University of Venda

SHORT TERM LOAD FORECASTING USING QUANTILE REGRESSION WITH AN APPLICATION TO THE UNIT COMMITMENT PROBLEM

Masters Dissertation

By

Moshoko Emily Lebotsa
16023607

January 2018

This dissertation is submitted in fulfilment for the requirements of the degree of Master of Science in Statistics at the University of Venda, Thohoyandou

UNIVERSITY OF VENDA
DEPARTMENT OF
STATISTICS

The undersigned hereby certify that they have read and recommend to the School of Mathematical and Natural Sciences for the acceptance of a dissertation entitled "**Short term load forecasting using quantile regression with an application to the unit commitment problem**" by **Moshoko Emily Lebotsa** in fulfillment of the requirements for the degree of **Master of Science**.

Dated: January 2018

Supervisor: _____
Dr. Caston Sigauke

Co-Supervisor: _____
Dr. Alphonce Bere

_____

Declaration

I, Moshoko Emily Lebotsa (student number: 16023607), declare that the dissertation titled: "Short term load forecasting using quantile regression with an application to the unit commitment problem" hereby submitted for qualification of Masters degree in Statistics at the University of Venda has not been submitted previously for any other degree at this university or any other institution. This dissertation is my own work and all significant references from the work of other authors have been duly acknowledged.

Signature:_____ Date:_____

Abstract

Generally, short term load forecasting is essential for any power generating utility. In this dissertation the main objective was to develop short term load forecasting models for the peak demand periods (i.e. from 18:00 to 20:00 hours) in South Africa using. Quantile semi-parametric additive models were proposed and used to forecast electricity demand during peak hours. In addition to this, forecasts obtained were then used to find an optimal number of generating units to commit (switch on or off) daily in order to produce the required electricity demand at minimal costs. A mixed integer linear programming technique was used to find an optimal number of units to commit. Driving factors such as calendar effects, temperature, etc. were used as predictors in building these models. Variable selection was done using the least absolute shrinkage and selection operator (Lasso). A feasible solution to the unit commitment problem will help utilities meet the demand at minimal costs. This information will be helpful to South Africa's national power utility, Eskom.

Keywords: *Mixed integer linear programming, peak demand, quantile semi-parametric additive models, short term load forecasting, unit commitment.*

Acknowledgements

Foremost, I would like to offer my heartfelt gratitude to Dr Caston Sigauke, my advisor, for the extended support, guidance and his many suggestions throughout this research.

To Dr Alphonce Bere, my co-advisor, I am grateful for dedicating your time to read my dissertation and for all the suggestions and valuable comments.

My sincere acknowledgement also goes to Mantombi Bashe and the rest of the team that hosted us for a research visit at Eskom. The sessions and presentations you managed to organise in such short time were very informative and helped me understand my research project better.

A very special gratitude goes out to all down to the National Research Foundation (NRF), thanks for the financial support provided to me during this period. Am grateful to Eskom and the South African Weather Service for the data provided.

Last but not least, I would like to thank my daddy, brothers and sisters for always being there for me; your patience, love and endless support kept me going throughout. To my friends and fellow classmates, I thank you for sharing this journey with me.

Thanks everyone for all your encouragement!

Table of contents

List of Figures

List of Tables

List of abbreviations

AIC	Akaike's Information Criterion
AICc	Corrected Akaike's Information Criterion
ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
BIC	Bayesian Information Criterion
CDD(s)	Cooling Degree Days
CDF	Conditional Distribution Function
CV	Cross Validation
AM	Ante Meridiem
DP	Dynamic Programming
GAM	Generalized Additive Model
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
HDD(s)	Heating Degree Days
LASSO	Least Absolute Shrinkage and Selection Operator
LP	Linear Programming
LR	Lagrange Relaxation
LTLF	Long Term Load Forecasting/Forecast(s)
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

MILP	Mixed Integer Linear Programming
MIP	Mixed Integer Programming
MTLF	Mediam Term Load Forecasting/Forecast(s)
MW(s)	Megawatt(s)
NESO	Net Energy Sent Out
OFQR	Optimal Forecasting Quantile Regression
OLS	Ordinary Least Square
PACF	Partial Autocorrelation Function
PM	Post Meridiem
QR	Quantile Regression
QSAM	Quantile Semi-parametric Additive Model
RC	Reduced Cost(s)
RMSE	Root Mean Square Error
SA	South Africa (African)
SAM	Semi-parametric Additive Model
SARIMA	Seasonal Autoregressive Moving Average
SAWS	South African Weather Service
SCUC	Security-Constrained Unit Commitment
Soln	Solution
SP	Stochastic Programming(s)
STD	Standard Deviation
STLF	Short Term Load Forecasting/Forecast(s)
TS	Tabu Search
UC	Unit Commitment

List of symbols

Symbol	Description
^{0}C	Degree Celsius
AdjR^2	Adjusted R-square
R^2	R-Square
p-value	probability value

Research Outputs

A list of research from this dissertation is given below

Peer Reviewed Journal Publications:

Lebotsa, M.E., Sigauke, C., Bere, A., Fildes, R. and Boylan, J. (2018). Short term electricity demand forecasting using partially linear additive quantile regression with an application to the unit commitment problem. Applied Energy, Vol 222: pp 104 - 118

National Conference Presentations:

Lebotsa, M.E., Sigauke, C., Bere, A., Fildes, R. and Boylan, J. (2017). Short term load forecasting using quantile regression with an application to the unit commitment problem. The 59th Annual Conference of the South African Statistical Association, 27 November - 30 December 2017, University of Free State, Ilanga Estate, Bloemfontein, South Africa.

Chapter 1

Introduction

The use of energy is growing rapidly in both developed and developing countries. It is estimated that if the current state of energy consumption continues to grow at the current rate, the world's energy consumption is expected to increase by 50 % before the year 2030 (Suganthi and Samuel, 2012). This has led to many countries seeking solutions or methods that will help with the management of demand, thus making the management of energy demand an important issue (Suganthi and Samuel, 2012).

Generally, energy is presented in many forms, for instance petroleum, electricity, natural gas, etc.(Ghalehkhondabi et al., 2016). However, the focus in this dissertation is only on electricity, hence the term electricity will be used instead of energy. Load demand is also used as a synonym to electricity demand (consumption) in the absence of load shedding as defined by Cottet and Smith (2003).

Load forecasting is defined as an important technique used in estimating the amount of power or electricity needed for future consumptions (Fan and

Hyndman, 2012). According to Pierrot and Goude (2011), load forecasting is essential for every electricity utility in order to maintain the demand and supply balance. The forecasts are also required for many purposes such as system security, rate design, revenue projection as well as for other scheduling activities (Hinman and Hickey, 2009; Hong and Fan, 2016; Xia et al., 2010). These are reasons why for many years, power generating companies have been depending on load forecasts to make important decisions in operating, maintaining, planning, controlling as well as the management of electric power systems (Fan and Hyndman, 2012; Taylor and McSharry, 2008). Other than electricity utilities, load forecasting techniques are useful in other areas as well. These include business utilities such as industrial and commercial companies, financial institutions, regulatory commissions and many more (Feinberg and Genethliou, 2005; Hong and Fan, 2016).

The objectives of any electricity generating utility is to be able to meet the demand of the consumers. Electricity is a type of energy that cannot be stored, i.e. should only be produced when needed (Almeshaiei and Soltan, 2011). So it is important to obtain accurate forecasts in order to avoid implications that come after, for example over forecasting or under forecasting (Fan and Hyndman, 2012). These implications include maintenance costs and allocation of resources which depend on the capacity of electricity demand in a particular area (Almeshaiei and Soltan, 2011). The accuracy of the forecasts depends on the methods used as well as the factors present in the region of study. Accurate forecasts are essential for effective management

(Taylor, 2010).

Electric load forecasts are usually assessed periodically, ranging from hours, days, and weeks, months, up to a year, or even longer. These periods are often classified into the following categories respectively; short term load forecasts (STLF), medium term load forecasts (MTLF), and lastly long term load forecasts (LTLF) (Almeshaiei and Soltan, 2011). Short term forecasts are from few hours up to a week and medium term forecasts are from weeks up to a year, whereas long term forecasts are for more than a year (Fan and Hyndman, 2012). All these three categories serve different purposes in the management of electric power systems. However, amongst the three forecasting categories, STLF is the one mostly performed by many utilities (Hinman and Hickey, 2009) because it is useful when it comes to daily operation and scheduling of power systems (Kurban and Filik, 2008).

Many statistical regression models have been used widely in the past including the linear regression, time-series regression and many more. Some researchers have gone as far as modifying available methods, i.e. hybrid models and combination of forecasts. The main objective is always to come up with a model that produces the most accurate forecasts.

1.1 Background

This dissertation follows on the study by Fan and Hyndman (2012), in which the relationship between load demand and other driving variables was ex-

amined by fitting a semi-parametric additive model (SAM). Similarly, this dissertation aims to investigate the relationship between load demand and other explanatory variables such as temperature, calendar effects, etc. The quantile regression (QR) model will be used to forecast the hourly electricity demand for peak period (i.e from 18:00 to 20:00) in South Africa (SA). SA is amongst the developing countries around the world faced with an electricity crisis. On the African continent, SA is considered as the most industrialised country with the highest electricity consumption (Odhiambo, 2010).

The proposed quantile regression models will be combined with semi-parametric additive models (SAM), forming new hybrid models called the quantile semi-parametric additive models (QSAM). The proposed hybrid model mentioned above is a modification of the quantile generalized additive model (quantGAM) developed by team Tololo in a Global Energy Forecasting Competition in 2014 (GEFCom2014) (Gaillard et al., 2016a).

The management of every electricity utility makes decisions of unit commitment based on the forecasts obtained (Hahn et al., 2009). Thus forecasts obtained from this dissertation will be used to solve a unit commitment (UC) problem. Hong and Fan (2016) defines UC problem as an optimization problem that determines an optimal number of generating units to be scheduled in order to meet the load demand at a particular time. Mixed integer programming (MIP) will be used to solve the UC problem. MIP is defined as "the problem of allocating a number m of resources among $1, 2, ..., n$ activities in such a way as to maximize the worth from all the activities" (Momoh,

2001).

1.2 Purpose of the dissertation

1.2.1 Aim

The main aim of this study is to develop a modelling framework which combines short term forecasting with optimal scheduling of power plants so that they generate needed electricity at minimal costs.

1.2.2 Objectives

The main objectives of this study are to:

- develop hybrid short term load forecasting models which combine quantile regression with semi-parametric additive models,
- forecast the hourly electricity demand during peak period,
- evaluate the accuracy of the forecasts,
- include the forecasts obtained in developed optimization models in order to find an optimal number of units to commit at minimal costs for each period.

1.3 Scope of the dissertation

Quantile regression (QR) is used to predict hourly peak demand using SA's electricity data from Eskom. Eskom was South Africa's national public power

utility company. The hourly historical load data is collected from January 2010 till end of March 2014. The QR models will be combined with SAMs because of the nonlinear relationship that exist between electricity demand and temperature. Then, mixed integer programming(MIP) will be used on the forecasts obtained to determine an optimal number of units to commit at each peak period.

1.4 Structure of the dissertation

In Chapter 2, a review of existing literature on STLF techniques as well as the summary of studies that used methods similar to the proposed ones are presented. Chapter 3 provides general theory of the statistical methods used in this dissertation, i.e. QR as well as that of QSAM including the unit commitment problem. Chapter 4 provides results obtained from the proposed models as well as a detailed discussion and interpretation of the results. The conclusion is given in Chapter 5.

Chapter 2

Literature review

2.1 Introduction

Short term load forecasting (STLF) has been receiving a lot of attention lately (Fan and Hyndman, 2012). Almost every year different models for short term load forecasting are developed, applied, reviewed and published. This chapter provides an overview of STLF, factors affecting it as well as the summaries of some studies that used the proposed methodologies to forecast electricity demand in different countries. Additionally, a review of the literature on the unit commitment (UC) problem is given.

2.2 An overview of STLF

Short term forecasts are used to estimate the load demand up to a week ahead of a schedule. STLF methodologies are useful to most electricity suppliers as they tend to generate more accurate forecasts in a short period of time. This is helpful when it comes to daily operations of electricity utility sys-

tems (Hahn et al., 2009; Hinman and Hickey, 2009). The forecasts obtained are then used to schedule generation as well as the transmission of units to consumers. Accurate and reliable forecasts are necessary for any power supplying company as they help in preventing implications such as under loading (i.e. load shedding, blackouts, etc.) or overloading (i.e. producing more capacity than needed), which are very costly to suppliers (Feinberg and Genethliou, 2005; Pierrot and Goude, 2011). Moreover, accurate forecasts also help with the unit commitment decisions by reducing production costs (Amina et al., 2012).

2.3 Factors influencing STLF

The electricity demand pattern is very complex in nature due to the presence of several conditions or factors in a region of study. Generally, this includes factors such as the economic, environmental, social, weather conditions, calendar effects and many more (Almeshaiei and Soltan, 2011; du-Plessis, 2007; Fahad and Arbab, 2014; Fan and Hyndman, 2012). Various studies have been conducted to examine such factors and the influence they have on the short term load forecasts. However, most of these studies focus on the factors that have a huge influence on electricity demand such as meteorological effects (Taylor and McSharry, 2008). For instance, most researchers use weather variables when modelling the demand pattern (Xia et al., 2010; Taylor, 2010). Hinman and Hickey (2009) further expand on the idea that including weather variables in a forecasting model improves accuracy. Fahad

and Arbab (2014) also highlight that knowledge about weather conditions help in reducing operational costs.

Generally, weather variables include wind speed, solar radiation, temperature, rainfall, precipitation, humidity, cloud cover, etc. (Chikobvu and Sigauke, 2013; Fahad and Arbab, 2014; Pierrot and Goude, 2011). However, many researchers use temperature in their forecasting models as it is one of the major drivers of electricity demand. One of the studies that use the variable temperature is that of Chikobvu and Sigauke (2013). These authors assess the impact of the temperature on the daily peak demand in SA and their results show that electricity demand is sensitive to cold weather. Another study that examines meteorological effects is that of Taylor and McSharry (2008), in which the seasonality patterns were from 10 European countries were studied.

Other than meteorological effects, time factors also have a huge impact on electricity demand. Time factors include type of the day, yearly seasonality, time of the day, etc. and are generally well known as calendar effects (duPlessis, 2007). Their impact in short term load forecasting is discussed in detail by Fahad and Arbab (2014). For instance, Fahad and Arbab (2014) use load curves to explain electricity consumption patterns for different hours during the day and based on the load curves they conclude that calendar effects mostly determine the daily life style of the consumers, i.e. highest peak load in the evening. Hinman and Hickey (2009) further highlight that the peak load demands are the most studied.

2.4 Weather station selection in STLF

Weather effects are generally known to have a major impact on load demand thus making selection of weather stations an important aspect to consider in load demand forecasting. Janicki (2017) highlights the fact that the selection also depends on the geographical location, climate as well as the industrial structure of a country or region of study.

One of the few papers that focused more on the weather stations selection is that of Hong et al. (2015). In this paper the authors propose a new weather station selection framework which they applied on two of the electricity utilities in the United State (US). Their focus was more on developing an algorithm that will help in determining the number of weather stations to consider as well as which stations to use for a particular electricity utility.

2.5 Variable selection techniques for STLF

Another important aspect in model building is to be able to identify the best set of input variables. For instance, over the past years many different variable selection methods have been used to select important variables to include when forecasting, generally. There are many different available methods that can be used to select important variables in a study. These include the dimension reduction, shrinkage as well as the subset selection methods (James et al., 2015). The most commonly used methods are the subset selection which include techniques such as the stepwise criterion and

many more.

In a study by Fan and Hyndman (2012) a stepwise variable selection technique was used to select a combination set of variables that were used in models that are used to predict 48 half-hourly load demand in Australia. In this paper the authors used the stepwise backward selection method which involves including all the variables at first and removing one at a time while keeping the rest in the model. Then they continue by checking the performance of the model with removed variable using mean absolute percentage error (MAPE). Thus if the model resulted with a lower MAPE value, it is then selected as the best model for that period.

Although stepwise variable selection methods have been and are still widely used by many researchers in various research fields, they have limitations. Some of the few limitations include multicollinearity, computational efficiency, etc., however these issues are mostly ignored in many studies. For instance, according to Olusegun et al. (2015) the stepwise selection criterion select variables based on the correlation between a response and a set of explanatory variables only. Somehow the correlation within the explanatory variables is not considered which can lead to having one or two variables with the same characteristics in the model, i.e. multicollinearity. Alternatively in order to get rid of multicollinearity that could possibly exist between explanatory variables, shrinkage methods can be used to overcome that.

2.6 Modelling STLF

Different models for short term load forecasting have been developed and used to predict electricity demand in different countries with the aim of finding the model that will produce the best or accurate load estimates.

2.6.1 Semi-parametric additive model.

Generally, the semi-parametric model deals with the nonlinear functional relationships in regression analyses. Semi-parametric models are flexible as they are able to incorporate both the parametric and the nonparametric components in the same framework (Ruppert et al., 2003). Semi-parametric regressions are popular methods suitable for aggregated data, i.e. regional or national.

Fan and Hyndman (2012) developed a short term load forecasting model based on the semi-parametric additive technique. This is used to investigate the relationship between the load demand and the explanatory variables in Australia. Calendar effects, temperature effects, historical data as well as lagged load observations were included in the model. Their model was used to forecast the half-hourly load demand up to a week ahead of schedule. The half-hourly demand model developed by Fan and Hyndman (2012) is as follows:

$$\log(y_{t,p}) = h_p(t) + f_p(\mathbf{w}_{1,t}, \mathbf{w}_{2,t}) + a_p(\mathbf{y}_{t-,p}) + n_t \tag{2.6.1}$$

where, $y_{t,p}$ denotes the demand at time t (measured at half-hourly intervals) during period p ($p = 1, 2, ..., 48$), $h_p(t)$ models all calender effects, $f_p(\mathbf{w}_{1,t}, \mathbf{w}_{2,t})$ models all temperature effects where $\mathbf{w}_{1,t}$ and $\mathbf{w}_{2,t}$ are vectors of recent temperatures at two locations, $a_p(\mathbf{y}_{t-,p})$ models the effects of recent demands and n_t denotes the model error at time t.

Their empirical results show that the model used performs better on both the historical data as well as in real-time on-site implementation. These authors further suggested that for future improvement in load forecasting, more temperature sites and variables should be included in the model. The variables will depend on the area of study.

Pierrot and Goude (2011) propose the use of a more general approach of semi-parametric technique called general additive model (GAM). The GAM model was first developed by Hastie and Tibshirani (1990). Pierrot and Goude (2011) used their proposed model to forecast the French hourly load data for over 5 years. The focus was to model effects that drove the French load consumptions and also compare the proposed model with the operational one. The effects used include weather conditions, economic growth, weekly and yearly seasonality, which are represented by different variables in the model given in equation (2.6.2). The proposed GAM model is given as:

$$L_t = f_1(L_{t-24}) + f_2(L_{t-168}) + f_3(T_t) + f_4(mean(T_{t-24}, T_{t-48})) + f_5(cc) + f_6(posan) + C + \varepsilon_t \quad (2.6.2)$$

where L_t, L_{t-24} and L_{t-168} are respectively the load to be forecasted, the day

and week lagged load. T_t, T_{t-24} and T_{t-48} denote the current temperature, day and two day lagged temperatures, *cc* is the cloud cover, *posan* is the position of the day throughout the year, C is the intercept and ε_t represents the error term. The empirical results show that the proposed model performs better and is able to capture the effects that the French's operational one could not capture (Pierrot and Goude, 2011).

Another study which applies the semi-parametric model is that of Goude et al. (2014). The authors propose a semi-parametric approach based on the generalized additive model which is used to estimate electricity demand in 2260 substations across France, on both short and middle term horizons. The results indicate good performance in both cases. The authors also prove that this approach is capable of capturing big and complex datasets without human intervention. For instance, in their study, the load data used is collected over 2260 substations for every 10 minutes.

The proposed semi-parametric approach for short term forecasts used by the studies mentioned above proved to be suitable for short term load forecasting in all studies discussed.

2.6.2 Quantile regression (QR) model

Quantile regression (QR) models are used extensively in forecasting by different energy sectors, e.g. wind power forecasting, price forecasting, etc. A recent study by Gibbons and Faruqui (2014), introduced an optimal forecast quantile regression (OFQR) model which was used to forecast the annual

peak demand of electricity in 32 zones in the United State (US). The main aim was to measure the performance of OFQR over the ordinary least square (OLS) regression, which is a standard method used by almost every utility. They found that the OFQR model provides more accurate forecasts compared to OLS and suggested that other utilities should adopt this approach.

In another study by Taieb et al. (2016), a QR model is used to forecast the uncertainty in electricity from smart meter data. A smart meter is "an electronic device that records and transmits electricity consumption information typically at intervals from 1 minute to one hour, hence generating a huge quantity of data" (Taieb et al., 2016). The data used in their study is collected from 3639 households in Ireland at both aggregated and disaggregated levels. The results obtained using the proposed QR model is compared with the other 3 benchmark methods.

2.7 Unit commitment

This section provides a short summary on unit commitment (UC) as well as the methods mostly used by other researchers to solve the UC problem.

2.7.1 Introduction

The main objective of UC is to determine which generating units to commit (switching on or off) in order to meet electricity demand with minimal costs of production and reserve requirements (Wright, 2013). So it is very important to find a reasonable solution to this optimization problem as this will help

with obtaining an optimal number of generators that would be scheduled accordingly. Such optimal solution is useful as it helps prevent losses and minimize fuel consumption.

There are a variety of optimization techniques which have been used to find a feasible solution to the UC problem (Hong and Fan, 2016). Lagrangian relaxation (LR), mixed integer linear programming (MILP), tabu search (TS) and dynamic programming (DP) are some of the latest optimization techniques reviewed and discussed in detail by Saravanan et al. (2013). Amongst these techniques, the mixed integer programming (MIP) and LR techniques are mostly applied by many researchers (Kurban and Filik, 2008; Wright, 2013; Wu, Shahidehpour, and Li, 2007).

2.7.2 Applications

Kurban and Filik (2008) propose a method that combine short term load forecasting with UC in order to reduce production costs in one of the thermal plants based in Kutahya region, Turkey. These authors propose two models which are used to forecast electricity demand and then use the forecasts obtained to find solutions to a UC problem using the LR method. The empirical results indicate that accurate load forecasting is essential for UC.

In another study, Wu et al. (2007) consider the stochastic model for the long term solution on security-constrained unit commitment (SCUC) problem. SCUC is an "extension of the basic UC problem that includes additional factors such as fuel, emission and transmission constraints" (Wright, 2013).

These factors were not included in the early stochastic problems (SP) (Dai et al., 2015). Wu et al. (2007) use the proposed model to capture the uncertainties of generation components (i.e. generation units and transmission lines) as well as the inaccuracy of the load forecasts.

2.8 Concluding remarks

There is limited literature on the use of quantile regression models, especially on electricity demand forecasting. One of the contributions we aim to make in this dissertation is to add the use of the proposed methodology to the existing literature focusing more on SA, while at the same time solving the unit commitment (UC) problem.

Chapter 3

Research Methodology

3.1 Introduction

This chapter describes the general theory of the proposed methods, i.e. the quantile regression (QR) models. The theory of QR together with semi-parametric additive models (SAMs) is covered including that of optimization techniques such as mixed integer programming (MIP) which is used to solve the unit commitment problem in this dissertation. In addition to this, the theory of the methods that are used to select the explanatory variables in this dissertation are also discussed.

3.2 Semi-parametric additive model

Semi-parametric regression model is used to develop nonlinear functional relationships of a response variable in regression analysis (Ruppert et al., 2003). Their advantage over ordinary linear regression (OLS) is the ability to han-

dle for nonlinearity in different data sets, i.e. they are very flexible models. They are developed by adding a nonlinear function to the linear regression and often are referred as generalized additive models (Keele, 2008). The general additive model (GAM) was first introduced by Hastie and Tibshirani (1990).

Generally, the OLS models are represented in this form:

$$y_i = \alpha + \beta_1(x_1) + ... + \beta_k(x_k) + \varepsilon \tag{3.2.1}$$

And the additive effects are in this form:

$$y_i = \alpha + f_1(x_1) + ... + \beta_k(x_k) + \varepsilon \tag{3.2.2}$$

Now adding the additive effects to Equation 3.2.1 we get:

$$y_i = \alpha + f_1(x_1) + ... + \beta_k(x_k) + \beta_1(x_{k+1}) + ... + \beta_j(x_j) + \varepsilon \tag{3.2.3}$$

Equation 3.2.3 represent general formula for semi-parametric additive models. Where, α a constant, y_i is the response variable, f_j are smooth functions assumed to have a nonlinear relationship with the response variable, x_k represent covariates and ε is the error term.

3.3 Quantile regression

Quantile regression (QR) is "gradually emerging as a unified statistical methodology for estimating models of conditional quantile functions" (Koenker, 2005). The QR model was first introduced by Koenker and Bassett (1978) as

an extension of the ordinary least square regression (OLSR), which is mostly used to establish a relationship between a response (dependent) variable and one or more independent variables. In general, OLSR determines the conditional mean of a response variable, whereas the QR model determines the conditional function of a response variable at a certain proportion instead of a conditional mean.

Advantages of QR model:

- Does not have limitations when it comes to conditional mean (Davino et al., 2014).
- Provides location, scale, and shape on the conditional distribution of the entire response variable (Davino et al., 2014).
- Approximates the whole distribution of a response variable conditional on the quantiles (proportions) (Davino et al., 2014).
- The model is robust to outliers, i.e. estimates are not sensitive to outliers (Koenker, 2005).

3.3.1 Quantile regression modelling framework

Let Y be a random variable (i.e. hourly electricity demand) and $\mathbf{X}$ be a set of explanatory variables (i.e. temperature, time, calendar effects, etc.). If $F_{Y|\mathbf{X}}$ denotes the conditional distribution function (CDF) of Y given $\mathbf{X}$, then

the conditional quantile q_α of order $\alpha \in [0,1]$ of Y knowing $\mathbf{X}$ is defined as the generalized inverse of $F_{Y|\mathbf{X}}$:

$$Q(\alpha) = q_\alpha(Y|\mathbf{X}) = F^{-1}_{Y|\mathbf{X}}(\alpha) = inf\{y \in \mathbb{R} : F_{Y|\mathbf{X}}(y) \geq \alpha\} \tag{3.3.1}$$

3.3.2 Parameter estimation

Now, the idea of quantile estimation arises from the observation that the median (i.e., $q_{0.5}(Y|\mathbf{X})$) minimizes the expected absolute error. More generally, it can be shown that the conditional quantile $q_\alpha(Y|\mathbf{X})$ is the solution of the following minimization problem:

$$q_\alpha(Y|\mathbf{X}) \in \arg\min_g \mathbb{E}[\rho_\alpha(Y - g(\mathbf{X}))|\mathbf{X}] \tag{3.3.2}$$

where, ρ_α is the pinball loss defined for all $u \in \mathbb{R}$ by $\rho_\alpha(u) = u(\alpha - \mathbf{1}_{u<0})$ (Gaillard et al., 2016a). The QR models minimization problem is solved by applying the linear programming technique through simplex method (Davino et al., 2014).

3.3.3 Model selection

The best model is selected based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The AIC was first developed by Akaike (1973). AIC score aims to find approximate model whereas the BIC score tries to find the true model. Both AIC and BIC reference scores are calculated using equations (3.3.3) and (3.3.4) respectively:

AIC:

$$\text{AIC} = 2k - 2\log(\ell(\theta)) \tag{3.3.3}$$

and

BIC:

$$\text{BIC} = k\log(n) - 2\log(\ell(\theta)) \tag{3.3.4}$$

where, k is the model degrees of freedom and n is the number of observations.

3.3.4 Error measure

The pinball loss function

The quantile loss function also known as the pinball loss function. It is given by:

$$L(\hat{Q}_{Y|X,Z}(\tau)) = \begin{cases} (\tau(y - \hat{Q}_{Y|X,Z}(\tau)) & \text{If} \quad y \geq \hat{Q}_{Y|X,Z}(\tau) \\ (1-\tau)(\hat{Q}_{Y|X,Z}(\tau) - y) & \text{If} \quad y < \hat{Q}_{Y|X,Z}(\tau) \end{cases} \tag{3.3.5}$$

where $\hat{Q}_{Y|X,Z}(\tau)$ is the quantile forecast and y is the observation used for forecast evaluation.

3.4 Quantile semi-parametric additive model

We now develop a hybrid model, the quantile semi-parametric additive model (QSAM) for hourly load demand forecasting for South Africa (SA). The model is developed by combining equation (3.2.3) for each quantile defined by equation (3.3.2). The peak demand period in this study will be from

18:00 to 20:00. The nonlinear relationship between load demand and the temperature will be modelled through regression splines. The developed model for this study will be as follows:

$$\log(y_{tp}(\alpha)) = a_0 + at + c_p(t) + g_p(t) + f_p(y_{(t-k)p}) + \varepsilon_t \tag{3.4.1}$$

where, $y_{tp}(\alpha)$ is the hourly demand at time t $(t = 1, 2, ..., n)$ for period p $(p = 1, 2, ..., 24)$, α is the quantile percentage or $\alpha \in [0, 1]$, $c_p(t)$ models all the calendar effects, $g_p(t)$ models the temperature effects, $f_p(y_{(t-k)p})$ models the lagged demand and ε_t denotes the error (residual) term at time t. If the error terms are autocorrelated, the general structure is as follows:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D\varepsilon_t = \theta_q(B)\Theta_Q(B^s)e_t \tag{3.4.2}$$

where, ϕ_p, Φ_P, θ_q and Θ_Q are polynomial functions of the orders p, P, q and Q respectively, B is the lag operator, ∇^d and ∇_s^D represent non-seasonal and seasonal difference operators respectively, s is the seasonal period and e_t is white noise (Box and Jenkins, 1976).

Calendar effects:

$$c_p(t) = \sum_{i=2}^{7}\beta_i(\alpha)DayType_i + \sum_{i=2}^{12}\gamma_i(\alpha)Month_i + b(\alpha)H_t + b_1(\alpha)H_{t-1} + b_2(\alpha)H_{t+1} \tag{3.4.3}$$

where, $DayType_i$ represents the day of the week, $Month_i$ represents the month of the year, H_t denotes the holiday effect, H_{t-1} and H_{t+1} respectively denote the day before and after the holiday.

Temperature effects:

$$g_p(t) = m_p(\alpha)T_t^- + n_p(\alpha)T_t^+ + q_p(\alpha)\bar{T}_t + r_p(\alpha)T_t^p \tag{3.4.4}$$

where, T_t^-, T_t^+ and $\bar{T}_t$ respectively denote the minimum, maximum and average temperature at time t and T_t^p denote the temperature at hour p (p is from 18:00 to 20:00).

Generally the above developed model can be represented as follows:

$$y_{tp}(\alpha) = \beta_{0p,\alpha} + \sum_{j=1}^{h_1} s_{jp,\alpha}(x_{tpj}) + \sum_{j=1}^{h_2} \beta_{jp,\alpha} z_{tpj} + \varepsilon_{tp,\alpha}, \quad \alpha \in (0,1) \tag{3.4.5}$$

where y_{tp} represent electricity demand on day t at hour p ($p = 18, 19$ and 20), with x_{tpj} as corresponding covariates (variables), $\beta_{jp,\alpha}$ are parameter estimates, $s_{jp,\alpha}$ are smoothing functions, $z_{tpj,\alpha}$ are linear variables and $\varepsilon_{tp,\alpha}$ is the quantile error term.

3.4.1 Variable selection

In this study the shrinkage selection methods for variable selection will be used. The least absolute shrinkage and selection operator (Lasso) as well as the ordinary ridge regression are some of the known shrinkage selection methods. However this study focuses on the Lasso including the Lasso via hierarchial interactions criterion as it is considered to have advantages over the ridge regression (Bien et al. (2013); among others). The Lasso technique first introduced by Tibshirani (1996) penalises the absolute size of the regression coefficients. It is also a useful variable selection method especially

when dealing with highly correlated predictors. The model is then given by Bien et al. (2013) as:

$$y_{t,p} = \beta_0 + \sum_j \beta_j x_{tj} + \sum_{j \neq k} \Theta_{jk} x_{tj} x_{tk} + \varepsilon_t. \tag{3.4.6}$$

In which pairwise interactions are also allowed between predictors. In this study, quadratic interactions are however not allowed, i.e. $x_{tj}x_{tk}$ where $j = k$.

For ordinary least square regression coefficients are calculated by minimising the sum of squares, however in the case of the Lagrangean form of the strong hierarchical Lasso they are calculated as (see Bien et al. (2013) for details):

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \delta \in \mathbb{R}^{p \times p}, \delta = \delta^T} w(\beta_0, \beta, \delta) + \lambda \sum_j \max \left\{ |\beta_j|, \|\delta_j\|_1 \right\} + \frac{\lambda}{2} \|\alpha\|_1 \tag{3.4.7}$$

where $w(\beta_0, \beta, \delta)$ is the loss function (quadratic loss function).

3.4.2 Parameter estimation

Coefficients for the developed model will be estimated by an optimization function which minimizes the sum of absolute residuals, i.e. the conditional quantile is a solution to an optimization problem. Generally, optimization techniques such as linear programming (LP) are used (Koenker and Hallock, 2001).

Equation (3.4.5) can also be written as follows:

$$\mathbf{Y} = s_\alpha(\mathbf{X}) + \mathbf{Z}^\mathbf{T} \beta_\alpha + \varepsilon_\alpha, \alpha \in (0, 1) \tag{3.4.8}$$

The parameter estimates of equation (3.4.8) are obtained by minimizing the following function:

$$Q(\beta, s(.)) = \sum_{i=1}^{n} \rho_{\alpha,i}(Y_i - Y_i Z_i^T \beta - s(X)) + \sum_{i=1}^{n} \int \theta_i (s_{\prime\prime}(t))^2 dt, \quad (3.4.9)$$

where $\rho_\alpha(u) = \alpha u - uI(u < 0)$ is the quantile loss function.

3.4.3 Model diagnostics

Before any further analysis the model will be assessed to check if the assumptions of the residuals will not be violated. For instance, we will check if the residuals are normally distributed. Other than normality we will also check if the residuals will be independent and that there is no serial autocorrelation left in them. Residuals are defined as the difference between the actual and the estimated values of a response variable (i.e. the hourly load demand) and statistically are represented by the equation (3.4.10):

$$\hat{e_i} = y_i - \hat{y_i} \quad (3.4.10)$$

where, y_i are the actual values of a response variable and $\hat{y_i}$ are estimated values of a response variable.

3.4.4 Model evaluations

The performance of the developed models are assessed by using the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE). Generally, the three measures mentioned above

are statistically represented by the equations (3.4.11), (3.4.12) and (3.4.13) respectively:

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|\hat{e}_i|, \tag{3.4.11}$$

$$\text{MAPE} = \frac{100}{m}\sum_{i=1}^{m}|\frac{\hat{e}_i}{y_i}|, \tag{3.4.12}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\hat{e}_i{}^2} \tag{3.4.13}$$

where, m is the number of observations in the test data set, y_i is the actual values of a response variable and $\hat{e}_i$ is the residual of the i^{th} observation.

3.4.5 Forecast combinations

Combining forecasts improves the forecast accuracy (Devaine et al. (2012); Gaillard et al. (2016b); among others). Let K denote the number of methods used to forecast the response variable, the combined forecasts will be calculated as follows:

$$f_{tp} = \sum_{k=l}^{K}\omega_{ktp}y_{ktp} \tag{3.4.14}$$

where ω_{ktp} is the weight. The pinball loss function in equation 3.3.5 is used as an evaluation criterion. In this study, the r package opera developed by Gaillard et al. (2016b) is used for combining the forecasts.

3.5 Unit commitment

3.5.1 Introduction

As stated by Wright (2013), the main objective of unit commitment (UC) is to minimize total cost of generating units over a particular period. There are variety of methods available that can be used to find an optimal solution to a UC problem. In this study, the bounded variable mixed integer linear programming (MILP) method will be used. According to Wright (2013) the bounded variable MILP is a special class of linear programming where some of the variables are restricted to integer values but the rest are ordinary continuous variables. This study intends to demonstrate how the forecasts obtained can be used in solving the UC problem.

3.5.2 Problem formulation

Generating units in this study are classified into 13 coal-fired, 1 nuclear, 4 gas, 2 pumped storage and 2 hydroelectric. These will be denoted as: $g_{1c}, g_{2c}, ..., g_{13c}, g_{14n}; g_{15g}, ..., g_{18g}, g_{19p}, g_{20p}, g_{21h}, g_{22h}$ respectively. Also note that we are going to use the total fuel cost to represent average production cost per megawatt produced. Generally, we minimize the following total fuel cost equation:

$$F\left(P_{Gi}^{tp}, x_i^{tp}\right) = \sum_{p=1}^{H} \sum_{i=1}^{m} \left[F_i\left(P_{Gi}^{tp}\right) x_i^{tp} + F_{si}(tp) x_i^{tp}\right] \tag{3.5.1}$$

where, P_{Gi}^{tp} represents load from generating unit $i,\ i = 1, ..., m$ at hour $p,\ p = 18:00, 19:00, 20:00$ on day $t,\ t = 1, ..., n$; P_G^{tp} national system load at hour p on day t; x_i^{tp} the 0-1 variable (in this study we will assume that during the peak period all units are up, i.e $x_i^{tp} = 1$ for all units); F_{si} the start up cost of unit i at hour p (in this study we will assume that the start up cost is zero); F_i the average production cost of unit i (cost/MW).

The total fuel cost is subjected to the following constraints;

Load balance:

$$\sum_{i=1}^{m} P_{Gi}^{tp} x_i^{tp} = P_D^{tp}, p = 18, 19, 20,\ t = 1, ..., n \tag{3.5.2}$$

where, P_D^{tp} is the system demand at hour p on day t,

Generator power output limits:

$$x_i^{tp} P_{Gi\min} \leq P_{Gi}^{tp} \leq x_i^{tp} P_{Gi\max}, p = 18, 19, 20,\ t = 1, ..., n,\ i = 1, ..., m \tag{3.5.3}$$

where, $P_{Gi(\min)}^{tp}$ is the lower limit of the unit power output and $P_{Gi(\max)}^{tp}$ is the upper limit of the unit power output,

Power reserve:

$$\sum_{i=1}^{m} P_{Gi\max} x_i^{tp} \geq P_D^{tp} + P_R^{tp}, p = 18, 19, 20,\ t = 1, ..., n \tag{3.5.4}$$

P_R^{tp} the power reserve at hour p on day t,

Minimum up/down time:

$$\left(U_{tp-1,i}^{\text{up}} - H_i^{\text{up}}\right)\left(x_i^{tp-1} - x_i^{tp}\right) \geq 0, p = 18, 19, 20,\ t = 1, ..., n,\ i = 1, ..., m \tag{3.5.5}$$

$$\left(U_{tp-1,i}^{\text{down}} - H_i^{\text{down}}\right)\left(x_i^{tp} - x_i^{tp-1}\right) \geq 0, p = 18, 19, 20,\ t = 1, ..., n,\ i = 1, ..., m \quad (3.5.6)$$

Chapter 4

Data analysis and discussion

4.1 Data sources

The 24 hourly historical load data used in this dissertation are obtained from Eskom's net energy sent out (NESO) and are measured in megawatts (MWs) at aggregate level (i.e. this include all South African (SA) economy sectors). The data consist of five years hourly electricity demand from the beginning of January 2010 until the end of March 2014. We also have the 24 hourly historical temperature data obtained from South African Weather Service (SAWS) climate database (measured in degrees Celsius (^{0}C)). The data were divided into two portions, whereby the first four years of the data are used to develop the hourly forecasting models and the remaining one are used to validate the accuracy of the forecasts obtained from the developed models.

4.2 Data explanatory analysis

In this dissertation electricity demand is considered as the response or dependent variable. Figure 4.1 illustrates 24 hourly electricity demand (consumption) profiles for particular days in South Africa (SA), June 31^{st} 2010 and December 29^{th} which are classified as winter and summer, respectively in the plot. The profiles also indicate that there are two peaks of electricity consumption which occur on a normal day in SA, the first one in the morning, i.e. between 8:00 and 12:00, and the second one in the evening, between 17:00 and 20:00. However in this dissertation we only focus on the higher peak period which is the second one during the day.

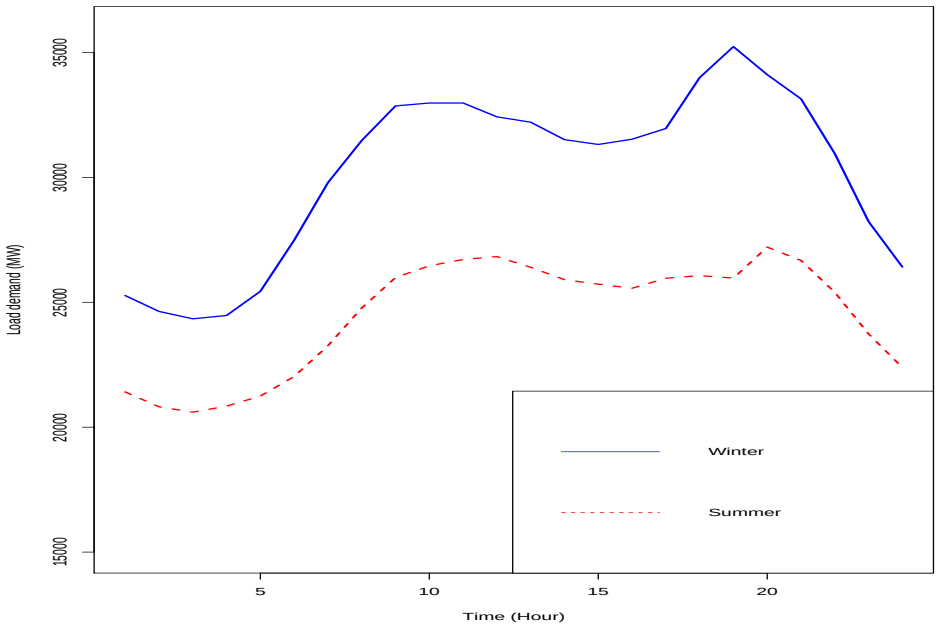


Figure 4.1: Daily electricity demand plot.

The South African daily electricity demand time series and density plots for the peak period considered in this dissertation are as shown in Figure 4.2. The daily demand shown here from the beginning of January 2010 till the end of March 2014. The panes on the left of Figure 4.2 show the time series plots while those on the right show the density plots of the demand for 18:00, 19:00 and 20:00 hours respectively. Note the seasonality patterns of electricity demand in SA, i.e. the plots indicate higher electricity demand in winter and lower demand in summer yearly.

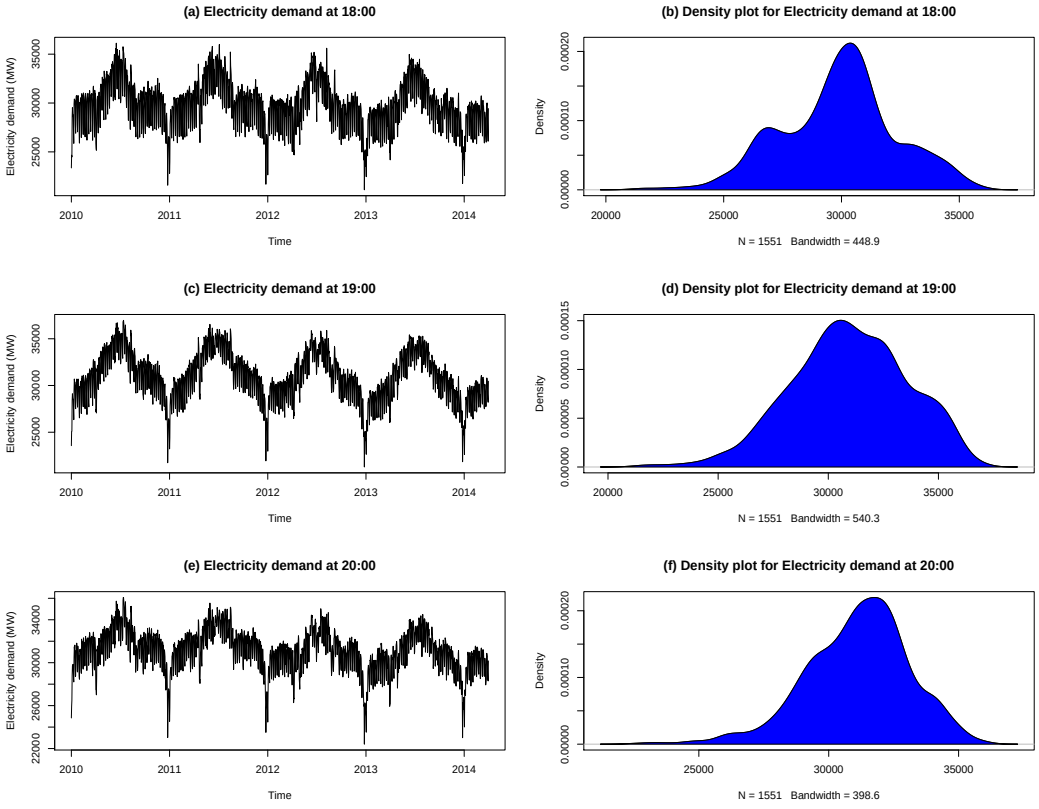


Figure 4.2: Daily electricity demand and density plots for peak period.

Looking at the plot in Figure 4.2(a) it is clear that the electricity demand series for the 18:00 hours is stationary, i.e. constant over time. The augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests were also applied to assess wether the series is stationary or not. An ADF test statistic value of -3.7538 with a p-value of 0.0213 and a KPSS Trend test statistic value of 0.20196 with a p-value of 0.01526 were obtained. Using 5% significance level on both tests it was concluded that the series at 18:00 hours is stationary thus confirming our visual inspection. Similar tests were also done for the data at 19:00 and 20:00 hours respectively and both did not confirm the data as stationary. However including the trend in the models will take care of the non-stationarity found in the series.

Table 4.1: Descriptive statistics of logged electricity demand during the peak period.

Hour	Mean	Std	Median	Minimum	Maximum	Skew	Kurtosis
Demand 18 pm	10.30	0.08	10.31	9.96	10.49	-0.41	0.37
Demand 19 pm	10.33	0.09	10.34	9.97	10.51	-0.51	0.40
Demand 20 pm	10.34	0.07	10.35	10.02	10.48	-0.81	1.50

Table 4.1 shows the summary statistics of the logged electricity demand for the peak period considered in this dissertation, i.e. from 18:00 to 20:00 hours. For all the hours the means are smaller than the medians indicating that the logged electricity demand distribution is slightly skewed to the left (i.e. the data is negatively skewed). This is consistent with the skewness values for these hours as shown in the table as well. The distributions of the logged electricity demand are further highlighted by the box and whisker

plots in Figure 4.3. The kurtosis values are all less than 3 meaning that the distributions are platykurtic, i.e. peaked distributions.

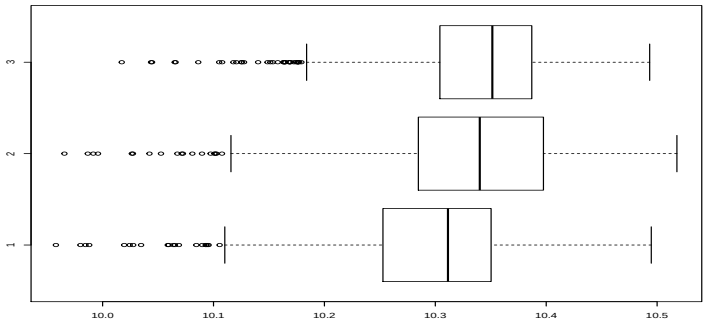


Figure 4.3: Box and whisker plots for electricity demand, where (1) 18:00 hours, (2) 19:00 hours and (3) 20:00 hours.

Figure 4.4 illustrates a relationship between electricity demand and the temperature at 18:00 hours and by visual inspection it is clear that there exists a nonlinear relationship between the two variables.

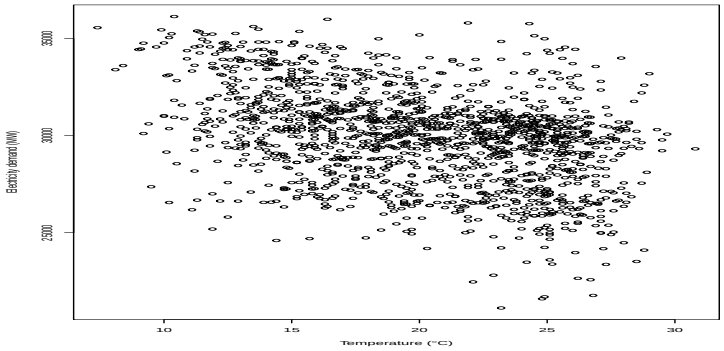


Figure 4.4: A plot for electricity demand versus the temperature (18:00 hours).

The explanatory or independent variables used in this dissertation in-

clude temperature, calender effects, holiday effects as well as lagged electricity demand. Although all the variables mentioned above are important in forecasting, temperature mostly has a major effect on short term load forecasting (Janicki, 2017). For instance, duPlessis (2007) highlighted that in South Africa (SA) weather conditions such as temperature influence about 80% of electricity demand. The variables used in this dissertation are defined as follows:

- $Load_p$, a response variable which represents electricity demand for peak hours in SA (i.e. 18:00, 19:00, 20:00) for a particular day,
- $Month_i$ accounts for the monthly seasonality patterns throughout the year, where $i = \{2, 3, ..., 12\}$ represent the yearly months from February till December with January as the base month.
- $Trend$ accounts for the linear trend in the series,
- $DayType_j$ accounts for the day of the week, where $j = \{2, 3, ..., 7\}$ represent the days from Tuesday till Sunday with Monday as base weekday,
- $Holiday$ represents public holiday in South Africa (SA) where,

$$Holiday = \begin{cases} 1, & \text{If it's a holiday} \\ 0, & \text{Otherwise} \end{cases}$$

- DBH is the day before the holiday,

$$DBH = \begin{cases} 1, & \text{If it's a day before the holiday} \\ 0, & \text{Otherwise} \end{cases}$$

- DAH is the day after the holiday,

$$DAH = \begin{cases} 1, & \text{If it's a day after the holiday} \\ 0, & \text{Otherwise} \end{cases}$$

- T, TC and TI simultaneously represent the overall, coastal as well as the interior temperatures,

- $MinT$, $MinTC$ and $MinTI$ represent the minimum temperatures of the overall, coastal as well as the interior for the past 24 hours,

- $MaxT$, $MaxTC$ and $MaxTI$ represent the maximum temperatures of the overall, coastal and interior for the past 24 hours,

- $AvgT$, $AvgTC$ and $AvgTI$ represent the average temperatures of the overall, coastal and interior for the past 24 hours,

- $MinL$ and $MaxL$ respectively represent daily minimum and maximum electricity demand,

- $AvgL$ represent the average electricity demand for the past 7 days,

- $Lag1$ and $Lag2$ respectively represent the lagged electricity demand for the same hour of the first and second previous days. Both the lags are differenced in order to make the series of the electricity demand stationary. According to Fan and Hyndman (2012) the inclusion of lagged demand effects reduce autocorrelation, however it does not completely remove it,

- ε_t is the error term and it follows a seasonal autoregressive moving average (SARIMA) model given below with e_t as a white noise process at time t, i.e. $e_t \sim \mathrm{N}(0, \sigma_e^2)$, hence the SARIMA model becomes

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D\varepsilon_t = \theta_q(B)\Theta_Q(B^s)e_t.$$

We then specify our general median quantile semi-parametric additive models for peak period as follows:

$$\begin{aligned}\log(Load(H18)_\alpha) = {} & c_\alpha + \beta_{0,\alpha}Trend + \sum_{i=2}^{7}\beta_{i,\alpha}DayType_i + \sum_{i=2}^{12}\gamma_{i,\alpha}Month_i \\ & + \delta_\alpha Holiday + \delta_{1,\alpha}DBH + \delta_{2,\alpha}DAH + \kappa_\alpha T(H18) \\ & + \lambda_\alpha MinT + \mu_\alpha MaxT + \omega_\alpha AvgT + Lag1(H18)_\alpha \\ & + Lag2(H18)_\alpha + \varepsilon_{t,\alpha},\end{aligned} \tag{4.2.1}$$

$$\begin{aligned}\log(Load(H19)_\alpha) = {} & c_\alpha + \beta_{0,\alpha}Trend + \sum_{i=2}^{7}\beta_{i,\alpha}DayType_i + \sum_{i=2}^{12}\gamma_{i,\alpha}Month_i \\ & + \delta_\alpha Holiday + \delta_{1,\alpha}DBH + \delta_{2,\alpha}DAH + \kappa_\alpha T(H19) \\ & + \lambda_\alpha MinT + \mu_\alpha MaxT + \omega_\alpha AvgT + Lag1(H19)_\alpha \\ & + Lag2(H19)_\alpha + \varepsilon_{t,\alpha}\end{aligned} \tag{4.2.2}$$

and

$$\log(Load(H20)_\alpha) = c_\alpha + \beta_{0,\alpha}Trend + \sum_{i=2}^{7}\beta_{i,\alpha}DayType_i + \sum_{i=2}^{12}\gamma_{i,\alpha}Month_i + \delta_\alpha Holiday + \delta_{1,\alpha}DBH + \delta_{2,\alpha}DAH + \kappa_\alpha T(H20) + \lambda_\alpha MinT + \mu_\alpha MaxT + \omega_\alpha AvgT + Lag1(H20)_\alpha + Lag2(H20)_\alpha + \varepsilon_{t,\alpha}. \tag{4.2.3}$$

where, c, β, β_2, β_3, β_4, β_5, β_6, β_7, γ_2, γ_3, γ_4, γ_5, γ_6, γ_7, γ_8, γ_9, γ_{10}, γ_{11}, γ_{12},δ, δ_1, δ_2, κ, λ, μ and ω are parameters to be estimated. The parameter α is the quantile percentage, i.e. 50% representing the median in this study.

Since the residuals, ε_t of the hourly load models are usually autocorrelated it is important that the autocorrelation is significantly reduced before the models are used for forecasting. In order to do that, we adopt a method by Marasinghe (2014) to reduce autocorrelation in the residuals. A summary of the procedure for modelling autocorrelation that possibly exists in the residuals of the models in this dissertation is as follows:

- Estimate the parameters in equations 4.2.1, 4.2.2 and 4.2.3,
- extract residuals, ε_t, from each model and test if autocorrelation exist in them, then fit a SARIMA(p,d,q)(P,D,Q)$_s$ if they are autocorrelated,
- subtract the fitted values of the residuals of the SARIMA(p,d,q)(P,D,Q)$_s$ from $Load_p$ to get $Load_p^*$,
- regress $Load_p^*$ on the covariates or dependent variables and

- check for residual autocorrelation in the new model. If the residuals are still autocorrelated repeat the process until the desired results are achieved.

4.3 Weather station selection

The temperature data initially consist of 28 weather stations across South Africa and out of the 28 stations only 8 of them were then considered for further analysis, i.e. stations with no data and those with large number of missing values were left out. On the selected stations, missing temperature values were then filled by imputing the data with possible temperature data values. The imputation was done in R using functions available inside the package called mice (van Buuren and Groothuis-Oudshoorn, 2011). In addition to this, the stations were also selected based on the method developed by Kruger and Sekele (2013) in which these authors classified the weather stations into clusters (groups). Similarly, weather stations used in this dissertation were then classified into the following clusters; coastal as well as the interior cluster. They were selected in such a way that they represent the different geographical parts of South Africa (SA). For each cluster, the data were aggregated to get the maximum, minimum and average daily temperature. This also mean that coastal and interior temperature data will be used as separate variables in the models hence the term non-aggregated will be used to represent these separate cluster variables. The overall aggregated temperature data were also considered for comparison. In other words the

analysis was done on the overall aggregated as well on the non-aggregated temperature data set.

4.4 Quantile semi-parametric additive models solution

4.4.1 Introduction

This section presents the procedure followed when building models which are then used to forecast electricity demand for the peak period (i.e. from hour 18:00 to 20:00) in SA. The developed quantile semi-parametric additive model (QSAM) which incorporates temperature, all the calendar, holiday as well as the demand effects for each period is used. Additionally the interactions effects between the independent variables are also considered.

4.4.2 Variable and model selection

Initially, all the calendar, lagged demand, weather and holiday effects were included and the least absolute shrinkage and selection operator (lasso) via hierarchical interactions technique was then applied to select significant variables. The Lasso was applied on the data with overall aggregated temperature scenarios as well as on the non-aggregated temperature scenarios, i.e. coastal and interior temperature scenarios as separate variables. Thus having two models for each period.

When using the Lasso technique to select variables a tuning parameter is required. The tuning parameter is used to control the amount of regu-

larization and is selected using cross-validation technique. Cross-validation evaluates by separating data into training and testing set.

Table 4.2: Variables selected using Lasso via hierarchical interactions with aggregated temperature at 18:00 hours

	Main effect	Month	DayType	Holiday	DBH	DAH	T(H18)	MaxT	MinL	MaxL	AvgL	Lag2(H18)
Month	0.004	-0.0174	0	0	0	0	0	0	0	0	0	0
Trend	0	0	0	0	0	0	0	0	0	0	0	0
DayType	-6e-04	0	-6e-04	0	0	0	0	0	0	0	0	0
Holiday	-3e-04	0	0	0	-0.0027	0	0	0	0	0	0	0
DBH	0	0	0	-0.0027	0	0	0	0	0	0	0	0
DAH	1e-04	0	0	0	0	1e-04	0	0	0	0	0	0
T(H18)	-0.0016	0	0	0	0	0	-2e-04	0	4e-04	0	0	7e-04
MaxT	0	0	0	0	0	0	0	0	0	0	0	0
MinL	0.0706	0	0	0	0	0	4e-04	0	-0.0023	0	0	-0.002
MaxL	-2e-04	0	0	0	0	0	0	0	0	-2e-04	0	0
AvgL	0	0	0	0	0	0	0	0	0	0	0	0
Lag2(H18)	0.0031	0	0	0	0	0	7e-04	0	-0.002	0	0	0

Tables 4.2 and 4.3 respectively show variables that were selected to predict electricity demand at 18:00 hours using data from both temperature scenarios. Variables that resulted with a value including the interactions as well were consider when building the models. Other tables showing variables selected for the remaining period, i.e. 19:00 and 20:00 hours are given in Appendix B. For each period two models were developed using the selected variables and the best one was then selected based on the bayesian information criterion (BIC) value. At the end only one model for each period was selected and then used to forecast daily electricity demand for each peak period respectively in South Africa.

For the first period, i.e. 18:00 hours, a BIC of 1.97 was obtained on the model with the overall aggregated temperature scenario whereas on the non-aggregated a BIC of 1.90 was obtained. Based on the obtained BIC values from both developed models, it is clear that the second model performs better

Table 4.3: Variables selected using Lasso via hierarchical interactions with non-aggregated temperature at 18:00 hours

	Main effect	DayType	Holiday	DBH	MinL	MaxL	Lag2(H18)
Month	-6e-04	0	0	0	0	0	0
Trend	0.0017	0	0	0	0	0	0
DayType	-0.0072	-0.0072	0	0	0	0	0
Holiday	-0.0016	0	0	-0.0016	1e-04	0	0
DBH	-3e-04	0	-0.0016	0	0	0	0
TI	0	0	0	0	0	0	0
MaxTI	-0.0022	0	0	0	0	0	0
MinL	0.0064	0	1e-04	0	-6e-04	0	0
MaxL	0.0658	0	0	0	0	0	-0.0021
Lag1(H18)	0.0037	0	0	0	0	0	0
Lag2(H18)	0.005	0	0	0	0	-0.0021	-5e-04

than the first one. Thus implying that the second model was then chosen as the perfect model to predict daily electricity demand for the 18:00 hours in South Africa. That is the model with non-aggregated temperature scenario.

For the second period, i.e. 19:00 hours, a BIC of 1.62 was obtained from both models, i.e. the model with aggregated temperature scenario and the one with non-aggregated temperature scenarios. The model with non-aggregated temperature scenario was then selected as the best model to predict electricity demand for 19:00 hours. While for the last period, i.e. 20:00 hours, a BIC of 1.29 for the model with overall aggregated temperature and a BIC of 1.28 for the model without one were then obtained. Similarly it is clear that the second model performs better and thus selected and used to forecast electricity demand for the last period.

In overall the models that included both the coastal and interior temperature as separate variables were then selected as the best models to predict

electricity demand for peak period. Also note that the models selected included interactions of independent variables and will be presented throughout the study as follows; M1 (QSAM with interactions at 18:00 hours), M2 (QSAM with interactions at 19:00 hours) and M3 (QSAM with interactions at 20:00 hours). The residuals of each of these selected models are further analysed in detail in the next section.

4.4.3 Residual analysis

The residuals obtained from the three developed models were then analysed to check if any of the model assumptions on the residuals were met, i.e. the residuals should not be serial autocorrelated and the Ljung-Box test was used to check for that. In case there still exists autocorrelation in the residuals after including all lagged demands as suggested by Fan and Hyndman (2012), another method by Marasinghe (2014) was also applied to reduce autocorrelation in the residuals of the models. This method is also discussed in details by Hyndman and Athanasopoulos (2013).

For Model M1, i.e. at 18:00 hours, the Ljung-Box test statistic of 951.26 with a p-value$<$0.0001 was obtained. Using a 5% significance level, this implies that there exist autocorrelation in the residuals. Despite the fact that a suggestion by Fan and Hyndman (2012) was employed when developing this model, it seems that the residuals still remain highly autocorrelated. In order to account for the autocorrelation we further used the method by Marasinghe (2014) by transforming the response variable by first extracting

the residuals and fitting a SARIMA$(2,1,0)(1,0,0)_7$. Then subtract the fitted values of the SARIMA from the response variable, thus making a new transformed response variable. The transformed variable was then used to estimate electricity demand for 18:00 hours. Furthermore the residuals of the model with the transformed demand variable were analysed to check the assumptions again. The test statistic of 37.089 with a p-value of 0.1746 were obtained indicating that there is no autocorrelation left in the residuals. Figure 4.5 further show the autocorrelation diagnostic plots of the final model for 18:00 hours.

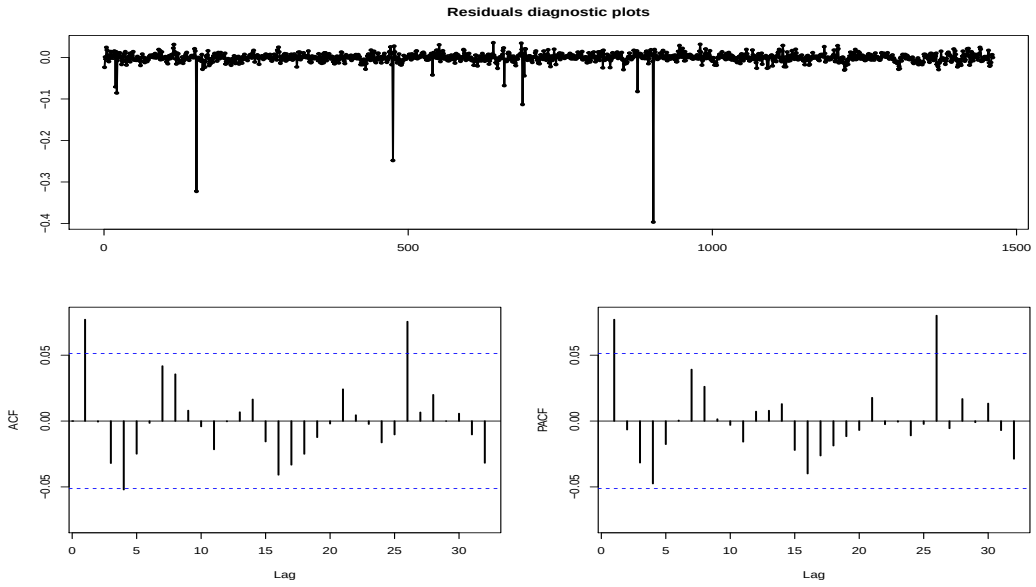


Figure 4.5: Residuals diagnostic plots for the 18:00 hour, where a) Time series plot of the residuals (top panel) b) ACF (bottom left panel) and c) PACF (bottom right panel).

For Model M2, i.e. at 19:00 hours, the Ljung-Box test statistic of 547.16

with a p-value<0.0001 was obtained. Using a 5% significance level this indicates that there exists autocorrelation in the residuals. The residuals were also extracted and a SARIMA$(2,2,0)(1,0,0)_7$ was fitted on them. Then the fitted values from this SARIMA model were subtracted from the response variable making a new transformed response variable. Using the transformed response variable on the final model and testing for autocorrelation again, this resulted with the Ljung-Box test statistic of 0.028941 with a p-value of 1. At 5% level of significance this indicates that autocorrelation no longer exists in the residuals of the model, thus the model was then used to forecast demand at 19:00 hours.

For Model M3, i.e. at 20:00 hours, the Ljung-Box test statistic of 649.89 with a p-value<0.0001 was obtained. Similarly, using a 5% significance of level this shows that there exists autocorrelation in the residuals. Therefore, the response variable was also transformed in this case as well by extracting the residuals and fitting a SARIMA$(2,1,0)(1,0,0)_7$ on them. Then after subtracted the fitted values of the SARIMA model from the response variable making a new transformed response variable. The final model for this period was then developed using the transformed variable and residuals of the new model were also analysed. With the new transformed response variable, the Ljung-Box test statistic of 17.139 with a p-value of 0.9709 were obtained. A p-value greater than 5% significance level indicates that there is no more autocorrelation left in the residuals.

4.4.4 Final models

The final developed Model M1, i.e. at 18:00 hours, show that on weekdays, the coefficient for variable $DayType_7$ which represents Sundays is negative, indicating that electricity demand on those days at this period is lower than the demand on the day which is acting as the "reference" day. Additionally on monthly basis, the coefficients are negative on $Month_2$ (February), $Month_3$ (March), $Month_9$ (September), $Month_{10}$ (October), $Month_{11}$ (November) as well as on $Month_{12}$ (December), indicating that electricity demand on these months is lower than the ones acting as "reference" months as well. Tables showing summary of coefficients of the variables included in the final model for this period are shown in Appendix B, Table 7.

4.4.5 Forecasting results (Developed models)

The training period used to evaluate the forecasts obtained from the developed model for all the periods considered in this dissertation, is from the beginning of January 2014 till the end of March 2014. To evaluate the performance of the developed models, the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) were used on the out-of-sample data for each period.

The forecasting results obtained for the first peak period, i.e. 18:00 hours, are shown in Figure 4.6 and summarized as follows; a MAE of 270.78, an RMSE of 339.37 as well as a MAPE of 0.956%. Whereas for the second and last periods, i.e. 19:00 and 20:00 hours, the forecasting results obtained are

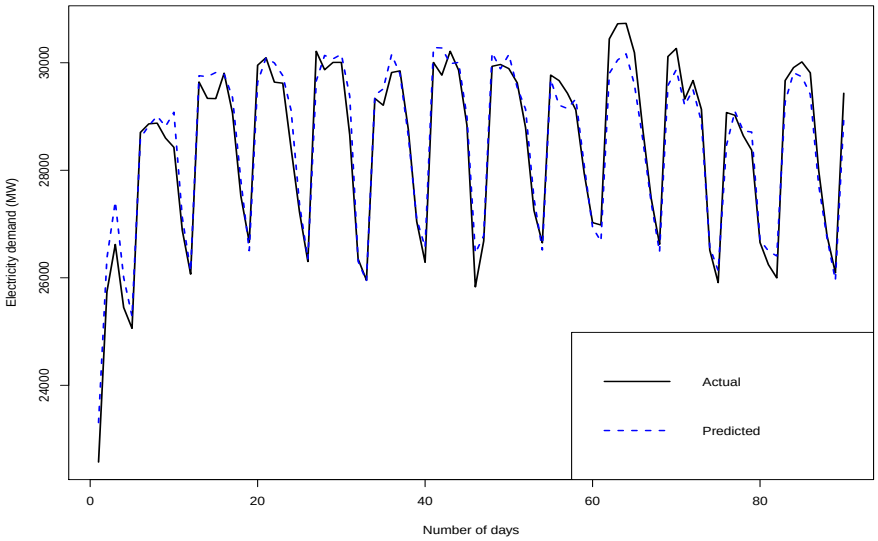


Figure 4.6: Actual and predicted electricity demand (MW) at 18:00 hours.

shown respectively by the Figures in the Appendix B. The summaries are as follows; for the second period, a MAE value of 324.42, an RMSE of 401.29 and MAPE of 1.116% whereas for the last period, a MAE value of 105.68, an RMSE of 179.48 and a MAPE 0.362% were respectively obtained from the out-of-sample data.

4.5 Benchmark models solution

4.5.1 Introduction

The previous section discussed the development of the quantile semi-parametric additive regression models for the peak period which included variables selected as well as significant interactions of variables. This section discusses

benchmark models which are then used to compare with the developed models. The models used as benchmarking models are similar to the developed presented in equations 3.4.4, 3.4.5 and 3.4.8 respectively, however in this case interactions between variables were not considered. For instance, to build benchmark models significant variables were selected using the lasso variable selection criterion for each period. Similarly variable selection was done on data with aggregated temperature scenario and on the one with non-aggregated temperature scenario (i.e. coastal and interior temperature as separate variables), thus having a total of 6 benchmark models (2 models per period).

Table 4.4 show coefficients of the variables selected by applying the lasso technique on the first period, i.e. 18:00 hours, whereby the first column represent variables which included aggregated temperature whereas the second column representing the one with non-aggregated temperature. For the remaining periods, i.e. 19:00 and 20:00 hours, the tables are shown in Appendix B. From Table 4.4 all the variables that resulted with some values on them were considered for further analysis, i.e. two models were then developed for each period. For the first one, 18:00 hours, a BIC of 1.870 were respectively obtained for both models. For the second period, i.e. 19:00 hours, a BIC of 2.210 and 2.172 were respectively obtained while for the last period considered in this study, i.e. 20:00 hours, a BIC of 1.336 was obtained from both models.

From all the 6 developed models, the best models were selected for each

Table 4.4: Variables selected by using the Lasso technique on the 18:00 hours.

Variable	With aggregated temperature	With non-aggregated temperature
(Intercept)	9.129580e+00	9.127886e+00
$Month$	-2.523436e-04	-2.504831e-04
$Trend$	9.644581e-06	9.845081e-06
$DayType$	-2.131795e-03	-2.102574e-03
$Holiday$	-2.673234e-03	-2.843848e-03
DBH	.	.
DAH	1.584870e-03	2.051343e-03
$T(H18)$	.	
TC		.
TI		.
$MinT$	.	
$MinTC$		.
$MinTI$		.
$MaxT$	-5.848554e-04	
$MaxTC$		.
$MaxTI$		-5.593705e-04
$AvgT$	.	
$AvgTC$		.
$AvgTI$		.
$MinL$	1.316121e-05	1.341944e-05
$MaxL$	3.128758e-05	3.134595e-05
$AvgL$	-3.471015e-06	-3.707278e-06
$Lag1$	2.580870e-06	2.697386e-06
$Lag2$	2.105118e-06	2.045764e-06

period based on the BIC values obtained from each. The second model which has a smaller BIC hence was selected as the benchmark model for 19:00 hours electricity demand while for 18:00 and 20:00 hours, second models were selected and used as the benchmark models for those hours. Thus ending up with 3 models representing each period and are presented as follows; BM1 (QSAM without interactions at 18:00 hours), BM2 (QSAM without

interactions at 19:00 hours) and BM3 (QSAM without interactions at 20:00 hours).

4.5.2 Forecasting results (Developed benchmarking models)

From the benchmark models, a MAE of 269.44, an RMSE of 345.15 and a MAPE of 0.956% were obtained for Model BM1 representing the first period, a MAE of 400.33, an RMSE of 516.79 and a MAPE of 1.400% were obtained for Model BM2 representing the second one whereas for Model BM3 representing the last period in this study a MAE of 88.34, an RMSE of 199.08 and a MAPE of 0.304% were obtained. Additionally, a summary showing estimates of the coefficients of variables included in the benchmark model for the first period is shown in Appendix B, Table 8.

4.6 Comparison analysis of models

In order to select the best peak period models for short term load forecasting between the developed (quantile semi-parametric additive models with interactions) and benchmarks (quantile semi-parametric additive models without interactions) models; the mean absolute error (MAE), root mean square error (RMSE) and mean absolute error (MAPE) were used. Moreover, forecasts obtained from both developed and benchmarks models were combined to improve the forecasting accuracy.

Table 4.5 summarizes the accuracy measures of the errors obtained for

Table 4.5: Evaluation of models using out-of-sample data at 18:00 hours

Accuracy measures	Model M1	Model BM1	Comb1
MAE	270.78	269.44	262.12
RMSE	339.37	345.15	330.76
MAPE(%)	0.956	0.956	0.922

the out-of-sample data from both models for the first period, i.e. for the developed model and the model referred to as benchmark. The accuracy measures of the combined forecasts, i.e. Comb1 column in 4.5, are also shown. In this case the accuracy measures of the combined forecasts have improved as compared to the ones obtained from both the models. This means that the combined forecasts are further used to solve the unit commitment problem for 18:00 hours.

Table 4.6: Evaluation of models using out-of-sample data at 19:00 hours

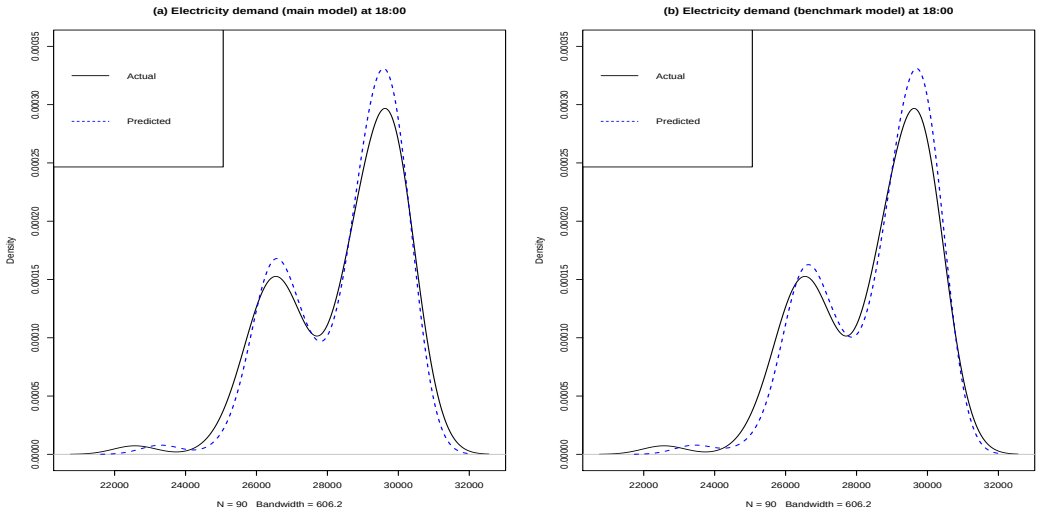
Accuracy measure	Model M2	Model BM2	Comb2
MAE	324.42	400.33	325.71
RMSE	401.29	516.79	401.34
MAPE(%)	1.116	1.400	1.127

Table 4.6 summarizes the accuracy measures of the errors obtained for the out-of-sample data from both models for the second period including that of the combined forecasts, i.e. as shown in column Comb2 in Table 4.6. However, in this case the developed model seems to be performing better than the benchmark and combined ones indicating that the interaction of variables has an improvement impact on this period.

Table 4.7 summarizes the accuracy measures of the errors obtained for

Table 4.7: Evaluation of models using out-of-sample data at 20:00 hours

Accuracy measures	Model M3	Model BM3	Comb3
MAE	105.68	88.34	89.42
RMSE	179.48	199.08	198.63
MAPE(%)	0.362	0.304	0.308



(a) (b)

Figure 4.7: Density plots at 18:00 hours

the out-of-sample data from both models for the last period considered in this study as well that of the combined forecasts, i.e. represented by column Comb3 in Table 4.7. In this case the benchmark model is performing better indicating that interaction of variables does not have an improvement effect on this period.

Figure 4.7 represent density plots for both the main model and the benchmark, respectively for the first period, i.e. 18:00 hours. Based on these density plots the daily minimum and maximum electricity at 18:00 hours are

22000 MW and 32000 MW respectively. This indicates or suggests that it is not likely to get electricity demand below 22000 MW or more than 32000 MW for this period. The density plots for the remaining peak period are shown in Appendix B.

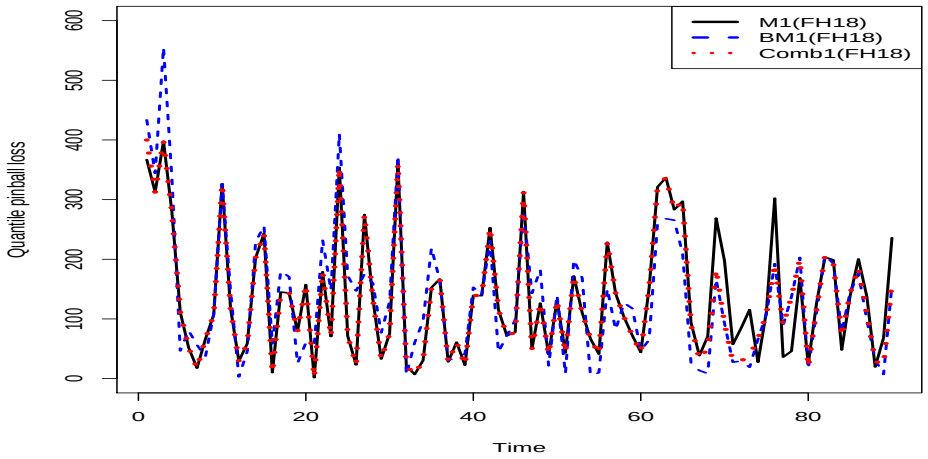


Figure 4.8: Quantile pinball loss on electricity demand (MW) at 18:00 hours.

The quantile pinball loss plots for 18:00 hours are shown in Figure 4.8. The quantile pinball loss values are represented as M1(FH18), BM1(FH18) and Comb1(FH18) in the plot and respectively are obtained from Model M1, Model BM1 and combined forecasts. The quantile pinball loss plots for 19:00 and 20:00 hours are shown in Appendix B.

In addition to quantile pinball loss plots, the average quantile loss values from all the obtained forecasts in this study are also shown in Table 4.8. All in all combined forecasts (Comb1) from the first period, forecasts from the developed model (Model M2) for the second period and the one from the

Table 4.8: Average quantile loss

18:00 hours		
Model M1	Model BM1	Comb1
135.3919	134.7196	131.0585
19:00 hours		
Model M2	Model BM2	Comb2
162.2114	200.1623	162.8558
20:00 hours		
Model M3	Model BM3	Comb3
52.83778	44.16944	44.70833

benchmark model (Model BM3) for the last period resulted with a smaller average quantile loss values. Forecasts from these are further used in solving the unit commitment problem, details are provided in the next section.

4.7 Unit commitment solution

The previous sections in this chapter focused more on building models and using them to forecast electricity demand for the peak periods in SA. This section focuses on using those obtained forecasts to find the solution to the unit commitment problem using the bounded variable mixed integer linear programming (MILP) model.

4.7.1 Solution to the unit commitment using MILP

Table 4.9 and 4.10 represent the average production cost, minimum and maximum electricity demand from 22 various generating units (stations) in

Table 4.9: The Minimum Average Production Cost (base load demand stations)

Unit	**Min Ave prod cost c_i(rands)**	**Min (MW)**	**Max (MW)**
g_{1c}	244.4	0	4116
g_{2c}	247.2	0	4110
g_{3c}	245.7	0	3990
g_{4c}	270.6	0	3708
g_{5c}	270.6	0	3654
g_{6c}	245.7	0	3600
g_{7c}	245.7	0	3600
g_{8c}	247.2	0	3000
g_{9c}	244.4	0	2400
g_{10c}	247.2	0	2000
g_{11c}	245.7	0	1600
g_{12c}	247.2	0	1200
g_{13c}	244.4	0	1000
g_{14n}	231.8	0	1931

SA, i.e. 13 coal-fired, 1 nuclear, 4 gas, 2 pumped storage and 2 hydroelectric. In this study fuel costs are represented by the average production costs. This study also assumes that the minimum electricity demand per generating unit equals to zero .

Table 4.11 shows a sample of the forecasts obtained from the selected model for each period. The highlighted forecasts representing Tuesday, 25 March 2014 on the table are then used in solving a unit commitment problem.

We treat the problem as a bounded variable mixed integer linear programming model

$$\text{Min} \quad Z \quad = \sum_{i=1}^{13} c_i g_{ic} + c_{14} g_{14n} + \sum_{i=15}^{18} c_i g_{ig} + \sum_{i=19}^{20} c_i g_{ip} + \sum_{i=21}^{22} c_i g_{ih}$$

Table 4.10: The Minimum Average Production Cost (peaking stations)

Unit	**Min Ave prod cost c_i(rands)**	**Min (MW)**	**Max (MW)**
g_{15g}	596.4	0	1338
g_{16g}	490.6	0	740
g_{17g}	490.6	0	171
g_{18g}	460.8	0	171
g_{19p}	231.2	0	1000
g_{20p}	231.2	0	400
g_{21h}	231.2	0	344
g_{22h}	231.2	0	240

Table 4.11: Out of sample forecasts

Date	**Comb1(FH18)**	**M2 (FH19)**	**BM3(FH20)**
Sun 23-Mar-2014	26411	27590	28541
Mon 24-Mar-2014	29281	29860	30545
Tue 25-Mar-2014	**29768**	**30235**	**30909**
Wed 26-Mar-2014	29739	29964	30740
Thu 27-Mar-2014	29442	29777	30596
Fri 28-Mar-2014	29731	28167	29025
Sat 29-Mar-2014	29442	27604	28475

where, Z represents the total fuel cost to be minimized.

Subject to

Load balance constraints:

$$\sum_{i=1}^{13} g_{ic} + g_{14n} + \sum_{i=15}^{18} g_{ig} + \sum_{i=19}^{20} g_{ip} + \sum_{i=21}^{22} g_{ih} = P_D^{ht}, \quad h = 18, 19, 20; \quad t = 25/03/2014$$

$$\text{with} \quad P_D^{18t} = 29768\text{MW}, \quad P_D^{19t} = 30235\text{MW}, \quad P_D^{20t} = 30909\text{MW}$$

Generally, about 10% to 15% of reserve margin is assigned by regulatory agencies in most utilities (Gross, 2015). Therefore, using a 15% reserve margin in this dissertation we get

$$P_R^{18t} = 0.15 \times 29814 = 4465.2\text{MW},$$

$$P_R^{19t} = 0.15 \times 30235 = 4535.25\text{MW},$$

$$P_R^{20t} = 0.15 \times 30909 = 4636.35\text{MW}$$

Power reserve constraints:

$$1.15 \left(\sum_{i=1}^{13} g_{ic} + g_{14n} + \sum_{i=15}^{18} g_{ig} + \sum_{i=19}^{20} g_{ip} + \sum_{i=21}^{22} g_{ih} \right) \leq 44003$$

$$0 \leq g_{1c} \leq 4116$$

.

.

.

$$0 \leq g_{13c} \leq 1000$$

$$0 \leq g_{14n} \leq 1931$$

$$0 \leq g_{15g} \leq 1338$$

.

.

.

$$0 \leq g_{18g} \leq 171$$

$$0 \leq g_{19p} \leq 1000$$

.

.

.

$$0 \leq g_{22h} \leq 240 \cdot$$

Table 4.12 shows the solution to the unit commitment problem for all hours considered as peak period in this study. For 18:00 hours, an optimal solution of R7,255,292.00 was found. The solution suggest that Eskom needs to produce a total of 33414 MW daily from fourteen of the twenty-two generating units considered in this study in order to incur the total fuel cost of

Table 4.12: Optimal Solution using bounded variable MILP

Variable	18:00 (R7,255,292.00)		19:00 (R7,370,734.00)		20:00 (R7,537,347.00)	
	Soln	RC	Soln	RC	Soln	RC
g_{1c}	4116	0	4116	0	4116	0
g_{2c}	4110	0	4110	0	4110	0
g_{3c}	3990	0	3990	0	3990	0
g_{4c}	0	23.4	0	23.4	0	23.4
g_{5c}	0	23.4	0	23.4	0	23.4
g_{6c}	3600	0	3600	0	3600	0
g_{7c}	3600	0	3600	0	3600	0
g_{8c}	1437	0	1904	0	2578	0
g_{9c}	2400	0	2400	0	2400	0
g_{10c}	0	0	0	0	0	0
g_{11c}	1600	0	1600	0	1600	0
g_{12c}	0	0	0	0	0	0
g_{13c}	1000	0	1000	0	1000	0
g_{14n}	1931	0	1931	0	1931	0
g_{15g}	0	349.2	0	349.2	0	349.2
g_{16g}	0	243.4	0	243.4	0	243.4
g_{17g}	0	243.4	0	243.4	0	243.4
g_{18g}	0	213.6	0	213.6	0	213.6
g_{19p}	1000	0	1000	0	1000	0
g_{20p}	400	0	400	0	400	0
g_{21h}	344	0	344	0	344	0
g_{22h}	240	0	240	0	240	0

R7,255,292.00 for this hour. Whereas for the second and the last one, i.e. 19:00 and 20:00 hours, optimal solutions of R7,372,381.00 and R7,538,994.00 were obtained respectively with a total of 33835 MW and 34509 MW daily, respectively. The results also suggest that the eight generating units that resulted with a solution of zero during the peak period can reduce the fuel cost by R1096.40. Additionally, the fourteen generating units are the optimal number of generating units (stations) required to produce electricity demand at minimal costs and this includes nine coal-fired, one nuclear, two pumped storage and two hydroelectric stations.

4.8 Concluding remarks

This chapter proposed short term load forecasting models for the peak periods, i.e. 18:00, 19:00 and 20:00 hours. In addition to this benchmark models were also presented and the results obtained were compared with those of the proposed models. After comparison, forecasts obtained from three of the best selected models (representing each hour) were then used to find an optimal number of generating units to minimize production costs of electricity demand.

Chapter 5

Conclusions

5.1 Introduction

In this dissertation, two applications were illustrated. Firstly, quantile semi-parametric additive models which were used to forecast electricity demand during the peak periods in South Africa were presented. Hours considered as peak periods were 18:00, 19:00 and 20:00 respectively. Secondly, this study also demonstrated how the forecasts obtained from the developed models can be used in finding a solution to a unit commitment problem. A solution to a unit commitment will guide decision makers at utilities with addressing challenges they are faced with daily, i.e. from management of resource planning to secure scheduling of units to commit (Hahn et al., 2009).

5.2 Summary

Chapter 1 discussed the main purpose of this research followed by Chapter 2 summarising the relevant literature on short term load forecasting methods

and the unit commitment problem. Chapter 3 then discussed in detail the methodology implemented in the dissertation whereas Chapter 4 presented a detailed analysis of the results, i.e. the modelling of daily electricity demand during the peak periods in South Africa with a combination of finding an optimal scheduling of power plants in generating units in order to generate needed electricity at minimal costs.

Quantile semi-parametric additive models were developed for each hour considered during the peak period. Fifteen variables were considered initially including their interactions as well. Then significant variables and interactions were selected using the least absolute shrinkage and selection operator (Lasso) method. Finally, for each hour during the peak period a model was fitted, residuals extracted and then checked if autocorrelation was present. In all the models developed, residuals were found to be serially correlated thus transformation on the response (dependent) variable was considered, i.e. the forecasts were obtained from a model with a transformed response variable. In addition to developing models for hours considered during the peak period, benchmark models were also considered for comparison. The benchmark models were built in the same way as the main models, however in this case interaction of variables were not taken into consideration.

This resulted with forecasts obtained for the hours considered as peak period in this dissertation, three from main and another three from the benchmark models. Then the best models were selected using the following evaluation methods; mean absolute error (MAE), root mean square error

(RMSE) and mean absolute percentage error (MAPE). Additionally combined forecasts were also considered for each hour and a MAE, an RMSE and a MAPE were obtained from them. The empirical results showed that the combination of forecasts method for 18:00 hours and developed model for 19:00 hours were the best, whereas for the 20:00 hours the benchmark one was best fitting, i.e. the model without interactions of variables. All the analysis up to this point were done using various add-on packages in R programming language.

Lastly, forecasts obtained from the best three selected models were further used in finding an optimal solution to the unit commitment problem and this was done using Lingo Version 14. A bound mixed integer linear programming method was used to find an optimal number of units (stations) to commit for each hour during the peak period. This information will be beneficial to the decision makers at most utilities such as Eskom, South Africa's national power utility especially with allocation of energy, scheduling of units to commit during the peak periods.

5.3 Limitations

One of the challenges faced in this dissertation was dealing with autocorrelated residuals, there is limited research on handling the autocorrelation of the residuals especially when using the quantile regression models. Additionally, a lot of assumptions were made about the data when solving the unit commitment problem this was due to difficulty in gaining access to necessary

data.

5.4 Future research

- In this dissertation daily aggregated electricity demand data was used. In future the daily electricity demand for interior as well as coastal regions must be included as well.
- Investigation on short term electricity demand forecasting on weekly or monthly basis should be considered.

References

H. Akaike. Information Theory as an Extension of the Maximum Likelihood Principle. In *In Second International Symposium on Information Theory*, pages 267 – 281, 1973.

E. Almeshaiei and H. Soltan. A methodology for electric power load forecasting. *Alexandria Engineering Journal*, 50:137 – 144, 2011.

M. Amina, V. S. Kodogiannis, I. Petrounias, and D. Tomtsis. A hybrid intelligent approach for the prediction of electricity consumption. *Electrical Power and Energy Systems*, 43:99 – 108, 2012.

J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierachical interactions. *The Annals of Statistics*, 41(3):1111 – 1141, 2013.

G. E. P. Box and G. M. Jenkins. *Time series analysis: forecasting and control*. Holden-day, 1976.

D. Chikobvu and C. Sigauke. Modelling influence of temperature on daily peak electricity demand in South Africa. *Journal of Energy in Southern Africa*, 24(4):63 – 70, 2013.

R. Cottet and M. Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98:839 – 849, 2003.

H. Dai, N. Zhang, and W. Su. A literature review of stochastic programming and unit commitment. *Journal of Power and Energy Engineering*, 3:206 – 214, 2015.

C. Davino, M. Furno, and D. Vistocco. *Quantile Regression: Theory and Applications*. John Wiley and Sons, 1st edition, 2014.

M. Devaine, P. Gaillard, Y. Goude, and G. Stoltz. Forecasting the electricity consumption by aggregating specialized experts; a review of sequential aggregation of specialized experts, with an application to slovakian and french country-wide one-day-ahead (half-) hourly predictions. *Machine Learning*, 90:231 – 260, 2012.

L. duPlessis. System optimisation and the impact of the short term load forecast. `www.eepublishers.co.za/article/system-optimisation-and-the-impact-of-the-short-term-load-forecast.html`, 2007.

M.U. Fahad and N. Arbab. Factor affecting short term load forecasting. *Journal of Clean Energy Technologies*, 2(4):305 – 309, 2014.

S. Fan and R. J. Hyndman. Short term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1): 134 – 141, 2012.

E. A. Feinberg and D. Genethliou. *In Applied Mathematics for Restructured Electric Power Systems: Optimization, Control and computational Intelligence*, chapter Load Forecasting, pages 269 – 285. New York: Springer, 2005.

P. Gaillard, Y. Goude, and R. Nedellec. Semi-parametric models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. http://doi:10.1016/j.forecast.2015.12.001, 2016a.

P. Gaillard, Y. Goude, and R. Nedellec. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32:1038 – 1050, 2016b.

I. Ghalehkhondabi, E. Ardjmand, G. R. Weckman, and W. A. Young II. An overview of energy demand forecasting methods published in 20052015. http://DOI:10.1007/s12667-016-0203-y, 2016.

C. E. Gibbons and A. Faruqui. Quantile Regression for Peak Demand Forecasting. Available at SSRN 2485657, 2014.

Y. Goude, R. Nedellec, and N. Kong. Local Short and Middle Term Electricity Load Forecasting With Semi-Parametric Additive Models. *IEEE Transactions on Smart Grid*, 5(1):440 – 446, 2014.

C. Gross. Power generation technology data for integrated resource plan of south africa. Technical update, ELECTRIC POWER RESEARCH INSTITUTE, 2015.

H. Hahn, S. Meyer-Nieberg, and S. Pickl. Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research*, 99: 902 – 907, 2009.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. London: Chapman and Hall, 1990.

J. Hinman and E. Hickey. Modelling and forecasting short-term electricity load using regression analysis. *Hall*, 2009.

T. Hong and S. Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 2016.

T. Hong, P. Wang, and L. White. Weather station selection for electric load forecasting. *International Journal of Forecasting*, 31:286 – 295, 2015.

R. J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice. `http://www.otexts.org/fpp`, 2013.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2015.

M. Janicki. Methods of weather variables introduction into short-term electric load forecasting models - a review. `http://DOI:10.15199/48.2017.04.18`, 2017.

L. Keele. *Semiparametric Regression for the Social Sciences*. John Wiley and Sons, 2008.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. W. Bassett. Regression quantiles. *Econometrica*, 46:33 – 50, 1978.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143 – 156, 2001.

A. C. Kruger and S. S. Sekele. Trends in extreme temperature indices in south africa: 1962 - 2009. *International journal of climatology*, 33:661 – 676, 2013.

M. Kurban and U. B. Filik. Unit commitment schedulling by using the autoregressive and artificial neural network models based short-term load forecasting. In *Probalistic methods applied to power systems*, pages 1 – 5. IEEE, 2008.

D. Marasinghe. Quantile regression for climate data. *All Theses*. paper 1909, 2014.

J. A. Momoh. *Electrical Power System Applications of Optimization*. Power Engineering, 2001.

N. M. Odhiambo. Energy consumption,prices and economic growth in three SSA countries: A comparative study. *Energy Policy*, 38:2463 – 2469, 2010.

A.M. Olusegun, H.G. Dikko, and S.U. Gulumbe. Identifying the limitation of stepwise selection for variable selection in regression analysis. *American Journal of Theoretical and Applied Statistics*, 4(5):414 – 419, 2015.

A. Pierrot and Y. Goude. Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, pages 593 – 600, 2011.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

B. Saravanan, S. Das, S. Sikri, and D. P. Kothari. A solution to the unit commitment problem – a review. *Front Energy*, 7(2):223 – 236, 2013.

L. Suganthi and A. A. Samuel. Energy models for demand forecasting a review. *Renewable and Sustainable Energy Reviews*, 16:1223 – 1240, 2012.

S. B. Taieb, R. Huser, R. J. Hyndman, and M. G. Genton. Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, 2016.

J. W. Taylor. Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204:139 – 152, 2010.

J. W. Taylor and P. E. McSharry. Short term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, 22:2213 – 2219, 2008.

R. Tibshirani. Regression shrinkage ad selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267 – 288, 1996.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1 – 67, 2011.

B. Wright. A review of unit commitment. `www.ee.columbia.edu/~lavaei/Projects/Brittany_Wright.pdf`, 2013.

L. Wu, M. Shahidehpour, and T. Li. Sochastic Security-Constrained Unit Commitment. *IEEE Transactions on Power Systems*, 22(2):800 – 811, 2007.

C. Xia, J. Wan, and K. McMenemy. Short, medium and long term load forecasting model and virtual load forecaster based on radial basis function neural networks. *Electrical Power and Energy Systems*, 32:743 – 750, 2010.

Appendices

Appendix A

```
#The following packages in R were used

mice, forecast, ggplot2, plaqr, quantreg, SparseM, tseries, splines, stats,
    glmnet,

#### Reading data into R
LoadSA <- read.csv()

#### Hourly load profile for 31/06/2010 and 29/12/2011
D1 <- LoadSA[546,1:24]
D2 <- LoadSA[1094,1:24]
ts.plot(t(d1), col="blue", lty=1, lwd=2, ylim=c(20000,36000), xlab="Time (
    Hours)",ylab="Load demand (MW)")
lines(t(d2), col="red", lty=2, lwd=2)
legend(x="bottomright", legend=c("Winter","Summer"), col=c("blue","red"),
    lty=1:2,lwd=1)

#### Daily load profiles & density plots the first peak period
Series <- ts(LoadSA[,18],frequency=365, start=2010)

par(mfrow=c(2,2))
plot(Series,ylab="Electricity demand (MW)",main="(a) Electricity demand at
    18:00")
plot(density(Series),main="(b) Density plot for Electricity demand at 18:00"
    )
polygon(density(ts1),col="blue",border="black")
par(mfrow=c(1,1))

#### Test for stationary in load demand
adf.test(ts1[Series],alternative="stationary")
kpss.test(ts1[Series],null="Trend")

#### Summary statistics for peak period
f <- log(LoadSA[,18:20])
describe(f)
boxplot(log(LoadSA$L18),log(LoadSA$L19),log(LoadSA$L20),horizontal = TRUE)

#### Creating lag variables for the first peak period
Lag1 <- diff(LoadSA$L18[365:1916],lag=1)
Lag2 <- diff(LoadSA$L18[364:1916],lag=2)

k <- 1551
n <- length(LoadSA$L18[1:1461])
h <- nrow(LoadSA) - n

##############################################################################
#                                                                            #
#                              MAIN MODEL                                    #
#                                                                            #
##############################################################################
#### Variable Selection using LASSO via interactions:
attach(LoadSA)

  **** WITH AGGREGATED
```

```
w1 <- cbind(Month,Trend,DayType,Holiday,DBH,DAH,T18,MinT,MaxT,AvgT,MinL,MaxL
    ,AvgL,Lag1,Lag2)
z <- log(L18)
f1 <- hierNet.path(w1,z)
cv1=hierNet.cv(f1,w1,z)
fit1=hierNet(w1,z,lam=cv1$lamhat.1se)
print(fit1)

M1 <- plaqr(formula= ts(log(L18))~ Month+DayType+Holiday+DAH+Lag2+MinL*Lag2,
     nonlinVars=~T18+MinL+MaxL+Lag2+T18*MinL+T18*Lag2,tau=0.5,data=LoadSA[1:
    n,])

  **** WITHOUT AGGREGATED
w2 <- cbind(Month,Trend,DayType,Holiday,DBH,DAH,TC18,TI18,MinTC,MinTI,MaxTC,
    MaxTI,AvgTC,AvgTI,Lag1,Lag2)
f2 <- hierNet.path(w2,z)
cv2=hierNet.cv(f2,w2,z)
fit2=hierNet(w2,z,lam=cv2$lamhat.1se)
print(fit2)

M2 <- plaqr(formula= ts(log(L18))~ Month+Trend+DayType+Holiday+DBH+Lag1+Lag2
    +Holiday*DBH+Holiday*MinL+MaxL*Lag2, nonlinVars=~MaxTI+MinL+MaxL+Lag1+
    Lag2,tau=0.5,data=LoadSA[1:n,])

#### Model selection
bic(M1)
bic(M2)

#### Residual analysis from the best model selected
Residuals <- residuals(best model selected)
Box.test(Residuals, lag=30, type="Ljung")

  **** If residuals are still autocorrelated then proceed
R1 <- Arima(Residuals,order=c(2,1,0),seasonal=list(order=c(1,0,0),period=7),
    lambda=0)

na.zero <- function(x){
  x[is.na(x)] = 0
  return(x)}

fit.Arima <- na.zero(residuals(R1))
N18 <- log(L18[1:n]) - fit.Arima

QSAM <- plaqr(formula= ts(log(N18))~ Month+Trend+DayType+Holiday+DBH+Lag1+
    Lag2+Holiday*DBH+Holiday*MinL+MaxL*Lag2, nonlinVars=~MaxTI+MinL+MaxL+
    Lag1+Lag2,tau=0.5,data=A1[1:n,])
Box.test(residuals(QSAM), lag=30, type="Ljung")
tsdisplay(residuals(QSAM),main="Residuals diagnostic plots")

summary(QSAM,se="ker")

Actual <- L18[(n+1):(n+h)]
Fitted <- predict(QSAM,LoadSA[(n+1):(n+h),])
Predicted <- round(exp(Fitted))
Error <- Actual - Predicted

### Out-of-sample evaluations
```

```
MAE <- function(error){mean(abs(error))}
MAPE <- function(actual,error){100*(sum(abs(error/actual))/length(error))}
RMSE <- function(error){sqrt(sum(error*error)/length(error))}

MAE(Error)
MAPE(Actual,Error)
RMSE(Error)

##############################################################################
#                                                                            #
#                              BENCHMARK MODEL                               #
#                                                                            #
##############################################################################

#### Variable Selection using LASSO:
  **** WITH AGGREGATED TEMPERATURE
y <- log(LoadSA[1:n,18])
x <- data.matrix(LoadSA[1:n,-18])
c1 <- cv.glmnet(x,y,alpha=1)
coef(c1,s="lambda.1se")

  **** WITHOUT AGGREGATED TEMPERATURE
Y <- log(CI1[1:n,8])
X <- data.matrix(CI1[1:n,-8])
c2 <- cv.glmnet(X,Y,alpha=1)
b2 <- coef(c2,s="lambda.1se")

#### Models (first period)
  **** WITH AGGREGATED TEMPERATURE
Qr1 <- plaqr(formula= ts(log(L18))~Month+Trend+DayType+Holiday+DAH,
    nonlinVars=~MaxT+MinL+MaxL+AvgL+Lag1+Lag2,tau=0.5,data=LoadSA[1:n,])
Qr2 <- plaqr(formula= ts(log(L18))~Month+Trend+DayType+Holiday+DAH,
    nonlinVars=~MaxTI++MinL+MaxL+AvgL+Lag1+Lag2,tau=0.5,data=LoadSA[1:n,])

#### Model selection
bic(Qr1)
bic(Qr2)

## Then predict same as for the main model

##############################################################################
#                                                                            #
#                              COMBINED FORECASTS                            #
#                                                                            #
##############################################################################
MLpol0 <- mixture(model = "MLpol", loss.type = "pinball")
X1 <- cbind(M1(forecasts),BM1(forecasts))
Y1 <- Actual1
weights <- predict(MLpol0, X1, Y1, type='weights')
z <- ts(predict(MLpol0, X1, Y1, type='response'), start=c(2014,1), freq=365)
```

Codes.R

Appendix B

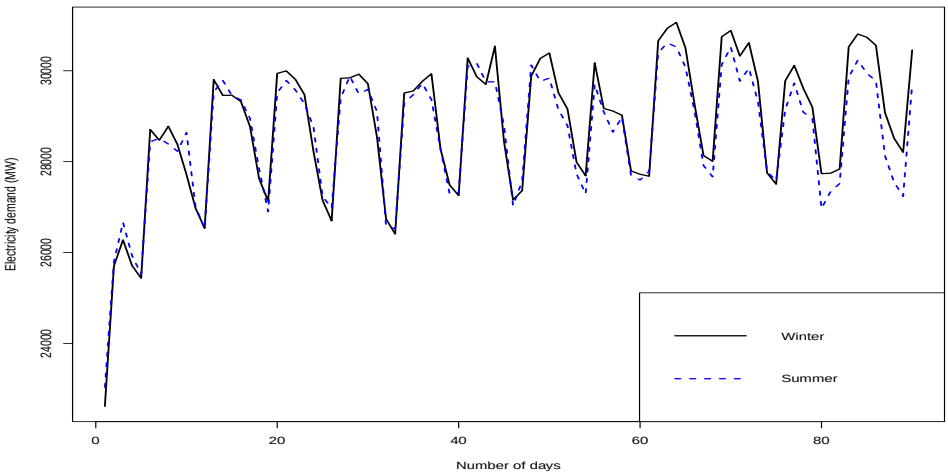


Figure 1: Actual and predicted electricity (MW) for the 19:00 hour.

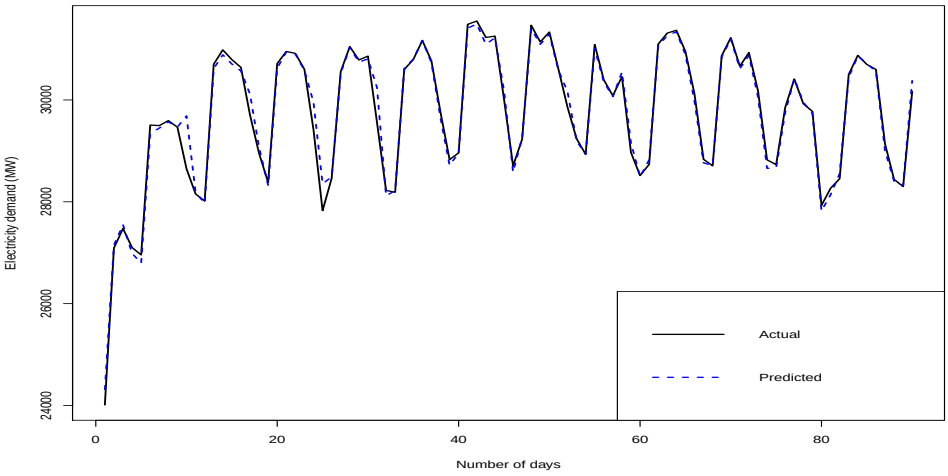
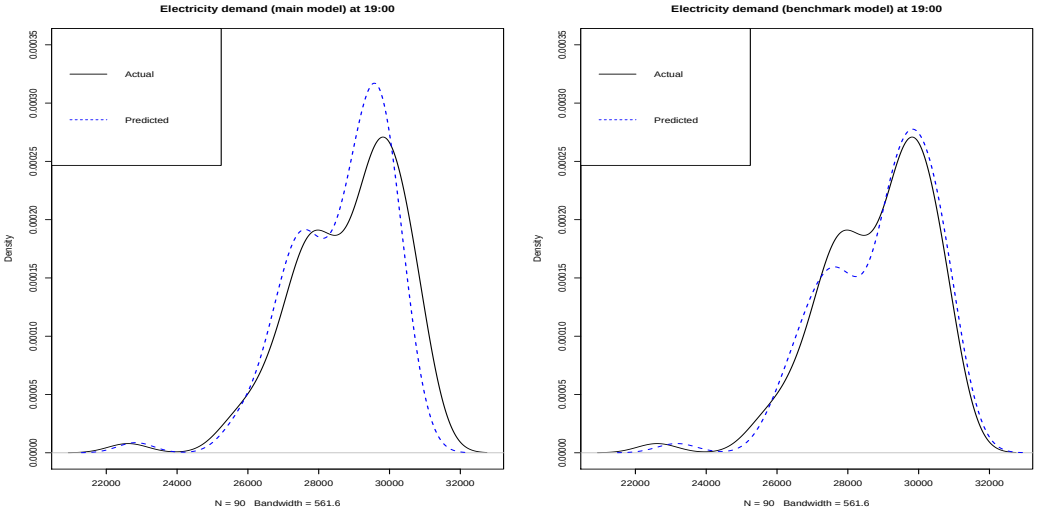
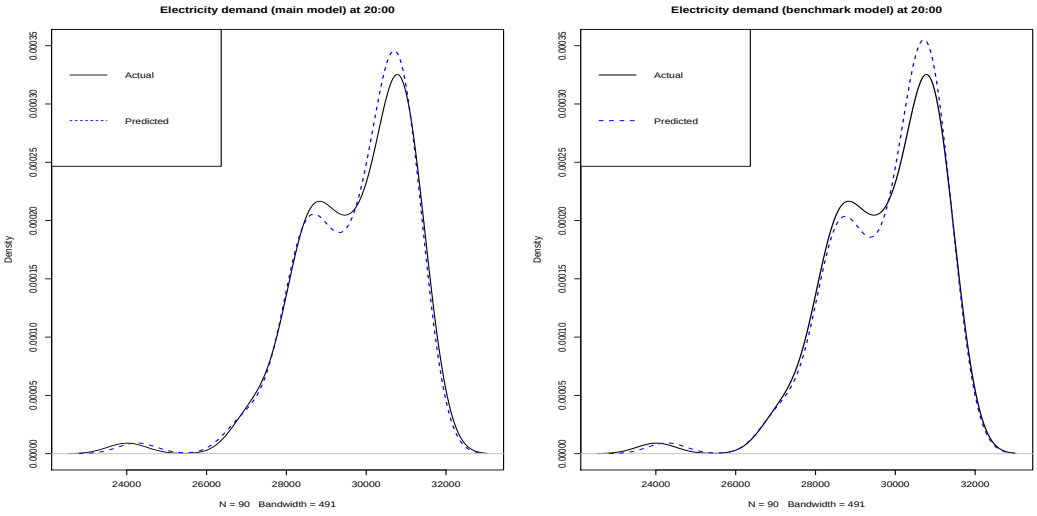


Figure 2: Actual and predicted electricity (MW) for the 20:00 hour.



(a) (b)

Figure 3: Density plots at 19:00 hours



(a) (b)

Figure 4: Density plots at 20:00 hours

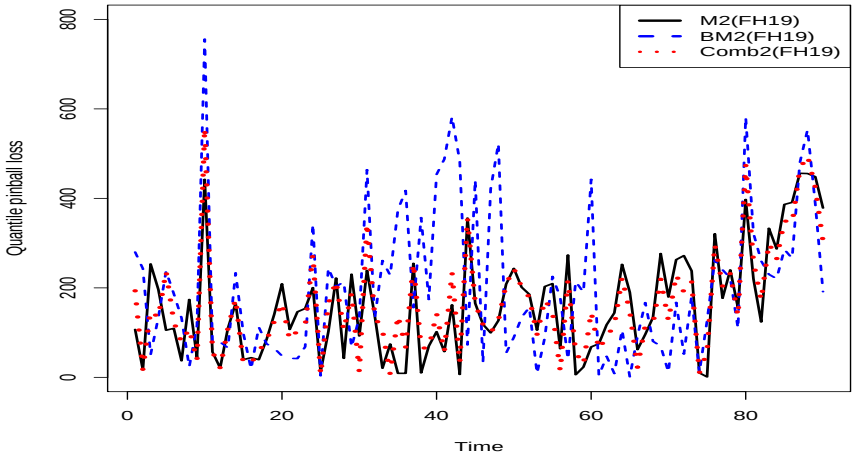


Figure 5: Quantile pinball loss on electricity demand (MW) at 19:00 hours.

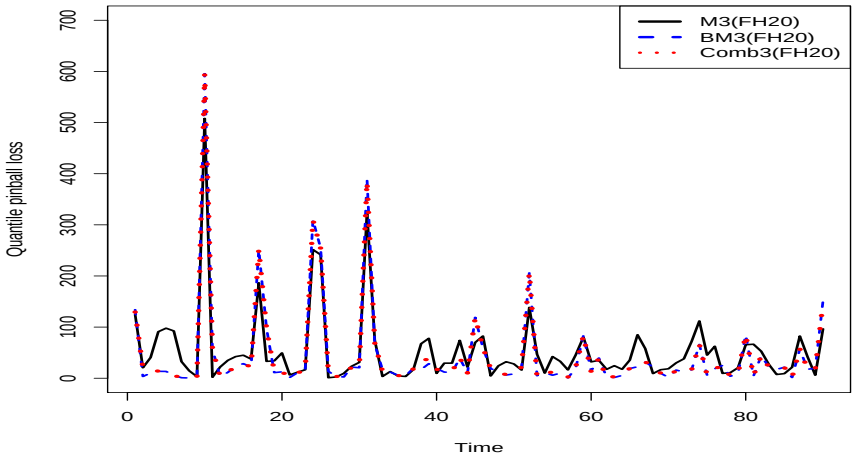


Figure 6: Quantile pinball loss on electricity demand (MW) at 20:00 hours.

Table 1: Variables selected using lasso via hierarchy interactions at 19:00 hours (with aggregated temperature)

	Main effect	Month	DayType	Holiday	DBH	DAH	T(H19)	MinL	MaxL	Lag1(H19)	Lag2(H19)
Month	0.0013	-0.0171	0	0	0	0	0	0	0	0	0
Trend	1e-04	0	0	0	0	0	0	0	0	0	0
DayType	-0.001	0	-0.001	0	0	0	0	0	0	0	0
Holiday	-2e-04	0	0	0	-0.0027	0	0	0	0	0	0
DBH	0	0	0	-0.0027	0	0	0	0	0	0	0
DAH	0	0	0	0	0	0	0	0	0	0	0
T(H19)	-0.0015	0	0	0	0	0	-2e-04	0	3e-04	0	8e-04
MinL	-1e-04	0	0	0	0	0	0	-1e-04	0	0	0
MaxL	0.07	0	0	0	0	0	3e-04	0	-0.0024	0	-0.0016
Lag1(H19)	3e-04	0	0	0	0	0	0	0	0	-3e-04	0
Lag2(H19)	0.0028	0	0	0	0	0	8e-04	0	-0.0016	0	0

Table 2: Variables selected using lasso via hierarchy interactions at 19:00 hours (with non-aggregated temperature)

	Main effect	Month	DayType	Holiday	DBH	DAH	TI(H19)	MinL	MaxL	Lag1(H19)	Lag2(H19)
Month	0.0013	-0.017	0	0	0	0	0	0	0	0	0
Trend	1e-04	0	0	0	0	0	0	0	0	0	0
DayType	-0.001	0	-0.001	0	0	0	0	0	0	0	0
Holiday	-2e-04	0	0	0	-0.0028	0	0	0	0	0	0
DBH	0	0	0	-0.0028	0	0	0	0	0	0	0
DAH	0	0	0	0	0	0	0	0	0	0	0
TI(H19)	-0.0016	0	0	0	0	0	-4e-04	0	3e-04	0	9e-04
MinL	-1e-04	0	0	0	0	0	0	-1e-04	0	0	0
MaxL	0.07	0	0	0	0	0	3e-04	0	-0.0024	0	-0.0016
Lag1(H19)	3e-04	0	0	0	0	0	0	0	0	-3e-04	0
Lag2(H19)	0.0028	0	0	0	0	0	9e-04	0	-0.0016	0	0

Table 3: Variables selected using lasso via hierarchy interactions at 20:00 hours (with aggregated temperature)

	Main effect	Month	Trend	DayType	Holiday	DBH	DAH	MinT	MinL	MaxL	AvgL	Lag1	Lag2
Month	0.0027	0	0	0	0	0	0	-1e-04	0	0.0016	0	0	0
Trend	-1e-04	0	0	0	0	0	0	0	0	-1e-04	0	0	0
DayType	-0.001	0	0	-9e-04	0	0	0	0	0	0	0	0	0
Holiday	-0.0017	0	0	0	0	-0.0017	0	0	0	0	0	0	0
DBH	-0.0015	0	0	0	-0.0017	0	0	0	0	0	0	0	0
DAH	-1e-04	0	0	0	0	0	-1e-04	0	0	0	0	0	0
MinT	-3e-04	-1e-04	0	0	0	0	0	1e-04	0	0	0	0	0
MinL	0.0042	0	0	0	0	0	0	0	-2e-04	0	-4e-04	0	0
MaxL	0.0518	0.0016	-1e-04	0	0	0	0	0	0	-0.0031	0	0	-1e-04
AvgL	0.0042	0	0	0	0	0	0	0	-4e-04	0	0	0	0
Lag1	0.0041	0	0	0	0	0	0	0	0	0	0	-7e-04	0
Lag2	0.0046	0	0	0	0	0	0	0	0	-1e-04	0	0	-9e-04

Table 4: Variables selected using lasso via hierarchy interactions at 20:00 hours (with non-aggregated temperature)

	Main effect	Month	DayType	Holiday	DBH	DAH	MinL	MaxL	AvgL	Lag1	Lag2
Month	0.0023	0	0	0	0	0	0	0.0013	0	0	0
DayType	-4e-04	0	-4e-04	0	0	0	0	0	0	0	0
Holiday	-0.0016	0	0	0	-0.0016	0	0	0	0	0	0
DBH	-0.0016	0	0	-0.0016	0	0	0	0	0	0	0
DAH	-1e-04	0	0	0	0	-1e-04	0	0	0	0	0
MinL	0.0044	0	0	0	0	0	-2e-04	0	-4e-04	0	0
MaxL	0.0527	0.0013	0	0	0	0	0	-0.0032	0	0	0
AvgL	0.0035	0	0	0	0	0	-4e-04	0	0	0	0
Lag1	0.0036	0	0	0	0	0	0	0	0	-5e-04	0
Lag2	0.0047	0	0	0	0	0	0	0	0	0	-9e-04

Table 5: Variables selected by using the lasso technique at 19:00 hours

Variable	With aggregated temperature	With non-aggregated temperature
(Intercept)	9.205880e+00	9.200660e+00
$Month$	·	·
$Trend$	·	1.987539e-06
$DayType$	·	·
$Holiday$	·	·
DBH	·	·
DAH	·	·
$Temp$	-5.960117e-04	
TC		-8.452663e-04
TI		-1.983896e-04
$MinT$	·	
$MinTC$		·
$MinTI$		·
$MaxT$	·	
$MaxTC$		·
$MaxTI$		·
$AvgT$	·	
$AvgTC$		·
$AvgTI$		·
$MinL$	·	-4.978459e-07
$MaxL$	3.623586e-05	3.694595e-05
$AvgL$	·	·
$Lag1$	·	·
$Lag2$	·	·

Table 6: Variables selected by using the lasso technique at 20:00 hours

Variable	With aggregated temperature	With non-aggregated temperature
(Intercept)	9.361552e+00	9.365561e+00
$Month$	2.519398e-04	2.092249e-04
$Trend$	·	·
$DayType$	·	·
$Holiday$	-1.212155e-02	-1.174636e-02
DBH	-9.660346e-03	-9.149687e-03
DAH	-1.679335e-03	-9.962666e-04
$Temp$	·	
TC		·
TI		·
$MinT$	·	
$MinTC$		·
$MinTI$		·
$MaxT$	·	
$MaxTC$		·
$MaxTI$		·
$AvgT$	·	
$AvgTC$		·
$AvgTI$		·
$MinL$	6.626079e-06	6.470429e-06
$MaxL$	2.419043e-05	2.425511e-05
$AvgL$	2.469696e-06	2.386712e-06
$Lag1$	3.589917e-06	3.431675e-06
$Lag2$	2.147117e-06	2.116686e-06

Table 7: Estimates from Model M1 at 18:00 hours

Parameter	Variable	Coefficient	Std. Error	t-value	p-value
c	$(Intercept)$	9.95017	0.01881	528.99748	0.00000
α	$Trend$	0.00001	0.00000	9.29686	0.00000
β_2	$DayType_2$	0.00491	0.00272	1.80517	0.07126
β_3	$DayType_3$	0.02527	0.00234	10.82081	0.00000
β_4	$DayType_4$	0.02507	0.00235	10.66533	0.00000
β_5	$DayType_5$	0.02200	0.00266	8.27421	0.00000
β_6	$DayType_6$	0.00851	0.00299	2.84752	0.00447
β_7	$DayType_7$	-0.00869	0.00288	-3.01732	0.00260
γ_2	$Month_2$	-0.00766	0.00123	-6.21604	0.00000
γ_3	$Month_3$	-0.00655	0.00137	-4.76925	0.00000
γ_4	$Month_4$	0.00076	0.00168	0.45147	0.65172
γ_5	$Month_5$	0.01996	0.00223	8.96022	0.00000
γ_6	$Month_6$	0.03718	0.00225	16.55738	0.00000
γ_7	$Month_7$	0.02277	0.00247	9.23673	0.00000
γ_8	$Month_8$	0.00410	0.00195	2.10026	0.03588
γ_9	$Month_9$	-0.00562	0.00162	-3.46315	0.00055
γ_{10}	$Month_{10}$	-0.00652	0.00158	-4.13696	0.00004
γ_{11}	$Month_{11}$	-0.00668	0.00145	-4.59720	0.00000
γ_{12}	$Month_{12}$	-0.00398	0.00131	-3.03837	0.00242
δ_1	$Holiday$	-0.05171	0.04682	-1.10446	0.26958
δ_2	DBH	0.00445	0.00175	2.54471	0.01104
κ_1	$bs(MaxTI)1$	-0.03522	0.00827	-4.25698	0.00002
κ_2	$bs(MaxTI)2$	-0.00151	0.00324	-0.46728	0.64037
κ_3	$bs(MaxTI)3$	-0.03593	0.00567	-6.33504	0.00000
η_1	$bs(MinL)1$	0.04950	0.01798	2.75244	0.00599
η_2	$bs(MinL)2$	0.08411	0.00905	9.29714	0.00000
η_3	$bs(MinL)3$	0.10140	0.01264	8.02131	0.00000
η_4	$bs(MaxL)1$	0.12357	0.01097	11.26324	0.00000
η_5	$bs(MaxL)2$	0.22623	0.00686	32.97892	0.00000
η_6	$bs(MaxL)3$	0.31064	0.00885	35.10343	0.00000
μ_1	$bs(Lag1)1$	0.03694	0.01617	2.28364	0.02254
μ_2	$bs(Lag1)2$	0.04344	0.00585	7.41994	0.00000
μ_3	$bs(Lag1)3$	0.07082	0.01039	6.81463	0.00000
μ_4	$bs(Lag2)1$	0.04015	0.01461	2.74906	0.00605
μ_5	$bs(Lag2)2$	0.04634	0.02026	2.28734	0.02232
μ_6	$bs(Lag2)3$	0.09278	0.03018	3.07371	0.00215
ν_1	$Holiday : Hbefore$	-0.06395	0.00837	-7.63779	0.00000
ν_2	$Holiday : MinL$	0.00000	0.00000	1.09093	0.27549
ν_3	$MaxL : Lag2$	0.00000	0.00000	-0.54683	0.58458

Table 8: Estimates from Model BM1 at 18:00 hours

Parameter	Variable	Coefficient	Std. Error	t-value	p-value
c	$(Intercept)$	9.93020	0.01790	554.71065	0.00000
θ_1	$Trend$	0.00001	0.00000	6.79858	0.00000
θ_2	$DayType_2$	0.01040	0.00404	2.57668	0.01008
θ_3	$DayType_3$	0.03342	0.00351	9.51568	0.00000
θ_4	$DayType_4$	0.03306	0.00361	9.16537	0.00000
θ_5	$DayType_5$	0.02952	0.00413	7.14917	0.00000
θ_6	$DayType_6$	0.01325	0.00437	3.03278	0.00247
θ_7	$DayType_7$	-0.01403	0.00407	-3.45112	0.00057
θ_8	$Month_2$	-0.00962	0.00151	-6.37318	0.00000
θ_9	$Month_3$	-0.00680	0.00166	-4.10489	0.00004
θ_{10}	$Month_4$	0.00170	0.00235	0.72599	0.46796
θ_{11}	$Month_5$	0.01850	0.00287	6.45321	0.00000
θ_{12}	$Month_6$	0.03316	0.00323	10.27816	0.00000
θ_{13}	$Month_7$	0.01785	0.00391	4.56090	0.00001
θ_{14}	$Month_8$	-0.00139	0.00284	-0.49045	0.62389
θ_{15}	$Month_9$	-0.00772	0.00207	-3.73252	0.00020
θ_{16}	$Month_{10}$	-0.00859	0.00216	-3.96988	0.00008
θ_{17}	$Month_{11}$	-0.00796	0.00190	-4.19150	0.00003
θ_{18}	$Month_{12}$	-0.00759	0.00179	-4.24735	0.00002
θ_{19}	$Holiday$	-0.00789	0.00286	-2.75517	0.00594
θ_{20}	DAH	-0.00923	0.00289	-3.19288	0.00144
θ_{21}	$bs(MaxTI)1$	-0.02906	0.00938	-3.09696	0.00199
θ_{22}	$bs(MaxTI)2$	-0.00761	0.00468	-1.62647	0.10407
θ_{23}	$bs(MaxTI)3$	-0.03268	0.00665	-4.91223	0.00000
θ_{24}	$bs(MinL)1$	0.07484	0.02093	3.57501	0.00036
θ_{25}	$bs(MinL)2$	0.07247	0.01098	6.60144	0.00000
θ_{26}	$bs(MinL)3$	0.08738	0.01514	5.77276	0.00000
θ_{27}	$bs(MaxL)1$	0.12827	0.01592	8.05715	0.00000
θ_{28}	$bs(MaxL)2$	0.20591	0.01011	20.37216	0.00000
θ_{29}	$bs(MaxL)3$	0.28472	0.01426	19.97241	0.00000
θ_{30}	$bs(AvgL)1$	-0.01191	0.01216	-0.97932	0.32759
θ_{31}	$bs(AvgL)2$	0.03218	0.00694	4.63294	0.00000
θ_{32}	$bs(AvgL)3$	0.04518	0.01088	4.15264	0.00003
θ_{33}	$bs(Lag1)1$	0.04315	0.02095	2.06012	0.03957
θ_{34}	$bs(Lag1)2$	0.04896	0.00870	5.62558	0.00000
θ_{35}	$bs(Lag1)3$	0.07188	0.01310	5.48530	0.00000
θ_{36}	$bs(Lag2)1$	0.06793	0.01997	3.40152	0.00069
θ_{37}	$bs(Lag2)2$	0.05096	0.01005	5.06951	0.00000
θ_{38}	$bs(Lag2)3$	0.10713	0.01404	7.62796	0.00000