis. Our analysis focuses on creating a predictive model where the explanatory variables, including lagged demand, would be

known. In view of this, the most significant autocorrelations exist at 24 hour displacements starting at *t-24.*

We can get a preliminary idea as to how current load is related to lagged load by considering Figure 16. 17 and 18 (The data points in these figures represent plots of each observation of load against its value lagged 48, 72 and 96 hours, respectively. For instance a single data point in Figure 16 plots the value of y_t and y_{t-48}). All figures demonstrate a linear association between current load and loads lagged 48, 72 and 96 hours, respectively.

Figure 16: Scatter Plot of Load against Observations Lagged 48 Hours.

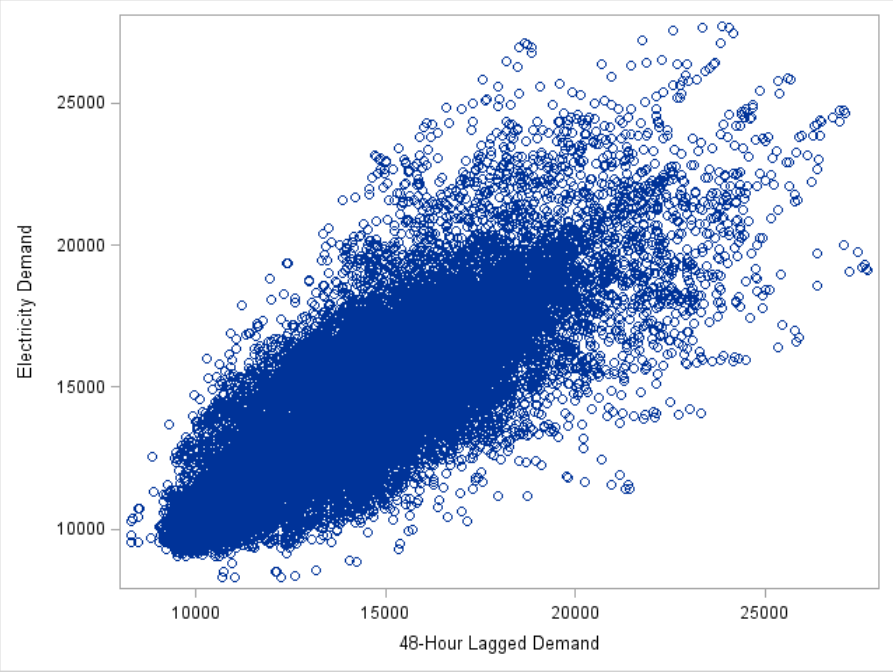


Figure 17: Scatter Plot of Load against Observations Lagged 72 Hours.

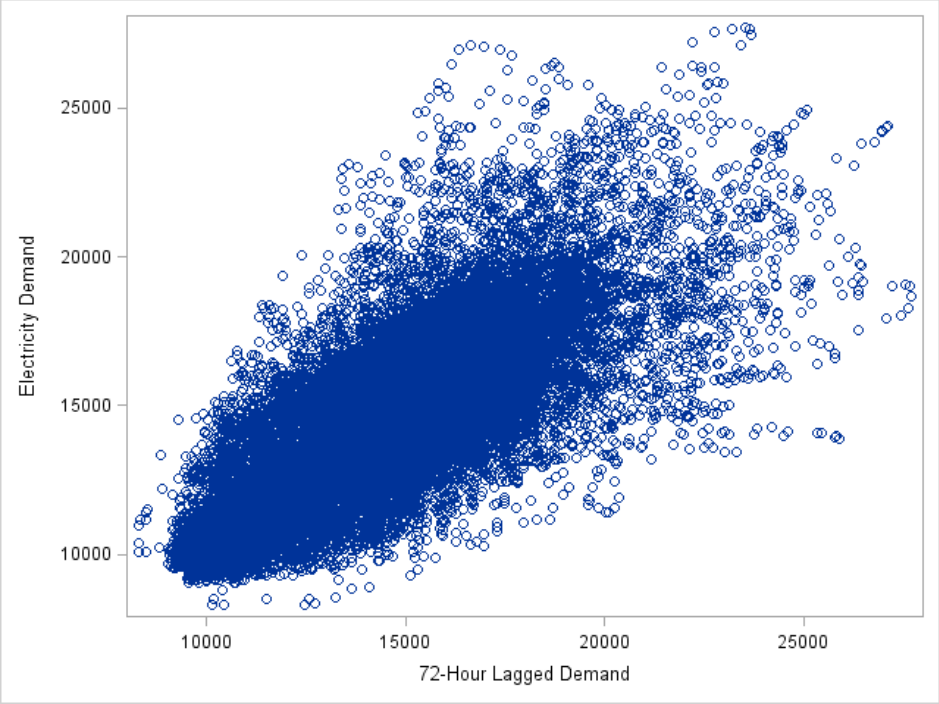


Figure 18: Scatter Plot of Load against Observations Lagged 96 Hours.

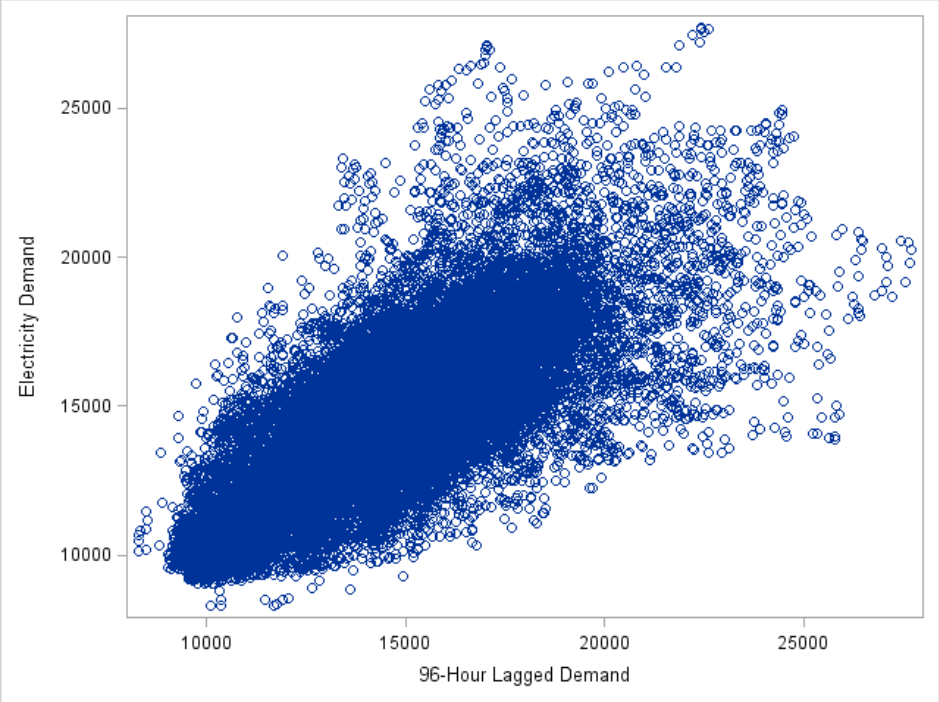


Table 4-2 reports the correlation coefficients for six different lags along with their probability values. These results suggest that it would be appropriate to model lagged load in the specification. Notice that these correlation coefficients are likewise represented by the peaks of the 24-hour periods for Figure 15.

Table 4-2: Correlation Statistics for Lagged Observations of Demand

Lag	Correlation Coefficient	P-Value
t-24	**.910**	**.000**
t-48	**.812**	**.000**
t-72	**.768**	**.000**
t-96	**.758**	**.000**
t-120	**.772**	**.000**
t-144	**.825**	**.000**
t-168	**.864**	**.000**

Using only those lagged values of demand in increments of 24 hours, $a(recent\ demand)$ in Equation (80) can be modeled as follows:

$$\begin{aligned} \alpha(\cdot) &= \alpha_1 y_{t-24} + \alpha_2 y_{t-48} + \alpha_3 y_{t-72} + \alpha_4 y_{t-96} + \alpha_5 y_{t-120} + \alpha_6 y_{t-144} \\ &\quad + \alpha_7 y_{t-168} \\ &= \sum_{i=1}^{7} \alpha_i y_{t-p} \end{aligned} \tag{83}$$

where

y_{t-p} = lagged demand at time t-p where $p = 24, 48, 72, 96, 120, 144, 168$.

α_i = a parameter for each lagged demand variable, $i = 1\ to\ 7$

If the expression for $a(recent\ demand)$ in Equation (83) is explicitly substituted into Equation (80) we have

$$y_t = \delta_0 + \sum_{j=1}^{4} \delta_j D_{jt} + v N_t + \sum_{m=1}^{11} \gamma_m M_{mt} + \sum_{i=1}^{7} \alpha_i y_{t-p} + f(weather) + \varepsilon_t \qquad (84)$$

4.4 The Functional Form of $f(weather)$

Weather components for load forecasting models currently used at ISONE are limited to temperature, humidity, wind speed and cloud cover. These four variables are known to be useful predictors for electricity usage (Soliman and Al-Kandari, 2010; Bunn and Farmer, 1985). They are also provided to ISONE from its three weather vendors to calculate the effective temperature (ET) and temperature-humidity index (THI) used for short term load forecasting. Aside from ISONE, it is also common to use observations of these variables to construct weighted indices as proxies for prevailing weather conditions (e.g., THI).

However, there is research that indicates the use of an index such as THI may not produce accurate electricity demand forecasts for the northeastern United States. Perhaps more so than any other region in the world, New England weather is among the most varied over such a small area (Zielinksi and Heim, 2003). It is affected by numerous sources of climate forces that operate on diurnal, annual and especially seasonal-scales. New England's proximity to the Atlantic Ocean as well as the coastal orientation of its populations has a strong influence on the degree of perceived comfort. These characteristics and patterns combine to create the sometimes chaotic tendencies of the region's climate.

There also exist several known patterns of high and low pressure systems, whose daily positions directly control temperature and precipitation in the region through their air

mass type. The development of global circulation systems, such as the subtropical Bermuda-Azores High and the subtropical Icelandic Low in the North Atlantic Ocean, can have a pronounced impact on regional climate. The former is known to bring warm and humid air to New England while the latter brings cool and humid air (Zielinski and Heim, 2003). In view of this, a single index constructed from observations of humidity and temperature may be a poor indicator for the actual weather condition.

In addition, there is growing evidence that historically consistent climatological patterns are changing across New England (Zielinksi and Heim, 2003, Clean Air – Cool Planet, 2005). This evidence prompts the search for a more flexible and robust method to incorporate weather information.

In this spirit, the functional form of f is generalized and given by

$$f(weather) = f_1(Temp_t) + f_2(Hum_t) + f_3(CC_t) + f_4(WS_t) \quad (85)$$

where

$Temp_t =$ drybulb temperature[26] in degrees Fahrenheit at hour t.

$Hum_t =$ relative humidity at hour t, measured by dewpoint.

$CC_t =$ amount of cloud cover at hour t, measured by the proportion of sky concealment.

$WS_t =$ wind speed in miles per hour at hour t.

The nonparametric representations for weather in Equation (85) allows for data-driven relationships to be estimated using observations of weather variables and load. Fitting this model through the use of penalized splines results in the following expression:

[26] Drybulb temperature in degrees Fahrenheit is a measurement of air temperature using a thermometer that is freely exposed to the air while being shielded from moisture.

$$f(weather) = \theta_1 Temp_t + \theta_2 Hum_t + \theta_3 CC_t + \theta_4 WS_t + \sum_{\kappa_T=1}^{K_T} u_{\kappa_T}^{T}(Temp_t - k_{\kappa_T}^{T})_+ + \sum_{\kappa_H=1}^{K_H} u_{\kappa_H}^{H}(Hum_t - k_{\kappa_H}^{H})_+ + \sum_{\kappa_{CC}=1}^{K_{CC}} u_{\kappa_{CC}}^{CC}(CC_t - k_{\kappa_{CC}}^{CC})_+ + \sum_{\kappa_{WS}=1}^{K_{WS}} u_{\kappa_{WS}}^{WS}(CC_t - k_{\kappa_{WS}}^{WS})_+ + \varepsilon_t \quad (86)$$

Where, for a general weather variable , the term $k_{\kappa_\omega}^{\omega}$ is a knot location associated with ω, indexed by $\kappa_\omega = 1, \dots, K_\omega$ and $u_{\kappa_\omega}^{\omega}$ is the corresponding knot coefficient.

Figure 19, 20, and 22 show plots of each weather variable against regional electricity from 2009 through 2011 and reveal separate relationships for each weather variable / load plot. Both temperature and humidity have a similar "U-shaped" relationship, while the plots of wind speed and cloud cover against load reveal a less defined relationship.

Figure 19: Electricity Load versus Temperature for 2009, 2010, and 2011.

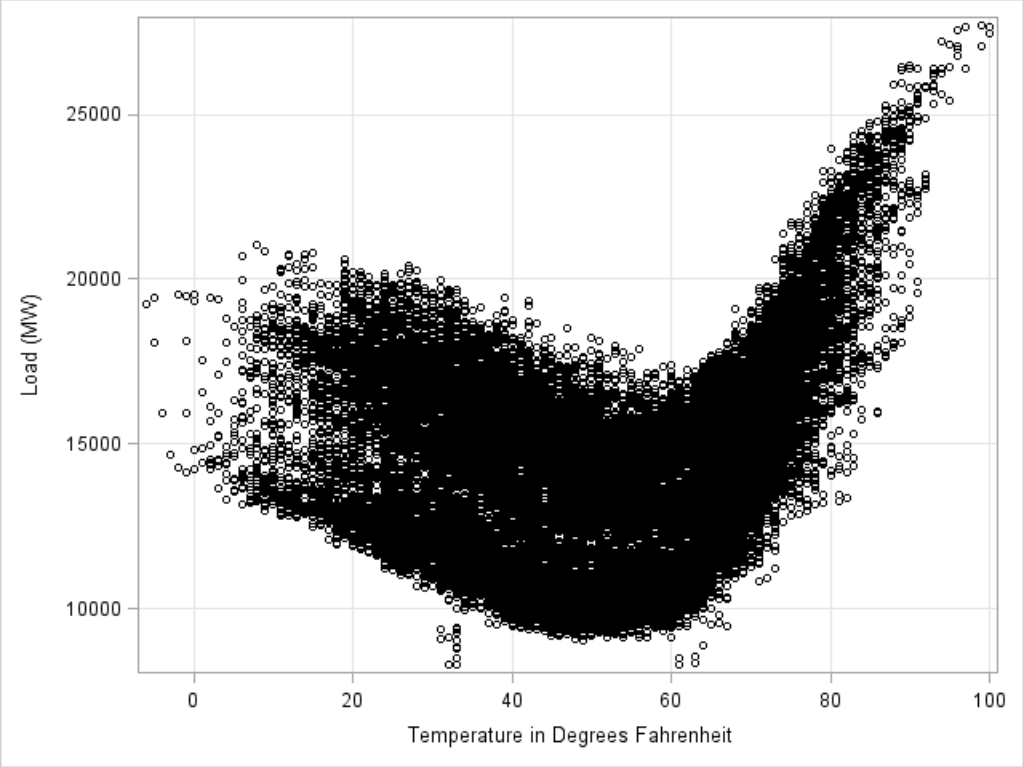


Figure 20: Electricity Load versus Humidity for 2009, 2010, and 2011.

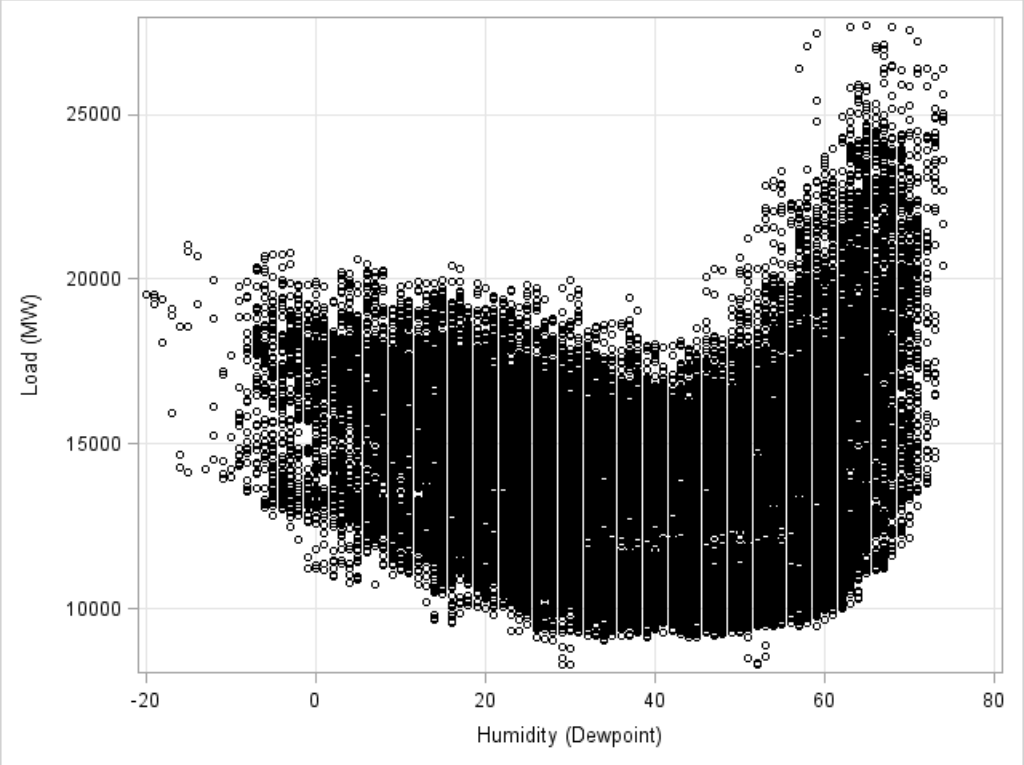


Figure 21: Electricity Load versus Wind Speed for 2009, 2010, and 2011.

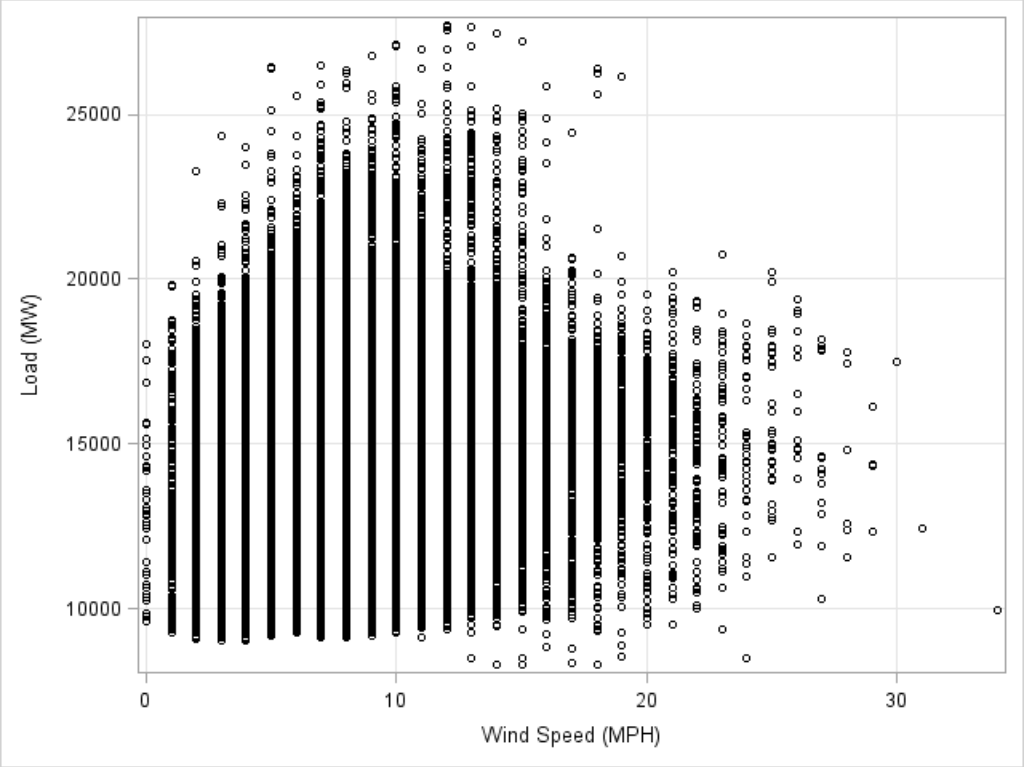
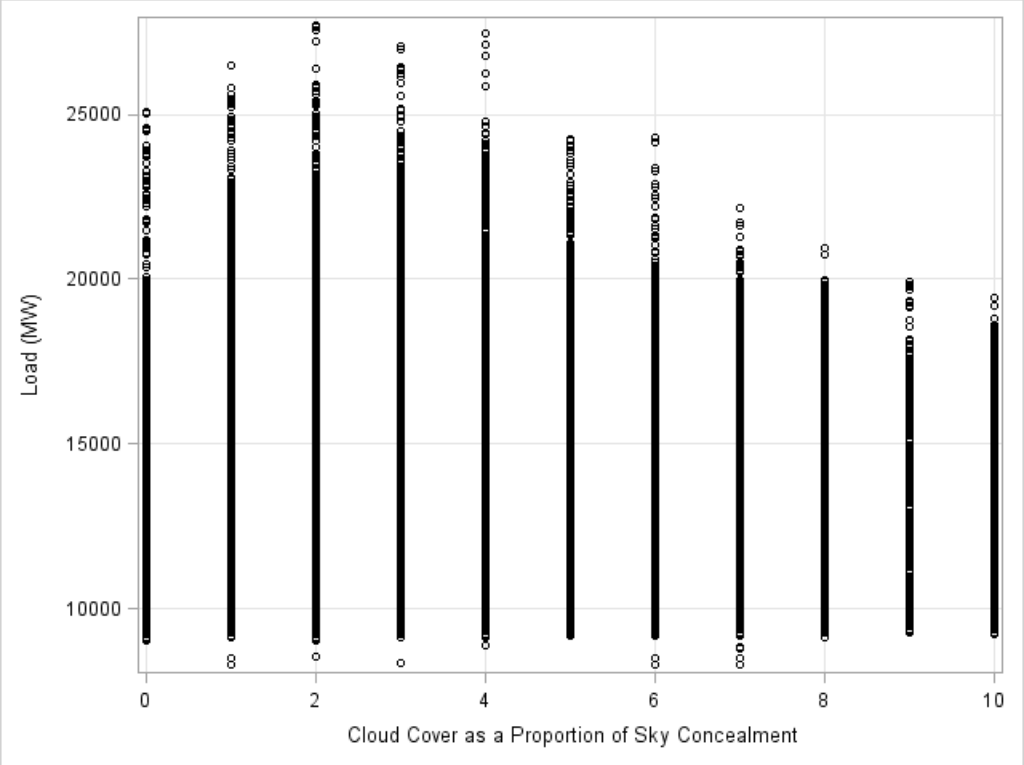


Figure 22: Electricity Load versus Cloud Cover for 2009, 2010, and 2011.



The lack of a clear relationship relating wind speed and load in Figure 21 and cloud cover and load in Figure 22 can be partially attributed to the manner in which the observed data were recorded. While temperature and humidity are recorded as degrees Fahrenheit, wind speed is measured by integers in terms of miles per hour while cloud cover is recorded as the proportion of sky concealment ranging from 0 to 10 and in unit increments 1. As such, both of these are essentially categorical variables that may hide variation in their load-weather relationships.

4.5 Semiparametric Forecasting Model

The analysis presented in the preceding three sections has investigated and addressed relationships between load and the short-term predictors. These relationships are complex, changing over time, and potentially highly nonlinear. In view of this, many advanced techniques have been applied to load forecasting as discussed in Chapter 2.

Alternatively, Fan and Hyndman (2011) propose a short-term load forecasting model that remains within the family of additive regression models and can accommodate nonlinear relationships through the use of regression splines. In addition to this recent work by Fan and Hyndman, a similar study by Engle, et al. (1986) uses a semiparametric specification for studying electricity demand. We take these ideas and extend them to models using penalized splines in the mixed-model framework discussed in Chapter 3.

Using the forms for $h(time)$, $\alpha(recent\ demand)$, and $f(weather)$ developed in the last three sections, a final semiparametric load forecasting model is constructed:

$$
\begin{aligned}
y_t = \delta_0 + \sum_{\mathrm{j}=1}^{4} \delta_{\mathrm{j}} \mathrm{D}_{\mathrm{jt}} + v\mathrm{N}_{\mathrm{t}} + \sum_{\mathrm{m}=1}^{11} \gamma_{\mathrm{m}} \mathrm{M}_{\mathrm{mt}} \\
+ \sum_{\mathrm{i}=1}^{7} \alpha_{\mathrm{i}} \mathrm{y}_{\mathrm{t-p}} + \theta_1 Temp_t + \theta_2 Hum_t + \theta_3 CC_t + \theta_4 WS_t \\
+ \sum_{\kappa_T=1}^{K_T} u_{\kappa_T}^{T} (Temp_t - k_{\kappa_T}^{T})_+ + \sum_{\kappa_H=1}^{K_H} u_{\kappa_H}^{H} (Hum_t - k_{\kappa_H}^{H})_+ \\
+ \sum_{\kappa_{CC}=1}^{K_{CC}} u_{\kappa_{CC}}^{CC} (CC_t - k_{\kappa_{CC}}^{CC})_+ + \sum_{\kappa_{WS}=1}^{K_{WS}} u_{\kappa_{WS}}^{WS} (CC_t - k_{\kappa_{WS}}^{WS})_+ + \varepsilon_t
\end{aligned}
\tag{87}
$$

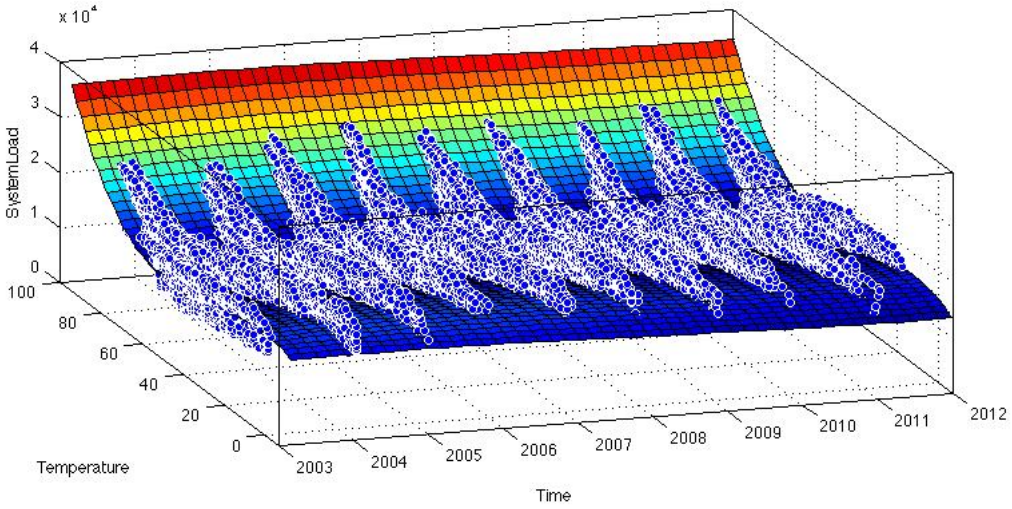
Fan and Hyndman (2011) successfully reduce prediction error around the peak periods by estimating 48 separate models for each half hour of the day. In their study, data were collected for the Australian Electricity Market (AEM). This market has half hour 'settlement' periods, and load is consequently metered at this frequency. The transmission grid in the corresponding region of Australia is also known for its extreme volatility where the within-day variation in electricity usage is high and demand patterns change over the course of a day. Similar load forecasting studies have shown that treating each half hourly or hourly period as separate is a good way to achieve better forecasts and also to partially mitigate the serial autocorrelation in the load series (Ramanathan et al., 1997; Fay et al., 2003). Accordingly, our data are hourly and we fit a separate model for each hour of the day.

4.5.1 Description of the Data

ISONE makes available a large amount of data related to historical electricity usage in New England and the zones that collectively form the regional transmission grid. Hourly

data sets containing observations of metered demand, price and weather conditions are available from March 1st, 2003 to date[27]. Plots of these hourly data allow for visual inspection of the load series. The surface representation of regional load against time and temperature in Figure 23 displays the seasonal patterns of New England electricity usage as well as confirms a consistently nonlinear weather relationship over time.

Figure 23: Surface Representation of Regional Load against Time and Temperature



While not clearly evident in Figure 23, there is also significant change in regional usage patterns from year to year. Figure 24, 25 and Figure 26 plot the annual load patterns over time for the years 2009, 2010, and 2011, respectively. These figures illustrate that the summer months (June, July and August) are when electricity usage is highest as well as most variable. 2009 appears to have a large summer peak occurring in late August, while 2010 and 2011 have summer peaks in July. In addition, 2010 has a large spike in the beginning of September and in late May. 2011 also has a similar spike in June.

[27] The excel files containing these historical data are constantly updated on the ISONE website.

Figure 24: Regional Load in New England, 2009.

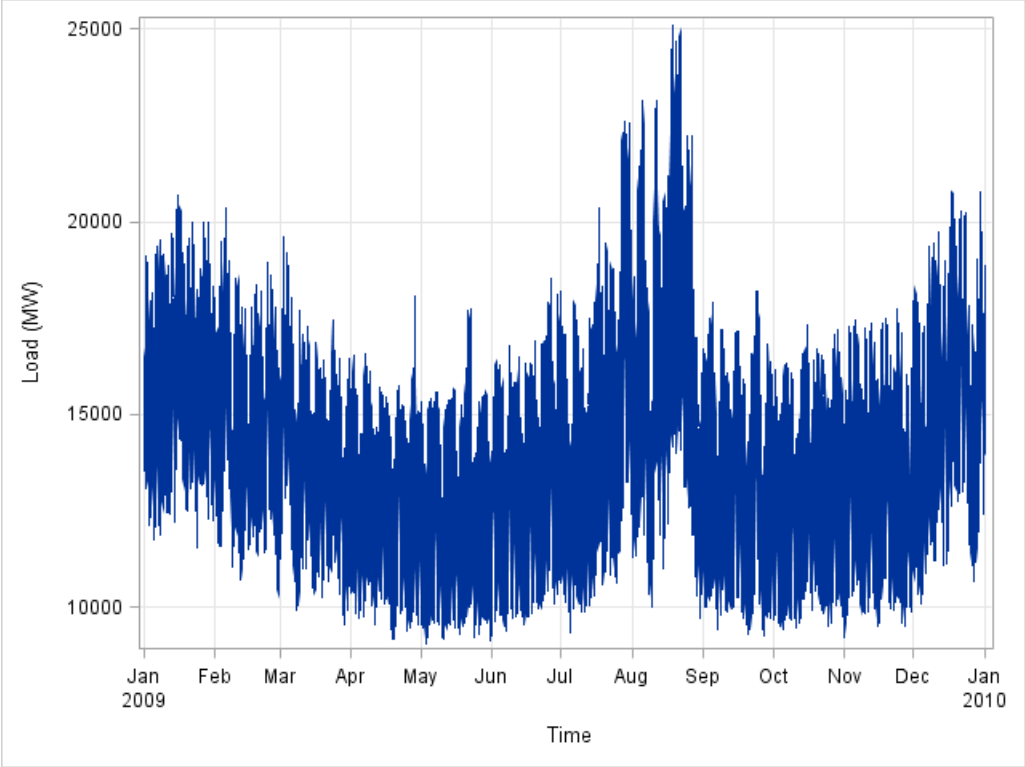


Figure 25: Regional Load in New England, 2010.

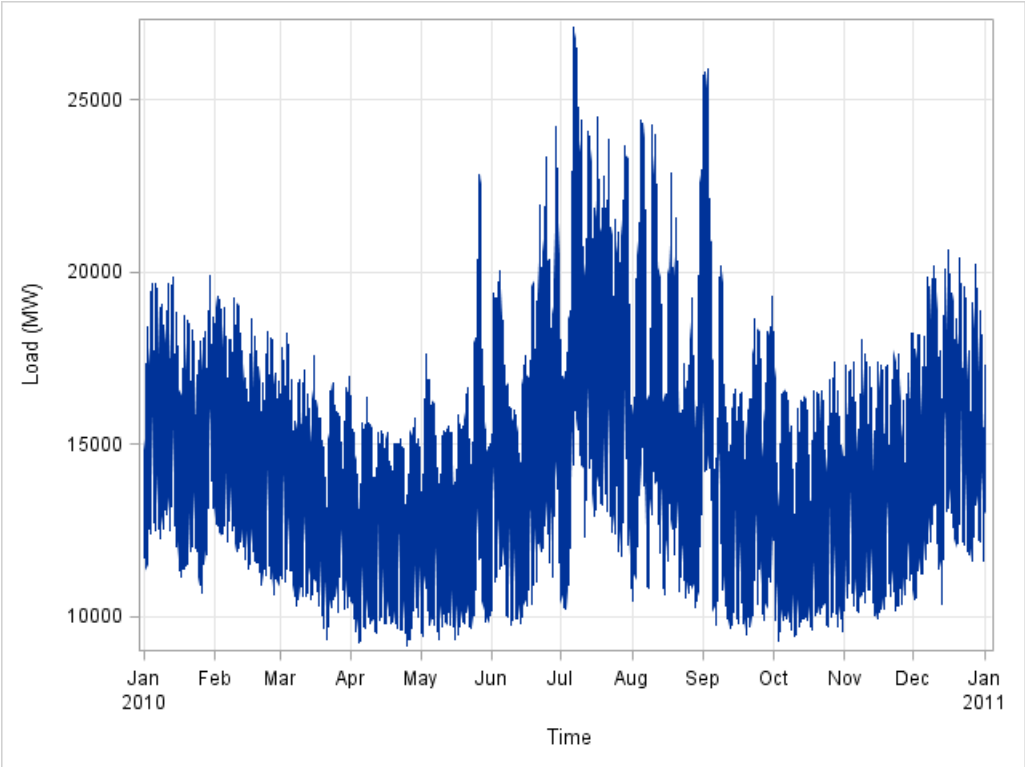
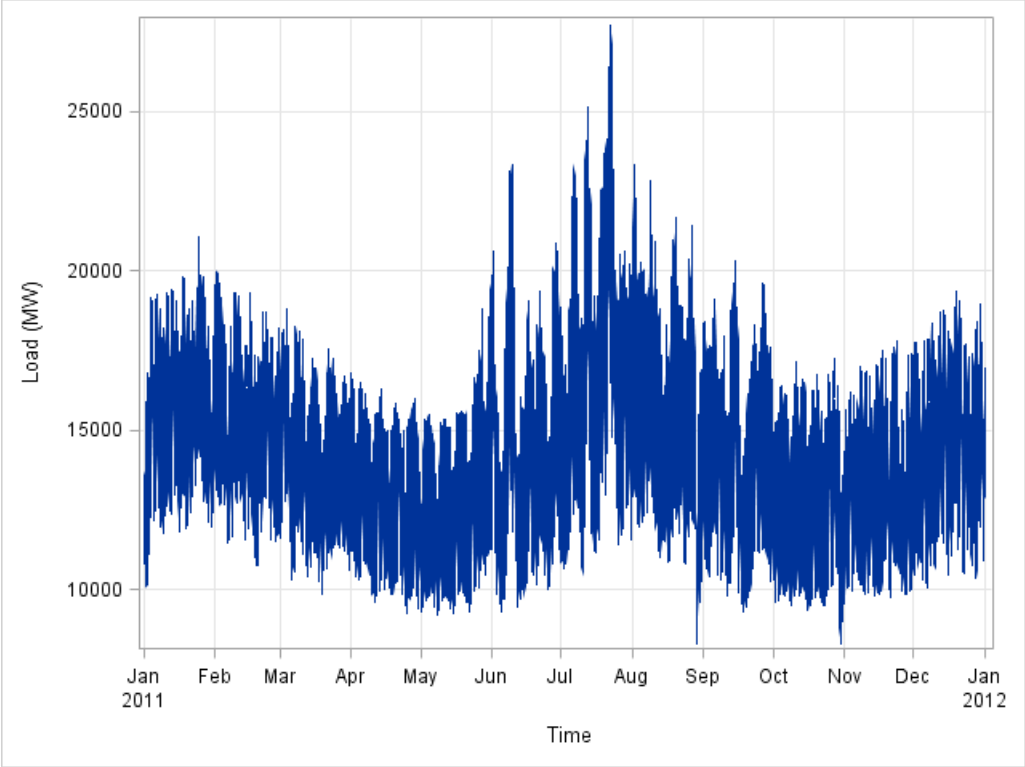


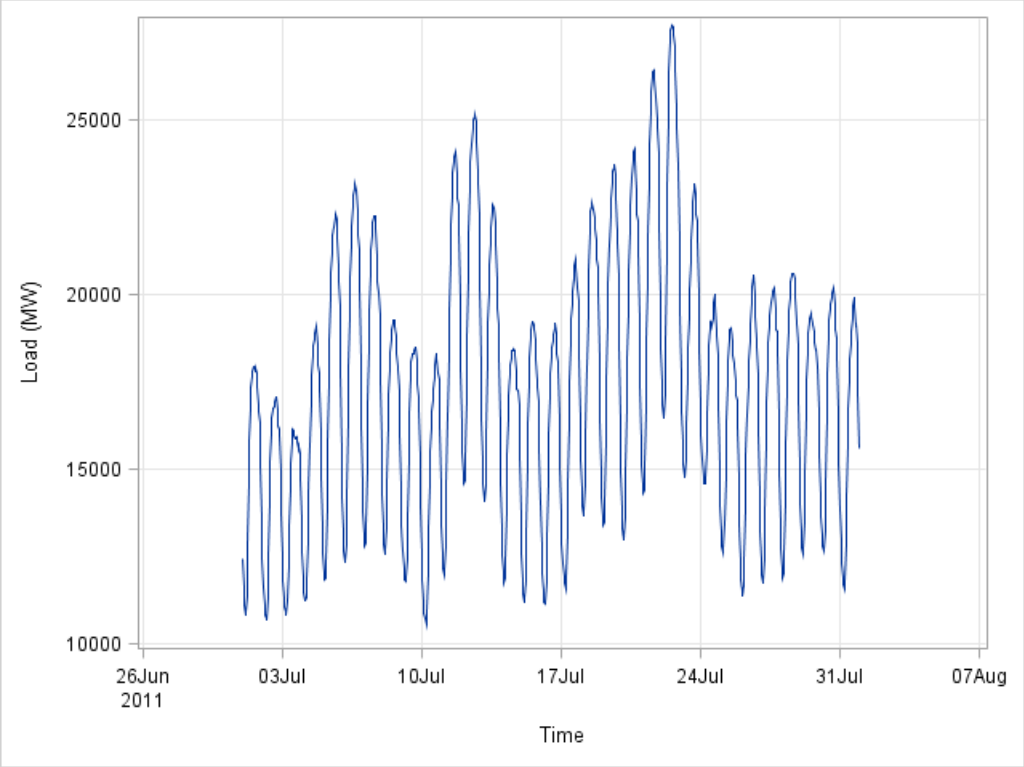
Figure 26: Regional Load in New England, 2011.



Also evident in the annual plots of load is the apparent propensity toward high winter usage in addition to the high summer usage. This is attributable to electric heating and extended hours of darkness. We note that the variation of usage over the winter is not as large as the summer, however.

The high variability that characterizes New England summer load is a daily phenomenon. Figure 27 displays daily load shapes for each day during July, 2011. The presence of large intraday variation as well as between-day variation is evident.

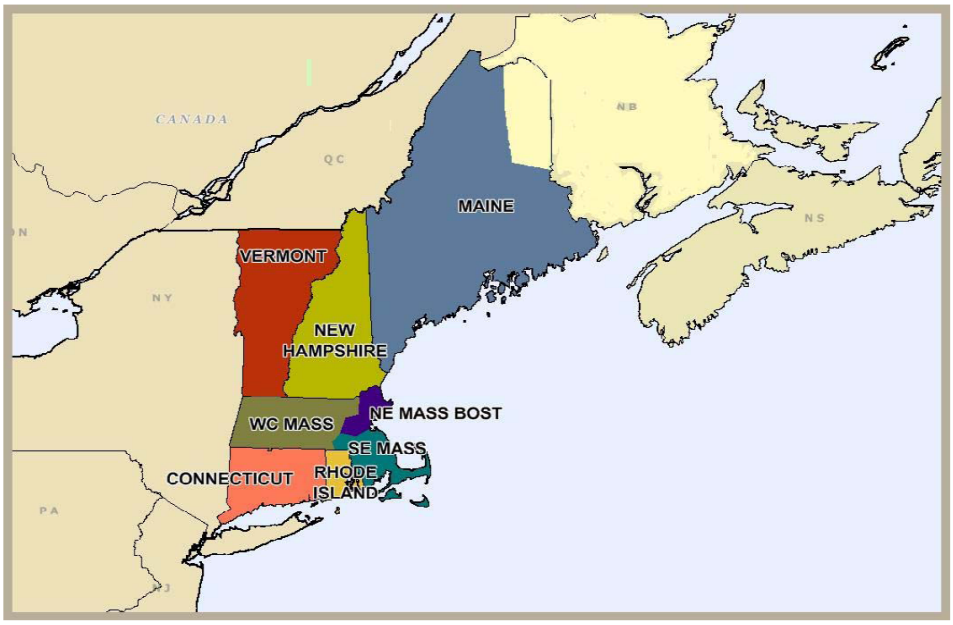
Figure 27: Regional Load during July, 2011.



The diminished use of electrical power during the Independence Day holiday is evident in Figure 27. In fact, this holiday has such a pronounced effect that its hourly values of electricity consumption are the smallest over the entire month of July.

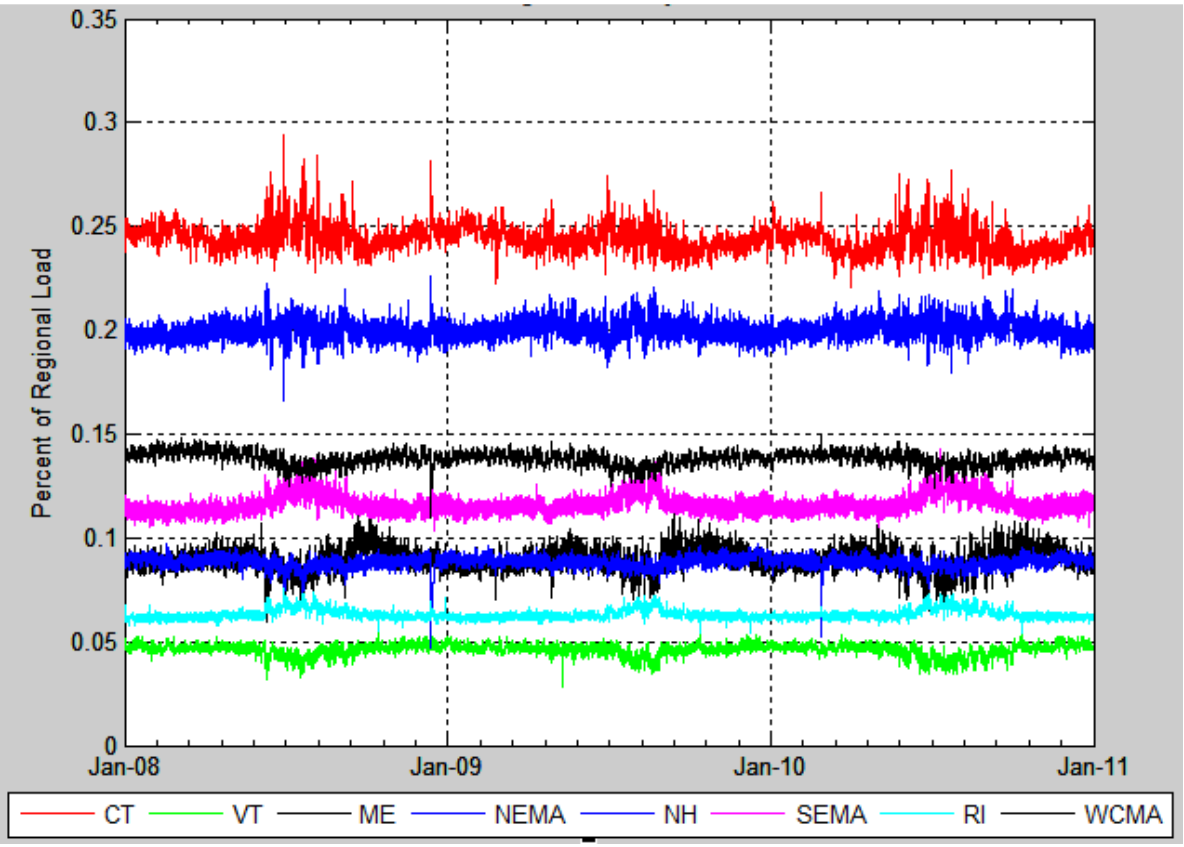
So far, we have focused on the New England regional load series. However, as already mentioned, this load is composed of eight different zones around the region. The individual observations of load for each of the eight load zones in New England are determined by metering. Weather observations are collected from eight weather stations across New England. The states of Connecticut, Rhode Island, Vermont, New Hampshire and Maine each represent a single load zone while Massachusetts is divided into three: northeast (NEMASS), southeast (SEMASS), and western and central (WCMASS). Figure 28 displays the locations of all eight load zones in the region.

Figure 28: Geographical Representation of New England Load Zones.



While Maine is the largest load zone by land mass, NEMASS and Connecticut generally have much higher loads than any other zone due to population density. Varying levels of commercial, industrial and residential electricity usage as well as the potential for drastically different climate conditions across the region can also contribute to zonal differences. Figure 29 displays each zone's load as a percent of regional load from January 1st, 2008 to January 1st, 2011 as a means to graphically present the heterogeneity of usage across load zones and over time.

Figure 29: Breakdown of Zonal Load as a Percentage of Regional Load



Over the course of each calendar year, NEMASS and Connecticut load zones account for the majority of regional load, while the other zones fluctuate relative to each other. This zonal variation is due to state specific weather factors, populations and demographics. None of these zonal specific characteristics are currently analyzed for ISONE load forecasting.

Regardless, the observations of regional weather conditions used for ISONE's load forecasts are calculated on a weighted basis from eight weather stations across the region.

These weights are static and determined by historical electricity sales data[28]. Table 4-3 reports each of the weather stations and their corresponding weight.

Table 4-3: ISONE Weather Stations

Variable	*State*	*Station Code*	*Closest Load Zone*	*Weight*
Boston	MA	BOS	NEMASS	0.208
Bridgeport	CT	BDR	--	0.277
Burlington	VT	BTV	Vermont	0.073
Concord	NH	CON	New Hampshire	0.212
Portland	ME	PWM	Maine	0.049
Providence	RI	PVD	Rhode Island & SEMASS	0.057
Windsor Locks	CT	BDL	Connecticut	0.043
Worcester	MA	ORH	WCMASS	0.084

Additional data were collected from the ISONE Forecasting Office. In addition to the hourly temperature and humidity observations available through data sets on the ISONE website, cloud cover and wind speed observations were obtained through this extended data set. However, while historical data back to 2003 are available on the website, the range of this extended data set spans from January 1st, 2009 to May 31st, 2012. As such, an hourly data set of load and weather observations is available for each load zone and the region during this time frame. Table 4-4 reports summary statistics for the regional data set while Appendix A contains replicated tables for each load zone.

Table 4-4: Summary Statistics for Regional Dataset

Variable	*Sample Size (n)*	*Mean*	*Standard Deviation*	*Median*	*Minimum*	*Maximum*
Load	29,928	14,620.00	2,855.00	14,766.00	8,296.00	27,707.00
Temperature	29,928	49.87	18.05	50.00	-6.00	100.00
Humidity	29,928	38.06	19.02	39.00	-20.00	74.00
Wind Speed	29,928	8.62	4.32	8.00	0.00	34.00
Cloud Cover	29,928	4.22	2.79	4.00	0.00	10.00

[28] Each ISONE data set contains a data dictionary and explains these calculations. For instance, see http://www.iso-ne.com/markets/hstdata/znl_info/hourly/index.html

4.5.2 Regional and Zonal Predictions

Currently, day-ahead forecasts made by ISONE are limited to regional demand values and forecasts are not made on a zonal basis. Given the potential for weather variation across the region and the high influence that weather conditions have on electricity usage, using a weighted temperature variable for a region like New England may not be ideal. We expect that producing short term forecasts for load in a specific zone can also provide desirable information for unit commitment processes as each zone is characterized by a different mix of generating units and thus different fuel types and supplies. In addition to improving reliability of the system as a whole, a deeper understanding of future zonal requirements can mitigate financial and economic losses that stem from prediction error.

Given the access to zonal data sets as well as the potential benefit from producing zonal forecasts, Equation (87) was fit for each hour and each load zone as well as the region as a whole. Collecting and aggregating the zonal forecasts produced a second set of forecasts for the region in addition to those outputted from the regional model. The extended data set mentioned in section 4.5.1 was divided into within-sample (training) and out-of-sample (forecasting) periods. Specifically, each forecasting model was trained using data from 2009 through 2010 and used to make forecasts for 2011. Data from 2012 were not included in this analysis as it did not constitute an entire year of data

CHAPTER 5 DISCUSSION OF RESULTS

This chapter discusses the results of fitting forecasting models for the region as well as for each load zone. Estimating hourly models produced a great deal of output as well as diagnostic measures. We present the results of fitting these models with the series of tables below. The relevant fit statistics, variance component estimates, and test statistics are presented and discussed. In section 5.2, forecasting performance is discussed while Appendix B contains the detailed results from each load zone's respective set of hourly models. The predictions provided by the zonal forecasting models are then aggregated across load zones to form a second regional forecast comparable to that provided by the regional forecasting model. The forecasting performance of both is then evaluated and compared.

5.1 Model Estimation Results

For each nonparametric component, knot placement and number were chosen following the default choices described in Wand (2002) and defined in Ngo and Wand (2004).

Specifically, a default rule for knot locations can be expressed by

$$\kappa_k = \left\{\frac{k+1}{K+2}\right\} th\ sample\ quantile\ of\ the\ unique\ \mathrm{x}_i's, \quad \text{for } k = 1, \ldots, K \tag{88}$$

where k here indexes the knot locations and x_i denotes observation i of (weather) variable x.

Analogously, a reasonable choice for the number of knot locations is suggested by Wand and given by

$$K = \max(5, \min\left(\frac{1}{4} * number\ of\ unique\ x_i's, 35\right)) \quad (89)$$

The components appearing in Equation (**89**) are intended to employ a minimum number of knots required to capture structural change (e.g., 5 in this case) but also to set an upper threshold where including additional knots may be unnecessary or inappropriate. That threshold is chosen to be the minimum of either ¼ the number of unique values of an observed variable (e.g., temperature) or 35.

Another method of selecting the number of knots as well as the knot locations is detailed in Ngo and Wand (2004). The SAS code published in their study provides an algorithm that selects knots at equally spaced intervals while limiting the total number of knots. Given a series of observations, the algorithm to determine the number of knots and their location is as follows:

1. Determine unique values fir the observed variable.
2. Calculate the number of unique values divided by 5 and set as the total number of knots.
3. If the calculated total number of knots is greater than or equal to 40, then set the total number of knots at 40.

4. Calculate the interval between knots by taking the number of unique values and dividing by the chosen total number of knots. Round to the nearest integer.
5. Output knot locations based on the calculated total number of knots and the calculated interval between each knot.

The SAS code for this algorithm can be found in Appendix C. This approach is more arbitrary and, unlike the default rules provided in Equations (88) and (89), does not explicitly take into consideration the distribution of the observed variable. In other words, whereas Equation (88) employs the sample quantiles of the observed variable to determine knot location, the algorithm provided in Ngo and Wand (2004) does not.

However, Wand (2002) and Ruppert, et al. (2003) discuss how knot number and placement is not a critical modeling choice for penalized splines as long as a sufficiently large number of knots is chosen to capture the underlying structure. As such, the SAS code provided in Ngo and Wand (2004) was used to determine knot location as well as the number of knots for each nonparametric term.

Both ML and REML estimation approaches estimation were attempted. As REML is asymptotically equivalent to ML, models that were initially estimated with ML resulted in the same estimates of λ (Equation 79, p. 69) as those estimated with REML. However, the REML method was ultimately chosen for estimation as it accounts for the degrees of freedom that are attributable to fixed effect terms within each model. It is this set of estimates that facilitates smoothing of weather predictors for each hourly model. As such, the resulting degree of smoothing for each weather predictor in each model is enumerated and characterized by its variance components. As stated in section 4.5.2, models were

estimated using 2009 and 2010 data (i.e., the training or within sample period). For each model, the estimated variance component associated with each weather variable is provided in Table 5-1.

Table 5-1: Estimated Variance Components and Smoothing Parameters for Each Hourly Regional Model.

Hour	$\widehat{\sigma}^2_{temp}$	$\widehat{\sigma}^2_{hum}$	$\widehat{\sigma}^2_{cc}$	$\widehat{\sigma}^2_{ws}$	$\widehat{\sigma}^2_{\varepsilon}$	λ_{temp}	λ_{hum}	λ_{cc}	λ_{ws}
1	5,074.11	23.77	1,603.96	316.65	64,911.11	3.58	52.26	6.36	14.32
2	4,258.03	24.40	1,150.10	735.49	54,667.34	3.58	47.34	6.89	8.62
3	3,446.45	17.27	1,452.47	287.63	49,940.65	3.81	53.77	5.86	13.18
4	3,485.75	22.83	2,086.24	228.35	48,718.09	3.74	46.20	4.83	14.61
5	3,708.99	15.68	1,915.13	359.68	44,610.78	3.47	53.33	4.83	11.14
6	3,941.04	19.42	4,566.87	292.43	52,941.15	3.67	52.22	3.40	13.46
7	4,235.41	38.37	6,435.55	0	100,801.40	4.88	51.26	3.96	0.00
8	4,289.13	471.98	0	577.07	132,351.60	5.55	16.75	0.00	15.14
9	4,824.26	682.31	5,209.49	0	113,130.70	4.84	12.88	4.66	0.00
10	5,675.81	782.17	616.55	0	109,386.30	4.39	11.83	13.32	0.00
11	7,124.34	961.81	1,083.45	0	108,453.60	3.90	10.62	10.01	0.00
12	4,082.52	1,156.83	0	396.34	130,403.90	5.65	10.62	0.00	18.14
13	5,341.91	1,438.57	0	1195.75	166,844.20	5.59	10.77	0.00	11.81
14	10,408.50	1,005.18	1,940.03	0	125,932.50	3.48	11.19	8.06	0.00
15	5,478.17	810.23	0	0	234,358.00	6.54	17.01	0.00	0.00
16	3,678.97	887.27	0	46.69	257,351.20	8.36	17.03	0.00	74.24
17	3,645.86	916.63	0	371.12	265,674.40	8.54	17.02	0.00	26.76
18	4,255.34	1,179.68	0	259.07	249,963.40	7.66	14.56	0.00	31.06
19	8,681.27	1,509.72	1,746.25	85.59	208,111.10	4.90	11.74	10.92	49.31
20	4,316.87	1,299.82	0	2,531.8	174,070.50	6.35	11.57	0.00	8.29
21	4,324.94	752.22	1,871.60	0	157,067.70	6.03	14.45	9.16	0.00
22	3,962.84	351.81	846.02	96.27	109,865.70	5.27	17.67	11.40	33.78
23	2,988.46	127.05	81.44	345.32	72,452.32	4.92	23.88	29.83	14.48
24	5,561.27	51.32	4,226.64	256.03	74,050.94	3.65	37.98	4.19	17.01

To illustrate the estimation results presented in Table 5-1, we consider the estimated variance components for hour 14 (i.e., at 2 pm) as an example. The data used to fit this model were observed at 2 p.m. when daily load is typically at its diurnal peak. For each nonparametric component of the hour 14 forecasting model, the estimated variance components yield the following smoothing parameter estimates:

$$\lambda_{temp} = \sqrt{\frac{\hat{\sigma}_{\varepsilon}^2}{\hat{\sigma}_{temp}^2}} = \frac{125{,}932.45}{10{,}408.53} = \ 3.48 \tag{90}$$

$$\lambda_{hum} = \sqrt{\frac{\hat{\sigma}_{\varepsilon}^2}{\hat{\sigma}_{hum}^2}} = \ \frac{125{,}932.45}{1{,}005.18} = \ 11.19 \tag{91}$$

$$\lambda_{cc} = \sqrt{\frac{\hat{\sigma}_{\varepsilon}^2}{\hat{\sigma}_{cc}^2}} = \ \frac{125{,}932.45}{1{,}940.03} = \ 8.05 \tag{92}$$

$$\lambda_{ws} = \sqrt{\frac{\hat{\sigma}_{\varepsilon}^2}{\hat{\sigma}_{ws}^2}} = \frac{125{,}932.45}{0.00} = \ \text{Not Calculated} \tag{93}$$

Because the MIXED procedure yielded a variance estimate of zero associated with wind speed in hour 14, there is no subsequent smoothing applied to this term in the model. In other words, wind speed enters into the hour 14 model linearly and the resulting parameter estimates for its spline functions are nothing more than the standard result from OLS.

Across all hourly models, temperature and humidity have smoothing parameter estimates that are consistent in magnitude. The same estimates for cloud cover and wind speed are larger during late night and early morning models, while models for other times of the day estimate much smaller smoothing parameters for these terms. Notably, the model estimated for hour 15 has both $\boldsymbol{\hat{\sigma}_{cc}^2}$ and $\boldsymbol{\hat{\sigma}_{ws}^2}$ as zero. This result may be due in part to the nature of the observed data themselves. As mentioned in Section 4.4, both wind speed and cloud cover are essentially recorded as categorical variables and may mask relevant

variation between load and weather. As a consequence, the corresponding smoothing parameter estimates are zero and these terms enter into their respective models as fixed rather than random.

In order to test for and validate the use of the mixed model approach, a model specification test was conducted for each hourly model. Specifically, a likelihood ratio test is available using PROC MIXED. These tests specify a null model where all explanatory variables are fixed and the only variance component is $\sigma_\varepsilon^2 I$. The alternative model specification to test is that which is actually estimated with PROC MIXED; specifically, where each nonparametric component is estimated with a penalized spline. The test statistic is asymptotically distributed as a Chi-Square with degrees of freedom equal to *q-1*, where *q* is the number of covariance parameters. The test statistic is calculated as -2 multiplied by the log likelihood from the null model minus -2 multiplied by the log likelihood from the model that was fitted.[29] The results of these tests for the regional models are reported in Table 5-2, while those for each zonal model are contained in Appendix B.

[29] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect025.htm

Table 5-2: Null Model Likelihood Ratio Tests for Each Hourly Regional Model

Hour	*Degrees of Freedom*	*Chi-Square Statistic*	*p-value*
1	4	1,928.01	0.00
2	4	1,241.94	0.00
3	4	1,224.43	0.00
4	4	1,193.04	0.00
5	4	1,211.25	0.00
6	4	1,180.91	0.00
7	3	885.53	0.00
8	3	833.37	0.00
9	3	998.32	0.00
10	3	1,052.20	0.00
11	3	1,075.64	0.00
12	3	1,092.52	0.00
13	3	1,077.37	0.00
14	3	966.01	0.00
15	2	998.50	0.00
16	3	966.88	0.00
17	3	950.15	0.00
18	3	954.20	0.00
19	4	959.19	0.00
20	3	996.72	0.00
21	3	979.27	0.00
22	4	1,115.96	0.00
23	4	1,246.78	0.00
24	4	1,279.14	0.00

Again, we consider hour 14 as an example. The calculated test statistic was 966.01 with 3 degrees of freedom. At a 5% level of significance, the Chi-Square critical value associated with 3 degrees of freedom is 12.83. The test statistic is much larger than the critical value, the corresponding probability value (0.00) is less than the level of significance (0.05) and the null hypothesis specifying a fixed effects-only model is rejected.

The varying degrees of freedom between hourly models directly correspond to the estimates of zero variance for several hourly models in Table 5-1. For each hourly model, the null hypothesis (i.e., fixed effects model) is rejected at virtually any level of significance. This is strong statistical evidence that, for those models with nonzero variance estimates, the covariance structure corresponding to fitting a semiparametric model is necessary. Since each hourly model has at least two variance estimates, a semiparametric specification is employed for all models.

5.2 Generating Forecasts of Demand

Once a model is estimated, it can be used to produce forecasts. An example of how forecasts are generated for each hourly model is described in this section. This section begins by providing and discussing the parameter estimates associated with the fixed effects components of a regional model. Table 5-3 contains the fixed effect parameter estimates and the corresponding standard errors, test statistics and probability values obtained for hour 14. Again, the data used for estimating these models begins January 1, 2009 and ends December 31, 2010.

Table 5-3: Fixed Effects Parameter Estimates for Hour 14 Regional Model

Model Component	*Parameter*	*Estimate*	*Standard Error*	*t-statistic*	*p-value*
Intercept	δ_0	12,330.98	2,006.66	6.15	0.00
Tuesday	δ_1	943.04	61.78	15.26	0.00
Wednesday	δ_2	310.80	67.26	4.62	0.00
Thursday	δ_3	318.40	67.86	4.69	0.00
Friday	δ_4	123.94	62.12	2.00	0.05
Nonworking / Holiday	v	-1,357.45	54.22	-25.04	0.00
January	γ_1	-10.11	70.42	-0.14	0.89
February	γ_2	-89.12	67.02	-1.33	0.18
March	γ_3	-156.30	74.51	-2.10	0.04
April	γ_4	-340.25	88.28	-3.85	0.00
May	γ_5	-260.00	95.17	-2.73	0.01
June	γ_6	173.40	104.90	1.65	0.10
July	γ_7	124.37	112.09	1.11	0.27
August	γ_8	193.55	115.01	1.68	0.09
September	γ_9	-83.24	100.66	-0.83	0.41
October	γ_{10}	-103.85	87.41	-1.19	0.24
November	γ_{11}	-101.40	83.09	-1.22	0.22
24 Hour Lagged Demand	α_1	0.30	0.01	20.53	0.00
48 Hour Lagged Demand	α_2	-0.01	0.02	-0.82	0.41
72 Hour Lagged Demand	α_3	0.03	0.02	1.93	0.05
96 Hour Lagged Demand	α_4	0.02	0.02	1.28	0.20
120 Hour Lagged Demand	α_5	-0.04	0.02	-2.67	0.01
144 Hour Lagged Demand	α_6	0.04	0.01	2.87	0.00
168 Hour Lagged Demand	α_7	0.04	0.01	2.88	0.00
Temperature	θ_1	-51.47	212.08	-0.24	0.81
Humidity	θ_2	-15.73	60.04	-0.26	0.79
Cloud Cover	θ_3	33.55	11.33	2.96	0.00
Wind Speed	θ_4	-0.71	3.31	-0.21	0.83

The parameter estimates reported result in the construction of a fitted model that is additive in nature. This characteristic allows us to generate separate forecasts for each of the components of the model in Equation (94). For exposition, we walk through the process of producing a forecast for 2 pm on Friday, July 20th, 2011.

Using the results in Table 5-3, the fitted equation to forecast calendar effects is given by

$$\begin{aligned}\widehat{h(\text{time})} \;=\; & 12{,}330.98 + 943.04\,D_{1t} + 310.80D_{2t} + 318.40\,D_{3t} + 123.94\,D_{4t} \\ & - 1{,}357.45\,N_t - 10.11\,M_{1t} - \;89.12\,M_{2t} - 156.30\,M_{3t} \\ & - 340.25\,M_{4t} - \;260.00\,M_{5t} + 173.40\,M_{6t} + 124.37\,M_{7t} \\ & + \;193.55\,M_{8t} - 83.24\,M_{9t} - 103.85\,M_{10t} - \;101.40\,M_{11t}\end{aligned} \tag{94}$$

Hour 14 (i.e., 2 pm) on July 20th, 2011 represents a single observation of data from the available data set. Notice that this time and date coincide with a post-sample (out-of-sample) time period. There is an estimated intercept of 12,330 MW for this hour and since the day of the week for this observation was Friday, all day-of-week binary variables become zero with the exception of D_{4t}. This dummy variable becomes one and its corresponding marginal effect is 123.94 MW. It is a working day and N_t becomes zero. Lastly, the month is July, the seventh month of the year. As such, monthly binary variables become zero with the exception of M_{7t}. This binary variable becomes one and its corresponding marginal effect is 124.37 MW. Using these estimates, the aggregate calendar effect at this time can be forecasted.

$$\begin{aligned}\widehat{h(\text{time})} \;&=\; 12{,}330.98 + 123.94(1) + 124.37(1) \\ &=\; \mathbf{12{,}579.29\ MW}\end{aligned} \tag{95}$$

The second component found in Equation (94) is that of $\alpha(recent\ demand)$. Substituting the coefficients in Table 5-3 into Equation (96) yields the following fitted equation:

$$\widehat{\alpha(.)} = 0.30\, y_{t-24} - 0.01\, y_{t-48} + 0.03\, y_{t-72} + 0.02\, y_{t-96} - 0.04\, y_{t-120} + 0.04\, y_{t-144} + 0.04\, y_{t-168} \quad (96)$$

Using the above fitted equation, observed values of lag demand can be used to forecast the effect recent demand has on load in at hour *t*. For the observation chosen, the lagged values of demand required to calculate $\widehat{\alpha(.)}$ are contained in Table 5-4 below.

Table 5-4: Lagged Demand Observations at 2 pm on July 20th, 2011.

Variable	*Load (MW)*
24 Hour Lagged Demand	23,531
48 Hour Lagged Demand	22,433
72 Hour Lagged Demand	20,101
96 Hour Lagged Demand	18,686
120 Hour Lagged Demand	19,219
144 Hour Lagged Demand	18,419
168 Hour Lagged Demand	22,591

It is straightforward to plug the observed demand values from Table 5-4 into fitted Equation (96) to forecast the effect recent demand has on load for the observation chosen.

$$\widehat{\alpha(.)} = 7{,}059.30 - 224.33 + 603.03 + 373.72 - 768.76 + 736.76 + 903.64 = \mathbf{8{,}683.36\ MW} \quad (97)$$

Note that the effect of the most recent lagged demand value is by far the greatest of any lagged demand value. This result supports the analysis done using the autocorrelation function in section 4.3.

Both $\widehat{h(time)}$ and $\widehat{\alpha(.)}$ require only those parameter estimates that pertain to the fixed effect model components in order to forecast their effect on load. However, the

calculation of $f(\widehat{weather})$ requires the inclusion of random effect predictions as each weather variable was treated as a random effect during the model fitting process.

Table 5-5 displays the corresponding prediction results.

Table 5-5: Random Effects Parameter Predictions for Hour 14 Regional Model

Model Component	*Parameter*	*Prediction*	*Standard Error*	*t-statistic*	*p-value*
$(\boldsymbol{Temp_t - (-2°)})_+$	u_1^T	0.00[30]	102.02	0.00	1.00
$(\boldsymbol{Temp_t - 3°})_+$	$\boldsymbol{u_2^T}$	0.00	102.02	0.00	1.00
$(\boldsymbol{Temp_t - 8°})_+$	$\boldsymbol{u_3^T}$	0.00	102.02	0.00	1.00
$(\boldsymbol{Temp_t - 13°})_+$	$\boldsymbol{u_4^T}$	0.00	102.02	0.00	1.00
$(\boldsymbol{Temp_t - 18°})_+$	$\boldsymbol{u_5^T}$	17.36	74.93	0.23	0.82
$(\boldsymbol{Temp_t - 23°})_+$	$\boldsymbol{u_6^T}$	-101.65	46.32	-2.19	0.03
$(\boldsymbol{Temp_t - 28°})_+$	$\boldsymbol{u_7^T}$	45.07	39.82	1.13	0.26
$(\boldsymbol{Temp_t - 33°})_+$	$\boldsymbol{u_8^T}$	50.01	35.53	1.41	0.16
$(\boldsymbol{Temp_t - 38°})_+$	$\boldsymbol{u_9^T}$	-54.79	34.37	-1.59	0.11
$(\boldsymbol{Temp_t - 43°})_+$	$\boldsymbol{u_{10}^T}$	43.54	33.23	1.31	0.19
$(\boldsymbol{Temp_t - 48°})_+$	$\boldsymbol{u_{11}^T}$	14.52	33.62	0.43	0.67
$(\boldsymbol{Temp_t - 53°})_+$	$\boldsymbol{u_{12}^T}$	24.44	31.55	0.77	0.44
$(\boldsymbol{Temp_t - 58°})_+$	$\boldsymbol{u_{13}^T}$	0.13	30.95	0.00	1.00
$(\boldsymbol{Temp_t - 63°})_+$	$\boldsymbol{u_{14}^T}$	44.63	30.59	1.46	0.15
$(\boldsymbol{Temp_t - 68°})_+$	$\boldsymbol{u_{15}^T}$	47.91	29.79	1.61	0.11
$(\boldsymbol{Temp_t - 73°})_+$	$\boldsymbol{u_{16}^T}$	125.72	30.11	4.18	0.00
$(\boldsymbol{Temp_t - 78°})_+$	$\boldsymbol{u_{17}^T}$	118.99	32.64	3.65	0.00
$(\boldsymbol{Temp_t - 83°})_+$	$\boldsymbol{u_{18}^T}$	-75.99	36.77	-2.07	0.04
$(\boldsymbol{Temp_t - 88°})_+$	$\boldsymbol{u_{19}^T}$	89.30	55.41	1.61	0.11
$(\boldsymbol{Temp_t - 93°})_+$	$\boldsymbol{u_{20}^T}$	73.37	88.90	0.83	0.41
$(\boldsymbol{Temp_t - 99°})_+$	$\boldsymbol{u_{21}^T}$	0.00	102.02	0.00	1.00
$(\boldsymbol{Hum_t - (-16)})_+$	$\boldsymbol{u_1^H}$	0.00	31.70	0.00	1.00
$(\boldsymbol{Hum_t - (-11)})_+$	$\boldsymbol{u_2^H}$	0.00	31.70	0.00	1.00
$(\boldsymbol{Hum_t - (-6)})_+$	$\boldsymbol{u_3^H}$	-4.63	31.50	-0.15	0.88
$(\boldsymbol{Hum_t - (-1)})_+$	$\boldsymbol{u_4^H}$	-1.08	27.98	-0.04	0.97
$(\boldsymbol{Hum_t - 4})_+$	$\boldsymbol{u_5^H}$	34.01	23.81	1.43	0.15
$(\boldsymbol{Hum_t - 9})_+$	$\boldsymbol{u_6^H}$	7.77	22.40	0.35	0.73
$(\boldsymbol{Hum_t - 14})_+$	$\boldsymbol{u_7^H}$	9.39	20.87	0.45	0.65
$(\boldsymbol{Hum_t - 19})_+$	$\boldsymbol{u_8^H}$	-30.89	20.40	-1.51	0.13
$(\boldsymbol{Hum_t - 24})_+$	$\boldsymbol{u_9^H}$	16.74	19.50	0.86	0.39
$(\boldsymbol{Hum_t - 29})_+$	$\boldsymbol{u_{10}^H}$	5.99	19.73	0.30	0.76
$(\boldsymbol{Hum_t - 34})_+$	$\boldsymbol{u_{11}^H}$	-26.92	19.75	-1.36	0.17
$(\boldsymbol{Hum_t - 39})_+$	$\boldsymbol{u_{12}^H}$	16.95	19.17	0.88	0.38
$(\boldsymbol{Hum_t - 44})_+$	$\boldsymbol{u_{13}^H}$	16.22	19.15	0.85	0.40
$(\boldsymbol{Hum_t - 49})_+$	$\boldsymbol{u_{14}^H}$	-9.62	19.10	-0.50	0.61
$(\boldsymbol{Hum_t - 54})_+$	$\boldsymbol{u_{15}^H}$	50.34	19.65	2.56	0.01
$(\boldsymbol{Hum_t - 59})_+$	$\boldsymbol{u_{16}^H}$	71.26	19.57	3.64	0.00
$(\boldsymbol{Hum_t - 64})_+$	$\boldsymbol{u_{17}^H}$	-1.55	20.80	-0.07	0.94
$(\boldsymbol{Hum_t - 69})_+$	$\boldsymbol{u_{18}^H}$	-18.14	27.72	-0.65	0.51
$(\boldsymbol{Hum_t - 73})_+$	$\boldsymbol{u_{19}^H}$	0.70	31.60	0.02	0.98
$(\boldsymbol{Hum_t - 74})_+$	$\boldsymbol{u_{20}^H}$	0.00	31.70	0.00	1.00
$(\boldsymbol{CC_t - 5})_+$	u_1^{CC}	-4.16	20.23	-0.21	0.84
$(\boldsymbol{CC_t - 9})_+$	u_2^{CC}	-26.63	41.81	-0.64	0.52
$(\boldsymbol{WS_t - 5})_+$	u_1^{WS}	0.00	0.00	0.00	0.00
$(\boldsymbol{WS_t - 11})_+$	u_2^{WS}	0.00	0.00	0.00	0.00
$(\boldsymbol{WS_t - 17})_+$	u_3^{WS}	0.00	0.00	0.00	0.00
$(\boldsymbol{WS_t - 23})_+$	u_4^{WS}	0.00	0.00	0.00	0.00
$(\boldsymbol{WS_t - 29})_+$	u_5^{WS}	0.00	0.00	0.00	0.00
$(\boldsymbol{WS_t - 31})_+$	u_6^{WS}	0.00	0.00	0.00	0.00

[30] Predicted parameter estimates extremely small but yet still nonzero.

Note that the predictions associated with extreme values of temperature and humidity drop out of the model. Given that the focus is currently limited to a forecasting model that generates mid-afternoon predictions, it is not an unjustifiable result that extreme values of temperature and humidity were not significant predictors of load at this time. Also, the random effects component pertaining to wind speed is dropped from the model entirely. This result is validated by the three degrees of freedom that were used for the hour 14 null model likelihood ratio test found in Table 5-2. In other words, the variance component related to wind speed was omitted during the model fitting process.

For context, our 2 p.m. sample observation was an hour where the temperature was 87 degrees Fahrenheit, the humidity was 64 degrees, the cloud cover was at about 30% (i.e., 3/10 or just 3), and the wind speed was measured as 13 miles per hour. Given these prevailing weather conditions, the basis functions used in the penalized spline for each weather predictor can be calculated.

Table 5-6 reports the enumerated value of each basis function corresponding to our sample observation.

Table 5-6: Spline Basis Functions Evaluated using Prevailing Weather Conditions at 2 pm, July 20th, 2011.

Temperature Basis Function $(Temp_t = 87)$	*Value*	*Humidity Basis Function* $(Hum_t = 64)$	*Value*	*Cloud Cover Basis Function* $(CC_t = 3)$	*Value*	*Wind Speed Basis Function* $WS_t = 13)$	*Value*
$(Temp_t - (-2°))_+$	89	$(Hum_t - (-16))_+$	80	$(CC_t - 5)_+$	0	$(WS_t - 5)_+$	8
$(Temp_t - 3°)_+$	84	$(Hum_t - (-11))_+$	75	$(CC_t - 9)_+$	0	$(WS_t - 11)_+$	2
$(Temp_t - 8°)_+$	79	$(Hum_t - (-6))_+$	70			$(WS_t - 17)_+$	0
$(Temp_t - 13°)_+$	74	$(Hum_t - (-1))_+$	65			$(WS_t - 23)_+$	0
$(Temp_t - 18°)_+$	69	$(Hum_t - 4)_+$	60			$(WS_t - 29)_+$	0
$(Temp_t - 23°)_+$	64	$(Hum_t - 9)_+$	55			$(WS_t - 31)_+$	0
$(Temp_t - 28°)_+$	59	$(Hum_t - 14)_+$	50				
$(Temp_t - 33°)_+$	54	$(Hum_t - 19)_+$	45				
$(Temp_t - 38°)_+$	49	$(Hum_t - 24)_+$	40				
$(Temp_t - 43°)_+$	44	$(Hum_t - 29)_+$	35				
$(Temp_t - 48°)_+$	39	$(Hum_t - 34)_+$	30				
$(Temp_t - 53°)_+$	34	$(Hum_t - 39)_+$	25				
$(Temp_t - 58°)_+$	29	$(Hum_t - 44)_+$	20				
$(Temp_t - 63°)_+$	24	$(Hum_t - 49)_+$	15				
$(Temp_t - 68°)_+$	19	$(Hum_t - 54)_+$	10				
$(Temp_t - 73°)_+$	14	$(Hum_t - 59)_+$	5				
$(Temp_t - 78°)_+$	9	$(Hum_t - 64)_+$	0				
$(Temp_t - 83°)_+$	4	$(Hum_t - 69)_+$	0				
$(Temp_t - 88°)_+$	0	$(Hum_t - 73)_+$	0				
$(Temp_t - 93°)_+$	0	$(Hum_t - 74)_+$	0				
$(Temp_t - 99°)_+$	0						

Using these evaluated functions, the load-weather relationship can be estimated and an aggregate demand forecast generated. Similar to the broader decomposition of load into its three main drivers (i.e., $h(time), \alpha(recent\ demand),\ f(weather)$), each weather variable is isolated and individually evaluated. Again, because our model is additive, the effect of temperature, humidity, cloud cover and wind speed is each calculated in the following equations.

$$
\begin{aligned}
f_1(\widehat{Temp}_t) \;=\; & -51.47\,(87) + 0\,(89) + 0\,(84) + 0\,(79) + 0\,(74) \\
& + 17.36\,(69) - 101.65\,(64) + 45.07\,(59) + 50.01\,(54) \\
& - 54.79\,(49) + 43.54\,(44) + 14.52\,(39) + 24.44\,(34) \\
& + 0.13\,(29) + 44.63\,(24) + 47.91\,(19) + 125.72\,(14) \\
& + 118.99\,(9) - 75.99\,(4) + 89.3\,(0) + 73.37\,(0) + 0\,(0) \\
= \; & \mathbf{714.52\ MW}
\end{aligned}
\tag{98}
$$

$$
\begin{aligned}
f_2(\widehat{Hum}_t) \;=\; & -15.73\,(64) + 0\,(80) + 0\,(75) - 4.63\,(70) - 1.08\,(65) \\
& + 7.77\,(55) + 9.39\,(50) + 30.89\,(45) + 16.74\,(40) \\
& + 5.99\,(35) + 26.92\,(30) + 16.95(25) + 16.22(20) \\
& - 9.62\,(15) + 50.34\,(10) + 71.26\,(5) - 1.55\,(0) - 18.14\,(0) \\
& + 0.7\,(0) + 0\,(0) \\
= \; & \mathbf{1,681.58\ MW}
\end{aligned}
\tag{99}
$$

$$
\begin{aligned}
\widehat{f_3(CC_t)} \;=\; & 33.55\,(3) - 4.16\,(0) - 26.63\,(0) \\
= \; & \mathbf{100.65\ MW}
\end{aligned}
\tag{100}
$$

$$
\begin{aligned}
\widehat{f_4(WS_t)} \;=\; & -0.71\,(13) + 0\,(8) + 0\,(2) + 0\,(0) + 0\,(0) + 0\,(0) + 0\,(0) \\
= \; & \mathbf{-9.23\ MW}
\end{aligned}
\tag{101}
$$

From Equations (98) through (101), it appears that mid-afternoon July humidity is the strongest contributor to regional load, followed by temperature. The predictive impact

of 30% cloud cover is an additional 100 MW of load on the system. This is the equivalent of an entire power plant. Finally, as the random effects associated with wind speed are omitted, the contribution of wind speed on the grid is limited to the estimated fixed effect parameter $\widehat{\theta_4}$ and corresponds to a load reduction of about 9 MW.

Aggregating these piecewise forecasts allows for calculation of the total predicted effect of weather upon regional load.

$$\begin{aligned} f(\widehat{weather}) &= f_1(\widehat{Temp}_t) + f_2(\widehat{Hum}_t) + f_3\widehat{(CC}_t) + f_4(\widehat{WS}_t) \\ &= 714.52 + 1{,}681.58 + 100.65 - 9.23 \\ &= \mathbf{2{,}487.52\ MW} \end{aligned} \tag{102}$$

Again, aggregating all piecewise forecasts allows a final prediction of New England regional load at 2 pm on July 20th, 2011.

$$\begin{aligned} \widehat{y_t} &= \mathbf{12{,}579.29\ MW + 8{,}683.36\ MW + 2{,}487.52} \\ &= \mathbf{23{,}750.17\ MW} \end{aligned} \tag{103}$$

The actual load, y_t, on the regional grid at this hour was 23,858 MW. The forecasted load, $\widehat{y_t}$, produced using actual observed weather and previous load under-forecasted the actual load by 107.83 MW or about 0.4% error. The example provided in this section fully describes the process that generates all other forecasts for all regional and zonal models. Their comparative performance is discussed in the next section.

5.3 Comparison of Forecast Performance

Following Fan and Hyndman (2010) as well as forecasting protocol at ISONE, forecasting performance is measured using Mean Absolute Error (MAE) and Mean Absolute Percent Error (MAPE). The MAE metric remains in terms of Megawatts and is advantageous for control room operators who may need to rapidly change resource commitment, while MAPE measures the average absolute error as a percent. The expressions that define each are given below.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{104}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t} \tag{105}$$

where again y_t is an actual observation of load at time t and $\hat{y}_t$ is its predicted value.

These metrics were calculated for both within-sample and out-of-sample periods defined in Section 4.5, and the overall forecasting performance of both models is provided in Table 5-7.

Table 5-7: Overall Forecasting Performance for Within-Sample and Out-of-Sample Periods.

Metric	***Within Sample***			***Out of Sample***		
	Regional Model	*Zonal Aggregation*	*Difference*	*Regional Model*	*Zonal Aggregation*	*Difference*
MAPE (%)	1.62%	1.66%	-0.04%	1.93%	1.97%	-0.03%
MAE (MW)	243.47	248.63	-5.16	283.04	287.36	-4.32

Both the regional model and zonal aggregation performed better during the within-sample time frame. However, the regional model performs better than the zonal aggregation during both within-sample and out-of-sample. In terms of both MAE and MAPE, the difference between each model's performance was slightly smaller during the out-of-sample period (e.g., -0.03% compared to -0.04%). In addition, the 1.93% and 1.97% overall out-of-sample MAPEs for both models are highly competitive with MAPE metrics reported by ISONE to FERC for approximately the same time frame[31].

It is important to note that the out-of-sample forecasts generated by each estimated model are referred to as *ex post* in that they are 'after the fact'. In other words, ex post forecasts are made using actual observations of predictor variables during this post-sample period. In contrast, *ex ante* forecasts are made using the available information of predictor variables at the time of forecasting. This may be actual values or predicted values. For instance, an ex ante forecast of load depends on weather forecasts as inputs to the model. The day-ahead forecast issued by ISONE at 10 a.m. every day relies on weather forecasts and the most recent observations of demand. As such, the ISONE day-ahead forecast is a 36 hour ex ante forecast.

The available data sets contain forecasted weather variables provided by each weather vendor. It is not clear, however, at which hour *t* these weather forecasts were provided to ISONE. In addition, the first 24-hour displacement of recent demand would not be available at the time of forecasting. Since there is a 36 hour forecasting horizon, forecast

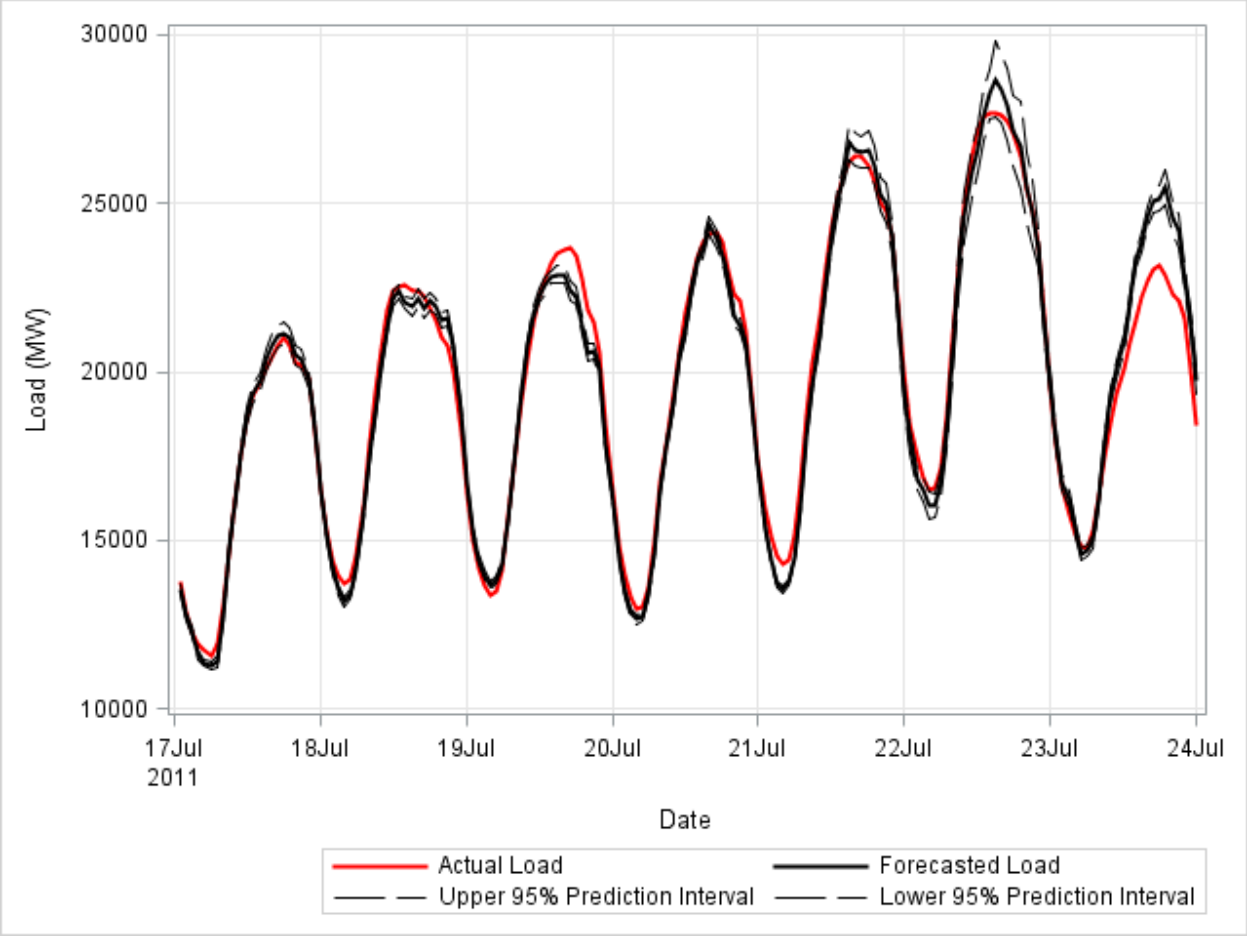
[31] http://www.iso-ne.com/regulatory/ferc/filings/2011/aug/ad10-5-00_8-31-11_joint_iso-rto_metrics_report.pdf

of regional demand at the appropriate hours (*t-24*) would have to be provided as model inputs in order to make a true ex ante forecast. As such, a true ex ante 36-hour forecast using these data is not currently possible.

Given that New England experiences the highest demand for electricity during the summer months, particular attention is paid to forecasting performance at this time. Out of the ten highest observations of daily peak demand that occurred during 2011, eight of them took place in the month of July and five of those eight were during the week spanning July 17th to July 24th. Figure 30 depicts the actual hourly demand for electricity during this week in New England. This figure also contains the forecasted values as well as 95% prediction intervals[32] produced by the regional model.

[32] Prediction intervals were calculated using the 'OUTPRED' option in PROC MIXED. Specifically, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect015.htm

Figure 30: Forecasted Load vs. Actual Load for the New England Region during Peak Load, 2011.



During 2011, the highest load on the New England electricity grid occurred on July 22nd and was metered at approximately 27,707 MW. The regional model generated a forecast for this hour that was slightly higher at 28,681 MW with an absolute percent error of 3.5%. The actual value of load at this hour as well as almost every hourly value of load during this week was well within the limits of the 95% prediction interval.

While forecasting performance at the highest values of summer demand was good, the large variation of load during these months resulted in poor forecasting performance compared to other months. For both periods and for both models, the largest MAPE occurred during July and August. These months also contain the largest percent differences between models for both periods. Specifically, there is a difference of -0.17

between model MAPE during August of the within-sample period and a difference of -0.23 during July of the out-of-sample period. Table 5-8 reports monthly forecasting performance for all periods and all models as measured by MAPE.

Table 5-8: MAPE (%) by Month

Month	*Within Sample*			*Out of Sample*		
	Regional Model	*Zonal Aggregation*	*Difference*	*Regional Model*	*Zonal Aggregation*	*Difference*
January	1.31%	1.36%	*-0.05*	1.52%	1.54%	*-0.03*
February	1.26%	1.31%	*-0.05*	1.45%	1.48%	*-0.03*
March	1.47%	1.53%	*-0.07*	1.46%	1.53%	*-0.07*
April	1.56%	1.54%	*0.02*	1.67%	1.81%	*-0.14*
May	1.62%	1.52%	*0.10*	1.40%	1.29%	*0.11*
June	1.66%	1.57%	*0.10*	2.07%	2.15%	*-0.08*
July	2.10%	2.21%	*-0.11*	2.32%	2.55%	*-0.23*
August	1.84%	2.01%	*-0.17*	3.10%	3.07%	*0.03*
September	1.84%	1.83%	*0.02*	1.61%	1.55%	*0.06*
October	1.37%	1.47%	*-0.09*	2.50%	2.46%	*0.04*
November	1.52%	1.57%	*-0.05*	2.08%	2.12%	*-0.04*
December	1.87%	1.94%	*-0.07*	1.97%	2.00%	*-0.03*

The largest MAPE for any month during the out-of-sample period was 3.10% for the regional model and 3.07% for the zonal aggregation. Both of these were in the month of August. Conversely, the best forecasting performance for any month in the same time frame occurred in May with MAPEs of 1.40% and 1.29% for the regional and zonal models, respectively. In general, the forecasting performance as measured by MAPE was consistent across months for both models. Figure 31 and Figure 32 depict these results graphically.

Figure 31: Within Sample MAPE(%) by Month, 2009 & 2010

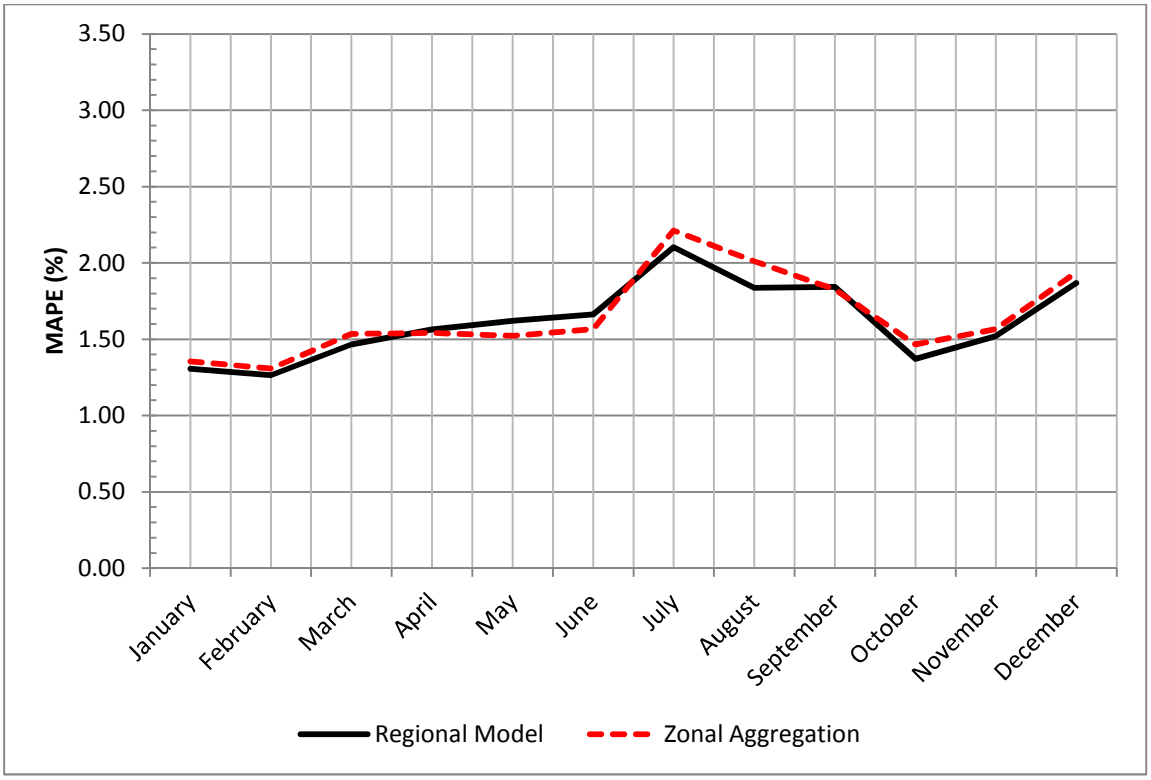
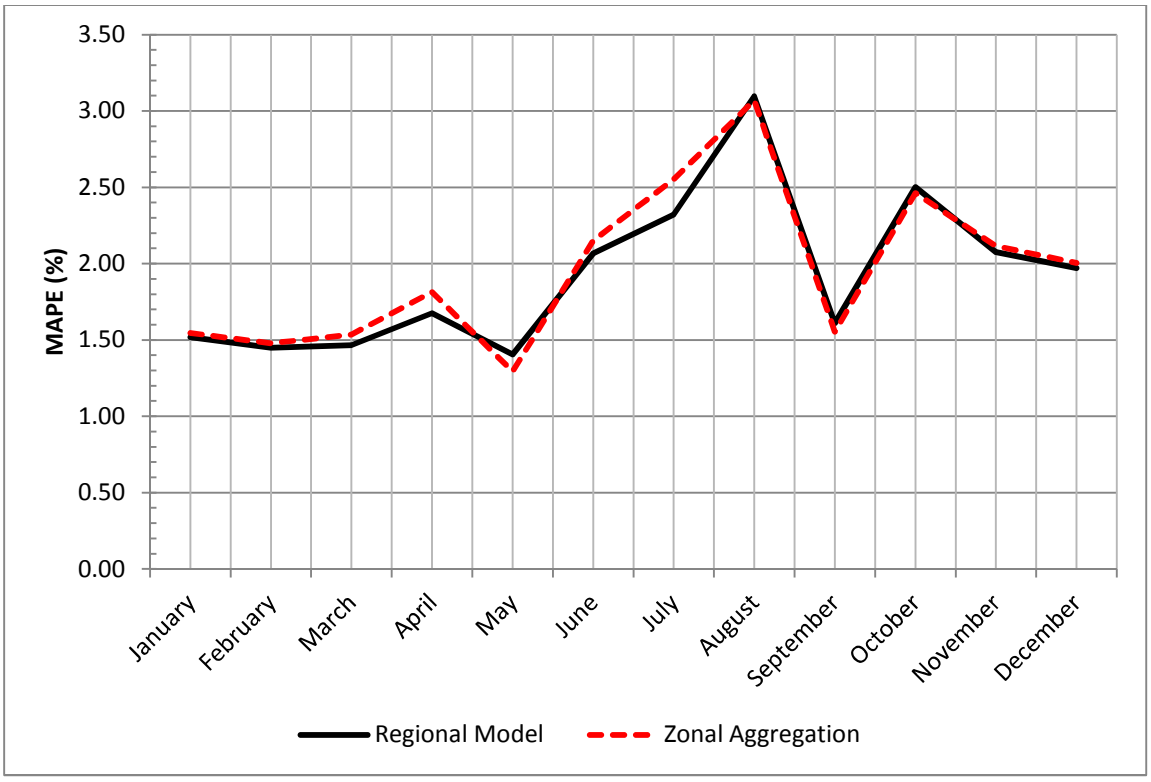


Figure 32: Out-of-Sample MAPE (%) by Month, 2011



An identical analysis was applied using the MAE metric. Not surprisingly, the monthly forecasting performances as told by MAE are comparable to those measured by MAPE. Table 5-9 reports the results.

Table 5-9: MAE (MW) by Month

Month	*Within Sample*			*Out of Sample*		
	Regional Model	*Zonal Aggregation*	*Difference*	*Regional Model*	*Zonal Aggregation*	*Difference*
January	211.04	217.83	*-6.78*	244.49	246.44	*-1.95*
February	190.96	197.13	*-6.18*	226.79	229.82	*-3.03*
March	207.27	217.13	*-9.86*	212.80	222.41	*-9.60*
April	210.75	207.23	*3.51*	221.44	240.97	*-19.53*
May	220.94	205.76	*15.18*	201.83	186.14	*15.69*
June	252.18	236.03	*16.15*	319.25	328.17	*-8.92*
July	353.13	372.01	*-18.88*	403.51	443.31	*-39.80*
August	307.96	334.47	*-26.51*	437.20	429.24	*7.96*
September	270.18	269.69	*0.49*	239.60	230.06	*9.54*
October	184.51	196.10	*-11.59*	311.47	307.34	*4.14*
November	208.77	214.97	*-6.21*	280.78	285.22	*-4.45*
December	294.14	304.55	*-10.40*	289.58	291.51	*-1.93*

An interesting result of this analysis indicates that forecasting performance is worst earlier in the year during 2009 and 2010 than it is in 2011. This conclusion is also supported by Figure 13 which indicates a trend of average monthly demand peaking more towards the end of the year. In addition, 2011 was known to be an unusually warm summer. Given the high correlation with temperature and humidity, forecasts made during an unusually warm season or year may be prone to error if the underlying model was trained using unsuitable past observations of weather. Figure 33 and Figure 34 show monthly MAE for both periods.

Figure 33: Within-Sample MAE (MW) by Month, 2009 & 2010

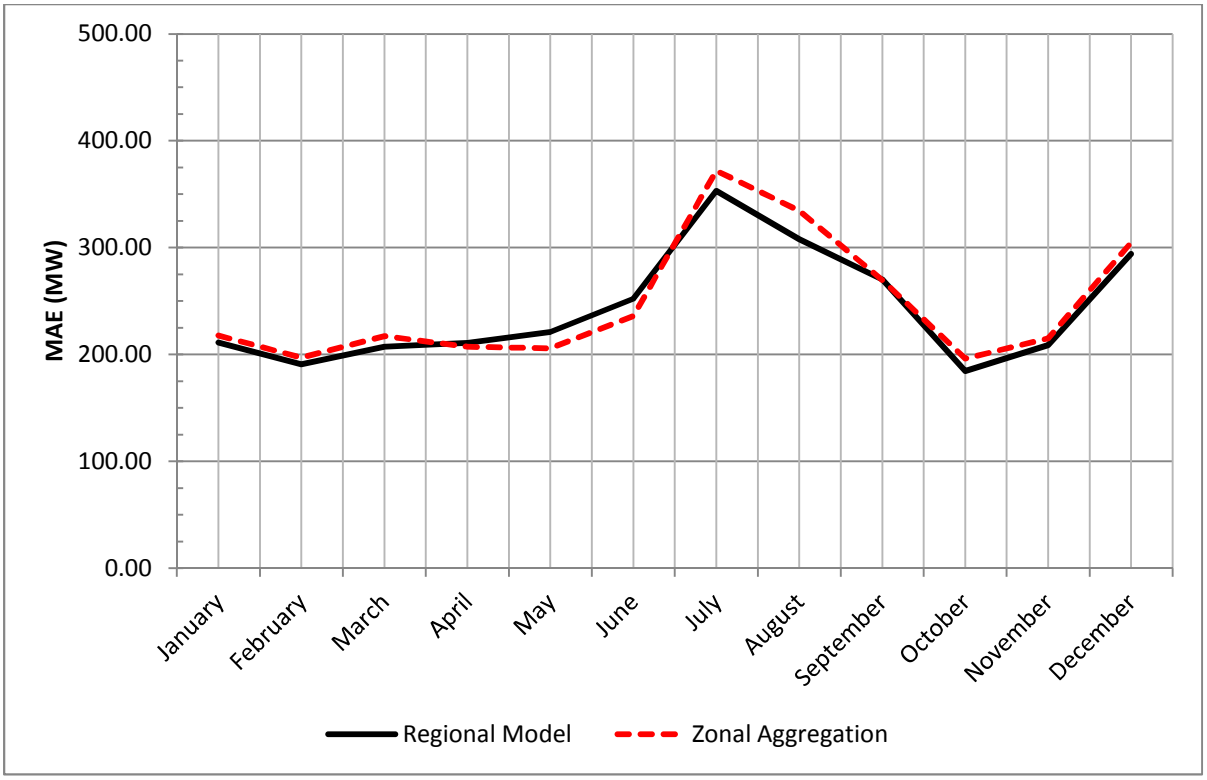
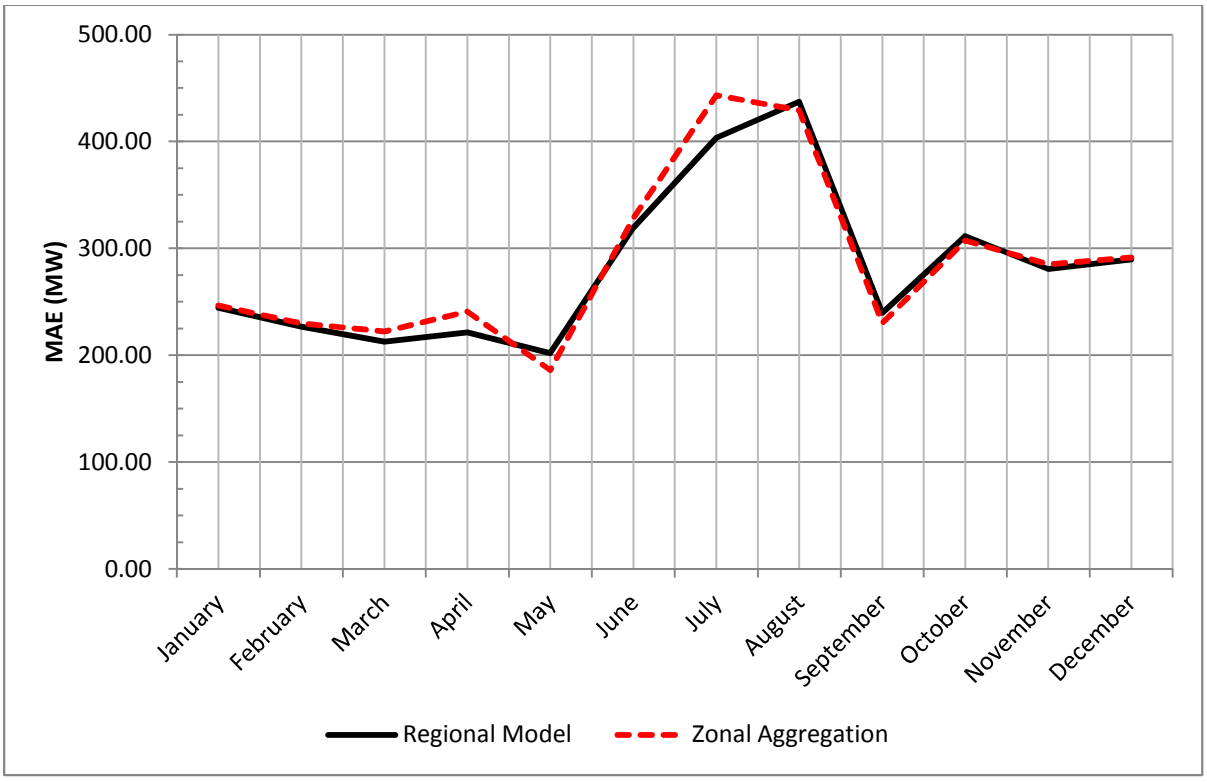


Figure 34: Out-of-Sample MAE (MW) by Month, 2011



Finally, an analysis was done on the forecasting performance of each hourly model. Table 5-10 reports MAPE performance for each hour. Comparable to the performance during the summer months, the afternoon hours exhibit the worst performance. This is expected and is a standard result with other load forecasting applications. However, the zonal aggregation consistently performed slightly better than the regional model at this time of day. During the out-of-sample period, the zonal aggregation produced forecasts with smaller MAPEs than those produced by the regional model contiguously from hour 12 to hour 19. Furthermore, the largest MAPE for any hour in the out-of-sample zonal aggregation does not exceed 2.5%.

Table 5-10: MAPE (%) by Hour.

Hour	*Within Sample*			*Out of Sample*		
	Regional Model	*Zonal Aggregation*	*Difference*	*Regional Model*	*Zonal Aggregation*	*Difference*
1	1.34%	1.42%	-0.09%	1.60%	1.75%	-0.15%
2	1.32%	1.42%	-0.10%	1.64%	1.76%	-0.12%
3	1.35%	1.42%	-0.07%	1.64%	1.76%	-0.12%
4	1.37%	1.41%	-0.03%	1.59%	1.70%	-0.11%
5	1.34%	1.39%	-0.05%	1.61%	1.68%	-0.07%
6	1.37%	1.44%	-0.07%	1.74%	1.78%	-0.04%
7	1.69%	1.73%	-0.04%	2.06%	2.10%	-0.04%
8	1.62%	1.67%	-0.04%	1.96%	2.00%	-0.03%
9	1.39%	1.44%	-0.05%	1.79%	1.86%	-0.06%
10	1.42%	1.43%	-0.01%	1.90%	1.91%	0.00%
11	1.50%	1.49%	0.01%	1.93%	1.94%	-0.01%
12	1.62%	1.61%	0.01%	2.00%	1.98%	0.03%
13	1.78%	1.78%	0.00%	2.09%	2.02%	0.07%
14	1.96%	1.98%	-0.02%	2.25%	2.19%	0.06%
15	2.12%	2.15%	-0.02%	2.37%	2.35%	0.01%
16	2.22%	2.24%	-0.02%	2.43%	2.39%	0.04%
17	2.25%	2.24%	0.01%	2.55%	2.50%	0.05%
18	2.08%	2.10%	-0.01%	2.48%	2.47%	0.01%
19	1.96%	1.95%	0.01%	2.29%	2.25%	0.04%
20	1.73%	1.72%	0.01%	1.96%	2.01%	-0.05%
21	1.60%	1.63%	-0.04%	1.77%	1.85%	-0.07%
22	1.40%	1.49%	-0.09%	1.63%	1.74%	-0.12%
23	1.28%	1.36%	-0.08%	1.61%	1.64%	-0.04%
24	1.27%	1.36%	-0.09%	1.53%	1.63%	-0.10%

While the average hourly forecasting errors may indicate some differences in central tendency between models, looking at the distribution of forecasting errors by hour reveals extremely similar results for both approaches. A convenient way of investigating the distribution of forecast errors is with a box-and-whisker diagram (or just boxplot). The following diagram explains the notation and symbols used in a boxplot produced by the VBOX statement within the SAS procedure PROC SGPLOT[33].

[33] http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#vbox-stmt.htm

Figure 35: SAS boxplot legend.

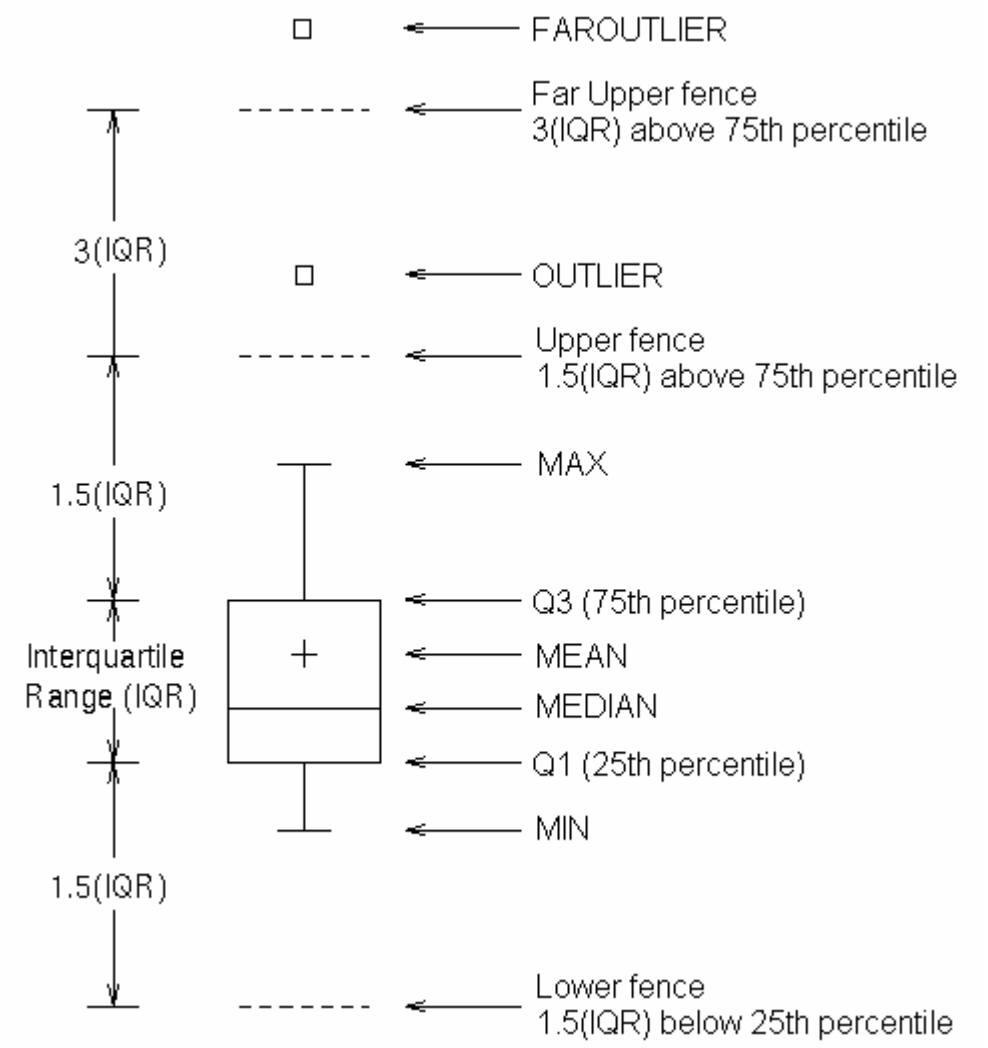


Figure 35 provides a legend for the meaning of the typical components in a boxplot: mean, quantiles, interquantile ranges, and outlier notation.

Boxplots of absolute percent error by hour illustrate that even the pattern of outliers for each hour is consistent across models for both within-sample and out-of-sample time periods. Figure 36 - 39 display these patterns for within-sample, zonal and regional and for out-of-sample, zonal and regional. Each black dot is an outlier and denotes an observed error which is 1.5 IQR above the 75th percentile. The center of each hourly boxplot provides a perspective for the placement of the percentiles (quantiles), mean,

minimum and interquartile range. Note: a 'diamond' symbol found in the following figures replaces the 'cross' indicator for mean value found in Figure 35: **SAS boxplot legend..**

Figure 36: Within-Sample Boxplots of Hourly Error Percent for the Regional Model.

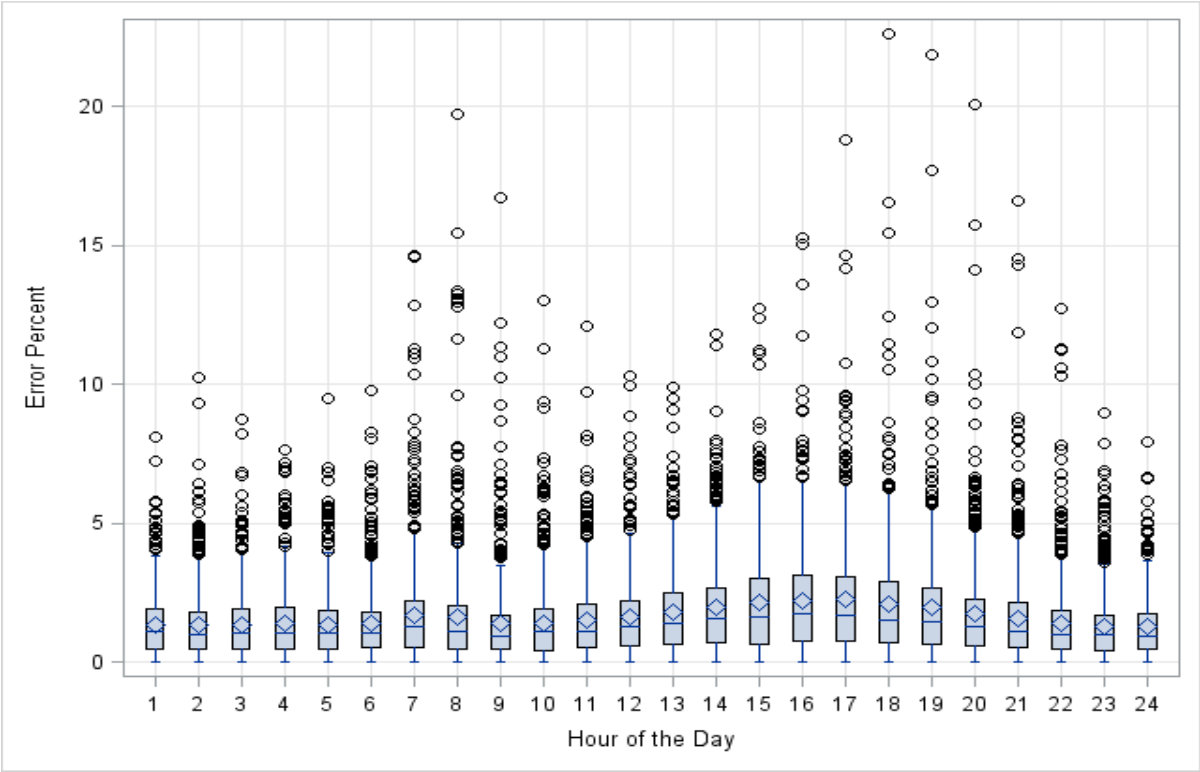


Figure 37: Within-Sample Boxplots of Hourly Error Percent for the Zonal Aggregation.

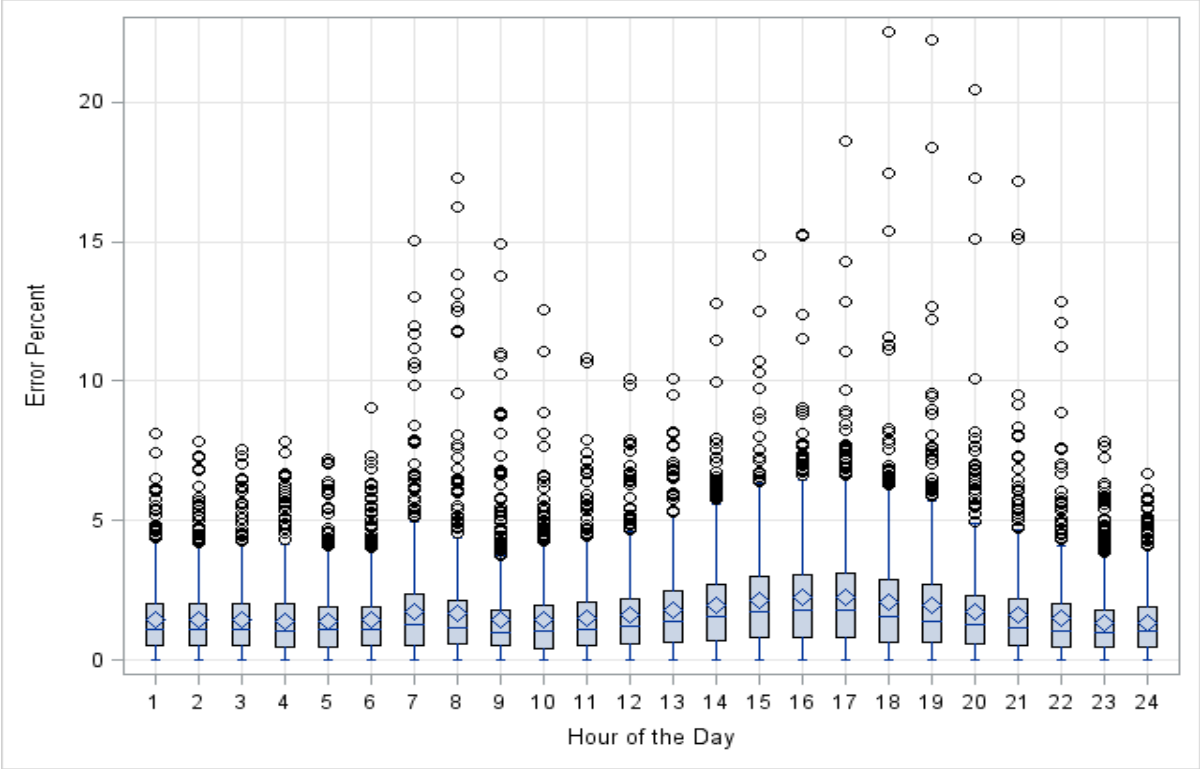


Figure 38: Out-of-Sample Boxplots of Hourly Error Percent for the Regional Model.

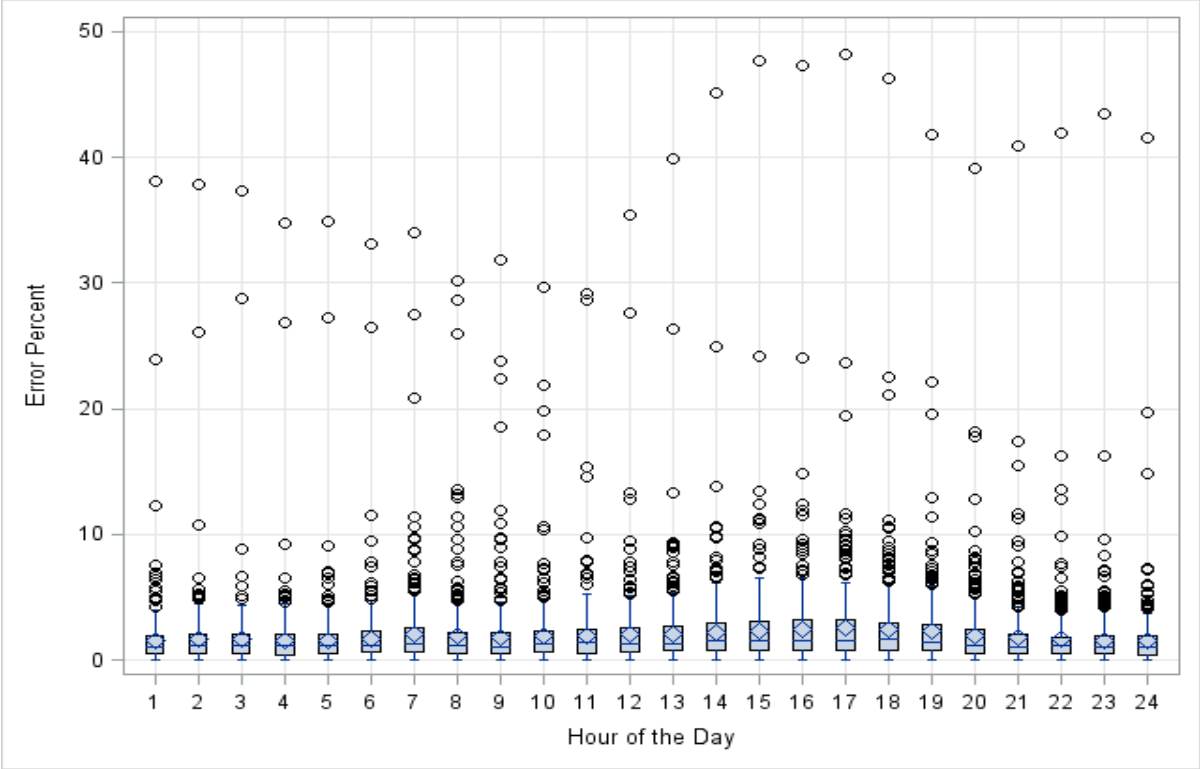
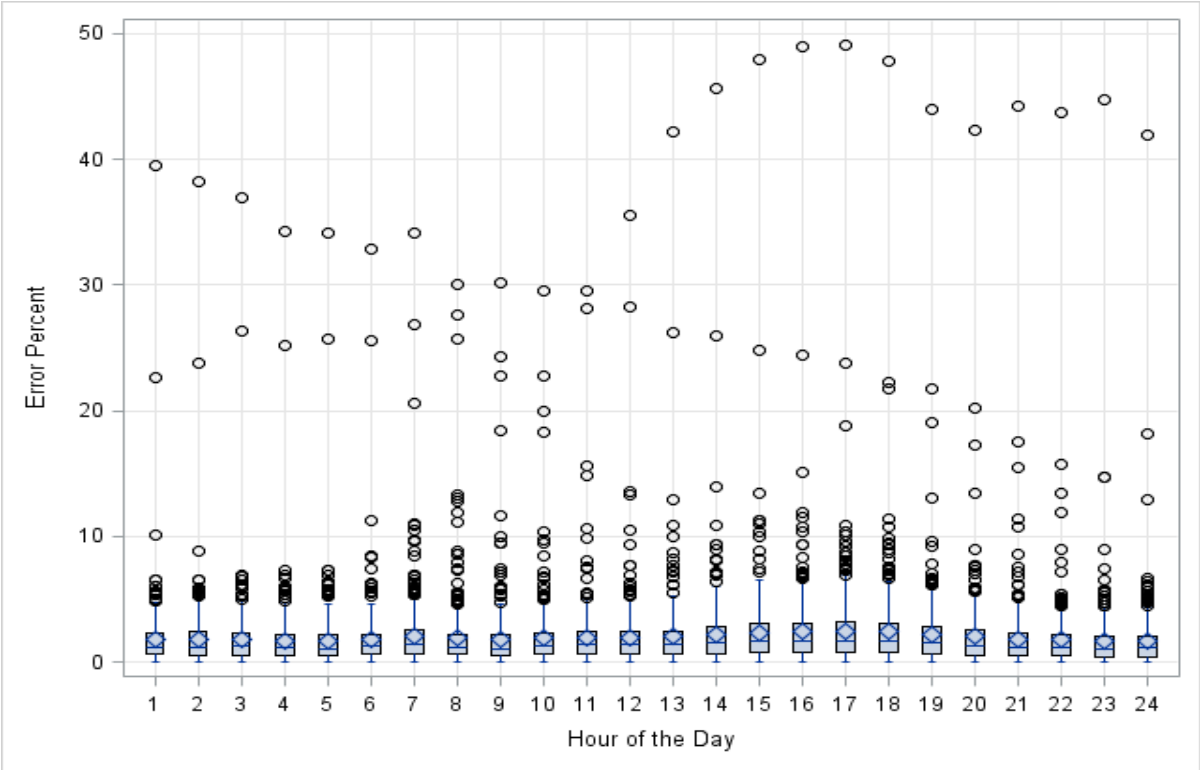


Figure 39: Out-of-Sample Boxplots of Hourly Error Percent for the Zonal Aggregation.



Comparing the error distribution among hourly models between the zonal aggregation and the regional model reveals little difference between the two approaches for both within-sample (Figures 36 and 37) and out-of-sample periods (Figures 38 and 39). For the figures pertaining to the within-sample period, the presence of far error outliers for both approaches is greatest during the afternoon peak demand period and the morning ramp up period when most commercial and industrial users begin their daily activities. This can be seen in Figures 36 and 37 during hours 14 through 19 and hours 6 through 9, respectively. A comparison between out-of-sample and within-sample error distributions, for both approaches, is difficult given the differing magnitudes. The increased magnitude of outliers in Figures 38 and 39 mask the density of outliers as compared to that of Figures 36 and 37. These extreme outliers in the out-of-sample period make a straightforward comparison with the within-sample errors difficult. However, the similarity in error distributions between approaches is even more pronounced in the out-of-sample period. This can be seen when contrasting Figure 38 and Figure 39. It appears that there is no significant difference between regional and zonal approaches at least in terms of absolute error among hourly models.

CHAPTER 6 CONCLUSION AND FUTURE RESEARCH

This paper has proposed a novel short term load forecasting approach using penalized splines in a semiparametric regression framework using data provided by the regional system operator for New England, ISONE. The equivalence between penalized splines and the special case of mixed model methodology allowed for estimation using existing software and the investigation of possible improved forecasting performance resulting from independent zonal models as opposed to the aggregate regional models currently used at ISONE. The conclusions inferred from these forecasting results are presented below. This paper concludes by outlining further areas of research and their expected results.

6.1 Key Findings

The primary focus of this research was to apply an emerging modeling methodology to the common problem of forecasting short term energy demand. The semiparametric, additive approach for load forecasting proposed by Fan and Hyndman (2011) performed extremely well when applied to Australian historical data. While not identically reproduced, the novel mixed model approach adopted in this thesis also performs well in generating short term forecasts of load in New England.

There exist subtle, yet significant, differences between the specifications of Fan and Hyndman (2011) and that adopted here. For example, the mixed-effects approach preferred by Ruppert, et al., (2003) is not implemented in Fan and Hyndman (2011). The latter authors prefer a simple approach using pre-specified knot locations and OLS to fit

hourly forecasting models. Furthermore, the semiparametric methodology proposed by Fan and Hyndman (2011) is somewhat simpler than that used in this research. To be more specific, Fan and Hyndman (2011) estimate smooth functions with simple cubic regression splines rather with penalized splines. They do not report any model specification tests that validate the use of cubic smoothing splines (as opposed to penalized splines). It is unclear which modeling choice performs better, or whether one is more appropriate than the other.

For both approaches, the same binary indicator variables are used for working and nonworking days. However, Fan and Hyndman (2011) have an additional term for the "time of year" effect. This additional term is treated as a smooth function and is estimated with a cubic regression spline. Fan and Hyndman (2011) also treat the relationships between load and lagged demand values as smooth functions, rather than entering the model linearly. Finally, Fan and Hyndman (2011) limit their weather variable to temperature in degrees Celsius but include transformations of temperature. For instance, the maximum and minimum temperature in the last 24 hours, and the difference in recorded temperature between two weather sites is used rather than temperature itself. Discrepancies in selecting predictor variables and how to incorporate those variables into the model represent differences in approaches.

However, these differences do not indicate that a particular semiparametric specification is more appropriate than the other. While the specific models used in Fan and Hyndman (2011) differ from those used in this research, each was chosen through separate variable selection processes. In fact, differences in available resources, population levels

and regulatory policies can cause clear differences in the fundamental characteristics of one electricity market to another. This heterogeneity among wholesale markets supports the notion that a model which performs well in one electricity market may not necessarily perform comparably in another. What does remain constant between the two specifications is the underlying approach towards model building; both models fundamentally decompose electricity load into calendar effects, lagged demand effects, weather effects and stochastic noise.

Explicit comparison of forecasting performance between the two studies is also difficult. There are several reasons for this, but in particular, a valid comparison of forecasting performance (e.g., MAPEs) is not possible as Fan and Hyndman (2011) estimate half-hourly models and our research presents hourly models. In other words, the specific MAPEs reported by Fan and Hyndman (2011) can't be directly contrasted to those generated by the models presented in this thesis. While an exact comparison may not be straightforward, Fan and Hyndman (2011) do state that the overall MAPE and MAE for their models were calculated to be 1.88% and 0.11 gigawatts, respectively. This metric is of comparable magnitude to the 1.62% within-sample and 1.93% out-of-sample MAPE calculated using the mixed modelapproach. As Fan and Hyndman (2011) state, this is satisfactory performance compared with state-of-the art load forecasting techniques.[34]

Subtle differences aside, it is clear that semiparametric regression-based short term load forecasting models can perform extremely well. This conclusion is supported by adequate forecasting performance for all models estimated. As discussed previously, ISONE

[34] Page 5 of Fan and Hyndman (2011).

makes use of several forecasting methods to produce the day-ahead forecast. Currently, neither the forecasts of load, nor the forecasting performance for any of the ISONE models is available. However, the annual performance of ISONE forecasts is reported in a 2010 FERC report to Congress[35]. In the report, the overall MAPE metrics for day-ahead load forecasting range from approximately 1.6% to 2.0% and span the years 2005 to 2010. ISONE's MAPE for the current out-of-sample period (e.g., 2011) is not available. However, given that ISONE uses several models to produce a single forecast, the 1.66% within-sample MAPE and 1.93% out-of-sample MAPE are extremely competitive.

It is clearly defined whether the forecasting performance reported in the FERC report are ex ante or ex post. However a discussion surrounding ISONE's weather vendors describes the potential increase in forecast error that stems from poor weather forecasts. This leads us to believe that the forecasting performance metrics reported for ISONE may be calculated using weather forecasts as predictor variables and therefore constitute ex ante forecasting performance. A logical comparison would be to estimate the semiparametric models with actual weather observations and then to use forecasts of weather to produce ex ante forecasts. This is discussed further in the next section.

One of the additional research objectives outlined in this thesis, and an always-current topic of interest in the forecasting literature, is to investigate whether or not a forecasting model based on aggregated data performs better than estimating multiple (zonal) forecasting models and then aggregating the forecasts. A main finding of this research is that an aggregation of zonal forecasts does not necessarily perform better than that which

[35] http://www.ferc.gov/industries/electric/indus-act/rto/metrics/iso-ne-rto-metrics.pdf

is produced from a single regional model. On an hourly, monthly and annual basis, both approaches perform very similarly and there is no clear winner. However, the single regional model does perform slightly better overall than the zonal aggregation in both within sample and out of sample periods.

6.2 Areas of Further Research

There were numerous opportunities for further research identified during the process of fitting these models and producing forecasts. A very clear progression from the ex post forecasting performance evaluated in Chapter 5 would be to produce ex ante forecasts with the hourly models fitted with PROC MIXED and evaluate their performance. If the idea of adding a semiparametric STLF model to the group of models currently used at ISONE was suggested, an analysis of ex ante forecasting performance would be required. After all, these are in fact what the models actually get used for. Note: in order for the fitted models in Chapter 5 to be used to produce ex ante forecasts, a previously forecasted value of hourly demand would have to be provided in addition to the forecasted weather. This is because the day-ahead forecast must be issued at 10 a.m. and, at the time of forecasting, the most recent observation of load would be lagged by 36 hours. As such, the observation for y_{t-24} used in making forecasts would have to be itself a previously forecasted value. Another possibility would be to estimate the models exactly as described in Chapter 4, but omitting y_{t-24}. These models would be very similar, and wouldn't require a previously forecasted value of demand. In either scenario, evaluating the semiparametric STLF model in terms of its ex ante forecasting performance is one area which warrants further research.

Another area relates to improving individual zonal forecasting performance. For instance, there is currently no analysis that focuses on how each load zone is paired with its representative weather station other than proximity. More detailed and granular weather data that are matched to zonal populations (and therefore load) might be one way of improving zonal forecasts. This idea is also mentioned in Fan and Hyndman (2011) as a potential research area.

While the weather variables used in fitting the semiparametric models (e.g., wind speed, cloud cover, temperature and humidity) are all used as inputs in the ISONE models, comfort indexes and aggregated proxies such as effective temperature and THI are also used by ISONE. Furthermore, the selection of weather variables used as model inputs changes over the course of a year. Specifically, effective temperature is used in the winter while THI is used in the summer. A more thorough analysis of weather variable selection could potentially lead to improved forecasting performance.

Within the framework of penalized splines and mixed models, a clear extension of the model specifications currently used would be to fit comparable models using other spline basis functions. For instance, radial or B-spline basis could be calculated for all of the smooth functions to be estimated. Other higher or lower degree polynomials of the Truncated Power Functions are another option for investigation.

Finally, an important extension to this research is exploration of the time-series properties of both load and weather variables at the frequency levels presented in this research. To date, this topic has received minimal attention. Nonstationarity in load or

weather variables or both may be an issue. If so, they may be cointegrated. This would suggest alternative modeling procedures that enhance forecasting performance.

CHAPTER 7 REFERENCES

Adams, Gail, P. Geoffrey Allen and Bernard J. Morzuch. "Probability Distributions of short-term electricity peak load forecasts." *International Journal of Forecasting* 7 (1991): 283-297.

Alfares, Hesham K. and Mohammad Nazeeruddin. "Electric Load Forecasting: Literature Survey and Classification of Methods." *International Journal of Systems Science* 33.1 (2002): 23-34.

Amjady, N. "Day-Ahead Price Forecasting Methods: An Evaluation Based on European Data." *IEEE Transactions on Power Systems* (2006): 887-896.

Barakat, E., J.M. Al-Qassim and S.A. Al-Rashed. "New Model for Peak Demand Forecasting Applied to Highly Complex Load Characteristics of a Fast Developing Area." *IEE Proceedings - C* 139 (1992): 136-149.

Begun, Janet, et al. "Information and Asymptotic Efficienty in Parametric-Nonparametric Models." *The Annals of Statistics* 11.2 (1983): 432-452.

Benjamin, Richard M. "An Electricity Transmission Primer for Energy Economists Part 1 and Part 2." *Politics & Energy eJournal* 2.21 (2013).

Bernard, Jean-Thomas and Michael R. Veall. "The Probability Distributions of Future Demand: The Case of Hydro Quebec." *Journal of Business & Economic Statistics* (1987): 417 424.

Black, Jonathan. “Load Hindcasting: A Retrospective Regional Load Predicition Method Using Reanalysis Weather Data.” Master's Thesis. 2011.

Brown, R.E., A.P. Hanson and D.L. Hagan. "Long Range Spatial Load Forecasting using Non-Uniform Areas." *IEEE Transmission and Distribution Conference Proceedings* (1999): 369-373.

Brumback, Babette A., David Ruppert and M. P. Wand. "Comment on Shively, Kohn and Wood." (1999).

Bunn, D. W. and E. D. Farmer. *Comparative Models for Electrical Load Forecasting*. New York: John Wiley & Sons, 1985.

Calderon, Christopher P., et al. "PSQR: A Stable and Efficient Penalized Spline Algorithm." Technical Paper. 2009.

Chandrashekara, A.S., T. Ananthapadmanabha and A.D. Kulkarni. "A Neuro-Expert System for Planning and Load Forecasting of Distribution Systems." *Electrical Power and Energy Systems Research* (1999): 309-314.

Chen, Ying, et al. "Short-term Load Forecasting: Similar Day-Based Wavelet Neural Networks." *IEEE Transactions on Power Systems* (2010): 322-330.

Chow, M., J. Zhu and H. Tram. "Application of Fuzzy Multi-Objective Decision making in Spatial Load Forecasting." *IEEE Transactions on Power Systems* (1998): 1185-1190.

Clean Air - Cool Planet. *Indicators of Climate Change in the Northeast 2005*. Report. University of New Hampshire. Portsmouth: Clean Air - Cool Planet, 2005.

Connor, J.T. "A Robust Neural Network Filter for Electricity Demand Prediction." *Journal of Forecasting* (1996): 437-458.

Corpening, S.L., N.D. Reppen and R.J. Ringlee. "Experience with Weather Sensitive Load Models For Short and Long-Term Forecasting." New York: IEEE Power Engineering Society, 1972. 1966-1972.

Crainiceanu, M. Ciprian and David Ruppert. "Restricted Likelihood Ratio Tests in Nonparametric Longitudinal Modles." *Statistica Sinica* (2004): 713-729.

Craven, P. and G. Wahba. "Smoothing Noisy Data with SPline FUnctions: Estimating the Correct Degree of SMoothing by Method of Generalized Cross-Validation." *Numerische Mathematik* 31 (1979): 377-403.

Cuzick, Jack. "Semiparametric Additive Regression." *Journal of the Royal Statistical Society* 54.3 (1992): 831-843.

Darbellay, Georges A. and Marek Slama. "Forecasting the short-term demand for electricity: Do neural networks stand a better chance?" *International Journal of Forecasting* (2000): 71-83.

Dash, P.K., et al. "Fuzzy and Neuro-Fuzzy Computing Models for Electric Load Forecasting." *Engineering Applications of Artificial Intelligence* (1995): 423-433.

de Boor, Carl R. *A Practical Guide to Splines*. New York: Springer, 1979.

Dominici, Francesca, Aidan McDermott and Trevor J. Hastie. "Improved Semiparametric Time Series Models of Air Pollution and Mortality." *Journal of the American Statistical Association* 99.468 (2004): 938-948.

Eilers, Paul H.C. and Brian D. Marx. "Splines, Knots and Penalties." Working Paper. 2004.

El-Keib, A.A., X. Ma and H. Ma. "Advancement of Statistical Based Modeling for Short-Term Load Forecasting." *Electric Power Systems Research* 35 (1995): 51-58.

Enders, Walter. *Applied Econometric Time Series*. John Wiley & Sons, 2004.

Engle, Robert F., et al. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Jounral of the American Statistical Association* (1986): 310 - 320.

Eubank, R. L. *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, 1988.

Fan, Shu and L Chen. "Short-term load forecasting based on adaptive hybrid method." *IEEE Transactions on Power Systems* 21.1 (2006): 1821-1830.

Fan, Shu and Rob J Hyndman. "Short-Term Load Forecasting Using Semi-parametric additive models." *2011 IEEE Power and Energy Society General Meeting*. n.d.

Fay, D., et al. "24-h electrical load data - a sequential or partitioned time series?" *Neurocomputing* 55.3-4 (2003): 469-498.

Federal Energy Regulatory Commission. "Performance Metrics: For Independant System Operators and Regional Transmission Organizations." Report to Congress. 2011. <http://www.ferc.gov/industries/electric/indus-act/rto/metrics/report-to-congress.pdf>.

Ferreira, V.H. and A.P. Alves da Silva. "Toward Estimating Autonomous Neural Network-Based Electric Load Forecasters." *IEEE Transactions on Power Systems* (2007): 1554-1562.

Gardner Jr., Everette S. “A Simple Method of Computing Predicition Intervals for Time Series Forecasts.” *Management Science* (1988): 541 - 546.

Gelper, S., R. Fried and C. Croux. "Robust Forecasting with Exponential and Holt-Winters Smoothing." *Journal of Forecasting* (2010): 285-300.

Gelper, Sarah and Roland Fried. "Robust Forecasting with Expoential and Holt-Winters Smoothing." Katholieke Unicersiteit Leuven, 2007.

Golub, Gene H. and Charles F. Van Loan. *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

Green, P.J. and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Suffolk: Clays Ltd, 1994.

Gross, G. and F.D. Galiana. "Short term load forecasting." *Proceedings of the IEEE* 75 (1987): 1558-1573.

Harvey, Andrew and Siem Jan Koopman. "Forecasting Hourly Electricity Demand Using Time-Varying Splines." *Journal of the American Statistical Association* 88.424 (1993): 1228-1236.

Hastie, Trevor and Robert Tibshirani. "Generalized Additive Models." *Statistical Science* 1.3 (1986): 297-318.

Hastie, Trevor. "Pseudosplines." *Journal of the Royal Statistical Society* 58.2 (1996): 379-386.

Heinemann, G.T., D.A. Nordman and E.C. Plant. "The Relationship Between Summer Weather and Summer Loads - A regression analysis." *IEEE Transactions of Power Apparatus and Systems* (1966): 1144-1154.

Hippert, H.S. and C.E. Pedreira. "Estimating Temperature Profiles for Short-Term Load Forecasting: Neural Networks Compared to Linear Models." *IEE Proceedings - Generation, Transmissing, and Distribution* (2004): 543-547.

Hippert, Henrique Steinherz, Carlos Eduardo Pedreira and Reinaldo Castro Souza. "Neural Networks for Short-Term Load Forecasting: a Review and Evaluation." *IEEE Transactions on Power Systems* (2001): 44 - 55.

Ho, K., et al. "Short Term Load Forecasting of Taiwan Power System using a Knowledge Based Expert System." *IEEE Transactions on Power Systems* (1990): 1214-1221.

Holt, C.C. "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages." *ONR Research Memorandum*. 1959. 52.

Huang, S.R. "Short-Term Load Forecasting using Threshold Autoregressive Models." *IEE Proceedings: Generation, Transmission, and Distribution* (1997): 477-481.

Hyndman, R., et al. "Prediction Intervals for Exponential Smoothing Using Two New Classes of State Space Models." *Journal of Forecasting* 24 (2005): 17-37.

Infield, D.G. and D.C. Hill. "Optimal Smoothing for Trend Removal in Short Term Electricity Demand Forecasting." *IEEE Transactions on Power Systems* 13 (1998): 1115-1120.

ISO New England, Inc. *ISO New England Company History*. 2012. 28th January 2011. <http://www.iso-ne.com/>.

—. "SOP-OUTSCH.0040.0010 - Create Demand Forecast." 17 January 2013. *www.iso-ne.com.* Standard Operating Procedure. 20 January 2013. <http://www.iso-ne.com/rules_proceds/operating/sysop/out_sched/sop_outsch_0040_0010.pdf>.

—. "WEM 101: Day-Ahead Energy Markets." n.d.

—. "WEM 101: Forecast and Scheduling - Reserve Adequacy Assessment." n.d.

Juberias, G, et al. "A New ARIMA Model for Hourly Load Forecasting." *IEEE Transmission and Distribution Conference Proceedings* 1 (1999): 314-319.

Kalaitzakis, K., G.S. Stavrakakis and E.M. Anagostakis. "Short-term Load Forecasting Based on Artificial Neural Networks Parallel Implementation." *Electric Power Systems Research* (2002): 185-196.

Kim, K.H., et al. "Implementaion of Hybrid Short-Term Load Forecasting System Using Artificial Neural Networks and Fuzzy Expert Systems." *IEEE Transactions on Power Systems* (1995): 1534-1539.

Liao, Shu-Hsien. "Expert System Methodologies and Applications - A Decade Review from 1995 to 2004." *Expert Systems with Applications* 2004.

Liu, K., et al. "Comparison of Very Short-Term Load Forecasting Techniques." *IEEE Transactions on Power Systems* 11.2 (1996): 877 - 882.

Matthewman, P.D. and H. Nicholson. "Techniques for Load Prediction in Electricity Supply Industry." *IEEE Transactions on Power Systes* 115 (1968): 1451-1457.

McCulloch, C.E. and S.R. Searle. *Generalized, Linear and Mixed Models*. New York: Wiley, 2001.

Mohamad, E.A., et al. "Results of Egyptian Unified Grid Hourly Load Forecasting Using an Artificial Neural Network with Expert System Interfaces." *Electric Power Systems Research* (1996): 171-177.

Moler, Cleve. “Least Squares.” *Numerical Computing with Matlab*. 2004. 1 - 27.

Mu, Qingqing, et al. "Short-term Load Forecasting Using Improved Similar Day Methods." *IEEE Transactions on Power Systems* (2010).

National Climatic Data Center. *Billion Dollar U.S. Weather/Climate Disasters*. 2011. December 2011. <http://www.ncdc.noaa.gov/oa/reports/billionz.html>.

Ngo, Long and M.P. Wand. "Smoothing with Mixed Model Software." *Journal of Statistical Software* (2004): 1 - 53.

Piras, A and B Buchenel. "A Tutorial on Short Term Load Forecasting." *International Journal of Engineering* 7.1 (1999): 41-48.

Rahman, S. and G. Shreshta. "A Priority Vector Based Technique for Load Forecasting." *IEEE Transactions on Power Systems* (1991): 1459-1465.

Rahman, S. and O. Hazim. "Load Forecasting for Multiple Sites: Development of an Expert System-Based Technique." *Electric Power Systems Research* (1996): 509-514.

Ramanathan, R., et al. "Short-run forecasts of electricity loads and peaks." *International Journal of Forecasting* 13 (1997): 161-174.

Ramanathan, Ramu, et al. "Short-run forecasts of electricity loads and peaks." *International Journal of Forecasting* (1997): 161-174.

Reinsch, Christian H. "Smoothing by Spline Functions." *Numerishe Mathematik* (1967): 177-183.

Robinson, G. K. "That BLUP is a Good Thing: The Estimation of Random Effects." *Statistical Science* 6.1 (1991): 15 - 32.

Rui, Y. and A.A. El-Keib. "A Review of ANN-based Short-Term Load Forecasting Models." University of Alabama, Tuscaloosa, n.d.

Ruppert, David, M.P. Wand and J Raymond Carrol. “Semiparametric Regression during 2003-2007.” *Electronic Journal of Statistics* (2009): 1193 - 1256.

Ruppert, David, M.P. Wand and R.J. Carrol. *Semiparametric Regression*. Cambridge, U.K.: Cambridge University Press, 2003.

Sigauke, Caston and Delson Chikobvu. "Daily Peak Electricity Load Forecasting in South Africa Using a Multivariate Non-Parametric Regression Approach." *Orion* 26 (2010): 97-111.

Smith, Michael. "Modeling and Short-Term Forecasting of New South Wales Electricity System Load." *Journal of Business and Economic Statistics* 18.4 (2000): 465-478.

Soliman, Soliman Abdel-hady and Ahmad M. Al-Kandari. *Electrical Load Forecasting*. Burlington, MA: Elsevier Inc., 2010.

Srinivasan, D., C.S. Chang and A.C. Liew. "Demand Forecasting Using Fuzzy Neural Computation with Special Emphasis on Weekend and Public Holiday Forecasting." *IEEE Transactions on Power Systems* (1992): 343-348.

Srinivasan, D., et al. "Parallel Neural Network-Fuzzy Expert System for Short-Term Load Forecasting: System Implementation and Performance Evaluation." *IEEE Transactions on Power Systems* (1999): 1100 - 1106.

Stephens, M.A. "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association* 69.347 (1974): 730-737.

Stone, Charles J. "Additive regression and other nonparametric models." *The Annals of Statistics* (1985): 686-705.

Taylor, James and Roberto Buizza. "Neural Network Load Forecasting with Weather Ensemble Predictions." *IEEE Transactions on Power Systems* 17 (2002): 626-632.

Taylor, James W. and Robert Buizza. "Using Weather Ensemble Predictions in Electricity Demand Forecasting." *International Journal of Forecasting* (2003): 57 - 70.

U.S. Energy Information Administration. *Electricity: Detailed State Level*. 4th January 2012. 22 June 2012. <http://www.eia.gov/electricity/data/state/>.

U.S. Energy Informatuon Administration. *Energy Explained*. 2012 йил 26-April. 2012 йил 10-June. <http://www.eia.gov/energyexplained/index.cfm?page=electricity_use>.

U.S. Environmental Protection Agency. *Emissions Frequently Asked Questions*. n.d. 26th January 2012. <http://www.epa.gov/climatechange/fq/emissions.html>.

Winters, P.R. "Forecasting Sales by Exponentially Weighted Moving Averages." *Management Science* 6 (1960): 324-342.

Yun, Z., et al. "RBF Neural Network and ANFIS-Based Short-Term Load Forecasting Approach in Real-Time Price Environment." *IEEE Transactions on Power Systems* (2008): 853 - 858.

Zhang, Guoqiang Peter. "Neural Networks for Classification: A Survey." *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews* (2000): 451-462.

Zhao, H., Z. Ren and W. Huang. "Short-Term Load Forecasting Considering Weekly Period Based on Periodical Auto Regression." *Proceedings of the Chinese Society of Electrical Engineers* 17 (1997): 211-213.

Zielinski, Gregory A. and Barry D. Keim. *New England Weather, New England Climate*. Lebanon: University Press of New England, 2003.

APPENDICES

A. SUMMARY OF DATA SETS

Table 7-1: Summary Statistics for Regional Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	14,619.75	2,854.91	8,296.00	27,707.00
Temperature	29,928.00	49.87	18.05	-6.00	100.00
Humidity	29,928.00	38.06	19.02	-20.00	74.00
Wind Speed	29,928.00	8.62	4.32	0.00	34.00
Cloud Cover	29,928.00	4.22	2.79	0.00	10.00

Table 7-2: Summary Statistics for NEMASS Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	2,973.27	564.95	1,911.00	5,716.00
Temperature	29,928.00	51.71	17.31	-2.00	102.00
Humidity	29,928.00	39.01	18.83	-19.00	75.00
Wind Speed	29,928.00	10.85	5.49	0.00	44.00
Cloud Cover	29,928.00	4.77	3.18	0.00	10.00

Table 7-3: Summary Statistics for SEMASS Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	2,048.60	388.33	768	3,715.00
Temperature	29,928.00	51.54	17.58	-1	101.00
Humidity	29,928.00	39.59	19.23	-19	77.00
Wind Speed	29,928.00	8.55	5.34	0.00	39.00
Cloud Cover	29,928.00	4.87	3.08	0.00	10.00

Table 7-4: Summary Statistics for WCMASS Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	1,735.45	388.00	889	3,700.00
Temperature	29,928.00	49.75	18.51	-7	99.00
Humidity	29,928.00	37.92	19.33	-20	76.00
Wind Speed	29,928.00	8.82	4.76	0.00	58.00
Cloud Cover	29,928.00	4.31	2.99	0.00	10.00

Table 7-5: Summary Statistics for Connecticut Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	3,603.14	767.75	1628	7,303.00
Temperature	29,928.00	51.14	18.54	-5	102.00
Humidity	29,928.00	38.67	19.26	-19	75.00
Wind Speed	29,928.00	7.47	5.09	0.00	71.00
Cloud Cover	29,928.00	4.62	2.92	0.00	10.00

Table 7-6: Summary Statistics for Rhode Island Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	942.59	203.13	370	1,967.00
Temperature	29,928.00	51.54	17.58	-1	101.00
Humidity	29,928.00	39.59	19.23	-19	77.00
Wind Speed	29,928.00	8.55	5.34	0.00	39.00
Cloud Cover	29,928.00	4.87	3.08	0.00	10.00

Table 7-7: Summary Statistics for Vermont Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	683.65	113.86	414	1,023.00
Temperature	29,928.00	46.55	20.09	-19	96.00
Humidity	29,928.00	36.10	19.62	-25	74.00
Wind Speed	29,928.00	7.10	5.45	0.00	36.00
Cloud Cover	29,928.00	4.60	3.54	0.00	10.00

Table 7-8: Summary Statistics for New Hampshire Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	1,326.09	266.17	540	2,467.00
Temperature	29,928.00	46.77	19.79	-22	100.00
Humidity	29,928.00	35.84	19.47	-26	75.00
Wind Speed	29,928.00	5.60	5.42	0.00	35.00
Cloud Cover	29,928.00	3.56	3.87	0.00	10.00

Table 7-9: Summary Statistics for Maine Data Set.

Variable	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Load	29,928.00	1,307.42	210.21	806	2,025.00
Temperature	29,928.00	47.08	17.84	-14	100.00
Humidity	29,928.00	37.12	19.05	-23	76.00
Wind Speed	29,928.00	7.48	5.24	0.00	53.00
Cloud Cover	29,928.00	3.64	3.73	0.00	10.00

B. ZONAL FORECASTING RESULTS

Load Zone: Connecticut.

Table 7-10: Overall Forecasting Results - Connecticut Load Zone.

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	78.289605	2.11%	12635.4002	98.51226	2.88%	29516.83448

Table 7-11: Hourly Forecasting Results - Connecticut Load Zone.

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	54.72	1.84%	5,679.08	71.84	2.59%	15,598.00
2	50.97	1.81%	5,006.41	69.28	2.61%	14,279.56
3	50.13	1.84%	4,648.85	67.52	2.63%	13,938.87
4	50.47	1.88%	4,553.98	65.36	2.57%	12,515.13
5	51.85	1.90%	4,927.63	66.08	2.56%	12,731.10
6	57.24	1.97%	5,887.98	71.94	2.62%	14,783.71
7	73.13	2.26%	9,745.47	92.33	3.03%	24,463.49
8	75.63	2.15%	11,661.47	98.70	3.01%	29,757.80
9	70.51	1.89%	10,007.17	96.10	2.80%	30,759.27
10	73.10	1.86%	10,146.05	98.47	2.76%	31,680.12
11	76.58	1.89%	10,971.05	102.89	2.79%	34,024.75
12	83.93	2.03%	13,188.43	107.86	2.90%	36,569.88
13	91.37	2.20%	15,118.78	110.73	2.96%	36,789.54
14	102.03	2.45%	18,996.90	116.52	3.09%	38,820.63
15	110.12	2.66%	22,202.60	127.57	3.36%	42,564.45
16	113.00	2.74%	25,269.11	125.13	3.34%	41,532.63
17	113.46	2.72%	24,432.39	134.64	3.50%	45,029.68
18	108.64	2.58%	22,823.05	136.10	3.46%	47,614.30
19	98.60	2.36%	18,786.33	130.23	3.29%	42,817.65
20	88.79	2.14%	15,513.48	116.05	2.98%	36,478.24
21	85.87	2.06%	15,875.40	105.19	2.76%	33,351.69
22	76.32	1.92%	12,732.26	96.48	2.65%	29,450.30
23	64.55	1.77%	8,418.39	83.18	2.50%	24,589.78
24	57.94	1.77%	6,657.34	74.13	2.46%	18,263.47

Table 7-12: Monthly Forecasting Results - Connecticut Load Zone.

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	67.19	1.67%	8,189.98	92.14	2.33%	14,755.45
February	67.28	1.78%	8,310.98	76.94	1.98%	9,667.72
March	58.30	1.69%	5,888.73	62.80	1.77%	6,378.16
April	69.46	2.13%	12,012.08	70.60	2.20%	9,031.76
May	67.83	2.02%	10,126.20	64.45	1.82%	10,380.24
June	89.69	2.35%	14,642.93	103.69	2.74%	18,609.34
July	120.27	2.84%	26,743.67	143.53	3.28%	34,508.22
August	109.69	2.64%	21,467.51	168.75	5.59%	114,172.34
September	84.87	2.36%	13,370.89	89.22	2.40%	14,390.44
October	57.18	1.80%	6,026.42	117.05	4.49%	69,338.45
November	61.57	1.86%	9,415.43	109.61	3.62%	34,179.79
December	83.59	2.14%	14,474.08	80.61	2.27%	15,519.03

Load Zone: Northeastern Massachusetts (NEMASS).

Table 7-13: Overall Forecasting Results - NEMASS.

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	59.153651	1.92%	7,434.12	63.93485	2.06%	9,406.97

Table 7-14: Hourly Forecasting Performance - NEMASS

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	38.06	1.50%	3,122.33	41.14	1.59%	3,658.44
2	34.88	1.44%	2,603.95	40.56	1.65%	3,376.87
3	35.42	1.49%	2,728.85	41.50	1.73%	3,604.99
4	33.27	1.42%	2,474.02	39.75	1.69%	3,260.05
5	33.24	1.42%	2,237.95	39.15	1.64%	3,074.10
6	36.97	1.50%	2,675.28	42.04	1.68%	3,494.50
7	44.49	1.66%	3,899.84	50.54	1.84%	5,253.08
8	49.67	1.71%	5,132.33	53.72	1.82%	6,705.84
9	49.71	1.63%	4,972.50	55.56	1.80%	6,752.23
10	54.46	1.72%	5,795.33	61.23	1.91%	7,763.43
11	59.70	1.82%	6,531.42	66.60	2.02%	8,686.41
12	64.91	1.94%	8,081.62	70.77	2.12%	10,168.79
13	73.40	2.17%	9,825.11	76.20	2.26%	11,664.53
14	80.67	2.39%	11,513.13	84.41	2.50%	14,782.80
15	87.63	2.60%	13,857.70	90.27	2.67%	17,094.33
16	91.18	2.70%	14,765.33	91.40	2.70%	17,517.57
17	93.85	2.76%	15,608.50	97.65	2.85%	20,211.98
18	89.20	2.60%	14,862.98	97.21	2.81%	19,325.51
19	80.19	2.37%	12,764.40	89.67	2.61%	15,960.27
20	70.74	2.10%	9,828.81	77.01	2.26%	12,936.64
21	67.74	2.02%	9,400.10	70.75	2.09%	11,289.62
22	59.82	1.86%	7,310.37	62.42	1.92%	8,851.78
23	48.73	1.63%	4,837.64	50.93	1.68%	6,253.28
24	41.75	1.53%	3,589.40	43.98	1.59%	4,080.32

Table 7-15: Monthly Forecasting Performance - NEMASS

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	46.71	1.44%	4,386.12	51.01	1.58%	4,998.04
February	43.43	1.41%	3,296.34	49.83	1.60%	4,266.95
March	50.79	1.74%	4,714.83	47.24	1.59%	3,745.69
April	49.76	1.79%	5,577.23	50.67	1.84%	5,029.08
May	55.32	1.94%	6,345.89	52.46	1.77%	7,634.02
June	62.24	1.99%	7,613.55	85.23	2.62%	18,396.11
July	100.35	2.92%	19,585.26	111.03	3.08%	20,729.06
August	88.23	2.56%	13,609.08	94.74	2.88%	17,768.97
September	55.77	1.82%	5,970.91	62.19	2.02%	7,488.90
October	50.70	1.84%	5,037.50	57.38	2.05%	8,895.07
November	43.34	1.55%	4,550.35	48.68	1.75%	5,362.37
December	59.46	1.88%	7,583.50	55.10	1.87%	8,028.40

Load Zone: Southeastern Massachusetts (SEMASS)

Table 7-16: Overall Forecasting Results - SEMASS.

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	47.66	2.29%	4,640.54	56.77	2.90%	8,222.79

Table 7-17: Hourly Forecasting Results - SEMASS.

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	38.27	2.24%	2,839.07	48.31	2.92%	5,928.39
2	36.48	2.23%	2,578.99	46.40	2.95%	5,347.61
3	34.60	2.17%	2,210.07	44.06	2.88%	4,809.65
4	33.40	2.12%	2,003.82	41.55	2.76%	4,367.80
5	32.41	2.02%	1,886.30	39.69	2.62%	4,378.59
6	33.61	1.99%	2,071.86	42.46	2.66%	4,920.37
7	41.30	2.22%	3,350.41	49.66	2.84%	6,736.55
8	42.33	2.11%	3,958.69	50.21	2.70%	7,884.62
9	40.35	1.90%	3,457.80	51.47	2.65%	8,051.07
10	41.26	1.87%	3,371.48	55.23	2.73%	8,639.01
11	42.98	1.89%	3,586.60	57.80	2.78%	9,194.87
12	47.28	2.04%	4,226.23	59.13	2.80%	9,449.41
13	51.33	2.21%	4,841.31	60.65	2.85%	9,797.92
14	56.77	2.43%	6,014.67	66.38	3.09%	10,717.14
15	62.05	2.68%	7,227.77	68.74	3.19%	10,953.16
16	64.70	2.81%	7,593.81	71.27	3.29%	11,297.95
17	64.27	2.76%	7,756.45	71.44	3.25%	11,253.18
18	64.96	2.76%	8,258.61	74.24	3.28%	12,292.93
19	61.04	2.62%	7,325.48	70.45	3.09%	10,827.53
20	58.00	2.49%	6,589.24	65.97	2.92%	9,646.76
21	55.83	2.42%	6,562.24	63.36	2.85%	9,167.63
22	53.28	2.40%	5,924.41	61.18	2.86%	8,719.81
23	46.76	2.30%	4,469.00	54.19	2.77%	7,100.46
24	40.56	2.20%	3,268.75	48.62	2.72%	5,864.62

Table 7-18: Monthly Forecasting Results - SEMASS.

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	37.76	1.68%	2,594.87	40.91	1.84%	3,170.77
February	32.83	1.54%	1,748.58	42.07	1.91%	3,116.29
March	40.43	2.05%	2,758.68	44.31	2.16%	3,115.44
April	37.72	2.03%	2,928.97	40.00	2.15%	2,873.61
May	45.11	2.35%	4,384.27	46.92	2.36%	4,984.58
June	47.42	2.25%	4,540.60	67.74	3.24%	7,495.85
July	77.22	3.38%	10,951.97	91.16	3.82%	13,626.14
August	70.13	3.08%	8,044.80	77.42	3.69%	11,831.96
September	50.34	2.44%	5,047.91	49.79	2.43%	4,645.84
October	37.25	1.99%	2,710.95	76.69	5.73%	31,294.73
November	38.89	2.06%	3,606.15	49.63	2.67%	5,628.22
December	53.71	2.44%	5,779.29	52.54	2.61%	6,000.83

Load Zone: Southeastern Massachusetts (WCMASS).

Table 7-19 : Overall Forecasting Results - WCMASS

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	41.951606	2.31%	3,942.08	48.60843	2.72%	6,481.79

Table 7-20: Overall Forecasting Results - WCMASS

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	28.80	2.01%	1,785.72	35.60	2.48%	3,256.02
2	26.72	1.97%	1,521.08	33.36	2.47%	2,832.29
3	24.45	1.87%	1,255.19	32.27	2.46%	2,600.99
4	24.26	1.88%	1,196.21	31.54	2.42%	2,464.94
5	23.81	1.82%	1,128.86	31.64	2.40%	2,454.37
6	25.22	1.83%	1,204.25	35.15	2.52%	2,867.70
7	33.82	2.22%	2,155.50	44.84	2.90%	4,538.65
8	36.08	2.15%	2,611.47	44.22	2.65%	5,010.69
9	34.37	1.91%	2,376.21	41.50	2.33%	4,187.23
10	37.04	1.95%	2,705.02	43.10	2.29%	4,166.84
11	41.47	2.10%	3,390.32	44.90	2.29%	4,698.87
12	44.39	2.21%	3,775.49	45.79	2.32%	5,267.27
13	49.57	2.47%	4,772.18	52.53	2.65%	7,287.12
14	55.32	2.74%	5,933.02	60.17	3.01%	9,899.06
15	59.33	2.98%	7,035.30	66.82	3.37%	11,865.85
16	64.22	3.21%	8,379.08	68.78	3.49%	12,270.55
17	64.40	3.18%	8,322.79	71.93	3.59%	12,675.74
18	62.52	3.02%	7,950.88	72.09	3.50%	12,668.69
19	58.59	2.81%	7,017.59	66.52	3.19%	11,278.02
20	53.12	2.54%	5,797.81	59.08	2.84%	9,445.75
21	48.94	2.33%	5,303.15	52.02	2.53%	7,586.10
22	41.96	2.12%	3,985.90	49.10	2.51%	6,625.20
23	36.48	2.05%	2,854.57	44.50	2.52%	5,537.04
24	31.95	2.03%	2,152.22	39.14	2.48%	4,077.95

Table 7-21: Overall Forecasting Results - WCMASS

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	33.65	1.82%	2,211.64	35.72	1.94%	2,707.31
February	28.20	1.61%	1,403.02	34.25	1.91%	1,874.51
March	33.10	1.99%	1,882.33	35.03	2.07%	2,086.08
April	40.32	2.56%	4,762.67	34.09	2.17%	2,350.99
May	36.60	2.27%	3,268.82	34.33	2.12%	1,965.39
June	49.76	2.66%	4,960.40	64.17	3.39%	8,168.49
July	67.32	3.16%	8,496.98	81.73	3.65%	13,306.58
August	71.94	3.37%	9,253.68	99.97	5.70%	31,315.58
September	41.20	2.32%	3,255.76	51.83	2.87%	5,127.30
October	29.39	1.85%	1,919.20	32.68	2.07%	1,920.24
November	30.20	1.86%	2,185.59	34.22	2.11%	2,833.14
December	39.25	2.13%	3,244.25	43.56	2.51%	3,439.74

Load Zone: Rhode Island.

Table 7-22: Overall Forecasting Results – Rhode Island.

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	19.92	2.06%	813.24	23.64	2.65%	2,096.11

Table 7-23: Hourly Forecasting Results – Rhode Island

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	14.76	1.88%	417.56	18.77	2.46%	1,140.67
2	14.12	1.89%	377.07	17.28	2.39%	967.16
3	13.30	1.84%	351.12	16.40	2.34%	829.02
4	13.02	1.83%	336.09	15.39	2.24%	760.36
5	12.58	1.75%	302.15	15.70	2.26%	762.04
6	13.26	1.76%	317.80	16.88	2.30%	872.41
7	16.87	2.05%	552.29	21.09	2.62%	1,323.58
8	18.76	2.08%	759.05	22.69	2.60%	1,691.74
9	17.99	1.87%	698.16	22.71	2.45%	1,811.78
10	18.26	1.79%	676.99	23.41	2.47%	2,170.60
11	19.48	1.84%	748.20	25.20	2.68%	2,665.45
12	21.68	2.01%	894.77	25.43	2.70%	2,887.59
13	23.67	2.18%	1,064.32	26.46	2.82%	3,094.51
14	25.78	2.37%	1,270.37	28.53	2.99%	3,194.55
15	27.61	2.55%	1,495.53	30.24	3.15%	3,200.89
16	28.32	2.65%	1,486.43	32.11	3.32%	3,340.93
17	27.91	2.60%	1,397.49	32.79	3.34%	3,350.03
18	27.44	2.51%	1,423.60	32.93	3.28%	3,353.93
19	25.23	2.34%	1,225.60	29.86	3.00%	2,835.03
20	22.35	2.06%	925.24	27.29	2.73%	2,583.39
21	21.31	1.97%	901.08	24.15	2.44%	2,367.45
22	20.09	1.93%	784.06	22.93	2.38%	2,042.22
23	18.28	1.92%	630.71	20.30	2.30%	1,615.80
24	16.02	1.87%	482.14	18.84	2.36%	1,445.61

Table 7-24: Monthly Forecasting Results – Rhode Island

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	15.89	1.59%	459.25	17.64	1.79%	693.59
February	14.19	1.49%	346.78	16.43	1.68%	455.28
March	16.48	1.83%	467.03	15.51	1.69%	426.74
April	17.61	2.04%	898.41	14.66	1.72%	408.29
May	16.89	1.94%	622.25	17.20	1.88%	773.59
June	21.00	2.13%	803.06	28.73	2.91%	1,470.02
July	28.71	2.56%	1,449.49	37.78	3.18%	2,144.13
August	30.58	2.75%	1,515.22	53.45	7.94%	14,737.70
September	22.01	2.27%	877.51	25.78	2.66%	1,132.81
October	17.58	2.09%	683.31	17.05	1.94%	860.44
November	16.56	1.90%	678.13	17.78	2.04%	775.11
December	20.47	2.07%	873.52	20.73	2.24%	968.55

Load Zone: Vermont.

Table 7-25: Hourly Forecasting Results – Vermont

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	12.141588	1.78%	272.56	14.20642	2.08%	365.25

Table 7-26: Hourly Forecasting Results – Vermont

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	10.48	1.86%	192.90	11.93	2.12%	258.62
2	10.23	1.89%	181.91	11.68	2.15%	246.97
3	9.88	1.86%	172.94	11.42	2.15%	241.76
4	9.77	1.85%	168.76	10.93	2.07%	220.58
5	9.47	1.75%	159.13	10.88	2.01%	221.48
6	10.26	1.78%	184.05	11.59	1.99%	265.98
7	12.79	1.98%	310.21	13.45	2.08%	377.65
8	12.72	1.80%	325.22	13.41	1.91%	391.04
9	11.78	1.61%	263.56	13.37	1.84%	353.37
10	11.91	1.58%	254.91	13.96	1.88%	355.39
11	12.63	1.66%	277.38	15.02	2.00%	378.78
12	13.48	1.76%	307.47	16.10	2.13%	420.51
13	13.46	1.78%	311.90	16.13	2.16%	411.04
14	13.73	1.83%	328.55	16.77	2.26%	446.75
15	14.03	1.90%	339.88	16.99	2.31%	459.94
16	14.32	1.94%	358.71	17.01	2.31%	459.80
17	14.68	1.96%	373.37	17.39	2.32%	480.92
18	14.57	1.90%	398.16	17.66	2.30%	518.73
19	13.87	1.80%	366.83	16.31	2.12%	478.19
20	12.77	1.67%	315.26	15.23	2.01%	439.50
21	12.36	1.64%	282.62	15.18	2.03%	419.49
22	11.70	1.66%	296.36	13.84	1.94%	343.08
23	10.28	1.57%	191.55	13.02	2.00%	313.64
24	10.22	1.71%	179.83	11.71	1.95%	262.89

Table 7-27: Hourly Forecasting Results – Vermont

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	14.25	1.88%	336.61	16.05	2.10%	384.88
February	13.07	1.80%	299.60	14.04	1.90%	311.80
March	10.98	1.62%	191.16	14.09	2.01%	314.90
April	10.68	1.71%	191.60	13.96	2.15%	309.93
May	10.41	1.70%	200.63	10.82	1.70%	215.89
June	9.42	1.43%	155.36	12.09	1.83%	258.11
July	12.68	1.81%	305.08	15.05	2.14%	357.50
August	11.35	1.62%	222.72	15.64	2.38%	559.59
September	10.71	1.67%	198.59	9.61	1.47%	150.34
October	9.01	1.38%	142.48	9.88	1.55%	166.70
November	14.37	2.21%	407.29	14.83	2.33%	433.23
December	19.00	2.55%	625.02	24.20	3.42%	905.01

Load Zone: New Hampshire.

Table 7-28: Overall Forecasting Results – New Hampshire.

Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	27.77	2.10%	1,618.72	33.33	2.57%	2,443.86

Table 7-29: Hourly Forecasting Results – New Hampshire.

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	22.21	2.10%	918.46	29.12	2.77%	1,718.43
2	22.11	2.23%	1,041.88	28.30	2.83%	1,713.67
3	20.68	2.12%	816.86	28.71	2.95%	2,244.27
4	20.37	2.13%	875.99	27.05	2.79%	1,583.03
5	20.56	2.09%	862.38	26.92	2.72%	1,502.11
6	21.73	2.05%	1,003.51	28.27	2.66%	1,650.62
7	28.11	2.36%	1,726.25	35.14	2.94%	2,506.33
8	28.04	2.15%	1,787.94	33.99	2.63%	2,509.42
9	25.86	1.87%	1,458.80	31.75	2.34%	2,408.70
10	26.11	1.81%	1,380.50	31.85	2.25%	2,373.51
11	27.48	1.86%	1,484.40	32.09	2.22%	2,336.46
12	28.34	1.90%	1,594.48	32.26	2.20%	2,508.20
13	29.62	1.99%	1,759.68	34.28	2.33%	2,647.13
14	32.29	2.17%	2,032.58	36.35	2.50%	2,965.93
15	34.30	2.34%	2,258.47	38.06	2.64%	3,176.53
16	35.82	2.45%	2,400.82	39.13	2.72%	3,327.40
17	37.88	2.53%	2,743.70	40.15	2.74%	3,446.33
18	36.77	2.42%	2,727.91	41.85	2.76%	3,599.37
19	35.05	2.30%	2,517.19	39.61	2.59%	3,214.57
20	31.73	2.08%	2,116.63	36.54	2.41%	2,771.98
21	28.62	1.91%	1,743.61	34.54	2.31%	2,470.35
22	26.41	1.86%	1,456.20	32.39	2.30%	2,100.13
23	24.03	1.87%	1,173.81	31.12	2.46%	1,955.47
24	22.36	1.94%	967.24	30.39	2.67%	1,922.77

Table 7-30: Monthly Forecasting Results – New Hampshire

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	23.63	1.64%	1,031.25	26.91	1.86%	1,254.68
February	27.54	2.21%	2,544.78	27.71	2.02%	1,582.21
March	23.11	1.81%	877.89	28.21	2.16%	1,181.03
April	23.29	1.94%	1,305.67	30.94	2.58%	1,707.95
May	22.87	1.90%	973.77	33.81	2.72%	1,663.03
June	25.72	1.98%	1,182.40	33.85	2.53%	2,009.62
July	38.80	2.67%	2,780.07	47.71	3.19%	3,522.14
August	36.20	2.49%	2,172.84	39.56	2.97%	4,550.28
September	29.38	2.29%	1,614.00	24.40	1.84%	974.28
October	19.86	1.63%	761.24	40.16	3.87%	6,165.56
November	25.83	2.09%	1,593.65	29.06	2.33%	2,089.26
December	36.29	2.56%	2,585.23	36.57	2.70%	2,446.35

Load Zone: Maine.

Table 7-31: Overall Forecasting Results – Maine.

.Overall	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
	26.91	2.04%	1322.81727	32.11	2.49%	1973.636392

Table 7-32: Hourly Forecasting Results – Maine.

Hour	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
1	20.71	1.94%	725.32	24.58	2.33%	1,194.86
2	20.08	1.95%	667.85	24.99	2.45%	1,164.18
3	20.29	2.01%	716.78	24.98	2.49%	1,155.42
4	20.33	2.01%	730.48	25.09	2.51%	1,110.39
5	20.25	1.95%	718.38	24.90	2.43%	1,178.77
6	22.23	1.99%	887.70	28.05	2.55%	1,463.00
7	27.16	2.19%	1,374.95	34.38	2.80%	2,276.76
8	26.34	1.96%	1,367.37	34.22	2.59%	2,288.56
9	26.06	1.85%	1,234.89	34.57	2.51%	2,164.57
10	26.14	1.80%	1,258.03	35.16	2.48%	2,124.95
11	28.14	1.91%	1,376.18	36.11	2.52%	2,192.61
12	29.84	2.03%	1,536.72	35.76	2.50%	2,156.88
13	31.30	2.15%	1,673.48	36.15	2.55%	2,134.62
14	31.77	2.21%	1,697.90	36.51	2.58%	2,192.89
15	33.57	2.36%	1,901.35	35.08	2.52%	2,105.00
16	32.99	2.32%	1,924.79	36.15	2.61%	2,359.33
17	35.50	2.44%	2,162.31	38.03	2.69%	2,745.49
18	34.80	2.35%	2,156.78	38.71	2.67%	3,012.02
19	33.09	2.24%	1,918.36	36.56	2.51%	2,778.52
20	29.41	1.99%	1,574.35	35.13	2.42%	2,625.28
21	26.28	1.81%	1,287.44	32.48	2.29%	2,391.86
22	25.06	1.83%	1,135.41	29.23	2.19%	1,745.33
23	23.08	1.85%	923.98	27.77	2.28%	1,476.00
24	21.50	1.89%	796.82	26.01	2.32%	1,329.97

Table 7-33: Hourly Forecasting Results – Maine.

Month	Within Sample			Out of Sample		
	MAE	MAPE	MSD	MAE	MAPE	MSD
January	23.52	1.67%	945.98	26.69	1.93%	1,102.11
February	25.13	1.78%	8,310.98	24.95	1.85%	1,015.74
March	24.87	1.69%	5,888.73	26.54	2.08%	1,137.43
April	23.04	2.13%	12,012.08	40.50	3.20%	2,679.09
May	26.59	2.02%	10,126.20	25.99	2.11%	1,185.74
June	23.29	2.35%	14,642.93	28.86	2.28%	1,264.66
July	30.45	2.84%	26,743.67	36.25	2.58%	2,033.43
August	30.80	2.64%	21,467.51	42.20	3.31%	5,044.98
September	30.15	2.36%	13,370.89	33.27	2.63%	1,784.49
October	23.24	1.80%	6,026.42	33.55	2.71%	1,878.41
November	29.88	1.86%	9,415.43	34.34	2.79%	2,483.04
December	31.41	2.14%	14,474.08	31.74	2.41%	1,992.03

C. SAS CODE AND ALGORITHMS

The following macro was used to calculate the default number of knots for each weather variable.

```
%macro default_knots(librefknots=,data=,knotdata=,varknots=,numknots=);

proc sort data=&data (keep=&varknots) out=q1;
      by &varknots;
run;
/*finds unique values*/
data q2;
      set q1;
      by &varknots;
      if first.&varknots;
run;

data &librefknots..&knotdata;
      set q2 nobs=n;
      knotsp=int(n/5);
      if knotsp>=40 then kmx=40; else
      if knotsp<40 then kmx=knotsp;
            %if &numknots ne %then %do
            ktemp=&numknots;
            if 1 <= ktemp <= 40 then kmx=ktemp;
            %end;

      kintrvl=round(n/kmx);
      knotsok=mod(_n_,kintrvl);
      knots=&varknots;
      if knotsok=0 or _n_=n-1 then output;


run;

%mend;
```

The following macro was used to create an analysis data set.

```
%macro make_data_set(data, out);

      data &out;
            set &data;

            time = _n_;

            lag24  = lag24(load);
            lag48  = lag48(load);
            lag72  = lag72(load);
            lag96  = lag96(load);
            lag120 = lag120(load);
            lag144 = lag144(load);
            lag168 = lag168(load);
```

```
templag = lag(temp);
tempsqr = temp**2;
tempcub = temp**3;

humlag = lag(hum);
humsqr = hum**2;
humcub = hum**3;

wslag = lag(ws);
wssqr = ws**2;
wscub = ws**3;

cclag = lag(cc);
ccsqr = cc**2;
xxcub = cc**3;

labor09 = holiday('labor',2009);
      format labor09 date9.;
labor10 = holiday('labor',2010);
      format labor10 date9.;
labor11 = holiday('labor',2011);
      format labor11 date9.;
labor12 = holiday('labor',2012);
      format labor12 date9.;

july409 = holiday("usindependence",2009);
      format july409 date9.;
july410 = holiday("usindependence",2010);
      format july410 date9.;
july411 = holiday("usindependence",2011);
      format july411 date9.;
july412 = holiday("usindependence",2012);
      format july412 date9.;

newyears09 = holiday("newyear",2009);
      format newyears09 date9.;
newyears10 = holiday("newyear",2010);
      format newyears10 date9.;
newyears11 = holiday("newyear",2011);
      format newyears11 date9.;
newyears12 = holiday("newyear",2012);
      format newyears12 date9.;

mlk09 = holiday("mlk",2009);
      format mlk09 date9.;
mlk10 = holiday("mlk",2010);
      format mlk10 date9.;
mlk11 = holiday("mlk",2011);
      format mlk11 date9.;
mlk12 = holiday("mlk",2012);
      format mlk12 date9.;

presidents09 = holiday("uspresidents",2009);
      format presidents09 date9.;
presidents10 = holiday("uspresidents",2010);
      format presidents10 date9.;
```

```
presidents11 = holiday("uspresidents",2011);
      format presidents11 date9.;
presidents12 = holiday("uspresidents",2012);
      format presidents12 date9.;


memorial09 = holiday("memorial",2009);
      format memorial09 date9.;
memorial10 = holiday("memorial",2010);
      format memorial10 date9.;
memorial11 = holiday("memorial",2011);
      format memorial11 date9.;
memorial12 = holiday("memorial",2012);
      format memorial12 date9.;

thanksgiving09 = holiday("thanksgiving",2009);
      format thanksgiving09 date9.;
thanksgiving10 = holiday("thanksgiving",2010);
      format thanksgiving10 date9.;
thanksgiving11 = holiday("thanksgiving",2011);
      format thanksgiving11 date9.;
thanksgiving12 = holiday("thanksgiving",2012);
      format thanksgiving12 date9.;

xmas09 = holiday("christmas",2009);
      format xmas09 date9.;
xmas10 = holiday("christmas",2010);
      format xmas10 date9.;
xmas11 = holiday("christmas",2011);
      format xmas11 date9.;
xmas12 = holiday("christmas",2012);
      format xmas12 date9.;

if weekday(date) in (1 7) then nonworking = 1;

else if date = labor09 then nonworking =1;
else if date = labor10 then nonworking =1;
else if date = labor11 then nonworking =1;
else if date = labor12 then nonworking =1;

else if date = july409 then nonworking =1;
else if date = july410 then nonworking =1;
else if date = july411 then nonworking =1;
else if date = july412 then nonworking =1;

else if date = newyears09 then nonworking =1;
else if date = newyears10 then nonworking =1;
else if date = newyears11 then nonworking =1;
else if date = newyears12 then nonworking =1;

else if date = mlk09 then nonworking =1;
else if date = mlk10 then nonworking =1;
else if date = mlk11 then nonworking =1;
else if date = mlk12 then nonworking =1;

else if date = presidents09 then nonworking  =1;
else if date = presidents10 then nonworking  =1;
```

```
            else if date = presidents11  then nonworking =1;
            else if date = presidents12  then nonworking =1;

            else if date = memorial09 then nonworking =1;
            else if date = memorial10 then nonworking =1;
            else if date = memorial11 then nonworking =1;
            else if date = memorial12 then nonworking =1;

            else if date = thanksgiving09 then nonworking =1;
            else if date = thanksgiving10 then nonworking =1;
            else if date = thanksgiving11 then nonworking =1;
            else if date = thanksgiving12 then nonworking =1;

            else if date = xmas09 then nonworking =1;
            else if date = xmas10 then nonworking =1;
            else if date = xmas11 then nonworking =1;
            else if date = xmas12 then nonworking =1;

            else nonworking = 0;

            if month(date) = 1  then month1 = 1; else month1   =0;
            if month(date) = 2  then month2 = 1; else month2   =0;
            if month(date) = 3  then month3 = 1; else month3   =0;
            if month(date) = 4  then month4 = 1; else month4   =0;
            if month(date) = 5  then month5 = 1; else month5   =0;
            if month(date) = 6  then month6 = 1; else month6   =0;
            if month(date) = 7  then month7 = 1; else month7   =0;
            if month(date) = 8  then month8 = 1; else month8   =0;
            if month(date) = 9  then month9 = 1; else month9   =0;
            if month(date) = 10 then month10 = 1; else month10 =0;
            if month(date) = 11 then month11 = 1; else month11 =0;

            weekday = weekday(date);

            if weekday(date) = 1 then monday    = 1; else monday    = 0;
            if weekday(date) = 2 then tuesday   = 1; else tuesday   = 0;
            if weekday(date) = 3 then wednesday = 1; else wednesday = 0;
            if weekday(date) = 4 then thursday  = 1; else thursday  = 0;
            if weekday(date) = 5 then friday    = 1; else friday    = 0;

      run;

%mend;
```

SAS program to prepare data sets and also to fit models.

```
/*
================================================================================
SAS PROGRAM : FIT FORECASTING MODELS

MASTERS THESIS - APPLIED ECONOMETRICS

TITLE       : ZONAL AND REGIONAL LOAD FORECASTING IN THE NEW ENGLAND
WHOLSEALE ELECTRICITY MARKET
SUBTITLE    : A SEMIPARAMETRIC REGRESSION APPROACH
WRITTEN BY  : JONATHAN T. FARLAND
DATE        : AUGUST 15th, 2013

================================================================================
*/

/*clear log and output windows*/
dm "log;    clear";
dm "lst;    clear";
dm "output; clear";

/*directory*/
%let dir = FOO;

/*libraries*/
libname macros "%superq(dir)\SAS\Macros";
libname rawdat "%superq(dir)\Data";
libname sasdat "%superq(dir)\Data\sasdat";
libname output "%superq(dir)\Output";

options symbolgen;
options spool;

/*seperate program containing macros used below*/
%include "&dir\SAS\Macros\load_forecasting_macros.sas" / source2;

/*specify the load zone as a macro variable*/
%let lz = region;

/*
LOAD ZONE                     VALUE
Entire Region               = region
North Eastern Massachusetts = nemass
South Eastern Massachusetts = semass
Western Massachusetts       = wcmass
Connecticut                 = ct
Rhode Island                = ri
Vermont                     = vt
New Hampshire               = nh
Maine                       = me
*/

/* Log and Output files */
%let log_path = &dir\Results\logs\&lz;
```

```
/*clear directory*/
proc datasets lib=work
      nolist kill;
quit; run;


/*initialize data set*/
data original;
      set sasdat.&lz;
run;

/*use a subroutine to make necessary variables for model fitting*/
%make_data_set(original, ds);

/*
============================================================================
Determining Knots for the weather variables
============================================================================
*/

/*the macro "default_knots" calculates the default number of knots
recommended by Ruppert et al (See "Smoothing with Mixed Model Software" with
Long Ngo and M.P.Wand)*/
%default_knots(librefknots=work,data=work.ds,knotdata=knots_temp,varknots=tem
p);
%default_knots(librefknots=work,data=work.ds,knotdata=knots_hum,varknots=hum)
;
%default_knots(librefknots=work,data=work.ds,knotdata=knots_cc,varknots=cc);
%default_knots(librefknots=work,data=work.ds,knotdata=knots_ws,varknots=ws);


/*create a generic constant to merge later on. specifically, 'm'*/
data ds2;
      set ds;
      m=1;
run;


data kt_temp;
      set work.knots_temp nobs=nk_temp;
      call symput('nkt_temp',nk_temp);
run;
proc transpose data=work.knots_temp prefix=knots_temp_ out=knotst_temp;
      var knots;
run;

data kt_hum;
      set work.knots_hum nobs=nk_hum;
      call symput('nkt_hum',nk_hum);
run;
proc transpose data=work.knots_hum prefix=knots_hum_ out=knotst_hum;
      var knots;
run;
```

```
data kt_cc;
      set work.knots_cc nobs=nk_cc;
      call symput('nkt_cc',nk_cc);
run;
proc transpose data=work.knots_cc prefix=knots_cc_ out=knotst_cc;
      var knots;
run;

data kt_ws;
      set work.knots_ws nobs=nk_ws;
      call symput('nkt_ws',nk_ws);
run;
proc transpose data=work.knots_ws prefix=knots_ws_ out=knotst_ws;
      var knots;
run;

/*merge all 'knot' data sets together */
data knotst;
      merge knotst_temp knotst_hum knotst_cc knotst_ws;
      m=1;
run;

/*
============================================================================
Creating the Z matrix
============================================================================
*/

data ds3;
      merge ds2 knotst;
      by m;

%let nk1=&nkt_temp;
%let nk2=&nkt_hum;
%let nk3=&nkt_cc;
%let nk4=&nkt_ws;

/*create truncated power functions of degree p=1 */

array Z1a (&nk1) Z1_1-Z1_&nk1;
array knots1a (&nk1) knots_temp_1-knots_temp_&nk1;
      do k=1 to &nk1;
            Z1a(k)=temp-knots1a(k);
            if Z1a(k) < 0 then Z1a(k)=0;
      end;

array Z2a (&nk2) Z2_1-Z2_&nk2;
array knots2a (&nk2) knots_hum_1-knots_hum_&nk2;
      do k=1 to &nk2;
            Z2a(k)=hum-knots2a(k);
            if Z2a(k) < 0 then Z2a(k)=0;
      end;
array Z3a (&nk3) Z3_1-Z3_&nk3;
array knots3a (&nk3) knots_cc_1-knots_cc_&nk3;
      do k=1 to &nk3;
            Z3a(k)=cc-knots3a(k);
            if Z3a(k) < 0 then Z3a(k)=0;
```

```
      end;
array Z4a (&nk4) Z4_1-Z4_&nk4;
array knots4a (&nk4) knots_ws_1-knots_ws_&nk4;
      do k=1 to &nk4;
            Z4a(k)=ws-knots4a(k);
            Z4a(k) = Z4a(k);
            if Z4a(k) < 0 then Z4a(k)=0;
      end;

drop knots1_1-knots1_&nk1 knots2_1-knots2_&nk2
     knots3_1-knots3_&nk3 knots4_1-knots4_&nk4 _name_;
run;

/*
================================================================================
Make Training Dataset
================================================================================
*/

/*select training period*/
%let training_beg = '01JAN09 00:00:00'dt;
%let training_end = '31DEC10 23:00:00'dt;

/*isolate training data set*/
data trn;
      set ds3;
      if datetime ge &training_beg and datetime le &training_end;
      training = 1;
run;

/*create two forecast data sets:
      (1) without load to make predictions with proc mixed and
      (2) with    load to calculate forecasting errors from predictions
*/

data fcst;
      set ds3;
      if datetime > &training_end;
      drop load; /*without dependant variable*/
run;

data fcst2;
      set ds3;
      if datetime > &training_end;
run;

/*stack training and forecasted*/

data ds4;
      set trn fcst;
      /*generate forecasting flag*/
      if training ne 1 then forecast = 1;
      else                  forecast = 0;

/*generate correct hour variable that ranges from 1 - 24 (as opposed to the
way the data came to us, e.g., 0-23)*/
      hour2 = hour+1;
```

```
      drop hour;
run;

/*
=============================================================================
Fit hourly models
=============================================================================
*/

ods listing;
ods html;

proc sort
      data = ds4;
      by hour2;
run;

ods output CovParms=work.varcomp FitStatistics=work.FitStatistics
LRT=work.RatioTest SolutionF=work.FParms SolutionR=work.RParms
Tests3=work.FixedTests Type1=work.ANOVA;
ods graphics on;
proc mixed data = ds4 noprofile method=REML
plots(maxpoints=50000)=residualpanel;
by hour2; /*estimate hourly models*/
model load = tuesday--friday month1-month11 nonworking lag24--lag168 temp hum
cc ws /  solution outp=work.yhat;
      random Z1_1-Z1_&nk1 / type=toep(1) s;
      random Z2_1-Z2_&nk2 / type=toep(1) s;
      random Z3_1-Z3_&nk3 / type=toep(1) s;
      random Z4_1-Z4_&nk4 / type=toep(1) s;
run;
ods graphics off;

/*stack fixed and random solutions*/
data parms;
      set Fparms
            Rparms;
      keep effect estimate;
run;
```