

---

Doctoral Dissertations

Student Theses and Dissertations

---

Spring 2016

## Modeling daily electricity load curve using cubic splines and functional principal components

Abdelmonaem Salem Jornaz

Follow this and additional works at: [https://scholarsmine.mst.edu/doctoral\\_dissertations](https://scholarsmine.mst.edu/doctoral_dissertations)



Part of the [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Department: **Mathematics and Statistics**

---

### Recommended Citation

Jornaz, Abdelmonaem Salem, "Modeling daily electricity load curve using cubic splines and functional principal components" (2016). *Doctoral Dissertations*. 2476.  
[https://scholarsmine.mst.edu/doctoral\\_dissertations/2476](https://scholarsmine.mst.edu/doctoral_dissertations/2476)

This thesis is brought to you by Scholars' Mine, a service of the Missouri S&T Library and Learning Resources. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact [scholarsmine@mst.edu](mailto:scholarsmine@mst.edu).

MODELING DAILY ELECTRICITY LOAD CURVE USING CUBIC SPLINES AND  
FUNCTIONAL PRINCIPAL COMPONENTS

by

ABDELMONAEM SALEM JORNAZ

A DISSERTATION

Presented to the Faculty of the Graduate School of the  
MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

2016

Approved

V.A. Samaranayake, Advisor

Robert Paige

Gayla Olbricht

Xuerong Wen

Gregory Gelles

© 2016

Abdelomnaem Salem Jornaz

All Rights Reserved

## ABSTRACT

Forecasting electricity load is very important to the electric utilities as well as producers of power because accurate predictions can cut down costs by avoiding power shortages or surpluses. Of specific interest is the 24-hour daily electricity load profile, which provides insight into periods of high demand and periods where the use of electricity is at a minimum. Researchers have proposed many approaches to modeling electricity prices, real-time load, and day-ahead demand, with varying success. In this dissertation three new approaches to modeling and forecasting the 24-hour daily electricity load profiles are presented. The application of the proposed methods is illustrated using hourly electricity load data from the Atlantic City Electric (AE) zone, which is part of the Pennsylvania, New Jersey, and Maryland (PJM) electricity market. The first approach that is proposed can be used to make short-term forecasts of electricity load. This approach employs a hybrid technique utilizing autoregressive moving average method (ARMA) and cubic spline models. The second approach is suitable for obtaining long-term forecasts of the daily electricity load and employs cubic splines with time varying coefficients. These coefficients are modeled as a multivariate time series using a vector autoregressive model with exogenous variables to forecast the average daily electricity load profile for a future month. The last approach uses functional principal components to model the daily electricity load profile for each day as a linear combination of three eigenfunctions, with the coefficients of the day-specific linear combinations modeled as univariate time series using transfer functions. The fitted models from the three approaches were applied to data from a subsequent year and the results show that these models perform quite well.

## ACKNOWLEDGMENTS

First of all, I would like to begin by thanking Allah (God), the almighty for providing me everything to finish this work and granting me capability to proceed successfully. Completion of my doctoral dissertation was due to support, encouragement, guidance and assistance of several people.

I sincerely and deeply thank Dr. V. A. Samaranayake for his patience, support, understanding, encouragement, and thoughtful guidance during my PhD study. I greatly appreciate Dr. Samaranayake for spending much of his time, especially weekends and evenings to help me to accomplish this research. I would also like to thank the members of my committee, Dr. Robert Paige, Dr. Gayla Ulbricht, Dr. Xuerong Wen, and Dr. Gregory Gelles.

My sincere thanks goes to the former chair of the Department of Mathematics and Statistics, Dr. Leon Hall, the current chair Dr. Stephen Clark, professors, staff and students, who provided a very enjoyable study environment.

I know that I could not have completed this work without the patience, understanding and sacrifices of my wife. Her support and encouragement was, in the end, what made this dissertation possible. I also would like to express gratitude to my brothers and sisters for their support and encouragement throughout my entire life. I also would like to thank my friends everywhere in this world.

Finally, I dedicate this dissertation to the memory of my parents, who are a constant source of inspiration. I hope they would have been proud. All the support they provided me over the years was the greatest gift of my life.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
ACKNOWLEDGMENTS .....	iv
LIST OF ILLUSTRATIONS .....	vii
LIST OF TABLES .....	x
SECTION	
1. INTRODUCTION .....	1
2. FACTORS AFFECTING ELECTRICITY LODA AND DATA .....	9
2.1. THE FACTORS EFFECTING ELECTRICITY LOAD .....	9
2.1.1. Economic Factors .....	9
2.1.2. Time Related Factors.....	10
2.1.3. Weather Related Factors .....	12
2.1.4. The Electricity Load Curve over Different Regions .....	14
2.2. DATA SOURCES .....	15
3. METHODOLOGY AND RELATED STATISTICAL TOOLS .....	19
3.1 SPLINE MODELS.....	19
3.1.1. Cubic Splines.....	21
3.1.2. Smoothing Splines.....	23
3.2 VECTOR AUTOREGRESSIVE MODEL WITH EXOGENOUS VARIABLES (VARX).....	24
3.2.1. Prediction Intervals for VARX Forecasts .....	25
3.3. TRANSFER FUNCTION MODELS .....	26
3.4. FUNCTIONAL DATA ANALYSIS (FDA) .....	27
3.4.1. Functional Principal Component Analysis (FPCA) .....	27
3.4.1.1 Prediction from FPCA .....	29
3.4.1.2 Prediction intervals from FPCA.....	29
3.5. A BRIEF DESCRIPTION OF THE THREE MODELING APPROACHES ..	30
3.5.1. Short-Term Approach.....	30
3.5.2. Long-Term Approach.....	31
3.5.3. The FPCA Approach .....	32

4. ANALYSIS AND EMPIRICAL RESULTS .....	33
4.1. SHORT TERM APPROACH.....	33
4.1.1. Predicting Long-Term Trend.....	34
4.1.2. Estimating Seasonal Variation in Data.....	37
4.1.3. Modeling the Hourly Load .....	41
4.1.3.1 The results of model 1.....	41
4.1.3.2 The results of model 2.....	46
4.2. LONG TERM APPROACH.....	55
4.2.1. Weekdays and Weekends Model .....	56
4.2.2. Combined Model .....	56
4.2.3. Modeling Hourly Temperature.....	57
4.2.4. Vector Autoregressive Model with Exogenous Variable (VARX).....	60
4.2.5. Simultaneous Prediction Intervals for VARX mode .....	62
4.3. FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS APPROACH .....	65
4.3.1. Removing the Trend.....	65
4.3.2. Creating a Basis and Smoothing the Data .....	66
4.3.3. Computing the Functional Eigenfunctions (Harmonics) and the Principal Component Scores.....	67
4.3.4. Modeling the Principal Components .....	70
4.3.5. The Prediction Interval for the Hourly Load Forecast .....	72
5. MODEL COMPARISONS AND CONCLUSIONs.....	75
5.1. THE COMPARISON OF THE LONG-TERM ELECTRICITY LOAD FORECASTS .....	75
5.1.1. The Comparison of the Long-Term Electricity Load per Month.....	75
5.1.2. The Comparison of the Long-Term Electricity Load per Hour .....	77
5.2. THE COMPARISON OF THE SHORT-TERM ELECTRICITY LOAD FORECASTS .....	79
5.2.1. The Comparison of the Short-Term Electricity Load per Month.....	80
5.2.2. The Comparison of the Short -Term Electricity Load per Hour .....	81
APPENDIX.....	85
BIBLIOGRAPHY.....	102
VITA .....	106

## LIST OF ILLUSTRATIONS

Figure	Page
1.1. Electric Power Markets in the US.....	2
2.1. Retail Sales of Electricity to Ultimate Customers by end-use Sector for New Jersey, March 2013.....	10
2.2. The Average of a 24-hour of Jan. Load Curve of Weekdays and Weekends 1993-2012 .....	11
2.3. The Average of a 24-hour of Jul. Load Curve of Weekdays and Weekends 1993-2012 .....	11
2.4. The Relationship between the Hourly Load and Hourly Temperature – South New Jersey 1993-2012 .....	12
2.5. The Average of 24-hour Load Curves and Temperature Jul. 2013 .....	13
2.6. The Average of 24-hour Load Curves and Temperature Jan. 2013.....	13
2.7. The Average of 24-hour Load Curves from Mar. 2013 for AE, PN, and the South of ERCOT .....	14
2.8. Map of PJM Regional Transmission Organization.....	16
2.9. Map Service of Atlantic City Electric (AE) – Study Area.....	17
2.10. The Hourly Load of AE over 20 Years 1993 – 2012.....	17
2.11. The Hourly Load of AE in 2013 .....	18
4.1. Residual Analysis of the Annual Regression Model .....	36
4.2. The Annual Average of Hourly Load and the Predicted Load 1993 - 2013.....	37
4.3. Analysis of Residuals for the Weekly ARMAX Model .....	39
4.4. Normality Diagnostics of the Residuals of the Weekly ARMAX Model .....	40
4.5. The Weekly Average of the Hourly Load and the Predicted Load - 2013 .....	40
4.6. The Comparison between the CV of the Predicted Monthly Average Load of Model 1 and Model 2 .....	50
4.7. The Comparison between the CV of the Predicted Hourly Average Load of Model 1 and Model 2 .....	51
4.8. The Comparison between the Actual Hourly Load, the Predicted Hourly Load of Model 1 and Model 2 – a Week in the Mid-Winter Season.....	52
4.9. The Comparison between the Actual Hourly Load, the Predicted Hourly Load of Model 1 and Model 2 – a Week in the Mid- Spring Season .....	52



4.10. The Comparison between the Actual Hourly Load, the Predicted Hourly Load of Model 1 and Model 2 – a Week in the Mid-Summer Season .....	53
4.11. The Comparison between the Actual Hourly Load, the Predicted Hourly Load of Model 1 and Model 2 – a Week in the Mid-Fall Season .....	53
4.12. Residual of the Weekly ARMA Model .....	58
4.13. Normality Diagnostics of the Residuals for the Weekly ARMA Model .....	59
4.14. The Comparison between the Actual Monthly Temperature, the Predicted Monthly Temperature - 2013 .....	59
4.15. The Comparison between the Actual Monthly Weekdays Load, the Predicted Monthly Load from the Weekdays Model with Nowcasting Temperature, and Forecasting Temperature - 2013 .....	61
4.16. The Comparison between the Actual Monthly Weekends Load, the Predicted Monthly Load of Weekdays Model with Nowcasting Temperature, and Forecasting Temperature - 2013 .....	61
4.17. The Comparison between the Actual Monthly Load, the Predicted Monthly Load from the Weekdays Model with Nowcasting Temperature, and Forecasting Temperature - 2013 .....	62
4.18. The Prediction Interval for the Weekdays Model – 2013 .....	63
4.19. The Prediction Interval for the Weekends Model – 2013 .....	64
4.20. The Prediction Interval for the Combined Model – 2013 .....	64
4.21. The Daily 24-hour Profile over 20 Years .....	66
4.22. The Comparison between the RMSE of the 8 Knots and the 12 Knots Model .....	67
4.23. The Mean Curve of the 24-hour Profile .....	68
4.24. The First Eigenfunction (Harmonic) Curve .....	69
4.25. The Second Eigenfunction (Harmonic) Curve .....	69
4.26. The Third Eigenfunction (Harmonic) Curve .....	70
4.27. The Comparison between the Actual Monthly Load and the Predicted Monthly Load - 2013 .....	72
4.28. The Prediction Interval of the FPCA - a Week in the Mid-Winter Season .....	73
4.29. The Prediction Interval from FPCA - a Week in the Mid-Spring Season .....	73
4.30. The Prediction Interval from FPCA - a Week in the Mid-Summer Season .....	74
4.31. The Prediction Interval from FPCA - a Week in the Mid-Fall Season .....	74
5.1. The CV of the Long-Term Load from the Three Models per Month Model 1, Model 2, and Model 3 .....	77

5.2. The <i>CV</i> of the Long-Term Load for the Three Models per Hour Model 1, Model 2, and Model 3 .....	79
5.3. The <i>CV</i> of the Short-Term Load for the Three Models per Month Model 1, Model 2, and Model 3 .....	81
5.4. The <i>CV</i> of the Short -Term Load for the Three Models per Hour Model 1, Model 2, and Model 3 .....	83

## LIST OF TABLES

Table	Page
4.1. The Results for the Regression Model for Annual Load - ANOVA Table .....	34
4.2. The Results for the Weekly ARMAX Model - Parameter Estimates .....	38
4.3. The Results for the Regression Model for the Winter Season.....	42
4.4. The Results for the Regression Model for the Spring Season .....	43
4.5. The Results for the Regression Model for the Summer Season .....	44
4.6. The Results for the Regression Model for the Fall Season.....	45
4.7. The Results for the Regression Model for the Winter and Spring Seasons.....	47
4.8. The Results for the Regression Model for the Summer and Fall Seasons.....	48
4.9. The Comparison between the Two Models .....	49
4.10. The Comparison between the Two Models for Each Season per Hour .....	54
4.11. The Results for the Monthly Temperature ARMA Model-Parameter Estimates ....	57
4.12. The Information Criteria Results of each VARX Model.....	60
4.13. The Results for the Daily PC's Scores Transfer Function Model - Fit Statistics....	71
5.1. The CV for the Long-Term Load of the Three Models per Month.....	76
5.2. The CV for the Long-Term Load of the Three Models per Hour .....	78
5.3. The CV for the Short-Term Load of the Three Models per Month .....	80
5.4. The CV for the Short -Term Load of the Three Models per Hour .....	82

## 1. INTRODUCTION

The electricity load over a given time interval is defined the actual amount of electricity consumed, in megawatts, over that period in a specific geographic area. The geographic area is usually defined as the zone (region) covered by an Independent Transmission Systems Operator (ISO) or a Regional Transmission Systems Operator (RTO). The ISO's and RTO's are, in general, consortiums of electric utilities and producers, who buy and sell electricity for use in the region covered by these entities. In short, such consortiums act as an "electricity market" and the price of electricity is governed by production costs as well as the demand. The electricity load is sometimes referred to as the real-time load as opposed to electricity demand, which is the total amount of electricity that utilities in the market make bids for, hours or days ahead of the actual usage. Electricity load is the actual amount of electricity that is consumed and is driven by several factors such as consumer behavior as well as economic conditions and weather particular to that region at the time of use. Moreover, load varies across different hours of the day and different days of the week, thus making the electricity load a time varying quantity. For example, the hourly electricity load observed over several years would form a time series with segments of 24-hour daily profiles (also known as load curves) each of which is non-constant across the hours of the day. In addition, these daily profiles would change based on the day of the week. Additional seasonal patterns, that accommodate the evolution of winter daily profiles to those during summer, add more complexity to the time series. Effects of weather conditions, in particular temperature, complicate matters further. Thus, finding reliable models to forecast electricity load can be a challenging task. Such models are of practical importance to those in the public and private utility sectors as well as to traders in the electricity market. Currently there are nine electricity markets in the United States. They are: The California Independent System Operator – CAISO, The Midcontinent Independent System Operator – MISO, The Independent System Operator of New England – ISO-NE, The New York Independent System Operator – NYISO, The Northwest Electricity Market, The Pennsylvania-New Jersey-Maryland - PJM, The Southeast Electricity Market, The

Southwest Electricity Market, The Southwest Power Pool – SPP, and The Electric Reliability Council of Texas – ERCOT. Utilities and generators of electricity trade this commodity based on forecast of the load. Therefore electric load modeling and forecasting has caught the interest of many researchers.

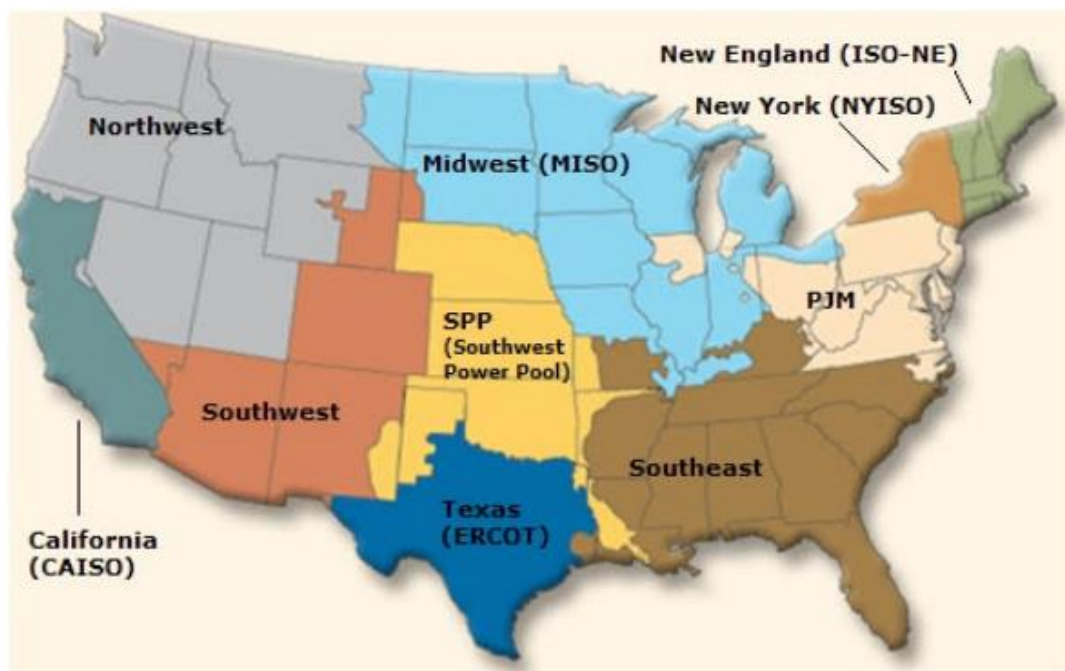


Figure. 1.1 Electric Power Markets in the US<sup>1</sup>

There is a long history of research work aimed at developing hourly electricity load models. Many of the TRO's and ISO's as well as utility companies have tended to use multiple regression models with many weather related inputs for short-term predictions, but research has recently progressed to include more sophisticated approaches. For early classical work the reader is referred to Bunn and Farmer (1985) which summarized approaches that were used to make short-term forecasts of the load. An important reference that classifies different methods of load forecasting is Alfares and Nazeeruddin (2002). The authors classified the various approaches into nine classes

<sup>1</sup> Federal Energy Regulatory Commission (FERC) website. <http://www.ferc.gov/market-oversight/mkt-electric/overview.asp>

which are: (1) multiple regression, (2) exponential smoothing, (3) iterative reweighted least-squares, (4) adaptive load forecasting, (5) stochastic time series, (6) Autoregressive Moving Average models with exogenous inputs (ARMAX models) with the optimal model selection using the genetic algorithm, (7) fuzzy logic, (8) neural networks, and (9) expert systems. Alfares and Nazeeruddin also commented that while the pure time series approach is widely used, hybrid approaches, which combine several techniques, have become more common. For example, El-Keib et al. (1995) presented a hybrid approach where exponential smoothing was augmented with power spectrum analysis and adaptive autoregressive modeling. On the other hand, Dash, Liew, and Rahman (1995) utilized an expert system modeled fuzzy neural network and a hybrid neural network to forecast electricity load. Other publications that employed hybrid approaches are: Kim, Park, Hwang, and Kim (1995) Chow and Leung (1996), and Choueiki, Mount-Campbell, and Ahalt (1997). Alkhatami (2015) also discussed various forecasting methodologies for load forecasting. He also mentioned that the complex methods give more accurate results. The approaches presented in this dissertation can also be considered as the amalgamation of two or more methods as will be seen later.

In this introduction, we will limit our discussion to more recent publications that have some connection to the approaches that will be developed later. Research work that influenced our approach to electricity load modeling is the publication by Nowicka-Zagrajek and Weron (2002), which proposed a two-steps procedure based on removing the trend and seasonal effects first and then fitting an autoregressive moving average (ARMA) model to the deseasonalized data to obtain day-ahead predictions. In contrast, Liu, Chen, and Harris (2006) developed a semi-parametric model for nonlinear time series data, with the model consisting of two components, namely a nonparametric component and parametric Autoregressive Integrated Moving Average (ARIMA) component, to forecast hourly electricity load. A generalization of the logistic Smooth Transition Autoregressive (STAR) model for short term forecasting was developed by Amaral, Souza, and Stevenson (2008), which is a combination of periodic models with a smooth transition between the regimes. Dordonnat, Koopman, Ooms, Dessertaine, and Collet (2008) present periodic state space model with different equations and different parameters for each hour to for the forecasting of hourly electricity load. Four methods of forecasting were compared by

Kosiorowski (2014), which concluded that the moving functional median is the appropriate approach for functional time series that contain outliers and nonstationary functional time series. In comparison, the other three approaches, functional autoregressive, fully functional regression, and the method proposed by Hyndman & Shang (2011), work for stationarity functional time series, and the prediction of the Hyndman & Shang method was the best overall. Annamareddi, Gopinathan, and Dora (2013) proposed a hybrid model based on a wavelet transform technique and double exponential smoothing to forecast the electricity load. Another hybrid method for predicting the electricity load using Support Vector Regression (SVR) and Krill-Herd (KH) algorithm was proposed by Baziar and Kavousi-Fard (2015). The first step used training data, and the KH algorithm was used to optimize the SVR parameters. Consequently, in the second step, the optimized SVR was used to forecast the electricity load.

Two key publications in load forecasting that motivated the second modeling approach to be presented in this dissertation is by Harvey and Koopman (1993), which proposed time-varying splines to model intra-weekly load, and Cho, Goude, Brossat, and Yao (2013) which proposed a hybrid approach using generalized additive model and curve linear regression to model weekly and daily electricity load. In this paper, the idea of time varying spline coefficients proposed by Harvey and Koopman was adopted, but with several important differences from their proposed approach. These differences are discussed in Section 3.5. The paper by Cho, Goude, Brossat, and Yao. (2013) used a generalized additive model to remove the trend and seasonal components from weekly data and then treated the residuals as a set of daily curves that are dependent on the previous days load curve or both the previous day's load curve and the current day's temperature curve. The authors then used a methodology related to functional canonical correlation analysis to reduce the modeling task to a univariate regression problem. The third approach proposed in this dissertation has some similarities with this model, but is relatively simpler to implement and uses functional principal component analysis to reduce the dimension of the problem. In addition, the method proposed in Cho *et al.* (2013) requires refitting of the model to obtain the load curve for a specific day of a given season. As the authors state, in order to forecast the load curve, say for "Wednesday, 2 April 2009," one needs "all pairs of load curves on Tuesdays and

Wednesdays in April” in the training data set. The third approach proposed herein does not require such refitting, but is somewhat less flexible in modeling the seasonal variation in the shape of the load curve when compared to the methodology presented in Cho *et al.* (2013). It is however, a relatively simpler alternative that works well for short-term and long-term predictions, whereas the method given in Cho *et al.* (2013) is recommended only for making short-term (one-day-ahead) forecasts.

One of the new techniques that has been used in the last a few years is the functional principal component analysis (FPCA), which was adopted for the third approach presented in this dissertation. There are a few studies that use the FPCA to forecast the electricity load or demand. Some examples are Hyndman and Shang (2009), Shang (2013), Kosiorowski (2014), and Cabrera and Schulz (2014). Both Shang (2013) and Cabrera and Schulz (2014) worked on forecasting electricity demand and Shang’s work is closest to what is presented in this dissertation and the differences between what is adopted in this work and Shang’s work is discussed in Section 3.5. Cabrera and Schulz (2014) do not work directly on demand but on the generalized quantile function based on daily electricity demand and thus is not of direct interest.

Several authors compared the performance of different types of models and one of the earliest and most comprehensive comparisons was provided by Willis and Northcote-Green (1984), which presents a comparison between fourteen methods used to forecast the electricity load, and concluded that the performance of the methods depends on the nature of the available data as well as several other factors. For example, the choice of a load forecasting method can be governed by computer resources that are available and the level of expertise of the users. Taylor and McSharry (2007) conducted an empirical comparison of some short-term forecasting methods. They used ten time series of intraday demand from ten European countries which were modeled using five different methods. They showed that the double seasonal Holt-Winters exponential smoothing method performed best, followed by the principal component analysis (PCA) and the ARIMA approaches. On the other hand, a new alternative exponential smoothing method and the periodic autoregressive (PAR) model produced disappointing results. The similar results were obtained by Taylor (2008), which suggested that the double seasonal Holt-Winters exponential smoothing method should be used for very short-term predictions,



such as ten to thirty minutes ahead. As pointed out previously, there is a distinction between electricity load and demand. Load (or real-time load) is the actual amount of electricity that was consumed over a given period (such as in one hour), whereas demand is the total amount of electricity that is traded at an electricity market for a given region for the same period in the next day. Even though demand and load are different quantities, approaches that are successful in modeling demand can also be successful in modeling real-time electricity load and vice-versa.

The electric utilities as well as electricity generators are interested in short term forecasting which includes forecasting a few minutes, hours, or days ahead. This is important for market organization reasons. Equally important is the long term forecasting of load from one year up to many years ahead. Commonly, regression models are used for modeling the short-term electricity demand; on the other hand, ARIMA is most commonly used model to forecast the long-term load. Many of the regression based models use several weather related variables, but such variables can also be utilized in other types of models are well.

The weather variables, mainly the temperature, have a significant effect on the electricity load and many publications used temperature in different forms such as minimum and maximum temperature, as wells as temperature and its quadratic term. Other temperature derived variables utilized in modeling the load are heating and cooling degree days (HDD and CDD respectively). The most common definition of CDD is the maximum of  $\{0, (\text{Temperature in Fahrenheit} - 65)\}$  and that of HDD is the maximum of  $\{0, (65 - \text{Temperature in Fahrenheit})\}$ . One of the papers that used several temperature derived variables is Valor, Meneu, and Caselles (2001). This paper employed the mean daily air temperature (in °C) which was calculated as the arithmetic mean of the maximum and minimum daily temperature, and alternatively, the average of the 48 half-hourly values. Small differences were obtained between the two approaches. The temperature data were collected from four weather stations distributed across Spain, and weighted using the population in the weather station zones. The authors defined the temperature variable as  $TI_t = \sum_{i=1}^4 \bar{T}_{ti} w_{ti}$ , where  $\bar{T}_{ti}$  is the mean daily temperature on day  $t$  at weather station  $i$ , and  $w_{ti}$  is a population weight of the area assigned to each station, which is calculated as  $w_{ti} = \frac{p_{ti}}{\sum_{i=1}^4 p_{ti}}$ , where  $p_{ti}$  is the total population on day  $t$  assigned to weather station  $i$ . The

authors also showed the presence of a significant trend related to demographic, social, and economic factors and incorporated seasonal factors such as holiday, weekly, and monthly effects into the model.

Momani (2013), using regression models, examined the relationship between the electricity demand and climate/non-climate factors, and showed that the consumption pattern is affected by demographic, technological, environmental, and national energy pricing. Cancelo, Espasa, and Grafe (2008) proposed two models for load forecasting, the short-term model based on weather and calendar data, and the long-term model based on population, gross state product (GSP), and price changes. Hor, Watson, and Majithia (2005) presented three basic models which proposed different parametrizations of temperature and humidity. Their three models included Gross Domestic Product (GDP), population growth, and some weather variables such as mean monthly wind speed, mean monthly sunshine hours, and monthly rainfall. In addition, they changed the temperature variables between the three models. In the first model, they included HDD, CDD, and enthalpy latent days (ELD)<sup>2</sup>, defined by  $ELD = \frac{1}{24} \sum_{d=1}^{N_d} \sum_{h=1}^{24} (\gamma_h) (Q - Q_b)$  where  $Q$  is the hourly value of enthalpy (in kilojoules per kilogram) and  $Q_b$  is the enthalpy at the reference temperature of 25.6°C,  $\gamma_h$  is indicator function that takes the value 1 if the hourly value of the temperature is above 25.6°C and a value 0 if the temperature is below this value or if  $Q - Q_b < 0$ . The second model included the monthly central England temperature (CET) and the mean monthly relative humidity. The third model included HDD, CDD, and the mean monthly relative humidity. The first and third models performed better than the second model.

In this dissertation, three approaches to building a load model were adopted, each targeting somewhat different goals. It was expected that some models will work better than the other models under certain conditions. The first approach was developed to capture the overall 24-hour electricity load profile specific for each season: Winter = {December, January, and February}, summer = {June, July, and August}, spring = {March, April, May} fall={September, October, and November}, and for weekdays and

---

<sup>2</sup> Enthalpy Latent Days (ELD): An alternative metric for assessing the cooling load was established by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE). The ELD accounts the possible humidity effects on air-conditioning demand.

weekends. This approach also accounted for the long-term trend due to economic factors and seasonal patterns in the total weekly load. In this method, the 24-hour load profile was captured through cubic splines, with data separated into seasons. The data was also separated into those from weekdays and those from weekends. Thus, each season and weekday/weekend combination had a separate model. A less flexible model that used dummy variables to separate effects due to weekdays and weekends was also built. The second approach to load modeling fitted cubic splines to the daily profile observed over each month and let the spline coefficients vary from month to month. The time series generated by the estimated spline coefficients were modeled using a Vector Autoregressive model with exogenous variables (VARX model). The VARX model contained lag variables to account for seasonality. The third approach employed functional principal components to model the daily profiles. The daily profile of each day was approximated as a linear combination of three eigenfunctions (Harmonics) and the associated coefficients of the linear combination (which are the eigencoefficients or scores) were modeled using transfer function time series.

The first two approaches to load modeling introduced herein were developed so as to provide the user with methodologies that are easy to implement with software that are widely available. The various steps used in each approach are based on regression and time series methodology that is easy to understand and implement by users who are not highly trained statisticians. The third approach, while resorting to new functional principal component techniques that may be unfamiliar to some users of classical statistical methods, can be implemented using freely available open-source software. Even though it may be somewhat challenging to the average modeler of electricity load, the underlying steps are simple enough that given sufficient time for familiarization, the third approach will also provide an appealing alternative to those without a high level of training in statistics.

## 2. FACTORS AFFECTING ELECTRICITY LODA AND DATA

### 2.1 THE FACTORS AFFECTING ELECTRICITY LOAD

There are many factors that affect the electricity load, some in the short term and others in the long term. Fahad and Arbab (2014) described the impact of various factors on the short term load, and grouped those factors grouped into four categories, namely time, weather, economy, and random disturbances. On the other hand, economic factors can also affect the load in the long term. Each of the broad categories these factors fall into are discussed in the following subsections.

**2.1.1 Economic Factors.** Several economic and macroeconomic factors influence the long term electricity load and researchers have utilized these to obtain long term forecasts. For example, Bianco et al. (2009) developed different regression models to forecast annual electricity consumption in Italy using gross domestic product (GDP), gross domestic product per capita (GDP per capita) and population. Gross domestic product was also used as a regressor variable in Mohamed and Bodger (2005), which in addition employed average price of electricity and population as selected variables to forecast obtain long term forecasts of the load.

Use of population figures as a predictor of electricity load seems obvious, because each individual in a population uses electricity to varying degrees but overall the total amount of electricity used is linked to the number of people living in the region under study. The price of electricity can also affect the load because higher prices would tend to depress per capita consumption of this commodity. On the other hand, the link between electricity load and gross domestic product may not be immediately apparent to a casual observer. The link between electricity load and gross domestic product is easily seen by examining Figure 2.1. It shows that the March 2013 New Jersey retail sales of electricity for commercial use is more than 50% of the retail sales for all sectors. Commercial electricity use is directly linked to industrial production and other such activities that contribute to gross domestic product, thus making GDP a predictor of electricity load.

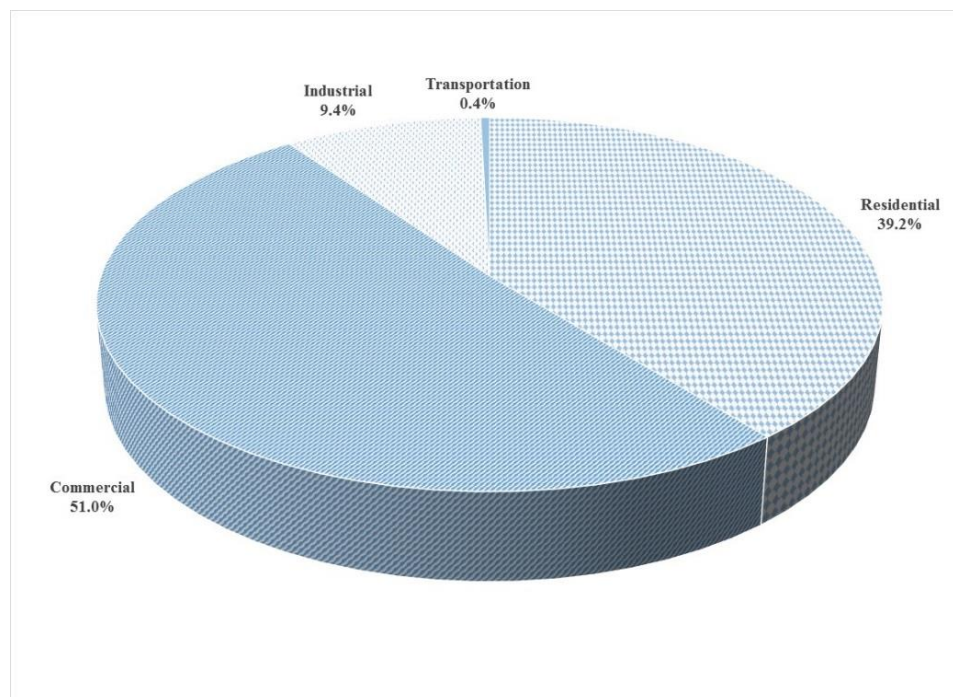


Figure. 2.1 Retail Sales of Electricity to Ultimate Customers by end-use Sector for New Jersey, March 2013

**2.1.2 Time Related Factors.** The changes in electricity load due to variations in human activity are time related. Such variations can be due to the 24-hour cycling between working, leisure, and sleeping periods. Other time related factors include day of the week, holidays, and seasonal changes in consumer behavior. Fig 2.2 gives the average load curve over 24 hours for January 2013 in the study area which covers South New Jersey. It shows that during weekdays there is a peak at 8:00 am when work usually starts, and a second peak at 7:00 pm when the most people are at home, cooking, using electronic devices such as Television. On the other hand, during weekends the first peak is at 10:00 am which is a shift of two hours compared with weekdays, but the second peak does not change. This may be because people get up at a later time on weekends and activities that consume electricity, such as cooking, are delayed. During summer months, a slightly different pattern is observed. For July 2013, (Fig 2.3) we see that there is no big difference between weekdays and weekends, but the load curve is lowest at 6:00 am during weekdays, whereas during weekends this low point occurs at 8:00 am.

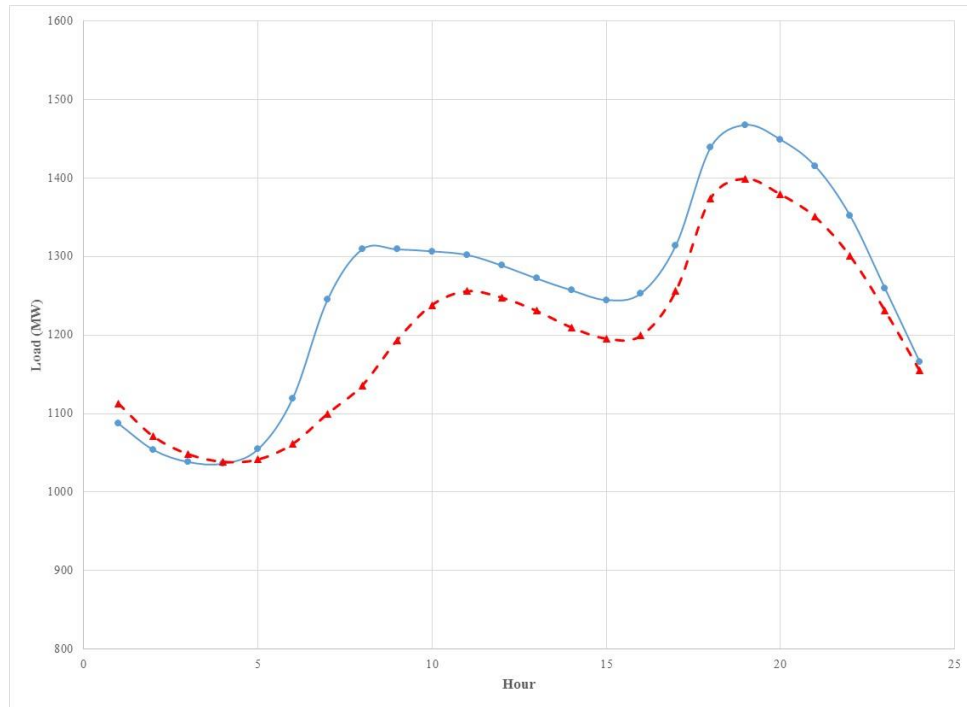


Figure. 2.2 The Average of a 24-hour of January Load Curve of Weekdays (blue solid) and Weekends (red dashed) 1993-2012

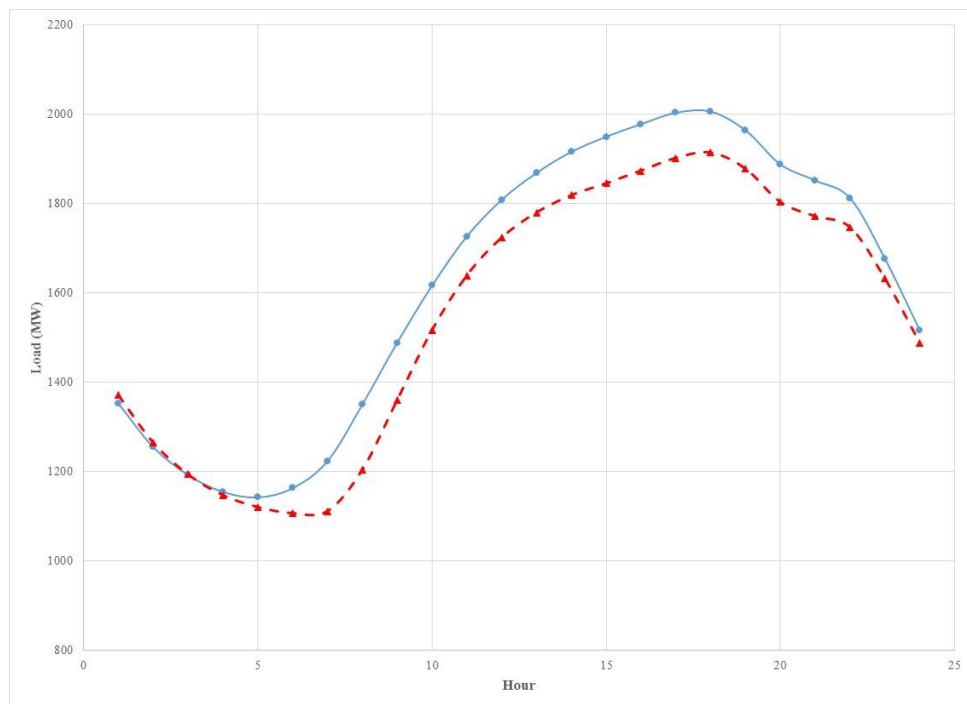


Figure. 2.3 The Average of a 24-hour of July Load Curve of Weekdays (blue solid) and Weekends (red dashed) 1993-2012

**2.1.3 Weather Related Factors.** The weather variables, such as temperature, humidity, precipitation, and wind speed, have a significant place in electricity load forecasting. Out of these factors, the temperature plays a major role in load forecasting. Valor et al. (2001) discussed the relationship between electricity load and daily air temperature, but used heating and cooling degree days instead temperature because of the nonlinear relationship between load and temperature. Fig 2.4 shows quadratic relationship between hourly load and hourly temperature, and the inflection point of the curve is around  $60^{\circ} F$ . This temperature is normally encountered during spring and fall months which are traditionally termed “shoulder” months by those engaged in load modeling. In addition, we can observe that the load peaks coincide with high temperature, and that occurs during summer months. Fig 2.5 shows clearly the almost synchronous movement of load and temperature (plotted using with different scales) for the summer month of July 2013. Such a highly linked relationship between hourly temperature and load is not seen during winter months as seen in Figure 2.6.

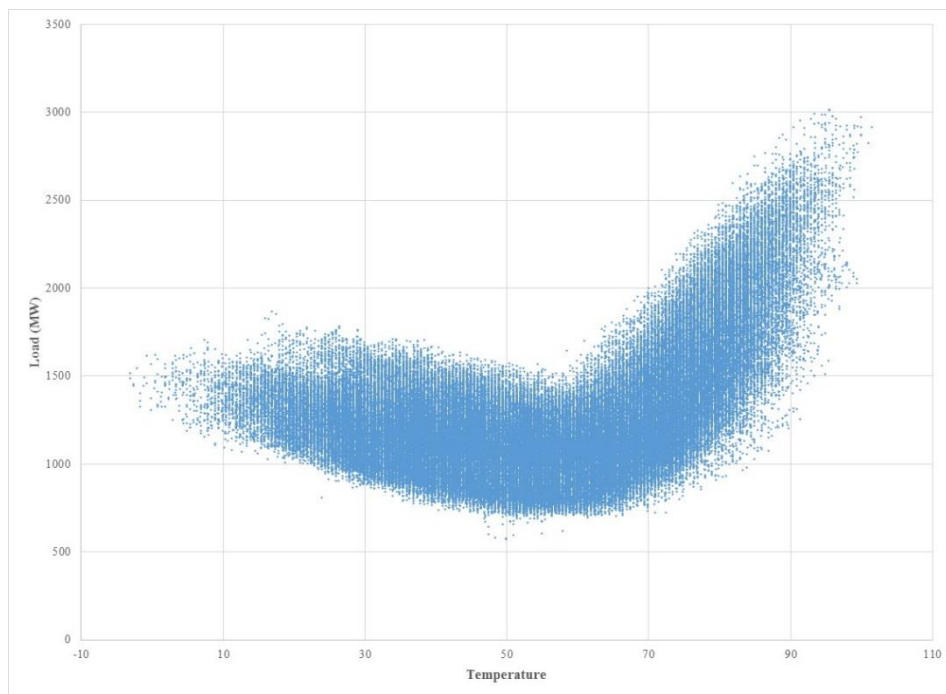


Figure. 2.4 The Relationship between the Hourly Load and Hourly Temperature  
– South New Jersey 1993-2012

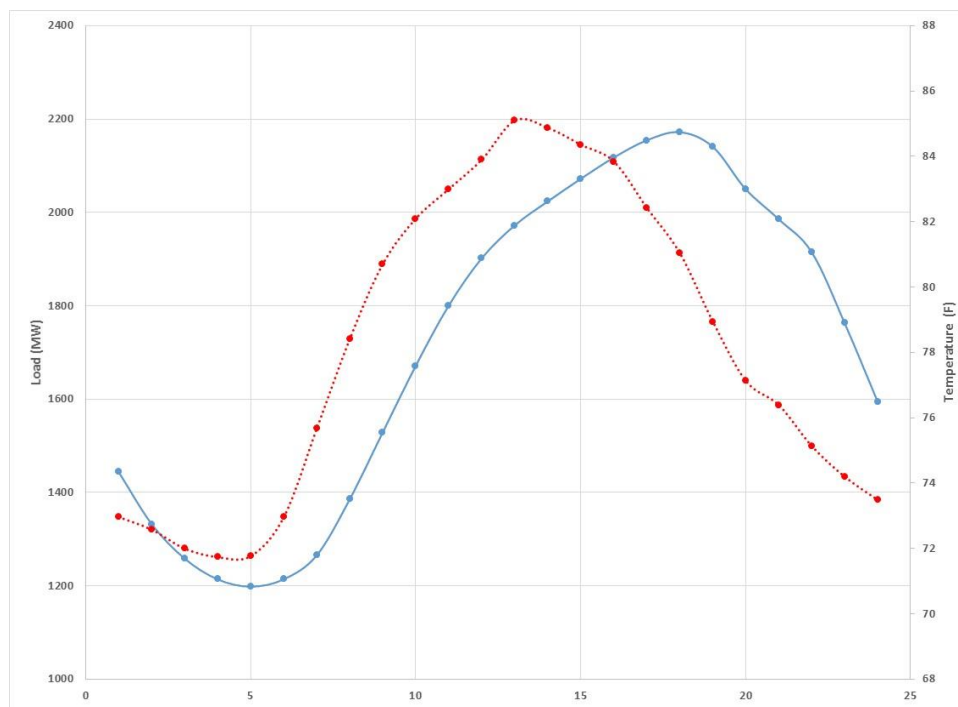


Figure. 2.5 The Average of 24-hour Load Curves (blue solid) and the Temperature (red dotted) Jul. 2013

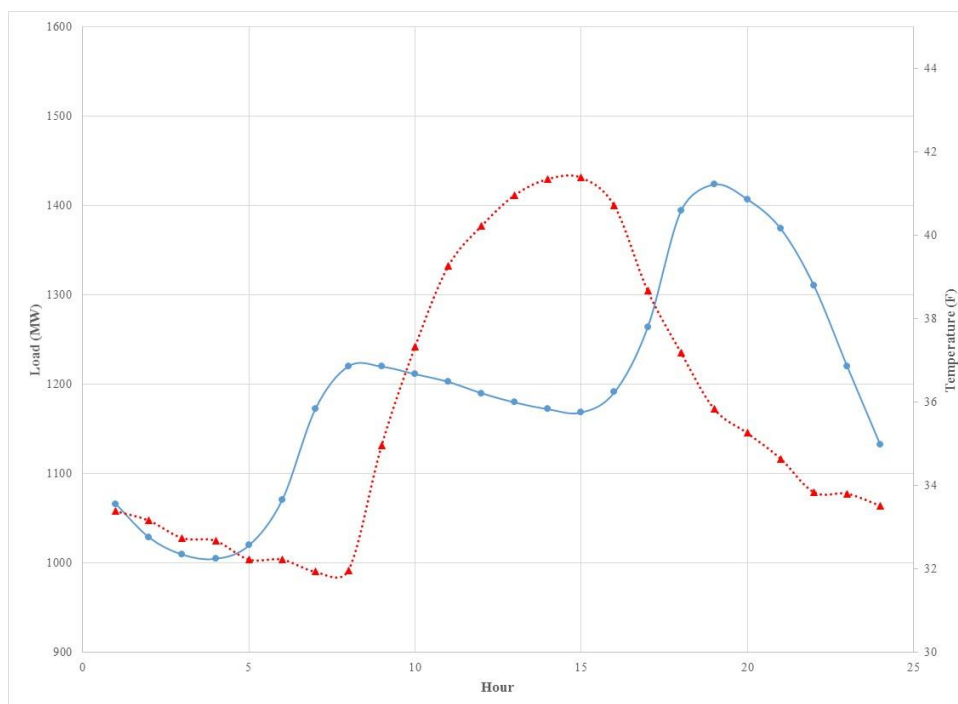


Figure. 2.6 The Average of 24-hour Load Curves (blue solid) and the Temperature (red dotted) Jan. 2013



**2.1.4 The Electricity Load Curve over Different Regions.** The 24-hour electricity load curve differs from region to region. The main reason for this difference are the weather variables, mainly temperature. Fig 2.7 shows the average of the 24-hour electricity load curve for three different regions, which are Atlantic City Electric zone (AECO or AE<sup>3</sup>), which is the study area, Pennsylvania Electric Company (PENELEC or PN<sup>4</sup>), and the south zone of the Electric Reliability Council of Taxes (ERCOT<sup>5</sup>).

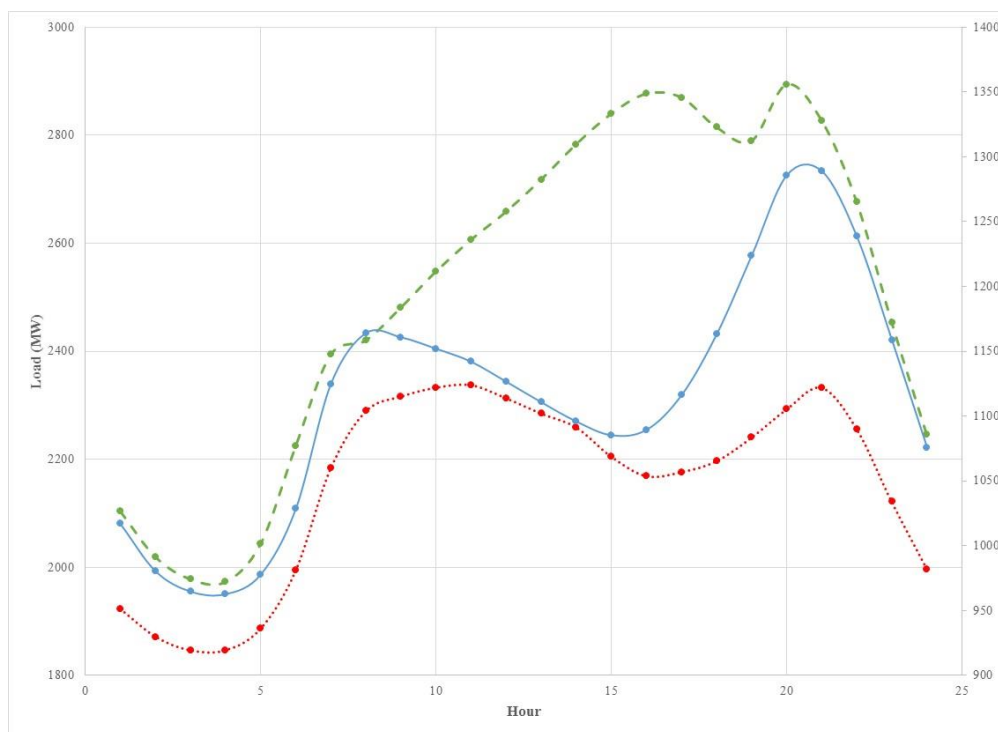


Figure. 2.7 The Average of 24-hour Load Curves from Mar. 2013 for AE (blue), PN (red), and the South of ERCOT (green)

<sup>3</sup> Atlantic City Electric, a subsidiary of Pepco Holdings, Inc. (PHI), delivers safe, reliable and affordable electric service to more than 545,000 customers in southern New Jersey.

<sup>4</sup> Pennsylvania Electric Company, is one of 10 electric utility operating companies in FirstEnergy Corp which is one of the nation's largest investor-owned systems, based on 6 million customers served within a 65,000 squared-mile area.

<sup>5</sup> Electric Reliability Council of Taxes (ERCOT), manages the flow of electric power to 24 million Taxes customers which connects more than 43,000 miles of transmission lines and 550 generation units.

## 2.2 DATA SOURCES

The historical load dataset was obtained from the Pennsylvania-New Jersey-Maryland RTO website (PJM<sup>6</sup>). The PJM market is the world's largest wholesale electricity market and serves all parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia (see Figure 2.8). The data used in this dissertation cover a sub region of PJM, namely the Atlantic City Electric zone (AE) in southern New Jersey which is shown in Fig 2.9. This dataset includes hourly observations measured in megawatts (MW) over 20 years from January 1, 1993 through December 31, 2012 as shown in Fig 2.10. This data were used for modeling purposes (i.e. as training data), and the data from January 1, 2013 through December 31, 2013, shown in Fig 2.11, were used for the computing forecasting error (i.e. as test data). The economic data was obtained from Federal Reserve Bank of St. Louis<sup>7</sup>. Moreover, the weather data was obtained from the National Oceanic Atmospheric Administration (NOAA<sup>8</sup>) based on four weather stations in different locations of the study area. These stations are located in Atlantic City, Millville, Mount Holly, and Wildwood. The temperature data used in this study were computed as weighted average of the individual station data, with sub-area populations used as weights.

The specific economic variables used in this study are: industrial production index in the US (IPI) which is an economic indicator that measures the amount of the output from manufacturing, mining, electric and gas industries; government employment in New Jersey (NJGOVTN), which is defined as the total body of employees in all government agencies apart from the military; home vacancy rate in New Jersey (NJHVAC), which is defined as the percentage of all available units in a rental property that are vacant or unoccupied at a particular time.

---

<sup>6</sup> PJM Interconnection is a regional transmission organization (RTO) that coordinates the movement of wholesale electricity in all or parts of 13 states and the District of Columbia, an area that includes more than 51 million people.

<sup>7</sup> The **Federal Reserve Bank of St. Louis** was established in 1914, after the creation of the [Federal Reserve System](#) in 1913. The Eighth Federal Reserve District is headquartered in St. Louis and has branches in [Little Rock, AR.](#), [Louisville, KY.](#), and [Memphis, TN.](#)

<sup>8</sup> The National Oceanic and Atmospheric Administration (NOAA) is an [American](#) scientific agency within the [United States Department of Commerce](#) focused on the conditions of the [oceans](#) and the [atmosphere](#). NOAA warns of dangerous [weather](#), charts seas and skies, guides the use and protection of ocean and coastal resources, and conducts research to improve understanding and stewardship of the [environment](#).

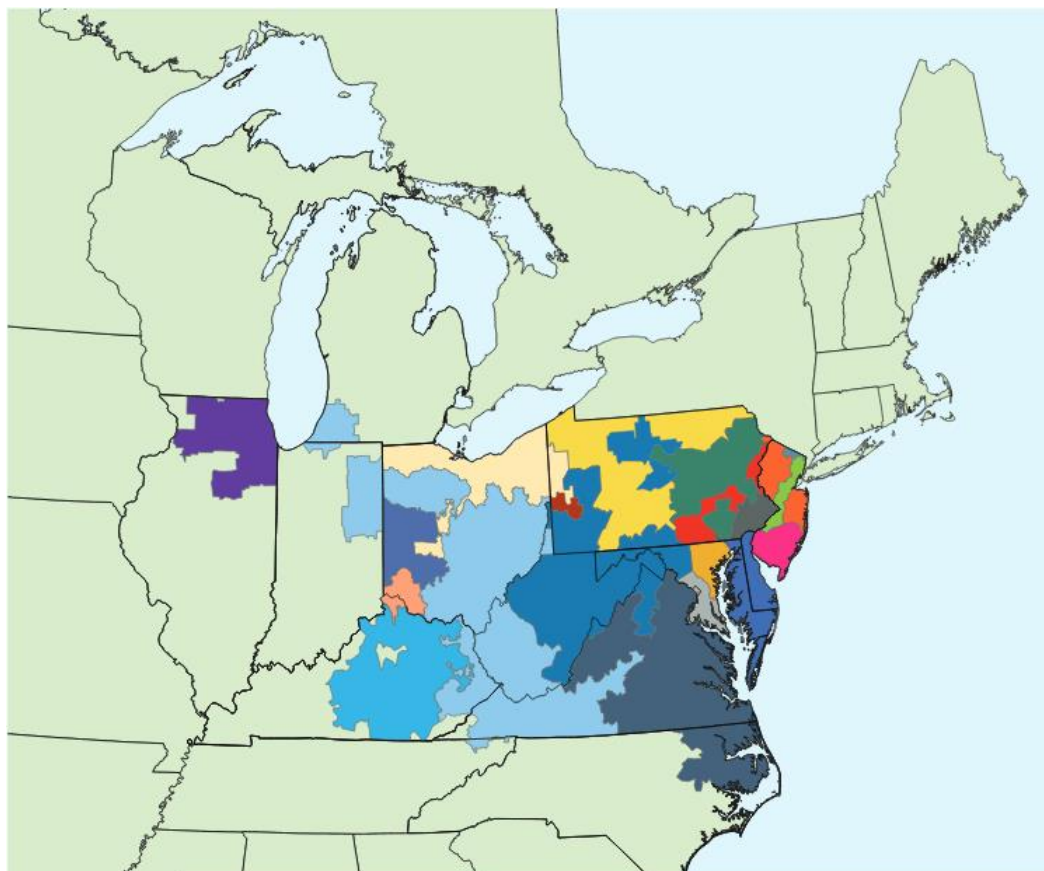






















Figure. 2.8 Map of PJM Regional Transmission Organization<sup>9</sup>

### Legend

#### ZONE

 Allegheny Power Systems	 Eastern Kentucky Power Cooperative
 American Electric Power Co., Inc.	 Jersey Central Power and Light Company
 American Transmission Systems, Inc.	 Metropolitan Edison Company
 Atlantic City Electric Company	 PPL Electric Utilities
 Baltimore Gas and Electric Company	 PECO Energy
 ComEd	 Pennsylvania Electric Company
 Dayton Power and Light Co.	 Potomac Electric Power Company
 Delmarva Power and Light Company	 Public Service Electric and Gas Company
 Dominion	 Rockland Electric Company
 Duke Energy Ohio and Kentucky	
 Duquesne Light	

<sup>9</sup> PJM website. <https://www.pjm.com/~media/about-pjm/pjm-zones.ashx>



Figure. 2.9 Map Service of Atlantic City Electric (AE) – Study Area<sup>10</sup>

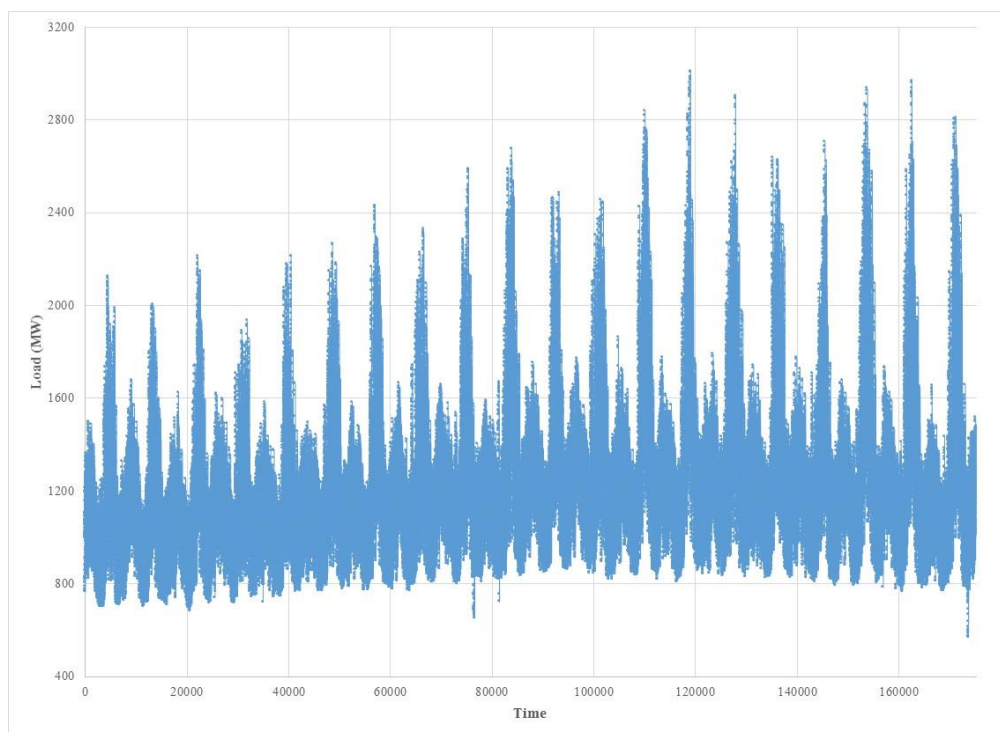


Figure. 2.10 The Hourly Load of AE over 20 Years 1993 – 2012

<sup>10</sup> Atlantic City Electric (AE) website. <http://www.atlanticcityelectric.com/connect-with-us/doing-business-with-us/builders-and-inspectors/resources/service-area-map/>

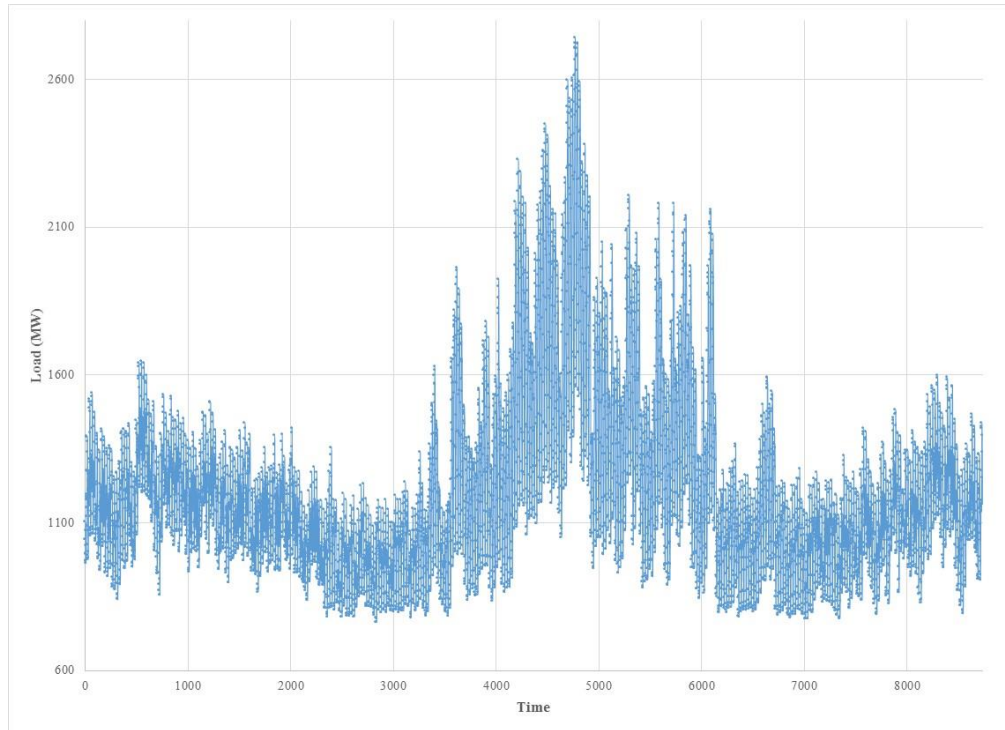


Figure. 2.11 The Hourly Load of AE in 2013

### 3. METHODOLOGY AND RELATED STATISTICAL TOOLS

A brief overview of the main statistical curve fitting approaches used in this research is given in this chapter. In total three approaches were used. The first approach to modeling electricity load utilizes a variation of cubic splines and hourly temperature to model hourly electricity load for weekdays and weekends in a given season. Model parameters were estimated for the mean load curve separately for weekdays and weekend days within each season. This approach can be considered suitable for short term prediction (because of the need to have good estimates of hourly temperature) and will be termed the short-term approach. The second approach fitted a separate spline model for each month across the 20-year time period over which the training data set was observed. The number of knots and the position of knots remained unchanged from month to month but the spline coefficients were allowed to vary over time. The resulting spline coefficient estimates were then modeled using a vector autoregressive model with temperature and its square as exogenous variables. The forecast values of these spline coefficients were then used to construct load profiles for future days. This approach is suitable for long-term forecasting (because only monthly temperature is needed and predictions are for the average load curve for a given month) and will be termed the long-term approach. The third approach employed functional principal component analysis (FPCA) to model the load profile for each day. The last approach will be termed the FPCA approach.

A brief overview of the statistical tools used in building models for each of the three approaches is given prior to describing the three main approaches taken to model the daily profile of electricity load. These include a short introduction to spline modeling and functional principal component analysis.

#### 3.1 SPLINE MODELS

A spline model is a piece-wise defined function where the individual segments are connected using continuity and smoothness restrictions. They are useful in describing the relationship between a response variable and one or more independent variables when the relationship requires a flexible model. Usually the segments of a spline function are low

order polynomials and the polynomial segments connect at a set of finite points known as knots. There are several types of spline functions and in order to provide a more formal introduction, the following notation is introduced. To make the explanations simpler, only the case where there is a single independent variable is considered.

Let  $X$  and  $Y$  be jointly distributed random variables such that

$$E[Y | X = x] = f(x) \quad (1)$$

where  $Y$  is the dependent variable and  $X$  is the independent variable, in the regression sense. The task is to estimate the function  $f$  based on observed data

$(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . Suppose that the observed data indicate that simple forms such as linear and quadratic functions would not fit the data well and using a high-order polynomial does not necessarily provide a better fit and pose challenges in interpreting the coefficients of the large number of polynomial terms. As an alternative approach, one can segment the domain  $[a, b]$  of the function  $f$  into  $k + 1$  segments  $K_i = [\kappa_{i-1}, \kappa_i]$   $i = 1, \dots, k+1$ , where  $\kappa_0 = a$ ,  $\kappa_{k+1} = b$  and defined the piece-wise function

$$f(x) = \sum_{i=1}^{k+1} g_i(x) I_{K_i}(x) \quad (2)$$

where  $g_i$  is a function defined on  $K_i$  for  $i = 1, 2, \dots, k+1$ , and the  $I_{K_i}$  are indicator functions that takes the value of one when the argument is in  $K_i$  and zero otherwise, for  $i = 1, 2, \dots, k+1$ . The points  $\kappa_i$ ,  $i = 1, 2, \dots, k$  are called knots. Knots can be pre-specified prior to estimating the function  $f$  or let that data determine their positions.

To ensure the continuity of  $f$ , the conditions  $g_i(\kappa_i) = g_{i+1}(\kappa_i)$  for  $i = 1, 2, \dots, k$  are imposed. In general, smoothness conditions are also imposed by letting one or more derivatives of the functions  $g_i$  match at the knots. Usually, the segment functions  $g_i$  are

selected to be low order polynomials and one of the commonly used are cubic polynomials. One reason given for this is that the human eye cannot detect discontinuity in the third derivative (Azzalini and Scarpa, 2012).

**3.1.1 Cubic Splines.** In cubic splines, each of the functions  $g_i$  takes the form

$$g_i(x) = a_i + b_i x + c_i x^2 + d_i x^3 \quad \text{for } i = 1, 2, \dots, k+1. \quad (3)$$

In the cubic spline setting, in addition to the continuity conditions  $g_i(\kappa_i) = g_{i+1}(\kappa_i)$  for  $i=1, 2, \dots, k$ , the following conditions are imposed on the functions  $g_i$  given in Equation (2):

$$g_i^{(m)}(\kappa_i) = g_{i+1}^{(m)}(\kappa_i) \quad \text{for } m = 1, 2 \text{ and } i = 1, 2, \dots, k+1 \quad (4)$$

where  $g_i^{(m)}$  denotes the  $m^{\text{th}}$  derivative of the function  $g_i$ . These are called *natural* or *simple* boundary conditions. The above conditions reduce the spline function given in Equation (3.1.1) to the following:

$$f(x) = a + bx + cx^2 + dx^3 + \sum_{i=1}^k e_i (x - \kappa_i)_+^3 \quad (5)$$

where  $x_+ = xI_{[0,\infty)}(x)$ . In other words, the function  $f$  is expressed as a linear combination of the basis functions  $u_1(x) = 1$ ,  $u_2(x) = x$ ,  $u_3(x) = x^2$ ,  $u_4(x) = x^3$ ,  $u_5(x) = (x - \kappa_1)_+^3$ ,  $\dots$ ,  $u_{k+4}(x) = (x - \kappa_k)_+^3$ .

Thus, if the number of segments  $k=4$ , there will be eight basis functions, one each for zero, first, second, and third powers of  $x$ , and functions of the form  $(x - \kappa_i)_+^3$  for the last  $k=4$  segments of the interval  $[a, b]$ . In general for cubic splines the number of basis functions, called the degrees of freedom, equals four degrees of freedom for each



segment of the interval  $[a, b]$  minus the number of constraints. So when  $k=4$ , the interval is divided into 5 segments, resulting in 20 degrees of freedom, but then there are 4 continuity restrictions and 8 restrictions equating first and second derivatives at each knot, yielding a net degrees of freedom of 8. If the second derivative restriction is removed, as was done in some parts of the analysis, the number of basis functions increases to 12.

An attractive feature of the cubic splines is that the coefficients  $a, b, c, d$ , and  $e_i$  in Equation (3.1.2) can be estimated using least squares. These spline functions with fixed degrees of freedom are sometimes called *regression splines*. Another special type of cubic spline is the *natural cubic splines*. Natural cubic splines have the extra condition that

$$g_1^{(m)}(a) = g_{k+1}^{(m)}(b) = 0 \text{ for } m = 2,$$

which is the same as saying that the function is linear beyond the region  $[a, b]$ . This is important if one is to avoid erratic behavior at the margins of the observed data. The above conditions can be part of the simple conditions mentioned earlier. If in addition  $g_1^{(m)}(a) = g_{k+1}^{(m)}(b) = 0$  for  $m=1$ , then we have what is termed as the *clamped* boundary conditions. This terminology arises out of the connection between a cubic spline and the shape taken by a thin flexible beam of wood with constant flexural stiffness when it is constrained to pass through points that corresponds to the knots in the cubic spline (Dancose and Angeles, 1990).

In this thesis, the daily profile of the hourly electricity load was modeled using cubic splines with the continuity restrictions and the first derivative restriction

$$g_i^{(1)}(\kappa_i) = g_{i+1}^{(1)}(\kappa_i) \text{ for } i=1, 2, \dots, k+1, \text{ but without the second derivative restriction}$$

$g_i^{(2)}(\kappa_i) = g_{i+1}^{(2)}(\kappa_i)$ . The lack of the second derivative restriction introduces more roughness into the model but allows the spline to better fit the data.

**3.1.2 Smoothing Splines.** Another type of a spline model that is extensively used is the smoothing spline. Smoothing splines come closes to the observed data points from

among all functions that satisfy a pre-specified “smoothness” criterion. Specifically, given a set of data  $(x_i, y_i), i = 1, 2, \dots, n$ , a smoothing spline is the function  $f$  that satisfies:

$$\text{Min}_{f \in \{h: [a, b] \rightarrow \mathbb{R}; h^{(2)} \text{ exists}\}} \left\{ \sum_{i=1}^n [y_i - f(x_i)]^2 + \lambda \int_a^b [f^{(2)}(u)]^2 du \right\}, \quad (6)$$

for some fixed  $\lambda$ . The quantity  $\int_a^b [f^{(2)}(u)]^2 du$  is called the roughness penalty and  $\lambda$  is called the smoothing parameter. Larger values of  $\lambda$  produces functions that are smoother than those for smaller values of  $\lambda$ . The value of  $\lambda$  may be optimized using the mean squared error obtained from fitting the estimated function to an independent data set.

An interesting result that connects smoothing splines to the natural cubic splines is given by the following theorem.

**Theorem 3.2.1:** Let  $f$  satisfy the minimization criteria given in (6). Then  $f$  is a natural cubic spline with knots at every unique value in the set  $\{x_i; i = 1, 2, \dots, n\}$ .

**Proof:** See Wang (2011, p. 6) for an outline.

While smoothing splines could have been used in modeling the electricity load data, the cubic spline with a relaxed smoothness criterion was employed in this study because of the ease of estimating the coefficients using standard regression software. When modeling the hourly electricity load data, the domain of the spline function  $f$  is the interval  $[0, 24]$ , representing the 24-hour period that defines each day. Letting  $t$  denote the day and  $x$  denote a time point in  $[0, 24]$ ,  $f_t(x)$  will represent the expected value of the average electricity load for the hour ending at time  $x$  during day  $t$ . If  $y_t(x_i)$  denotes the observed electricity load at time  $x_i$  on day  $t$ , it is assumed that

$$y_t(x_i) = y_{t,i} = f_t(x_i) + \varepsilon_{t,i} \quad (7)$$

where  $\varepsilon_{t,i} \sim \text{independent}(0, \sigma_i^2)$  for  $t=1, 2, \dots, N$  and  $i=1, 2, \dots, 24$ . Also note that the data set used in this study reports the average electricity load for the hour ending at time  $x_i = 1, 2, \dots, 24$ .

### 3.2 VECTOR AUTOREGRESSIVE MODEL WITH EXOGENOUS VARIABLES (VARX)

The Vector Autoregressive formulation (VAR) is a natural extension of the univariate autoregressive representation and is commonly used to model multivariate time series. An extension of the VAR model is the Vector Autoregressive Model with Exogenous Variables (VARX) formulation that allows one or more exogenous regressor variables to enter into the model. The structure of the VARX model allows a linear function of past lags of the time series and past lags of the other variables to explain the current value of the time series. The Vector Autoregressive model with exogenous variable VARX( $p, s$ ) is written as:

$$y_t = \delta + \sum_{i=1}^p \Phi_i y_{t-i} + \sum_{i=0}^s \Theta_i^* x_{t-i} + \varepsilon_t, \quad (8)$$

where,  $y_t = (y_{1t}, \dots, y_{kt})'$ ,  $t=1, 2, \dots, N$ , denote a  $k$ -dimensional time series, whose components are univariate time series of interest,  $x_t = (x_{1t}, \dots, x_{rt})'$  is an  $r$ -dimensional vector time series of exogenous variables,  $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{kt})'$  is a vector white noise process such that  $E(\varepsilon_t) = 0$  and  $E(\varepsilon_t \varepsilon_t') = \Sigma$  is a finite positive definite matrix,

$E(\varepsilon_t \varepsilon_s') = 0$  for  $t \neq s$ ,  $\Phi$  is a  $k \times k$  matrix, and  $\Theta_i^*$  is a  $k \times r$  matrix. Note that  $\delta$  in Equation (8) denote an intercept vector.

For example, a VARX(1,0) model for a 3-dimensional time series with two input variables is

$$y_t = \delta + \Phi_1 y_{t-1} + \Theta_0^* x_t + \varepsilon_t, \quad (9)$$

where,  $y_t = (y_{1t}, y_{2t}, y_{3t})'$  and  $x_t = (x_{1t}, x_{2t})'$ .

The parameters of a VAR model are estimated using the maximum likelihood estimation (MLE) method. The likelihood function is determined using the Kalman filtering approach that provides a state-space formulation of the model. In the VARX case, however, this approach produces a very large state-space model and hence the parameters of the VARX model are estimated using a two-step procedure. First the deterministic portion of the model that contains the constant term, any linear and quadratic trend components, seasonal dummies if present, and the coefficients of the exogenous variables, are estimated. Then fixing these parameters at the estimated value, the rest of the parameters are determined using the MLE method.

**3.2.1 Prediction Intervals for VARX Forecasts.** The VARX model was employed in this work to obtain forecasts of spline coefficients that could then be used to construct the 24-hour load profile for a future day. In addition to point predictions, simultaneous forecast intervals were also obtained for the average load at each hour of future months. That is, assuming that the estimation was done over  $M$  observed months, simultaneous prediction intervals were obtained for the average load  $Y_{M+h,i}$  for each of the hours  $i = 1, 2, \dots, 24$ , for the future month  $M+h$ ,  $h = 1, 2, \dots, 12$ . Note that what is referred to as the average load for a give hour is the load for that hour averaged over all days of that month.

If there are  $d$  basis functions used in modeling the 24-hour load curve, then there are  $d$  estimated coefficients of the spline function. Also assume that the spline function has  $k$  knots and that only the equality of the first derivative at the knots is imposed. Then a set of  $(1-\alpha)100\%$  simultaneous confidence intervals about each  $Y_{M+h,i}$ , using Scheffe's method is given by:

$$Y_{n+h,m} \pm \sqrt{d \times F_{(1-\alpha; d, n-d)}} \times s(x'V_{M+h}x) + s_e, \quad (10)$$

where,  $x = [1, i, i^2, i^3, (i - \kappa_1)^2, \dots, (i - \kappa_k)^2, (i - \kappa_1)^3, \dots, (i - \kappa_k)^3]^T$ ,  $S_e$  is the estimated standard deviation of the terms  $Y_{m,i} - \hat{f}_m(i)$ ,  $m = 1, 2, \dots, M$ ,  $i = 1, 2, \dots, 24$ , and  $V_{M+h}$  denotes the estimated variance-covariance matrix of the forecast coefficient vector.

### 3.3 TRANSFER FUNCTION MODELS

A transfer function model is a statistical model that describes the behavior of a time series as a function of its past values as well as one or more independent variables and their lags. Whereas the ARIMA model is purely a univariate time series model, transfer function model deals with more than one time series.

The simple transfer function model is

$$Y_t = \mu + v(B)X_t + E_t = \mu + \frac{\omega(B)B^b}{\delta(B)}X_t + \frac{\theta(B)}{\phi(B)}\varepsilon_t, \quad (11)$$

$t = 1, 2, \dots, N$ , where  $\mu$  is a constant,  $\omega(B) = \omega_0 + \omega_1 B + \omega_2 B^2 + \dots + \omega_s B^s$  and  $\delta(B) = 1 - \delta_1 B - \dots - \delta_r B^r$  are finite-order polynomial in  $B$  with degree  $s$  and  $r$ , and  $\omega_0 \neq 0$ . Obviously,  $\omega(B)$  and  $\delta(B)$  are assumed to have no common factors. Also,  $\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ ,  $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$  are polynomial in  $B$  of degree  $q$  and  $p$  respectively, and  $\{\varepsilon_t\}$  is a sequence of independent and identically distributed random variables with mean zero and variance  $\sigma^2$  ( $\varepsilon_t \sim iid(0, \sigma^2)$ ). Usually  $\varepsilon_t$  is assume to be normal. The parameter  $b$  is called the *time delay* (or dead time) of the system; when  $b > 0$  the transfer function model is useful for predicting the turning points of  $Y_t$  given those of  $X_t$ . A transfer function model can have more than one input variables. For example such a model with two input variables is given by

$$Y_t = \mu + \frac{\omega_1(B)B^{b_1}}{\delta_1(B)}X_{1t} + \frac{\omega_2(B)B^{b_2}}{\delta_2(B)}X_{2t} + \frac{\theta(B)}{\phi(B)}\varepsilon_t, \quad (12)$$

The above type of formulation was used in this research to model the eigencoefficients or scores that are generated by FPCA of the 24-hour load curve for each day of the 20 year observation period.

### 3.4 FUNCTIONAL DATA ANALYSIS (FDA)

Functional data, that are usually high-dimensional data, can be observed in different fields and in many forms. Usually, functional data arises when repeated measurements are taken over a given domain. The functional data analysis is a set of modern techniques that allow you to perform statistical analysis on set of curves or multidimensional shapes. The main benefit of FDA is that we look at the functional data as a whole, and it does not require us to select a single dependent variable for study. We can use functional versions of common analysis methods (e.g. PCA, ANOVA)

**3.4.1 Functional Principal Component Analysis (FPCA).** The principal component analysis (PCA) is a statistical technique that utilizes the variation in a multivariate dataset to identify a set of orthogonal linear combinations of the original variables that can be arrange in the order of the amount of total variation they explain as a percentage to the total variation explained by all the variables. This allows one to pick only a few of the linear combinations, which are called principal components, that explain a significant amount of variation in the data, thus reducing the dimension of the problem. Therefore, PCA is used often as a dimension reduction technique. That is, PCA allows one to reduce the number of dimension without loss much information. The first principal component explains the largest amount of the variance. The second explains the nest largest and so on.

The idea of extension of the PCA to functional principal component analysis (FPCA) is a natural one. In FPCA, the data vectors are replaced by functions, and scalar products in vector space by scalar products in  $L^2$  space. The first step in the FPCA is to use observed values of a function taken at discrete points in its domain to obtain a smooth function, using B-splines or some such smoothing technique. Then the smoothed function is utilized to carry out the FPCA. To illustrate the FPCS methodology, assume that the observed data points  $(x_i, y_{t,i}), i=1, 2, \dots, n, t=1, 2, \dots, N$  represent the

electricity loads,  $y_{t,i}$ , at time points  $x_i$  observed each hour  $i$  of day  $t$ . Assume that this data are generated by an underlying function  $f_t$  which is the realization of a stochastic process. It is assumed that the observed data obeys the following relationship:

$$y_i = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}, \quad t = 1, 2, \dots, N, \quad (13)$$

where the  $\varepsilon_{t,i}$  are independent identically distributed random variables with zero mean and unit variance and  $\sigma_t(x_i)$ 's are multiplicative factors that enable the noise components  $\sigma_t(x_i)\varepsilon_{t,i}$  to have different variances across  $t$  and  $i$ . The smoothed functions, say  $S_t$ , obtained for each day  $t$  using B-splines or some other technique, is assumed to be “close” to the respective functions  $f_t$ . Given a realization  $f_t$  of a random function defined over a compact domain  $[0, \tau]$ , we can write

$$f_t(x) = \mu(x) + \sum_{k=1}^{\infty} \beta_{t,k} \phi_k(x), \quad \text{for } x \in [0, \tau] \quad (14)$$

where  $\mu$  is the population mean function,  $\beta_{t,k}$  is the  $k^{\text{th}}$  principal component score, and  $\phi_k$  is the  $k^{\text{th}}$  functional principal component (Shang 2013). Note that in the context of load curve modeling,  $\tau = 24$ . Similar to the dimension reduction performed in regular PCA, the first few functional principal components that explain a major portion of the total variation can be selected. Assuming that  $K$  components are selected, one can write,

$$f_t(x) = \mu(x) + \sum_{k=1}^K \beta_{t,k} \phi_k(x) + e_t(x) \quad \text{for } t = 1, 2, \dots, N, \quad (15)$$

where  $\mu(x)$  is the mean function estimated by  $\mu(x) = \frac{1}{N} \sum_{t=1}^N S_t(x)$ ,  $\{\phi_1, \dots, \phi_K\}$  is the set of the first  $K$  functional principal components,  $\{\beta_{t,1}, \dots, \beta_{t,K}\}$  is the set of uncorrelated principal component scores for day  $t$ ,  $e_t(x)$  is the residual function with mean zero, and  $K$  the number of functional principal components used ( $K$  should be less than the number of nonzero eigenvalues of the (empirical) covariance operator). In this dissertation  $K = 3$  (number of hours per day) and  $N = 7302$  (number of days in twenty years).

**3.4.1.1 Prediction from FPCA.** The forecasting of  $h$ -step-ahead function  $f_{N+h}$  is achieved through forecasting the principal component scores  $\beta_{t,k}$  for  $k = 1, \dots, K$  and  $t = N+1, \dots, N+h$ . The forecast of the electricity load at time  $x$  is given by (as shown in Hyndman and Shahid Ullah, 2007)

$$Y_{N+h|T}(x) = E[Y_{N+h}(x) | I, \Phi] = \mu(x) + \sum_{k=1}^K \beta_{N+h|T,k} \phi_k(x), \quad (16)$$

where  $I = \{Y_1(x_i), \dots, Y_N(x_i)\}$  is the past curves,  $\Phi = \{\phi_1(t), \dots, \phi_K(t)\}$  the fixed functional principal components, and  $\beta_{N+h|N,k}$  denotes the  $h$ -step-ahead forecast of  $\beta_{t,k}$  using univariate time series (transfer function model in this dissertation).

**3.4.1.2 Prediction intervals from FPCA.** To build a prediction interval of the step-ahead value  $X_{N+h|N}(t)$ , we need to find the forecast variance. Since the functional principal components and the error term are orthogonal (and each component is orthogonal to the other components), the overall forecast variance can be approximated by the simple sum of component variances, and it is given by

$$\text{Var}[Y_{N+h}(x) | I, \Phi] \approx \sigma_\mu^2(x) + \sum_{k=1}^K u_{N+h|N,k} \phi_k^2(x) + \nu(x) + \sigma_{N+h}^2(x), \quad (17)$$



where  $\sigma_\mu^2(x)$  the variance of the mean function which can be obtained from the smoothing method used,  $u_{N+h|N,k}$  is the variance of the  $k^{\text{th}}$  principal component scores, obtained from the time series model, which defines as  $u_{N+h|N,k} = \text{Var}(\beta_{N+h,k} | \beta_{1,k}, \dots, \beta_{N,k})$ ,  $\phi_k^2(x)$  is the variance of the  $k^{\text{th}}$  functional principal component,  $v(x)$  is the variance of the model error which is estimated by averaging  $\hat{e}_t^2(x)$  for each  $x$ , and  $\sigma_{N+h}^2(x)$  is the variance or the observational error Hyndman and Shahid Ullah (2007).

The  $100(1-\alpha)\%$  point-wise prediction interval of  $Y_{N+h}(x)$  is given as

$$Y_{N+h|N}(x) \pm Z_\alpha \sqrt{\text{Var}[Y_{N+h}(x) | I, \Phi]}, \quad (18)$$

where  $Z_\alpha$  is the  $(1-\alpha/2)$  standard normal quantile, and assuming the various source of error are all normally distributed.

### 3.5 A BRIEF DESCRIPTION OF THE THREE MODELING APPROACHES

A description of the three approaches used to model the daily electricity load profile is given in the following. Results of implementing these approaches are given in Chapter 4.

**3.5.1 Short-Term Approach.** The first approach to modeling electricity load is suitable for short term forecasting and uses cubic splines to model the daily (24-hour) electricity use profile. The data set was separated into subsets, each representing one of the four seasons, namely winter, spring, summer, and fall. These subsets were further divided into data corresponding to weekdays and data corresponding to weekends. The hourly load data were modeled separately for each season by type of day combination, only after removal of the long term trend and the seasonal component.

The long-term trend was removed by first computing the hourly electricity load averaged over each of the 20 years in the training data set, and then using regression analysis to estimate the trend as a function of population and economic variables. The seasonal effects were then estimated from de-trended weekly data using a seasonal time

series approach. Details of the way the trend and seasonal variation was removed from the hourly data are given in Chapter 4.

An important point that should be stressed is that the removal of the seasonal variation only eliminated the effect of seasons on the overall average load level for a given season. It did not eliminate the differences in the shape of the daily load profiles across seasons. Each season potentially would still exhibit different load profiles.

Once the trend and the seasonal patterns in the load levels were removed from the hourly data, spline modeling was carried out. Each season by type of day combination employed the same number of knots but at different positions. Each of the above load models depend on temperature and since forecasts of temperature may not be reliable in the medium and long-term, these models are more appropriate for short term forecasting, such a few hours ahead or at most a few days ahead.

**3.5.2 Long-Term Approach.** This approach focuses on long term forecasting, again using cubic splines, but with the number of knots and their positions remaining the same irrespective of the season or the type of day. For each month in the sample, a separate spline model was fitted, resulting in a set of spline coefficient estimates particular for that month. These spline coefficients, which formed a vector, were estimated using hourly load data for all the days for a given month. The estimated spline model coefficients are time-varying and the vectors of these estimates were modeled as a multivariate time series using a VARX formulation.

The above approach was motivated by the publication Harvey and Koopman (1993), but differs significantly from the approach suggested by these authors. They suggest using a multivariate random walk formulation to model the time-varying behavior of the spline coefficients. They also suggest incorporating this behavior into a state-space model and then carrying out the estimation of all parameters using this model. First, from initial inspection of the spline estimates obtained from the month-by-month modeling of the data, it became clear that the spline coefficients do not behave in a random walk fashion. Second, for the average practitioner, the method proposed in this dissertation is easier to understand and implement, even though somewhat cumbersome. In addition, Harvey and Koopman modeled the seasonality using the state-space approach. In this dissertation, stochastic seasonality is modeled through seasonal lag

terms in the VARX model, which is easier to understand and control. In addition, Harvey and Koopman used trigonometric functions to model the seasonal effects in their example where data from Puget Sound Power and Light over approximately 70 months was modeled. The approach presented here does not use a deterministic cyclical model and used a model that allows stochastic seasonality. The time series of the estimated spline coefficients also indicate that the stochastic seasonality approach is better because no deterministic sinusoidal seasonality is apparent in the data.

**3.5.3 The FPCA approach.** The third approach employs FPCA to model detrended data. The resulting principal component scores were then modeled as time series using a transfer function formulation to forecast the principal components score for future days and use these scores to forecast the next year's load. The idea of using FPCA to model and forecast electricity load or demand data is not new. For example Shang (2013) introduced the FPCA methodology construct a model for forecasting very short-term electricity demand. Five principal components were used and a penalized least squares method was developed by the author for updating point forecast. Shang does not, however, incorporate exogenous variables, such as temperature, in modeling the principal component scores. The above author employs pure ARIMA processes to model and forecast the scores. The research presented herein, on the other hand, incorporates temperature, as well as dummy variables indicating weekends/weekdays when modeling the FPCA scores and employ the more flexible transfer functions to do so. Shang (2013) conducted the FPCA on each day of the week separately because of the belief that the demand profile is different for each day of the week. The approach taken in the research presented herein assumes that at least for the five weekdays, any changes in the profile will be reflected by different combination of FPCA scores. Shang also used bootstrap methods to build forecast intervals while less computationally intensive parametric methods are used in this research to construct prediction intervals. Moreover, the method proposed herein provides good load forecasts beyond short-term.

## 4. ANALYSIS AND EMPIRICAL RESULTS

Additional details of the three modeling approaches and the results obtained by fitting these models to the AE region of the PJM electricity market data are given in this chapter. In section 4.1, the results for the Short-Term Models are given. Section 4.2 contains results for the Long-Term Models and Section 4.3 focuses on the results for the FPCA approach.

Let the time series  $\{Y_t(x)\}$  denote the average electricity over one hour period ending at time  $x$  on day  $t$ , where  $t = 1, 2, \dots, N$  and  $x \in [0, 24]$ . When the load  $Y_t(x)$  is reported on the hour, say at hour 1, hour 2, etc., then one can represent the load by  $Y_t(x_i)$ ,  $i=1, 2, \dots, 24$ . The hourly load data from the AE zone in the PJM market is reported on the hour and as such, for simplicity of notation,  $Y_{t,i}$  can be used in place of  $Y_t(x_i)$ , with the latter expression used when a functional form for  $Y$  is needed, as in the case of FPCA.

### 4.1 SHORT TERM APPROACH

In this approach to modeling electricity load, it will be assumed that  $Y_{t,i}$  is a composite of structural components consisting of a long-term trend  $\tau_t$ , a seasonal component  $S_t$ , a weekly cycle  $w_t$ , a load function  $f(x_i)$ ,  $i=1, 2, \dots, 24$ , representing the hourly load, and an irregular stochastic component  $u_{t,i}$ . Thus  $Y_{t,i}$  can be expressed as

$$Y_{t,i} = \tau_t + S_t + w_t + f(x_i) + u_{t,i},$$

with  $t = 1, 2, \dots, N$  and  $i = 1, 2, \dots, 24$ . Each component in the above model was estimated separately.

The long-term trend was modeled using classical regression with select economic variables as regressors. The seasonal component was modeled using an ARMAX formulation with the average weekly temperature and its square as exogenous variables,

and the 24-hour load function was modeled by using a separate set of cubic splines for each season and for weekdays and weekends. Different spline models were used for each season because the 24-hour load within a season has almost the same pattern but differs across seasons. The weekdays were modeled separately within each season because they have quite different load patterns as well.

**4.1.1 Predicting Long-Term Trend.** The first step was modeling the hourly average electricity load per year,  $\tau_l^* = \frac{1}{N_l} \sum_{t,i}^{N_l} Y_{t,i}$ , using classical regression analysis. Note that in the above expression for the average load,  $l$  denotes the year with  $l = 1, 2, \dots, 20$ , and  $N_l$  denotes the total number of hours in that year. A stepwise selection method was used to determine the independent variables to be included in the model. Out of more than 20 economic plus population variables and the average monthly temperature, the following variables were selected: government employment in New Jersey (NJGOVTN), industrial production index in the US (IPI), home vacancy rate in New Jersey (NJHVAC), and the average temperature of September (Temp\_Sep). The following results were obtained by multiple linear regression analysis. Tables 4.1.a, 4.1.b, and 4.1.c show the results of ANOVA table, the model fit statistics, and the parameter estimates, respectively, of the regression model for the annual load.

Table 4.1. The Results for the Regression Model for Annual Load

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	4	117546	29386	310.41	<.0001
<b>Error</b>	15	1420.06099	94.67073		
<b>Corrected Total</b>	19	118966			

Table 4.1. The Results for the Regression Model for Annual Load (continued)

<b>Root MSE</b>	9.72989	<b>R-Square</b>	0.9881
<b>Dependent Mean</b>	1237.83958	<b>Adj R-Sq</b>	0.9849
<b>Coeff Var</b>	0.78604	<b>AIC</b>	95.2545

Table 4.1. The Results for the Regression Model for Annual Load (continued)

<b>Parameter Estimates</b>							
<b>Variable</b>	<b>Label</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>	<b>Variance Inflation</b>
<b>Intercept</b>	Intercept	<b>1</b>	-422.01120	91.16702	-4.63	0.0003	0
<b>NJGOVTN</b>	NJGOVTN	<b>1</b>	1.60946	0.10941	14.71	<.0001	2.42944
<b>NJHVAC</b>	NJHVAC	<b>1</b>	-36.77562	4.97535	-7.39	<.0001	1.04120
<b>IPI</b>	IPI	<b>1</b>	2.76858	0.34269	8.08	<.0001	2.39863
<b>Sep_M</b>	Temp_Sep	<b>1</b>	7.24667	1.27613	5.68	<.0001	1.07907

The regression model for the annual data is:

$$\hat{\tau}_i^* = -422.01 - 36.78 \text{ NJHVAC} + 1.61 \text{ NJGOVTN} + 2.77 \text{ IPI} + 7.25 \text{ Temp\_Sep}.$$

The selected independent variables explain 98.5% of the variation in the average annual load and the root mean square error (RMSE) is 9.7, which is small. Moreover, no serious multicollinearity among the independent variables was detected. The residual analysis is shown in the Figure 4.1, and there were not much of a concern because the model has performed a good job of predict the annual load.

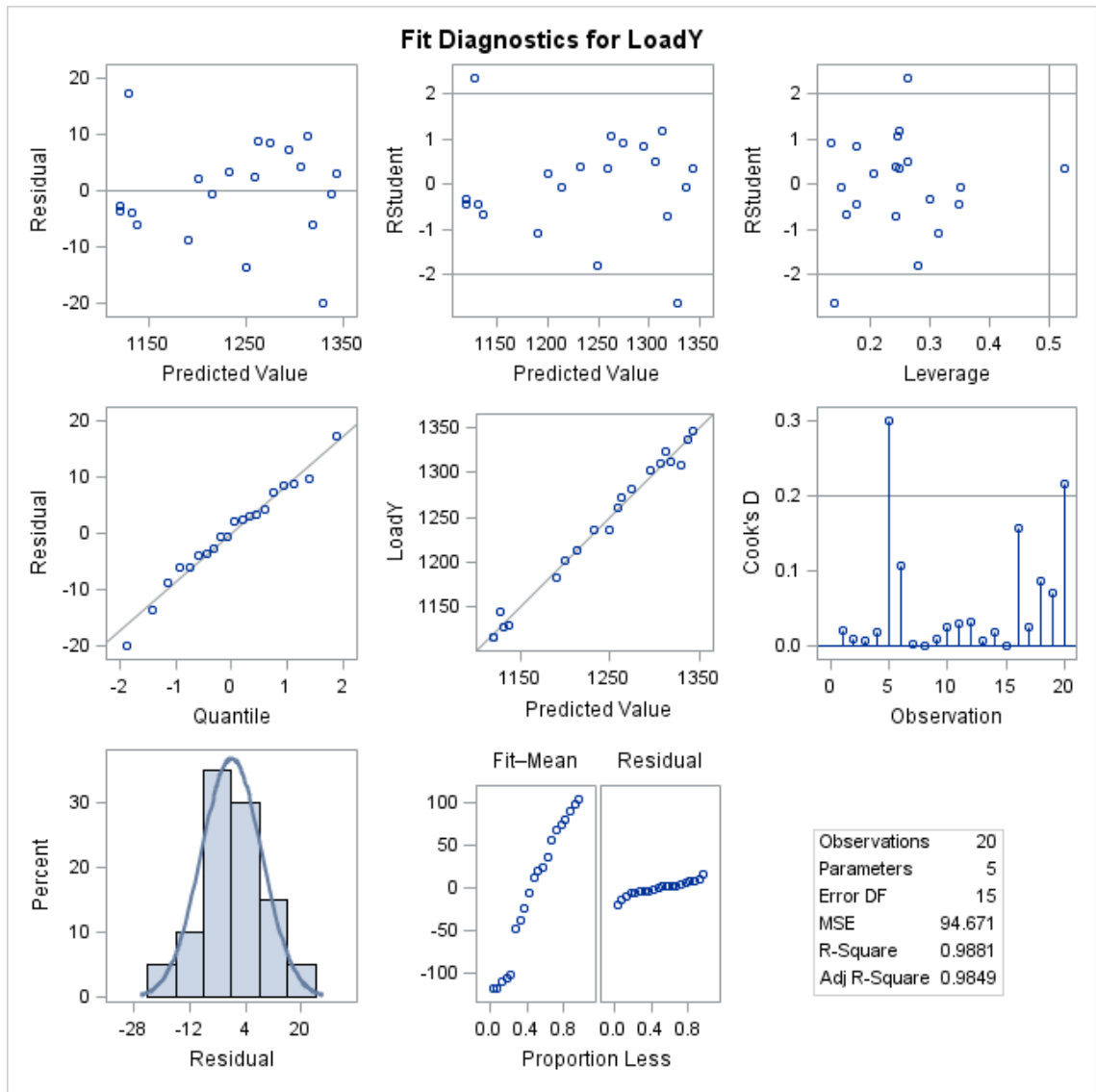


Figure 4.1. Residual Analysis of the Annual Regression Model

Figure 4.2 displays the average annual load (average is per hour) for the 20 years of training data and one year of test data and the average annual load predicted using the estimated regression model. The display shows very good in-sample agreement between the actual and predicted load and a reasonable agreement between the two for the test year.

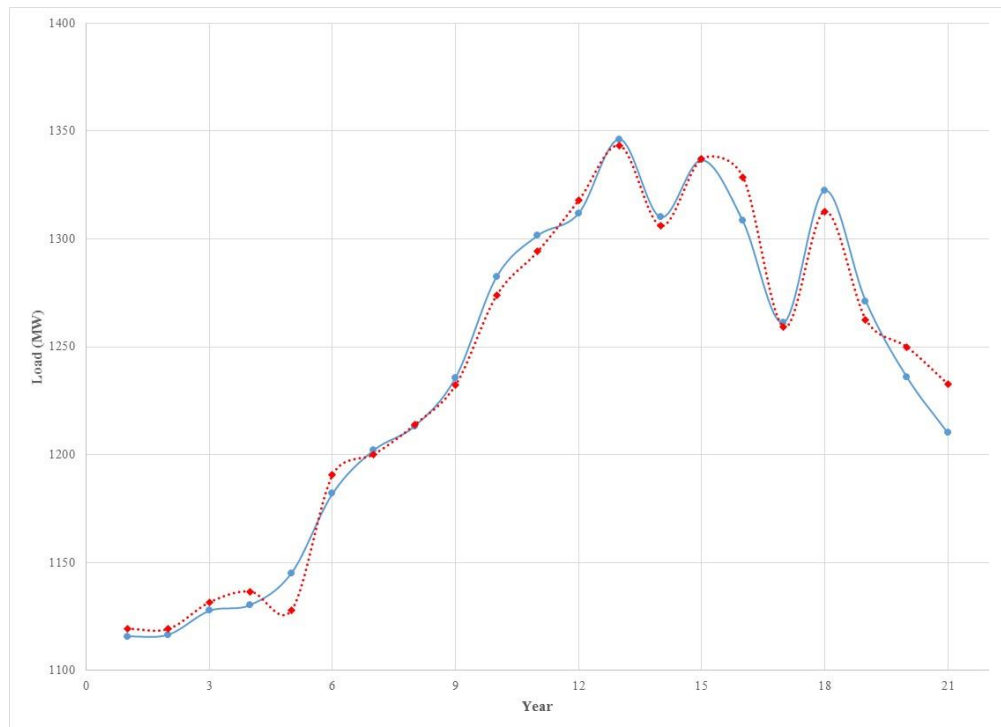


Figure 4.2. The Annual Average of Hourly Load (blue solid) and the Predicted Load (red dotted) 1993 - 2013

**4.1.2 Estimating Seasonal Variation in Data.** The trend estimates for each year were transposed onto a weekly series, and a 52-week moving average was applied to this series to smooth the predictions from a step function to a smooth one (see Figure A.1 in the appendix). The smoothed trend,  $\tilde{\tau}_w^*$ , for week  $w$ , was removed from weekly load data for that week by subtraction, with the process repeated for all weeks, resulting in a smoothed weekly series. The de-trended weekly time series was then modeled using the subset ARMAX model. The Tables 4.2.a and 4.2.b present the results of the ARMAX model and model fit statistics, respectively.



Table 4.2. The Results for the Weekly ARMAX Model

<b>Maximum Likelihood Estimation</b>					
<b>Parameter</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Approx Pr &gt;  t </b>	<b>Lag</b>
<b>MU</b>	1023.0	28.44590	35.96	<.0001	0
<b>MA1,1</b>	0.75131	0.05985	12.55	<.0001	1
<b>MA1,2</b>	-0.05634	0.01887	-2.99	0.0028	50
<b>MA1,3</b>	0.71204	0.03706	19.22	<.0001	52
<b>MA1,4</b>	-0.44034	0.04773	-9.22	<.0001	53
<b>MA1,5</b>	-0.10499	0.02477	-4.24	<.0001	54
<b>AR1,1</b>	1.12835	0.05735	19.67	<.0001	1
<b>AR1,2</b>	-0.22648	0.03295	-6.87	<.0001	2
<b>AR1,3</b>	0.76161	0.02714	28.06	<.0001	52
<b>AR1,4</b>	-0.67333	0.03045	-22.11	<.0001	53
<b>T</b>	-50.35659	0.95800	-52.56	<.0001	0
<b>T<sup>2</sup></b>	0.53344	0.0092772	57.50	<.0001	0

Table 4.2. The Results for the Weekly ARMAX Model (continued)

<b>Constant Estimate</b>	10.07384
<b>Std Error Estimate</b>	41.47282
<b>AIC</b>	10765.41
<b>SBC</b>	10824.81
<b>Number of Residuals</b>	1043

The ARMAX model for the weekly average of de-trended load is:

$$\hat{S}_w^* = 1023 + 1.13L_{(w-1)} - 0.23L_{(w-2)} + 0.76Z_{(w-52)} - 0.67Z_{(w-53)} + 0.75Z_{(w-1)} - 0.06Z_{(w-50)} \\ + 0.71Z_{(w-52)} - 0.44Z_{(w-53)} - 0.11Z_{(w-54)} - 50.36T + 0.53T^2,$$

where  $L_{(w-lag)}$  denotes the autoregressive lag term,  $Z_{(w-lag)}$  denotes the moving average lag terms, and  $T$  denotes the weekly average temperature. The residuals do not show any major autocorrelations and the test for white noise is shown in the Figure 4.3 on the bottom right hand corner shows no evidence that the residuals are anything other than white noise.

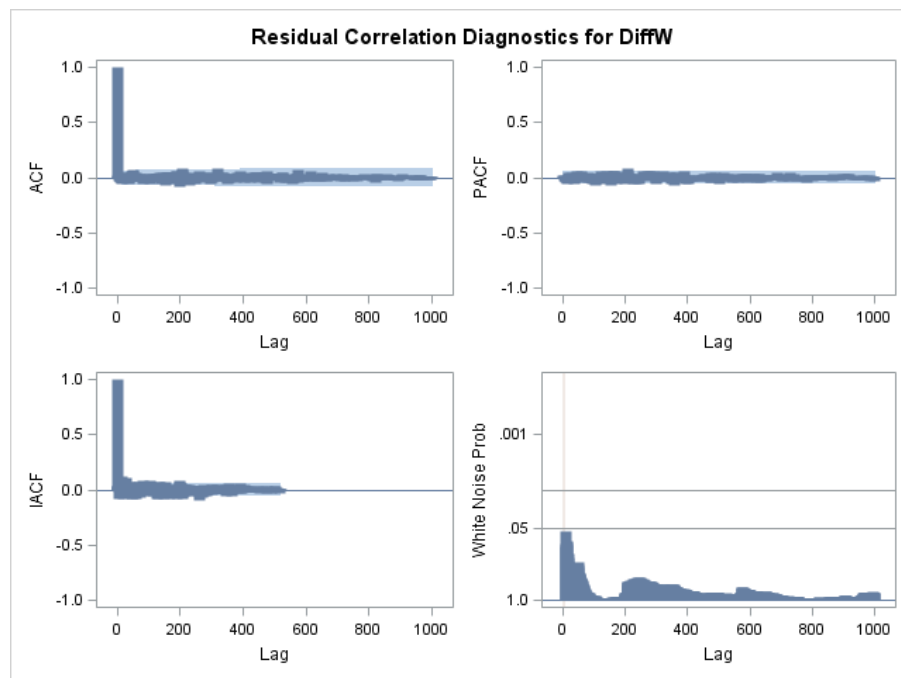


Figure 4.3. Analysis of Residual for the Weekly ARMAX Model.

The check for normality of the residuals given in Figure 4.4 show some deviation from normality, but this is not much of a concern because the model performed an adequate job of extracting the seasonal component as indicated by white noise residual.

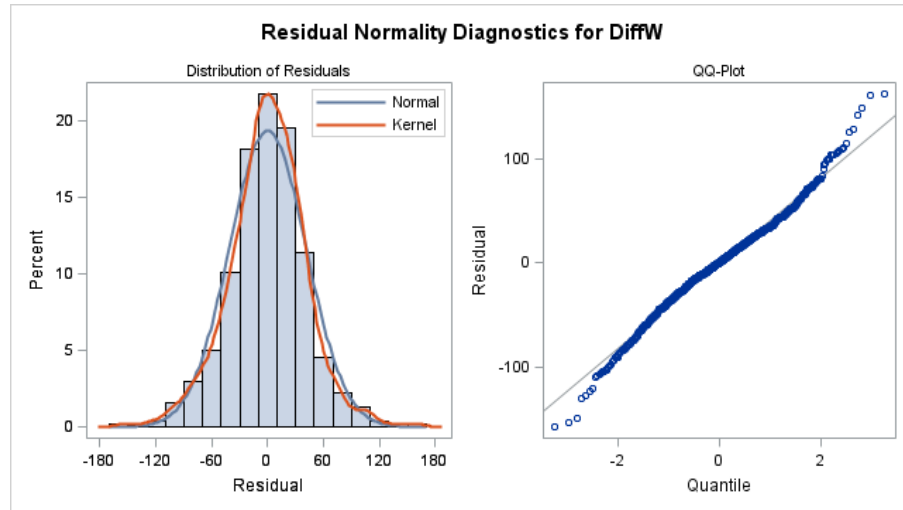


Figure 4.4. Normality Diagnostics of the Residuals of the Weekly ARMAX Model

Figure 4.5 shows the weekly average load for the weeks for the test year and the actual weekly averages. These out-of-sample checks show that the seasonal (weekly) model provides a satisfactory estimation of the seasonal component.

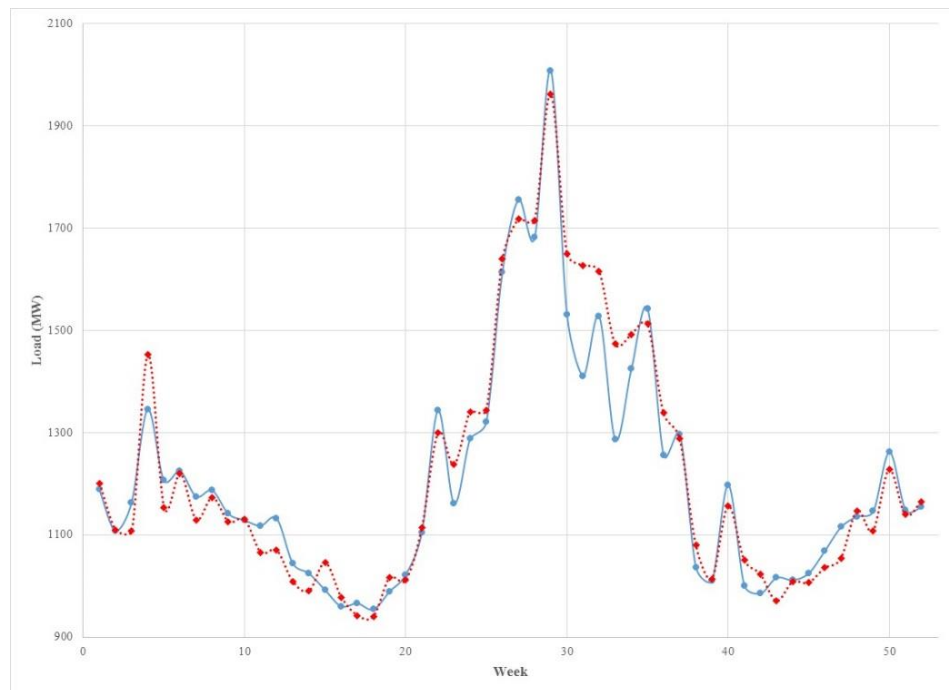


Figure 4.5. The Weekly Average of the Hourly Load (blue solid) and the Predicted Load (red dashed) 2013

**4.1.3 Modeling the Hourly Load.** At this point, the weekly smoothed trend  $\tilde{\tau}_w^*$  and the estimated seasonal component  $\hat{S}_w^*$  was removed from the hourly data  $Y_{t,i}$  and a new de-trended and de-seasonalized time series,  $Y_{t,i}^*$ , was obtained. The new times series was modeled using a cubic spline with different spline estimates obtained for each season, weekday, and weekend combination. Temperature and its interaction with spline coefficients were also fitted.

Two scenarios were applied here. The first one modeled each season and each day type separately. We denoted this as Model 1. The second scenario modeled each season separately and ignored the day type but added a dummy variable for the day type. This approach is denoted as Model 2.

**4.1.3.1 The Results of Model 1.** The general spline model used is:

$$Y_{t,i}^* = b_0 + b_1i + b_2i^2 + b_3i^3 + b_4(i - \kappa_1)^2 + b_5(i - \kappa_2)^2 + b_6(i - \kappa_3)^2 + b_7(i - \kappa_1)^3 + b_8(i - \kappa_2)^3 + b_9(i - \kappa_3)^3 + b_{10}T + b_{11}T * i + b_{12}T * i^2 + b_{13}T * i^3 + b_{14}T * (i - \kappa_1)^2 + b_{15}T * (i - \kappa_2)^2 + b_{16}T * (i - \kappa_3)^2 + b_{17}T * (i - \kappa_1)^3 + b_{18}T * (i - \kappa_2)^3 + b_{19}T * (i - \kappa_3)^3 ,$$

where  $i$  is the hour,  $\kappa$ 's are the knots that change according to season, and  $T$  is temperature. The following tables present the regression model results for each season. The tables 4.3, 4.4, 4.5, and 4.6 present the results of the regression model for each season and each type of day. The tables for weekdays and weekends for a given season are paired together for easy comparison.

Table 4.3. The Results for the Regression Model for the Winter Season

Winter - Weekdays			Winter - Weekends		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	-43.94700	0.0003	Intercept	33.16159	0.0438
$h$	43.00126	<.0001	$h$	-68.74540	<.0001
$h^2$	-28.47236	<.0001	$h^2$	10.86316	0.0456
$h^3$	3.90525	<.0001	$h^3$	-0.34335	0.5244
$(h - 6)^2$	-72.28854	<.0001	$(h - 5)^2$	4.53027	0.1994
$(h - 14)^2$	-42.29815	<.0001	$(h - 12)^2$	21.97808	<.0001
$(h - 17)^2$	-103.8987	<.0001	$(h - 17)^2$	-42.63870	<.0001
$(h - 6)^3$	-1.95948	<.0001	$(h - 5)^3$	-0.98781	0.0461
$(h - 14)^3$	8.92152	<.0001	$(h - 12)^3$	2.63976	<.0001
$(h - 17)^3$	-9.58272	<.0001	$(h - 17)^3$	-0.96957	0.0194
temp	-2.89175	<.0001	temp	-2.59769	<.0001
$T^*h$	-0.71365	<.0001	$T^*h$	-0.15635	0.0027
$T^*h^2$	0.17868	<.0001	$T^*(h - 5)^2$	0.06344	<.0001
$T^*h^3$	-0.01462	<.0001	$T^*(h - 12)^2$	-0.56360	<.0001
$T^*(h - 6)^2$	0.26047	<.0001	$T^*(h - 17)^2$	-0.67866	0.0005
$T^*(h - 14)^2$	0.23657	<.0001	$T^*(h - 12)^3$	0.06180	<.0001
$T^*(h - 17)^3$	0.02458	<.0001	$T^*(h - 17)^3$	-0.03642	0.0005

<b>RMSE</b>	72.9606
<b>Adj-R2</b>	0.7671
<b>AIC</b>	260089.07

<b>RMSE</b>	74.3877
<b>Adj-R2</b>	0.7065
<b>AIC</b>	100750.97

Table 4.4. The Results for the Regression Model for the Spring Season

Spring - Weekdays			Spring - Weekends		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	93.65097	<.0001	Intercept	-7.13119	0.6187
$h$	-150.22285	<.0001	$h$	-48.84602	<.0001
$h^2$	68.15250	<.0001	$h^2$	13.06001	<.0001
$h^3$	-8.97118	<.0001	$h^3$	-0.74388	<.0001
$(h - 4)^2$	115.05545	<.0001	$(h - 9)^2$	-13.54758	0.0001
$(h - 8)^2$	61.51265	<.0001	$(h - 18)^2$	124.02630	0.0021
$(h - 19)^2$	-124.08785	<.0001	$(h - 20)^2$	231.54827	0.0011
$(h - 4)^3$	-3.47963	0.0009	$(h - 9)^3$	2.62643	<.0001
$(h - 8)^3$	13.74708	<.0001	$(h - 18)^3$	-66.42324	<.0001
$(h - 19)^3$	9.48140	<.0001	$(h - 20)^3$	65.77067	<.0001
temp	-2.50879	<.0001	temp	-1.45289	<.0001
$T^* h^2$	-0.41258	<.0001	$T^* h^2$	-0.22542	<.0001
$T^* h^3$	0.07054	0.0001	$T^* h^3$	0.02324	<.0001
$T^* (h - 4)^2$	-1.05263	<.0001	$T^* (h - 9)^2$	-0.43314	<.0001
$T^* (h - 8)^2$	-1.02330	<.0001	$T^* (h - 18)^2$	-2.77329	<.0001
$T^* (h - 19)^2$	1.75668	<.0001	$T^* (h - 20)^2$	-5.14219	<.0001
$T^* (h - 4)^3$	0.06608	<.0001	$T^* (h - 9)^3$	-0.03102	<.0001
$T^* (h - 8)^3$	-0.14706	<.0001	$T^* (h - 18)^3$	1.26328	<.0001
$T^* (h - 19)^3$	-0.19409	<.0001	$T^* (h - 20)^3$	-1.22318	<.0001

<b>RMSE</b>	73.5242
<b>Adj-R2</b>	0.7496
<b>AIC</b>	271284.41

<b>RMSE</b>	84.0868
<b>Adj-R2</b>	0.6247
<b>AIC</b>	111701.59

Table 4.5. The Results for the Regression Model for the Summer Season

Summer - Weekdays			Summer - Weekends		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	-1031.8216	<.0001	Intercept	-1121.8108	<.0001
$h$	20.83889	0.6311	$h$	92.24404	0.0765
$h^2$	23.34519	0.0028	$h^2$	-4.50357	0.5201
$h^3$	-2.15146	<.0001	$h^3$	0.23011	0.4594
$(h - 9)^2$	-58.35710	<.0001	$(h - 7)^2$	-50.56548	0.0002
$(h - 14)^2$	-52.55395	0.0212	$(h - 12)^2$	79.89184	<.0001
$(h - 19)^2$	-55.43427	0.1113	$(h - 20)^2$	-45.09066	0.0476
$(h - 9)^3$	13.63412	<.0001	$(h - 7)^3$	0.75179	0.4427
$(h - 14)^3$	-10.04996	<.0001	$(h - 12)^3$	-1.41147	0.3334
$(h - 19)^3$	-3.42800	<.0001	$(h - 20)^3$	-5.65133	<.0001
temp	12.30673	<.0001	temp	13.92763	<.0001
$T^*h$	-1.67881	0.0056	$T^*h$	-2.61695	<.0001
$T^*h^2$	-0.23403	0.0326	$T^*h^2$	0.15085	0.0283
$T^*h^3$	0.03886	<.0001	$T^*(h - 7)^2$	1.43653	<.0001
$T^*(h - 14)^2$	1.07409	0.0002	$T^*(h - 20)^2$	0.72256	0.0058
$T^*(h - 19)^2$	1.20714	0.0053	$T^*(h - 7)^3$	-0.14516	<.0001
$T^*(h - 9)^3$	-0.17607	<.0001	$T^*(h - 12)^3$	0.15275	<.0001
$T^*(h - 14)^3$	0.09527	<.0001			

<b>RMSE</b>	121.682
<b>Adj-R2</b>	0.8661
<b>AIC</b>	303082.91

<b>RMSE</b>	138.2348
<b>Adj-R2</b>	0.8152
<b>AIC</b>	124226.62

Table 4.6. The Results for the Regression Model for the Fall Season

Fall - Weekdays			Fall - Weekends		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	-124.73508	<.0001	Intercept	-21.40477	0.3987
$h$	49.54582	0.0001	$h$	-27.84008	0.0282
$h^2$	-20.69335	<.0001	$h^2$	7.44891	0.0016
$h^3$	2.88465	<.0001	$h^3$	-0.32170	0.0803
$(h - 6)^2$	-62.12202	<.0001	$(h - 8)^2$	-14.30060	0.3538
$(h - 15)^2$	106.77851	<.0001	$(h - 10)^2$	5.40145	0.5837
$(h - 20)^2$	190.28792	<.0001	$(h - 18)^2$	-147.68332	<.0001
$(h - 6)^3$	-1.41533	<.0001	$(h - 8)^3$	-1.67366	0.6195
$(h - 15)^3$	-17.40696	<.0001	$(h - 10)^3$	5.28608	0.1368
$(h - 20)^3$	4.95084	<.0001	$(h - 18)^3$	5.18245	<.0001
temp	-0.55227	0.0901	temp	-1.22595	0.0026
$T^* h$	-0.95896	<.0001	$T^* h$	-0.87935	<.0001
$T^* h^2$	0.05788	0.0049	$T^* h^2$	0.00681	<.0001
$T^* (h - 6)^2$	0.31932	<.0001	$T^* (h - 8)^2$	0.68040	0.0036
$T^*(h - 15)^2$	-1.29167	<.0001	$T^*(h - 18)^2$	1.72536	<.0001
$T^*(h - 20)^2$	-3.25219	<.0001	$T^*(h - 8)^3$	-0.14824	0.0002
$T^* (h - 6)^3$	-0.02245	<.0001	$T^*(h - 10)^3$	0.11613	0.0105
$T^*(h - 15)^3$	0.23814	<.0001	$T^*(h - 18)^3$	-0.10069	<.0001

<b>RMSE</b>	92.4018
<b>Adj-R2</b>	0.7446
<b>AIC</b>	282449.54

<b>RMSE</b>	100.7796
<b>Adj-R2</b>	0.6528
<b>AIC</b>	115156.88



**4.1.3.2 The Results Model 2.** The general spline model is:

$$\begin{aligned}
 Y_{t,i}^* = & b_0 + b_1i + b_2i^2 + b_3i^3 + b_4(i - \kappa_1)^2 + b_5(i - \kappa_2)^2 + b_6(i - \kappa_3)^2 + b_7(i - \kappa_1)^3 + b_8(i - \kappa_2)^3 \\
 & + b_9(i - \kappa_3)^3 + b_{10}T + b_{11}T * i + b_{12}T * i^2 + b_{13}T * i^3 + b_{14}T * (i - \kappa_1)^2 + b_{15}T * (i - \kappa_2)^2 \\
 & + b_{16}T * (i - \kappa_3)^2 + b_{17}T * (i - \kappa_1)^3 + b_{18}T * (i - \kappa_2)^3 + b_{19}T * (i - \kappa_3)^3 + w,
 \end{aligned}$$

where  $h$  is an hour,  $\kappa$  is a knot that changes according to the season,  $T$  is temperature, and  $w$  is a weekend dummy variable. The Tables 4.7 and 4.8 present the regression model results for each season. Note that non-significant terms were dropped from the models and the results are for the reduced models.

Table 4.7. The Results for the Regression Model for the Winter and Spring Seasons

Winter			Spring		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	-16.55248	0.1207	Intercept	-28.71512	0.3281
$h$	23.25116	0.0014	$h$	43.44180	0.1746
$h^2$	-20.66445	<.0001	$h^2$	-18.69310	0.0568
$h^3$	2.96367	<.0001	$h^3$	2.56499	0.0042
$(h - 6)^2$	-52.98180	<.0001	$(h - 5)^2$	17.48659	<.0001
$(h - 15)^2$	-16.66555	<.0001	$(h - 8)^2$	42.40157	<.0001
$(h - 17)^2$	-112.22096	<.0001	$(h - 19)^2$	-123.9761	<.0001
$(h - 6)^3$	-1.83272	<.0001	$(h - 5)^3$	-13.11806	<.0001
$(h - 15)^3$	13.03173	<.0001	$(h - 8)^3$	11.98398	<.0001
$(h - 17)^3$	-12.65773	<.0001	$(h - 19)^3$	9.16725	<.0001
temp	-2.76643	<.0001	temp	-0.64997	0.2414
$T^* h$	-0.64320	<.0001	$T^* h$	-2.15721	0.0003
$T^* h^2$	0.13092	<.0001	$T^* h^2$	0.50905	0.0049
$T^* h^3$	-0.00834	<.0001	$T^* h^3$	-0.04807	0.0027
$T^*(h - 6)^2$	0.12037	0.0070	$T^*(h - 8)^2$	-0.96311	<.0001
$T^*(h - 15)^2$	0.13705	0.0009	$T^*(h - 19)^2$	1.72975	<.0001
$T^*(h - 17)^3$	0.01452	<.0001	$T^*(h - 5)^3$	0.18007	<.0001
Weekend	-49.01851	<.0001	$T^*(h - 8)^3$	-0.14305	<.0001
			$T^*(h - 19)^3$	-0.18822	<.0001
			Weekend	-59.90003	<.0001

<b>RMSE</b>	75.7805
<b>Adj-R2</b>	0.7410
<b>AIC</b>	363556.63

<b>RMSE</b>	79.6872
<b>Adj-R2</b>	0.7054
<b>AIC</b>	386694.58

Table 4.8. The Results for the Regression Model for the Summer and Fall Seasons

Summer			Fall		
Variable	Parameter Estimate	Pr >  t	Variable	Parameter Estimate	Pr >  t
Intercept	-1131.9170	<.0001	Intercept	-84.22522	<.0001
$h$	115.72585	<.0001	$h$	40.45021	0.0004
$h^2$	1.12711	0.2755	$h^2$	-17.21285	<.0001
$h^3$	-0.77157	<.0001	$h^3$	2.41275	<.0001
$(h - 9)^2$	-83.52495	<.0001	$(h - 6)^2$	-51.38721	<.0001
$(h - 14)^2$	-46.86818	0.0578	$(h - 15)^2$	120.69047	<.0001
$(h - 19)^2$	-45.76871	0.1566	$(h - 20)^2$	205.64318	<.0001
$(h - 9)^3$	13.02938	<.0001	$(h - 6)^3$	-1.34485	<.0001
$(h - 14)^3$	-11.66866	<.0001	$(h - 15)^3$	-18.15278	<.0001
$(h - 19)^3$	-3.20285	<.0001	$(h - 20)^3$	4.97629	<.0001
temp	13.81380	<.0001	temp	-0.70591	0.0145
$T^*h$	-2.73699	<.0001	$T^*h$	-0.99218	<.0001
$T^*h^3$	0.02435	<.0001	$T^*h^2$	0.06113	0.0008
$T^*(h - 9)^2$	0.30996	0.0564	$T^*(h - 6)^2$	0.30423	<.0001
$T^*(h - 14)^2$	1.08035	0.0008	$T^*(h - 15)^2$	-1.37326	<.0001
$T^*(h - 19)^2$	1.11159	0.0059	$T^*(h - 20)^2$	-3.36051	<.0001
$T^*(h - 9)^3$	-0.17592	<.0001	$T^*(h - 6)^3$	-0.02131	<.0001
$T^*(h - 14)^3$	0.11706	<.0001	$T^*(h - 15)^3$	0.24565	<.0001
Weekend	-55.87929	<.0001	Weekend	-71.82423	<.0001

<b>RMSE</b>	128.1746
<b>Adj-R2</b>	0.8496
<b>AIC</b>	428670.67

<b>RMSE</b>	97.1472
<b>Adj-R2</b>	0.7165
<b>AIC</b>	399798.26

As seen from the above results, the models for each season is different from the others, reflecting changes in the daily load profiles across seasons. The comparison between the two models (Model 1 which fitted separate regression models to weekdays and weekends and Model 2 which used a dummy variable to account for differences between weekdays and weekends) based on Akaike Information Criteria (*AIC*) and Root Mean Square Error (*RMSE*) is presented in Table 4.8. In addition, the Figures 4.6 and 4.7 show the comparison between the two models based on the Coefficient of Variation (*CV*) for each month and each hour, respectively.

Table 4.9. The Comparison between the Two Models

Season	Day Type	RMSE		AIC	
		Model 1	Model2	Model 1	Model2
Winter	Weekdays	72.9606	72.9968	260089.07	260119.13
	Weekends	74.3945	74.7430	100751.11	100862.34
Spring	Weekdays	73.5242	73.4836	271284.41	271249.53
	Weekends	84.0868	84.2725	111701.59	111757.18
Summer	Weekdays	121.682	121.754	303082.91	303118.57
	Weekends	138.232	138.465	124225.17	124267.61
Fall	Weekdays	94.4033	92.4018	282449.53	282449.54
	Weekends	100.777	101.231	115155.18	115268.54

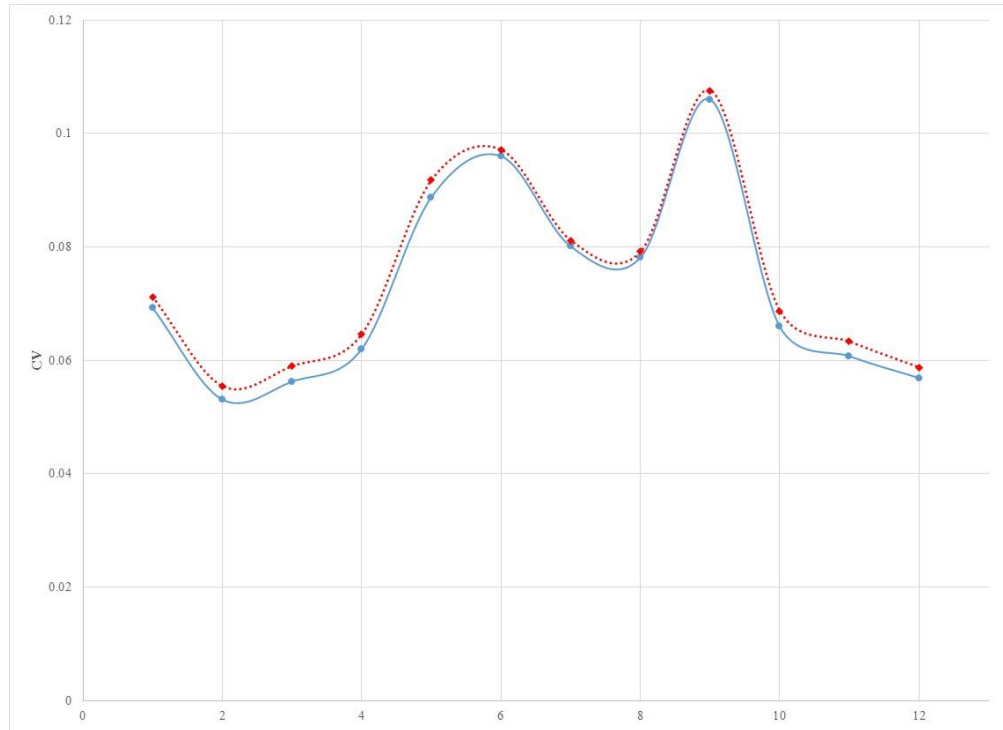


Figure 4.6. The Comparison between the *CV* of the Predicted Monthly Average Load of Model 1 (blue solid) and Model 2 (red dotted).

Figure 4.6 shows very close agreement between the two models when compared using the *CV* by month. There is a slight drop in the *CV* for Model 1 suggesting a slight gain in accuracy when the weekdays and weekends are modeled separately. Figure 4.7 given below shows the *CV* for the two models by each hour of the day. Again, Model 1 shows a slight advantage with the *CV* for Model 2 showing higher values for hours below 10 am. This may be because the load profile for weekends shows a two-hour shift in the morning and the use of the dummy variable is not sufficient to account for this difference in the shape of the load profile.

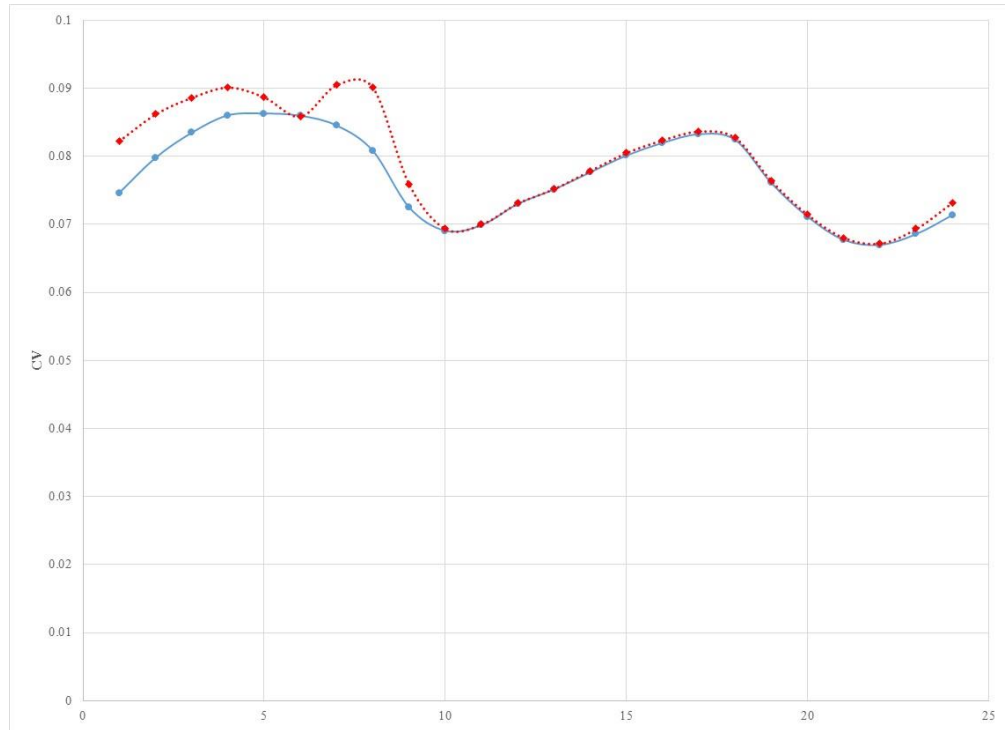


Figure 4.7. The Comparison between the CV of the Predicted Hourly Average Load of Model 1 (blue solid) and Model 2 (red dotted).

The Figures 4.8, 4.9, 4.10, and 4.11 provide a comparison between the two models with the actual load of the test year for four different weeks. Each figure shows a week from the middle of each season. The forecasts based on each model are very close to one another, which suggests that adding a dummy variable for the day type instead of building the extra models for the type of day provides satisfactory forecast overall, but for all seasons except summer, Model 2 yields forecasts that fall below the actual load during the weekends (last two days in the graph), especially in the morning period. However, except for the weekends, both models underestimate the afternoon peak in winter and spring. For the summer season (figure 4.10), the afternoon peak is overestimated by both models on Fridays.

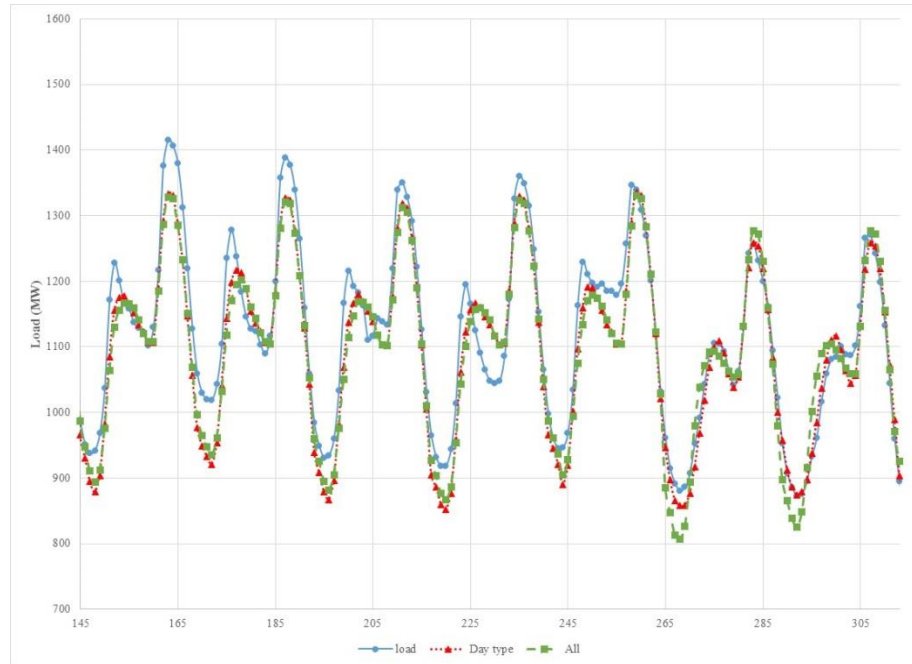


Figure 4.8. The Comparison between the Actual Hourly Load (blue solid), the Predicted Hourly Load of Model 1 (red dotted), and Model 2 (green dashed) – a Week in the Mid-Winter Season

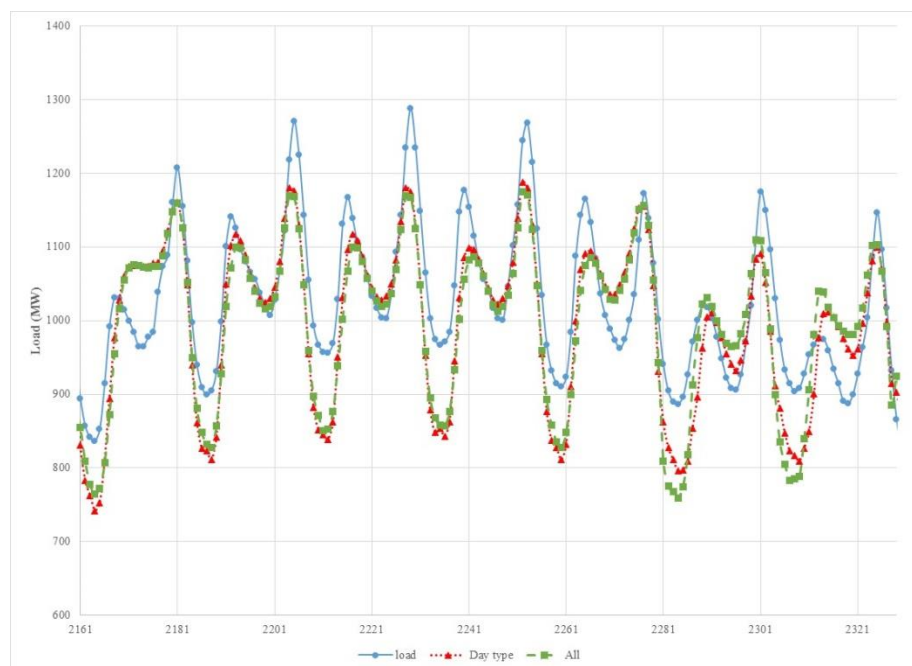


Figure 4.9. The Comparison between the Actual Hourly Load (blue solid), the Predicted Hourly Load of Model 1 (red dotted), and Model 2 (green dashed) – a Week in the Mid-Spring Season

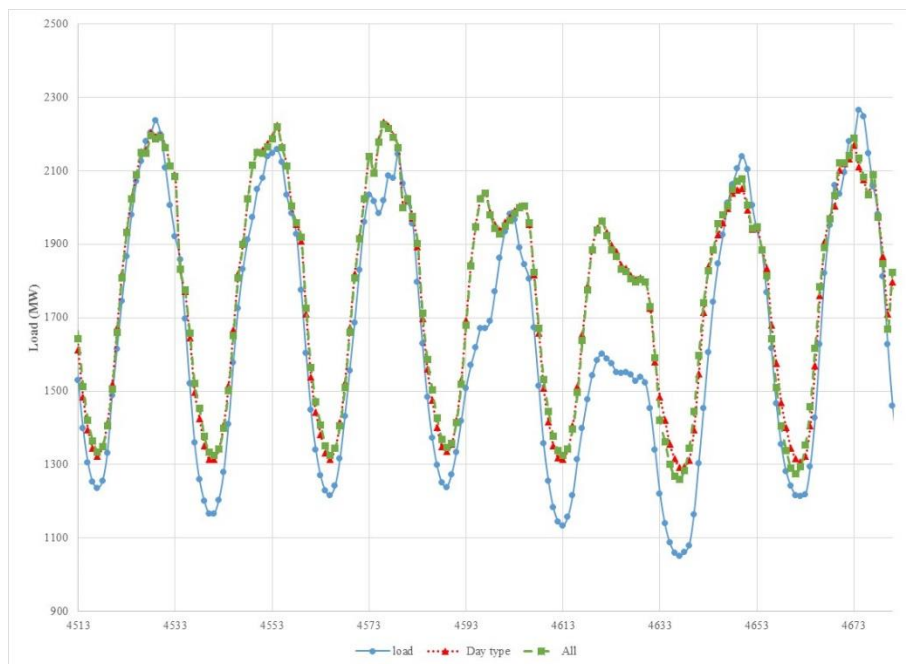


Figure 4.10. The Comparison between the Actual Hourly Load (blue solid), the Predicted Hourly Load of Model 1 (red dotted), and Model 2 (green dashed) – a Week in the Mid-Summer Season

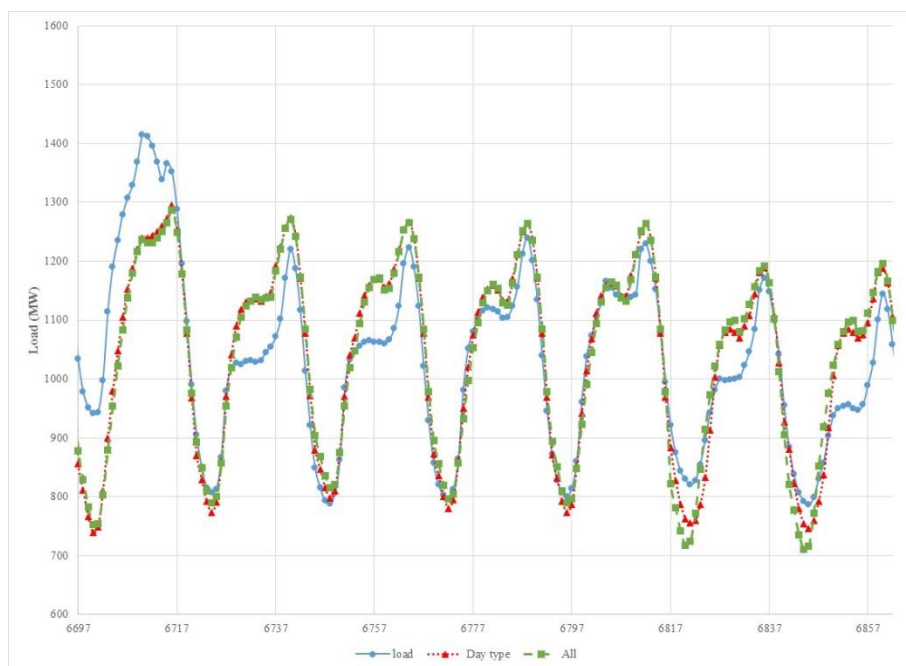


Figure 4.11. The Comparison between the Actual Hourly Load (blue solid), the Predicted Hourly Load of Model 1 (red dotted), and Model 2 (green dashed) – a Week in the Mid-Fall Season



Table 4.10. The Comparison between the CV of the Two Models for Each Season per Hour

Hour	Winter		Spring		Summer		Fall	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
1	0.0640	0.0678	0.0774	0.0818	0.0874	0.0899	0.0825	0.0921
2	0.0610	0.0634	0.0764	0.0798	0.0925	0.0938	0.0770	0.0856
3	0.0632	0.0648	0.0753	0.0798	0.0931	0.0930	0.0777	0.0861
4	0.0654	0.0666	0.0786	0.0795	0.0926	0.0910	0.0810	0.0881
5	0.0627	0.0631	0.0771	0.0773	0.0908	0.0892	0.0813	0.0854
6	0.0590	0.0594	0.0774	0.0785	0.0868	0.0877	0.0807	0.0809
7	0.0655	0.0747	0.0866	0.0934	0.0910	0.0924	0.0881	0.0928
8	0.0619	0.0764	0.0763	0.0897	0.0964	0.0902	0.0879	0.0945
9	0.0556	0.0642	0.0719	0.0804	0.1043	0.0978	0.0875	0.0877
10	0.0630	0.0658	0.0777	0.0814	0.1037	0.1060	0.0921	0.0910
11	0.0715	0.0712	0.0827	0.0837	0.0997	0.1009	0.0959	0.0970
12	0.0759	0.0744	0.0901	0.0886	0.1021	0.1018	0.1050	0.1072
13	0.0758	0.0759	0.0987	0.0964	0.1066	0.1044	0.1185	0.1201
14	0.0734	0.0746	0.1047	0.1036	0.1013	0.0997	0.1319	0.1322
15	0.0707	0.0723	0.1066	0.1071	0.0957	0.0951	0.1386	0.1382
16	0.0686	0.0691	0.1103	0.1114	0.0928	0.0929	0.1436	0.1448
17	0.0682	0.0695	0.1144	0.1147	0.0897	0.0906	0.1496	0.1504
18	0.0716	0.0731	0.1181	0.1167	0.0869	0.0862	0.1562	0.1550
19	0.0603	0.0616	0.1155	0.1153	0.0851	0.0854	0.1341	0.1349
20	0.0587	0.0603	0.1058	0.1049	0.0829	0.0826	0.1158	0.1172
21	0.0611	0.0619	0.0986	0.0974	0.0789	0.0781	0.1070	0.1094
22	0.0615	0.0619	0.0895	0.0849	0.0795	0.0775	0.0964	0.0982
23	0.0602	0.0603	0.0797	0.0743	0.0917	0.0884	0.0860	0.0884
24	0.0603	0.0601	0.0775	0.0754	0.0854	0.0836	0.0810	0.0858

## 4.2 LONG TERM APPROACH

The long-term model was developed to predict beyond a few hours or days, but the ability to detect short-term variation from day to day was sacrificed. In short, the long-term model predicts the average daily load profile for a whole month rather than for a specific day. The model uses a unique cubic spline curve for every month over the 20 years. The number of knots and the position of knots were kept constant from month to month, but the knot positions were selected through a trial and error process. This process used *AIC* and *RMSE*. Each set of knot positions yielded a set of 240 *AIC* values and 240 *RMSE* values. One set of knot positions were compared to another set of knot positions and the set that had a majority of *AIC* and *RMSE* values lower than the other set was selected. The process was repeated with another choice of knot positions in a recursive manner. In the future a more exhaustive and automated process could be implemented.

Because the estimation of the spline coefficients was done separately for each month, the estimated coefficients varied from month to month. Examination of these estimates showed that the coefficient themselves are slowly varying time series. This provided a way to forecast spline coefficients for future months through time series modeling of the estimated coefficients. In general, this approach is similar to that proposed by Harvey and Koopman (1993), but differs in the fact that what is proposed in this dissertation is a long-term forecast model that predicts average load profile for a month at a time, and uses a two-step approach where splines coefficients are estimated separately for each month and modeling the resulting coefficient estimates in a subsequent step. This approach allows the estimation of the spline coefficients using simple regression models and enables the fitting of more complicated time series models without introducing too much complexity into the likelihood function.

Since weekdays and weekends exhibit different load profiles, two load models were estimated for each month; one for weekdays and the other for weekends. Since the splines models are fitted separately for each month, one can use the load data  $\{Y_{t,i} \mid t \in \text{Month } m\}$ . Instead of using 24-hour load data for each day, the average load for each hour  $i$  over the month  $m$  was used. These average loads will be denoted by  $Y_{m,i}$ .

**4.2.1 Weekdays and Weekends Model.** Since the 24-hours curve differs between weekdays and weekends, a specific model was used for each day type. The weekdays' spline model employed is:

$$Y_{m,i} = b_{0,m} + b_{1,m}i + b_{2,m}i^2 + b_{3,m}i^3 + b_{4,m}(i-7)^2 + b_{5,m}(i-15)^2 + b_6(i-17)^2 + b_{7,m}(i-7)^3 + b_{8,m}(i-15)^3 + b_{9,m}(i-17)^3,$$

for  $i = 1, 2, \dots, 24$  and  $m = 1, 2, \dots, 12$ .

The weekends' spline model is:

$$Y_{m,i} = b_{0,m} + b_{1,m}i + b_{2,m}i^2 + b_{3,m}i^3 + b_{4,m}(i-9)^2 + b_{5,m}(i-16)^2 + b_6(i-18)^2 + b_{7,m}(i-9)^3 + b_{8,m}(i-16)^3 + b_{9,m}(i-18)^3,$$

with  $i$  and  $m$  defined as before.

For each of the above models, the ten model coefficients were estimated for each month. The ten coefficients for each model were treated as a vector autoregressive process. When modeling this vector time series, temperature and squared temperature were added as exogenous variables. The SAS<sup>®</sup> procedure VARX was employed to model each of these vector time series.

In addition to the separate models for weekdays and weekends, a combined model was also estimated using all seven days of the week. The estimated coefficients from this combined analysis were also modeled using the VARX procedure.

**4.2.2 Combined Model.** In this approach, the following general spline model was fitted.

$$L_{m,i} = b_{0,m} + b_{1,m}i + b_{2,m}i^2 + b_{3,m}i^3 + b_{4,m}(i-7)^2 + b_{5,m}(i-16)^2 + b_6(i-18)^2 + b_{7,m}(i-7)^3 + b_{8,m}(i-16)^3 + b_{9,m}(i-18)^3,$$

for  $i = 1, 2, \dots, 24$  and  $m = 1, 2, \dots, 12$ .

**4.2.3 Modeling Hourly Temperature.** Since the VARX models employed to forecast the spline coefficients contain average monthly temperature as an exogenous variable, use of these models to predict the coefficients for a future month requires the future values for temperature. Thus a model for predicting the average monthly temperature would be needed. The ARMA subset model was used to forecast the monthly average temperature. The following table presents the results.

Table 4.11. The Results for the Monthly Temperature ARMA Model

<b>Maximum Likelihood Estimation</b>					
<b>Parameter</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Approx Pr &gt;  t </b>	<b>Lag</b>
<b>MU</b>	54.76245	0.22485	243.55	<.0001	0
<b>MA1,1</b>	0.21638	0.05905	3.66	0.0002	1
<b>MA1,2</b>	-0.48601	0.05205	-9.34	<.0001	5
<b>MA1,3</b>	0.22560	0.05731	3.94	<.0001	12
<b>AR1,1</b>	0.48979	0.0075249	65.09	<.0001	1
<b>AR1,2</b>	-0.49774	0.0042286	-117.71	<.0001	5
<b>AR1,3</b>	0.14351	0.0075218	19.08	<.0001	12

Table 4.11. The Results for the Monthly Temperature ARMA Model (continued)

<b>Constant Estimate</b>	47.33881
<b>Variance Estimate</b>	8.580602
<b>Std Error Estimate</b>	2.929267
<b>AIC</b>	1214.316
<b>SBC</b>	1238.68
<b>Number of Residuals</b>	240

The ARMA model the average monthly temperature is:

$$T_m = 54.7625 + 0.49T_{(m-1)} - 0.50T_{(m-5)} + 0.14T_{(m-12)} + 0.22Z_{(m-1)} - 0.49Z_{(m-5)} + 0.23Z_{(m-12)},$$

where  $T_{(m-lag)}$  denotes the autoregressive lag term and  $Z_{(w-lag)}$  denotes the moving average terms. The residual analysis shows that the estimated model extracted almost all the information about the monthly average temperature present in the data because of the residuals passed the test for white noise. The residuals do not show any major autocorrelations and the test for white noise shown in the Figure 4.12 on the bottom right hand corner shows no evidence that the residuals are anything other than white noise.

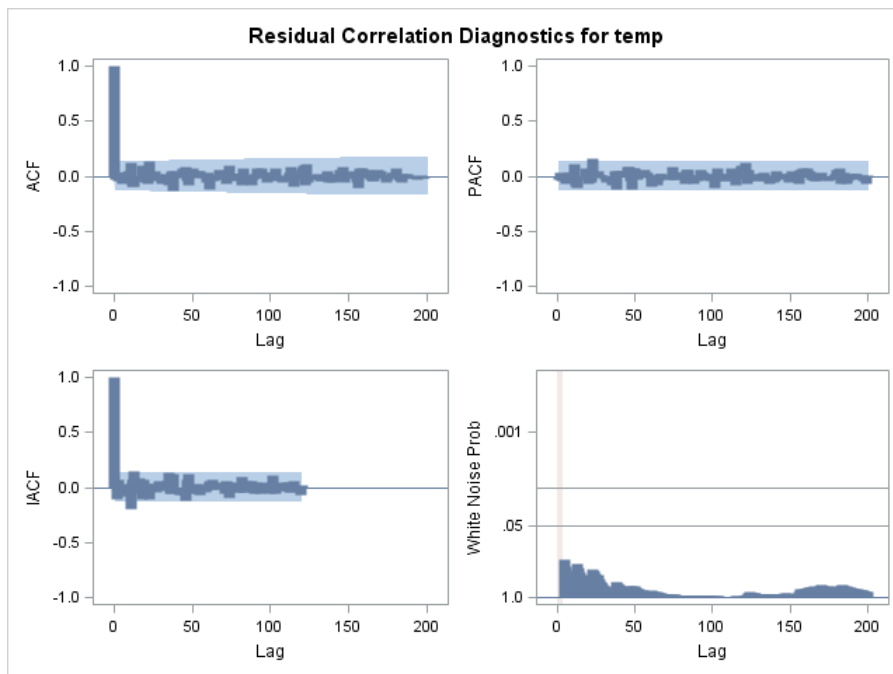


Figure 4.12. Residual of the Weekly ARMA Model

The check for normality of the residuals given in Figure 4.13 show a little deviation from normality, but this is not much of a concern because the model has did an adequate job of predicting the monthly average temperature.

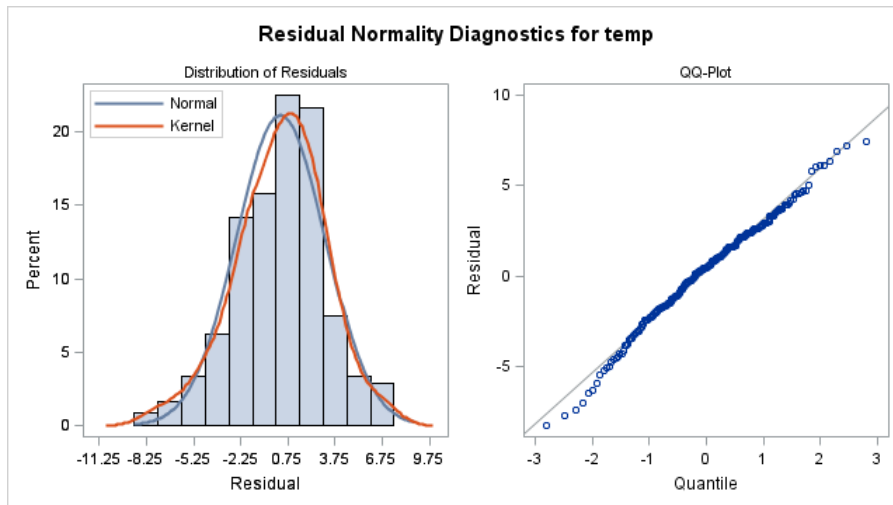


Figure 4.13. Normality Diagnostics of the Residual for the Monthly ARMA Model

Note that this model uses monthly average temperature data from the twenty previous years to forecast the monthly average temperature for twelve months for the next year. This is forecasts from one step to eleven steps ahead. As Figure 4.14 shows, the model provides reasonably accurate forecasts.

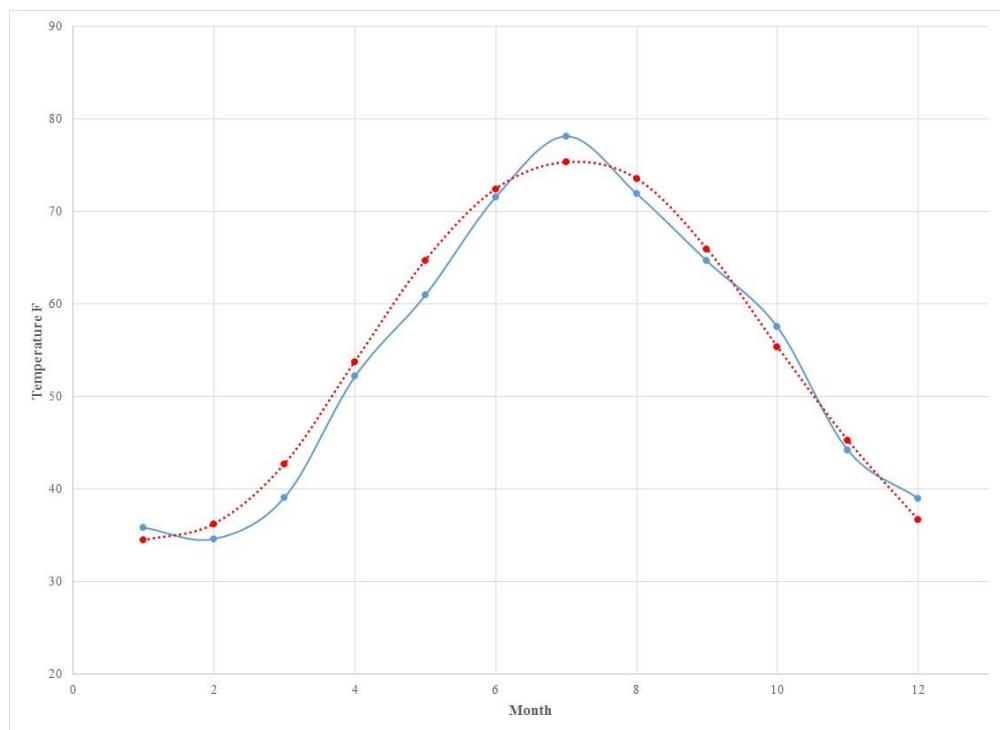


Figure 4.14. The Comparison between the Actual Monthly Temperature (blue solid), the Predicted Monthly Temperature (red dotted) - 2013

**4.2.4 Vector Autoregressive Model with Exogenous Variable (VARX).** The monthly spline coefficients were treated as a vector time series and the possibility of cross-correlation was allowed. Graphs of these coefficients are given in Figures A.3 through A.12 in the appendix. Also, the average monthly temperature and its square were included as input variables. The information criteria for the weekday, weekend, and combined model are given in table 4.10. In addition, the figures 4.15, 1.16 and 4.17 display a comparison between the actual monthly load weekday, weekend, and combined model of the test year, respectively, and the their predictions using VARX model included nowcasting (using the actual temperature) and forecasting (using predicted temperature) separately. These figures show very good predictions with some difference between nowcasting and forecasting temperature for a few months which are July, August, and September.

Table 4.12 The Information Criteria Results of each VARX Model

<b>Weekdays Model</b>		<b>Weekends Model</b>		<b>Combined Model</b>	
<b>Information Criteria</b>		<b>Information Criteria</b>		<b>Information Criteria</b>	
<b>AICC</b>	-3.44664	<b>AICC</b>	3.328577	<b>AICC</b>	-7.59225
<b>HQC</b>	-1.71372	<b>HQC</b>	5.140705	<b>HQC</b>	-6.0795
<b>AIC</b>	-4.31484	<b>AIC</b>	1.934671	<b>AIC</b>	-8.08213
<b>SBC</b>	2.132762	<b>SBC</b>	9.881713	<b>SBC</b>	-3.1186
<b>FPEC</b>	0.013985	<b>FPEC</b>	7.538935	<b>FPEC</b>	0.000315

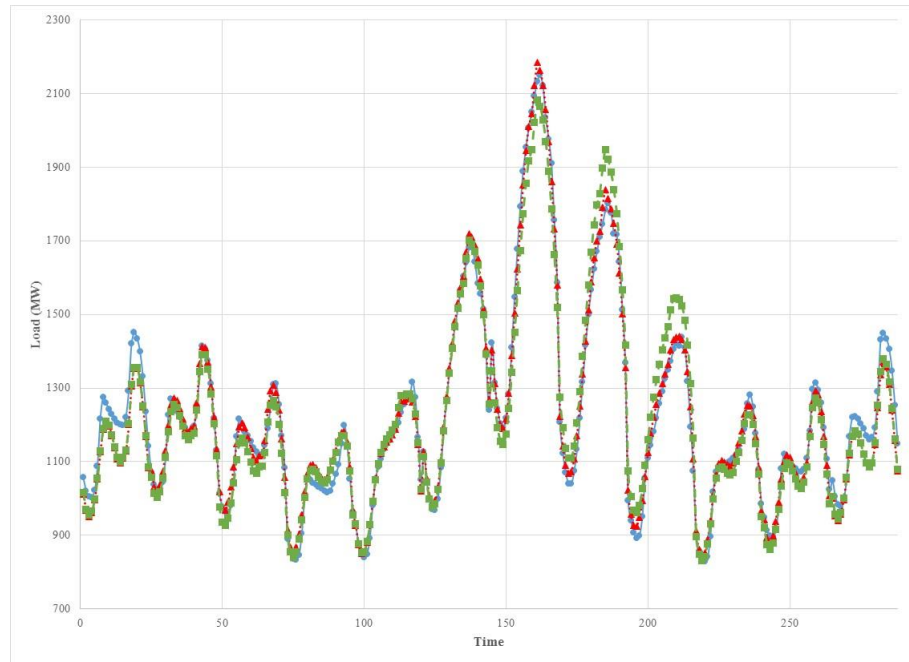


Figure 4.15. The Comparison between the Actual Monthly Weekdays Load (blue solid), the Predicted Monthly Load from the Weekdays Model with Nowcasting Temperature (red dotted), and Forecasting Temperature (green dashed) – 2013

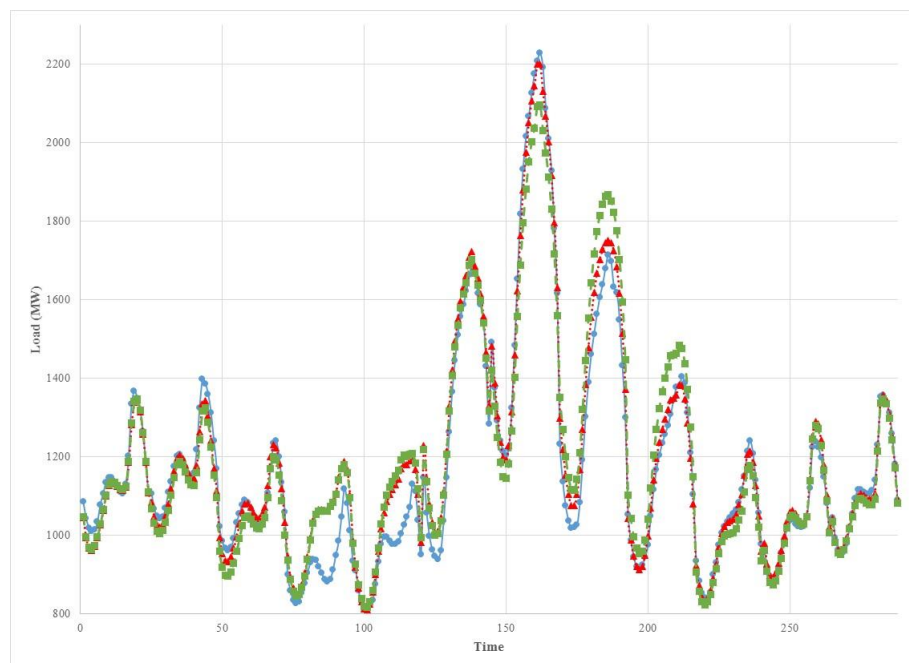


Figure 4.16. The Comparison between the Actual Monthly Weekends Load (blue solid), the Predicted Monthly load of Weekends Model with Nowcasting Temperature (red dotted), and Forecasting Temperature (green dashed) – 2013



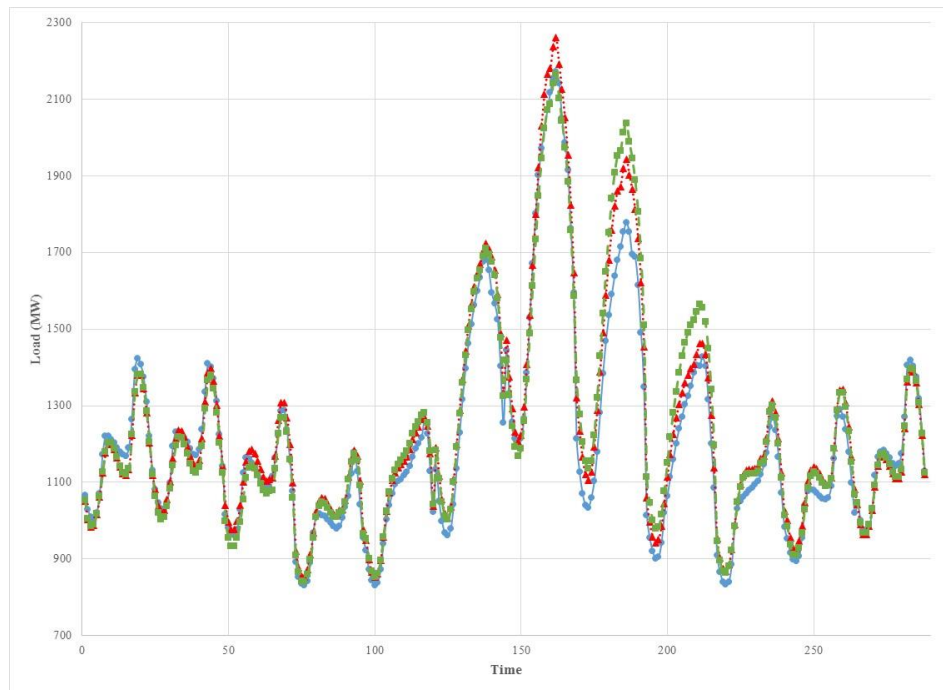


Figure 4.17. The Comparison between the Actual Monthly Load (blue solid), the Predicted Monthly Load from the Whole Model with Nowcasting Temperature (red dotted), and Forecasting Temperature (green dashed) – 2013

**4.2.5 Simultaneous Prediction Intervals for VARX mode.** Scheffe's method was used to calculate the prediction interval of the monthly average forecast load across the 24-hour time period representing the average hourly load for each month of the test year. The mathematical formula used for building the prediction intervals is given in Chapter 3. As Figures 4.18 – 4.10 shows, the prediction intervals are reasonable in the sense that they do not seem to spread out that much for later months of the year.

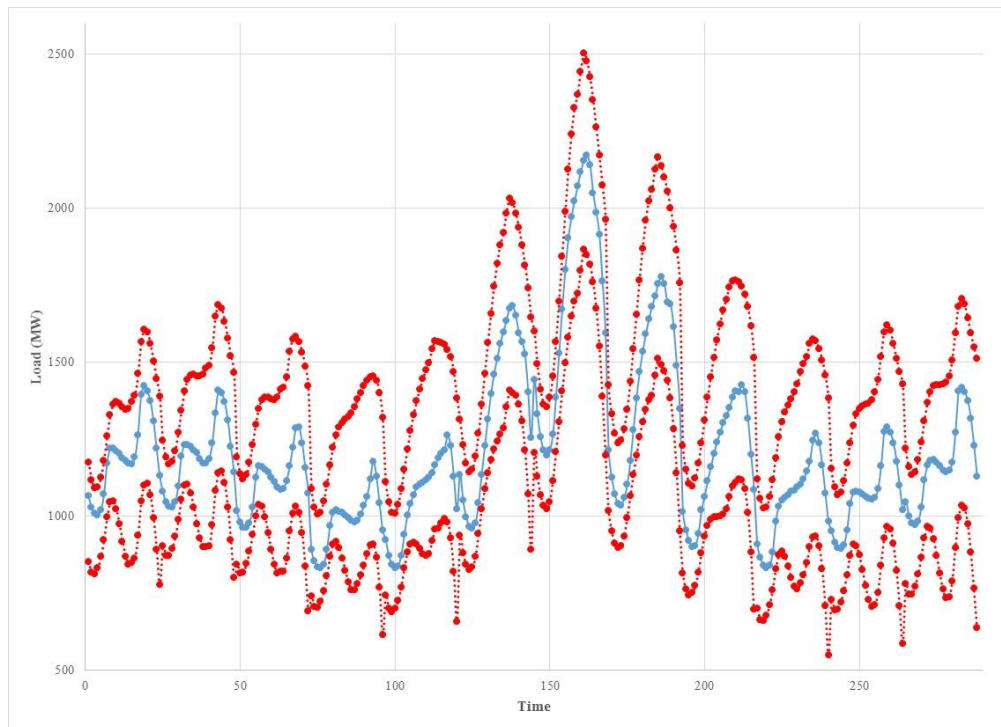


Figure 4.18. The Prediction Interval for the Weekdays Model – 2013

Figure 4.19 giving the prediction intervals for weekend load profiles show prediction bands that are quite wide and unusable. Part of the reason for this may be the small number of data points belonging to the weekends in each month. This lack of data can inflate the standard error of the spline coefficients, and since the predicted load for later hours are determined by not only the coefficients of the intercept, linear, quadratic, and cubic terms, but also by terms involving knots, the standard errors of this multitude of parameter estimates can inflate the width of the prediction intervals. Figure 4.20 provides the intervals for predictions for combined weekend and weekday load profiles.

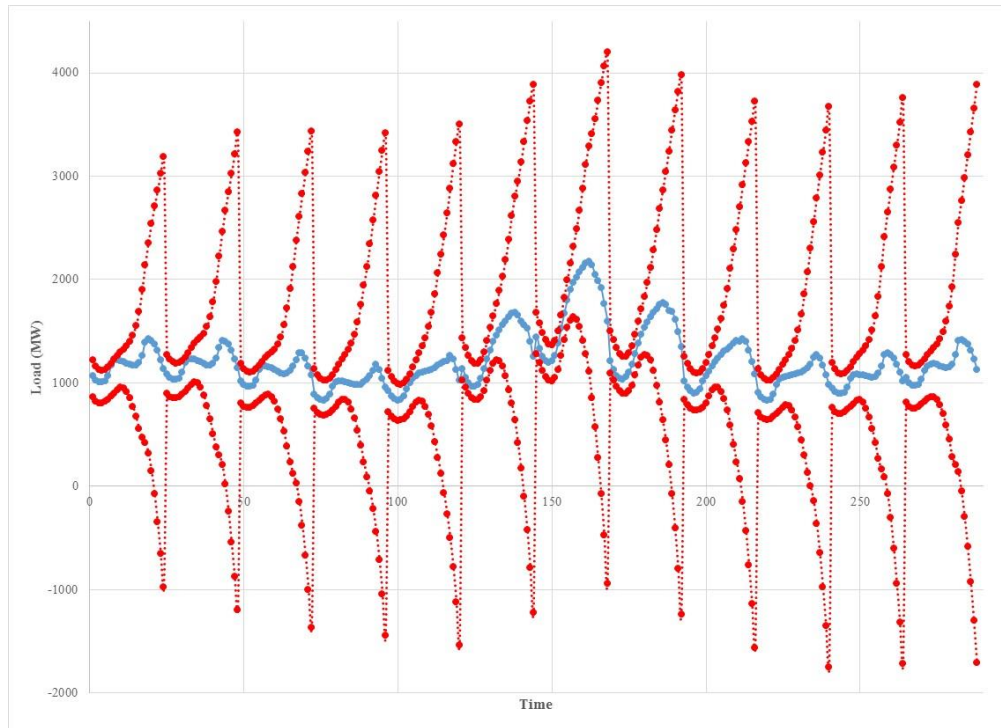


Figure 4.19. The Prediction Interval for the Weekends Model – 2013.

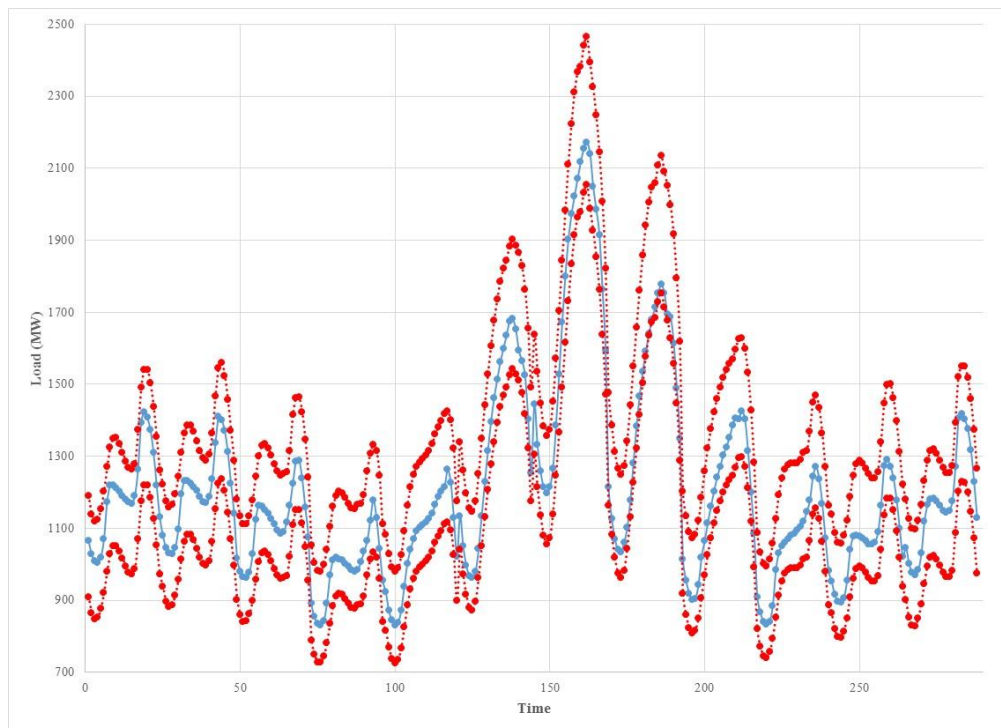


Figure 4.20. The Prediction Interval for the Combined Model – 2013.

### 4.3 FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS APPROACH

In the first approach (short term models), we observed that the knots' locations in the spline curve differ according to the season or the type of day, so we used a specific spline curve for each season and each weekday or weekend. This means that eight models were built, with two models (weekday and weekend) for each of the four seasons. On the other hand, the second approach (long term models) used the same sets of knots at the same positions within each model irrespective of the season. A third approach is to assume that there are underlying fundamental “basis” curves, linear combinations of which can provide the seasonal and type of day variations. This is the FPCA approach.

**4.3.1 Removing the Trend.** The FPCA could be performed on the original data without removing the long-term trend. It is possible that the trend would get incorporated into one or more of the principal component scores, which in turn could be modeled using time series approaches. This trend, however, can distort the estimate of the covariance operator which is used to determine the functional principal components. In addition, preliminary analysis showed that removing long-term trend prior to performing FPCA results in principal component scores that are easier to model using time series approaches. Therefore, the long-term trend was removed from the series by using the same technique that was employed in the first approach, which is to predict the average annual load using the linear regression analysis with select economic variables and subtracting the annual predicted average load from the hourly load.

Figure 4.21 shows the daily load profiles for each day in the training sample. The graph shows an overall pattern, with some deviations from the norm clearly visible. Figures A.13 through A.16 in the appendix provides the observed daily load patterns by each season. Several steps are necessary to carry out FPCA on this data and then model the resulting PCA scores. These steps are described in subsections 4.3.2 through 4.3.4.

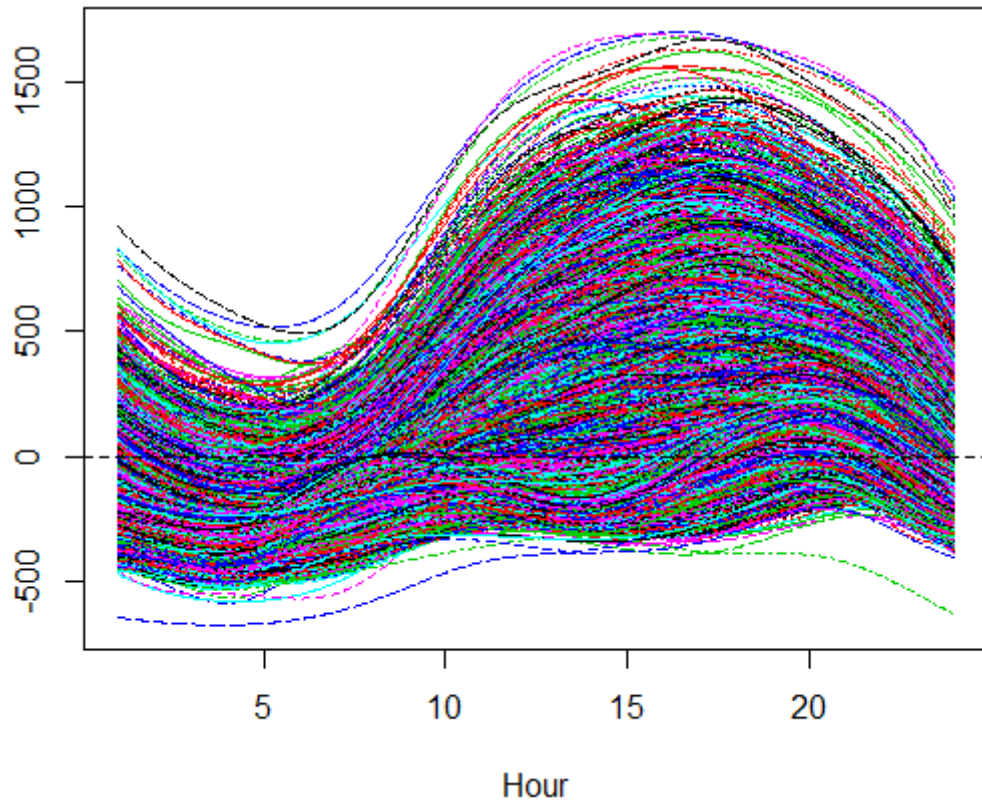


Figure 4.21. The Daily 24-hour Profile over 20 Years

**4.3.2 Creating a Basis and Smoothing the Data.** The daily load data has to be smoothed into a functional object prior to submitting to FPCA. This was achieved through the use of B-splines. The first two approaches to modeling electricity load used only three knots because adding more knots will create complex models, but for the FPCA approach the principal aim is to smooth the data and so 4, 6, 8, 12 knots were employed to fit cubic splines. The resulting smoothed data were subjected to FPCA and in-sample predictions were made. The resulting root mean square error (RMSE) was employed to choose an optimal number of knots from the above list. Results showed that the difference in RMSE between using 8 or 12 knots was small, so 8 knots<sup>11</sup> were used for the analysis.

<sup>11</sup> Number of basis = number of knots + order – 2. Ramsay, J. O and others (2009).

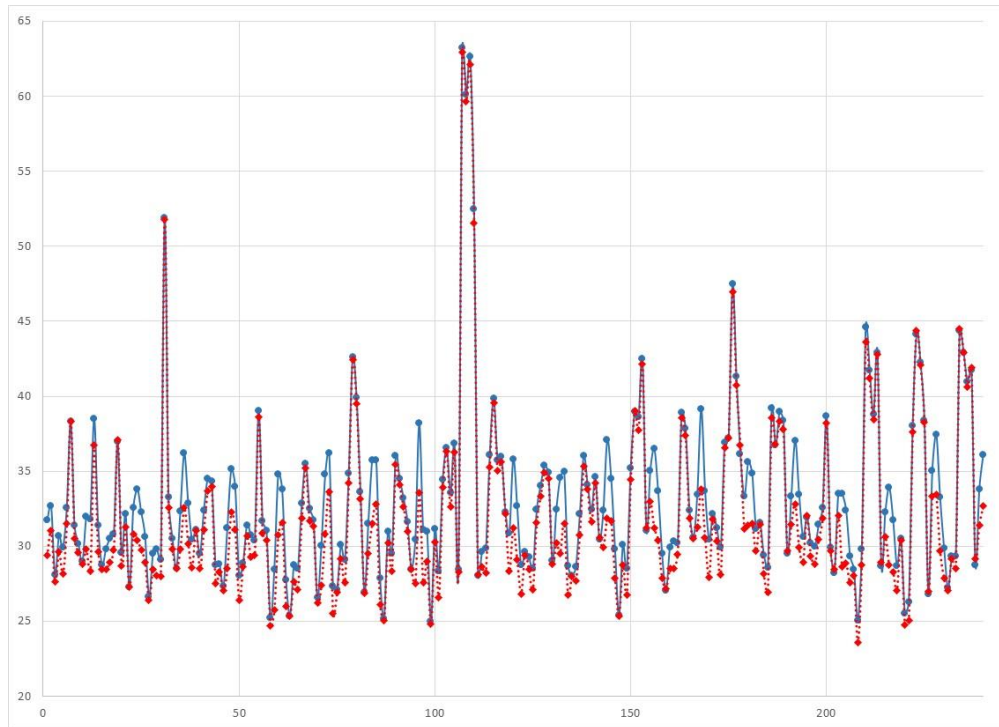


Figure 4.22. The Comparison between the RMSE of the 8 Knots Model (blue solid) and the 12 Knots Model (red dotted).

**4.3.3 Computing the Functional Eigenfunctions (Harmonics) and the Principal Component Scores.** After smoothing the data by using cubic splines with eight knots and three basis functions, the eigenfunctions, the principal component scores, and the mean function was computed for the de-trended data using R software. The proposed model is:

$$f_t(x) = \mu(x) + \sum_{k=1}^K \beta_{t,k} \phi_k(x) + e_t(x), \quad t = 1, 2, \dots, N,$$

where  $N = 7302$  (number of days) and  $k = 1, 2, \dots, 24$  (number of hours per day).

The first principal component explained most of the variation in the data; specifically 92.7% of the variation. The second and third principal components explained 4.8% and 1.2% of the variation respectively, so the first three principal components explained 98.7% leaving only 1.3% unexplained. The reason that the first three principal components were used was not only because of the total percentage of variation explained combination, but

also because when in sample predictions were made using only the first two principal components, the prediction did not fit the actual hourly load for winter months very well, especially around the mid-day hours after first peak.

Figure 4.23 shows the mean function, which can be taken as the overall load profile over the 24-hour period of an “average” day, irrespective of the season or the type of day. Figure 4.24 shows the first eigenfunction (Harmonic). It clearly emphasizes the late afternoon and early evening peaks. The second eigenfunction shown in Figure 4.25 contributes to the morning peak, which may reflect people using electricity for morning chores as well as offices powering up to get ready for the day’s activities. The third eigenfunction accentuates the morning and evening peaks. It is worth noting that the mean curve resembles the general shape of the electricity load for weekdays during the late spring, summer and early fall periods. The two peaks are quite prominent during winter weekdays.

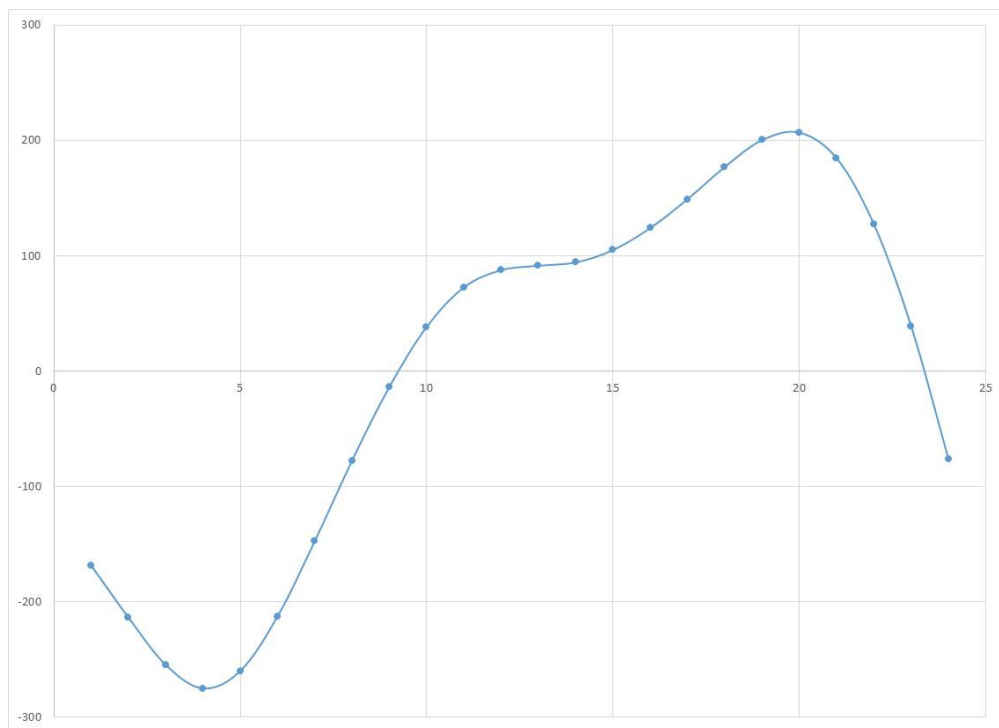


Figure 4.23. The Mean Curve of the 24-hour Profile

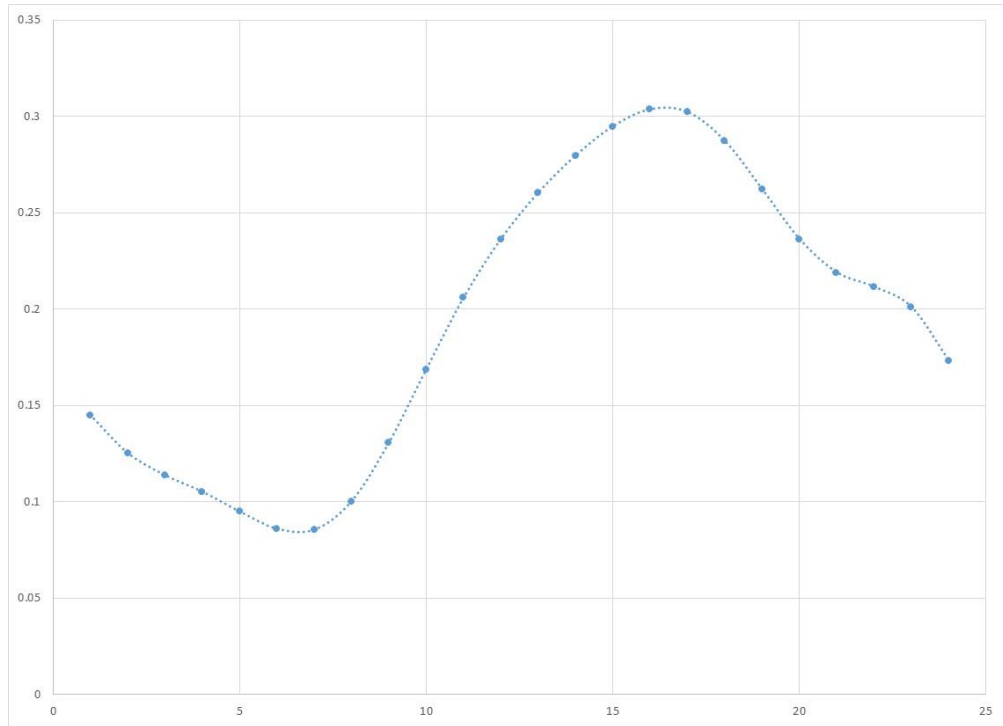


Figure 4.24. The First Eigenfunction (Harmonic) Curve

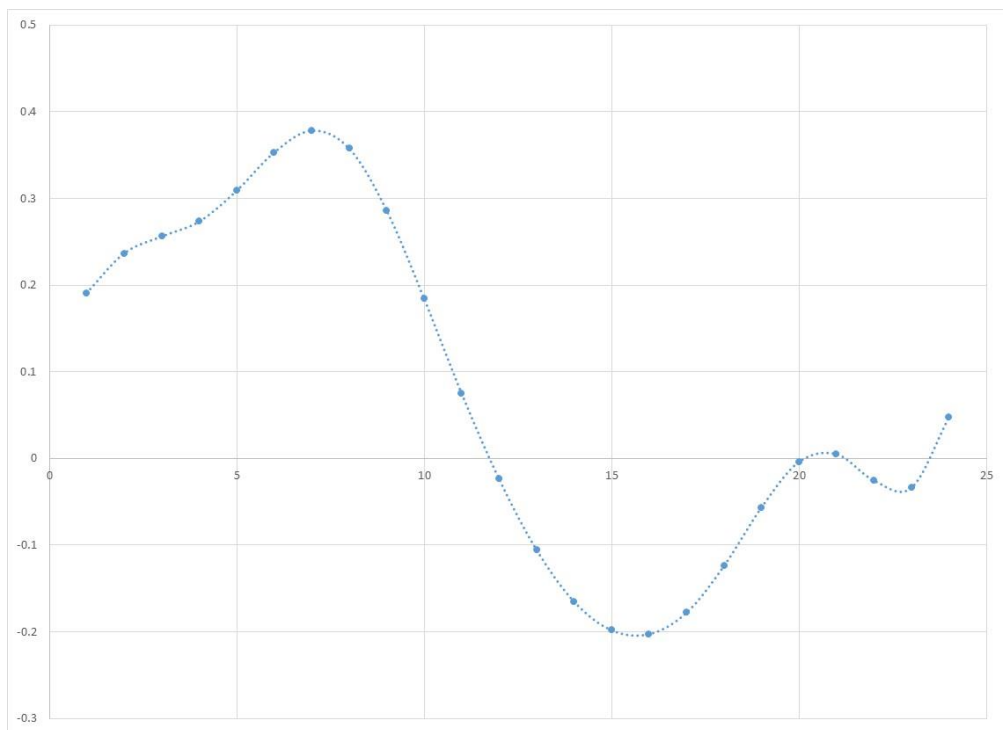


Figure 4.25. The Second Eigenfunction Curve



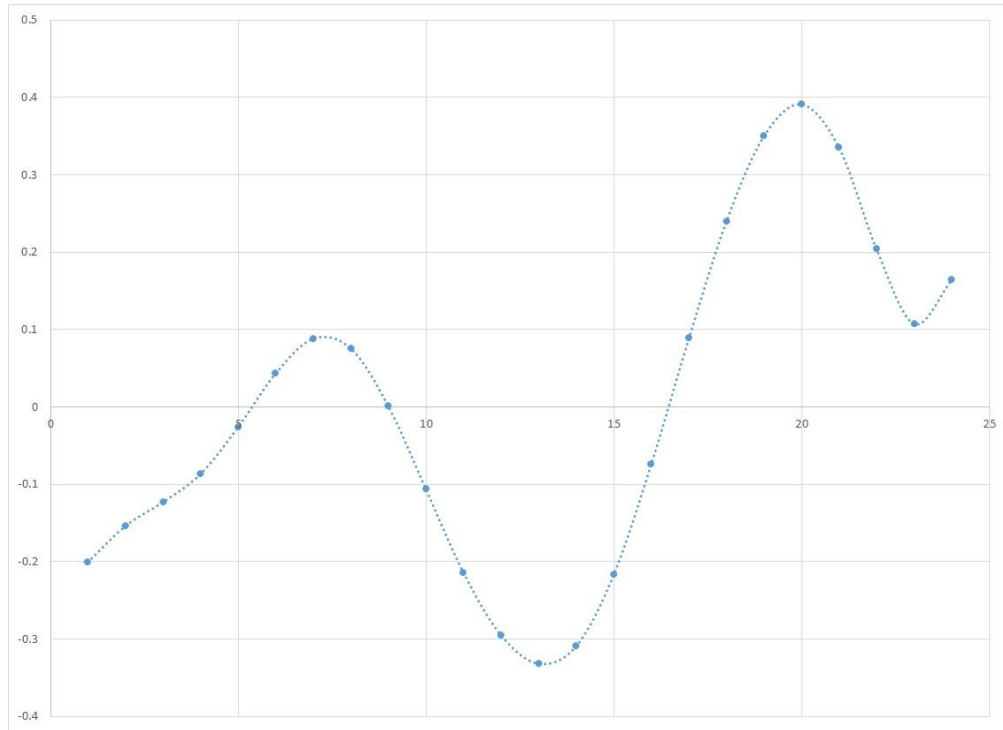


Figure 4.26. The Third Eigenfunction Curve

**4.3.4 Modeling the Principal Components.** It was difficult to model the third principal component scores using an ARIMA time series structure, but the transfer function models provided good fit to the data from the first three principal component scores with independent variables included as input. The models that used to predict the principal component scores are:

- i. The first principal component (PC1) model

$$\begin{aligned}
 Y_t = & 1595.81 + (1 - 1.39B + 0.30B^2 + 0.08B^7 - 0.85B^{364} + 1.27B^{365} - 0.43B^{366})Y_t \\
 & + (1 - 0.83B - 1.67B^2 + 0.05B^7 - 0.80B^{364} + 0.77B^{365})Z_t + (-64.37 - 1.79B^2)T_t \\
 & + (0.47 - 0.02B)T_t^2 + (6.81 + 0.69B + 0.39B^2)CDD_t \\
 & - 215.94w_t - 0.03w_t * T_t^2 + 0.23w_t * CDD_t,
 \end{aligned}$$

where  $t = 1, 2, \dots, N$ ,  $T$  denotes the daily average temperature,  $CDD_t$  denote the total of the Cooling Degree Day values for each hour on day  $t$ , and  $w_t$  is the weekend dummy for

day  $t$ . Note that  $Z_t$ 's are moving average components that will be estimated by the transfer function procedure.

ii. The second principal component (PC2) model is

$$\begin{aligned}
 Y_t = & 1346.72 + (1 - 0.98B - 0.004B^2 - 0.005B^7 - 0.99B^{364} + 0.98B^{365})Y \\
 & + (1 - 0.78B - 0.17B^2 - 0.98B^{364} + 0.79B^{365} + 0.15B^{366})Z_t \\
 & + (-33.23 - 6.66B)T_t + 0.24T_t^2 + (-0.50 + 1.49B + 0.07B^2)CDD_t \\
 & - 122.75w_t + 0.20w_t * CDD_t.
 \end{aligned}$$

iii. The third principal component (PC3) model is

$$\begin{aligned}
 Y_t = & -59.94 + (1 - 1.30B + 0.04B^7 - 0.76B^{364} + 0.75B^{365})Y_t \\
 & + (1 - 1.08B + 0.05B^2 + 0.04B^7 - 0.66B^{364} + 0.73B^{365} - 0.08B^{366})Z_t \\
 & + (0.10 + 7.32B - 2.73B^2)T_t + (-0.04 - 0.04B + 0.02B^2)T_t^2 \\
 & + (0.93 + 0.83B)CDD_t - 91.49w_t - 0.02w_t * T_t^2 - 0.21w_t * CDD_t.
 \end{aligned}$$

Table 4.13. The Results for the Daily PC's Scores Transfer Function Model Fit Statistics

Fit Statistics	PC1 Model	PC2 Model	PC3 Model
Constant Estimate	0.466297	0.171331	-0.01215
Variance Estimate	43444.1	14236.11	8629.324
Std Error Estimate	208.4325	119.3152	92.89415
AIC	98747.74	90693.18	86948.07
SBC	98906.34	90824.2	87099.77
Number of Residuals	7300	7300	7300

The above models were used to forecast the scores for each principal component score for one year ahead, then the forecast scores were multiplied by the eigenfunctions, and the mean function was added to obtain the forecasts of the 24-hour load profiles for each day of the test year (2013). These 24-hour profiles were then averaged by month to produce the composite graph given in Figure 4.27. The figure shows that the average of the point forecasts fit nicely with the average monthly forecasts, suggesting that the FPCA approach can be used successfully to conduct long-term forecasting of average monthly profiles.

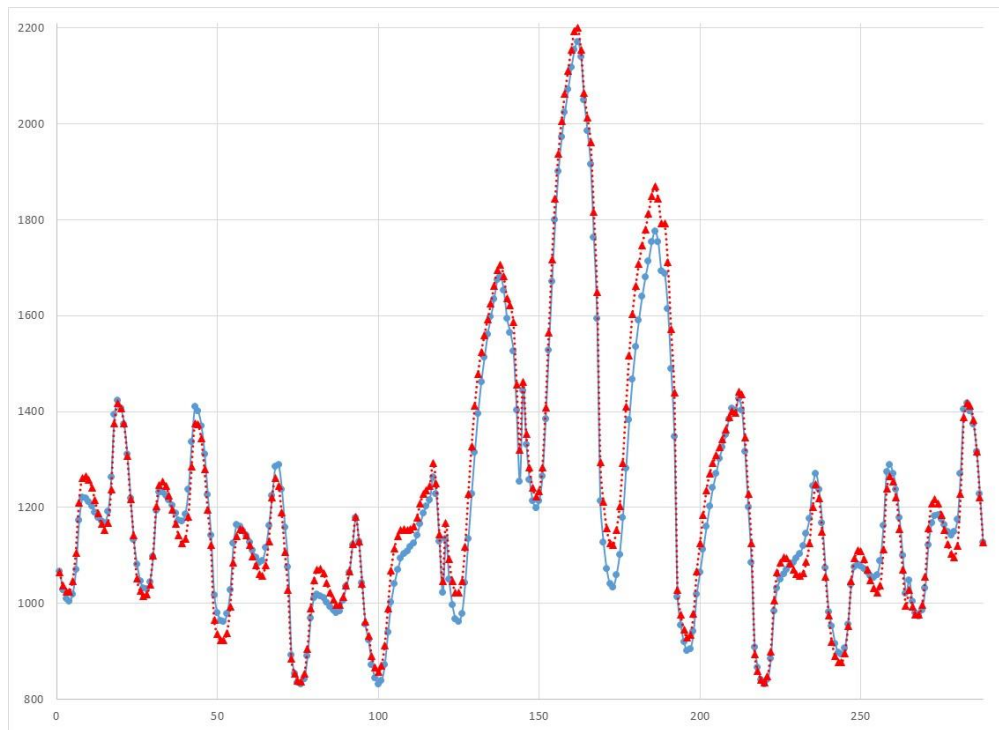


Figure 4.27. The Comparison between the Actual Monthly Load (blue solid) and the Predicted Monthly Load (red dotted) – 2013.

**4.3.5 The Prediction Interval for the Hourly Load Forecast.** The mathematical details on the procedure to compute prediction intervals for hourly load forecasts constructed based on the FPCA method are given in Section 3.4 and is not repeated here to conserve space. The resulting prediction intervals are given in figures 4.28 – 4.30.

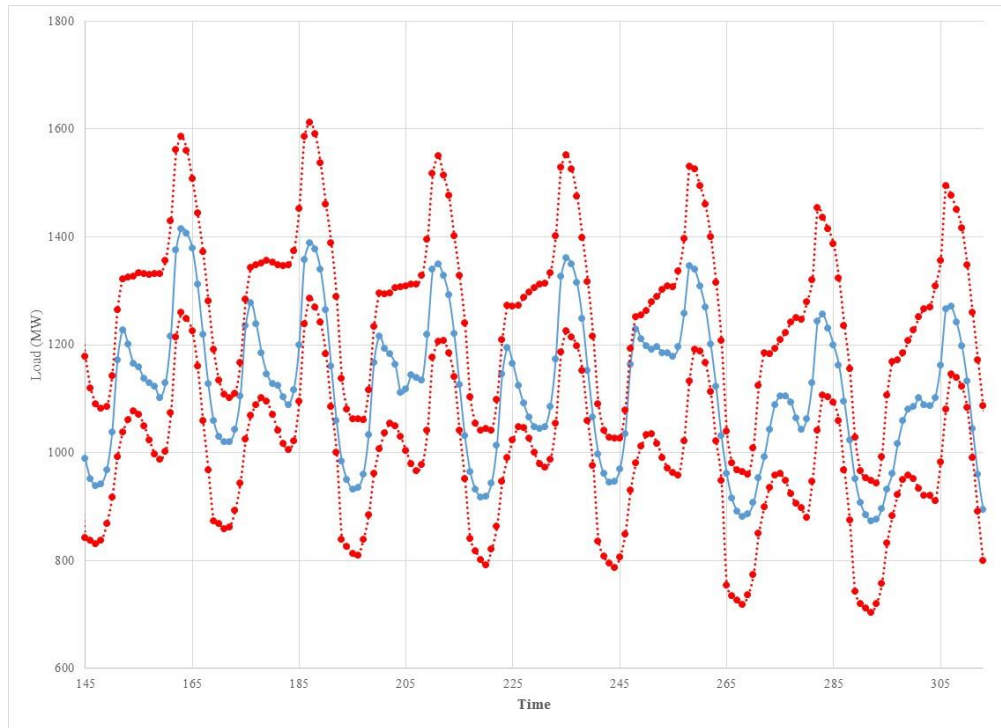


Figure 4.28. The Prediction Interval from FPCA - a Week in the Mid-Winter Season

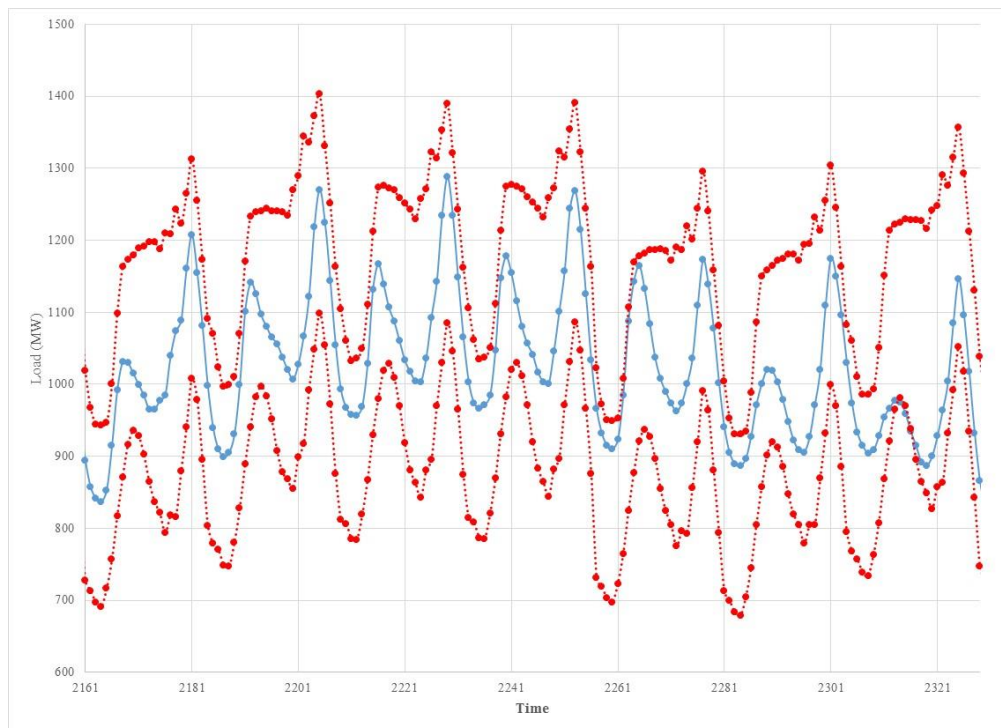


Figure 4.29. The Prediction Interval from FPCA - a Week in the Mid-Spring Season

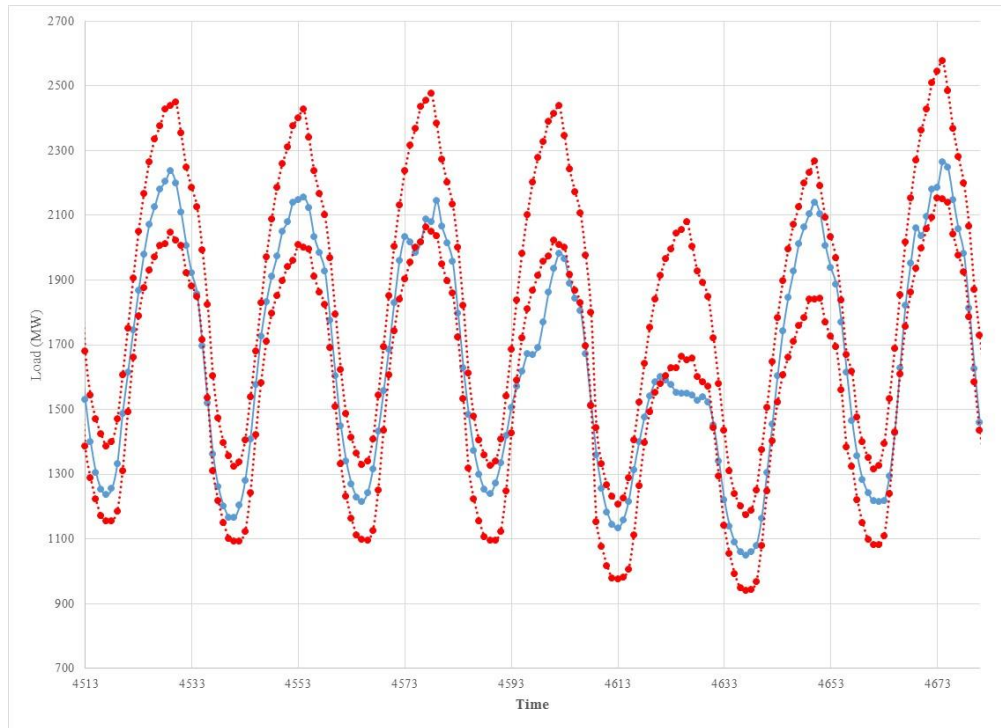


Figure 4.30. The Prediction Interval from FPCA - a Week in the Mid-Summer Season

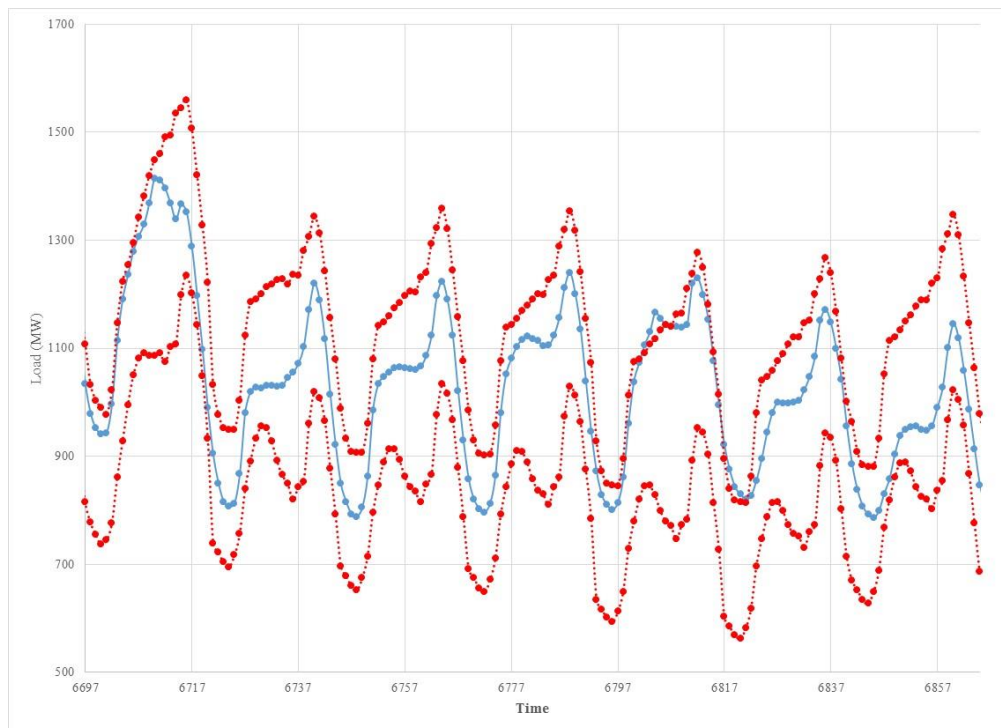


Figure 4.31. The Prediction Interval from FPCA - a Week in the Mid-Fall Season

## 5. MODEL COMPARISONS AND CONCLUSIONS

In this dissertation three approaches to modeling the 24-hour daily electricity load profile were proposed. The first approach provides two models, each of which can be used to obtain short-term forecasts. The second model can be used to obtain only long-term forecasts because it models the monthly average of 24-hour load profile. On the other hand, the third approach can be used to forecast the long-term and shot-term electricity load.

The models arising out of the three approaches can be compared based on the coefficient of variation, which is prediction error normalized by the average load, say  $\bar{X}$ .

### 5.1 THE COMPARISON OF THE LONG-TERM ELECTRICITY LOAD FORECASTS

The coefficient of variation (*CV*) was used to compare between the three approaches. *CV* is defined as follows:

$$CV = \frac{RMSE \text{ for each model}}{\bar{X}}$$

where *RMSE* is the root of the mean square error which is defined as

$$RMSE_m = \sqrt{\frac{\sum_{i=1}^{24} (Y_{m,i} - \hat{Y}_{m,i})^2}{24}} \text{ for each month, or } RMSE_i = \sqrt{\frac{\sum_{m=1}^{12} (Y_{m,i} - \hat{Y}_{m,i})^2}{12}} \text{ for each hour.}$$

#### 5.1.1 Comparison of the Long-Term Electricity Load Forecasts by Month.

The Table 5.1 and Fig 5.1 show a comparison between the coefficients of variation (*CV*) for the three models for each month of the forecasted year (2013). In general, the third approach presents the lowest *CV* over six months, the first approach presents the lowest *CV* over four months, and the second approach presents the lowest *CV* only over two months. Note that the first approach requires knowledge of the hourly temperature, the

second approach requires the knowledge of the monthly average temperature, and the third approach requires the knowledge of the daily average temperature. The results given below were obtained using the actual temperature so as to determine the prediction accuracy of the approaches without introducing the error due to predicting the temperature. Use of one-step prediction of temperature, however, does not make an appreciable difference in the forecast accuracy of the second approach.

Table 5.1. The CV for the Long-Term Load of the Three Models by Month

<b>Month</b>	<b>Approach 1 Combined Model</b>	<b>Approach 2 Combined Model</b>	<b>Approach 3</b>	<b>The Best Approach</b>
<b>1</b>	0.021717	0.028516	0.012271	<b>Approach 3</b>
<b>2</b>	0.029635	0.013771	0.023851	<b>Approach 2</b>
<b>3</b>	0.037026	0.020604	0.030913	<b>Approach 2</b>
<b>4</b>	0.045212	0.033577	0.030429	<b>Approach 3</b>
<b>5</b>	0.02484	0.031073	0.039578	<b>Approach 1</b>
<b>6</b>	0.020292	0.044454	0.034879	<b>Approach 1</b>
<b>7</b>	0.065309	0.030394	0.018977	<b>Approach 3</b>
<b>8</b>	0.061778	0.091617	0.068852	<b>Approach 1</b>
<b>9</b>	0.02358	0.042546	0.026656	<b>Approach 1</b>
<b>10</b>	0.032066	0.040796	0.028044	<b>Approach 3</b>
<b>11</b>	0.04732	0.044529	0.027615	<b>Approach 3</b>
<b>12</b>	0.030604	0.020227	0.017272	<b>Approach 3</b>

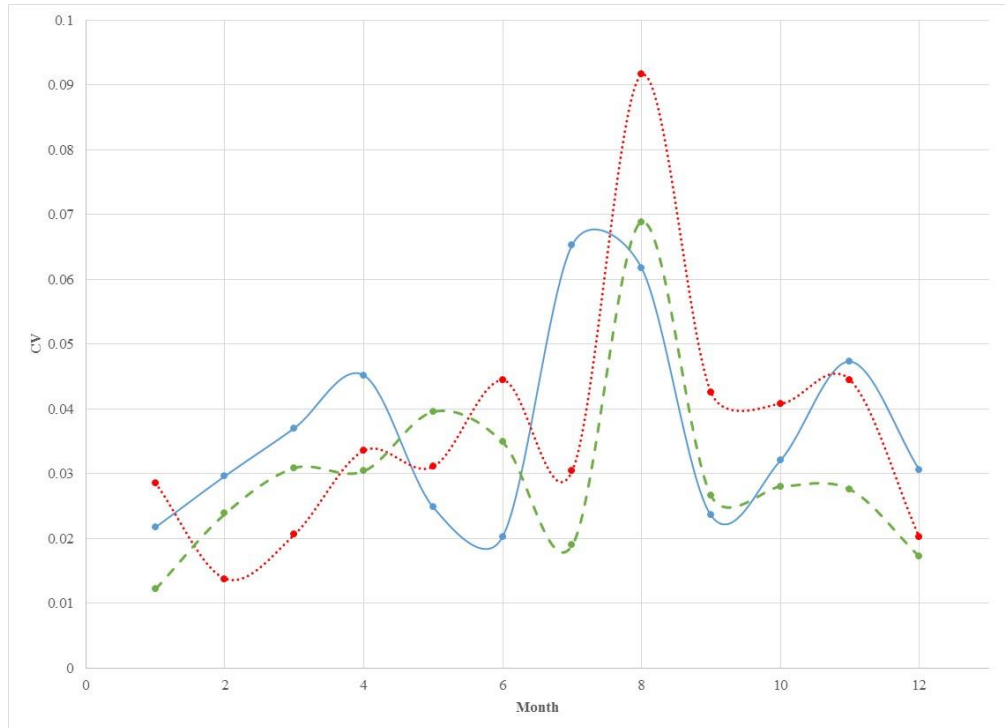


Figure 5.1. The *CV* of the Long-Term Load for the Three Models by Month; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed)

### 5.1.2 The Comparison of the Long-Term Electricity Load Forecasts by Hour.

The Table 5.2 and Fig 5.2 show a comparison between the coefficients of variation (*CV*) for the three models for each month of the forecasted year (2013). The third approach presents the lowest *CV* over sixteen hours, the second first presents the lowest *CV* over eight hours, and the second approach does not present the lowest *CV* comparing with the others, but the *CV* of the first approach is lower than the *CV* of the second approach for the first twelve hours. Again, the results presented here are based on actual temperatures rather than forecasted temperature values.



Table 5.2. The CV for the Long-Term Load of the Three Models per Hour

<b>Month</b>	<b>Approach 1 Combined Model</b>	<b>Approach 2 Combined Model</b>	<b>Approach 3</b>	<b>The Best Approach</b>
<b>1</b>	0.051816	0.042275	0.031913	<b>Approach 3</b>
<b>2</b>	0.057126	0.045251	0.02916	<b>Approach 3</b>
<b>3</b>	0.060465	0.040975	0.029096	<b>Approach 3</b>
<b>4</b>	0.065162	0.036733	0.027331	<b>Approach 3</b>
<b>5</b>	0.064392	0.036701	0.024557	<b>Approach 3</b>
<b>6</b>	0.059132	0.033509	0.021589	<b>Approach 3</b>
<b>7</b>	0.06166	0.037209	0.020239	<b>Approach 3</b>
<b>8</b>	0.050848	0.038925	0.018126	<b>Approach 3</b>
<b>9</b>	0.048837	0.039595	0.025467	<b>Approach 3</b>
<b>10</b>	0.05447	0.041731	0.035711	<b>Approach 3</b>
<b>11</b>	0.04962	0.042578	0.041575	<b>Approach 3</b>
<b>12</b>	0.04692	0.045453	0.043292	<b>Approach 3</b>
<b>13</b>	0.044297	0.050384	0.043676	<b>Approach 3</b>
<b>14</b>	0.039047	0.053643	0.043588	<b>Approach 1</b>
<b>15</b>	0.031179	0.053296	0.043524	<b>Approach 1</b>
<b>16</b>	0.026686	0.046265	0.043591	<b>Approach 1</b>
<b>17</b>	0.026812	0.043568	0.040448	<b>Approach 1</b>
<b>18</b>	0.034718	0.04454	0.034168	<b>Approach 3</b>
<b>19</b>	0.024569	0.042784	0.027898	<b>Approach 1</b>
<b>20</b>	0.021575	0.048052	0.026039	<b>Approach 1</b>
<b>21</b>	0.026337	0.039071	0.028585	<b>Approach 1</b>
<b>22</b>	0.028592	0.039543	0.031547	<b>Approach 1</b>
<b>23</b>	0.040992	0.050775	0.034813	<b>Approach 3</b>
<b>24</b>	0.04405	0.043335	0.034866	<b>Approach 3</b>

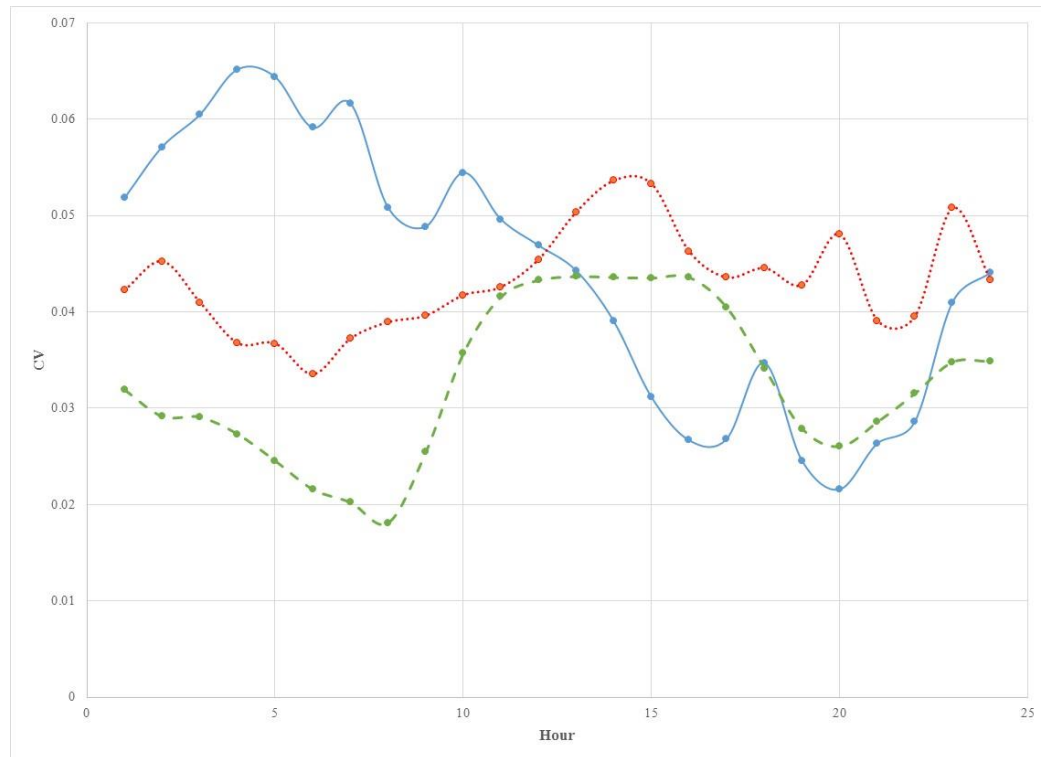


Figure 5.2. The CV of the Long-Term Load for the Three Models by Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed)

## 5.2 THE COMPARISON OF THE SHORT-TERM ELECTRICITY LOAD FORECASTS

The coefficient of variation (CV) was used to compare between the effectiveness of the three approaches for short-term prediction. The formula for RMSE would change from what was previously used to the following:

The RMSE for each month is defined by  $RMSE_m = \sqrt{\frac{\sum_{t=1}^{N_l} \sum_{h=1}^{24} (Y_{m,t,i} - \hat{Y}_{m,t,i})^2}{24 \times N_l}}$  and for each

hour is defined as  $RMSE_i = \sqrt{\frac{\sum_{t=1}^{N_l} \sum_{m=1}^{12} (Y_{m,t,i} - \hat{Y}_{m,t,i})^2}{12 \times N_l}}$ , where  $N_l$  is the number of days for

the given month.

**5.2.1 The Comparison of the Short-Term Electricity Load per Month.** The Table 5.3 and Fig 5.3 provide comparison between the coefficients of variation (*CV*) for the three models for each month of the forecasted year (2013). In general, the third approach presents the lowest *CV* over eight months, the first approach presents the lowest *CV* over four months.

Table 5.3. The *CV* for the short-term load of the three models per month

<b>Month</b>	<b>Approach 1 Combined Model</b>	<b>Approach 2 Combined Model</b>	<b>Approach 3</b>	<b>The Best Approach</b>
<b>1</b>	0.080001	0.095418	0.047696	<b>Approach 3</b>
<b>2</b>	0.052844	0.060892	0.053321	<b>Approach 1</b>
<b>3</b>	0.069404	0.072015	0.067545	<b>Approach 3</b>
<b>4</b>	0.070818	0.076308	0.078164	<b>Approach 1</b>
<b>5</b>	0.122593	0.16125	0.07567	<b>Approach 3</b>
<b>6</b>	0.065271	0.165652	0.068956	<b>Approach 1</b>
<b>7</b>	0.09346	0.151676	0.064555	<b>Approach 3</b>
<b>8</b>	0.104277	0.154339	0.10457	<b>Approach 1</b>
<b>9</b>	0.144395	0.211306	0.075331	<b>Approach 3</b>
<b>10</b>	0.074403	0.110849	0.069476	<b>Approach 3</b>
<b>11</b>	0.078837	0.090176	0.058889	<b>Approach 3</b>
<b>12</b>	0.059224	0.083436	0.056806	<b>Approach 3</b>

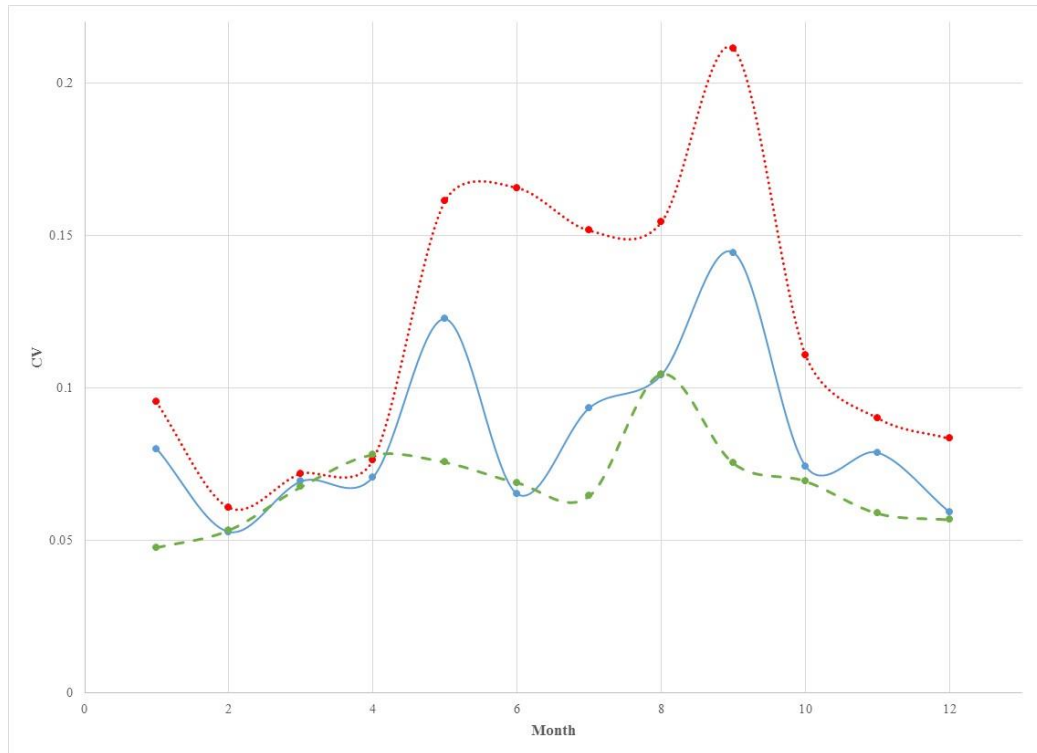


Figure 5.3. The CV of the Short-Term Load for the Three Models by Month; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed)

### 5.2.2 The Comparison of the Short-Term Electricity Load per Hour. The

Table 5.4 and Fig 5.4 show a comparison between the coefficients of variation (*CV*) for the three models for each month of the forecasted year (2013). The third approach presents the lowest *CV* across all 24 hours. Additional illustrations comparing the three methods using *CV* for short-term predictions are given in Figures A.17 through A.28 in the appendix. Figures A.29 through A.32 provide a comparison of the load profiles obtained by each method to each other and the actual load.

Table 5.4. The CV for the short-term load for the three models per hour

<b>Month</b>	<b>Approach 1 Combined Model</b>	<b>Approach 2 Combined Model</b>	<b>Approach 3</b>	<b>The Best Approach</b>
<b>1</b>	0.091626	0.115105	0.0672	<b>Approach 3</b>
<b>2</b>	0.096508	0.113674	0.064623	<b>Approach 3</b>
<b>3</b>	0.099608	0.109262	0.061986	<b>Approach 3</b>
<b>4</b>	0.103005	0.104262	0.059061	<b>Approach 3</b>
<b>5</b>	0.10163	0.101208	0.055824	<b>Approach 3</b>
<b>6</b>	0.096505	0.098419	0.053089	<b>Approach 3</b>
<b>7</b>	0.095408	0.106685	0.063386	<b>Approach 3</b>
<b>8</b>	0.086231	0.111089	0.066964	<b>Approach 3</b>
<b>9</b>	0.084465	0.110969	0.063511	<b>Approach 3</b>
<b>10</b>	0.090368	0.118429	0.069211	<b>Approach 3</b>
<b>11</b>	0.091015	0.130623	0.076302	<b>Approach 3</b>
<b>12</b>	0.093075	0.143651	0.081231	<b>Approach 3</b>
<b>13</b>	0.096096	0.155636	0.084231	<b>Approach 3</b>
<b>14</b>	0.097644	0.164279	0.085673	<b>Approach 3</b>
<b>15</b>	0.097152	0.168072	0.086282	<b>Approach 3</b>
<b>16</b>	0.096581	0.166599	0.085947	<b>Approach 3</b>
<b>17</b>	0.095026	0.162091	0.082814	<b>Approach 3</b>
<b>18</b>	0.092855	0.152938	0.074873	<b>Approach 3</b>
<b>19</b>	0.084613	0.142942	0.068646	<b>Approach 3</b>
<b>20</b>	0.078168	0.134138	0.062681	<b>Approach 3</b>
<b>21</b>	0.074998	0.122618	0.059756	<b>Approach 3</b>
<b>22</b>	0.07317	0.119154	0.060396	<b>Approach 3</b>
<b>23</b>	0.078469	0.120694	0.064392	<b>Approach 3</b>
<b>24</b>	0.082857	0.116402	0.068482	<b>Approach 3</b>

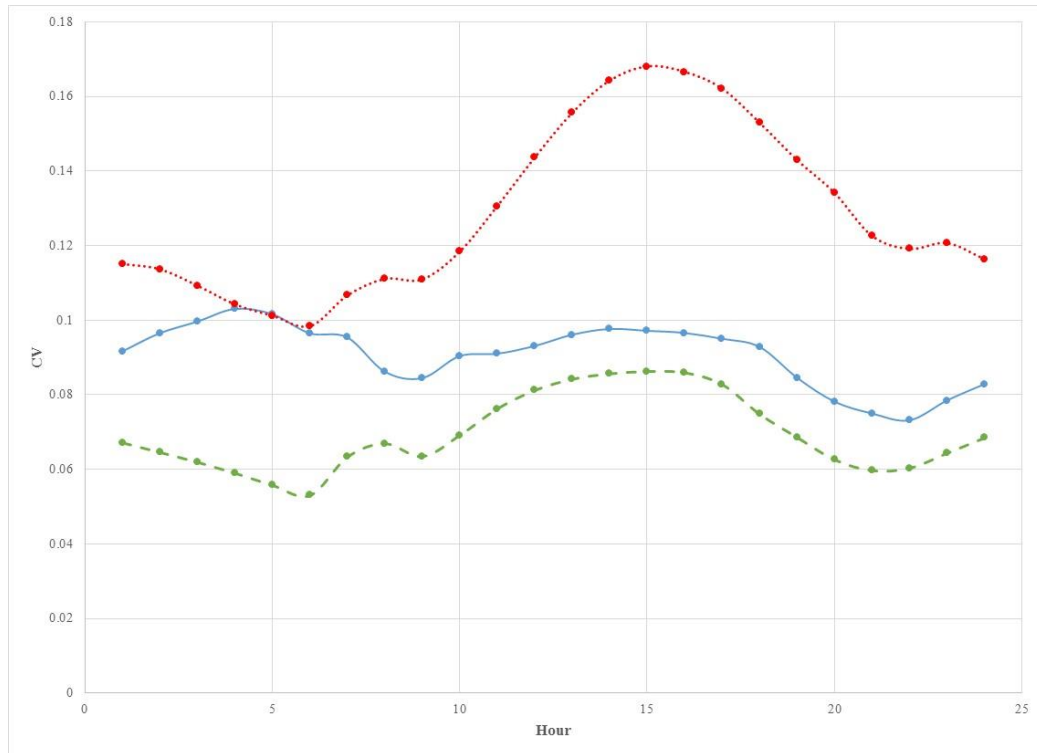


Figure 5.4. The CV of the Short-Term Load for the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed)

The three approaches presented in this thesis add to the considerable literature available for modeling and forecasting electricity load profile. One advantage of two of the three approaches is that they can be implemented by standard statistical software packages and do not require extensive programming. These two methods yield good predictions for either the short-term or the long-term. The third method uses a technique that is fairly new in the area of electricity load forecasting. It has some appealing features such as decomposing the electric load profile into eigenfunctions that correspond to the components that go into making a given days electricity load profile.

The models presented herein can be extended in several ways. The first approach can be modified by fitting separate models for each month rather than by season. In addition, separate models can be obtained for each day-of-the-week rather than for weekdays and weekends. Further improvements can be made by carrying out the de-trending and seasonal adjustments in one step rather than in two. The second approach can be improved if a methodology can be developed to obtain monthly estimates of spline

coefficients simultaneously rather than one month at a time. This may be possible through a state-space approach suggested by Harvey and Koopman (1993), but care must be taken to ensure that the model is estimable when the amount of data is not large. The third approach has several avenues for improvement. Currently, no exogenous variables were included as inputs when conducting the FPCA estimation. Variables such as temperature and weekend/weekday dummies are some possibilities as variables for inclusion. Another modification is to allow the eigenfunctions to change gradually from year to year by including slowly varying economic factors when conducting the FPCA.

## APPENDIX

In this appendix, additional figures and tables relevant to material presented in Chapters 4 and 5 of the main text are given.

### 1. Graphs Related to the Short-Term Approach

Figure A.1 displays the predicted values obtained from regressing average annual load values against economic variables (in blue) and the weekly data obtained by applying a moving average (in red). Details of the procedure are given in sub-section 4.1.2. Figure A.2 provides a comparison of actual and predicted weekly load.

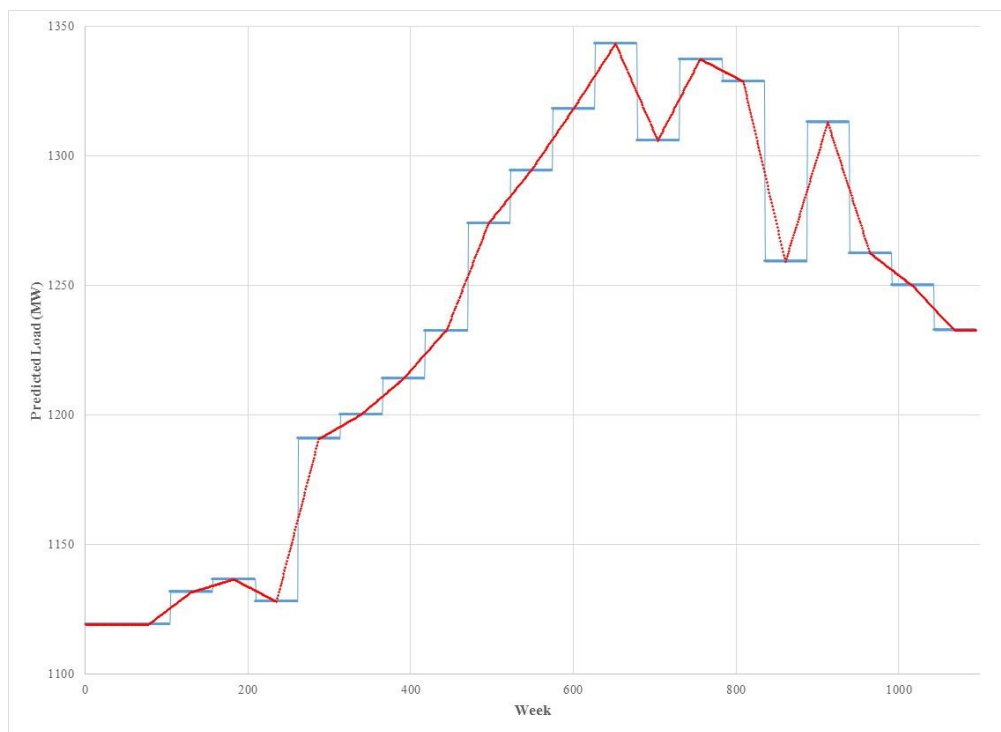


Figure A.1. The Weekly Step of the Annual Prediction (blue)

and the Moving Average (red)



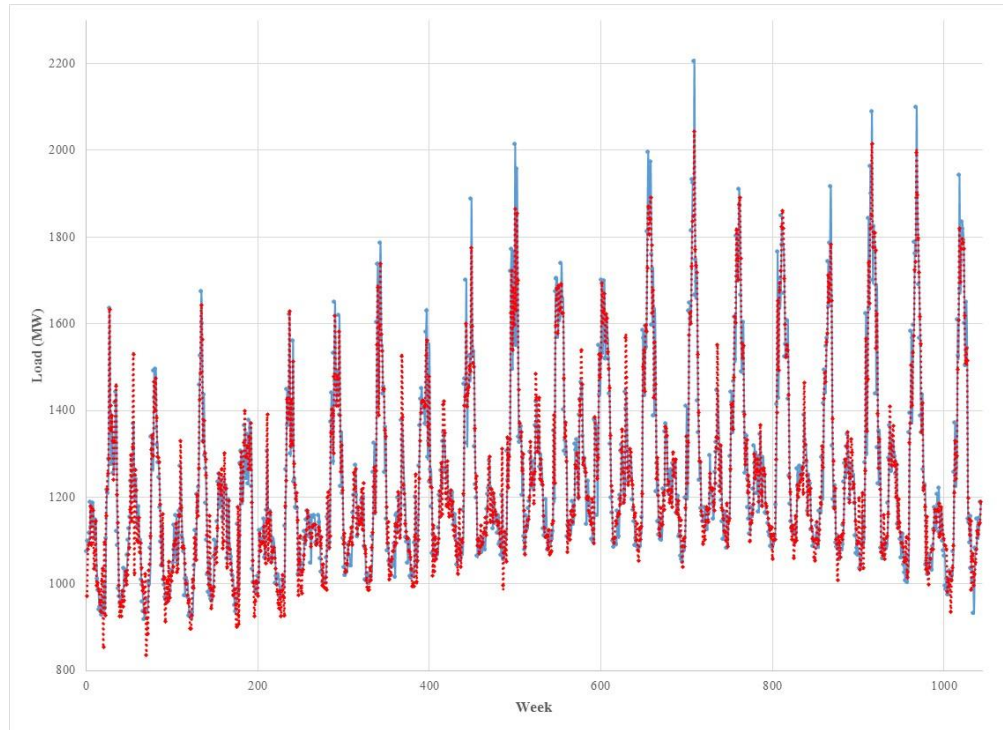


Figure A.2. The Weekly Average of the Hourly Load (blue solid) and the Predicted Load (red dashed) over 20 years

## 2. Graphs Related to the Long-Term Approach

Figures A.3 through A.12 displays the spline coefficients estimates obtained by fitting cubic splines to hourly load data for each month. The graphs clearly show that a random-walk behavior suggested by Harvey and Koopman (1993) is not present and the spline coefficients behave more like seasonal time series. Figures A.4 through A.7 show an anomalous behavior of four of the ten spline coefficients around month 110, which corresponds to approximately November, 2001 through February 2002. The exact reason for this phenomenon is not apparent. Investigation of the load curves during this period show some anomalous behavior during weekdays but not weekends.

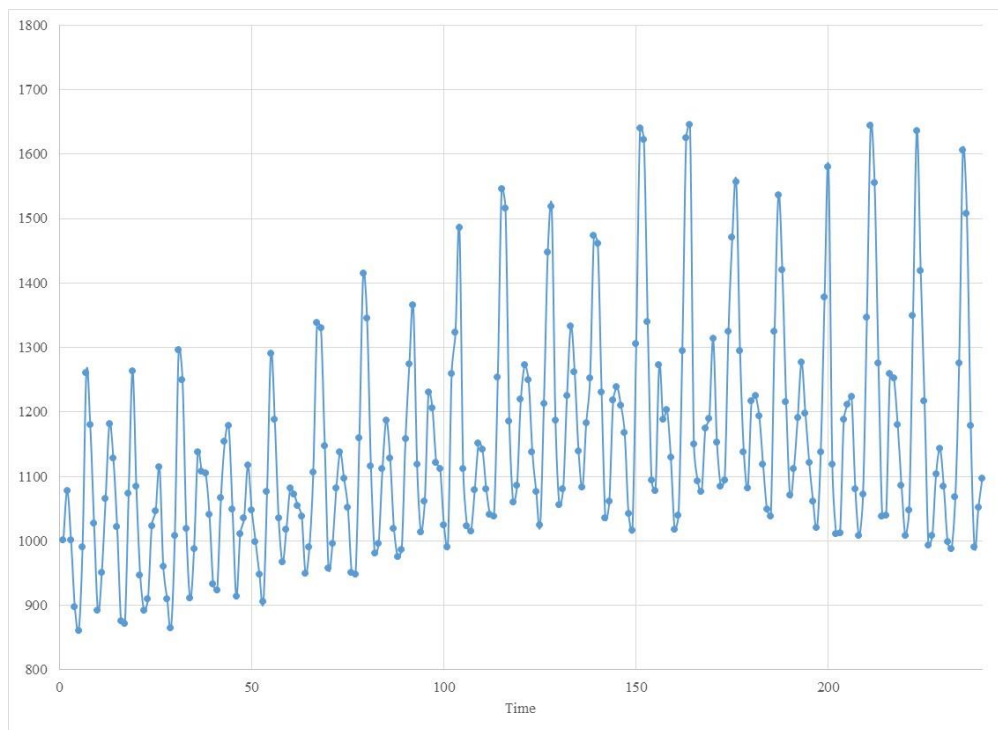


Figure A.3. The Time Series of the Spline Coefficient ( $b_0$ )

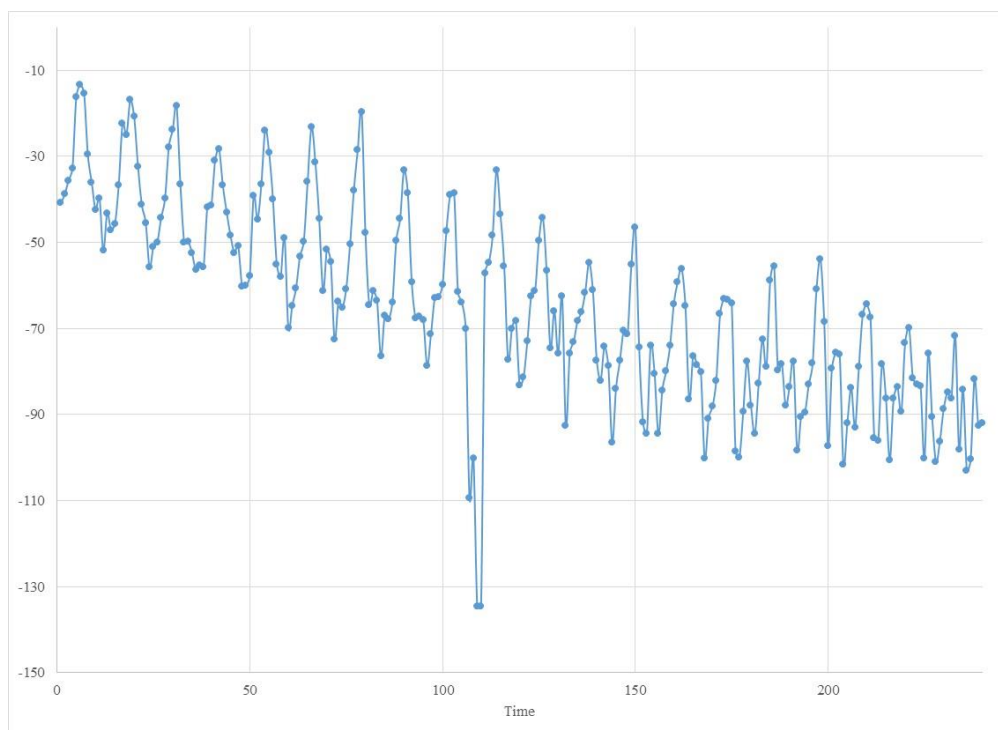


Figure A.4. The Time Series of the Spline Coefficient ( $b_1$ )

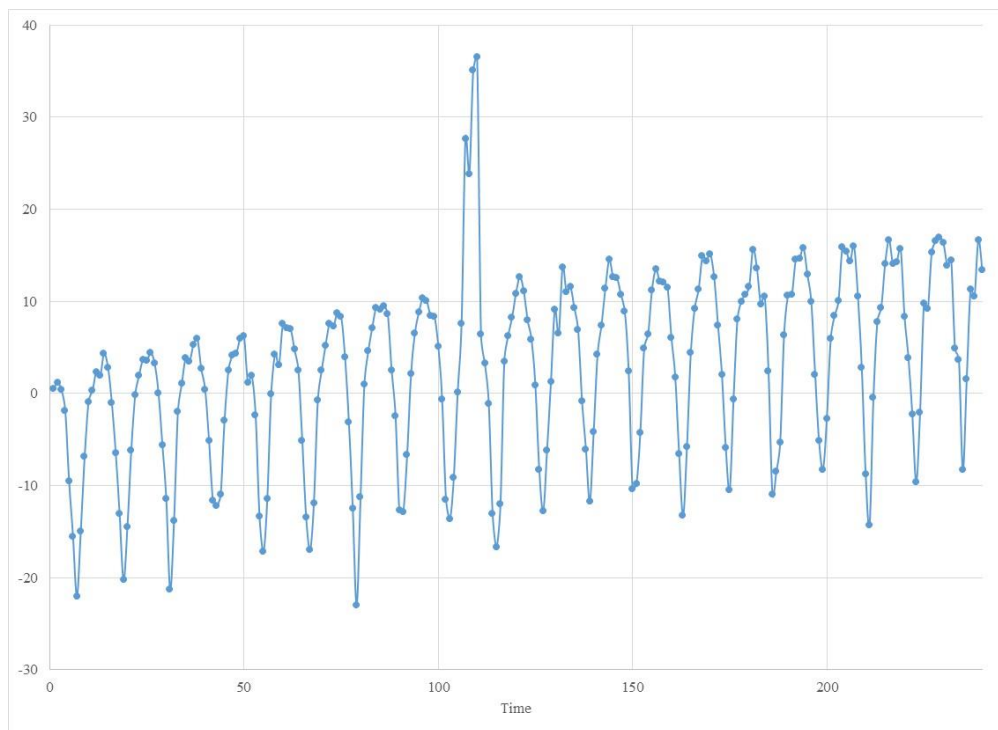


Figure A.5. The Time Series of the Spline Coefficient ( $b_2$ )

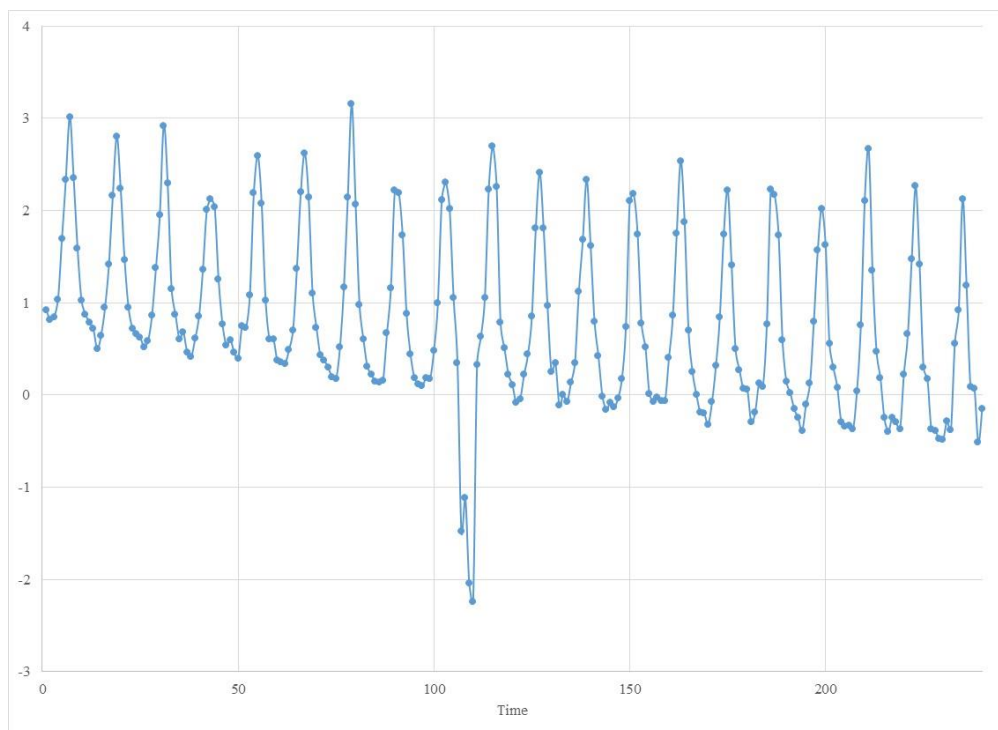


Figure A.6 The Time Series of the Spline Coefficient ( $b_3$ )

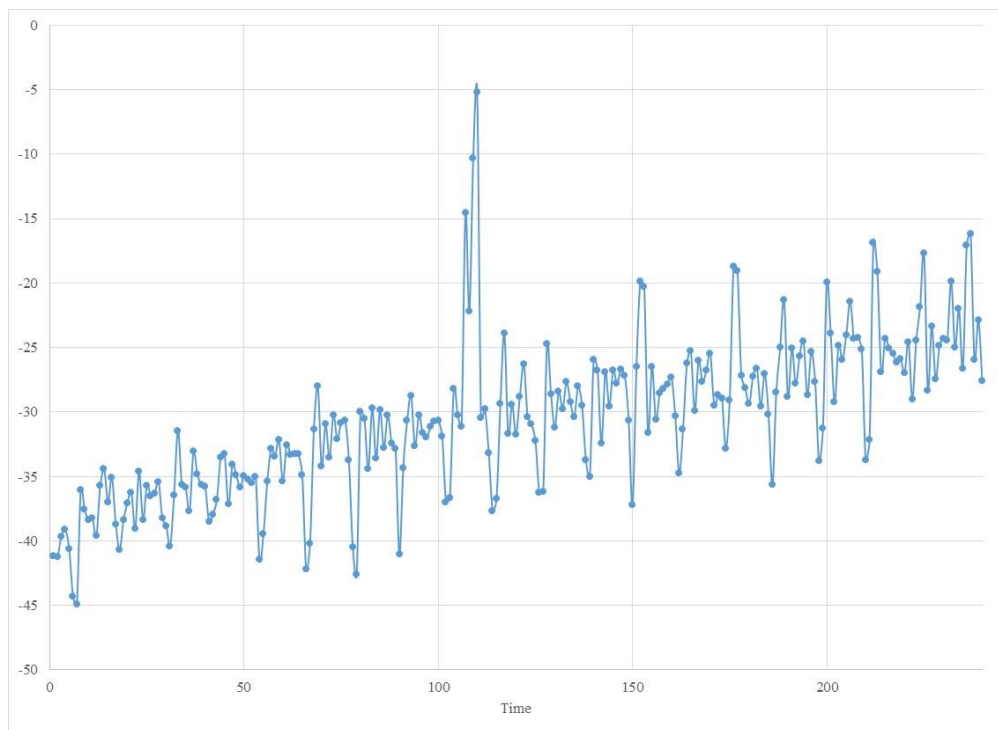


Figure A.7 The Time Series of the Spline Coefficient ( $b_4$ )

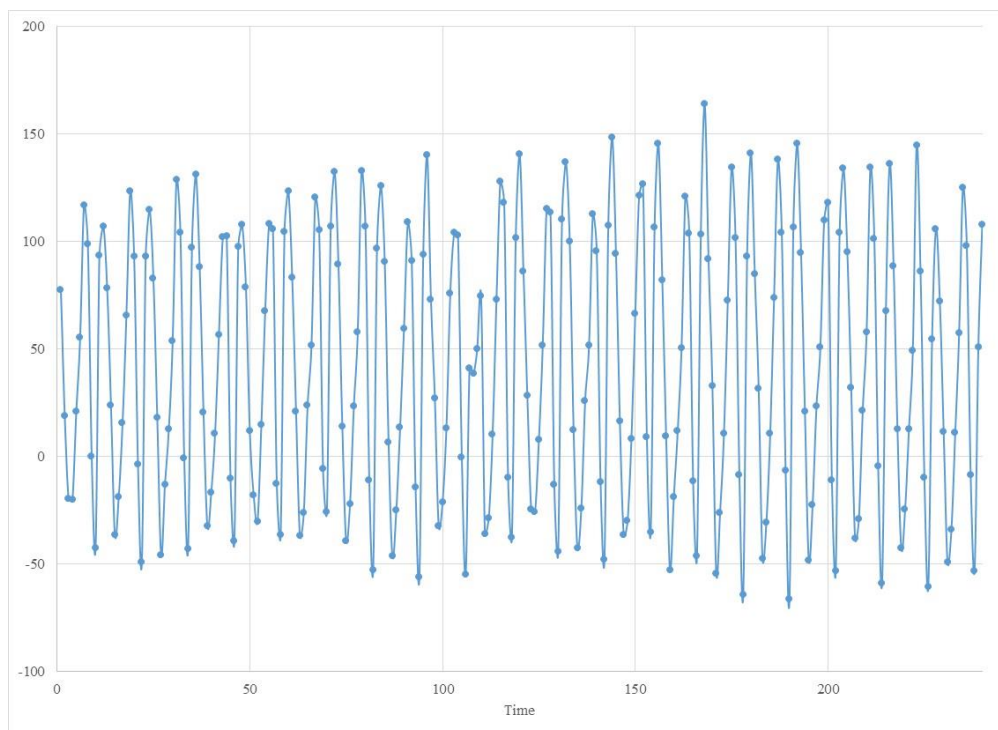


Figure A.8. The Time Series of the Spline Coefficient ( $b_5$ )

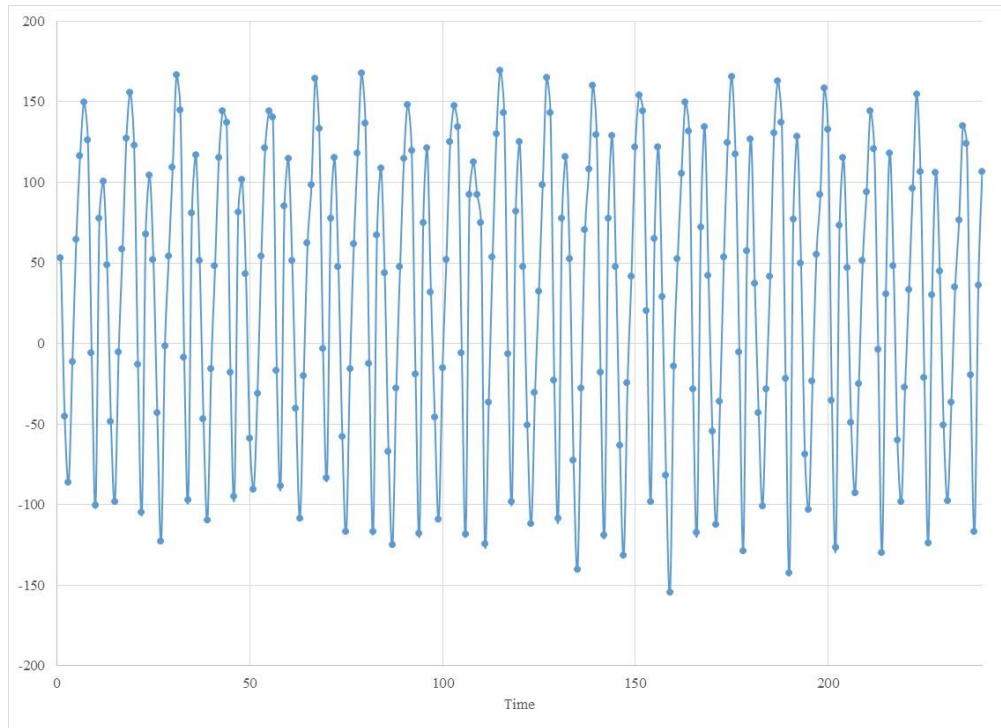


Figure A.9. The Time Series of the Spline Coefficient ( $b_6$ )

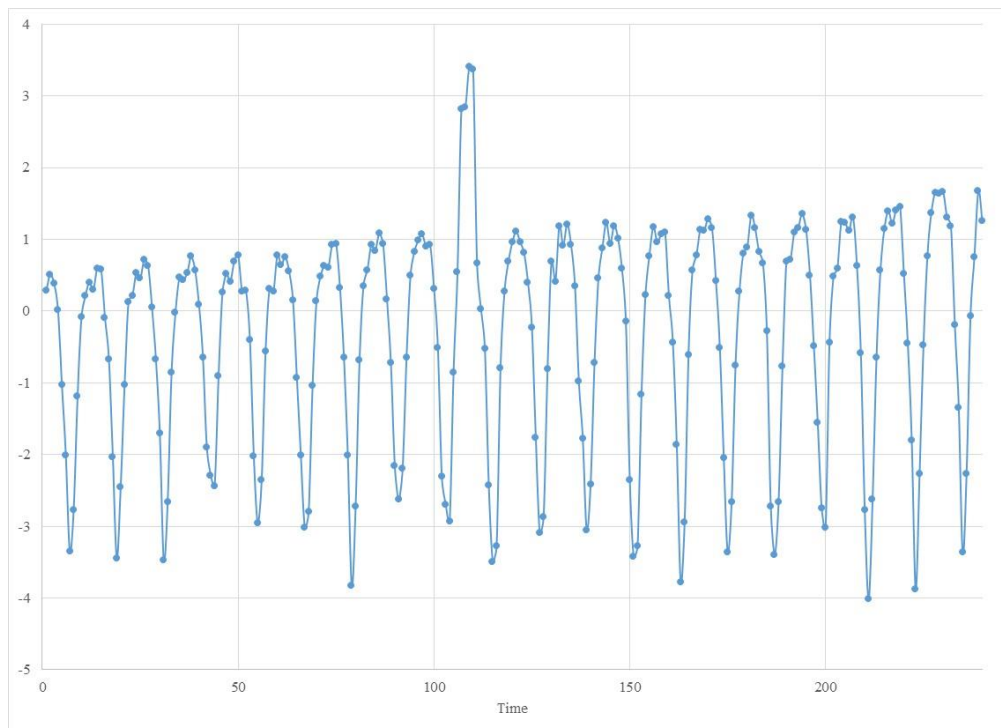


Figure A.10. The Time Series of the Spline Coefficient ( $b_7$ )

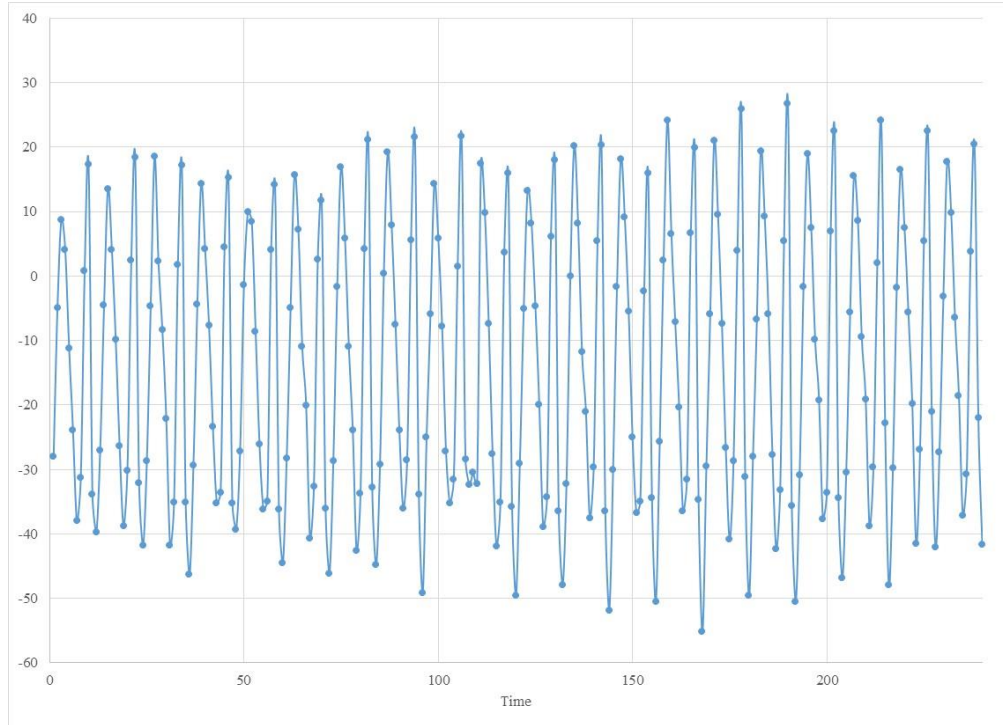


Figure A.11. The Time Series of the Spline Coefficient ( $b_8$ )

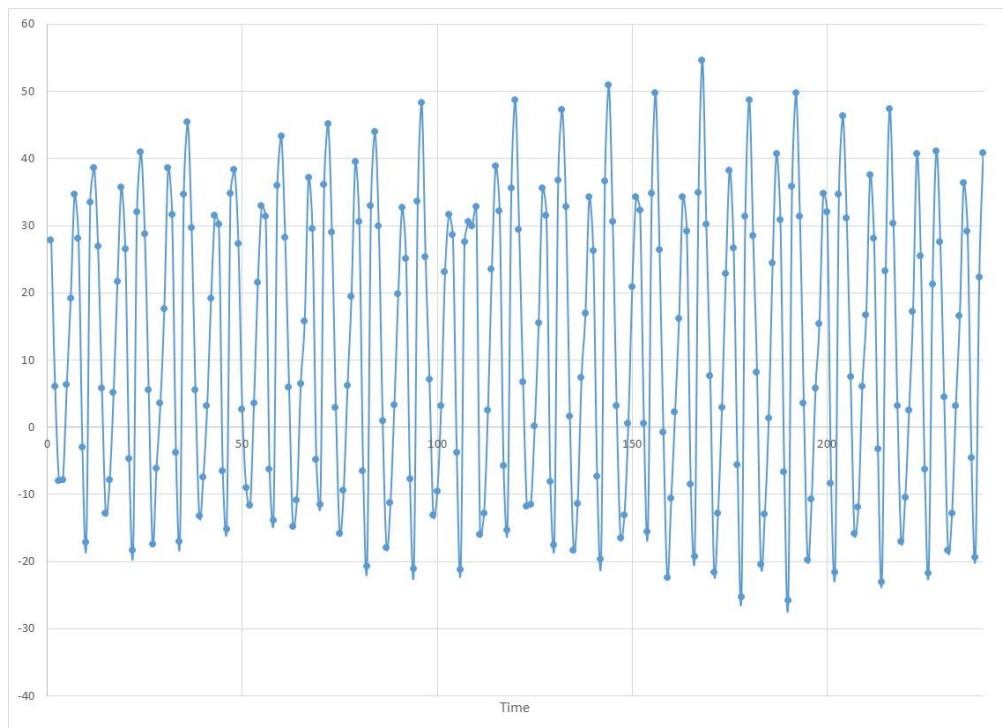


Figure A.12. The Time Series of the Spline Coefficient ( $b_9$ )

### 3. Graphs Related to the Functional Principal Component Analysis Approach

These graphs show the de-trended daily observed load curve for each season.

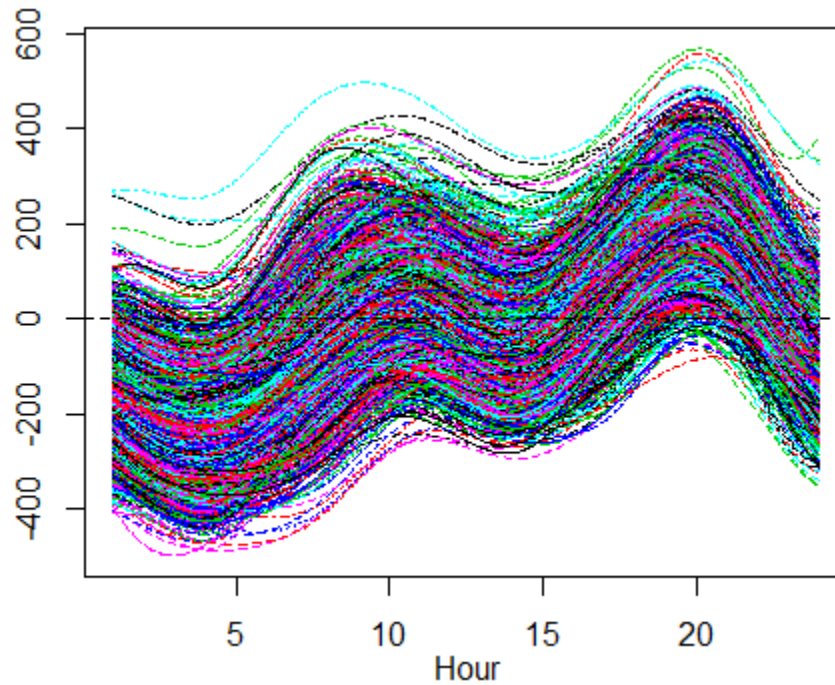


Figure A.13. The Daily 24-hour Profile over 20 Years – Winter Season

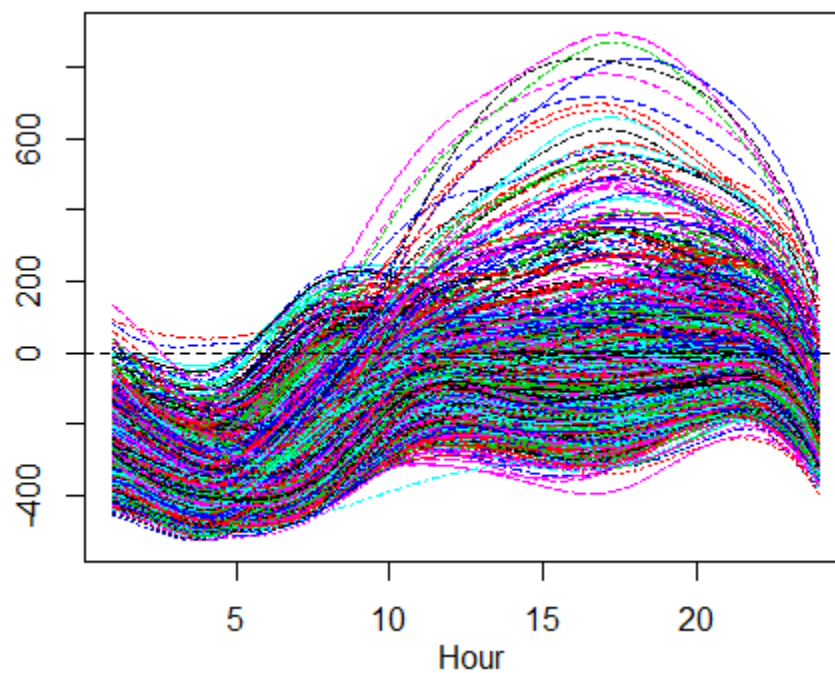


Figure A.14. The Daily 24-hour Profile over 20 Years – Spring Season

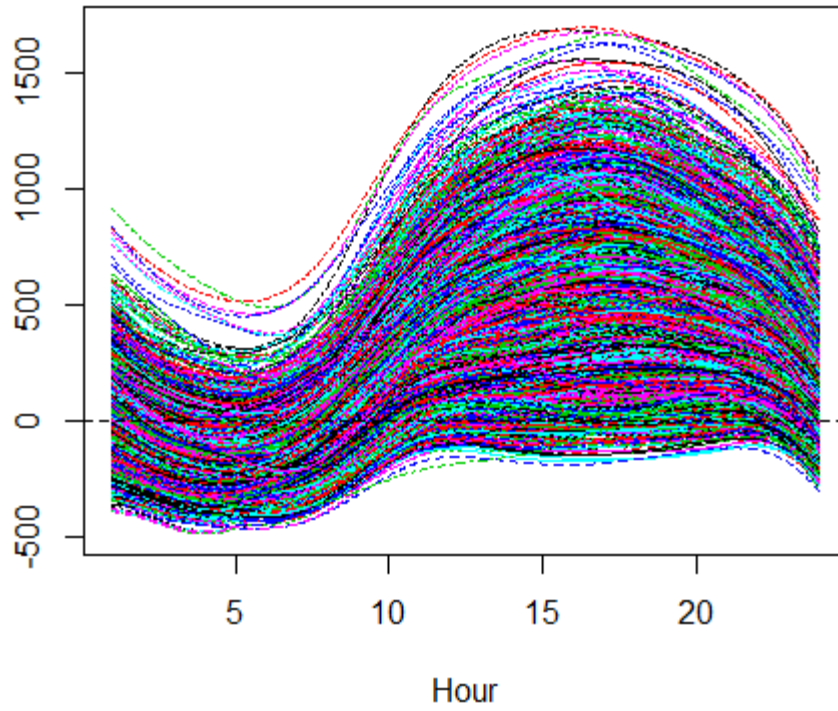


Figure A.15. The Daily 24-hour Profile over 20 Years – Summer Season

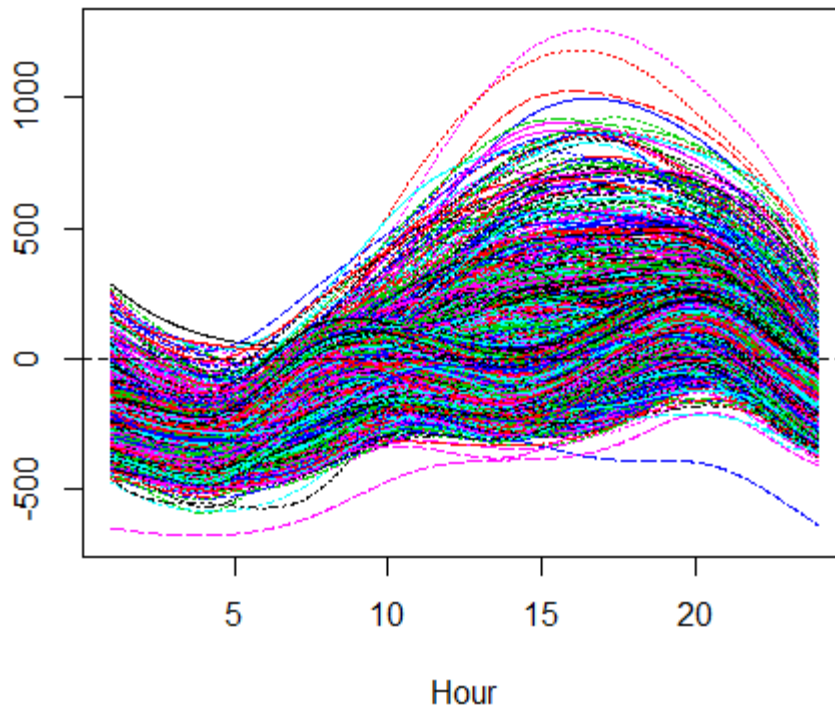


Figure A.16. The Daily 24-hour Profile over 20 Years – Fall Season



#### 4. Graphs Related to the Comparison between the Three Approaches

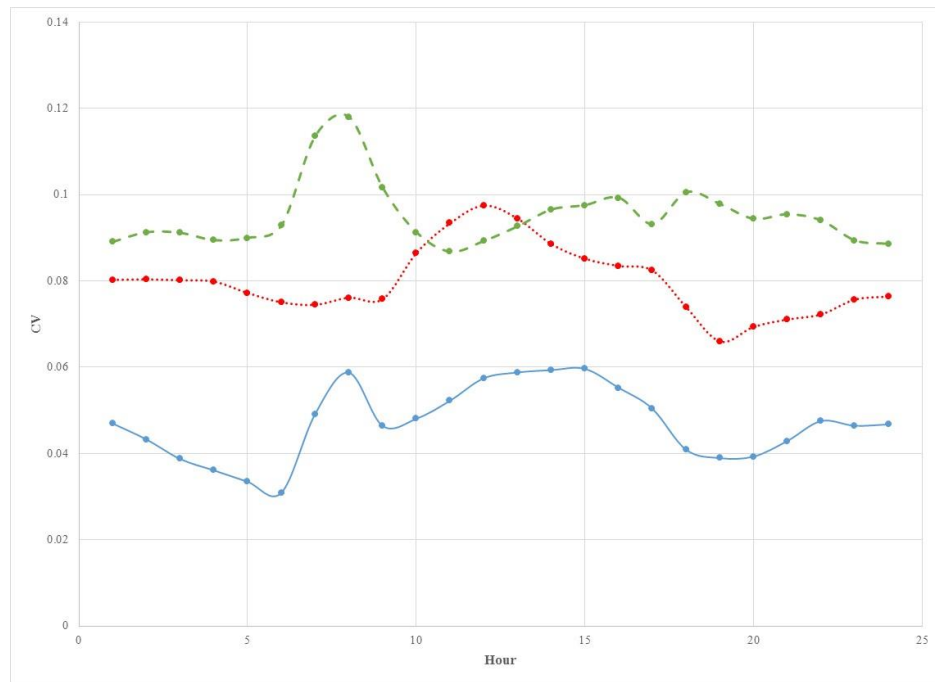


Figure A.17. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Jan.

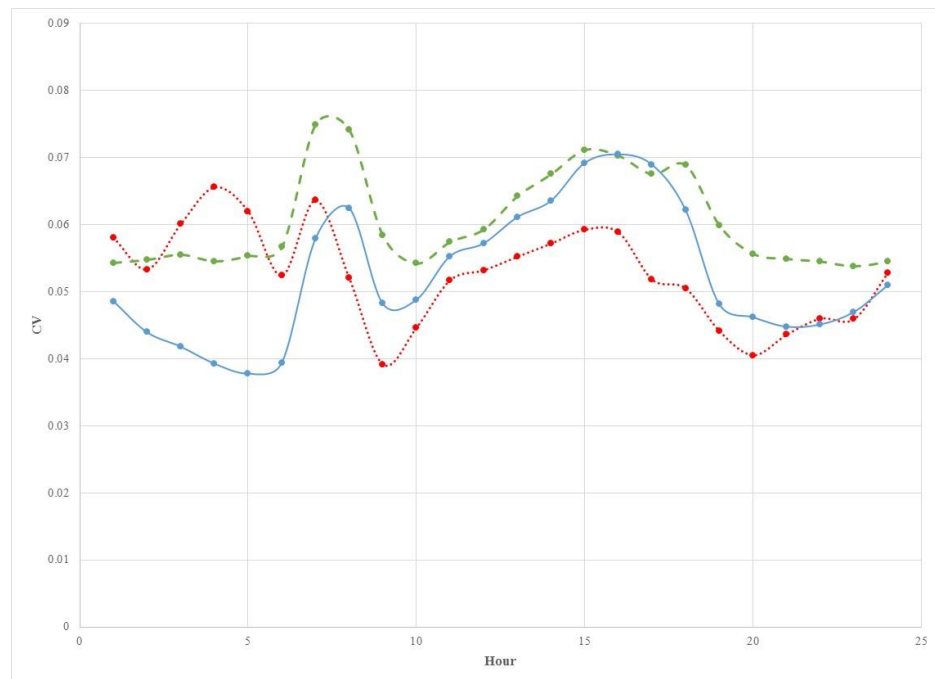


Figure A.18. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Feb.

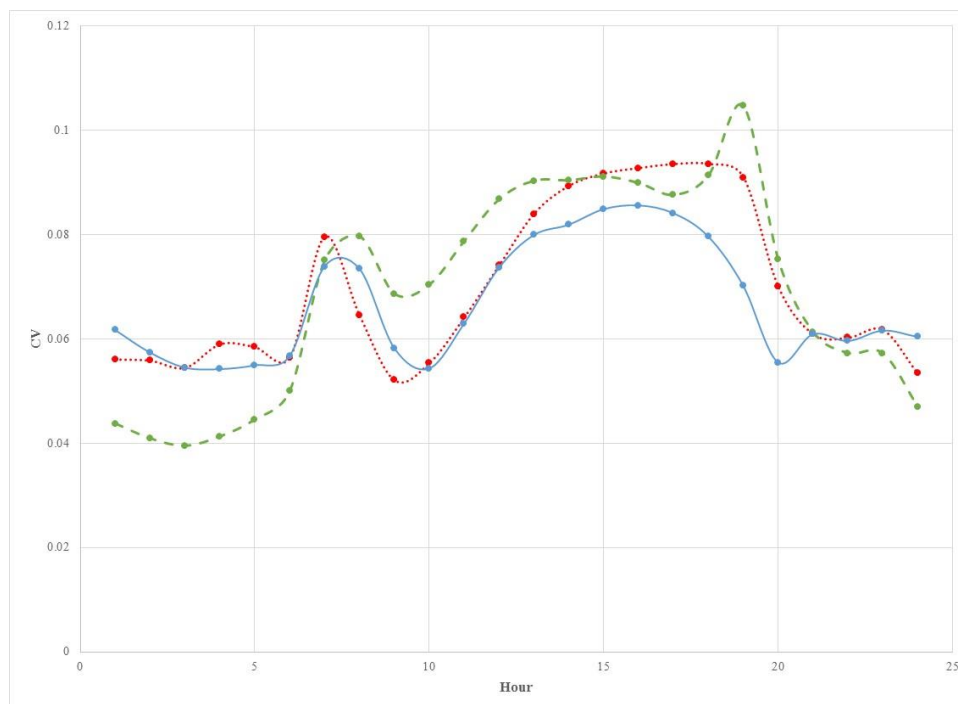


Figure A.19. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Mar.

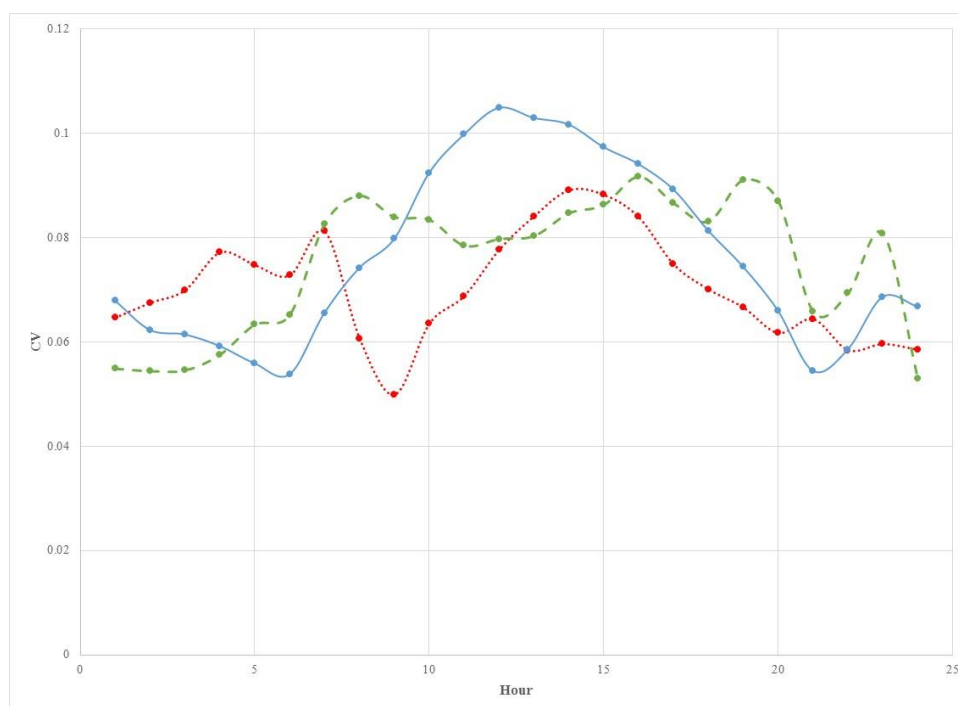


Figure A.20. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Apr.

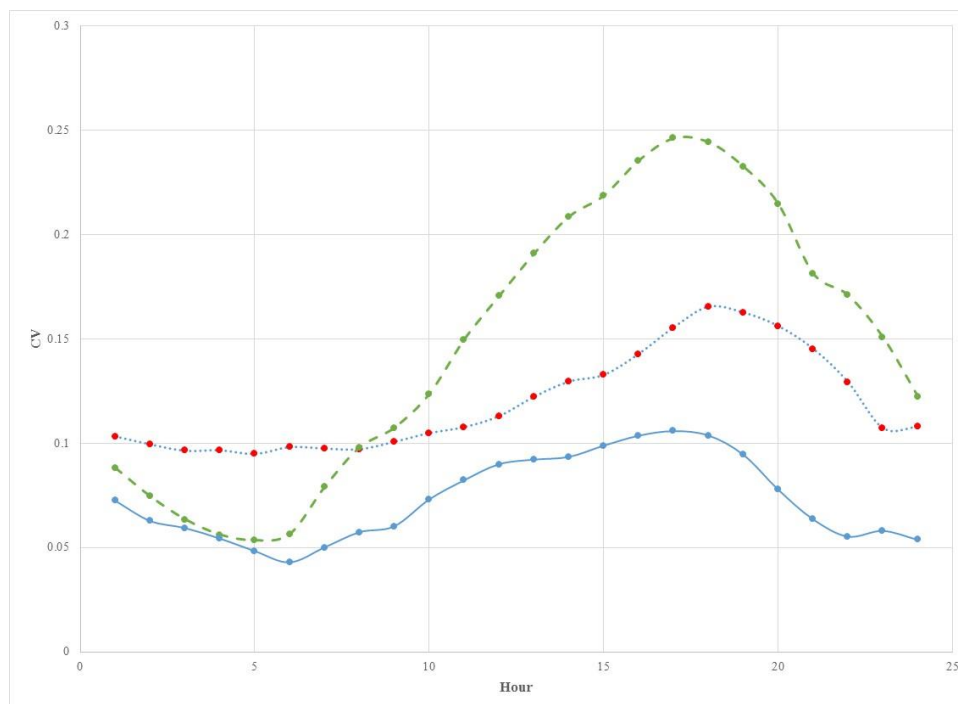


Figure A.21. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – May.

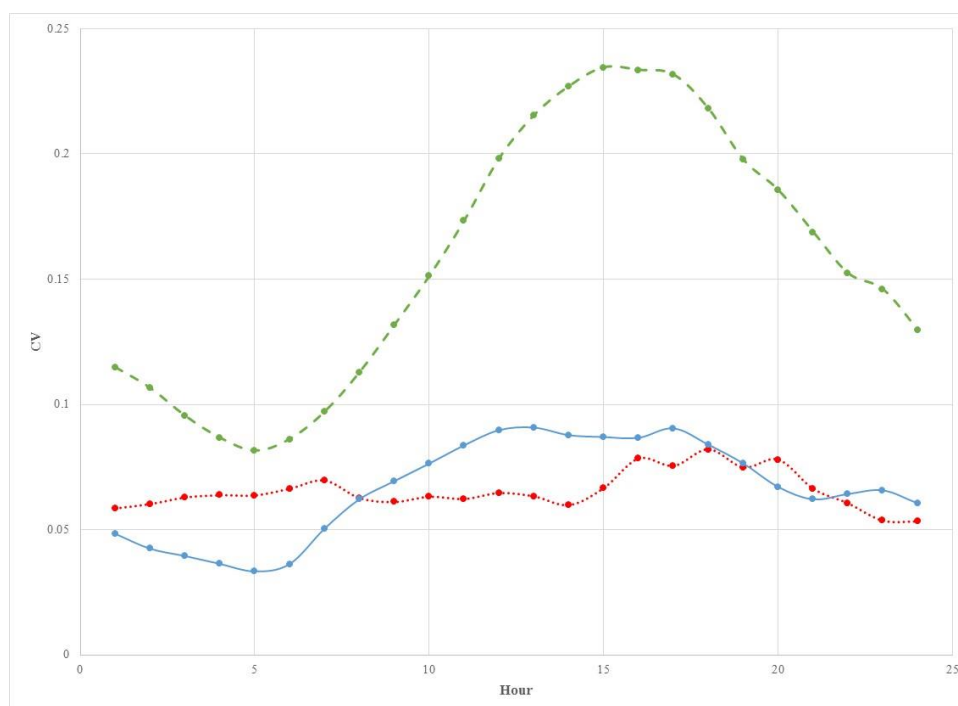


Figure A.22. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Jun.

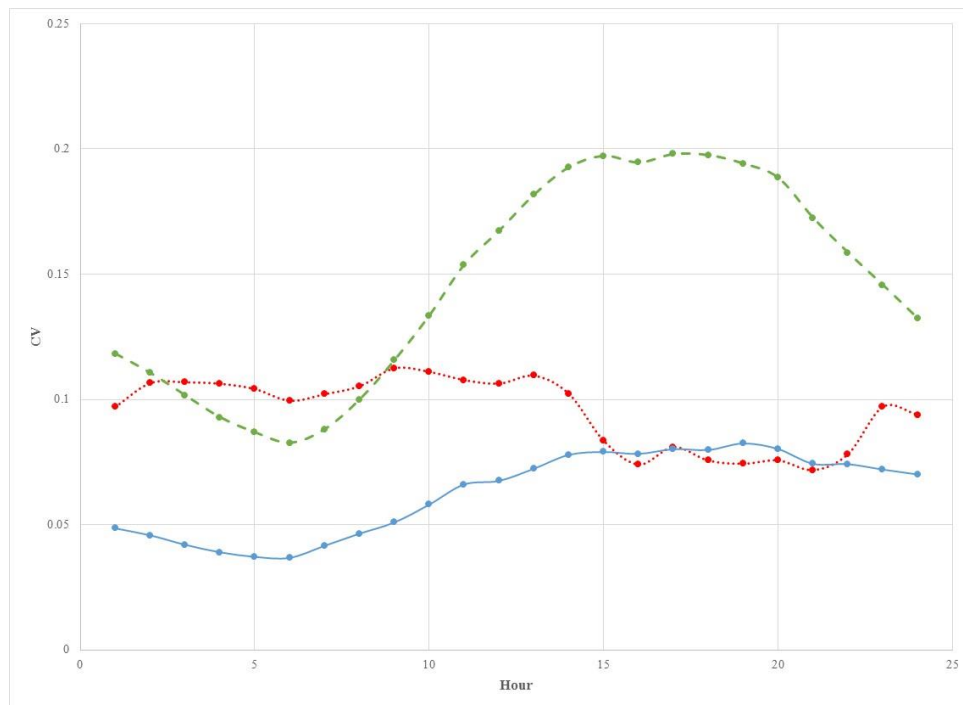


Figure A.23. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Jul.

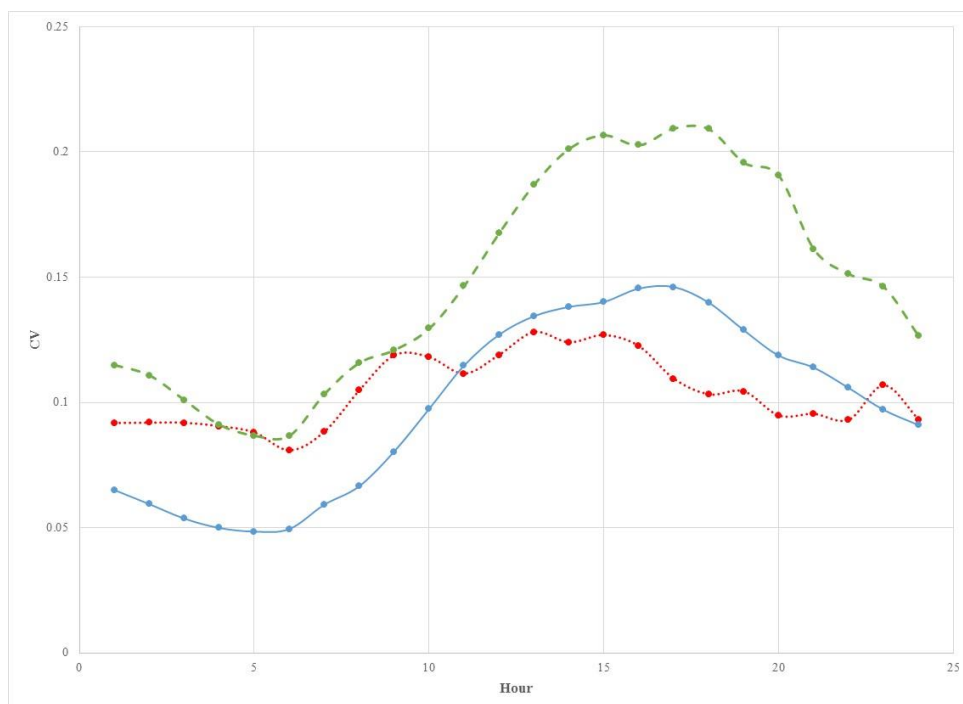


Figure A.24. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Aug.

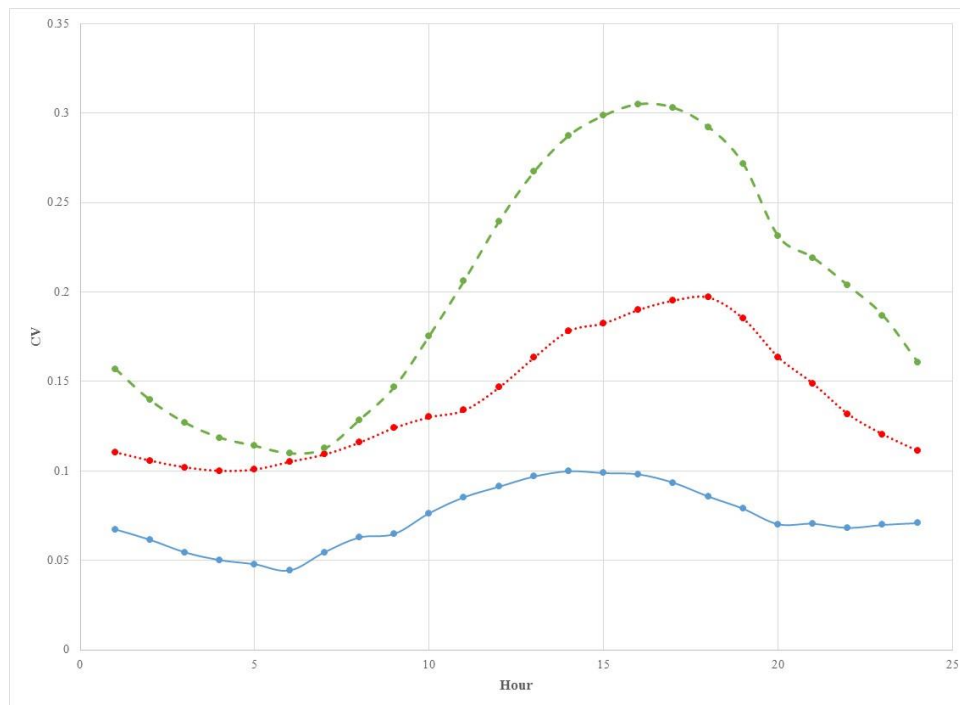


Figure A.25. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Sep.

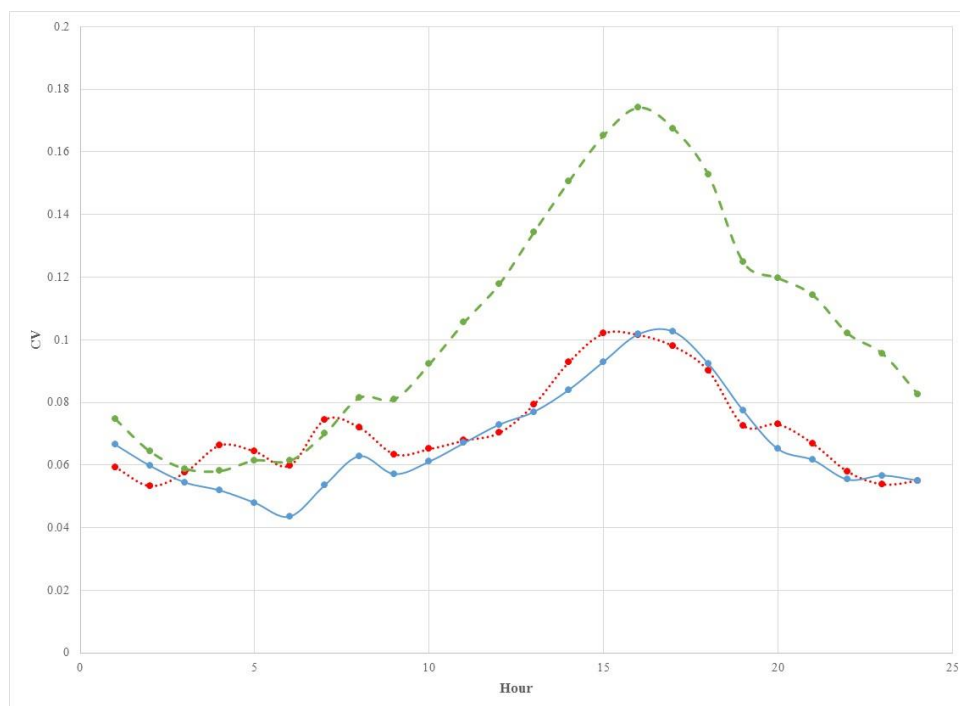


Figure A.26. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Oct.

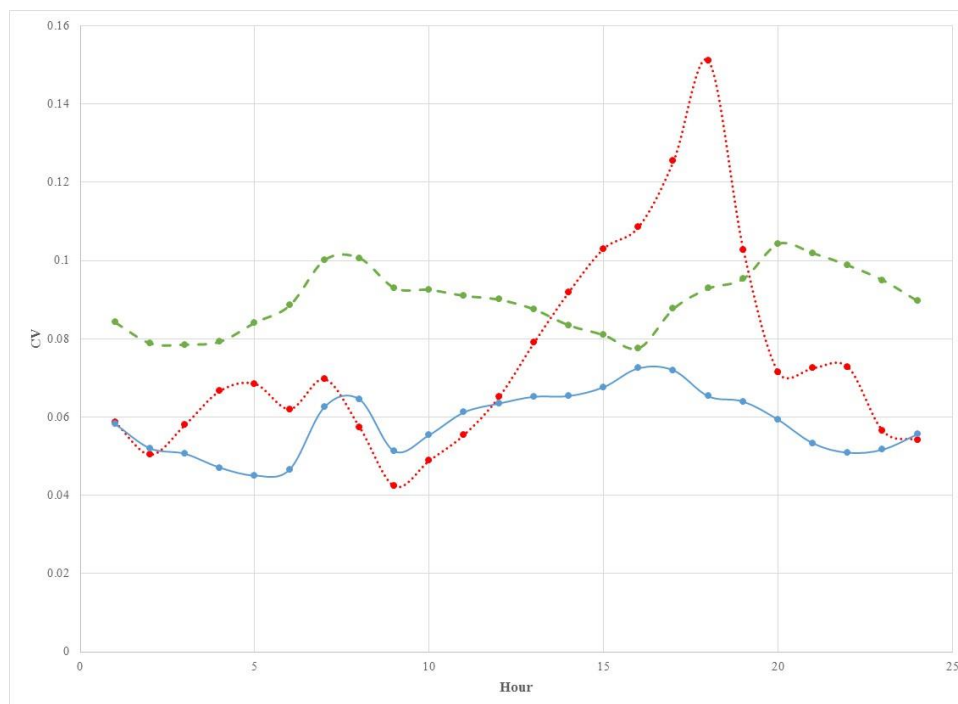


Figure A.27. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Nov.

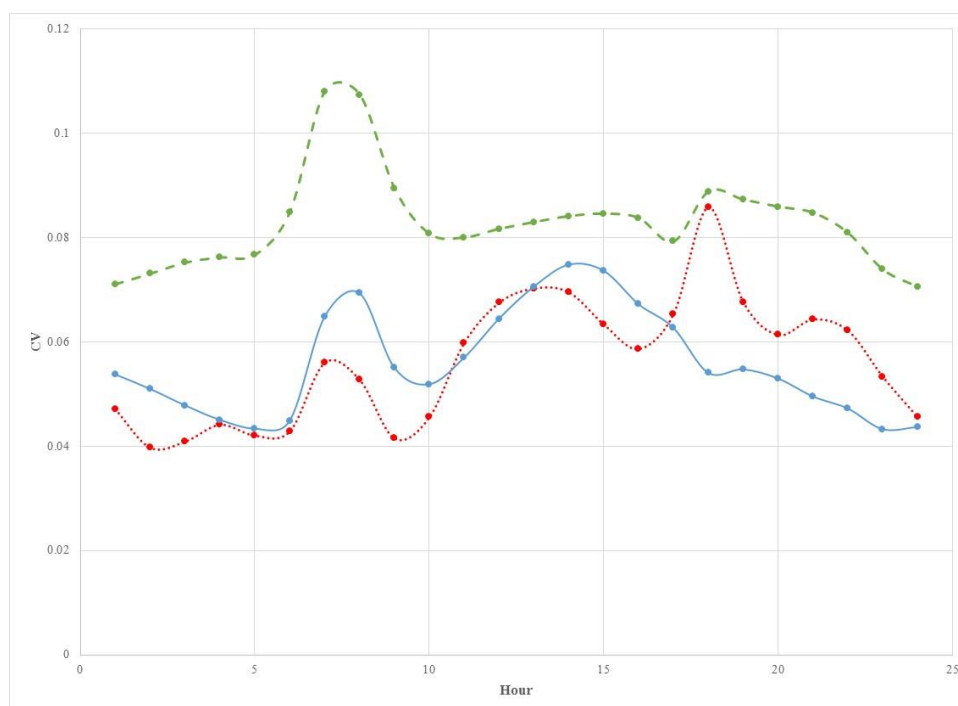


Figure A.28. The CV of the Electricity Load of the Three Models per Hour; Model 1 (blue solid), Model 2 (red dotted), Model 3 (green dashed) – Dec.

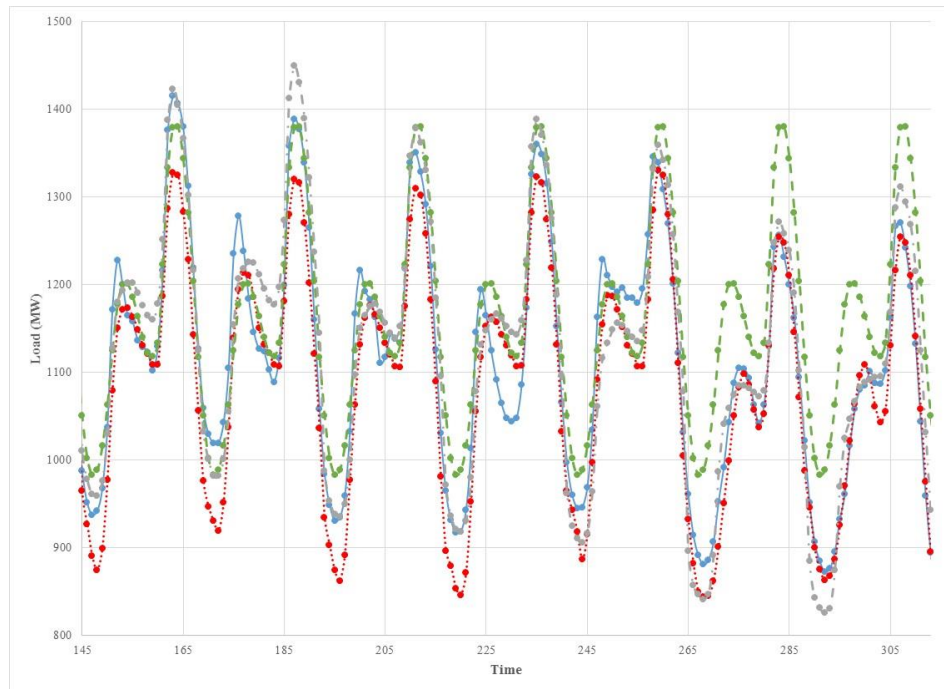


Figure A.29. The Comparison between the Actual Load (blue - solid) and the three Approaches: Approach 1 (red), Approach 2 (green), and Approach 3 (gray) - a Week in the Mid-Winter Season

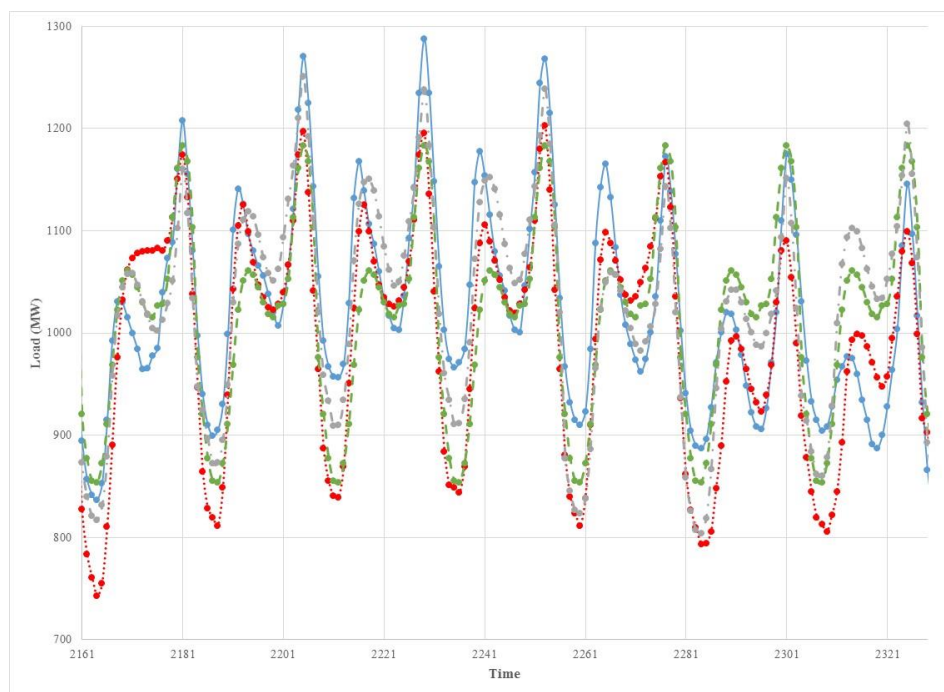


Figure A.30. The Comparison between the Actual Load (blue - solid) and the three Approaches: Approach 1 (red), Approach 2 (green), and Approach 3 (gray) - a Week in the Mid-Spring Season

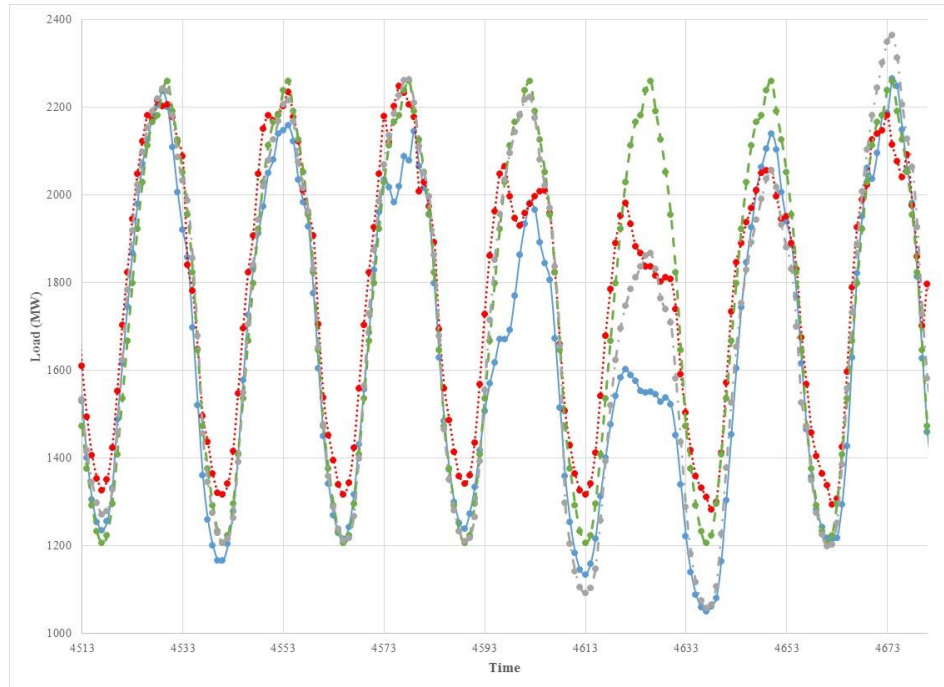


Figure A.31. The Comparison between the Actual Load (blue - solid) and the three Approaches: Approach 1 (red), Approach 2 (green), and Approach 3 (gray) - a Week in the Mid-Summer Season

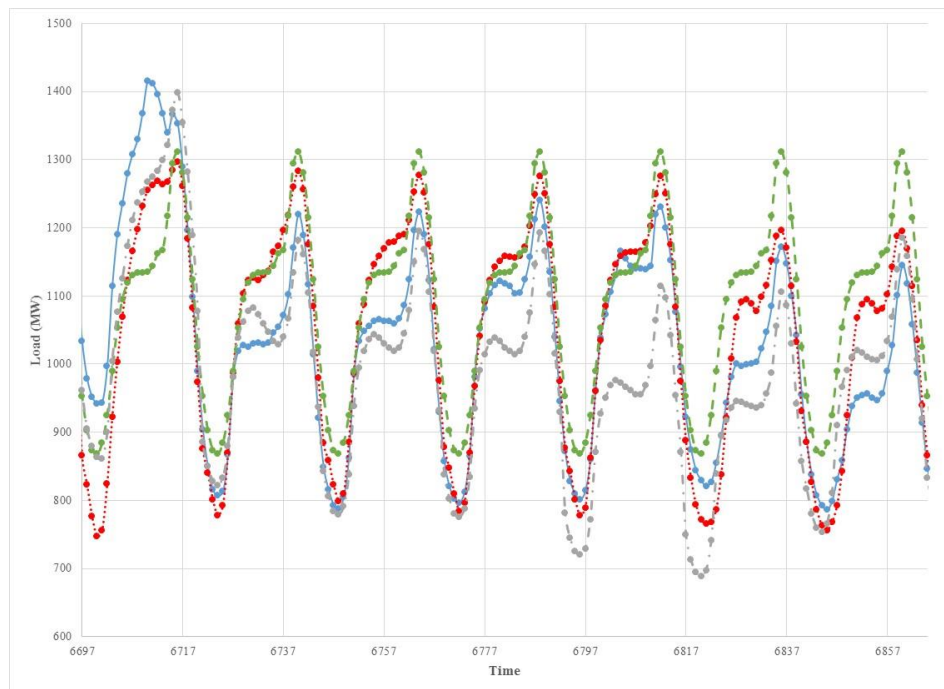


Figure A.32. The Comparison between the Actual Load (blue - solid) and the three Approaches: Approach 1 (red), Approach 2 (green), and Approach 3 (gray) - a Week in the Mid-Fall Season



## BIBLIOGRAPHY

1. H. K. Alfares and M. Nazeerudin, "Electric load forecasting: Literature survey and classification of methods," *International Journal of System Science*. 33 (1), pp. 23-34, 2002.
2. M. Alkhatami, "Introduction to electric load forecasting methods," *Journal of Advanced Electrical and Computer Engineering*. 2 (1), pp. 1-12, 2015.
3. L. F. Amaral, R. C. Souza, and M. Stevenson, "A smooth transition periodic autoregressive (ATPAR) model for short-term load forecasting," *International Journal of Forecasting*. 24, 603-615, 2008.
4. S. Annamareddi, S. Gopinathan, and B. Dora, "A simple hybrid model for short-term load forecasting," *Journal of Engineering*. 2013 (5), 23-34, 2013.
5. P. Arumugam and V. Anithakumari, "Seasonal time series and transfer function modeling for naturalrubber forecasting in India," *International Journal of Computer Trends and Technology*. 4 (5), 1366-1370, 2013.
6. A. Azzalini and B. Scarpa, *Data Analysis and Data Mining*. Oxford University Press. 2012.
7. A. Baziar and A. Kavousi-Fard, "Short term load forecasting using a hybrid model based on support vector regression," *International Journal of Scientific & Technology Research*. 4 (5), 189-195, 2015.
8. V. Bianco, O. Manca, and S. Nurdini, "Electricity consumption forecasting in Italy using linear regression models," *Energy*. 34, 1413-1421, 2009.
9. D. Bunn and E. Farmwe, *Comparative Models for Electrical Load Forecasting*. New York: Wiley. 1985.
10. B. L. Cabrera and F. Schulz, "Forecasting generalized quantiles of electricity demand: a functional data approach," *SFB 649, Economic Risk Berlin*, 2014.
11. J. R. Cancelo, A. Espasa, and R. Grafe, "Forecasting the electricity load from one day to one week ahead for the Spanish system operator," *International Journal of Forecasting*. 24 (4), 588-602, 2008.
12. H. Cho, Y. Goude, X. Brossat, and Q. Yao, "Modeling and forecasting daily electricity load curves: A hybrid approach," *Journal of the American Statistical Association*. 108 (501), 7-21, 2013.
13. M. H. Choueiki, C. A Mount-Campell, and S. Ahalt, "Implementing a weighted least squares procedure in training a neural network to solve the short-term load forecasting problem," *IEEE Transactions on Power Systems* 12 (4), 1689-1694, 1997.

14. T. W. S. Chow and C. T. Leung, "Nonlinear autoregressive neural network model for short-term load forecasting," *IEE Proceedings: Generation Transmission, and Distribution*. 142 , 500-506, 1996.
15. S. Dancose and J. Angeles, "Modeling and simulation of flexible beams using cubic splines and zero-order holds," *Progress in System and Control Theory*. 4 (2), 553-564, 1990.
16. P. K. Dash, A. C. Liew, and S. Rahman, "Comparision fuzzy neural network for the generation of daily average and peak load profiles," *International Journal of System Science*. 26, 2091-2106, 1995.
17. V. Dordonnat, S. J. Koopman, M. Ooms, A. Dessertaine, and J. Collet, "An hourly periodic state space model for modeling French national electricity load," *Tinbergen Institute Discussion*. Paper No. 2008-008/4.
18. M. U. Fahad and N. Arabab, "Factor affecting short term load forecasting," *Journal of Clean Energy Technologies* 2 (4), 305-309, 2014.
19. J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting* 22 (3), 443-473, 2006.
20. A. Harvey and S. J. Koopman, "Forecasting hourly electricity demand using time-varying splines," *Journal of the American Statistical Association*. 88 (424), 1228-1236, 1993.
21. J. Hinman and E. Hickey, "Modeling and forecasting short-term electricity load using regression analysis," 2009, *The Calm before the Storm Conference*.
22. C. L. Hor, S. J. Watson, and S. Majithia, "Analyzing the impact of weather variables on monthly electricity demand," *IEEE Transactions on Power Systems* 51 (10), 4942-4956, 2005.
23. R. J. Hyndman and M. Shahid Ullah, "Robust forecasting of mortality and fertility rates: a functional data approach," *Computational Statistics and Data Analysis* 20 (4), 2078-2085, 2007.
24. R. J. Hyndman and H. L. Shang, "Functional time series forecasting," *Journal of the Korean Statistical Society*. 38 (3), 199-211, 2009.
25. K. H. Kim, J. K. Park, K. J Hwang, and S. H Kim, "Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert system," *IEEE Transactions on Power Systems* 10 (3), 1534-1539, 1995.
26. D. Kosiorowski, "Functional regression in short-term prediction of economic time series," *Statistics in Transition*, 15 (4), 611-626, 2014.
27. J. M. Liu, R. Chen. L. Liu, and J. Harris, "A semi-parametric time series approach in modeling hourly electricity loads," *Journal of Forecasting*. 25, 537-559, 2006.
28. H. Lütkepohl, *New Introduction to Multiple Time series Analysis* 2<sup>nd</sup> Ed. Springer. 2007

29. Z. Mohamed and P. Bodger, "Forecasting electricity consumption in New Zealand using economic and demographic variables," *Energy*. 30, 1833-1843, 2005.
30. M. A. Momani, "Factors affecting electricity demand in Jordan," *Energy and Power Engineering*, 5, 50-58, 2013.
31. J. Nowicka-Zagrajek and R. Weron, "Modeling electricity loads in California: ARMA models with hyperbolic noise," *Signal Processing*. 82, 1903-1915, 2002.
32. A. Pielow, R. Sioshansi, and M. Roberts, "Modeling short-run electricity demand with long-term growth rates and consumer price elasticity in commercial and industrial sectors," *Energy*. 46 (1), 533-540, 2012.
33. PJM. Historical load data. <http://www.pjm.com/markets-and-operations/ops-analysis/historical-load-data.aspx> (Accessed 2014).
34. J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies*. Springer. 2002.
35. J. O. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer. 2009.
36. J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied Regression Analysis: A Research Tool 2<sup>nd</sup> ed.* Springer. 1998.
37. H. L. Shang, "A survey of functional principal component analysis," *Advance in Statistical Analysis*. 98 (2), 121-142, 2013.
38. H. L. Shang, "Functional time series approach for forecasting very short-term electricity demand," *Journal of Applied Statistics*. 40 (1), 152-168, 2013.
39. H. L. Shang and R. J. Hydman, "Nonparametric time series forecasting with dynamic updating," *Mathematical and Computer in Simulation*. 81 (7), 1310-1324, 2011.
40. S. A. Soliman and A. M. Al-Kandari, *Electricity Load Forecasting, Modeling and Model Construction*. ELSEVIER, 2010.
41. J. Taylor, "An evaluation of method for every short-term load forecasting using minute-by-minute British data," *International Journal of Forecasting*. 24, 645-658, 2008.
42. J. W. Taylor and P. E. McSharry, "Short-term load forecasting Methods: An evaluation based on European data," *IEEE Transaction on Power Systems*. 22 (4), 2213-2219, 2007.
43. The Federal Reserve Bank of St. Louis. Economic Data. <https://research.stlouisfed.org/>
44. [fred2/categories/27312](https://fred2/categories/27312). (Accessed 2014).
45. U.S. Energy Information Administration (eia). Electric Power Monthly with Data for March 2015.

46. E. Valor, V. Meneu, and V. Caselles, "Daily air temperature and electricity load in Spain," *Journal of Applied Meteorology and Climatology*. 40 (8), 1413-1421, 2001.
47. Y. Wnag, *Smoothing Splines: Methods and Applications*. Chapman & Hall/CRC, 2011.
48. E. J. Wegman and I. W. Wright, "Splines in statistics," *Journal of American Statistical Association*. 78 (382), 351-365, 1983.
49. H. L. Willis and J. E. D. Nothcote-Green, "Comparison tests of fourteen distribution load forecasting methods," *IEEE Transaction on Power Apparatus and Systems*. 103 (6), 1190-1197, 1984.
50. S. Wold, "Spline functions in data analysis," *Technometrics*. 16 (1), 1-11, 1974.

## VITA

Abdelmonaem Salem Jornaz was born in Tripoli, Libya on September 15, 1974. He received his Bachelor of Statistics from University of Tripoli, Libya in March 1997. He has also earned a Computer Programing Certificate (1 year Certificate) from Al-Shumoukh Institute, Tripoli, Libya in June 1999. He received his Masters of Science in Statistics from University of Tripoli, Libya in March 2004. He received his Masters of Science in Applied Mathematics from Missouri University of Science and Technology, Rolla, Missouri in May 2014. In May 2016, he received his Ph.D. in Mathematics with emphasis in Statistics from Missouri University of Science and Technology, Rolla, Missouri.

He worked as a statistical analyst at Tripoli Company of Water and Wastewater during the period 2001 – 2007 and as a faculty member at Azzaytuna University, Tarhunah, Libya, from 2007. During this period he worked as a part time lecturer at three universities and four institutes, during this time, he helped some graduate students from other fields with the statistical analyses for their theses.

He started his graduate program at Missouri University of Science and Technology in January 2011. He started work as graduate teaching assistance at Missouri University of Science and Technology in August 2012. He received a GTA Teaching Excellence Award from Mathematics and Statistics department for 2013-2014, and honorable mention for the same award two other years, 2012-2013 and 2014-2015.