Probabilistic long-term electricity demand forecasting in South Africa

PHD THESIS

Supervisors:
Prof. Pravesh Debba

Prof. V.S.S. Yadavalli

Author:
Paul Moloantoa Mokilane

A thesis submitted for degree of the Doctor of Philosophy

in the

Faculty of Science,
School of Statistics and Actuarial Science,
University of the Witwatersrand
Johannesburg

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Abstract

Electricity demand forecasts inevitably differ from the actual electricity demand, hence the task of a forecaster is to make some educated guesses of future electricity demand and to effectively communicate them. Forecasts come with uncertainties whose quantification is as important as the forecasts themselves. Probabilistic modelling accounts for uncertainties in estimation, prediction or in forecasting. Uncertainties in future electricity demand emanate from changing weather conditions, market penetration of renewable sources of electricity, power saving appliances, electric vehicles, escalating cost of electricity, transition of economic sectors away from reliance on energy-intensive towards a more diverse range of services-oriented activity or vice versa and unpredictable economic growth.

Short-term electricity demand forecasts are used to ensure system stability. They are used for planning the day-to-day running of the electricity generation system. Medium-term forecasts are used for maintenance planning. They are used to schedule maintenance in such a way that electricity demand is met during maintenance period. Long-term forecasts are used for capital planning. They assist in determining whether the current generation infrastructure will still generate enough electricity to meet future demand.

Literature has shown that a universally best electricity demand forecasting technique is non-existent, it can therefore be argued that a good scientific approach to electricity demand forecasting will produce good forecasts. This study involves forecasting electricity demand for long-term planning purposes using both frequentist and Bayesian modelling approaches. Long-term forecasts for hourly electricity demand from 2007 to 2023 are done with in-sample forecasts from 2007 to 2012 and out-of-sample forecasts from 2013 to 2023 and 2013 to 2015 is used to validate the models.

Quantile regression (QR) in the frequestist modeling paradigm is used to forecast hourly electricity demand at various percentiles of the distribution of electricity demand. The findings are that the future distributions of hourly electricity demand and peak daily demand would be more likely to shift towards lower demand over the years until 2023.

Bayesian structural time series (BSTS) in the Bayesian modeling paradigm is used to forecast hourly distribution of electricity demand. Accurate trend specification is important in long-term forecasting, otherwise erroneous forecasts could be obtained, especially in South

Africa where it is difficult to determine if the demand trend would continue a downward trajectory, stabilise or would revert to an upward trajectory. The findings are that future South African hourly demand until 2023 are less likely to exceed the highest historical hourly demand of 36 826 kW. The South African electricity demand from Eskom are more likely to maintain the downward trend until 2023.

The developed generalised additive mixed quantile averaging (GAMMQV) model indicates that demand for electricity is unlikely to exceed the highest (36 826 kW) hourly electricity demand until 2023. The probability of electricity exceeding 36 826 kW was below 0.15 from 2013 to 2023. The GAMMQV model gave better point forecasts than BSTS and QR.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, to any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

Signature.....................................

Date: March 27, 2019

Dedication

This work is dedicated to all the working, part-time PhD students who dare not give up despite pressure at work and family responsibilities. They always source some motivation and strength somewhere to continue working on their PhDs.

Acknowledgements

The research was conducted during a contract research project funded by Eskom, the CSIR Parliamentary grant and some funding from the Thuthuka fund of the National Research Fund (NRF). I would therefore, like to thank Eskom, the CSIR and the NRF for funding the research. I would also like to thank the following Eskom staff members for supplying the data; Moonlight Mbata and Ntokozo Sigwebela. I would like to thank the following CSIR staff members for their technical support: Nontembeko Dudeni-Tlhone, Nosizo Sebake, Renee Koen, Jenny Holloway, Sibusisiwe Makhanya and Siphamandla Sibiya.

I wish to express my sincere appreciation to Professor Jacky Galpin who was part of my supervision team earlier in my research, but retired later on. Her guidance during the development of my proposal and her contribution on quantile regression paper is highly appreciated. I would like to thank Dr. Caston Sigauke for his guidance, advice and for making time to critique my work. My sincere thanks to my supervisors, Professors: Pravesh Debba and Venkata S.S. Yadavalli, for their stimulating suggestions, guidance, as well as their encouragement which assisted me through the research.

To my employer CSIR for not only financing the study, but for also giving me permission to use the CSIR resources and time to work on the thesis, I am deeply indebted to the organisation and I also want to express my sincere appreciation to Renee Koen for language editing and proofreading this document.

I would also like to thank my wife Kamohelo and my three children, Tebello, Teboho and Thato who gave me huge support throughout my research work.

Contents

List of Figures

List of Tables

List of abbreviations and acronyms

QR	Quantile Regression
OLS	Ordinary Least Squares
BSTS	Bayesian Structural Time Series
AR	Autoregressive
CI	Confidence Interval
Eskom	Electricity Supply Commission
PI	Prediction Interval
GAM	Generalised Additive Model
GAMM	Generalised Additive Mixed Model
GAMMQV	Generalised Additive Mixed Quantile Averaging Model
MA	Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average
GCV	Generalized Cross Validation
ARMA	Autoregressive Moving Average
KW	Kilo Watts
GPD	Generalized Pareto Distribution
i.i.d.	independent and identically distributed
SARIMA	Seasonal Autoregressive Integrated Moving Average
DSARMA	Double Seasonal Autoregressive Moving Average
TSARMA	Triple Seasonal Autoregressive Moving Average
MLE	Maximum Likelihood Estimation
USTS	Univariate Structural Time Series
SARIMAX	Regression Seasonal Autoregressive Integrated Moving Average
P-P plot	Probability-Probability plot
CSIR	Council for Scientific and Industrial Research
Q-Q plot	Quantile-Quantile plot
MLR	Multiple Linear Regression
MAPE	Mean Average Percentage Error
GDP	Gross Domestic Product
MCMC	Markov Chain Monte Carlo
ICC	Intraclass Correlation Coefficient

ANN	Artificial Neural Networks
AI	Artificial Intelligence
IPP	Independent Power Producers
SURE	Seemingly Unrelated Regression Equations

List of notation and special symbols

$F_X(x)$	distribution function
$f_X(x)$	density function
$\ell(\theta, x)$	maximum likelihood function
$s_h j$	cubic spline
log	natural logarithm
Q_τ	quantile function
μ_0	intercept coefficient of the location parameter
μ_i	coefficient of the location parameter
σ	scale parameter
σ_0	intercept coefficient of the scale parameter
σ_1	slope coefficient of the scale parameter
σ_2	quadratic/parabolic coefficient of the scale parameter
α	shape parameter
ϕ_τ	probability of exceeding the threshold
λ	smoothing parameter
$\pi(\theta)$	prior distribution
$\pi(\theta \mid x)$	posterior distribution
$\pi(x \mid \theta)$	likelihood function
θ	target parameter

Research outputs

A list of research outputs from this thesis are,

Peer reviewed publications

1. Mokilane, P., Galpin, J., Yadavalli, V.S.S., Debba, P., Koen, R. and Sibiya, S. (2018). Density forecasting for long-term electricity demand in South Africa using quantile regression, South African Journal of Economic and Management Sciences, vol. 21, pp. 1-14.

2. Mokilane, P., Debba, P., Yadavalli, V.S.S., and Siguake, C. (2019). Bayesian structural time series approach to long-term probabilistic forecasting of electricity demand in South Africa. Applied Mathematics and Information Sciences Journal, vol. 13, no. 2, pp. 189-198.

3. Mokilane, P., Debba, P., Yadavalli, V.S.S., and Siguake, C. (2018). Generalised additive mixed model approach to long-term electricity demand forecasting in South Africa. Industrial Engineering and Operations Management (IEOM), pp. 1618-1629, http://ieomsociety.org/southafrica2018/proceedings/.

Chapter 1

Introduction

1.1 Background

A power utility company like the South African "Electricity Supply Commission" (Eskom) that manages the electricity grid for the country must ensure the balance between electricity supply and demand at all times for the country. Unfortunately, electricity demand does not follow electricity supply because when the utility company is unable to generate enough electricity, consumers do not automatically adjust their electricity demand accordingly to align with electricity supply. Electricity cannot be massively stored for later use, hence electricity is generated and delivered to consumers for immediate consumption. Eskom generates approximately 95% of the total electricity consumed in South Africa from coal and nuclear while municipal-owned power plants and independent power producers (IPPs) generate the remaining 5% (Deloitte, 2017). IPPs mostly generate electricity from renewable sources of energy, such as wind and solar. Electricity is generated from wind through wind power via turbines while electricity from solar is generated from solar power via photovoltaic cells and other technologies. Until technology is discovered to massively store surplus electricity when abundant electricity can be generated, for use in times when supply is limited, full reliance on renewable electricity would be a risky option despite most governments globally wanting to reduce the greenhouse gas emissions from fossil fuel power plants for environmental imperatives. Utility companies do not have full control over the amount of electricity they generate from renewable sources of energy because wind turbines rely entirely on air flow, and the wind does not always flow at the steady rate and solar cells do not always receive a constant level of sunlight. The amount of electricity generated from renewable sources of electricity is therefore beyond the power generation company's control. The higher percentage of electricity generated in South Africa can easily be managed as it is generated from coal and nuclear that are not directly dependent on wind speed and level of sunlight.

Electricity load is defined as the amount of electricity that balances the amount generated with that drawn from the electricity grid (Cottet and Smith, 2003). In the absence of black-outs and load-shedding, electricity load is equivalent to electricity demand (Cottet and Smith, 2003). In this study, the hourly electricity demand is defined as the amount of electricity (load) in kW sent out every hour by Eskom to meet consumers' demand in South Africa. Therefore, Electricity demand from Eskom is used as a proxy for electricity demand for South Africa in this study.

Access to electricity in this study is defined as the percentage of South African households who have access to electricity for lighting. Electricity was originally used solely for lighting according to Hong and Shahidehpour (2015). Access to electricity for lighting is therefore used as a measure of access to electricity by the South African households in this study. The 1996 census showed that only 57.6% of the South African households had access to electricity (Statistics South Africa, 1998). The 2001 census showed that this percentage went up to 70.2% (Lehohla, 2005). The 2007 Community Survey indicated that 80.1% of the South African households had access to electricity (Statistics South Africa, 2008). The 2011 census showed that this percentage went up to 84.7% (Statistics South Africa, 2011). These censuses and surveys indicate that a high percentage of new households that were connected to the electricity grid between 1996 and 2007, which would imply that residential electricity demand would have increased during the same period. The percentage of new households connected to the grid stabilised between 2007 and 2011 and, therefore, residential electricity demand would be expected to have stabilised during this period. The shrinking South African economic growth between 2007 and 2015 could have contributed to a decline in electricity demand. South Africa experienced an average economic growth rate of approximately 5% in real terms between 2004 and 2007. However, during the period 2008 to 2012 the average economic growth dropped to just above 2% (Deloitte, 2017). The real GDP continued declining from about 5% in 2006 to below 1% in 2016 (Deloitte, 2017).

The penetration of other sources of electricity, such as solar and wind, could also have contributed to a decline in demand by customers from Eskom. In addition, because of the lack of capacity in the generation of electricity experienced by Eskom in 2007 (Inglesi and Pouris, 2010), some companies and households had to find other sources of electricity, which would have resulted in a decline in electricity demand from Eskom. Deloitte (2017) has also indicated that South Africa experienced a series of highly disruptive power outages in early 2008 as Eskom began a programme of load-shedding as a response to a shortage of generating capacity. Load-shedding events were re-introduced in early 2014 and they happened regularly between November 2014 and September 2015. Unfortunately, the actual amount of electricity demand in South Africa is unknown because of the unavailability of certain types of data, such as the amount of renewable electricity that companies and some households generate for themselves. The combined effect of all changes in demography, weather conditions, technology using electricity, economy, usage patterns and growing renewable sources of electricity, future electricity tariffs (in the monopolised South African electricity market) contribute to uncertainties when trying to forecast future electricity demand, especially when data on these drivers of demand are not sufficiently collected.

Uncertainties occur in estimation, prediction or in forecasting. When statisticians develop predictions (forecasts) for an uncertain future, they need to quantify uncertainties around these forecasts for decision makers. Uncertainties in future electricity demand patterns could emanate from uncertainties regarding future changes in technologies making use of electricity, rates of population growth, general randomness in individual usage of electricity, seasonal effects, prevailing economic patterns, change in weather conditions, escalating costs of electricity, use of power saving electrical appliances, growing sources of renewable electricity especially if data is not sufficiently collected on this, market penetration of electric vehicles, in addition to model uncertainties (e.g. whether the demand trend will follow an upward or downward trajectory in the long-term). The inherent uncertainties in predictions imply that forecasts should ideally be probabilistic; in other words, they should take the form of probability distributions over future quantities or events (Gneiting and Katzfuss, 2014). Probabilistic forecasts could take the form of quantiles, prediction intervals, fan charts, scenarios or density forecasts. Tay and Wallis (2000) define density forecasts of the realisation of a random variable at some future time as estimates of the probability distribution of the possible future values of that variable. Probabilistic forecasts are essential ingredients of optimal decision-making (Gneiting and Katzfuss, 2014). It is important to quantify uncertainties around the electricity demand forecasts, especially for long-term planning purposes in order to avoid building unnecessary infrastructure while at the same time ensuring that future electricity demand is met.

Hong et al. (2014) argue that forecasting is by nature a stochastic problem, but that most of the utilities are still developing and using point forecasts. Hong et al. (2014) state that it would be better to use probabilistic forecasts that provide estimates of the full distribution of possible future values as a way of quantifying uncertainties in the forecasts.

In the late 1880 s, when lighting was the sole end use of electricity, the forecasting of electricity demand was straightforward (Hong and Shahidehpour, 2015). Power generating companies would count the number of light bulbs they installed and planned to install and they would then roughly estimate the level of demand in the evening. As electric appliances such as electric irons, radios, television sets, geysers, stoves and washing machines were invented and commonly used in many households, the complexity of forecasting electricity demand grew. The penetration of air conditioners into homes and offices to regulate temperature within human comfort zones, resulted in temperature becoming an important driver of electricity demand. These drivers of electricity demand add complexity to electricity demand forecasting and create uncertainties around the forecasts. Uncertainty in forecasts can be quantified as an interval around the forecast in which repeated forecasts will fall. The quantification of uncertainties are important (Hyndman and Fan, 2010). Hyndman and Fan (2010) argue that an overestimate of the long-term electricity demand could result in substantial wasted investment due to the construction of excess power facilities, while an underestimate of demand could result in insufficient generation and unmet future demand. Electricity forecasting methods have evolved from counting light bulbs and engineering approaches which were based on the use of charts and tables, to manually forecasted demand, to sophisticated forecasting techniques. The availability of powerful computers and statistical software today enables forecasters to produce more accurate forecasts through sophisticated forecasting methods.

Electricity demand forecasts can be developed for short, medium or long-term horizons. They could be provided as point forecasts, which give one value at each step within the forecast horizon, or as probabilistic forecasts which give a full distribution of future values at each step within the forecast horizon and therefore allow for the assessment of uncertainties around the forecasts. The quantification of uncertainties around the forecasts is even more important for long-term forecasting, as Sigauke and Chikobvu (2011) indicated that long-term decision-making in the electricity sector involves planning under substantial uncertainty.

According to McSharry et al. (2005), short-term forecasts are future values ranging from five minutes to one week ahead, and are used to ensure system stability, medium-term forecasts are mainly used for maintenance scheduling and they range from one week to six months ahead, while long-term forecasts are used for capital planning and they range from six months to many years ahead.

In the literature to date, short-term electricity demand forecasting has attracted substantial attention because of its importance for power system control, unit commitment and electricity markets. Medium and long-term forecasting have not received much attention, despite their value for system planning and budget allocation (Hyndman and Fan, 2010). The literature on probabilistic load forecasting is very limited, and load forecasting is still dominated by short-term point forecasting.

There is a vast body of literature on long-term forecasting of annual electricity demand as well as peak demand in South Africa as shown in Inglesi-Lotz (2011); Koen et al. (2017); Sigauke (2014); Sigauke and Chikobvu (2011); Ziramba (2008) and Boraine and Yadavalli (2003). Maharaj and Yadavalli (2017) modelled the stochastic behaviour of the price dynamics of electricity spot prices with a three-regime Markov switching model and they concluded that seasonality and uncertainties play important roles in predicting hourly spot prices. Little has been published on forecasting the full profile of electricity demand and the popular forecasts are for short-term point or only peak demand forecasting.

1.2 Problem statement

Regardless of the limitations of point forecasts, especially in the long-term forecasting horizon, the literature has shown that point forecasts are still popularly used for electricity planning, and not only in South Africa. Probabilistic forecasts for long-term electricity demand are, however, recommended because of the uncertainties inherent in future electricity demand and the far reaching consequences that could result from making decisions based on a single forecast. Uncertainties in electricity demand forecasts are quantifiable in probabilistic forecasting approaches and they enhance optimal decision making.

The South African electricity demand trend showed a fluctuating trajectory. It followed an upward trajectory between 1997 and 2007, then stabilised between 2008 and 2011, and started declining in the last four years until 2015, as shown in Figure 1.1. There are uncertainties pertaining to whether electricity demand trend will continue declining or if the trajectory will change in future. It is, therefore, important for the developed models to correctly capture the direction of the future electricity demand trend or, rather, that the trend should contribute minimally to the developed models, otherwise the long-term forecasts of electricity demand would vastly deviate from the future actual electricity demand if the actual trend takes the opposite direction to the anticipated direction.

The literature has shown that researchers are still searching for a universally best electricity demand forecasting technique. The choice of a modelling technique must be data-driven, objective-driven and it should be country specific. The electricity demand data, especially hourly electricity demand patterns, are complicated and attract different modelling techniques. The demand data could attract hierarchical modelling techniques because demand could be seen to be embedded within the day, day within month and month within the year and, if the forecaster wants to take this data structure into consideration, then a hierarchical modelling approach is used. It could also attract stochastic time series models like the seasonal autoregressive integrated moving average (SARIMA) model because the data can be seen to consist of a sequence of observations collected over equally spaced time periods with no missing data, and therefore that the observations are correlated in time. It could also attract regression based approaches where hourly demand could be modelled separately and model parameters estimated either separately (ignoring the intraday correlation) or simultaneously (taking intraday correlation into consideration). It could also attract generalised additive models (GAMs) or semi-parametric models which are considered as statistical learning techniques. None of these different approaches would be wrong, but the forecaster must understand the theory behind the modelling technique of choice and its limitations because, a good scientific approach to forecasting need to be taken in order to produce robust forecasts.

There is still a gap in literature on long-term electricity demand forecasting using probabilistic approaches in South Africa.

1.3 Study aim and objectives

The aim of this study is to develop a modelling framework which can be used to accurately forecast the long-term hourly electricity demand in South Africa in such a way that uncertainties are quantifiable in the process. The quantification of uncertainties in long-term electricity demand forecasts are as important as the forecasts themselves. Therefore, probabilistic approach to electricity demand forecasting is important. The density functions representing the distribution of the hourly electricity demand in each year within the forecasting horizon are forecasted in this study and the objectives are:

- to forecast the density functions of the hourly electricity demand over an entire year for all years from 2013 to 2023,
- to apply a quantile regression (QR) model to forecast hourly electricity demand in South Africa between 2013 and 2023 in the presence of uncertainties,
- to apply a Bayesian structural time series (BSTS) model to forecast the distribution of hourly electricity demand in South Africa between 2013 and 2023 with the inclusion of a nonlinear trend. The local level trend which is assumed to follow a random walk is used in this model,
- to develop and apply a generalised additive mixed model (GAMM) with the inclusion of a nonlinear trend to forecast the distribution of hourly electricity demand in South Africa between 2013 and 2023. The demand trend is extracted from the electricity demand data, modelled and its future evolution is forecasted. The within hour correlation structures are harnessed in this model,
- to investigate the variabilities in the forecasts and evaluate uncertainties around the point forecasts,
- to study the shifts in demand distributions, that is, whether the distribution of electricity demand is shifting towards higher or lower demand over the years and to evaluate the shifts in morning and afternoon peak demand distributions in South Africa, and
- to find and recommend the model that will give robust long-term probabilistic forecasts of hourly electricity demand in South Africa.

1.4 Scope of study

The study intends to use the hourly electricity demand data provided by Eskom which were recorded from 1997 to 2015 , in order to forecast hourly electricity demand from 2013 to 2023. The electricity demand from 2013, 2014 and 2015 are used as the validation data. These are the amounts of electricity required to meet the South African customers' demand every hour and hence electricity demand is measured at the distribution point. Important drivers of demand, as identified from literature and previous studies, such as calendar and time variables, Fourier series or harmonic terms related to seasonal patterns and lagged demand variables were considered in the modeling. The harmonic terms were used to capture the cycles inherent in the demand data while the lagged demand variables were used to capture the autocorrelations inherent in the demand data. The harmonic terms and lagged variables were used in the QR and BSTS models.

The study used both parametric and semi-parametric modelling approaches. In parametric modelling, the electricity demand model is completely defined by a set of parameters, it uses the assumptions of the distribution to estimate the parameters of the distribution assumed for the electricity demand data. In nonparametric modelling, the model is not completely defined by the fixed set of parameters and the distributional assumptions for the data are not made while the semi-parametric model is the hybrid of the two. Under the parametric approach, QR and BSTS were considered, while the GAMM was considered under a semi-parametric modelling approach. The model parameters in the parametric approaches are estimated in both frequentist and Bayesian modelling paradigms. In the frequentist approach, the data is considered random and the model parameters fixed, QR is used as the modelling approach in this framework. The data is considered fixed and the model parameters random in the Bayesian approach in which BSTS is located.

The entire distribution of the hourly electricity demand is forecasted using a QR model, which result in uncertainties around the forecasts being quantified. In this modelling framework, electricity demand is forecasted at various percentiles of the electricity demand distribution. The densities of hourly demand and the peak demand over the years are forecasted using QR.

The entire distribution of electricity demand is forecasted using a BSTS model. In a Bayesian modelling framework, the model parameters are estimated with their full distributions. Uncertainties inherent in the forecasts are incorporated in a coherent fashion and, hence, the entire distribution of future hourly demand is forecasted.

GAMM is a semi-parametric modelling approach because it has both parametric and nonparametric components. The relationship between electricity demand and its drivers is captured by smooth functions (basis functions) and some parametric terms. The GAMM model has been extended to the GAMM quantile averaging (GAMMQV) model. The hourly electricity demand is forecasted at $0.01, 0.02, 0.03, \ldots, 0.99$ quantiles of the demand distribution using the GAMMQV model. Finally, comparisons between the three models considered in this study are made.

1.5 Modelling framework

The hourly electricity demand could be modelled through a Bayesian or a frequentist approach. Both approaches are considered in this study. In the frequentist approach, the attention is restricted to phenomena that are repeatable under identical conditions and the probability of an event is defined to be the limiting relative frequency with which the outcome would occur in n repetitions as n goes to infinity. The model parameters in this framework are considered to be fixed while the data are random. Under this framework, QR is an attractive approach for probabilistic forecasting in which the demand is forecasted at $0.01,\ 0.02,\ 0.03, \ldots,\ 0.99$ quantiles of the demand distribution, hence each hour of the day would have 99 forecasted future hourly demand from 99 regression models, instead of forecasting just a single expected hourly demand as in the case of ordinary least square regression (OLS).

A Bayesian approach is a natural approach to probabilistic forecasting, where the data are considered to be fixed and the model parameters random. Under this paradigm, the model parameters are estimated with their full distributions. There are two approaches to the Bayesian modelling, the Full and Empirical Bayes. In Full Bayes modelling, the prior parameters or the distribution of the prior parameters are assumed known, while in Empirical Bayes modelling, the observed data are used to estimate the prior parameters and then proceed as though the priors were known. In both Bayes approaches, the prior distribution represents what is known about the parameters before observing the data, which can be based on past experience (i.e. *expert information*) gained from previous related studies. If such expert information is not available, then non-informative priors are used which give the data more weight in the posterior distribution. The next step is to collect data and form a likelihood function, and then the prior is combined with the likelihood function to form the posterior distribution. A BSTS modelling approach was considered as ideal for forecasting the South African electricity demand because the demand trend has shown changes in the recent past and therefore it may be difficult to make assumptions regarding the shape of the future electricity demand's trend. BSTS modelling approach makes it possible to determine the contribution of each component of the model. Therefore, it makes it possible to investigate the origin of the problem if results appear to be counter-intuitive.

A generalised additive mixed model (GAMM) is an additive and functional modelling technique which captures the impact of predictive variables through smooth functions which could be nonlinear. The goal of GAMM in this study is to model electricity demand (response variable) using its drivers (independent variables), which are expressed in the form of some smooth functions (splines). Smoothing splines are real functions that are piecewise-defined by polynomial functions called basis functions. The places, where the polynomial pieces connect are called knots. In GAMM, penalized regression splines are used in order to regularize the smoothness of the spline. GAMM assumes that model errors are identically and independently distributed. This assumption is not fulfilled in the case of time series regression. Present values of the time series are correlated with past values, so errors of the model are correlated. They are said to be autocorrelated. This implies that estimated regression coefficients and residuals of the model might be biased, which also implies that confidence intervals would be wrong. This problem in GAMM is avoided by modeling the within hours correlation structures using AR and ARMA. Harnessing the correlation structures within-hours could improve the forecasts vastly. The nonlinear trend is included with the future values forecasted using the QR model. The future trend values were forecasted at $0.01,\ 0.02,\ 0.03, \ldots,\ 0.99$ quantiles of the distribution.

1.6 Significance of the study

The economic growth of any country is dependent on its energy security. For planning purposes, especially under conditions of limited resources, it is important to ensure the accuracy of the long-term electricity demand forecasts. This will not only enable the power generating company to avoid substantial wasted investment due to the construction of excess power facilities, but will also ensure that future electricity demand is sufficiently met. Due to the fact that uncertainties around the forecasts are quantifiable in probabilistic forecasting, this will be an essential ingredient for optimal decision making in the presence of uncertainties.

The current electricity demand from Eskom follow a downward trend which is partly caused by the growing renewable sources of electricity. Due to unavailability of sufficient data on other sources of electricity like renewable sources of electricity, the study discourages any long-term demand forecasting using point forecasts as there are too much uncertainties that need to be quantified for planning purposes. Point forecasting gives only the mean of the distribution of electricity demand, and the probabilities of exceeding certain values cannot be calculated, and hence this could lead to poor decision making.

1.7 Contributions

As indicated in Chapter 7, the contributions of this study are the re-contextualisation of the existing QR, BSTS and the GAMM modelling techniques, which resulted in the application of these modelling techniques to a totally new context (the probabilistic forecasting of the long-term electricity demand). The study highlights the importance of probabilistic approaches to long-term electricity demand forecasting because of the uncertainties surrounding possible future patterns of electricity demand in South Africa. Uncertainties make it difficult to defend any long-term point electricity demand forecasts which are still popularly used in South Africa. Probabilistic forecasts are less popular compared to point forecasts because they are not communicated in such a way that the end users (planners) who might not be statisticians would understand and appreciate their benefits. The study paid attention to the effective communication of the probabilistic forecasts and the contributions are as follows:

- the literature on long-term electricity demand in South Africa using probabilistic approaches is either non-existent or not documented to the best of our knowledge,
- in search of the best statistical model for forecasting the long-term electricity demand in South Africa, both parametric and non-parametric approaches to statistical modelling were used due to the complexity of the electricity demand data. Under a parametric approach, the frequentist and Bayesian modelling paradigms were considered. Therefore, in this study, both approaches to statistical modelling (non-parametric and parametric) and both opposing paradigms to statistical modelling (frequentiest and Bayesian) were explored,
- QR is presented in this study as an alternative modelling approach to long-term probabilistic forecasting of electricity demand in South Africa. In the frequentist framework, QR was used to forecast the long-term hourly electricity demand at $0.01, 0.02, 0.03, \ldots, 0.99$ quantiles of the demand distribution, enabling variabilities in the forecasts to be evaluated and uncertainties around the forecasts to be assessed while probabilities of exceeding certain electricity demand can be calculated,
- the study presents the development and application of a GAMM model with non-linear trend in long-term electricity demand forecasting in South Africa. GAMM can either be represented as a nonparametric or as a semi-parametric model. The model allows flexible specification of the dependence of the response variable on the covariates in terms of both smooth and parametric functions. In order to capture the correlation inherent in the time series, the correlation structures within different hours were harnessed using autoregressive (AR) and autoregressive moving average (ARMA). The nonlinear trend with forecasted values at different quantiles of its distribution were included as part of the fixed effects component of the GAMM model. The GAMM model was extended to a generalised additive mixed quantile averaging model (GAMMQV),
- a Bayesian structural time series (BSTS) model is presented as another alternative approach to long-term probabilistic forecasting of the hourly electricity demand. In the Bayesian framework, a BSTS model was used to forecast the long-term distribution of hourly electricity demand in South Africa. Electricity demand trend in South Africa has shown changes over the recent past, because it showed an upward trajectory between 1997 and 2007, stabilised between 2008 and 2011 and started declining in the last four years until 2015. Therefore, there are some uncertainties with regard to whether it will continue declining or whether the trajectory will again change in future. It is therefore important to get the direction of the trend correct otherwise the long-term forecasts could vastly deviate from the future actual electricity demand. The BSTS model gives the forecaster a mechanism to assess the contribution of each component of the model, hence in the case of getting counter-intuitive

forecasts, the origin of the problem could easily be located. Due to uncertainties with the future direction of electricity demand trend, a good forecasting model with trend minimally contributing to the overall model was chosen. Such a modelling approach to long-term electricity demand forecasting has so far not yet been documented in the literature,

- for the long-term forecasting of the electricity demand in South Africa, it is important to accurately capture the evolution of the demand trend because the long-term forecasts could vastly deviate from the actual demand if the trend is not accurately captured. On the basis that the long-term demand trend in South Africa is unpredictable, three models distinctly capturing the long-term demand trend were considered. The trend in QR model was determined by the fitted models. In BSTS the local level trend was assumed to follow a random walk and included in the model as one of the model components, while the trend in GAMMQV was derived from the data and its evolution was forecasted using QR,

- a recommended forecasting model was developed from the QR, BSTS and GAMMQV models by averaging their forecasts at the 50 th percentiles of the demand distributions using QR. The three models considered are from different modelling approaches, QR is the frequentist modelling approach, BSTS Bayesian approach and the GAMM based model (GAMMQV) is a semi-parametric model. This averaging approach helps to avoid choosing one model over other considered models which could be equally as good,

- the electricity demand time lags in time series analysis are important because values within the time series tend to correlate with previous values in the series, and this is called autocorrelation. Using regression based techniques which exclude missing data, like BSTS and QR, it is a challenge to not only include the time lags but also to determine the number of time lags to be included in the long-term forecasting model. The number of time lags and time lags to be included were determined by using the autocorrelation plot in Figure 1.4. Data at every 24 th point are highly correlated as shown in Figure 1.4. At 168 th data point, the correlation cycle stops and the new cycle starts, hence the number of time lags to be included in the model is the number of data points at every 24 th point in the cycle (i.e. if $n = 24$ then the time lags are at 24, 48, 72, 96, 120, 144 and 168). Each cycle in our demand for electricity data represents a week. With one time lag, only one future value can be forecasted, while with two time lags only two future values can be forecasted. Therefore, with n time lags n future values can be forecasted. If the first lag is n then the next 6 lags are ($n+24$, $n+48$, $n+72$, $n+96$, $n+120$ and $n+144$). Any week in the series can be chosen depending on the number of future values to be forecasted. The time lags in this study were chosen in such a way that the leap years coincided. To be able to forecast demand for electricity up to year 2023 given our demand for electricity data, the first time lag must be $n = 70128$ and the next six time lags are (70152, 70176, 70200, 70224, 70248, 70272). This approach makes it possible to include time lags and to determine the number of time lags to be included in the long-term forecasting model, and such an approach has not yet been documented in the literature.

1.8 Data

The hourly electricity demand data for South Africa for the period 1997–2015 was provided by Eskom. Figure 1.1 depicts the time series of the actual hourly electricity demand over the 1997–2015 year period. During this period, the highest electricity demand reached was 36 826 kW in 2011 at hour 18h00, whereas the minimum was 13 533 kW in 1998.

The historical electricity demand data from 1997 to 2015 indicated that demand in South Africa increased steadily between 1997 and 2007 (Figure 1.1). During this period, South Africa experienced accelerated economic growth and a large number of new households were being connected to the electricity grid, as government wanted to make electricity accessible to all South Africans. The electricity demand from Eskom stabilised between 2007 and 2012 and started declining in the latest four years until 2015 (Figure 1.1). The decline in electricity demand from Eskom could partly be attributed to the shrinking South African economic growth between 2007 and 2015, the growth in renewable sources of electricity and the steep increases in electricity tariffs in the recent past. The National Energy Regulator of South Africa (NERSA) approved several sharp increases in annual tariffs resulting in electricity prices more than doubled in real terms rising by a cumulative 114% between 2008 and 2013 (Deloitte, 2017). According to Deloitte (2017) a 1% increases in the price of electricity is expected to result in a 0.2% to 0.5% decrease in electricity demand.

After carrying out a number of transformations of hourly electricity demand data, from the Box and Cox (1964) family, logarithm was found to be the best fit to the South African electricity demand data. The logarithmic transformation is convenient for turning a highly skewed variable into one that is more approximately normal (Benoit, 2011). Therefore, the logarithmic hourly electricity demand data was modelled.

The typical South African electricity demand profile has two daily peaks, especially noticeable in winter, namely a morning demand peak at around 06h00 and an afternoon demand peak at around 18h00. The daily profile of electricity demand in South Africa takes different shapes and the noticeable ones are given in Figure 1.2. The ideal forecasting model would accurately capture the different shapes of the daily profiles in Figure 1.2. The shape of the daily profiles in South Africa cannot be determined by the month, week or season in which they fall because within a month, week or season, the daily profiles take on different shapes. The nonlinear relationships between electricity demand and some drivers of electricity demand motivated the use of a non-parametric modelling approach, among other techniques. For example, Figure 5.2 (a) provides an example of the nonlinear relationship between electricity demand and "day of the week". The demand in this study is forecasted at every hour and it is important that the forecasted hourly demand in the day captures the

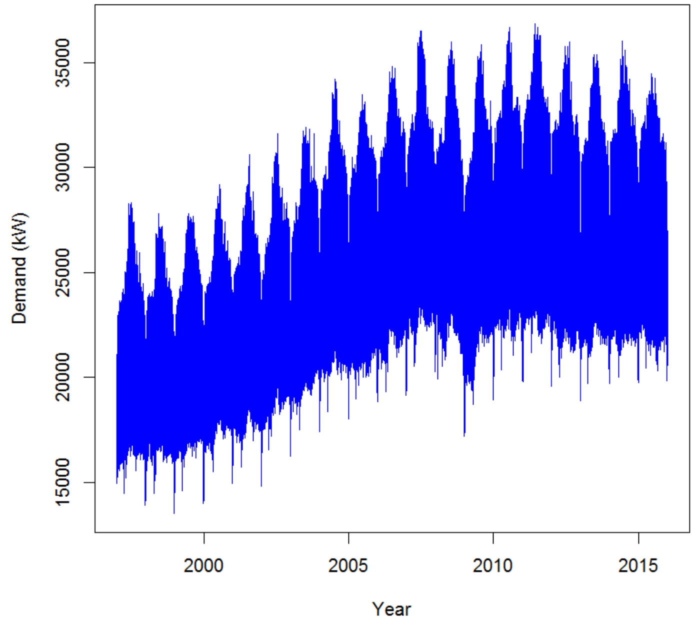


Figure 1.1: Electricity demand between 1997 and 2015 in South Africa

daily profiles accurately.

The distribution of hourly electricity demand over the year in South Africa is bimodal as shown in Figure 1.3. It is important for the considered models to forecast the density functions of electricity demand of each year in the forecast horizon accurately because the study intends to forecast the long-term electricity demand using density functions representing the distribution of the hourly electricity demand over the year. The statistical techniques that assume normality might not be ideal for the South African electricity demand data unless the proper transformation of the electricity demand variable is made. The QR model which does not make any assumption about the distribution of the electricity demand is attractive for modelling the South African electricity demand data. The GAMM model that captures the nonlinear relationship between the electricity demand and some of it's drivers is also attractive. The BSTS, which is a Bayesian approach, is a logical approach for probabilistic modelling. Figure 1.3 represents the typical density function of hourly electricity demand over the year representing the distribution of demand intended to be forecasted from 2013 to 2023. The forecasted density function of electricity demand over the year is generated from the forecasted hourly electricity demand.

The hourly electricity demand is forecasted from 2013 to 2023. Data on actual demand from 2013 to 2015 were withheld in order to validate the models. The electricity demand from 2007 to 2012 are used as the training data, whereas the forecasts from 2013 to 2023 are out of sample forecasts. According to Pierrot and Goude (2011) exogenous variables that are found to affect electricity demand significantly in the literature can be broadly classified into:

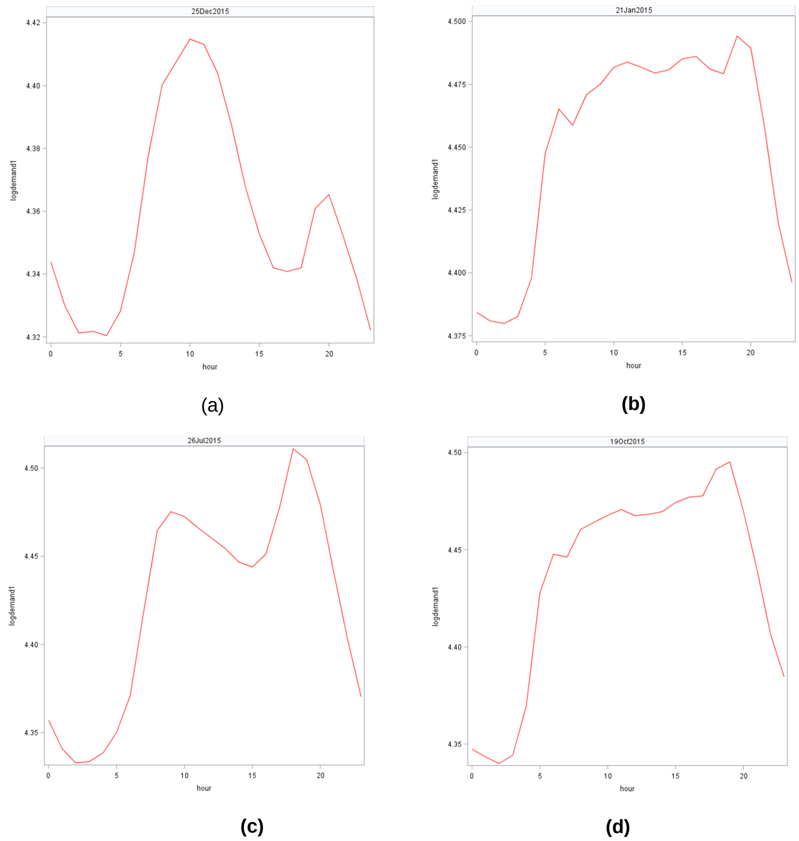


Figure 1.2: Different daily profiles of electricity demand in South African data

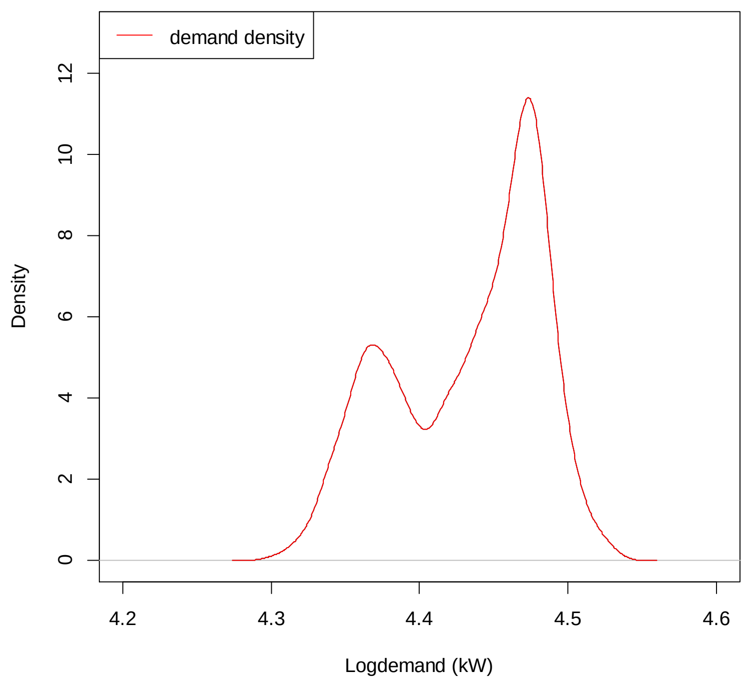


Figure 1.3: The typical density function of electricity demand in South Africa over a year

- social variables e.g. year cycles (public holidays), weekly cycles (difference between days of the week), daily cycle (night and day) or within-daily cycles (difference between hours of the day),
- natural variables e.g. temperature, cloud cover and day length, and
- economical variables e.g. gross domestic product (GDP).

From the economic point of view, according to Deloitte (2017) the electricity demand effects can be decomposed into *activity*, *structural* and *efficiency or intensity* effects. The *activity* effect relates to the change in demand emanating from changes in economic activity. The *structural* effect relates to changes in demand as the result of changes in economic activity mix within the economy and the *intensity* effect is the changes in demand as result of changes in energy use per unit of sectoral activity. Unfortunately, economic variables have not been used in this study because they are not available at the hourly level like the data on electricity demand. If temperature were to be used in this study, then the long-term future hourly temperature would need to be forecasted or simulated since models used in this study are regression based. The simulated or forecasted temperature would bring its own uncertainties.

Various time-related variables are used as covariates (as listed in Table 1.1 and 1.2), namely: day, public holidays, months, weekends, December break and seasons. The lagged demand variables are included in the model where applicable to test whether suspected lagged demand effects from the high degree of diurnal activity in electricity usage are significant, that is, whether South African consumers of electricity typically exhibit consistent daily patterns of usage as in Farland (2014). For example, in the afternoon when people return from work, around 18h00, they start cooking, watch television and take bathe and at this time the household electricity demand is expected to increase. The harmonic terms are used, where applicable, to capture the cycles inherent in the electricity demand data.

The correlation between the hourly electricity demand data is depicted in Figure 1.4. The highest correlation is at every 24 th data point indicating that electricity demand at 24 hours apart have the highest correlation.

Table 1.1: Dichotomous independent variables considered

Variable	Created variables
Newyear	1 if day is 01 Jan ; 0 otherwise
Easter	1 if day is Easter ; 0 otherwise
Humanrights	1 if day is 21 Mar ; 0 otherwise
FreedomDay	1 if day is 27 Apr ; 0 otherwise
WorkersDay	1 if day is 01 May ; 0 otherwise
YouthDay	1 if day is 16 Jun ; 0 otherwise
HeritageDay	1 if day is 24 Sep ; 0 otherwise
ReconciliationDay	1 if day is 16 Dec ; 0 otherwise
ChristmasDay	1 if day is 25 Dec ; 0 otherwise
GoodwillDay	1 if day is 26 Dec ; 0 otherwise
Month1	1 if Month is Jan ; 0 otherwise
Month2	1 if Month is Feb ; 0 otherwise
Month3	1 if Month is Mar ; 0 otherwise
Month4	1 if Month is Apr ; 0 otherwise
Month5	1 if Month is May ; 0 otherwise
Month6	1 if Month is Jun ; 0 otherwise
Month7	1 if Month is Jul ; 0 otherwise
Month8	1 if Month is Aug ; 0 otherwise
Month9	1 if Month is Sep ; 0 otherwise
Month10	1 if Month is Oct ; 0 otherwise
Month11	1 if Month is Nov ; 0 otherwise
Month12	1 if Month is Dec ; 0 otherwise
Sun	1 if day is Sun ; 0 otherwise
Mon	1 if day is Mon ; 0 otherwise
Tues	1 if day is Tue ; 0 otherwise
Wed	1 if day is Wed ; 0 otherwise
Thurs	1 if day is Thu ; 0 otherwise
Fri	1 if day is Fri ; 0 otherwise
Sat	1 if day is Sat ; 0 otherwise
Fribtwn	1 if holiday pre-Fri ; 0 otherwise
Monbtwn	1 if Mon pre-holiday ; 0 otherwise
Decclosure	1 if period 16 Dec-01 Jan ; 0 otherwise
Lngwknd	1 if day falls in Long weekend; 0 otherwise
WinterschoolholiDay	1 if school-closure Jun/Jul ; 0 otherwise

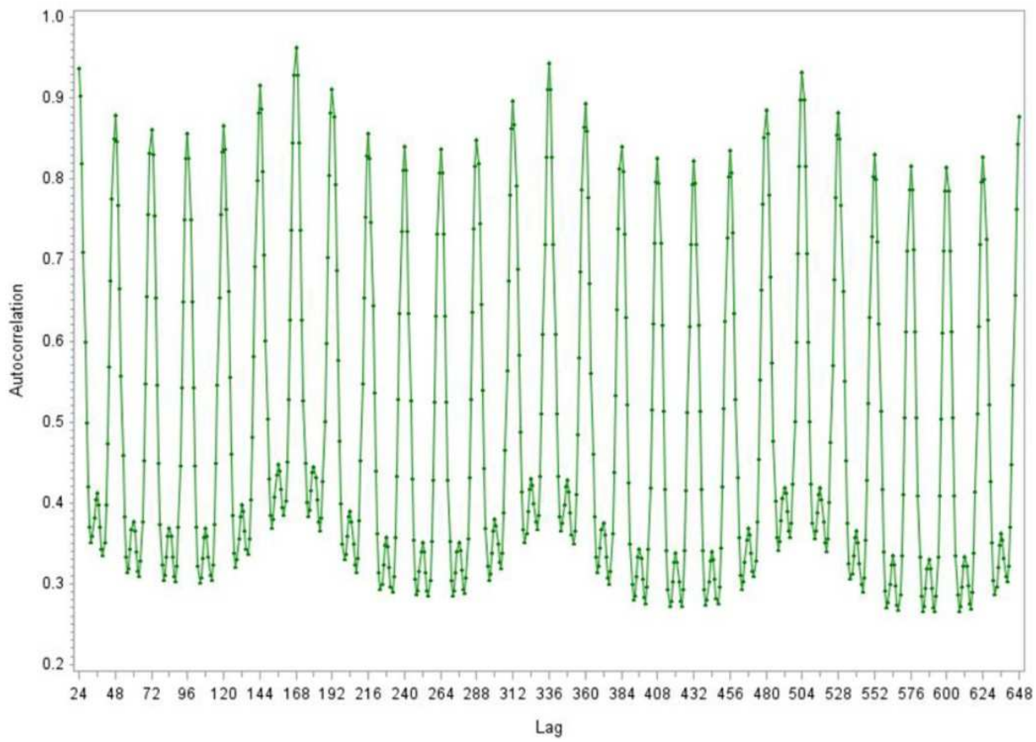


Figure 1.4: Autocorrelation of hourly electricity demand

1.9 The work done on forecasting load profile by the CSIR team

This study builds on previous load profile forecasting work done by a team from the CSIR. Before the inception of this study, univariate structural time series (USTS), seasonal autoregressive integrated moving average (SARIMA) and multilevel models were investigated for forecasting the hourly electricity demand for South Africa by the CSIR team. The multilevel model is in the family of hierarchical models which are used to analyse longitudinal data. Electricity demand data are collected over time and exhibit correlated hourly measurements nested within the day, day within week, week within month and month within year. The multilevel model takes the hierarchical structure that is inherent in the data into consideration. Koen et al. (2014) indicated that the calculated Intra-class Correlation Coefficient (ICC) of 0.65 confirmed that there is sufficient correlation among the hourly electricity demand within a day to justify a hierarchical modeling approach. The forecasts from the multilevel model picked up variations and some annual fluctuations, but underestimated annual peaks (Koen et al., 2014).

The univariate structural time series (USTS) was used to forecast the hourly electricity demand for 2013. The USTS model is a time series with a regression component and hence allows for exogenous variables to be included in the model. The model forecasted the 2013 hourly demand well. The USTS like other time series and regression based models is used

Table 1.2: Continuous independent models variables considered

Variable	Type of variable	Created variables
Demand	Dependent	Logdemand
Lag70128	Independent	1 st time lag
Lag70152	Independent	2 nd time lag
Lag70176	Independent	3 rd time lag
Lag70200	Independent	4 th time lag
Lag70224	Independent	5 th time lag
Lag70248	Independent	6 th time lag
lag70272	Independent	7 th time lag
Sin6	Independent	Sine term of Fourier series with Period 6
Cos6	Independent	Cosine term of Fourier series with Period 6
Sin12	Independent	Sine term of Fourier series with Period 12
Cos12	Independent	Cosine term of Fourier series with Period 12
Sin18	Independent	Sine term of Fourier series with Period 18
Cos18	Independent	Cosine term of Fourier series with Period 18
Sin24	Independent	Sine term of Fourier series with Period 24
Cos24	Independent	Cosine term of Fourier series with Period 24

to predict the mean of the demand distribution and therefore likely to under-predict or over-predict the extreme demand.

In their contract research report, the CSIR team used a SARIMA model as a benchmark for the USTS and multilevel models. The USTS outperformed the multilvel and SARIMA models in terms of accurately predicting demand over one year while multilevel outperformed SARIMA. The models considered by the CSIR team were for point forecasting and the disadvantage of point forecasts is that uncertainties around the forecasts cannot be quantified and the forecasts represent the expected values of electricity demand at each hour and they tended to under-predict some peak demands.

The CSIR team also used multiple linear regression models to forecast the total electricity consumed in the year in different South African economic sectors using scenario forecasts (Koen et al., 2017). Scenario forecasts could also be classified as another form of probability forecasts. The CSIR team forecasted the following sectorial electricity demands for a range of different underlying economic scenarios, “Commerce and manufacturing”, “Agriculture”, “Transport”, “Domestic” and “Mining” for a range of different underlying economic scenarios. In order to explain the apparent demand trend in the data, the total electricity consumed in each year from 1970 to 2017 was analysed separately for economic sector with attention given to the blue graphs representing the historical demand in Figures 1.5 and 1.6. Figure 1.5 and Figure 1.6 are taken from Koen et al. (2017), a report produced by the CSIR team for Eskom. The CSIR report showed that the total electricity consumed in the year in South Africa is mainly driven by "Commerce and manufacturing" sector as indicated by the resemblance of Figures 1.5 and 1.6 (a) represented by the blue graphs depicting the actual electricity consumption. According to Deloitte (2017) the growth in the energy-intensive "mining" and "Commerce and manufacturing" industries remains an important determinant of overall national demand. Deloitte (2017) indicated that the data compiled by the Department of Energy (DoE) on South Africa's national energy balances in 2012 shows that "mining" and "manufacturing" are responsible for 62% of the total electricity consumed in South Africa hence, the resemblance of the electricity consumption graphs of the "Commerce and manufacturing" to national consumption in Figures 1.5 and 1.6 (a).

Electricity consumption in South Africa only consistently increased in the "Agriculture" and "domestic" sectors. The demand has been declining in the last few years until 2017 in the "commerce and manufacturing" as well as in the "Mining" sector as depicted in Figure 1.6 (a) and Figure 1.6 (e) respectively. These two sectors dominate in terms of electricity consumption in South Africa hence the decline in total demand in South Africa. (Deloitte, 2017) has also argued that the South African economy is in transition away from reliance on energy-intensive mining and manufacturing sectors towards a more diverse range of services-oriented activity. The residential sector's electricity consumption sharply increased from 16% in 1993 to 23% in 2013 according to Deloitte (2017). The downward trend in hourly demand in the last four years until 2015 could be explained in the main by the shrinking "mining" and "commerce and manufacturing" sectors, the transition of the South African economy from energy-intensive sectors to less energy-intensive service oriented economy, growth of renewable sources of electricity and the less aggressive connection of households to the electricity grid compared to earlier years because in 1996 there was only 57.6% of the households on the electricity grid and the percentage went up to 84.7% in 2011 (the electricity consumption in the domestic sector stabilised after 2011).

1.10 Thesis Structure

The introduction giving the full background of the study, the significance of the study, the scope of the study, the contributions made and the data used in the study are presented in Chapter 1. The literature study in Chapter 2 discusses the modelling approaches in similar studies in the literature, starting from the very basic to the advanced models, and highlighting the limitations of each modeling approach. In Chapter 3, the probabilistic forecasting of the long-term electricity demand in South Africa using QR is presented. QR is built from a linear regression model fitted at every quantile of the demand distribution. In Chapter 4, long-term electricity demand forecasting in the presence of uncertainties using BSTS is presented. In the BSTS model, the coefficients of the drivers of electricity demand are estimated with their full distributions. Chapter 5 presents the long-term electricity demand

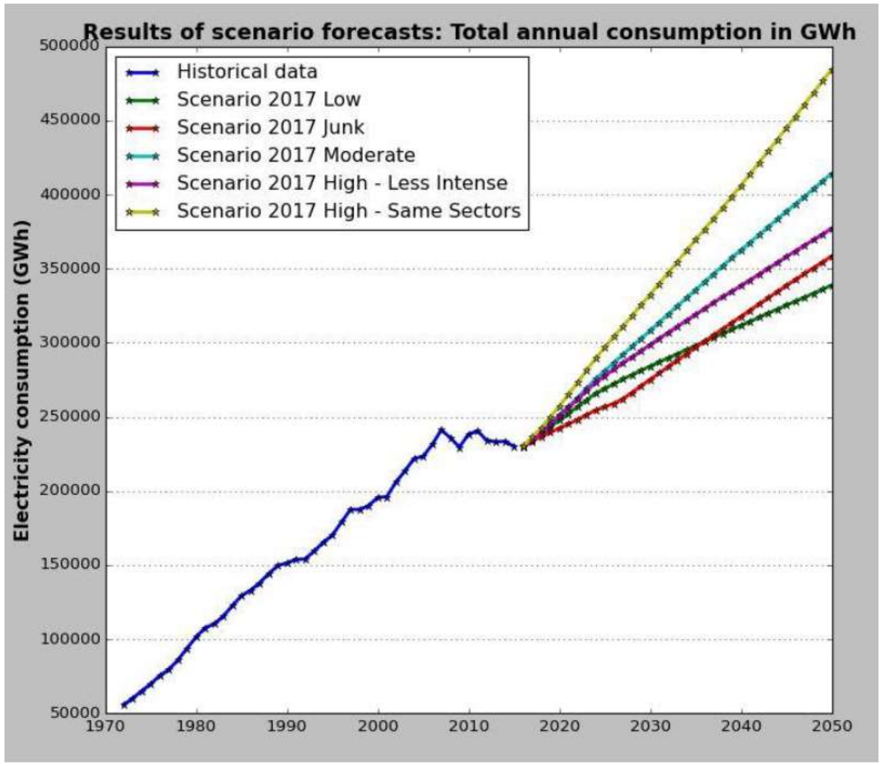


Figure 1.5: National consumption and forecasts of electricity in South Africa

forecasting in the presence of uncertainties using the GAMMQV model. The GAMM model is discussed as a semi-parametric model and this model is extended to GAMMQV. Chapter 6 presents the synthesis in which the main concepts in the study are discussed, the comparisons of the models considered in this study are made, and the averaging of the three models considered in this study using QR is discussed. Upon receiving the additional data in December 2018 for 2016, 2017 and up to the end of June 2018 after results were published in two Journals and in the conference proceedings, the models were also assessed using the additional data and this assessment is presented in Chapter 6. The summary of findings, conclusions and recommendations are presented in Chapter 7. Conclusions are also presented at the end of the following chapters, Chapter 2, Chapter 3, Chapter 4, Chapter 5 and Chapter 6. The literature study in Chapter 2 gives an account of some work done on electricity demand forecasting with emphasis on three models used in this study, and this follows in the next section.

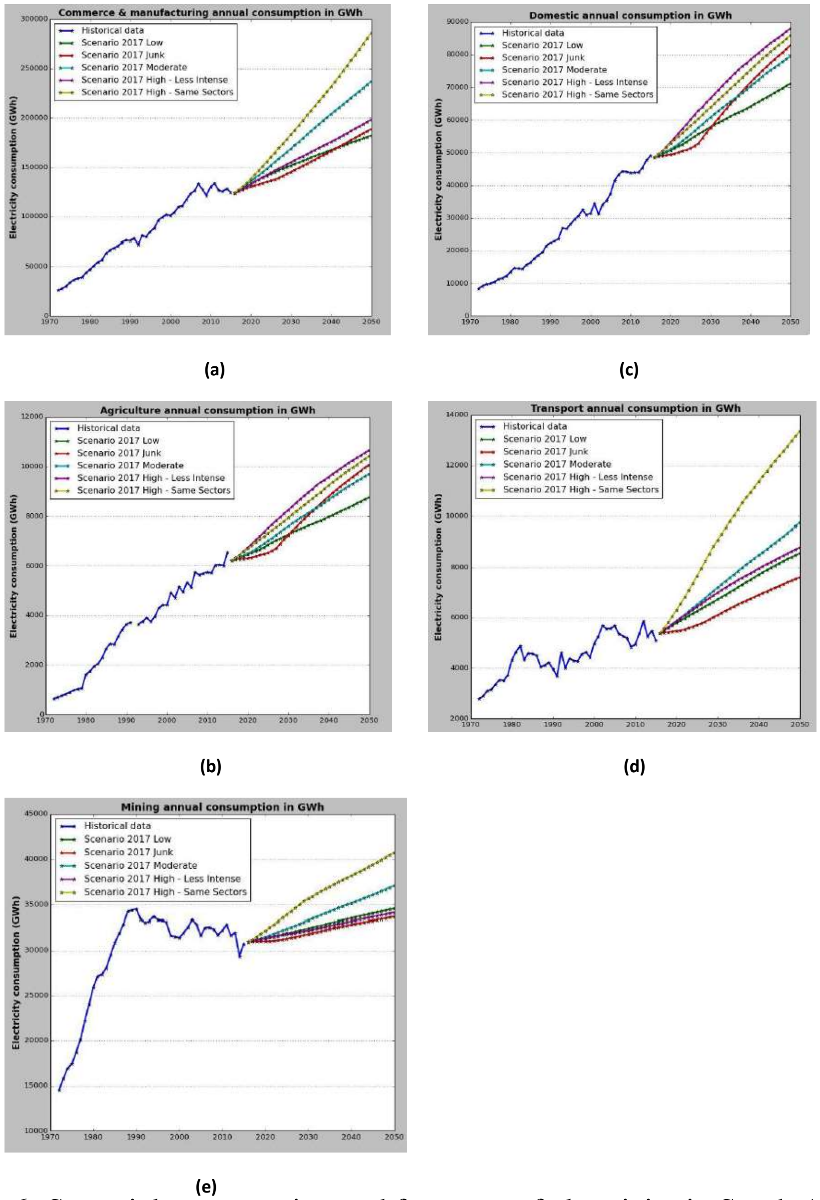


Figure 1.6: Sectorial consumption and forecasts of electricity in South African

Chapter 2

Literature Review

2.1 Background

Any long-term electricity demand forecasts will inevitably differ from the actual electricity demand. Hong and Fan (2016) argue that researchers have long been pursuing the most accurate forecasts which resulted in people hoping to find the best technique of all, but that this is a myth. In search of a better forecasting technique, some researchers such as Mori and Takahashi (2011); Sigauke (2014) and Xiong et al. (2014) combined different techniques to form hybrid models. Zhang et al. (2018) used a hybrid model based on improved empirical mode decomposition, ARIMA and wavelet neural network model optimised by fruit fly optimisation algorithm in order to improve the short-term forecasts. According to Hong and Fan (2016) most of the hybrid techniques are of minimal value for load forecasting practice. Hong and Fan (2016) pointed out that researchers and practitioners must understand that a universally best technique does not exist and they argue that it is the data and jurisdictions that determine the appropriateness of the technique to be used.

Electricity demand forecasts range from short, to medium, to long-term and they could either be in point form which involves forecasting one electricity demand value per hour in the forecast horizon representing the mean of the demand distribution, or they could be in the form of probabilistic forecasts which involves forecasting all possible electricity demand values for each hour in the forecast horizon representing the demand distributions which could be in the form of quantiles, prediction intervals, fan charts, scenarios or density functions. Probabilistic forecasts give the full distribution of future values and they make it possible to assess uncertainties around the forecasts. For long-term forecasting, it is important to use models that quantify uncertainties around the forecasts. Sigauke and Chikobvu (2011) indicated that decision-making in the electricity sector involves planning under uncertainty. Gneiting and Katzfuss (2014) argue that the inherent uncertainties in predictions imply that forecasts should be probabilistic, that they should take the form of probability distributions over future quantities or events. Despite the value of probabilistic forecasts, especially in the long-term, more attention is currently still given to point forecasts by decision-makers and planners.

Maciejowska and Weron (2015) and Weron and Misiorek (2004) indicate that forecasting models could be classified into two broad streams which are statistical (for example, multiple regression, Autoregressive (AR), Autoregressive integrated moving average (ARIMA), Autoregressive Generalised Autoregressive Conditional Heteroscedasticity (AR-Garch), Jump diffusion, factor models, regime switching models, multilevel model, mixed model, semi-parametric additive regression models) and artificial intelligence (AI) techniques (such as, artificial neural networks (ANN), fuzzy techniques and support vector machines). A critical review of literature on AI-based load forecasting is given by Hippert et al. (2001). Among researchers who used AI-based models, Song et al. (2005) used fuzzy linear regression to forecast the loads during holidays and Chen, BoJuen and Ming-Wei, Chang (2004) used support vector machine to forecast the daily peak loads of the next 31 days while gradient boosting has been applied to load forecasting by Taieb and Hyndman (2014). Bedi and Toshniwal (2019) used a deep learning based framework to forecast electricity demand by taking care of long-term historical dependencies.

The statistical methods differ with the "AI" based models in that the former forecast the current value of a variable by using mathematical combination of the previous values of that variable, and sometimes, the previous values of exogenous factors (Weron and Misiorek, 2004). They also offer insights into the relationship between the load and its driving factors. Weron and Misiorek (2004) pointed out that the conclusion made on the reviews of "AI" based forecasting systems is that much work still needs to be done before they are accepted as established forecasting techniques. "AI" based models are considered as "black-box" models.

The statistical modelling approaches to electricity demand could also be broadly classified, namely into parametric and non-parametric approaches. The number of parameters to be estimated are fixed in parametric modelling approaches whereas, in the non-parametric approach, they could increase as more data become available (e.g. in a Gaussian Process). The relationship between the response variable and covariates could be represented by nonlinear smooth functions in non-parametric approaches, such as in the generalised additive models (GAMs) or the semi-parametric additive models which are hybrids between parametric and non-parametric models.

The electricity demand data consist of a sequence of observations collected over equally spaced time periods (e.g., hourly or half-hourly). The data are typically serially correlated. The statistical modelling approaches to electricity demand could broadly be classified into the following streams:

1. approaches that consider electricity demand data as a stochastic time series. The order in which data come is important and no data values should be missing. The inherent internal structure which characterises the time series data is decomposable into

seasonal, trend, cyclic and irregular components. The components are estimated and adding all relevant component estimates could immensely improve the forecasts. For example, the seasonal autoregressive integrated moving average model (SARIMA) rather than an ARIMA model could provide improved forecasts,

2. time series with regression component approaches. These approaches consider separate equations for each intra-day period (hours). The model parameters are estimated separately for each equation. The inherent internal structure can be decomposed into seasonal, trend and irregular components. In this approach exogenous variables can be included into the model. The intra-day correlation is ignored in the process, for example the univariate structural time series (USTS). This approach can be extended to multivariate structural time series which models hours simultaneously and incorporates correlation structures,

3. regression approaches that consider separate equations for each intra-day period (hour) but estimate model parameters from the system of equations which take intra-day correlations into consideration, like seemingly unrelated regression equations (SURE). Harnessing the correlation structures could improve the forecasts. This approach poses the problem of the dimensionality curse, resulting from estimating a large number of parameters, which could result in model over-fitting if too many parameters are included in the model. Over-fitting means that the model has small in-sample residuals and huge out-of-sample errors which could result in poor out of sample forecasting. This problem could be overcome by using dimension reduction techniques, such as principal component analysis (PCA). Taylor et al. (2006) suggest PCA as a standard statistical method that can be used to reduce the dimension of a multivariate data set where variables are highly correlated,

4. approaches that consider data as arranged in hierarchies, being it geographical or spatial hierarchies, or temporal hierarchies. The hierarchical models use structured or unstructured correlations. In 2017 there was a global energy forecasting competition (GEFCOM2017) which attracted more than 300 students and professionals from over 30 countries for solving hierarchical probability load forecasting problems (Hong et al., 2019). The hierarchical probability forecasting models are new modelling approaches to electricity demand forecasting. Lai and Hong (2013); Hong (2008) and Fan et al. (2009) have used hierarchical methods to forecast electricity demand (e.g. a multilevel model),

5. simple regression based approaches which ignore the dependency of data points, like a multiple linear regression (MLR) model.

6. regression approaches that represent the relationship between electricity demand and drivers of electricity demand by functions which can be nonlinear like generalised additive models (GAMs) and the Gaussian process (GP). These represent non-parametric modeling approaches, but can also be extended to semi-parametric approaches by including a parametric component, and

7. approaches that assume that electricity demand is generated by a certain known distribution like the generalised Paredo distribution, that is used to predict extreme electricity demand via extreme value theorem (EVT) techniques. In this case the forecaster believes that the extreme electricity demand data are generated by a certain known distribution, and the data can be represented by a mathematical equation with a small set of unknown parameters.

Statistical models are attractive for forecasting electricity demand because physical interpretation may be attached to their components, and hence allow forecasters to understand their behaviour (Weron and Misiorek, 2004). The most basic statistical approach to forecasting electricity demand is exponential smoothing. Exponential smoothing is a widely used univariate approach to load forecasting as shown in Alfares and Nazeeruddin (2002); Gelper et al. (2010); Taylor and McSharry (2007) and Winters (1960). The exponential smoothing model is rarely a top candidate in real-world short-term load forecasting ($STLF$) practice despite its success in some academic papers according to Hyndman and Athanasopoulos (2013) and Hyndman et al. (2008). The approach is to first model electricity demand based on previous values of the series and then use the model to predict the future electricity demand. According to Hyndman and Athanasopoulos (2013) and Hyndman et al. (2008), in these methods, weights which decrease exponentially over time are assigned to the past observations. The exponential smoothing model is written as in equation (2.1);

$$F_{t+1} = \alpha y_t + (1-\alpha)F_t \tag{2.1}$$

where F_{t+1} is the electricity demand forecasts at time $t+1$; α is the smoothing constant; y_t is the actual electricity demand at time t; F_t is the demand forecasts at time t.

Exponential smoothing techniques are ideal for short-term point forecasting. However, this approach, according to Farland (2014), neglects the relationship between load and exogenous variables such as calendar, economic and weather variables that could potentially improve the forecasts if included in the model. It has been shown in the literature by Hyndman and Fan (2010) and Sigauke (2014) that exogenous variables such as temperature are important drivers of electricity demand. These techniques are used to forecast the mean of the distribution of electricity demand and therefore are likely to under-predict some peak electricity demand. In the case where more than one forecast is required, this approach could give the forecasted values more weight than the observed values. The approach does not assume any internal structure which characterises the time series that can be decomposed into seasonal, trend, cyclic and irregular components. The estimation of the inherent structure could immensely improve the forecasts.

Stochastic time series models assume some internal structure decomposable into seasonality, trend, cyclical and irregular components. The Box and Jenkins (Box and Jenkins, 1976) approach has widely been used in the estimation of the structural components by

Barakat et al. (1992); Huang (1997); Juberias et al. (1999); Liu et al. (1996) and Zhao et al. (1997). The stochastic nature of electricity demand as a function of time has frequently been modelled with seasonal autoregressive integrated moving average (SARIMA) and the state space models (Taylor et al., 2006). Due to the level of development of the SARIMA models, they are often used as sophisticated benchmarks for alternative proposed models according to Taylor et al. (2006) and Soares and Medeiros (2005).

Time series models were developed by Box and Jenkins (1976). Muñoz et al. (2010) indicated that the SARIMA model developed by Box and Jenkins (1976) is the most used in short-term forecasting of the electricity load to date. Generally, the SARIMA model can be written as in equation (2.2);

$$\phi_p(L)\Phi_P(L^s)\nabla^d{\nabla_s}^D x_t = \theta_q(L)\Theta_Q(L^s)\varepsilon_t \tag{2.2}$$

where x_t is the electricity load at time t; ε_t is a white noise process at time t, s is the seasonal length, L is the lag operator, $\nabla^d = (1-L)^d$ and ${\nabla_s}^D = (1-L^s)^D$ are nonseasonal and seasonal difference operators of order d and D respectively, and ϕ_p, Φ_P, θ_q and Θ_Q are polynomial functions of order p , P, q and Q respectively. Taylor (2010) has shown that equation (2.3) could be rewritten as single seasonal autoregressive moving average (SARMA) model as given in equation (2.3).

$$\phi_p(L)\Phi_p(L^s)(x_t - a - bt) = \theta_q(L)\Theta_Q(L^s)\varepsilon_t \tag{2.3}$$

where a and b are constant term and the coefficient of the linear deterministic trend term. The other terms of equation (2.3) are defined as in equation (2.2). Taylor and McSharry (2007) and Taylor (2010) further proposed a SARMA model which is able to capture the intraday and intraweek seasonal cycle in the electricity data and the model is given in equation (2.4). Taylor (2010) called this model a double seasonal ARMA (DSARMA).

$$\phi_p(L)\Phi_{P_1}(L^{s_1})\Omega_{P_2}(L^{s_2})(x_t - a - bt) = \theta_q(L)\Theta_{Q_1}(L^{s_1})\Psi_{Q_2}(L^{s_2})\varepsilon_t, \tag{2.4}$$

where $\Phi_{(P_1)}$ and Θ_{Q_1} are the polynomial functions that model the intra-day seasonality while Ω_{P_2} and $\Psi_{(Q_2)}$ model the intra-week seasonality. The DSARMA model was extended to a Triple SARMA (TSARMA) model by Taylor (2010) in order to model the intra-year seasonality in electricity data and this model is given in equation (2.5)

$$\phi_p(L)\Phi_{P_1}(L^{s_1})\Omega_{P_2}(L^{s_2})\Gamma_{P_3}(L^{s_3})(x_t - a - bt) = \theta_q(L)\Theta_{Q_1}(L^{s_1})\Psi_{Q_2}(L^{s_2})\Lambda_{Q_3}(L^{s_3})\varepsilon_t \tag{2.5}$$

Taylor (2010) indicates that Γ_{P_3} and Λ_{Q_3} represents lag polynomial functions for modelling the intra-year seasonality. The other terms of equation (2.3) are defined in equations (2.2), (2.3) and (2.4). They are generally good for short-term point forecasting. These time series models strictly use previous values of the same time series to forecast the future expected demand for electricity and they are therefore likely to under-predict peak demand for electricity. The variance of the disturbance term is assumed to be constant in the SARIMA models.

Mostly, electricity data exhibit not only non-constant mean and variance but they also exhibit multiple seasonality corresponding to daily, weekly and yearly periodicity. The assumption of homoscedasticity in time series models including SARIMA is inappropriate, hence the autoregressive conditional heteroscedasticity (ARCH) and the generalised autoregressive conditional heteroscedasticity (GARCH) models were developed to capture the non-constant variance and to accommodate the possibility of serial correlation in the series (Sigauke, 2014). However, most stochastic time series models do not use exogenous variables such as weather, economic and calendar variables which could improve the forecasts.

Univariate structural time series (USTS) and regression seasonal autoregressive integrated moving average (SARIMAX) are time series models with regression components. The USTS does not only decompose a time series into trend, seasonal, cycle and irregular components, but it can also include calendar variables, economic and weather variables in the model. Harvey (1990) used USTS models to forecast short-term hourly electricity demand. According to Scott and Varian (2014) it is formulated as a state space model, hence there are two parts to a USTS model, namely, the observation and transition equations. The observation equation is represented by equation (2.6);

$$y_t = Z_t^T \alpha_t + \varepsilon_t \quad \varepsilon_t \sim N(0, H_t) \quad t = 1, 2, 3, \ldots, 24 \tag{2.6}$$

where y_t is the demand for electricity at hour t, α_t is a vector of latent variables (a set of unknown parameters); H_t and Z_t are structural parameters where H_t is a constant diagonal matrix with diagonal elements ${\sigma_\varepsilon}^2$ and Z_t is the observation vector. The transition equation is represented by equation (2.7);

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \quad \eta_t \sim N(0, Q_t) \tag{2.7}$$

where T_t, R_t and Q_t are structural parameters and η_t is a lower dimension than α_t. Q_t is a constant diagonal matrix with diagonal elements ${\sigma_u}^2$, ${\sigma_\omega}^2$ and ${\sigma_\varphi}^2$. $\eta_t = (V_t, \varphi_t, \omega_t)$ are the independent components of the Gaussian random noise. The USTS model with a regression component can be written as in equation (2.8);

$$y_t = \mu_t + \gamma_t + \beta^T X_t + \varepsilon_t \tag{2.8}$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + \varphi_t \tag{2.9}$$

$$\delta_t = \delta_{t-1} + v_t \tag{2.10}$$

$$\gamma_t = -\sum_{s=1}^{s-1} \gamma_{t-s} + \omega_t \tag{2.11}$$

where y_t is the response series; μ_t the trend; γ_t is the seasonal component; ε_t is the irregular component, $\beta^T X_t$ the regression component, the local level term which defines how the latent state evolves over time and it is referred to as an unobserved trend; level $=\mu_t$ plus slope δ_t, and seasonal; $S-1$ dummy variables with time varying coefficients that are expected to sum up to zero.

A good USTS model is obtained when a separate parametric intraday model for each hourly period is fitted and the model parameters are estimated separately. The intraday correlation is ignored in the process. Hyndman and Fan (2010) argue that better estimates can be obtained if each hourly period is treated separately since the demand patterns change throughout the day. Individual models for each time of the day have been applied by Fan and Chen (2006); Fay et al. (2003) and Ramanathan et al. (1997). According to the recent unpublished CSIR study, the USTS model outperformed SARIMA and Multilevel models in forecasting the long-term hourly demand for electricity in South Africa. Although, this model is expected to give better forecasts than most of the point forecasting models, it is still likely to underestimate the peak demand for electricity as it predicts the mean of the demand for electricity distribution. This model can be extended to multivariate structural time series or to a Bayesian structural time series (BSTS) which predicts the full distribution of the hourly electricity demand.

According to Brodersen et al. (2015) and Scott and Varian (2014) BSTS is formulated as a state space model. The observation equation is given in equation (2.6) and transition is represented by equation (2.7).

The natural way to represent sparsity in Bayesian modelling is through a spike and slab prior on the regression coefficients (Scott and Varian, 2014). If τ denotes a vector of ones and zeros where a value of one indicates that the variable is selected into the model (nonzero coefficient) and *zero* indicates that the variable is not selected into the model, and if β_τ is the subset of elements of β where, $\tau_k = 1$ if $\beta_k \neq 0$ and $\tau_k = 0$ if $\beta_k = 0$, then the spike and slab prior can be written as in equation (2.12);

$$p(\tau, \beta, \sigma_\varepsilon^2) = p(\tau)p(\sigma_\varepsilon^2|\tau)p(\beta_\tau|\tau, \sigma_\varepsilon^2) \tag{2.12}$$

where the marginal distribution $p(\tau)$ is the spike. The probability of choosing a given variable is assumed to follow a Bernoulli distribution given in equation (2.13);

$$p(\tau) = \prod_{j=1}^{J} \pi_j^{\tau_j} (1-\pi_j)^{1-\tau_j} \tag{2.13}$$

where π_j is the prior probability of regressor j being included in the model. For practical consideration all π_j are of the same value π. If there are p non-zero predictors then $\pi = p/k$, where k is the dimension of x_t. For specific values of j, some variables can be forced to be included or excluded by setting $\pi_j = 1$ for inclusion and $\pi_j = 0$ for exclusion. For the *slab* part, a normal prior is used for β which leads to inverse gamma prior for σ_ε^2 and it is given by $\frac{1}{\sigma_\varepsilon^2} \sim G(\frac{v_\varepsilon}{2}, \frac{s_\varepsilon}{2})$ where v_ε is the prior sample size (the shrinkage) which can be thought of as a prior sample size for learning the variance parameter of each coefficient, and s_ε is the prior sum of squares in equation (2.14):

$$\frac{s_\varepsilon}{v_\varepsilon} = (1-R^2)s_y^2 \tag{2.14}$$

where s_y^2 is the marginal standard deviation of the response and R^2 is the expected variation explained by the regression model. The BSTS software supplies the default values of $R^2 = 0.5$, $v_\varepsilon = 0.01$ and $\pi_j = 0.5$. The default spike and slab priors are used as the priors for the regression component of the model. The full Bayes approach is used with a minor violation as s_y^2 is data-determined. The detailed formulation of BSTS is given by Brodersen et al. (2015) and Scott and Varian (2014).

Stochastic time series and exponential smoothing models do not accommodate exogenous variables, with the exception of the time series models with a regression component such as the regression seasonal autoregressive integrated moving average (SARIMAX), USTS and BSTS models. The regression based modelling approaches can incorporate exogenous drivers of electricity demand like economic, weather and calendar variables.

Regression based models are used to explain the relationships between the demand for electricity and drivers of the demand for electricity. Regression based methods have been extensively used in electricity load forecasting. Alfares and Nazeeruddin (2002) indicate that the regression based load forecasting models have been used to analyse the statistical relationship between total load, weather conditions and time-of-year effects. A multiple linear regression (MLR) model with m independent variables can be written as in equation (2.15);

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_m x_{m,i} + \varepsilon_i, \tag{2.15}$$

where, for example, $x_{2,i}$ is an observation of explanatory variable two at time period i and β_2 is a regression coefficient associated with explanatory variable two. The population parameters $\beta_0, \beta_1, \ldots, \beta_m$ are estimated by minimising the sum of squared errors. Papalexopoulos and Hesterberg (1990) proposed a regression based approach to STLF while Hong et al. (2014) developed a linear regression model for long-term load forecasting (LTLF). These models are classified as the ordinary least square (OLS) regression models.

The OLS based regression model must satisfy the following stringent assumptions: linearity, no multicollinearity, constant variance (homoscedasticity) and strict exogeneity of explanatory variables, that is, $E(\varepsilon_i | x_1, \ldots, x_n) = 0$, according to Koenker and Bassett Jr (1982). The electricity demand data are collected over time and therefore it is considered as time series data, the observations are not independent and the variance is heteroscedastic. Therefore, some underlying assumptions of the OLS regressions would be violated especially the linearity and independence of errors. A quantile regression (QR) model does not require any distribution assumptions regarding the population and can estimate the parameters non-parametrically. For a random variable Y (an hourly electricity demand) with probability distribution function in equation (2.16);

$$F(y) = \text{Prob}(Y \leq y) \tag{2.16}$$

the τ th quantile of Y is defined as the inverse function in equation (2.17),

$$Q_Y(\tau) = \inf[y : F(y)\text{Prob}(Y \geq \tau)] \quad 0 < \tau < 1 \tag{2.17}$$

The QR was developed as an extension of ordinary least-square regression for estimating rates of change in all parts of the distribution of the response variable (Cade and Noon, 2003). QR offers a comprehensive strategy for completing the regression picture. Gibbons and Faruqui (2014) applied QR methodology in forecasting the annual peak electricity demand. QR intends to find the coefficients α_0 and α_i which minimise $\sum_i^n$ f$(y_i - (\alpha_0 + \alpha_i\, x_i))$.

A linear model for the τ th quantile is given in equation (2.18);

$$y_i = {x_i}^T \beta_\tau + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{2.18}$$

where ${x_i}^T$ is the transposed (indicated by T) design matrix (matrix of covariates), τ is the regression coefficients and the τ th quantile of ε_i is assumed to be zero. The standard QR model is given in equation (2.19);

$$Q_Y(\tau | X) = {x_i}^T \beta_\tau \tag{2.19}$$

where

$$Q_Y(\tau | X) = \inf[y : F_\tau(y | x_i) \geqslant \tau] \tag{2.20}$$

in equation (2.20) is the conditional τ th quantile of the response (y_i) given the covariate (x_i) and $Q_Y(\tau|X)$ is a non-decreasing function of τ for any given x. β_τ is the vector of parameters and is the marginal change in the quantile because of the marginal change in x_i. In the estimation of a QR model for a given quantile, we follow the ideas of Koenker (2005) and Yue and Rue (2011) who follows the standard approach of Koenker and Bassett Jr (1978) to estimate the QR model. QR minimises the tilted absolute function $\rho_\tau(.)$, which asymmetrically weighs residuals from the model to a degree that depends upon τ.

$$\rho_\tau(\varepsilon) = \begin{cases} (1-\tau)\varepsilon & \text{if } \varepsilon < 0 \quad 0 < \tau < 1 \\ \tau\varepsilon & \text{if } \varepsilon \geqslant 0 \end{cases}$$

$\rho_\tau(\varepsilon)$ is a continuous piecewise linear function and non-differentiable at $\varepsilon = 0$ but differentiable everywhere else (has directional derivative in all directions) (Yue and Rue, 2011). This check-function ensures that all ρ_τ are positive and the scale is based on the probability τ. A linear model is estimated by solving equation (2.21):

$$\hat{\beta}_\tau = \operatorname{argmin}_{\beta_\tau \in R^p} \sum_{i}^{n} \rho_\tau(y_i - x_i{}^T \beta_\tau) \tag{2.21}$$

This concept is extendable to any quantile, such as the 75 th, 90 th and 99 th percentile. The QR estimator for β_τ at quantile τ minimises the objective function in equation (2.22):

$$Q_Y(\beta_\tau) = \min_{\beta_\tau \in R^p} \sum_{i:y_i \geq x_i{}^T \beta_\tau}^{n} \tau|y_i - x_i{}^T \beta_\tau| + \sum_{i:y_i < x_i{}^T \beta_\tau}^{n} (1-\tau)|y_i - x_i^T \beta_\tau| \tag{2.22}$$

This is a non-differentiable function and there is no closed-form solution for β; instead these parameters can be found using a linear programming algorithm (Gibbons and Faruqui, 2014). The minimisation is done for each subsection defined by ρ_τ, where the estimate of the τ th quantile function is achieved with the parametric function $x_i^T \beta_\tau$. The detailed formulation of quantile regression is available in Koenker and Bassett Jr (1978). The main features that characterise QR and differentiate it from other regression methods are:

- QR computes several different regression curves corresponding to various percentile points of the distribution and thus provides a more complete picture of the relationship between the response variable and the covariates,
- heteroscedasticity can be detected and, if the data are heteroscedastic, median regression estimators can be used instead of mean regression estimators, and
- median regression is more robust to outliers than other regression methods that use mean estimators, and it is semi-parametric, therefore it avoids assumptions about the parametric distribution of the error process.

Nonlinear relationships can be accommodated in linear regression models, but the functional form of the relationship must be determined beforehand. An alternative approach that introduces flexibility regarding the specification of a functional form is non-parametric regression (Farland, 2014) and (Ruppert et al., 2003). Hyndman and Fan (2010) proposed a semi-parametric additive model in the regression framework, but with some nonlinear relationships and serially correlated errors. The proposed models allow nonlinear and non-parametric terms using the framework of additive models.

The semi-parametric additive model that was developed by Hyndman and Fan (2010) is given in equation (2.23):

$$\ln(y_{t,p}) = h_p(t) + f_p(w_{1,t}, w_{2,t}) + \sum_{j=1}^{J} c_j z_{j,t} + n_t \tag{2.23}$$

where $y_{t,p}$ is the hourly demand for electricity at time t during period p ($p = 1, 2, 3, \ldots, 48$); $h_p(t)$ models all calendar effects, $f_p(w_{1,t}, w_{2,t})$ models all temperature effects; $w_{1,t}$ is the vector of recent temperature taken in town 1 while $w_{2,t}$ is the vector of recent temperature taken in town 2; $z_{j,t}$ is the economic variables at time t; its impact on demand is measured by c_j; n_t is the model error at time t. This model was used for the long-term probabilistic forecasting of peak electricity demand. Although, the model ignores the intraday correlation inherent in the time series, it produced accurate long-term forecasts (Hyndman and Fan, 2010).

Ruppert et al. (2003) assert that semi-parametric regression models based on penalized splines can be couched in the mixed model framework. The mixed model representation of penalized splines allows for a seamless fusion between parametric mixed models and smoothing and this model is referred to as a semi-parametric mixed model (Ruppert et al., 2003). The more contemporary application of mixed models is the analysis of longitudinal datasets (Ruppert et al., 2003). A GAM can be written as a nonparametric or as a semi-parametric model. The GAM model allows the flexible specification of the dependence of the electricity demand on the drivers of electricity demand by specifying the model only in terms of smooth functions, rather than by a parametric relationship. This flexibility needs the smooth functions to be represented in some way and to choose how smooth they should be.

GAM is classified as a statistical learning technique in electricity demand forecasting (Sigauke, 2017). A generalised additive mixed model (GAMM) is an extension of the GAM model. GAMM is an additive and a functional modelling technique where the impact of predictive variables is not only captured through the parametric effects but also through smooth functions which could be nonlinear. The goal of applying GAMM in this study is to model the electricity demand using its drivers, some of which are expressed in the form of parametric and others which are expressed in the form of smooth functions (splines). Smoothing splines are real functions that are piecewise-defined by polynomial functions called basis functions. The places where the polynomial pieces connect are called knots. In GAMM, penalized regression splines are used in order to regularize the smoothness of a spline. The GAMM model is generally written as in equation (2.24):

$$y_{th} = \beta_{0h} + \sum_{j=1}^{p} \beta_{hj} x_{thj} + \sum_{j=1}^{p} s_{hj}(x_{thj}) + z_{th} b_t + u_{th} \tag{2.24}$$

where y_{th} is an electricity demand on day t at hour h, $\beta_{hj} x_{thj}$ is the linear part of the model, β_{hj} is the corresponding fixed parameter, β_{0h} is the constant parameter of the linear component of the model; s_{hj} are smooth functions of covariates x_{thj}; z_{th} is a row of random effects in the model matrix; b_t is a vector of random effects coefficients and u_{th} are residuals. $t = 1,2,\ldots,n$; $h = 1,2,\ldots,24$ and $j = 1,2,\ldots,p$, and $x_{th1}, x_{th2}, \ldots, x_{thp}$ are p covariates. Each smooth function (s_{hj}) in equation (2.24) is treated as having a parametric (unpenalised) component, which can be absorbed into $x_{thj}\ \beta_{hj}$, and a random effects (penalised) component, which can be absorbed into $z_{th} b_t$. The random effects component of the smooth function has an associated Gaussian distribution assumption. To estimate the smooth function (s_{hj}), s_{hj} is required to be represented in such a way that equation (2.24) becomes linear, which is achievable by choosing the basis functions in equation (2.26). The GAMM is estimated using penalised cubic splines according to Goude et al. (2014) and Wood (2006).

$$\min_{s_{hj}} [\sum_{t=1}^{n} [y_{th} - X_t \beta - Z_{th} b_t - \sum_{j=1}^{p} s_{hj}(x_{thj})]^2 + \sum_{j=1}^{p} \lambda_j [\int (f^{''}(x))^2] \mathrm{d}x \tag{2.25}$$

According to Wood (2017), the degree of smoothness is controlled by the penalty parameter $\Lambda = \lambda_j, j = 1,2,\ldots,p$, determining the roughness of the function fitted to the data. The generalised cross-validation criterion (GCV) is used to optimise this function. The smooth function s_{hj} is the sum of basis functions, $b_{hi}(x)$ and their regression coefficients β is written as in equation (2.26):

$$s_{hj}(x) = \sum_{i=1}^{q} \beta_{hi} b_{hi}(x) \tag{2.26}$$

where q in equation (2.26) denotes the basis dimensions.

A hierarchical approach could be considered for electricity demand data as the data are recorded repeatedly over time. The hourly data is nested within the day, which is nested within a month and the month within the year. If this hierarchical structure is considered, then the correlation of electricity demand within clusters should be evaluated using the intraclass correlation coefficient (ICC). This approach is based on the premise that demand within the same cluster are more correlated to each other than demand between clusters. The hierarchical approach performed favourably well compared to a SARIMA model (Koen et al., 2014). The literature on hierarchical techniques in electricity load forecasting is very limited (Hong and Fan, 2016).

Some of the comprehensive reviews of electricity demand forecasting models which are commonly used in energy sector are given by Alfares and Nazeeruddin (2002); Hong and Fan (2016); Jebaraj and Iniyan (2006) and Suganthi and Samuel (2012).

The following researchers have done extensive work in forecasting total electricity demand as well as peak electricity demand in South Africa, Amusa et al. (2009); Koen et al. (2017); Inglesi-Lotz (2011); Koen et al. (2014); Sigauke (2014); Sigauke and Chikobvu (2011) and Ziramba (2008). All the models used to forecast the demand for electricity in South Africa to date are not for long-term probabilistic forecasting. Most of them produce point forecasts and are mostly for short-term demand (Sigauke, 2014). None of these models attempted to explore the possible intraday correlation structures which could be inherent in the time series.

2.2 Model parameter estimation

The approaches mostly used for parameter estimation in demand forecasting are maximum likelihood (MLE) and Bayesian approaches.

The MLE is a useful tool for parameter estimation. The MLE of θ is the value of θ that makes the observed data most probable or most likely. In this approach, the data are considered random and the model parameters fixed. If the random variables $x_1, x_2, \ldots, x_n$ are assumed to be identically and independently distributed, their joint density is the product of their marginal densities and their likelihood is given by equation (2.27):

$$\ell(\theta, x) = \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad \theta = \theta_1, \theta_2, \ldots, \theta_p \quad \text{and} \quad X = x_1, x_2, \ldots, x_n \tag{2.27}$$

Instead of maximising the likelihood function, it is easier to maximise its natural logarithm function which is given by equation (2.28);

$$\log \ell(\theta, x) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2.28}$$

The values of the model parameters are the ones that minimise the errors, and the MLE does not provide for uncertainties around the estimated parameters.

The Bayesian approach to parameter estimation accounts for uncertainties because the parameters are estimated with their full distributions. In this approach, the data are considered fixed and the model parameters random. There are two approaches to Bayesian modelling, namely Full Bayesian and Empirical Bayes. In Full Bayesian modelling, the prior parameters or the distribution of the prior parameters are assumed known, while in the Empirical Bayes modelling the observed data are used to estimate the prior parameters and then proceeds as though the priors were known. In Bayesian analysis, the prior distribution represents what is known about the parameter before observing the data. The prior information can be based on past experience (expert information) gained from previous, related studies. If such expert information is not available, then non-informative priors are used. The next step is to collect data and form the likelihood function, and then the prior is combined with the likelihood function to form the posterior distribution which is given by equation (2.29):

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\int_{\theta} \pi(x|\theta)\pi(\theta)} \mathrm{d}\theta \tag{2.29}$$

where θ is the parameters to be estimated; x are the data; $\pi(x|\theta)$ is the likelihood function, $\pi(\theta)$ is the prior distribution and $\int_{\theta} \pi(x|\theta)\pi(\theta)\mathrm{d}\theta$ is a normalising constant, therefore the posterior distribution in (2.28) can be written as:

$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$.

The posterior distribution summarises all the available information about the parameters after observing the data. The Markov Chain Monte Carlo (MCMC) is used to sample from the posterior distribution. The posterior mean is estimated by equation (2.30):

$$E(\theta|x) = \int_{\theta} \theta\pi(\theta|x)\mathrm{d}\theta \tag{2.30}$$

which serves as the point estimate (most likely value) of parameter θ given the data. The posterior predictive distribution could be used to predict the future values of x given by equation (2.31);

$$\pi(x_{n+1}|x) = \int_{\theta} \pi(x|\theta)\pi(\theta|x)\mathrm{d}\theta \tag{2.31}$$

A Bayesian model is a natural probabilistic approach to modelling and it has been applied successfully by Cottet and Smith (2003) where they use a multi-equation approach to estimate and forecast intraday electricity load. In their approach, they used a first-order vector autoregression (VAR) to model the errors and a Bayesian model selection methodology to explore the intraday correlation structure. Sigauke (2014), uses Bayesian parameter estimation in forecasting the short-term extreme electricity demand in South Africa.

Researchers like Hyndman and Fan (2010) and Ramanathan et al. (1997) considered methods that fit separate models to the data from each half-hourly or hourly period in order to forecast future electricity demand. Their argument is that the electricity demand patterns change throughout the day, and therefore, better estimates can be obtained if each half-hourly or hourly period is treated separately. Soares and Medeiros (2005) argue that considering separate models for each hour of the day avoids modelling complicated intraday patterns in the hourly load, which is commonly called the load profile. The interday correlation in this approach is often catered for by using a Fourier series.

2.3 Concluding remarks on literature study

Generally, any stochastic time series model like SARIMA would perform poorly in forecasting the long-term electricity demand in South Africa because the data have "structural breaks" (i.e. points at which the long-term trend changes direction) which means that, from such a point, the fitted trend would continue the historical trajectory and consequently deviates from the actual future trend resulting in forecasts vastly deviating from the actual electricity demand in the long-term. The trend becomes important in this modelling framework because the forecasting model is built by decomposing the time series into trend, cyclic, seasonal and irregular components and if the historical trend is upwards then the predicted trend will continue an upward trajectory even after the point where demand started dropping. The stochastic time series models also do not include exogenous variables that are found in literature to affect electricity demand significantly such as temperature, calendar and economic variables.

Simple regression models would perform poorly in forecasting the long-term electricity demand because the demand data points are not independent of each other and this would result in the violation of the underlying assumptions of linearity and independence of errors underpinning the regression model. Regression model assumes that individual data points are determined solely by the underlying parameters and some random noise, so that all information in a given data point can be used towards estimating the underlying parameters, since regression models attempt to make educated guesses about the underlying parameters based on the observable data (Sóskuthy, 2017). Regression models are appropriate for modelling data with no inherent grouping, temporal or spatial structures. A regression model with random effects or an appropriate error model would be more appropriate for modelling the South African electricity demand data.

Time series with regression model component such as USTS, BSTS and SARIMAx would give better forecasts of the electricity demand compared to a stochastic time series or simple regression model because in such modelling framework, data are treated as time series which take the dependence of data points into consideration and this is added to the regression part that uses exogenous drivers of electricity demand. Time series with regression models are expected to give better short-term demand forecasts than stochastic time series models, but the forecaster should think hard about the trend if the model is used for forecasting demand in the long-term. In this modelling framework, a linear relationship between the demand for electricity and its drivers is assumed, but unfortunately in the South African electricity demand data some drivers of demand have a nonlinear relationship with demand.

Regression based models like GAMMs which take the grouping, temporal or spatial structures into consideration by adding random effects, or error models, or both in the modelling and relaxes the assumption of linearity between predictors and electricity demand would give better forecasts of electricity demand in South Africa. This represent a semi-parametric approach to modelling because both parametric and nonparametric predictor variables can be included in the model. Nonparametric predictors are represented by smooth functions.

The electricity demand model choice must be data-driven, objective-driven and must perhaps be country-specific. It should be country-specific because some electricity demand drivers affect demand differently for different countries. For example, in third world country where few people own air conditioners but the majority own electric heaters, extremely cold temperatures could substantially affect electricity demand while extremely hot temperatures would not affect electricity demand, whereas in first world country, where the majority of people own air conditioners, both extreme low and high temperatures could substantially affect demand for electricity. The forecasts should be objective-driven because if the researcher wants to quantify the uncertainties around the forecasted demand for electricity, probabilistic approaches like QR or a Bayesian approach should be used, otherwise point forecasting techniques should be considered where the mean of the electricity demand distribution is to be forecasted.

A good scientific approach to electricity demand forecasting will produce good forecasts (i.e. if the model is data-driven). The forecaster must understand the theory behind the modelling technique of choice and its limitations. The hourly demand for electricity data has a complex pattern and attracts different modelling techniques. It can attract hierarchical modelling techniques as hourly demand for electricity could be seen to be embedded within the day, day embedded within a month and a month within the year. The data could also attract stochastic time series approaches like SARIMA. It could also attract a regression based approach where hours are modelled separately and the model parameters estimated separately (ignoring the intraday correlation) or simultaneously (taking intraday correlation structures into consideration). These approaches would all have different advantages and disadvantages, but their limitations and the theory behind them must be well understood.

Some researchers argue that since the electricity demand patterns change throughout the day, better estimates can be obtained if each half-hourly or hourly period is treated separately (Hyndman and Fan, 2010). This approach ignores the intraday correlation which some researchers argue that it could improve the forecasts if taken into consideration.

There is extensive literature on short-term point forecasting of the electricity demand, while the literature on long-term probabilistic forecasting of the demand for electricity is very limited. No paper has been found in which probabilistic forecasts for the long-term electricity demand is given, using South African data or an attempt is made to forecast the demand for electricity taking intraday correlation into consideration. This study takes the view that long-term point forecasts should be discouraged because uncertainties around the forecasts can not be fully quantified and the best that can be done would be to generate confidence intervals which may be too narrow. The probabilities of exceeding certain demand levels for electricity cannot be calculated since single demand values are forecasted at each step in the forecast horizon. The contributions of Chapter 2 are:

1. due to the complexity of the hourly electricity demand data, the literature has shown that a universally best electricity demand forecasting technique is either non-existent or not yet found,

2. the literature has shown that the model choice must be data-driven, objective-driven and it should perhaps be country specific,

3. all the point forecasting models predict the mean of the electricity demand distribution and hence they inevitably under-predict the peak electricity demand, and

4. a good scientific approach to demand forecasting will generate good forecasts.

In the regression modelling approach, QR as an extension of linear regression model, is used to forecast a long-term electricity demand in South Africa using probabilistic methods and this approach is presented in Chapter 3.

Chapter 3

Quantile Regression Approach

Quantile Regression Approach[1]

3.1 Background

The QR model was developed as an extension of an OLS regression for estimating rates of change at specified percentiles of the electricity demand distribution (Cade and Noon, 2003). The OLS regression models the relationship between the drivers of electricity demand (X) and the conditional mean of electricity demand Y given $X = x$. Koenker and Bassett Jr (1978) argue that what the regression curve does is to give a summary for the averages of the distributions corresponding to the set of Xs. One could go further and compute several different regression curves corresponding to various percentage points of the distribution and thus get a more complete picture of the set (Koenker and Bassett Jr, 1978). Ordinarily, this is not done, and so OLS regression often gives a rather incomplete picture. In forecasting electricity demand, OLS regression models the mean of electricity demand.

QR offers a comprehensive strategy for completing the regression picture. It has been applied in ecology (Cade and Noon, 2003) and has been widely used in financial economics where data are volatile and extremes are important to study. Gibbons and Faruqui (2014) applied a QR model to forecast the annual peak electricity demand. Cornec (2014) proposes QR as a way of estimating the distribution of the forecasts, and uses the dispersion of the estimated quantiles for calculating an uncertainty index.

The demand for electricity is said to be at τ th quantile of the demand distribution if it is above proportion τ, and below proportion (1-τ) of the whole electricity demand data. Therefore, quantiles can be defined in terms of order statistics, that is, $y_1 \leq y_2 \leq \cdots \leq y_n$. In this study, QR is used to forecast the future hourly electricity demand at 1 st, 2 nd, ..., 99 th percentiles of the electricity demand distribution, and variabilities in the electricity demand are assessed in the process since the hourly demand is forecasted 99 times at each

[1] A modified version of this chapter has been published in South African Journal of Economic and Management Sciences (SAJEMS) in March 2018.

step in the forecast horizon. The features that characterise QR and differentiate it from other regression methods are:

- QR computes several different regression curves corresponding to the various percentile points of the distribution and thus provides a more complete picture of the relationship between the response variable and the covariates,
- heteroscedasticity can be detected and, if the data are heteroscedastic, median regression estimators can be used instead of mean regression estimators, and
- median regression is more robust to outliers than other regression methods that use mean estimators, and it is semi-parametric, therefore, avoiding assumptions about the parametric distribution of the error process.

3.2 Methodology

The QR model does not impose an assumption of normality, allowing, for example, for fat-tailed distributions, which are useful for forecasting extreme events. In electricity demand forecasting, QR can be used to model the median, the 1 st, 5 th, 10 th, 90 th, 95 th and 99 th percentiles or all quantiles, to describe the full distribution of electricity demand at each hour in the forecast horizon. Generally, QR attempts to find the coefficients α_0 and α_i of the electricity demand model which minimise $\sum_i^n \mathrm{f}(y_i - (\alpha_0 + \alpha_i\, x_i))$. QR does not require any distribution assumptions regarding electricity demand and can estimate parameters non-parametrically (Koenker and Bassett Jr, 1982). A linear model for the τ th quantile is given in equation (3.1):

$$y_i = X_i{}^T \beta_\tau + \varepsilon_i \quad i = 1,\ 2, \ldots, n \tag{3.1}$$

where $X_i{}^T$ is the transposed design matrix (matrix of the drivers of the electricity demand), β_τ is the regression coefficient and the τ th quantile of ε_i is assumed to be zero. The standard QR model is given in equation (3.2):

$$Q_Y(\tau|X_i) = X_i{}^T \beta_\tau \tag{3.2}$$

where

$$Q_Y(\tau|X_i) = \inf[y : F_\tau(y|X_i) \geq \tau] \tag{3.3}$$

and equation (3.3) is the conditional τ th quantile of the electricity demand (Y_i) given drivers of electricity demand (X_i), and $Q_Y(\tau|X_i)$ is a non-decreasing function of τ for any given driver of electricity demand (x). β is the vector of parameters and represents the marginal change in the quantile because of the marginal change in x_i.

In estimating the QR model for a given quantile, we follow the ideas of Koenker (2005) and Yue and Rue (2011) who used the standard approach of Koenker and Bassett Jr (1978) to estimate their QR model. QR minimises the tilted absolute function $\rho_\tau(.)$, which they called the check-function (Maistre et al., 2017), which asymmetrically weights residuals from the electricity demand model to a degree that depends upon τ.

$$\rho_\tau(\varepsilon) = \begin{cases} (1-\tau)\varepsilon & \text{if } \varepsilon < 0, \quad 0 < \tau < 1 \\ \tau\varepsilon & \text{if } \varepsilon \geq 0 \end{cases}$$

$\rho_\tau(\varepsilon)$ is a continuous piecewise linear function that is non-differentiable at $\varepsilon = 0$ but differentiable everywhere else (Yue and Rue, 2011). When the predicted electricity demand is less than the actual electricity demand it implies a positive residuals, and negative residuals are obtained when the predicted demand is greater than the actual electricity demand. Consequently, when the residuals are positive there is under-prediction of the actual electricity demand, and when the residuals are negative there is an over-prediction of the actual electricity demand. Quantile regression minimises a sum that gives asymmetric penalties $(1-\tau)|\varepsilon|$ for over prediction, and $\tau|\varepsilon|$ for under prediction. For example, if $\tau = 0.1$, there is higher a penalty for over-prediction $(0.9)|\varepsilon|$ and a lower penalty for under-prediction $(0.1)|\varepsilon|$. The check-function ensures that all ρ_τ are positive and the scale is based on the probability τ. A linear model is estimated by solving equation (3.4):

$$\hat{\beta}_\tau = \text{argmin}_{\beta_\tau \in R^p} \sum_i^n \rho_\tau(y_i - x_i{}^T \beta_\tau). \tag{3.4}$$

This is extendable to any quantile, such as the 75 th, 90 th and 99 th percentiles. The QR estimator for β_τ at quantile τ minimises the objective function in equation (3.5):

$$Q_Y(\beta_\tau) = \min_{\beta_\tau \in R^p} \sum_{i: y_i \geq x_i{}^T \beta_\tau}^n \tau |y_i - x_i{}^T \beta_\tau| + \sum_{i: y_i < x_i{}^T \beta_\tau}^n (1-\tau)|y_i - x_i^T \beta_\tau|. \tag{3.5}$$

This is a non-differentiable function and there is no closed-form solution for β_τ; these parameters can be found using a linear programming algorithm (Gibbons and Faruqui, 2014). The minimisation is done for each subsection defined by ρ_τ, where the estimate of the τ th quantile function is achieved with parametric function $x_i^T \beta_\tau$.

The future hourly electricity demand was forecasted at 0.01, 0.02, 0.03 , . . . , 0.99 quantiles of the electricity demand distribution using QR, hence each hour of the day in the forecast horizon would have 99 forecasted electricity demand values, instead of forecasting just a single overall hourly electricity demand as in the case of OLS.

Uncertainties in the forecasts in this study are captured by the interval between 0.01 and 0.99 quantiles of the distribution of demand, as this is the interval into which 98% of the possible future hourly demand is expected to fall. The wider the interval, the more uncertain we are about the forecasted hourly demand as the variability between the forecasts would be very high. The forecasts at the 50 th percentile (median) are important because they could be used as point forecasts, namely, our best guess of electricity demand at that certain hour.

The density functions give the full distribution of hourly demand. The probability of hourly demand between two demand points say "*a*" and "*b*" is the area under the demand density function between the two points. The area could be calculated by integrating the density function between the two points. The probability of exceedance relates to the probability of electricity demand exceeding the specified demand and this could also be calculated by integrating the density function from the specified electricity demand and upwards. The forecasted density demand functions between 2013 and 2023 can therefore be compared in this way. If the density curves shift towards higher electricity demand over the years, then an increase in future hourly electricity demand is expected. If the shift is towards smaller electricity demand over the years, then a decline in future demand is expected.

The South African electricity demand has two daily peaks, especially noticeable in winter, namely a morning demand peak at around 06h00 and an afternoon demand peak at around 18h00. As the winter peak represents the highest annual demand, this is important for planning electricity generation. Therefore, it is important to examine the demand densities of both morning and afternoon peak forecasts over the years. The morning peak density demand is generated from all possible demand forecasts at 06h00, 07h00 and 08h00, whereas the afternoon peak density demand is generated from all possible demand forecasts at 18h00, 19h00 and 20h00. The probability of the future peak electricity demand exceeding a certain value can be calculated by integrating the density functions.

The performance of the QR model is evaluated by comparing the predicted electricity demand density functions with the actual electricity demand density functions. The predicted demand density is generated from all demand forecasts from 1 st to 99 th percentile of the demand distribution. If the forecasted electricity density function closely tracks the actual demand for electricity density then it shows that the model is forecasting well and it is, therefore, reliable. The model is also evaluated by observing the closeness of the actual electricity demand distribution to the predicted lower and upper 99% interval. If the interval is narrow, then the predictions exhibit sharpness. The mean absolute percentage error (MAPE) is used to compare the forecasts at the 50 th percentile with the actual electricity demand; this is mainly to determine how far the point forecasts are from the actual electricity demand.

3.3 Results

3.3.1 Quantile regression model estimates

For illustration purposes, estimated coefficients of the predictor variables of the five QR models (at 0.1, 0.25, 0.5, 0.75, 0.90 quantile levels) are given in Table 3.1. At the 5% level of significance, some variables in Table 3.1 are significant at the certain quantile of the demand distribution, but insignificant at other quantiles. Some variables have strong effects at the lower quantiles while others have strong effects at the higher quantiles, but they could also have the same effect across quantiles. A predictor variable is only insignificant in QR if it is insignificant at all quantiles of the distribution of electricity demand. Table 3.1 shows that the effect of variable "Tues" (Tuesday) differs considerably across the distribution of electricity demand, having a strong effect at the lower quantiles and a weak effect at the higher quantiles of the distribution of electricity demand, while on the other hand variable "month6" (June) has a stronger effects at the higher quantiles and weaker effects at the lower quantiles of the distribution of electricity demand. The predictor "newyear", for example, has the same effect at all quantiles of the electricity demand distribution.

3.3.2 Model assessment

The densities of the actual and predicted electricity demand between 2013 and 2015, illustrated in Figure 3.1, were used to assess the reliability of the forecasts. If the distribution of the predicted electricity demand is close to that of the actual electricity demand, then the forecasts are considered to be reliable.

The sharpness of the forecasts refers to how tightly the predicted distribution covers the actual distribution. The closer the forecasted electricity demand at the 1 st and the 99 th percentile points are to the actual demand, the better, and this indicates the sharpness of the forecasts (Figure 3.2).

The MAPE was calculated from the untransformed electricity demand data. The MAPE between the hourly electricity demand forecasts at the 50 th percentile and the actual hourly electricity demand over the three year period were below 5% and the overall MAPE was 2.77%, as shown in Table 3.1 indicating that the demand forecasts deviate from the actual demand by less than 5% on average. Lewis (1982) indicates that a MAPE of less than 10% can be classified as a highly accurate forecast. The QR model therefore provides good demand forecasts at the 50 th percentile.

Table 3.1: Model of log demand

Variables	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
Intercept	4.097 ***	4.254 ***	4.443 ***	4.582 ***	4.610 ***
newyear	−0.079 ***	−0.072 ***	−0.064 ***	−0.050 ***	−0.029 ***
Humanrights	−0.018 ***	−0.018 ***	−0.015 ***	−0.010 ***	−0.009 **
freedomDay	−0.021 ***	−0.016 ***	−0.013 ***	−0.014 ***	−0.017 ***
WorkersDay	−0.025 ***	−0.028 ***	−0.029 ***	−0.016 ***	−0.013 ***
YouthDay	−0.035 ***	−0.029 ***	−0.015 ***	−0.010 **	−0.010 ***
HeritageDay	−0.034 ***	−0.025 ***	−0.022 ***	−0.017 ***	−0.010
Christmasday	−0.054 ***	−0.043 ***	−0.033 ***	−0.020 ***	−0.009 **
goodwillday	−0.034 ***	−0.036 ***	−0.034 ***	−0.029 ***	−0.026 ***
month1	−0.021 ***	−0.013 ***	−0.005 ***	−0.002 ***	−0.001 *
month2	−0.008 ***	−0.001	0.004 ***	0.004 ***	0.003 ***
month4	−0.007 ***	−0.007 ***	−0.005 ***	−0.002 **	0.001
month5	−0.009 ***	−0.006 ***	−0.004 ***	−0.004 ***	−0.001
month6	−0.004 *	0.003 ***	0.005 ***	0.006 ***	0.007 ***
month7	−0.008 ***	−0.002	−0.001	0.001	0.002
month8	−0.018 ***	−0.008 ***	−0.007 ***	−0.006 ***	−0.005 ***
month9	0.004 ***	0.004 ***	0.004 ***	0.003 ***	0.003 ***
month10	0.002 ***	0.003 ***	0.002 ***	0.002 ***	0.003 ***
month11	0.003 ***	0.004 ***	0.005 ***	0.005 ***	0.005 ***
sin6	0.003 ***	0.004 ***	0.005 ***	0.007 ***	0.008 ***
cos6	−0.003 ***	−0.003 ***	−0.003 ***	−0.001 ***	−0.001 ***
sin12	−0.029 ***	−0.031 ***	−0.033 ***	−0.034 ***	−0.034 ***
cos12	−0.004 ***	−0.005 ***	−0.006 ***	−0.009 ***	−0.010 ***
sin24	−0.034 ***	−0.035 ***	−0.036 ***	−0.037 ***	−0.036 ***
cos24	−0.039 ***	−0.039 ***	−0.040 ***	−0.041 ***	−0.041 ***
lag70128	−0.018	−0.039 ***	−0.048 ***	−0.044 ***	−0.056 ***
lag70152	−0.050 ***	−0.059 ***	−0.069 ***	−0.086 ***	−0.074 ***
lag70176	−0.089 ***	−0.089 ***	−0.121 ***	−0.131 ***	−0.136 ***
lag70200	−0.017	−0.019 *	0.005 ·	0.024 **	0.026 **
lag70224	0.414 ***	0.439 ***	0.438 ***	0.422 ***	0.395 ***
lag70248	−0.034 ***	−0.051 ***	−0.050 ***	−0.039 ***	−0.031 ***
lag70272	−0.133 ***	−0.144 ***	−0.157 ***	−0.177 ***	−0.161 ***
Sun	−0.007 ***	−0.007 ***	−0.007 ***	−0.007 ***	−0.007 ***
Mon	0.005 ***	0.003 ***	0.002 ***	0.003 ***	0.003 ***
Tues	0.004 ***	0.003 **	0.002 *	0.001	0.001
Wed	0.008 ***	0.007 ***	0.006 ***	0.005 ***	0.004 ***
Thurs	0.007 ***	0.006 ***	0.004 ***	0.003 ***	0.003 ***
Fri	0.009 ***	0.007 ***	0.005 ***	0.005 ***	0.005 ***
fribtwn	−0.067 ***	−0.035 ***	−0.030 ***	−0.024 ***	−0.023 *
monbtwn	−0.016 ***	−0.015 ***	−0.010 ***	−0.010 ***	−0.005 *
lngwknd	−0.015 ***	−0.013 ***	−0.012 ***	−0.012 ***	−0.011 ***

$p > 0.05$; $^{\star}p < 0.05$; $^{\star\star}p < 0.01$; $^{\star\star\star}p < 0.001$

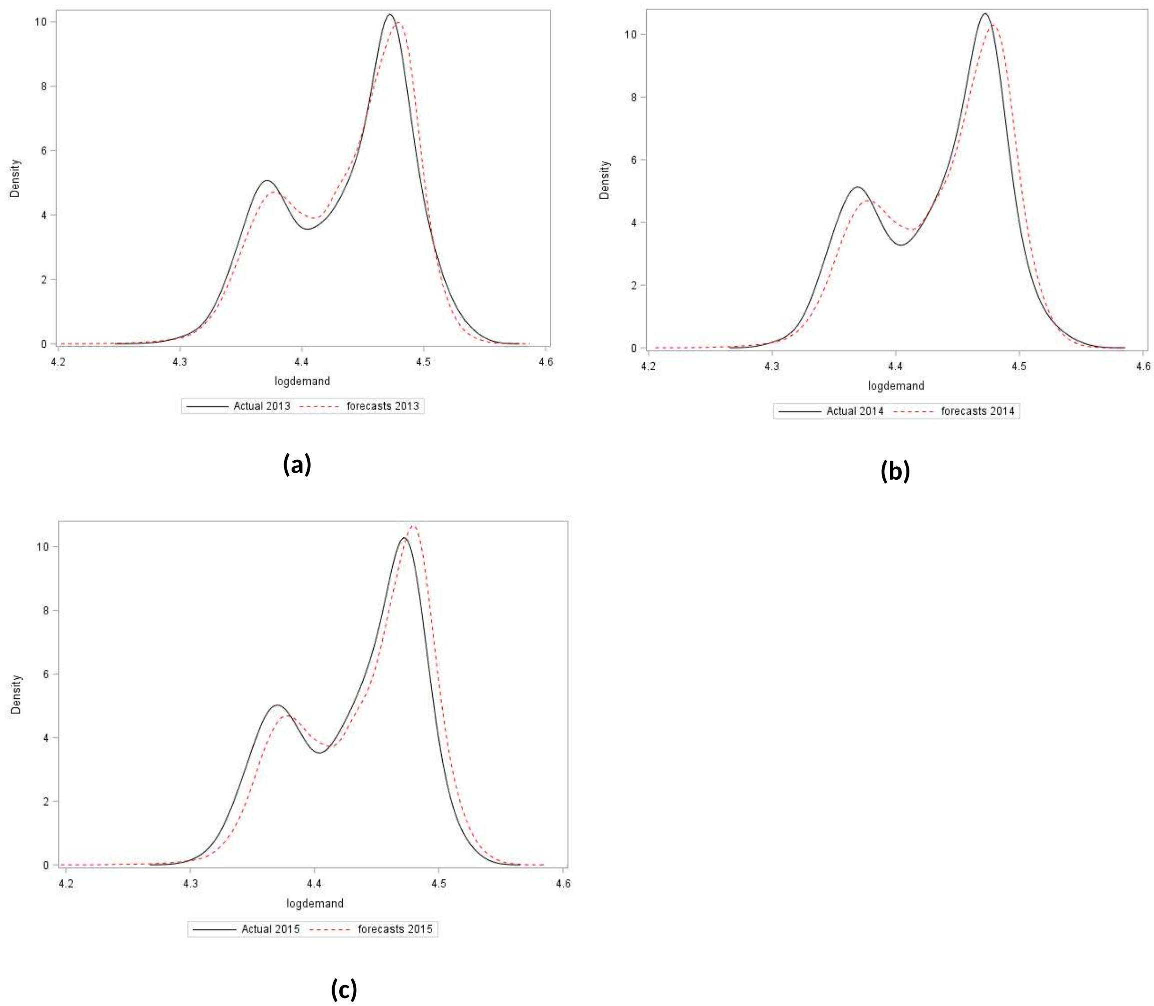


Figure 3.1: Comparison of actual and forecasted demand distributions: (a) 2013 actual against forecasted density demand; (b) 2014 actual against forecasted density demand; (c) 2015 actual against forecasted density demand

Table 3.2: Mean absolute percentage error (Forecasts at 50 th percentile)

Hours	**2013**	**2014**	**2015**	**Average**
0	2.160	2.962	3.464	2.862
1	1.784	2.789	3.372	2.649
2	1.688	2.698	3.353	2.579
3	1.584	2.537	3.123	2.414
4	1.813	2.298	2.779	2.297
5	3.781	3.816	3.916	3.838
6	4.408	4.181	4.225	4.272
7	3.750	3.803	4.144	3.899
8	2.751	2.771	2.955	2.826
9	2.443	2.442	2.265	2.384
10	1.991	1.863	2.064	1.972
11	1.826	1.838	2.219	1.961
12	1.770	2.112	2.552	2.145
13	2.036	2.493	2.898	2.475
14	2.048	2.578	2.986	2.537
15	1.827	2.318	2.802	2.316
16	2.158	2.219	2.802	2.393
17	3.837	3.218	3.740	3.598
18	5.111	4.448	4.120	4.560
19	2.776	2.390	2.443	2.536
20	2.400	2.004	2.385	2.263
21	2.319	2.022	2.612	2.318
22	2.242	2.519	3.125	2.629
23	2.168	2.874	3.340	2.794
Average	2.528	2.716	3.070	2.772

3.3.3 Probabilistic forecasts of daily profiles over the years

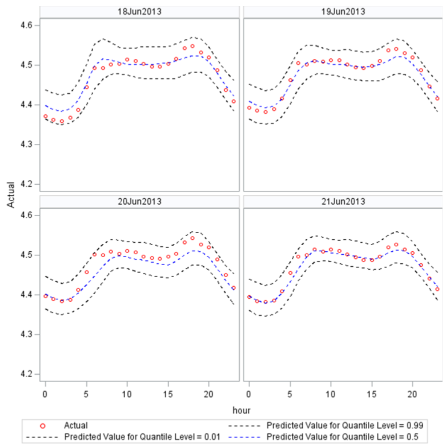
For illustration purposes, four days in June were selected and their results from 2013 to 2015 are discussed here. The four days in June were selected in such a way that the day with the highest hourly electricity demand of the year was included. The different panels in Figure 3.2 give the hourly electricity demand of each of the four days in 2013. For each day, the black graph represents the hourly electricity demand at the 1 st percentile level of the electricity demand distribution i.e. these are the points below which 1% of all possible future hourly electricity demand is expected to fall. The top black line represents future hourly electricity demand at the 99 th percentile of the electricity demand distribution i.e. these are the points above which 1% of the future hourly electricity demand would fall and below which 99% of all possible future hourly electricity demand is expected to fall. The blue line represents the forecasted hourly electricity demand at the 50 th percentile and the red circles represent the actual hourly electricity demand .

The 18 June 2013 is the day that contained the highest hourly electricity demand and this happened at 18h00. The demand on this day at 18h00 was forecasted to lie between 30 342 kW (logdemand = 4.482) and 37 115 kW (logdemand = 4.570) while the actual electricity demand was 35 394 kW. In 2014 the highest demand of 36 039 kW (logdemand = 4.557) was on 12 June at 18h00. The electricity demand was forecasted to lie between 29 973 kW (logdemand = 4.477) and 36 294 kW (logdemand = 4.560) while in 2015, the highest electricity demand of 34 481 kW (logdemand = 4.538) was on 11 June at 18h00. The electricity demand was forecasted to lie between 30 917 kW (logdemand = 4.490) and 36 880 kW (logdemand = 4.567). As indicated above, the lower predicted bounds are the forecasts at the 1 st percentile while the upper bounds are the forecasted electricity demand at the 99 th percentile of the demand distribution. Therefore, this is the interval within which 98% of the forecasted electricity demand were expected to fall on these days at 18h00 and the actual value was indeed within this interval over all of these days.

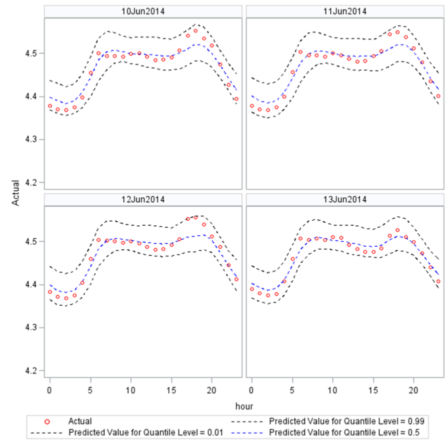
3.3.4 Density forecasts of electricity demand over years

The distribution of the hourly electricity demand in South Africa is bimodal as shown in Figure 3.3. By comparing the density functions of the electricity demand over the years, insight into expected shifts in patterns can be obtained. The forecasted electricity demand distributions obtained from the QR models for the period investigated suggest that the hourly electricity demand from Eskom is more likely to shift towards lower demand over the years until 2023 (Figure 3.3). The apparent year to year decline in electricity demand between 2013 and 2015 could be attributed to the increase in the number of households and companies generating their own electricity through renewable energies, the shrinking economic growth and the increase in electricity prices , amongst other, random, factors.

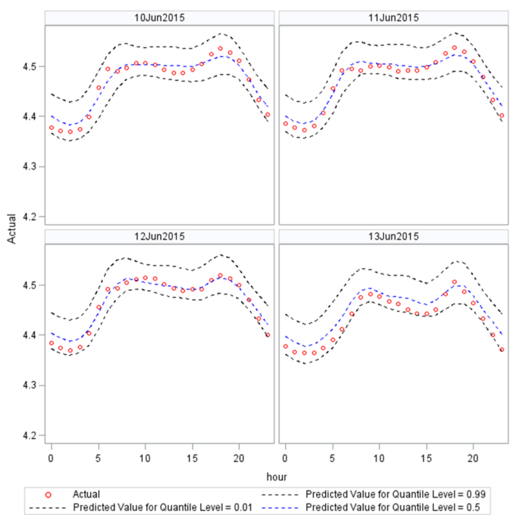
In addition, the forecasted hourly density of electricity demand (in Figures 3.3 to 3.5) could



(a)



(b)



(c)

Figure 3.2: Distribution of forecasted daily electricity demand: (a) 2013 actual hourly demand against forecasted; (b) 2014 actual hourly demand against forecasted; (c) 2015 actual hourly demand against forecasted

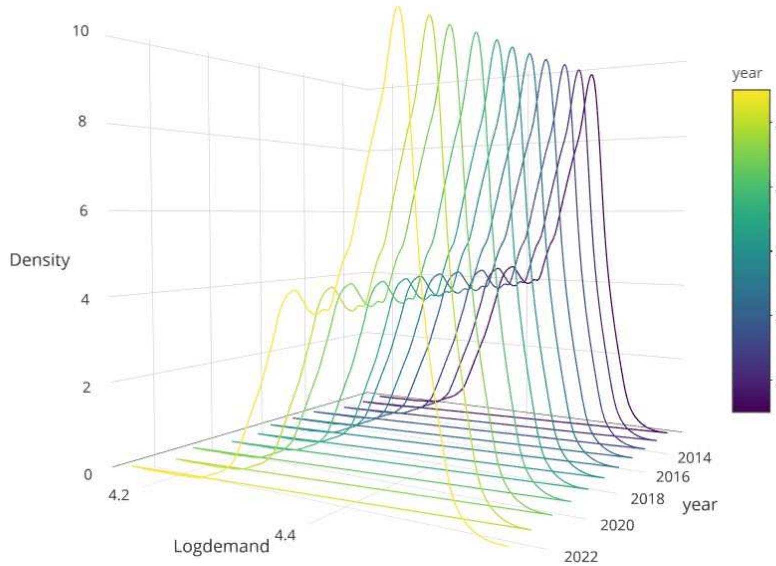


Figure 3.3: Distribution of forecasted hourly demand between 2013 and 2023

be used to calculate probabilities of exceedance, for example, the probability of hourly electricity demand exceeding 30 342 kW (logdemand = 4.482). This probability of exceedance can be calculated as the area under the density curve comprising all electricity demand ranging from 30 342 kW and above, and the area can be computed by integrating the density functions in Figure 3.3. The probability of electricity demand exceeding 30 342 kW is less likely in 2023 than it is in any previous year while 2015 had the highest probability of electricity demand exceeding 30 342 kW in the period between 2013 and 2023 as shown in Figure 3.3 (the 2015 density function of electricity demand (yellow graph) shifted more to the higher electricity demand compared to other years' demand densities).

The forecasts obtained from QR can be used to investigate the expected future peak electricity demand and the probability of exceedance of certain levels of demand over the years. The annual peak electricity demand is very important for planning purposes, as these represent the maximum that would need to be supplied in an hour and if the power generating company could meet the daily peak hourly electricity demand, it could meet any hourly demand in the day. Figure 3.4 and Figure 3.5 suggest that the morning and afternoon peak electricity demand distributions are likely to shift towards lower electricity demand over the years until 2023.

3.4 Concluding remarks on QR

The daily electricity demand in South Africa generally has two peaks, more noticeable during winter than summer seasons. The morning electricity demand peak occurs at around 06h00 and the afternoon demand peak at around 18h00, as discussed in Chapter 1. OLS models would most likely underestimate the peaks as they model the mean of the demand distribution while QR models the demand at all percentiles of the demand distribution and

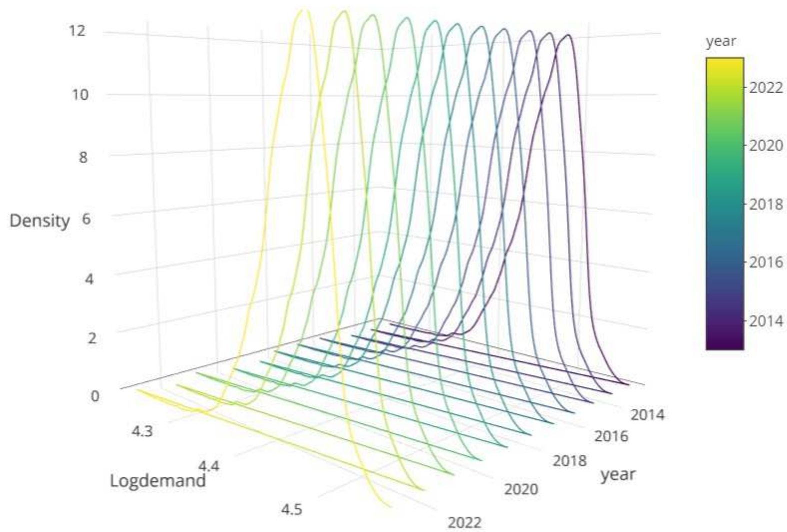


Figure 3.4: The morning peak demand distributions

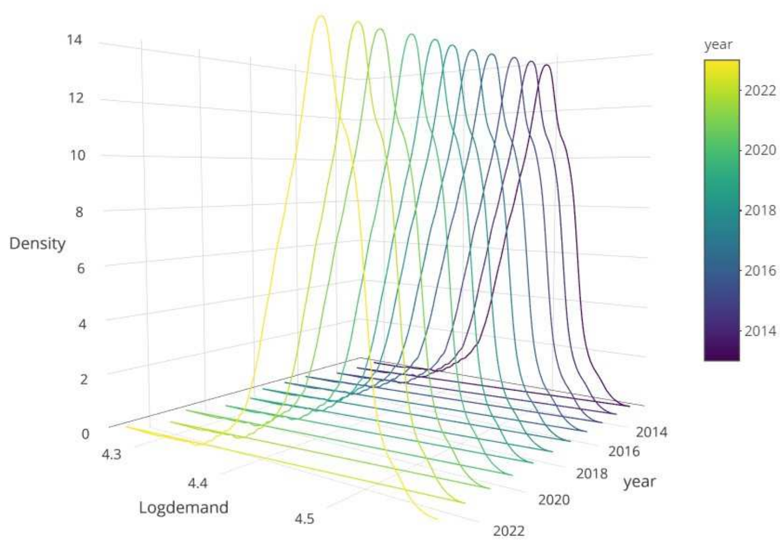


Figure 3.5: The afternoon peak demand distributions

therefore can provide better peak forecasts. In addition, QR gives the full hourly demand distribution, hence the uncertainties around the forecasts are quantifiable. While the best guess of the future hourly electricity demand can be obtained from the forecasted demand at the 50 th percentile, QR gives forecasts at all percentiles of the distribution, allowing the potential variabilities in the forecasts to be evaluated by comparing the 50 th percentile forecasts with the forecasts at other percentiles. Additional planning information, such as expected pattern shifts and probable peak values, could also be obtained from the forecasts produced by the QR model, while such information would not easily be obtained from the point forecasting approaches.

The first important finding is that the demand forecasts at the 50 th percentile from the QR model closely estimate the actual hourly electricity demand (see the red circles and blue line in Figure 3.1 and the MAPE values in Table 3.2). The second finding is that the distributions of the hourly electricity demand and the peak daily electricity demand in South Africa are shifting towards lower demand over the years, as shown in Figures 3.3, 3.4 and 3.5.

The forecasted demand distribution closely matched the actual demand distribution between 2013 and 2015, as shown in Figure 3.1. Therefore, the forecasted electricity demand distribution from QR is expected to continue representing the actual electricity demand distribution beyond 2015.

More work still needs to be done on the selection of predictor variables to be included in the QR model because variables have different effects at different quantiles of the electricity demand distribution. Predictors in QR are only insignificant if they are insignificant at all quantiles of the electricity demand distribution. The determination of significant variables is cumbersome because the effect of each variable is checked at all quantiles of the demand distribution (in this study at 99 points of the distribution). The contributions of Chapter 3 are:

1. that QR can be used as an extension of an OLS regression to generate probabilistic forecasts of the long-term electricity demand in South Africa. QR has not been used before to forecast the long-term hourly electricity demand to the best of our knowledge,

2. that variabilities in the forecasts are evaluated and uncertainties around the forecasts are assessed as the full distribution of demand is forecasted in this modelling approach, and

3. that the probabilities of exceedance can be calculated, such as the probability of future peak demand exceeding certain levels of demand because electricity demand is calculated at all quantiles of the demand distribution.

As already indicated in Chapter 1, QR is in the frequentist modelling framework under which the model parameters are considered to be fixed and the data random, whereas the

Bayesian approach in Chapter 4 assumes that the model parameters are random and the data are fixed. The model parameters are estimated with their full distribution represented by density functions.

Chapter 4

Bayesian Structural Time Series Approach

Bayesian Structural Time Series Approach[1]

4.1 Background

The literature on BSTS is very limited and nonexistent for long-term electricity demand forecasting to the best of our knowledge; however, a BSTS modelling approach would be ideal for the long-term forecasting of electricity demand in South Africa. This is because the electricity demand trend has shown some changes in the recent past and therefore it may be difficult to make assumptions regarding the shape and direction of the future trend of electricity demand. The BSTS model is modular, therefore if results appear to be counter-intuitive, then the contribution of each model component can be analysed separately and the origin of the problem can easily be located.

The BSTS model was used to forecast the long-term electricity demand in South Africa, since it was considered a good option for forecasting demand in the presence of uncertainties. As indicated in Chapter 1, such uncertainties could among others emanate from increased technologies making use of electricity, population growth, general randomness in individual usage of electricity, seasonal effects, prevailing economic patterns, change in weather conditions, escalating costs of electricity, changes in use of power saving electrical appliances, the growing sources of renewable energies and market penetration of electric vehicles.

Unlike the QR model that considers the likelihood of the data given the model (frequentist approach), the BSTS model considers the likelihood of the model given the data (Bayesian approach). The important building blocks of the BSTS model are:

- the Kalman filter, which is a technique for time series decomposition. Different state

[1] A modified version of this chapter has been accepted by an ISI rated journal: Applied Mathematics and Information Sciences (AMIS) on 23 November 2018 and published online on 1 March 2019.

variables like trend, seasonality, regression and other state variables are added in this step,

- the Spike and slab method, the method by which most important regression predictors are selected in this step,
- Bayesian model averaging, by which the results and calculated predictions are combined.

4.2 Methodology

The posterior distribution of electricity demand in this study can be represented by equation (4.1):

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\int_\theta \pi(x|\theta)\pi(\theta)} \mathrm{d}\theta \tag{4.1}$$

where θ is the set of coefficients for the electricity demand driver variables to be estimated, x are the drivers of electricity demand, $\pi(x|\theta)$ is the function generating the electricity demand data given the model parameters (the likelihood function), $\pi(\theta)$ is the prior distribution representing information on the model parameters before the demand for electricity data were observed and $\int_\theta \pi(x|\theta)\pi(\theta)\mathrm{d}\theta$ is a normalising constant. Since this form of the normalising constant makes it difficult to compute the posterior in closed-form. Therefore, the posterior distribution in equation (4.1) can be written as, $\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$. The posterior distribution summarises all available information about the parameters of the drivers of electricity demand after observing the electricity demand data. The posterior mean of the parameters of electricity demand are estimated by equation (4.2):

$$E(\theta|x) = \int_\theta \theta\pi(\theta|x)\mathrm{d}\theta, \tag{4.2}$$

which serves as the point estimate of the parameters θ given the data. The posterior predictive distribution of the electricity demand could be used to predict future values of electricity demand (x) with equation (4.3):

$$\pi(x_{n+1}|x) = \int_\theta \pi(x|\theta)\pi(\theta|x)\mathrm{d}\theta. \tag{4.3}$$

The complicated integrals in equation (4.1) make it difficult to compute the posterior distribution, hence the Markov Chain Monte Carlo (MCMC) method is used to estimate the model parameters through sampling from the posterior distribution. The idea is to attain a chain whose equilibrium distribution matches that of the hourly electricity demand posterior distribution. The sampling stops when the chain reaches its equilibrium.

According to Brodersen et al. (2015) and Scott and Varian (2014) the BSTS is formulated as a state space model. State space models are attractive because they are modular as the independent state components can be combined by concatenating the observation vectors Z_t (Scott and Varian, 2014). There are two pieces to this model, that is, the observation equation that links the observed hourly electricity demand (y_t) with the unobserved latent state (α_t). The observation equation is represented in equation (4.4).

$$y_t = Z_t^T \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \quad t = 1, 2, \ldots, 24 \tag{4.4}$$

where y_t is the electricity demand at hour t; α_t is a vector of latent variables (a set of unknown parameters); H_t and Z_t are structural parameters where H_t is a constant diagonal matrix with diagonal elements ${\sigma_\varepsilon}^2$ and Z_t is the observation vector. The transition equation is given in equation (4.5):

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t), \quad t = 1, 2, \ldots, 24 \tag{4.5}$$

where T_t, R_t and Q_t are structural parameters and η_t is a lower dimension than α_t. Q_t is a constant diagonal matrix with diagonal elements ${\sigma_u}^2$, ${\sigma_\omega}^2$ and ${\sigma_\varphi}^2$. $\eta_t = (v_t, \varphi_t, \omega_t)$ are the independent components of Gaussian random noise. The state space model representation of the BSTS model used in this study is presented in equation (4.6):

$$y_t = \mu_t + \gamma_t + \beta^T X_t + \varepsilon_t \tag{4.6}$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + \varphi_t \tag{4.7}$$

$$\delta_t = \delta_{t-1} + v_t \tag{4.8}$$

$$\gamma_t = -\sum_{s=1}^{s-1} \gamma_{t-s} + \omega_t \tag{4.9}$$

where y_t is the electricity demand at hour t.

The μ_t is the local level trend of the electricity demand series. The trend is assumed to follow a random walk with φ_t representing some errors where $\varphi_t \sim N(0, {\sigma_\varphi}^2)$. The prior is on ${\sigma_\varphi}^2$.

The γ_t is the seasonal component of the electricity demand series. The seasonal model can be thought of as a regression on n seasons, dummy variables with coefficients constrained to sum to one. If there are S seasons then the state vector γ_t is of dimension $S-1$. ω_t are seasonality errors, $\omega_t \sim N(0, {\sigma_\omega}^2)$ and the prior is on ${\sigma_\omega}^2$.

The ε_t is the irregular component, $\beta^T X_t$ is the regression component, the local level term which defines how the latent state evolves over time and it is often referred to as an unobserved trend; the level = μ_t plus slope δ_t and seasonal $S-1$ dummy variables with time varying coefficients that are expected to sum up to zero, and y_t is the response series representing electricity demand at hour t.

The ε_t is the irregular component (noise), are assumed to follow an autoregressive process of order one AR(1) and it is distributed as $\varepsilon_t \sim N(0, \sigma^2)$. The prior coefficients for the AR(1) component are set at $\beta = 0$, $\sigma^2 = 1$, and stationarity is ensured by setting $-1 < \phi_1 < 1$.

In this model, a time series component captures the general trend (μ_t), seasonality (γ_t) and irregular patterns (u_t) in the data. The priors for the time series component of the model are specified during the creation of the state specification of the model.

The $\beta^T X_t$ represents the regression component of the BSTS model where β represents the vector of regression coefficients of the electricity demand model. The concept of sparse modelling asserts that if there are many possible predictors for the electricity demand, some of them are expected to have zero coefficients. The natural way to represent sparsity in Bayesian modelling is through a spike and slab prior on the regression coefficients (Scott and Varian, 2014). In this way, possible drivers of electricity demand are selected. The prior is on the set of β coefficients.

Since the BSTS model in this study has a regression component, the spike and slab priors are used to select the model variables. The spike part governs the probability of a given variable to be selected into the model (having a non-zero coefficient). The slab part shrinks the non-zero coefficients towards the prior expectation (often zero).

Let τ denotes a vector of *ones* and *zeros* where *one* indicates that the variable is selected into the model (non-zero coefficient) and *zero* indicates that the variable is not selected into the model. If β_τ is the subset of the elements of β where $\tau_k = 1$ if $\beta_k \neq 0$ and $\tau_k = 0$ if $\beta_k = 0$, then the spike and slab prior for the electricity demand model can be written as in equation (4.10);

$$p(\tau, \beta, \sigma_\varepsilon^2) = p(\tau)p(\sigma_\varepsilon^2|\tau)p(\beta_\tau|\tau, \sigma_\varepsilon^2) \tag{4.10}$$

where σ^2 is the error variance, the marginal distribution $\rho(\tau)$ is the spike representing the probability of choosing a given variable for inclusion into the electricity demand model and it is assumed to follow a Bernoulli distribution given in equation (4.11):

$$p(\tau) = \prod_{j=1}^{J} \pi_j^{\tau_j} (1-\pi_j)^{1-\tau_j} \qquad j = 1, 2, \ldots, J \tag{4.11}$$

where π_j is a prior probability of variable j being included into the demand model. As a practical consideration, all π_j are assumed to have the same value π. If there are h non-zero predictors, then $\pi = h/k$, where k is the dimension of x_t. For the specific values of j, some variables can be forced to be included or excluded by setting $\pi_j = 1$ for inclusion and $\pi_j = 0$ for exclusion. For the slab part, a normal prior is used for β which leads to an inverse gamma prior for σ_ε^2 and it is given by equation (4.12):

$$\frac{1}{\sigma_\varepsilon^2} \sim G\left(\frac{v_\varepsilon}{2}, \frac{s_\varepsilon}{2}\right) \tag{4.12}$$

where v_ε in equation (4.13) is the prior sample size (shrinkage parameter) which can be thought of as a prior sample size for learning the variance parameter of each coefficient and s_ε is the prior sum of squares.

$$\frac{s_\varepsilon}{v_\varepsilon} = (1 - R^2) s_y^2 \tag{4.13}$$

where s_y^2 is the sample variance and R^2 is the amount of variation explained by the regression model.

In this study, the spike and slab priors for the regression component are set in such a way that all potential independent variables are given equal chances (50%) of being included in the model ($\pi_j = 0.5$). The prior for the overall variation explained by the model is set at 0.5 ($R^2 = 0.5$) and a shrinkage parameter is set at 1% ($v_\varepsilon = 0.01$). The full Bayes approach is used with a minor violation as s_y^2 is data-determined. The detailed formulation of BSTS is given by Brodersen et al. (2015) and Scott and Varian (2014).

4.3 Results

4.3.1 Model variable selection

The fitted BSTS model in this study has a regression component, therefore, predictor variables in the regression component are selected using the spike and slab prior. This is in general a powerful way of reducing a large set of possibly correlated variables into a smaller set to be used in the model, which generally addresses the problem of multicollinearity and helps to ensure that the resultant model is parsimonious. The first step in variable selection is to estimate the coefficients of all potential regression models given the predictor variables under consideration. The averages of the coefficients of each considered predictor

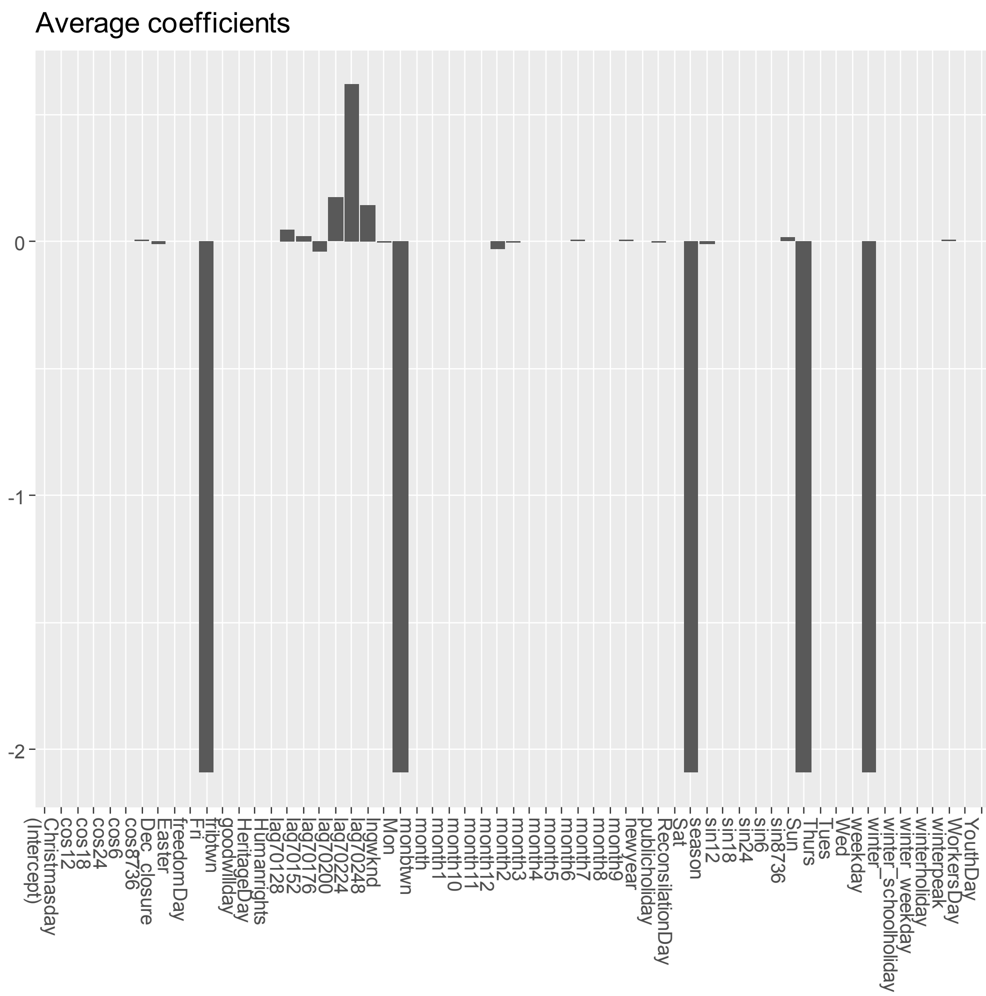


Figure 4.1: Average coefficients of parameters calculated from all possible models

variable from all possible models are then calculated and plotted in Figure 4.1. The regression model is dominated by eight variables (Friday, Monday, Saturday, Sunday, lag70200, lag70224 and lag70248) because their average coefficients are by far greater than the rest (Figure 4.1). Other variables had coefficients either close to *zero* or *zero* which could likely be insignificant.

Figure 4.2 depicts the predictor variables with their proportions within all possible models into which they would be selected. There were 20 variables selected into all possible models indicating that, given the predictor variables under consideration, 20 of them would be included in all possible models. In this study, only variables that were selected into all possible models (inclusion probability of one in Figure 4.2) are considered for modelling.

4.3.2 Estimated model parameters

The coefficients of the predictor variables in Bayesian models such as BSTS are estimated with their full distributions and their most likely values correspond to the value where

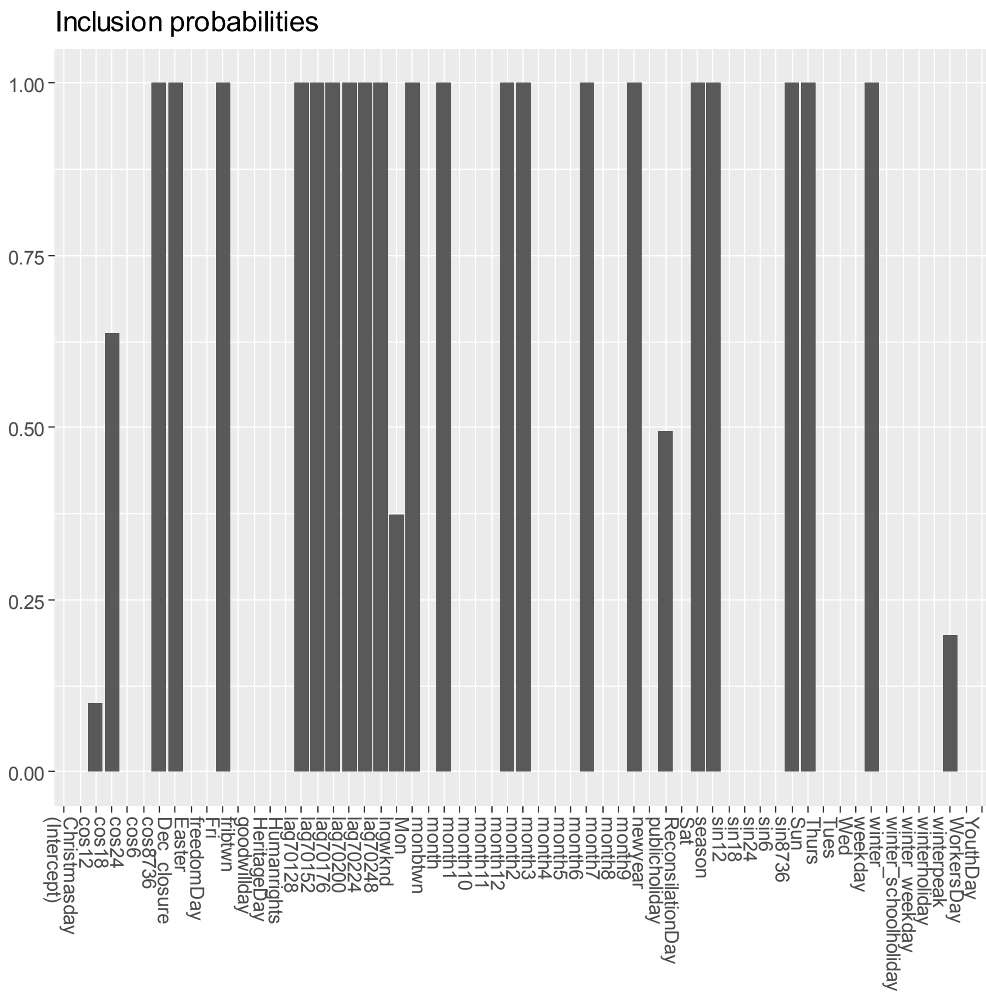


Figure 4.2: Parameters inclusion probabilities into the model

the density function reaches the peak. The estimated variances of each model component and the AR(1) parameter are depicted in Figure 4.3. The model variances are estimated with their full distributions and uncertainties around the estimated parameters are therefore quantified. The most likely variance of the observations (variance of the electricity demand) (σ_{ε}) in Figure 4.3 is about 0.004 and it is nearly normally distributed, while the estimated variances of other components of the model are either skewed towards the right or left.

Figure 4.4 gives the estimated coefficients of the predictor variables of electricity demand associated with the regression component of the BSTS model. The coefficients are estimated with their full distributions. Unlike in the QR model in Chapter 3, parameters in BSTS are estimated with their full distributions because, under the Bayesian framework, model parameters are considered to be random. For example, the most likely value for β_{lag70248} in Figure 4.4 is 0.11 but the parameter is estimated with its full distribution.

4.3.3 Model components and their contributions

The considered individual BSTS model components are given in Figure 4.5. The model has four components, with the local level assumed to follow a random walk. The other components are a seasonal, an autoregressive components of order one which represents the stationary component of the model, and a regression component. The regression component contributed the most to the overall model, followed by the trend component (Figure 4.5 (a) and 4.5 (c)). The distribution graph for the regression component in Figure 4.5 (c) takes on high values compared to other components of the model, indicating that it contributed the most to the forecasted electricity demand, followed by the nonlinear trend component. Given the data under consideration, it would be ideal for the trend to contribute minimally to the overall model, because the historical electricity demand data in the last 19 years until 2015 did not show a simple one directional trend, and therefore there may be uncertainties regarding the future direction of the trend. If the trend contributed more to the overall model and it was incorrectly specified then the long-term electricity demand forecasts would vastly deviate from the actual electricity demand. One of the advantages of the BSTS model is that it can provide information about the contribution of each component to the model, which makes it easier to investigate the origin of the problem in the case of getting suspicious results. A good model with the local level trend component contributing minimally to the model was preferred in this study.

4.3.4 Forecasts assessment

The forecasted daily profiles at the 50 th percentile of the distribution of electricity demand in Figure 4.6 could be used to assess the accuracy of the forecasts. If the forecasted daily profiles capture the actual daily profiles over the years and the forecasts at the 50 th percentile are close to the actual electricity demand, then it would indicate that the model was forecasting well. The forecasted daily profiles consistently compared well with the actual daily profiles of demand as shown in Figure 4.6. The MAPE between the actual and the forecasts at the 50 th percentile of the distribution of electricity demand were mostly below 4% in all hours, as shown in Table 4.1, indicating that the forecasts at the 50 th percentile

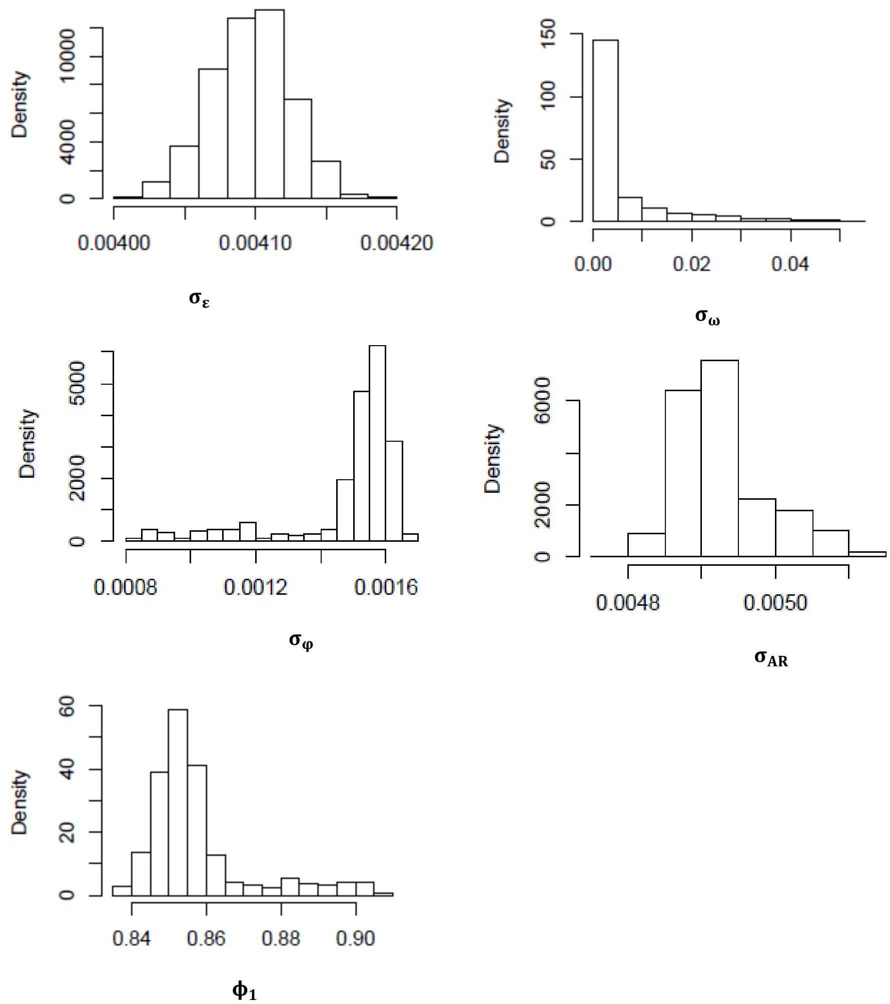


Figure 4.3: Variances of different model components

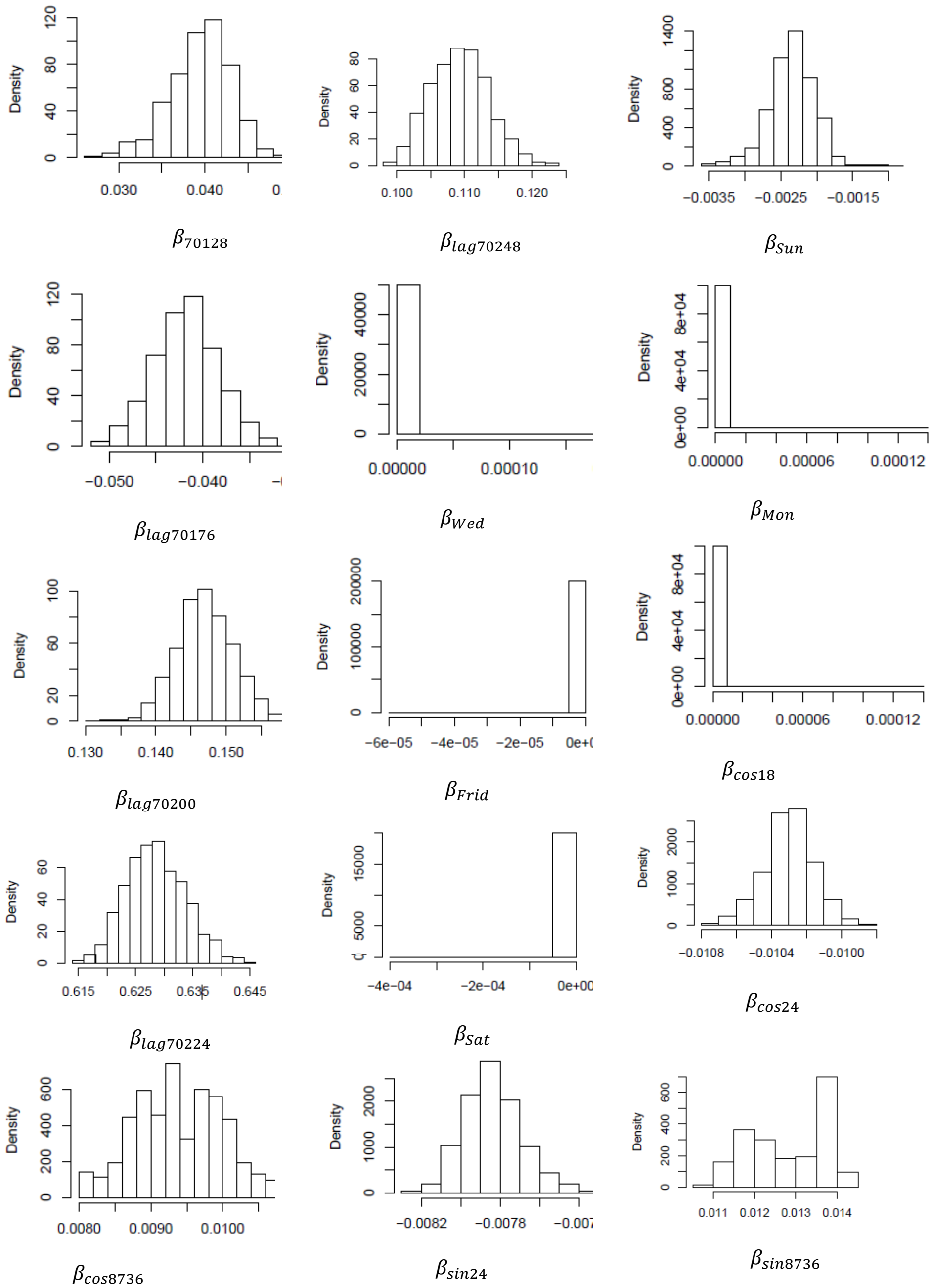
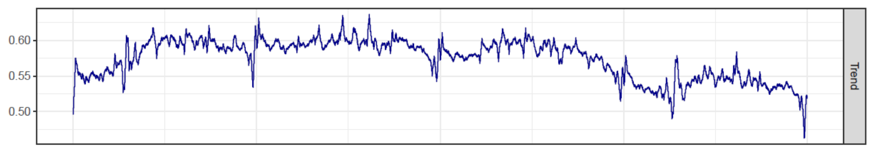
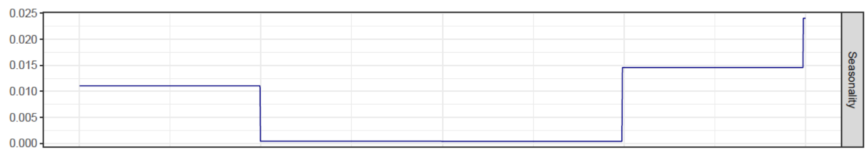


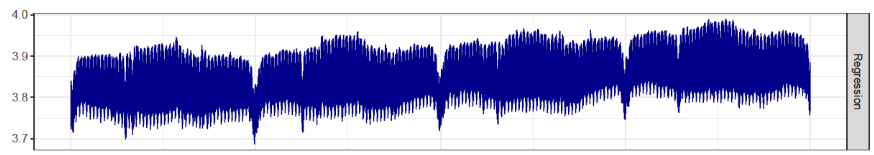
Figure 4.4: Regression model component: estimated parameters



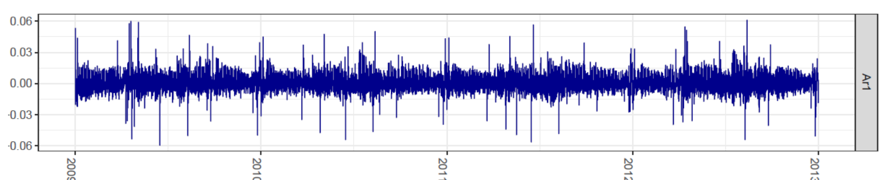
(a)



(b)



(c)



(d)

Figure 4.5: Model components: (a) trend; (b) seasonal; (c) regression; (d) AR(1)

Table 4.1: Mean absolute percentage error (Forecasts at 50 th percentile)

Hours	**2013**	**2014**	**2015**	**Average**
0	3.714	2.190	4.093	3.332
1	3.531	2.221	4.349	3.367
2	3.456	2.320	4.841	3.539
3	3.603	2.324	4.702	3.543
4	4.282	2.569	3.809	3.553
5	6.331	4.691	4.162	5.061
6	5.970	4.699	4.722	5.130
7	4.135	2.673	4.266	3.692
8	4.299	2.337	3.361	3.332
9	4.303	2.315	3.036	3.218
10	4.480	2.367	3.437	3.428
11	4.226	2.337	3.798	3.454
12	3.927	2.362	4.340	3.543
13	3.840	2.463	4.753	3.685
14	3.829	2.717	5.065	3.870
15	4.127	2.881	4.921	3.976
16	4.501	3.012	4.443	3.985
17	4.523	3.187	4.020	3.910
18	4.099	3.333	3.399	3.611
19	3.951	2.997	3.125	3.358
20	4.593	2.771	3.442	3.602
21	4.653	2.479	3.845	3.659
22	4.272	2.256	3.815	3.448
23	3.941	2.189	3.912	3.347
Average	4.274	2.737	4.069	3.693

of the demand distribution deviate by less than 4% from the actual demand on average. As indicated previously, Lewis (1982) indicates that a forecast with MAPE of less than 10% can be classified as highly accurate forecast.

The MAPE in Table 4.1 were calculated from the untransformed hourly electricity demand data from 2013 to 2015. There is no reason to doubt that the selected BSTS model will continue to forecast electricity demand accurately beyond 2015.

4.4 Discussion

For discussion purposes, four days in June were selected in such a way that the day with the highest hourly electricity demand of the year was included. The highest demand in 2013 was 35 393 kW on 18 June, 36 039 kW on 12 June 2014 and 34 481 kW on 11 June 2015

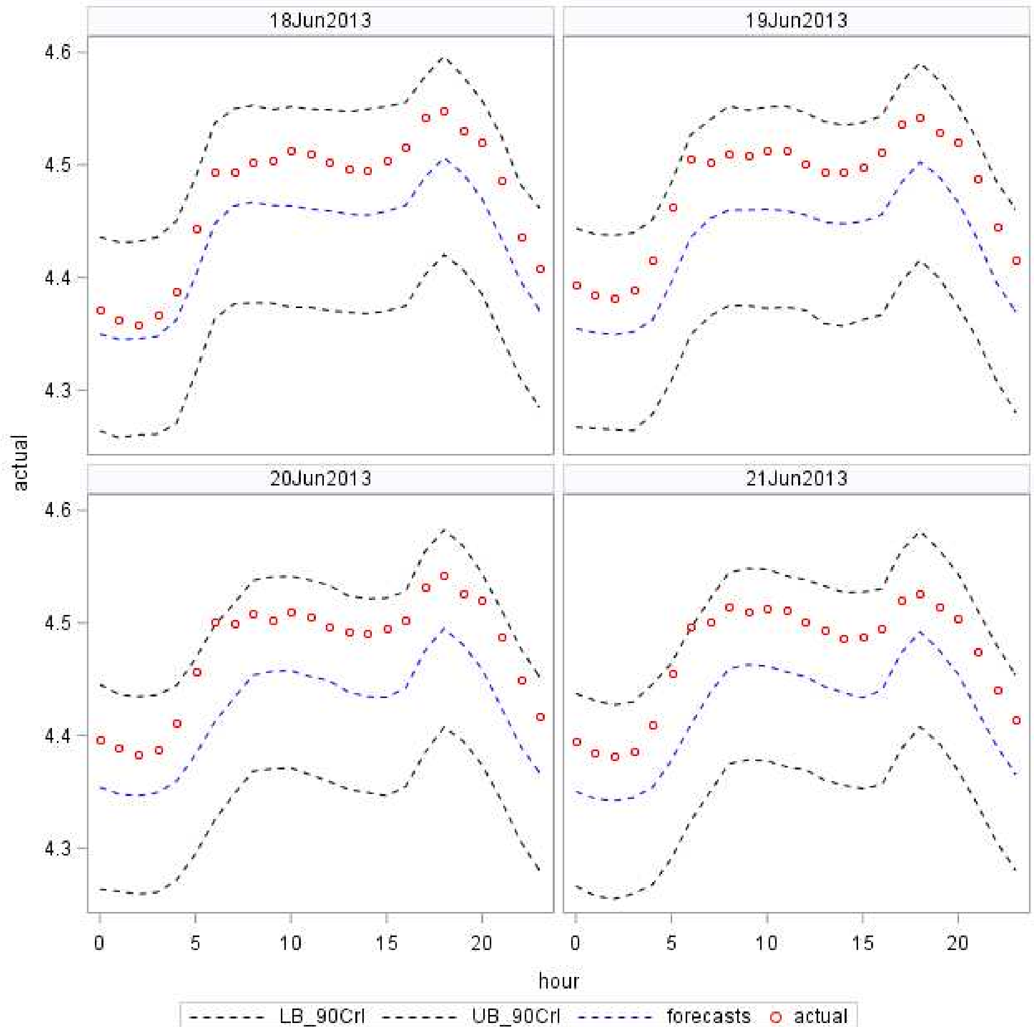


Figure 4.6: Comparison of actual and forecasted hourly demand at 50 th percentile with 90% credible interval: 2013 actual hourly demand against forecasted

and they were all at 18h00. The black graphs in Figures 4.6, 4.7 and 4.8 represent the 90% credible interval for the forecasted demand.

The demand on 18 June 2013 at 18h00 was forecasted to lie between 26 316 kW (logdemand = 4.4202) and 39 496 kW (logdemand = 4.5966) with a 90% probability. The actual demand was 35 393 kW (logdemand = 4.5489) while the forecasted demand at the 50 th percentile was 32 126 kW (logdemand = 4.5065). The blue graph represents the forecasted hourly electricity demand at the 50 th percentile of the demand distribution while the red circles represent the actual hourly demand for electricity. The black graphs represent the lower and upper 90% credible intervals for the hourly electricity demand.

By comparing the density functions of electricity demand over the years, insight into expected shifts in demand patterns can be obtained, that is, whether the distribution of elec-

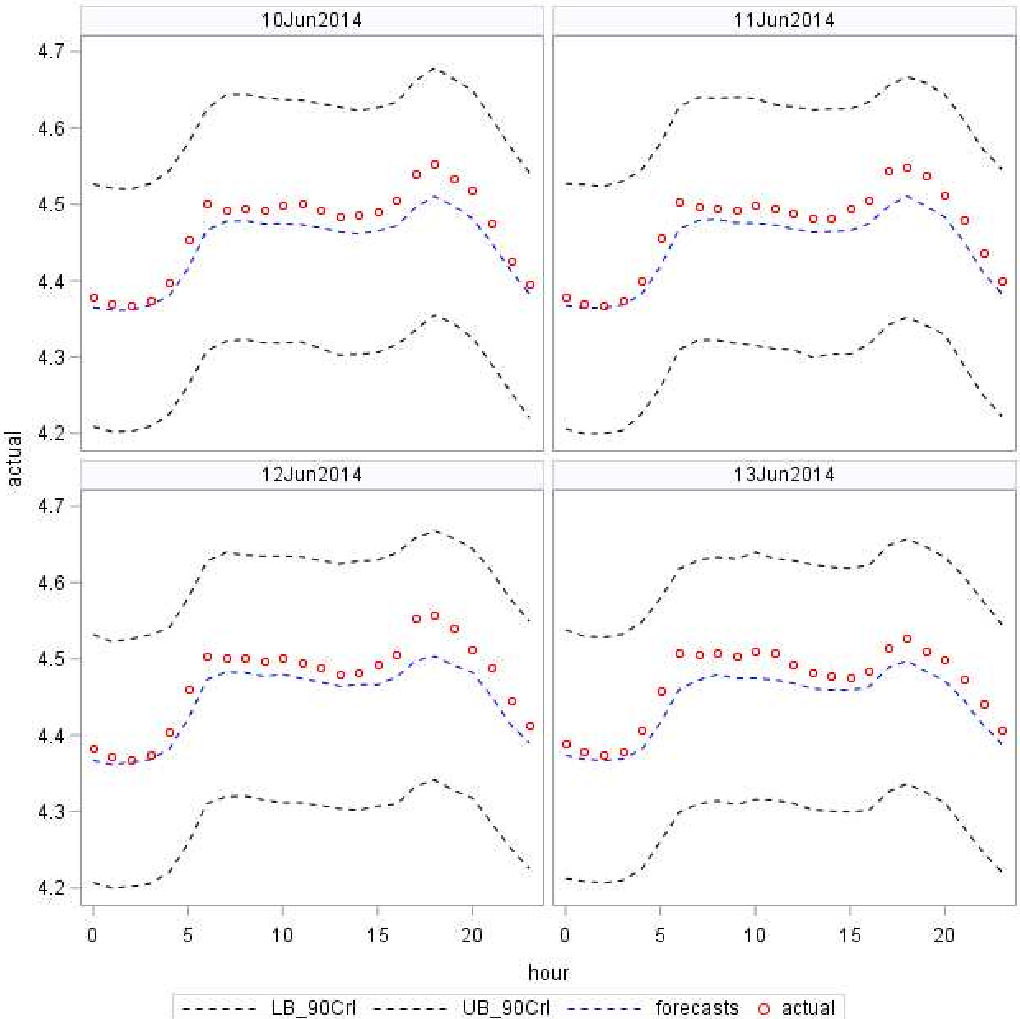


Figure 4.7: Comparison of actual and forecasted hourly demand at 50 th percentile with 90% credible interval: 2014 actual hourly demand against forecasted

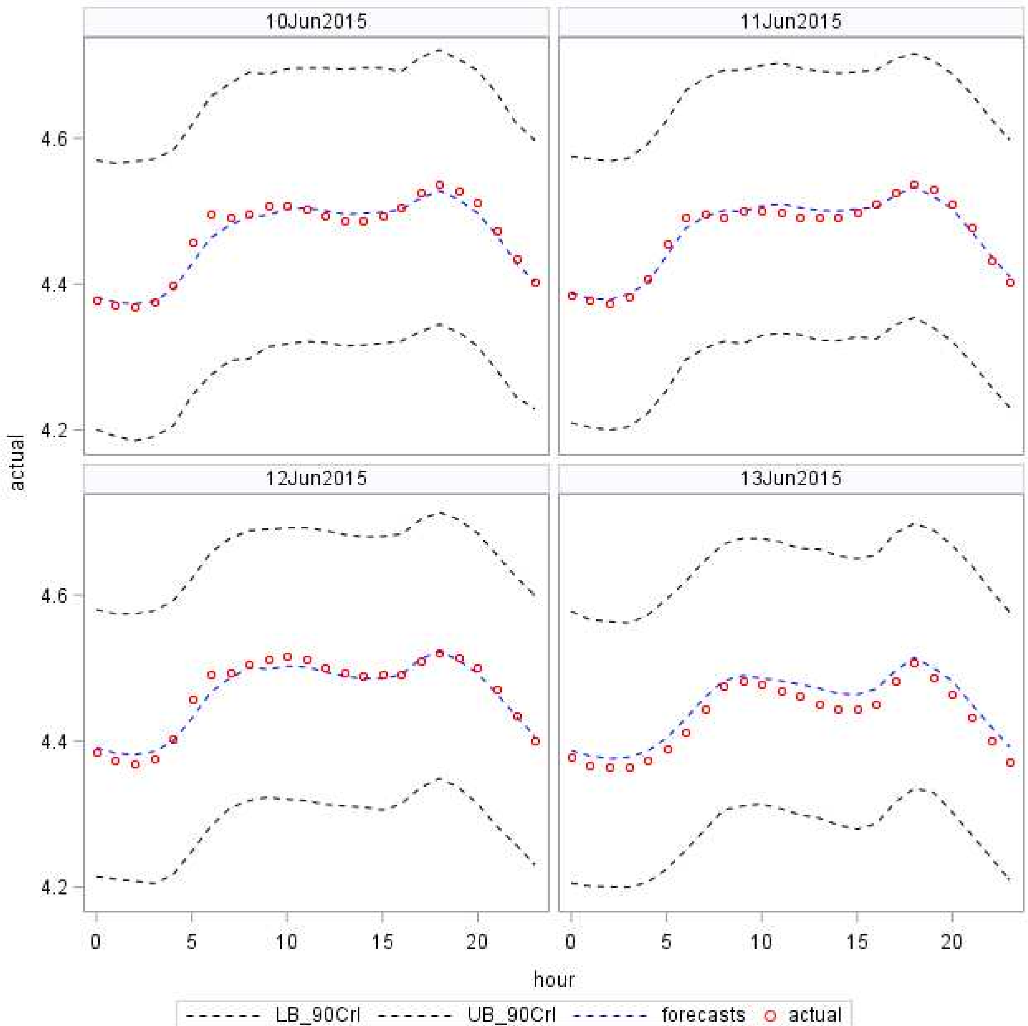


Figure 4.8: Comparison of actual and forecasted hourly demand at 50 th percentile with 90% credible interval: 2015 actual hourly demand against forecasted

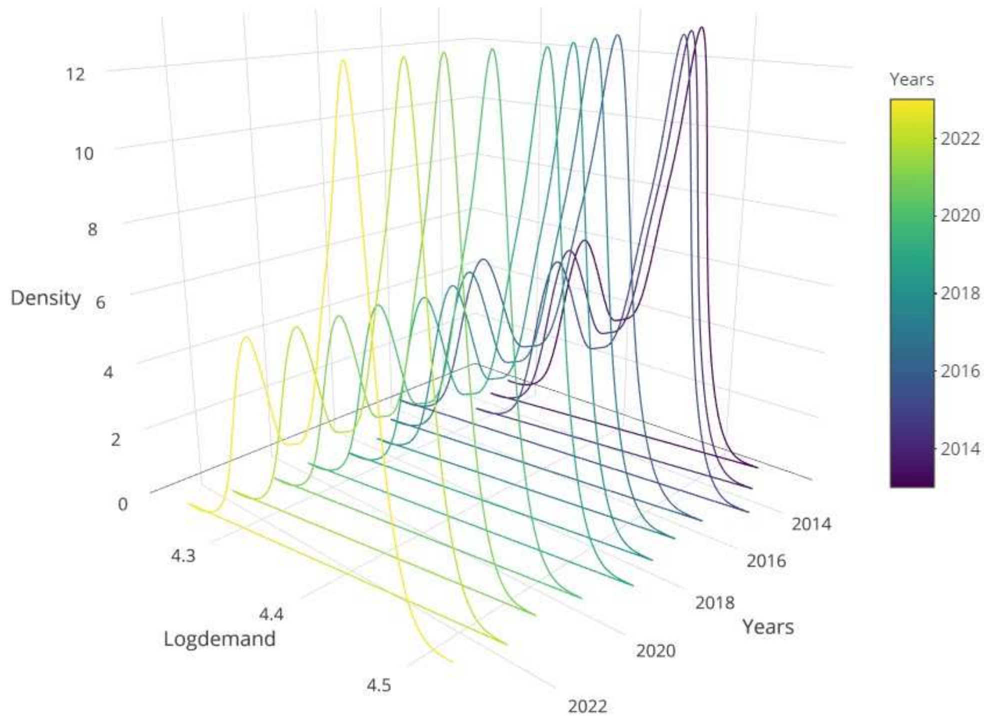


Figure 4.9: Comparisons of density functions of electricity demand over years

tricity demand in Figure 4.9 is expected to shift towards higher or lower demand over the years until 2023. The forecasted distribution of electricity demand in Figure 4.9 for the period investigated suggest that the demand is likely to shift towards lower demand levels over the years until 2023. There are some signs of electricity demand shifting towards lower demand in the historical data in the last four years until 2015 which could be attributed to the growing use of renewable sources of electricity in South Africa, the sluggish economic growth, the steep increases in electricity tariffs in the recent past and the market penetration of energy efficient appliances and the transition of the South African economy from energy-intensive "manufacturing" and "mining" to a less energy-intensive service oriented economy among others as indicated in Chapter 1 and Chapter 3.

The density functions in Figure 4.9 could also be used to calculate the probabilities of exceeding certain demand levels by integrating the density functions. For example, it would be possible to calculate the probability of the hourly demand exceeding the highest historical hourly demand in future and, therefore, to determine if Eskom will be able to continue meeting future demand with the existing generation infrastructure.

The annual peak electricity demand is very important for planning purposes, as this represents the maximum that would need to be supplied in an hour, and if the power generating company could meet the daily peak hourly demand, it could meet any hourly demand. Figures 4.10 and 4.11 suggest that the distributions of the morning and afternoon peak electricity demand are more likely to shift towards lower demand levels over the years until

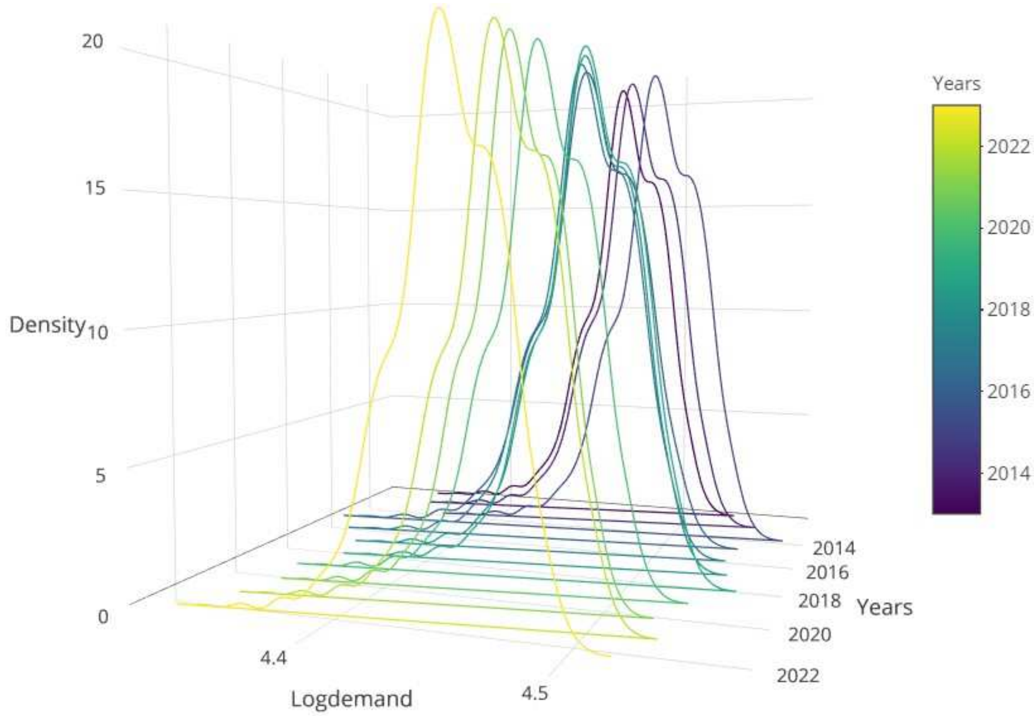


Figure 4.10: Afternoon peak density functions of electricity demand from 2013 to 2023

2023.

Figure 4.12 represents the full forecasted distribution of the hourly electricity demand in South Africa with its 90% credible interval. The actual hourly demand from 2007 to 2012 (the training data) are represented by the black graph; while the blue graph represents the forecasted posterior mean of the distribution of hourly demand between 2013 and 2023. The interval represented by the green graphs is the 90% credible interval of the forecasted hourly demand.

Figure 4.12 shows that the demand during the period between 2013 and 2023 will lie between 15 849 kW (logdemand = 4.2) and 39 810 kW (logdemand = 4.9) with a 90% probability. The blue graph in Figure 4.12 indicates that the hourly demand on average is forecasted to lie between 18 000 kW and 35 000 kW between 2013 and 2023 in South Africa. Intuitively, uncertainties are expected to increase the further the forecasts are in future years, as can be seen in Figure 4.12 by the widening of the credible intervals (the interval between the green graphs in Figure 4.12). Therefore, for long-term forecasting, the BSTS model captures uncertainties well because the uncertainties are behaving as they are expected to, that is, the further the forecasts are made into the forecasting horizon, the more uncertainty increases.

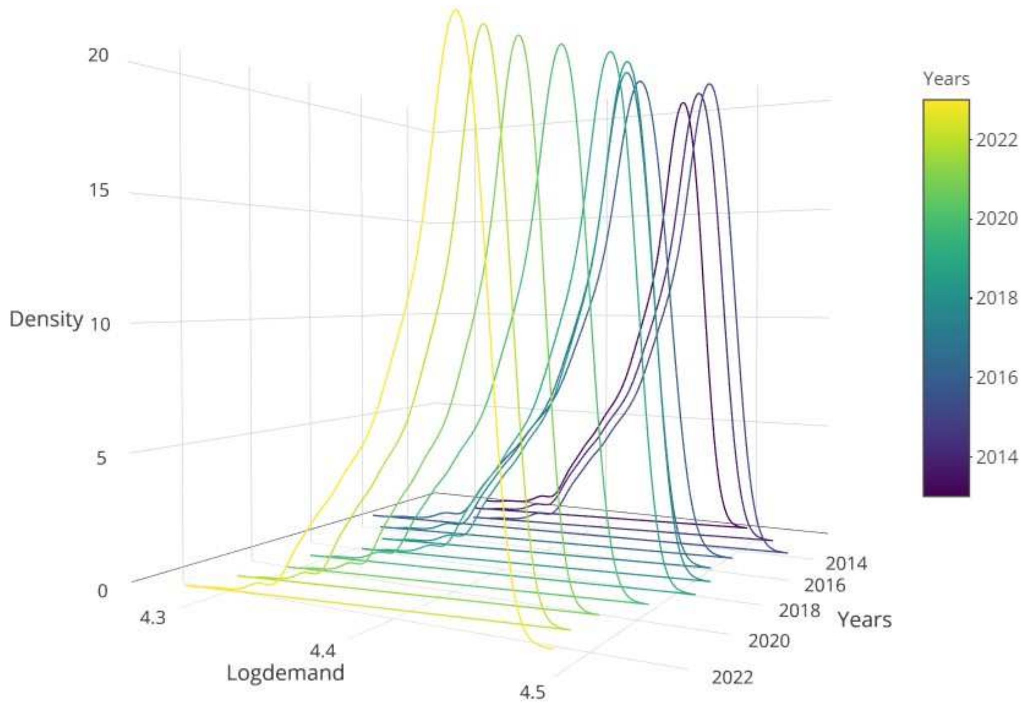


Figure 4.11: Morning peak density functions of electricity demand from 2013 to 2023

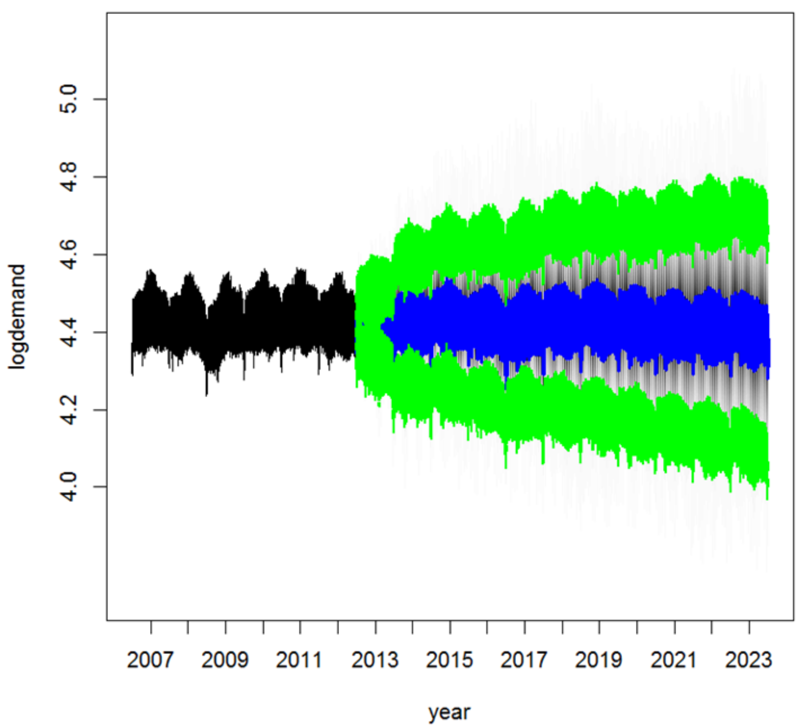


Figure 4.12: Predicted distributions of demand with 90% credible interval

4.5 Concluding remarks on BSTS

The forecasts of the hourly electricity demand showed that it was unlikely for the future demand from Eskom to exceed the highest historical demand of 36 826 kW until 2023. Electricity demand is likely to shift towards smaller demand over the years until 2023 and it is estimated to lie between 15 849 kW (logdemand = 4.2) and 39 810 kW (logdemand = 4.6) during the period between 2013 and 2023 with a 90% probability.

One of the advantages of the BSTS model is that it provides measures to assess the contributions of the individual components (such as trend, seasonal and regression components) of the model. For the South African electricity demand data, it would be ideal for the trend component to contribute less to the overall model as there are uncertainties with regard to whether the trend would continue the apparent decline trajectory as shown in the last four years until 2015, or if it would stabilise, or would revert to the old upward trajectory going forward. If the trend component contributes more to the model and it were captured inaccurately, then the long-term demand for electricity forecasts will vastly deviate from the actual electricity demand and this could have serious planning implications.

The BSTS model uses the spike and slab priors to minimise correlations between variables in the model and this is a powerful way of solving the problem of multicollinearity and ensuring that the resultant model is always parsimonious.

One of the main advantages of the probabilistic forecasts is that the planner could assess the future chances of exceeding the current electricity generation capability of the system. If there are such chances of exceeding the maximum that the system can generate, then the probability of exceeding that maximum can be calculated and an informed planning decision could be made based on the calculated exceedance probabilities.

The contributions of Chapter 4 are:

1. the application of the BSTS model for generating the probabilistic forecasts of the long-term electricity demand in South Africa. The BSTS has not yet been used to forecast the long-term electricity demand to the best of our knowledge,

2. the use of an autocorrelation plot to determine the time lags and the number of time lags in long-term electricity demand forecasting. The literature is silent on how to determine the time lags and the number of time lags to be included in the forecasting model,

3. the long-term forecasting of electricity demand using the South African data with the forecasts, uncertainties quantified, and

4. the inclusion of the local level trend, assumed to follow a random walk.

Both QR and BSTS are parametric modelling approaches because in these modelling frameworks the electricity demand functions are completely defined by the fixed number of model parameters that are known beforehand. On the other hand, the GAMM based model in Chapter 5 is a semi-parametric model because the relationships between electricity demand and its drivers are represented by smooth functions whose shapes are fully determined by the data and also by a set of parameters representing the parametric relationship.

Chapter 5

Generalised Additive Mixed Quantile Averaging Model Approach

Generalised Additive Mixed Quantile Averaging Model Approach[1]

5.1 Background

The literature on statistical models used for forecasting electricity demand is dominated by parametric modelling approaches. In parametric modelling, the model is completely specified by a fixed number of and typically a small set, of parameters. The literature is limited on the application of non-parametric modeling for forecasting the long-term electricity demand. In non-parametric modelling, the relationship between the outcome variable and covariates is defined by functions whose shapes are fully determined by the data, and the number of parameters is determined by the size of the data. A semi-parametric model is a hybrid between parametric and non-parametric models. Semi-parametric models have been applied in electricity demand (Farland, 2014; Hyndman and Fan, 2010; Ruppert et al., 2003 and Sigauke, 2017).

A generalised additive model (GAM) is in the family of semi-parametric models. The difference between GAM and linear regression models is that GAM allows for smooth terms alongside parametric terms and hence it is appropriate for fitting nonlinear relationships between outcome and predictor variables. Nonlinear relationship between outcome and predictor variables could be modelled through generalised linear regression (GLM) but the functional form of the relationship must then be determined beforehand. Sigauke (2017) indicated that GAM is classified as a statistical learning technique in forecasting electricity demand that it has not yet been discussed in literature in South Africa. Hastie and Tibshirani (1986) defined GAM as a generalised linear model with a linear predictor involving a

[1]A modified version of this chapter is published in the Scopus 2018 proceedings of the International Conference on Industrial Engineering and Operations Management (IEOM). It received the “Best Track Paper Award” in the Energy track at the 2018 First African International Conference for Industrial Engineering and Operations Management.

sum of smooth functions of covariates. A generalised additive mixed model (GAMM) is a GAM with a mixed part. GAMMs incorporate random effects alongside parametric and smooth terms and therefore it is an extension of the GAM model.

The GAMM model in this study is used to flexibly model the nonlinear relationships between electricity demand and the drivers of electricity demand and to harness the within hour correlation structures.

The GAMM model is characterised by the following:

- it relaxes the assumption of linearity between predictors and a response variable,
- the relationship between electricity demand and some drivers of electricity demand may be nonlinear,
- there is no need in GAMM to determine the functional form of the relationship beforehand, like in the case of GLM, if the relationship is not linear,
- categorical predictors and interactions can be included in the GAMM model, since it uses other distributions than normal for the response, and
- GAMM uses mixed approaches to include autocorrelation estimates or hierarchical sampling structures.

5.2 Methodology

GAMM is an additive and a functional modelling technique where the impact of predictive variables is not only captured through parametric effects but also through smooth functions which could be nonlinear. The goal of GAMM in this study is to model electricity demand (outcome variable) using its drivers (covariates), some of which are expressed in parametric form and others in a non-parametric form represented by smooth functions (splines). This means that the dependence of electricity demand on covariates are specified in terms of smooth functions and also in terms of parametric relationships. Smoothing splines are real functions that are piecewise-defined by polynomial functions called basis functions. The places where the polynomial pieces connect are called knots. In GAMM, penalised regression splines are used in order to regularise the smoothness of a spline. The GAMM model is generally written as in equation (5.1):

$$y_{th} = \beta_{0h} + \sum_{j=1}^{p} \beta_{hj} x_{thj} + \sum_{j=1}^{p} s_{hj}(x_{thj}) + z_{th} b_t + u_{th} \tag{5.1}$$

where y_{th} is the electricity demand on day t at hour h, $\beta_{hj} X_{thj}$ is the linear part of the model, β_{hj} is the corresponding fixed parameter, β_{0h} is the constant parameter of the linear component of the model, s_{hj} are smooth functions of covariates x_{thj}; z_{th} is a row of random effects model matrix, b_t is a vector of random effects coefficients and u_{th} are residuals.

$t = 1, 2, \ldots, n$, $h = 1, 2, \ldots, 24$ and $j = 1, 2, \ldots, p$ and $x_{th1}, x_{th2}, \ldots, x_{thp}$ are p covariates. Each smooth, $s_{hj}(x_{thj})$ is treated as having a parametric (unpenalised) component, which can be absorbed into the linear part of the model, and a random effects (penalised) component, which can be absorbed into $z_{th}b_t$. The random effects component of the smoothing function has an associated Gaussian distribution assumption.

The variables "Month", "day of week" and "day of year" were included in the model as smooth functions while the variables "fribtwn", "monbtwn" and "lngwknd" in Table 5.1 were included as part of the linear component of the model together with the trend variables. The GAMM in equation (5.1) was estimated using penalised cubic splines (Goude et al., 2014 and Wood, 2006).

$$\min_{s_{hj}} \sum_{t=1}^{n} [y_{th} - \beta_{0h} - \sum_{j=1}^{p} \beta_{hj} x_{thj} - \sum_{j=1}^{p} s_{hj}(x_{thj})]^2 + \sum_{j=1}^{p} \lambda_j [\int (f''(x))^2 \mathrm{d}x] \tag{5.2}$$

According to Wood (2017), the degree of smoothness is controlled by the penalty parameter $\Lambda = \Lambda_j$,$j = 1, 2, \ldots, p$, determining the roughness of the function estimate to the data. The generalised cross-validation criterion (GCV) is used to optimise this function. The smooth function s_{hj} is the i th basis function at hour h and β_{hi} are unknown parameters. Therefore s_{hj} can be written as in equation (5.3);

$$s_{hj}(x) = \sum_{i=1}^{q} \beta_{hi} b_{hi}(x) \tag{5.3}$$

where q in equation (5.3) denotes the number of basis functions (dimensions). The GAMM approach assumes that the model errors are identically and independently distributed. This assumption is not fulfilled in the case of time series regression. The present values of time series are serially correlated with past values and hence the errors of the model are also correlated. In such cases, the errors are said to be autocorrelated. This implies that estimated regression coefficients and residuals of the model might be biased, which also implies that the confidence intervals would be incorrect. This problem is avoided by harnessing the correlation structures within hours by using appropriate autoregressive moving average (ARMA) models. Harnessing the correlation structures in this way could improve the forecasts. The correlation structures within hours were found to differ. Some hours have simple correlation structures (autoregressive (AR)) while others have ARMA(p,q) correlation structures. Errors in equation (5.1) are therefore modelled using the ARMA(p,q) given in equation (5.4);

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_{t-p} u_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{5.4}$$

where u_{t-1}, $u_{t-2}, \ldots, u_{t-p}$ are the AR component with coefficients $\phi_1, \phi_2, \ldots, \phi_{t-p}$ and $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are the MA components with coefficients $\theta_1, \theta_2, \ldots, \theta_q$. The p in equation (5.4) represents the number of AR terms while the q represents the number of MA terms.

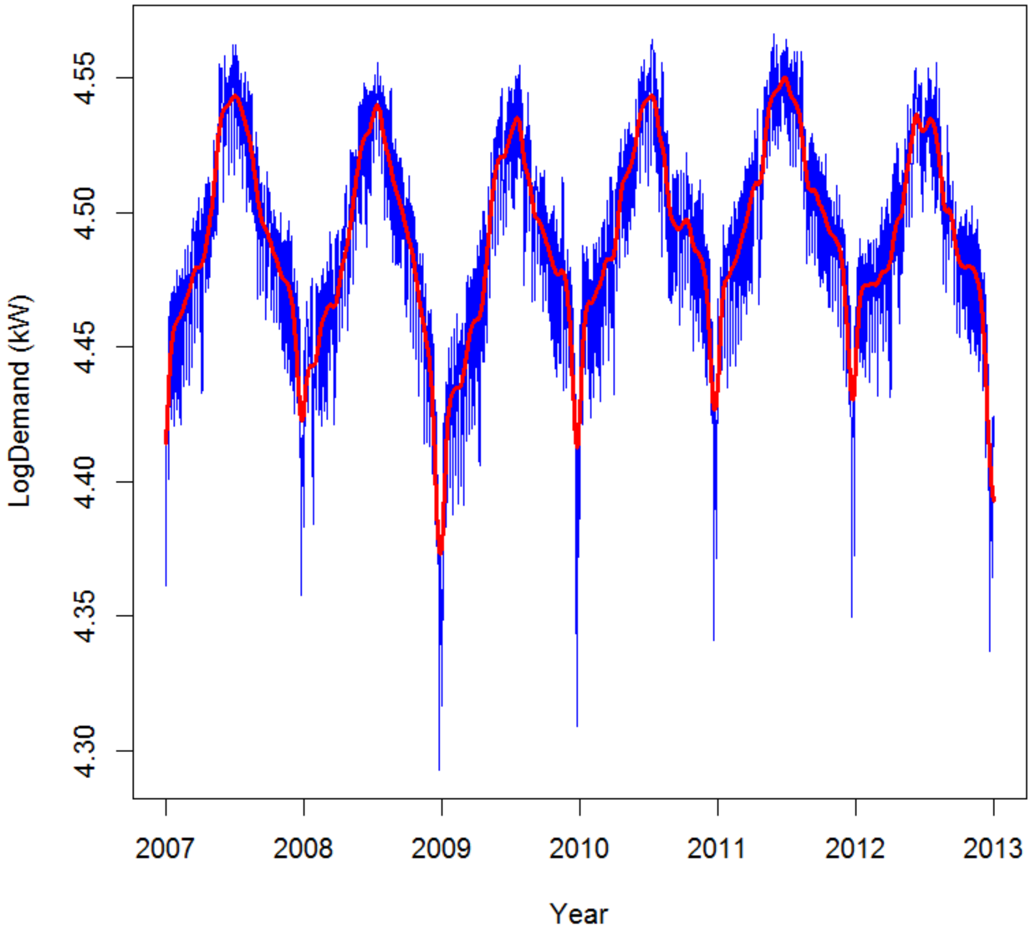


Figure 5.1: The nonlinear trend fitted to the demand data at hour 18

The GAMM model was extended by including a fitted trend to the parametric component of the model. A nonlinear trend was fitted to the demand for electricity data as shown in Figure 5.1 (demand for electricity at hour 18h00). The red line represents the fitted nonlinear demand trend between 2007 and 2012. The fitted values were then extracted to form the derived variable referred to as "trendfitted". The extraction of the fitted values to form the derived "trendfitted" was done for all other hours as well, since the GAMM model was fitted for each hour separately.

The challenge was that only trend values up to 2015 could be extracted to derive the "trendfitted" variable as the electricity demand data were not available beyond 2015. Therefore in order to forecast electricity demand beyond 2015, the "trendfitted" values needed to be forecasted. The future values of "trendfitted" were forecasted from 2013 to 2023 using various time-related variables such as "day", "public holidays", "months", "weekends" and "December break", using quantile regression. The trend was forecasted at 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95 quantiles of its distribution. The derived variables from the forecasted trend at different quantiles of its distribution were referred to as "predquant5", "predquant10", "predquant25", "predquant50","predquant75","predquant90", "predquant95" respectively. The prediction of the future values of the covariates to be used to forecast the response variable have been done before. For example, Hyndman and Fan (2010) used bootstrapping method to forecast the future values of temperature in order to use it as one of the covariates in the model that they used for forecasting the long-term peak demand for electricity.

The GAMM model was then fitted to predict the hourly demand for electricity using calendar variables and the derived trend variables as drivers of demand for electricity. The correlation structures were harnessed in this model using the AR and ARMA processes. A separate model for each hour was fitted and this is consistent with approaches of other researchers including Fan and Chen (2006); Fay et al. (2003); Hyndman and Fan (2010) and Ramanathan et al. (1997) who considered methods that fit separate models to the data from each half-hourly period in order to forecast future demand for electricity. Their argument is that, the demand patterns change throughout the day, and therefore, better estimates can be obtained if each half-hourly period is treated separately. Soares and Medeiros (2005) argue that considering separate models for each hour of the day avoids modelling complicated intraday patterns in the hourly load, which is commonly called the load profile. In this study the assumption is made that demand is driven by electricity influencing activities at specific hours across days, for example, at around 18h00 when people usually return home, some of the common activities include, cooking, watching television, bathing and ironing clothes for the next day. These activities are carried out every day at around the same time, hence the within hour correlation is expected to be stronger than the between hour correlation, therefore, modelling hours separately is likely to give better forecasts especially if within hour correlation structures are harnessed. The fitted model was referred to as a "GAMM with trend".

Liu et al. (2017) fitted a forecasting model from the averaging of point forecasts from sister regression models, and this concept was extended to GAMM in this study. Therefore, the simple GAMM model with no trend referred to as a "GAMM" and "GAMM with trend" models were fitted and averaged. The forecasts of the developed "GAMM with trend" and the "GAMM" models were combined to develop a model referred to as a "GAMM quantile averaging" model (GAMMQV). The GAMMQV model is given in equation (5.5):

$$y_i = \beta_0 + \beta_1 \text{GAMM} + \beta_2 \text{GAMM}_{\text{trend}} + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{5.5}$$

where GAMM are the forecasts from the GAMM model, $\text{GAMM}_{\text{trend}}$ are forecasts from the GAMM model with trend, and β_1 and β_2 are the corresponding regression coefficients. The τth quantile of ε_i is assumed to be zero and the corresponding quantile regression model is given in equation (5.6);

$$Q_Y(\tau|\text{GAMM}, \text{GAMM}_{\text{trend}}) = \beta_0 + \beta_1\text{GAMM} + \beta_2\text{GAMM}_{\text{trend}} + \varepsilon_i \textit{ and } i = 1, 2, \ldots, n \tag{5.6}$$

where

$$Q_Y(\tau|\text{GAMM}, \text{GAMM}_{\text{trend}}) = \inf[y : F_\tau(y|\text{GAMM}, \text{GAMM}_{\text{trend}}) \geq \tau] \tag{5.7}$$

in equation (5.7) is the conditional τth quantile of the response (y_i) given the covariate (GAMM and $\text{GAMM}_{\text{trend}}$), and Q_τ $(Y|\text{GAMM}, \text{GAMM}_{\text{trend}})$ is a non-decreasing function of τ for any given covariate. β is the vector of parameters and represents the marginal change in the quantile because of the marginal change in covariate.

In estimating the QR model for a given quantile, the ideas of Koenker (2005) and Yue and Rue (2011) who used the standard approach of Koenker and Bassett Jr (1978) to estimate their QR model were used. QR minimises the tilted absolute function $\rho_\tau(.)$, which they called the check-function Maistre et al. (2017), which asymmetrically weighs residuals from the model to a degree that depends upon τ.

$$\rho_\tau(\varepsilon) = \begin{cases} (1-\tau)\varepsilon & \text{if } \varepsilon < 0, \quad 0 < \tau < 1 \\ \tau\varepsilon & \text{if } \varepsilon \geq 0 \end{cases}$$

where $\rho_\tau(\varepsilon)$ is a continuous piecewise linear function and non-differentiable at $\varepsilon = 0$ but differentiable everywhere else (Yue and Rue, 2011). This check-function ensures that all ρ_τ are positive and the scale is based on the probability τ. The model parameters are estimated by equation (5.8);

$$\hat{\beta}_\tau = \text{argmin}_{(\beta_\tau \in R^p)} \sum_i^n \rho_\tau(y_i - (\beta_0 + \beta_1\text{GAMM} + \beta_2\text{GAMM}_{\text{trend}})) \tag{5.8}$$

$\rho_\tau(\varepsilon)$ is a continuous piecewise linear function. The detailed formulation of quantile regression is available in Koenker and Bassett Jr (1978).

The objectives of the study were; to forecast the demand for electricity distributions in South Africa over the long-term horizon (10 years), to forecast the distributions of the morning (demand at 06h00, 07h00, 08h00) and afternoon (demand at 18h00, 19h00, 20h00) peak demand over the years in South Africa and to investigate the shifts in the distributions of demand for electricity over the years until 2023.

Table 5.1: Parametric model coefficients for the hour 18 GAMM model

Variables	Estimate	Std Error	t value	$Pr(>\|t\|)$
Intercept	-1.365	0.934	-1.462	0.144
fribtwn	-0.018	0.008	-2.163	$0.031^{\star}$
monbtwn	-0.007	0.006	-1.168	0.243
lngwknd	-0.042	0.003	-13.138	$< 0.001^{\star\star\star}$
predquant5	6.760	0.608	11.103	$< 0.001^{\star\star\star}$
predquant10	-5.877	0.891	-6.595	$< 0.001^{\star\star\star}$
predquant25	-6.424	1.212	-5.298	$< 0.001^{\star\star\star}$
predquant50	3.645	1.795	2.030	$0.042^{\star}$
predquant75	6.574	1.366	4.811	$< 0.001^{\star\star\star}$
predquant90	-5.910	1.844	-3.204	$0.001^{\star\star}$
predquant95	2.532	1.222	1.223	$0.038^{\star}$

$p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

5.3 Results and discussion

5.3.1 Estimated model parameters at 18h00

5.3.1.1 Estimated model parameters for the parametric part at hour 18h00

For discussion purposes, Table 5.1 represents the estimated coefficients of the parametric variables of the GAMMQV model fitted for hour 18. At the 0.05 significance level, the Friday after a holiday variable (Fribtwn) significantly affects electricity demand in South Africa while the long weekend variable (Lngwknd) highly influences electricity demand.

5.3.1.2 Nonparametric estimated model parameters

The estimated degrees of freedom (edf) is the equivalent degrees of freedom for each of the fitted smooth functions of the predictor variables. If q basis functions were used within a smooth function and the maximum likelihood estimation is used, then there would be q degrees of freedom (df) for the smooth function. The penalisation of the smooth function would most of the time reduce the edf to a number smaller than$q$. If the smooth function has one edf for a variable then it means that the relationship between demand for electricity and the predictor variable of interest is represented by a straight line. In Table 5.2 the edf for all predictor variables were greater than one indicating that the fitted smooth functions are not represented by straight lines. The edf for all predictors are equal to the degrees of freedom and they are all significant at the 0.001 significance level.

It can be seen in Figure 5.2 (a) that electricity demand is lower during weekends (days 1 and 7). A Friday before a public holiday and a long weekend behave like Saturday and

Table 5.2: Approximated significance of smooth terms for nonparametric part of hour 18 model

smooth terms	**edf**	**Ref.df**	**F**	**p-value)**
DayOfweek	5.910	5.910	172.890	< 0.001 ***
DayOfyear	7.981	7.981	21.740	< 0.001 ***
Month	10.725	10.725	13.270	< 0.001 ***

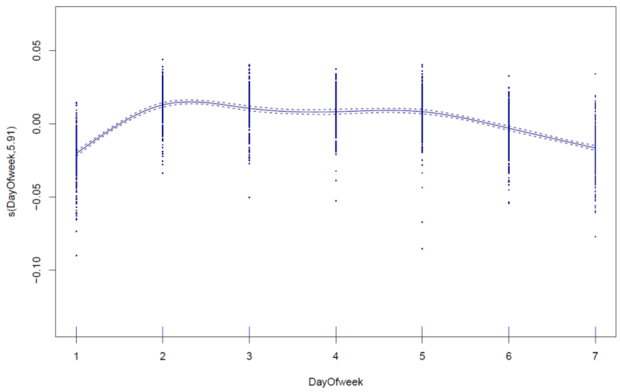
$p > 0.05$; $\quad ^{\star}\ p < 0.05$; $\quad ^{\star\star}\ p < 0.01$; $\quad ^{\star\star\star}\ p < 0.001$

Sunday in terms of demand for electricity in South Africa. A similar analysis was done for the other hours.

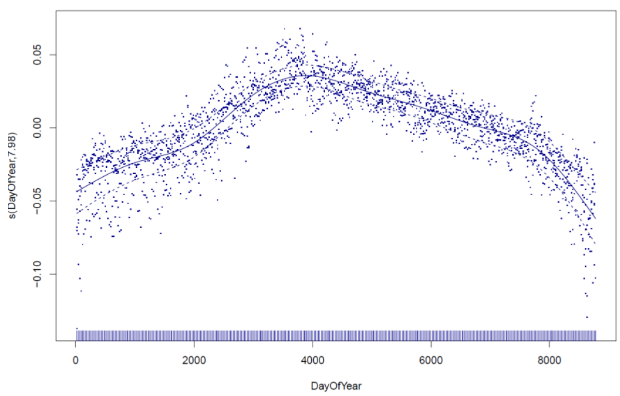
For illustrative purposes, the visual representation of the fitted smooth functions are given in Figure 5.2 which represents the nonlinear relationship between electricity demand and predictor variables at 18h00. Figure 5.2 (a) depicts the plot of smoothed "day of the week" variable and electricity demand. The day *one* in Figure 5.2 (a) represents Sunday, *two* represents Monday and *seven* represents Saturday. The pieces of the function between the days of the week represent the basis functions or splines and they are connected at the knots, which are represented with vertical lines in all graphs in Figure 5.2. Figure 5.2 shows that the relationship between electricity demand and "day of week" in South Africa is nonlinear. Electricity demand in South Africa is lower on weekends (day 1 and day 7). Figure 5.2 (b) shows the plot of the smoothed "day of year" variable and that electricity demand increases significantly during the winter period in South Africa compared with other seasons. Figure 5.2 (c) indicates the non-linear relationship between month and demand that was used in the GAMM model.

5.3.2 Model diagnostics for GAMMQV

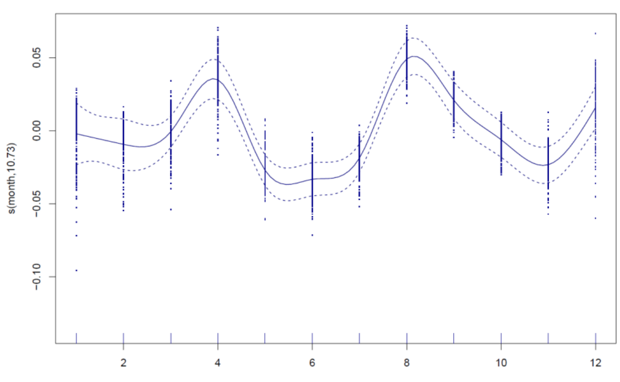
For illustrative purposes, the model diagnostic results for hour 18 (corresponding to 18h00) are given in Figure 5.3. Figure 5.3 shows that a GAMMQV model is a good fit to the data. Figure 5.3 (d) shows a plot of the fitted hourly electricity demand values versus the actual hourly electricity demand values which are scattered around a diagonal line showing a good fit. The clustering of residuals around zero in Figure 5.3 (c) is an indication that the errors are independent of each other and that there is no apparent relationship between them, while Figure 5.3 (b) shows that the distribution of the residuals was skewed to the left. The amount of variation explained by the GAMMQV model is 87.7% ($r^2 = 0.877$).



(a)



(b)



(c)

Figure 5.2: Relationship between electricity demand and its drivers

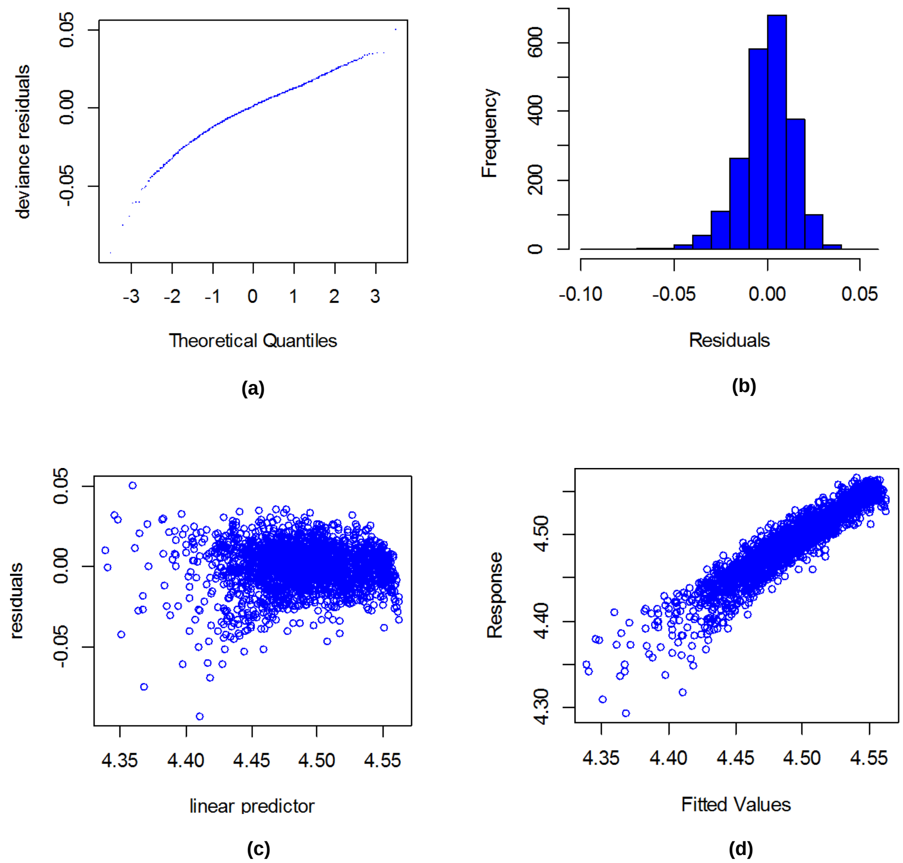


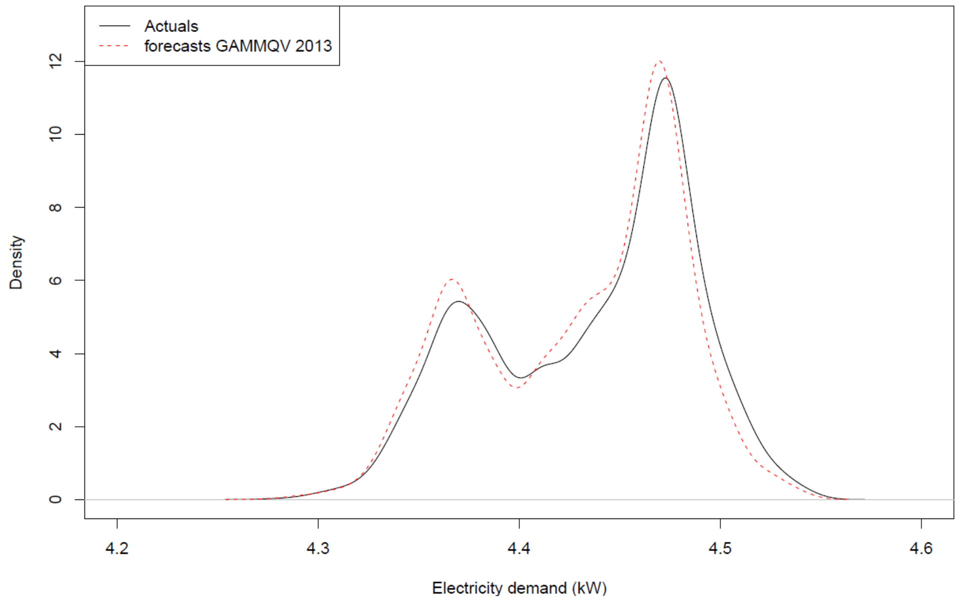
Figure 5.3: Diagnostic plots the hour 18 model: (a) normal quantile−quantile; (b) residuals histogram plot; (c) linear predictor values against residuals; (d) fitted values plotted against response values

5.3.3 GAMMQV assessment and comparisons

The forecasts of electricity demand are validated as follows:

- by comparing the distributions of the actual and the forecasted electricity demand (Figure 5.4),
- by comparing the daily profiles of the actual and forecasted electricity demand (Figure 5.5), and
- by assessing the closeness of the point forecasts of electricity demand to the actual electricity demand using the mean absolute percentage errors (MAPE) in Table 5.3.

The density functions of the electricity demand generated from the GAMMQV are compared to those of the actual electricity demand (Figures 5.4 to 5.6). Figures 5.4, 5.5 and 5.6 show that the density functions of the forecasted demand from the GAMMQV model (red graphs in Figures 5.4, 5.5 and 5.6) closely match the density functions of the actual electricity



(a)

Figure 5.4: Comparisons of actual and forecasted demand densities for 2013

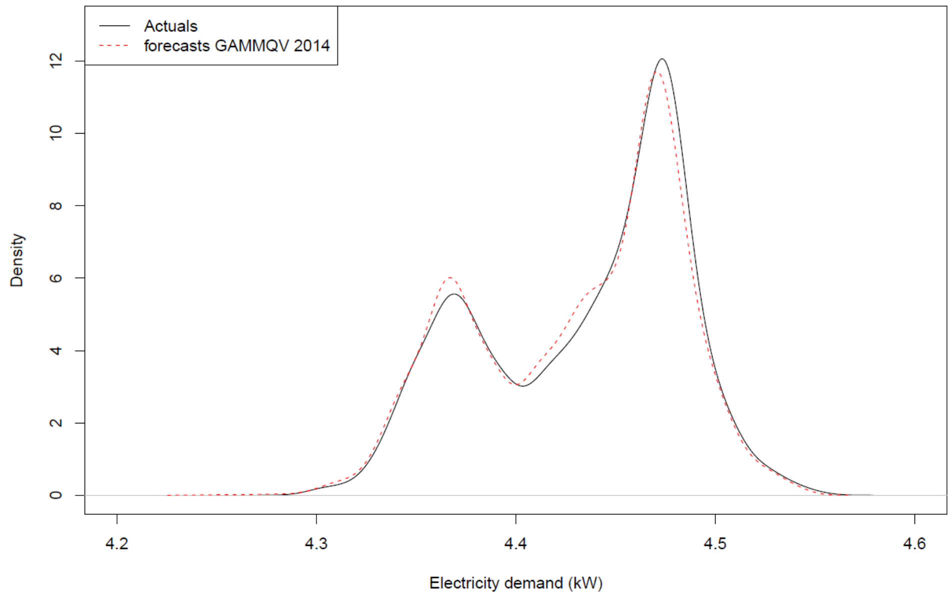
demand in 2013, 2014 and 2015 respectively.

The MAPE between the actual and forecasts at the 50 th percentile of the demand distributions in 2013, 2014 and 2015 were below 4% in all hours, as shown in Table 5.3 indicating that the demand forecasts deviated from the actual demand by less than 4% on average. Lewis (1982) indicates that a MAPE of less than 10% can be classified as highly accurate forecast. The GAMMQV model, therefore, produced accurate forecasts.

The MAPE and density functions in Table 5.3 and in Figures 5.4, 5.5 and 5.6 respectively indicated that the GAMMQV gives robust forecasts. Therefore, there is no reason to doubt that the GAMMQV will continue giving robust long-term forecasts of hourly electricity demand beyond 2015.

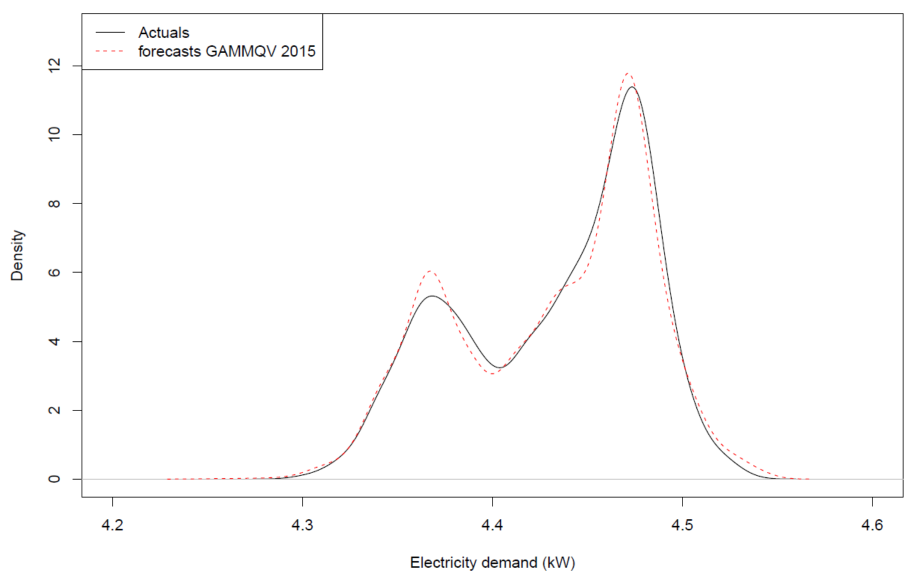
5.3.4 Forecasts

The three year out of sample forecasts were further validated by comparing the actual and the forecasted daily electricity demand profiles. For discussion purposes, and comparison with forecasts from the other modeling approaches in this study, four days in June were selected in such a way that the day with the highest hourly electricity demand of the year was included. The highest electricity demand in 2013 was 35 393 kW on 18 June, it was 36 039 kW in 2014 on 12 June and 34 481 kW in 2015 on 11 June and they were all at 18h00. Figures 5.7 (a), 5.7 (b) and 5.7 (c) show that the forecasted electricity demand (blue line) well captured the actual electricity demand (red circles) in 2013, 2014 and in



(b)

Figure 5.5: Comparisons of actual and forecasted demand densities for 2014

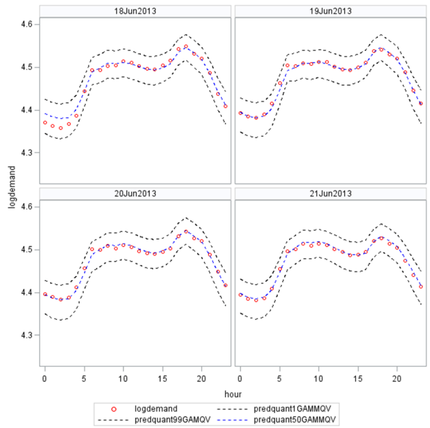


(c)

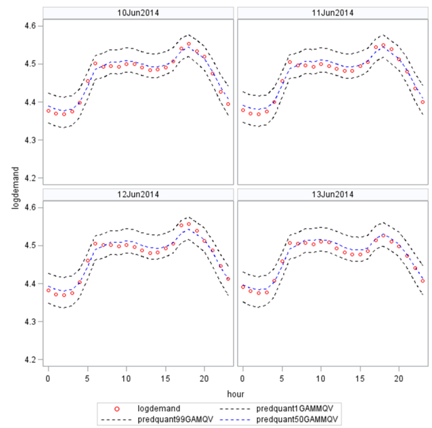
Figure 5.6: Comparisons of actual and forecasted demand densities for 2015

Table 5.3: Mean absolute percentage error (Forecasts at 50 th percentile)

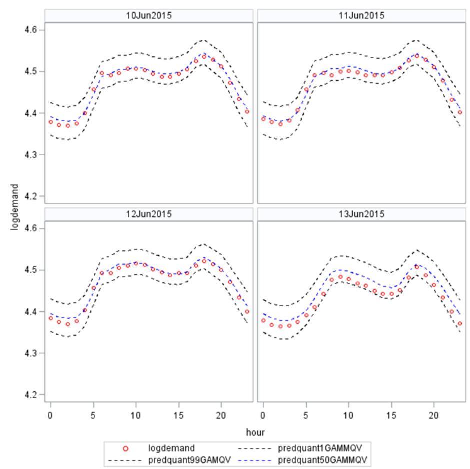
Hours	GAMMQV			
	2013	2014	2015	Average
0	1.706	1.914	2.633	2.084
1	1.701	1.942	2.657	2.100
2	1.688	1.896	2.653	2.079
3	1.901	1.927	2.587	2.138
4	2.294	2.217	2.641	2.384
5	3.183	3.152	3.181	3.172
6	3.146	3.124	3.490	3.254
7	2.018	2.172	2.886	2.359
8	1.742	2.029	2.681	2.151
9	1.791	2.031	2.284	2.035
10	1.826	1.969	2.507	2.100
11	1.893	2.022	2.592	2.169
12	1.934	2.041	2.608	2.194
13	2.028	2.102	2.583	2.238
14	2.156	2.207	2.646	2.337
15	2.381	2.226	2.649	2.419
16	2.513	2.308	2.709	2.510
17	2.397	2.510	3.013	2.640
18	2.111	2.677	2.777	2.522
19	1.872	2.304	2.504	2.227
20	1.900	1.946	2.925	2.257
21	1.806	1.806	2.907	2.173
22	1.664	1.835	2.718	2.072
23	1.726	1.873	2.659	2.086
Average	2.057	2.176	2.729	2.321



(a)



(b)



(c)

Figure 5.7: Comparisons of actual and forecasted daily profiles: (a) 2013 actual hourly demand against forecasted; (b) 2014 actual hourly demand against forecasted; (c) 2015 actual hourly demand against forecasted

2015 respectively. The black lines in Figure 5.7 represent electricity demand forecasts at the 1 st and 99 th percentiles of the electricity demand distribution.

Considering the electricity demand forecasts on 18 June 2013, at 18h00, the forecasted demand was 34 602 kW (logdemand = 4.5391) while the actual demand was 35 393 kW (logdemand = 4.5489). The forecasted electricity demand at the 1 st percentile was 32 839 kW (logdemand = 4.51639) while the forecast at the 99 th percentile was 37 717 kW (logdemand = 4.57654). Therefore, electricity demand at 18h00 on 18 June 2013 is predicted by the model to fall between 32 839 kW and 37 717 kW with a 98% probability.

The probability of electricity demand exceeding the highest hourly demand (of 36 826 kW) since 1997 was 0.15 for all the years starting from 2013 until 2023 (Table 5.4). The elec-

Table 5.4: Forecasts at percentile points of the demand distribution (2013-2023)

Years	30 perc	40 perc	50 perc	60 perc	70 perc	80 perc	85 perc
2013	34 973	35 232	35 472	35 731	36 031	36 380	36 583
2014	35 101	35 336	35 567	35 824	36 146	36 526	36 754
2015	35 126	35 361	35 592	35 850	36 171	36 551	36 779
2016	35 112	35 349	35 581	35 839	36 159	36 535	36 761
2017	35 114	35 350	35 582	35 839	36 160	36 538	36 765
2018	35 154	35 383	35 612	35 869	36 196	36 584	36 819
2019	35 235	35 445	35 668	35 924	36 267	36 678	36 931
2020	35 099	35 336	35 568	35 826	36 146	36 523	36 750
2021	35 090	35 329	35 563	35 820	36 138	36 513	36 737
2022	35 108	35 343	35 575	35 832	36 154	36 533	36 761
2023	35 105	35 340	35 572	35 829	36 151	36 531	36 759

tricity demand at the 85 th percentile in 2023 is 36 759 kW which implies that 85% of the forecasted electricity demand will lie below this value while only 15% will fall above it. The probability of electricity demand exceeding the highest demand since 1997 is below 0.15 in all the years except in 2019 (Table 5.4). The electricity demand of 36 826 kW falls in the top 15 th percentile of the distribution of electricity demand. Therefore, the GAMMQV indicates that the probability of exceeding the highest hourly electricity demand ever assured by Eskom since 2013 was not very likely to be reached again before 2023.

5.3.5 Comparison of electricity demand distributions over years

Electricity demand for each hour was forecasted at various quantiles of the demand distribution using the fitted GAMMQV model. The forecasted values were used to generate the density functions in Figure 5.8, Figure 5.9 and Figure 5.10. By comparing the electricity demand density functions over the years, insight into the expected shifts in electricity demand patterns can be obtained, that is, whether the distribution of electricity demand in Figure 5.8 is expected to shift towards higher or lower demand. The forecasted distribution of electricity demand in Figure 5.8 obtained from the GAMMQV model for the period investigated suggest that electricity demand from Eskom is likely to shift towards lower demand over the years until 2023.

As highlighted in previous chapters, the daily peak of electricity demand is very important for planning purposes, as this represents the maximum that would need to be supplied in an hour, and if the power generating company could meet the daily peak hourly electricity demand, it could meet any hourly electricity demand. The distribution of the morning and afternoon peak electricity demand suggest a shift towards lower peak demand over the years until 2023 as indicated in Figures 5.9 and 5.10.

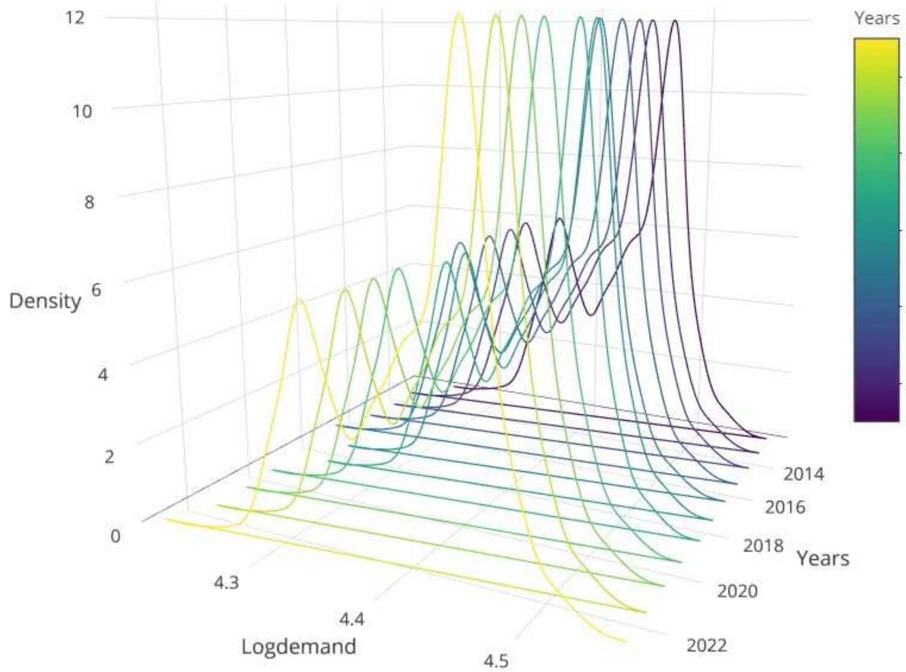


Figure 5.8: Demand distribution between 2013 and 2023

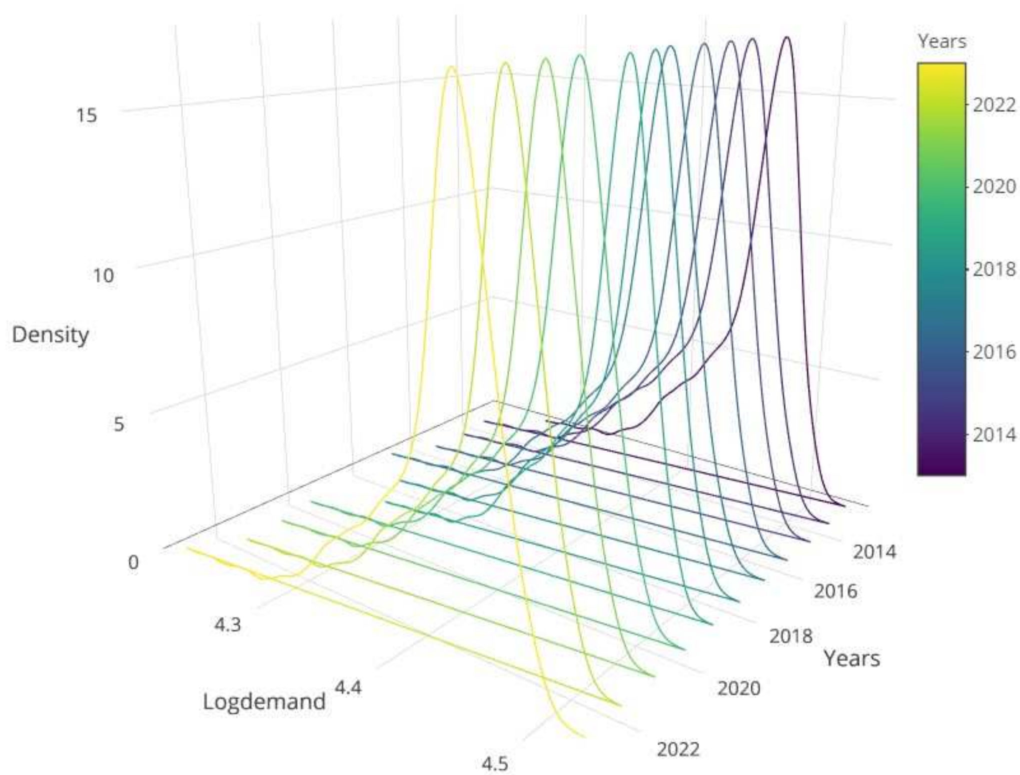


Figure 5.9: Morning peak demand distribution between 2013 and 2023

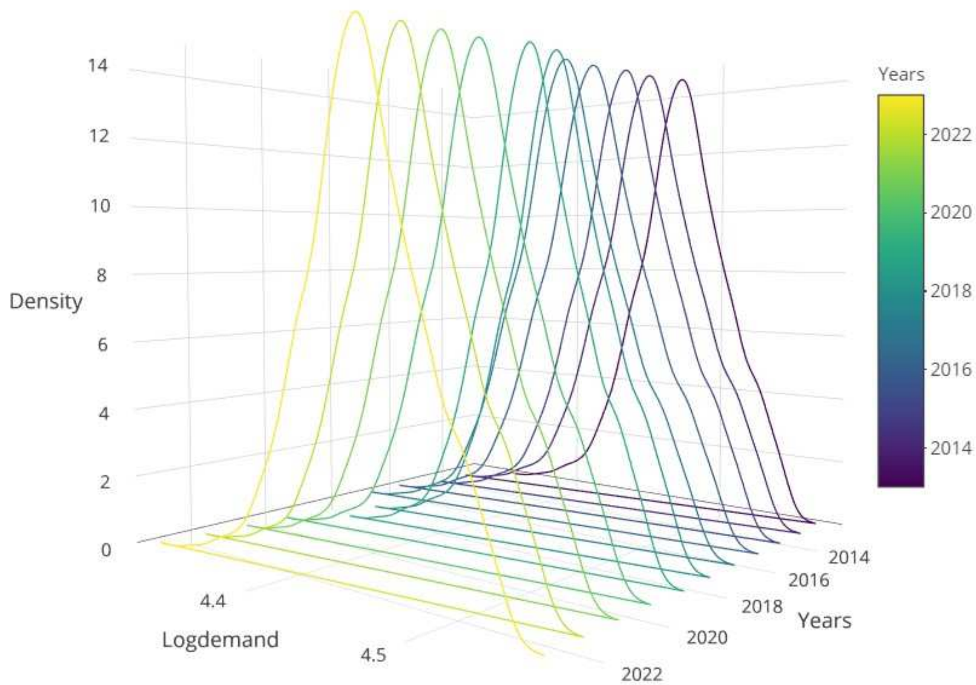


Figure 5.10: Afternoon peak demand distribution between 2013 and 2023

5.4 Concluding remarks on GAMMQV

The forecasted distribution of electricity demand closely matched that of the actual electricity demand between 2013 and 2015; therefore, there is no reason to doubt that the GAMMQV model will continue to forecast electricity demand accurately beyond 2015. The GAMMQV model indicates that the distribution of electricity demand from Eskom is expected to shift towards lower demand over the years. The model indicates that the distributions of both morning and afternoon peak electricity demand are also expected to shift towards lower demand in future. A decline in electricity demand is apparent in the last four years until 2015, which could be attributed to the growing renewable sources of electricity in South Africa, the sluggish economic growth, the steep increases in electricity tariffs and the market penetration of energy efficient appliances, among others.

The forecasted daily profiles from the GAMMQV model accurately captured the actual daily profiles in the long-term. The MAPE shows that the GAMMQV model gives good point forecasts as well. The probabilities of exceeding certain electricity demand levels can be obtained from the quantile forecasts of the GAMMQV model. The trend values beyond 2012 were forecasted at 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95 quantiles of the distribution.

Chapter 5 contributed the following:

1. the GAMM was extended to a GAMMQV for probabilistic forecasting of the long-

term electricity demand in South Africa,

2. the within hour correlation structures were harnessed in forecasting the long-term hourly electricity demand,

3. South African data was used to generate probabilistic forecasts of the long-term electricity demand, and

4. a nonlinear trend was included in the model with its future evolution forecasted using QR.

The synthesis in Chapter 6 deals with the comparisons of the models considered in this study. The main concepts used in this study are discussed in detail in Chapter 6. The choice of a model over other models was considered as a challenge because GAMMQV is better than other models in terms of predicting electricity demand at the 50 th percentile of the demand distribution with narrow prediction intervals, but BSTS predicts demand at the 50 th percentile well with wide prediction intervals. Therefore, to avoid choosing one model over others which could be equally good or even better, model averaging is used, and this is discussed in Chapter 6.

Chapter 6

Synthesis

6.1 Background

Forecasts of the long-term electricity demand are mainly used for capital planning and could also be used for maintenance planning. Such planning decisions are always made under uncertainties about future changes. As indicated in the introduction, the decisions made from the long-term forecasts could have far reaching consequences because decisions for infrastructure expansion could result in the construction of unnecessary power generation facilities, while decisions not to expand on the current infrastructure could result in failure to meet future electricity demand, which could have serious consequences on the economy and result in unintended job losses, among others. Electricity demand forecasts can either be in point or probabilistic form. However, since demand planning decisions are always made under uncertainties, forecasts of the long-term electricity demand must ideally be probabilistic.

6.1.1 Types of forecasts

Point forecasts of electricity demand can be defined as the forecasts of the mean or median of the distribution of electricity demand. In point forecasting, a single forecast is made at each step in the forecast horizon, for example, in forecasting the daily electricity demand at an hourly level, 24 forecasts are made for each hour of the day. Uncertainties are therefore not quantifiable in point forecasting since just one forecast is made at each step within the forecast horizon. Point forecasts can be made with confidence intervals which are often too narrow to fully capture uncertainties emanating from the economic growth, the market penetration of electric vehicles, the increasing cost of electricity and the market penetration of electricity saving appliances, among others in the long-term. They are more useful for short-term forecasting.

Probabilistic forecasts of electricity demand are forecasts of the full distribution of electricity demand at each step of the forecast horizon. They can be in the form of prediction intervals, quantiles, fan charts, scenarios or density forecasts. Uncertainties are quantifiable in probabilistic forecasting because the range of possible electricity demand values are forecasted at each step in the forecasting horizon, that is, a range of forecasts are made at each step, such as an hour. Such probabilistic forecasts are more useful in medium and long-term forecasting.

6.1.2 Forecasting horizon

Short-term forecasts of electricity demand are mainly used to manage the stability of the electricity system. They are used for planning the day-to-day running of the electricity generation system. The quicker it takes to fit the model and to generate accurate results, the better the model is because in short-term demand forecasting, quick decisions are required to be made, for example, the amount of electricity to be generated to meet demand in the next hour. In short-term forecasting, probabilistic forecasts are mostly used for predicting peak electricity demand over a short period. The forecast horizon for the short-term forecasting ranges from five minutes to one week ahead.

Medium-term forecasts of electricity demand are mainly used for maintenance planning. The forecasts are used to schedule the maintenance of power stations in such a way that overall electricity demand is still met during the maintenance period. In medium-term forecasting, probabilistic forecasts of electricity demand are useful. The forecast horizon for medium-term ranges from one week to six months ahead.

Long-term forecasts of electricity demand are mainly used for capital planning. These forecasts are made to assess whether the current generation infrastructure will generate enough electricity to meet future demand in the long-term. The long-term forecasts assist planners to make informed decisions about the possible need for infrastructure expansions. It is therefore important that the distribution of the long-term electricity demand are forecasted with some acceptable levels of accuracy, and ideally with probabilistic forecasts. The long-term forecast horizon ranges from six months to years ahead.

6.1.3 Modeling framework

The two distinct statistical modelling paradigms (the frequentist and Bayesian approaches) are considered in this study. The frequentist approach assumes that the model parameters are fixed and the data are random. In this framework, the most likely values of the model parameters are the ones minimising residuals.

The Bayesian approach assumes that the model parameters are random and data are fixed. In this modelling framework, the equilibrium posterior distribution which is believed to be the same or closer to the actual posterior distribution of the electricity demand is obtained.

The model parameters of the obtained posterior distribution are estimated with their full distributions. The electricity demand can be forecasted with their intervals in both frequentist and Bayesian approaches.

6.1.4 Confidence, prediction and credible intervals

The confidence interval (CI) is an interval associated with model parameters and it is a frequentist concept. Therefore, model parameters are assumed to be non-random but unknown, and the CI is computed from the data. The CI is random because the data are random hence the interpretation of the 95% CI means that 95% of the CI are expected to include the estimated model parameter. According to Carstens et al. (2018) one of the disadvantages of the point forecasts is that the CI either contains the value, or does not and the probability is either zero or one. The values within the CI is therefore not associated with a probability distribution. The CI is calculated as:

$$\hat{y} \pm t_{n-2} s_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}} \tag{6.1}$$

where x_i is the i th independent variable, $\bar{x}^2$ is the mean of the independent variable, y is the electricity demand, $\hat{y}$ is the fitted value of electricity demand and s_y is the standard deviation, which is calculated as;

$$s_y = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}} \tag{6.2}$$

A prediction interval (PI) on the other hand is an interval associated with a random variable yet to be observed, with a specified probability of the random variable lying within the interval. The PI can arise in Bayesian or frequentist statistics. In frequentist statistics, the PI can be obtained from QR or scenarios or can be calculated using equation (6.3).

$$\hat{y} \pm t_{n-2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}} \tag{6.3}$$

The PI is always wider than the CI as shown in equations (6.1) and (6.3). It is noted that equation (6.1) is similar to equation (6.3), except that the variability is higher in equation (6.3) since there is an added one in equation (6.3).

A Bayesian CI, also known as a credible interval, is an interval associated with the posterior distribution of the model parameter. In the Bayesian perspective, model parameters are treated as random variables, and they have probability distributions. Thus a Bayesian CI is like a PI, but associated with a parameter rather than an observation. In this study, the PI in the frequentist approach is done using quantiles, while credible intervals are used in the Bayesian approach.

6.1.5 Models considered in this study

In a QR model, a regression model is fitted at every specified quantile of the distribution of demand and forecasts are made at each specified quantile. Therefore, the forecasts at each quantile represent the mean of electricity demand falling at that quantile. The prediction interval (that is, the intervals between the 1 st and 99 th percentile of the demand distribution) are narrow in QR. Some of the extreme electricity demand values are likely to fall outside the prediction intervals, especially in the long-term. The selection of the predictor variables in QR is a challenge because for a predictor variable to be insignificant, it must be insignificant at all quantiles of the demand distribution. This is because electricity demand is forecasted at every quantile of the distribution of electricity demand.

In the BSTS modelling framework, the model parameters are considered random and they are therefore estimated with their full distributions. The local level trend which is assumed to follow a random walk is used in the fitted BSTS model. Electricity demand is forecasted with wide credible intervals. The extreme electricity demand values always fall within the credible intervals. The model is simple to fit but it takes a long time to converge. The variables are selected into the model via the spike and slab prior.

A GAMM approach incorporates random effects, parametric and smooth terms. The impact of the predictor variables in GAMM models are not only captured through fixed effects, but also through smooth functions which could be nonlinear. The correlation structures inherent within hours can easily be taken into consideration in GAMM modelling approach. In this framework, electricity demand is forecasted at the mean of the demand distribution. The GAMM model was also extended to include a nonlinear trend. The trend was not assumed in the GAMM model but it was fitted and the future evolution of the trend was forecasted. The GAMM and GAMM with trend models were averaged to form GAMMQV. The PI widths are narrow and the modelling takes longer because each of the 24 hours are modelled separately.

6.1.6 Comparisons of models based on their prediction intervals

The PI widths for QR and GAMMQV in this study are taken as the intervals between the 1 st and 99 th quantiles of the distribution for each forecasted value of the electricity demand in South Africa. The PIs for both QR and GAMMQV are narrow compared to the credible intervals of the BSTS model. The electricity demand in QR and GAMMQV is therefore forecasted with narrow PIs and, consequently, some extreme electricity demand values in the long-term are likely to fall outside the PI. For example, based on the GAMMQV model, electricity demand on 18 June 2013 at 18h00 is forecasted to lie between 32 839 kW (logdemand = 4.51639) and 37 717 kW (logdemand = 4.57654) with a 98%

probability (Figure 6.1).

In contrast, the electricity demand forecasts on 18 June 2013 at 18h00 based on BSTS model is expected to lie between 26 316 kW (logdemand = 4.4202) and 39 496 kW (logdemand = 4.5966) with a 90% probability.

The credible interval widths of the BSTS model are taken as the interval between the lower and upper 90% forecasts. This is referred to as the 90% credible interval. The BSTS 90% credible intervals are wider than the 98% PI of the GAMMQV and therefore any extreme demand values are more likely to fall within the credible intervals.

It is logical to believe that the further the forecasts are in future, the more uncertain the forecaster would be about them. Uncertainties in the BSTS model are consistent with this expectation, and therefore, BSTS is an ideal model to generate the long-term probabilistic forecasts because it is able to capture uncertainties (Figure 4.12) while the forecasted electricity demand at the 50 th percentile of the demand distribution are close to the actual electricity demand. Therefore, based on the fact that the forecasted electricity demand at the 50 th percentile of the demand distribution in BSTS are close to the actual electricity demand and the consistency with the expectation that uncertainties grow the further into the future one predicts demand, the BSTS model would be a logical model choice for forecasting the long-term hourly electricity demand in South Africa. The prediction intervals for the BSTS model in Figures 6.1, 6.2 and 6.3 become wider over the years while those of the GAMMQV do not widen the further the forecasts are made into the future.

6.1.7 Comparisons of models based on their point forecasts

The actual and forecasted demand were transformed back from the logarithmic scale to their original scale and MAPE values were then calculated from the data on their original scale. Almost all MAPE values in Table 6.1 for the three considered models are far below 4% for 2013, 2014 and 2015, indicating that the forecasts have not deviated from the actual demand by 4% on average. The 2014 and 2015 were abnormal years because Eskom reintroduced load-shedding in early 2014 and the load-shedding happened regularly between November 2014 and September 2015. While the demand data was adjusted by Eskom with an amount estimated to be lost due to load-shedding, the considered models in this study gave robust demand forecasts in this abnormal period. The MAPE for the GAMMQV model are lower than those of the QR and BSTS in almost all hours in 2013, 2014 and 2015 indicating that GAMMQV estimated the actual electricity demand better than the other two models.

Figures 6.4 to 6.6 represent the weeks in 2013, 2014 and 2015 in which the highest hourly electricity demand of the year fell. The three models considered in this study captured the daily profiles of electricity demand well from 2013 to 2015, therefore, it can be assumed that they will continue capturing the daily profiles well beyond 2015. It is evident from the

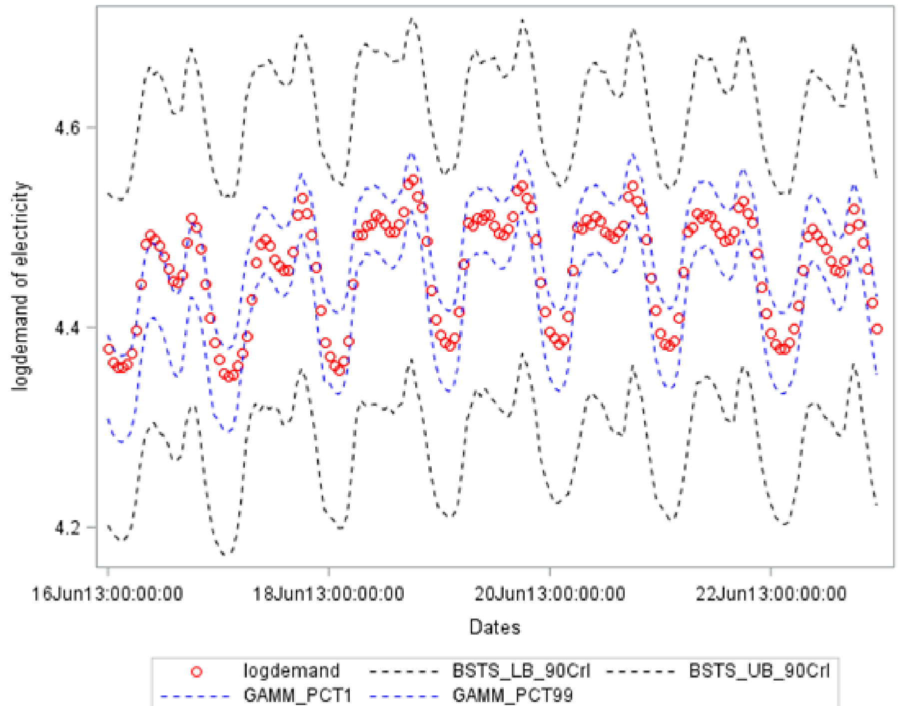


Figure 6.1: Prediction intervals BSTS vs GAMMQV 2013

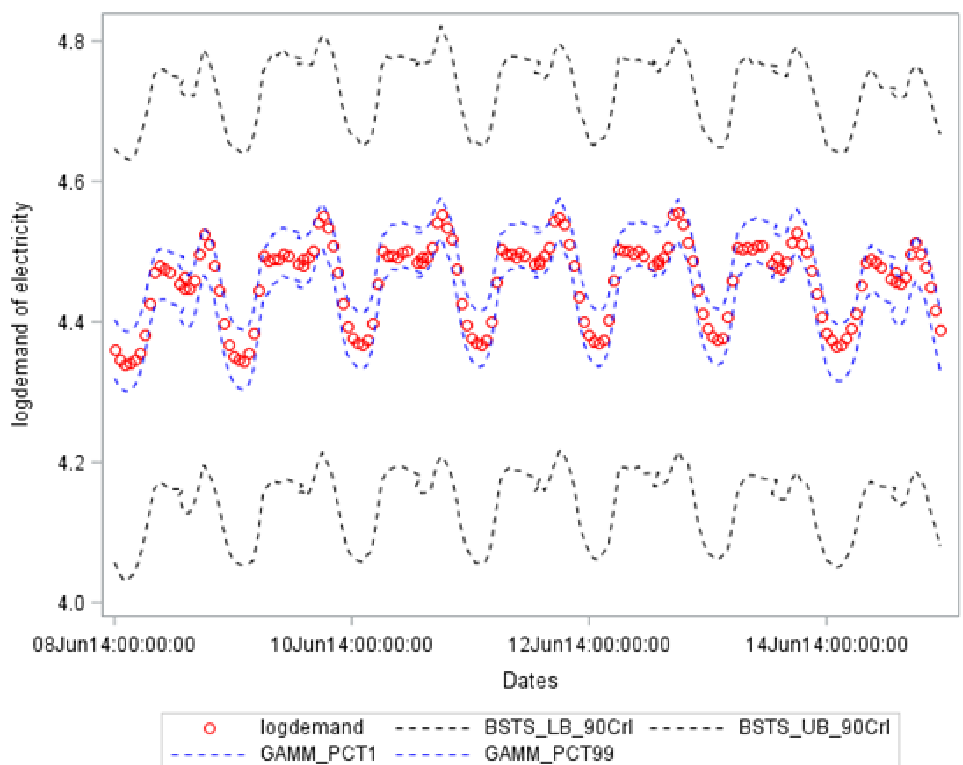


Figure 6.2: Prediction intervals BSTS vs GAMMQV 2014

Table 6.1: Comparisons of models based on their MAPE

Hrs	2013			2014			2015		
	QR	BTST	GQV	QR	BTST	GQV	QR	BTST	GQV
0	2.160	3.714	1.706	2.962	2.190	1.914	3.464	4.093	2.633
1	1.784	3.531	1.701	2.789	2.221	1.942	3.372	4.349	2.657
2	1.688	3.456	1.688	2.698	2.320	1.896	3.353	4.841	2.653
3	1.584	3.603	1.901	2.537	2.324	1.927	3.123	4.702	2.587
4	1.813	4.282	2.294	2.298	2.569	2.217	2.779	3.809	2.641
5	3.781	6.331	3.183	3.816	4.691	3.152	3.916	4.162	3.181
6	4.408	5.970	3.146	4.181	4.699	3.124	4.225	4.722	3.490
7	3.750	4.135	2.018	3.803	2.673	2.172	4.144	4.266	2.886
8	2.751	4.299	1.742	2.771	2.337	2.029	2.955	3.361	2.681
9	2.443	4.303	1.791	2.442	2.315	2.031	2.265	3.036	2.284
10	1.991	4.480	1.826	1.863	2.367	1.969	2.064	3.437	2.507
11	1.826	4.226	1.893	1.838	2.337	2.022	2.219	3.798	2.592
12	1.770	3.927	1.934	2.112	2.362	2.041	2.552	4.340	2.608
13	2.036	3.840	2.028	2.493	2.463	2.102	2.898	4.753	2.583
14	2.048	3.829	2.156	2.578	2.717	2.207	2.986	5.065	2.646
15	1.827	4.127	2.381	2.318	2.881	2.226	2.802	4.921	2.649
16	2.158	4.501	2.513	2.219	3.012	2.308	2.802	4.443	2.709
17	3.837	4.523	2.397	3.218	3.187	2.510	3.740	4.020	3.013
18	5.111	4.099	2.111	4.448	3.333	2.677	4.120	3.399	2.777
19	2.776	3.951	1.872	2.390	2.997	2.304	2.443	3.125	2.504
20	2.400	4.593	1.900	2.004	2.771	1.946	2.385	3.442	2.925
21	2.319	4.653	1.806	2.022	2.479	1.806	2.612	3.845	2.907
22	2.242	4.272	1.664	2.519	2.256	1.835	3.125	3.815	2.718
23	2.168	3.941	1.726	2.874	2.189	1.873	3.340	3.912	2.659

GQV= GAMMQV, Hrs=Hours

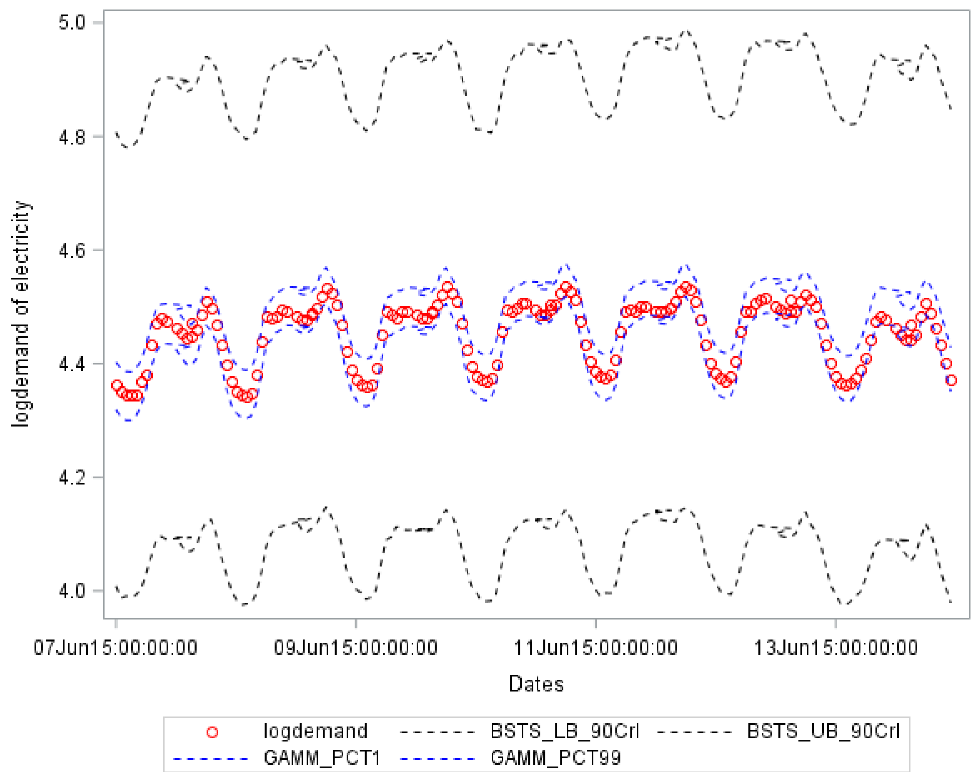
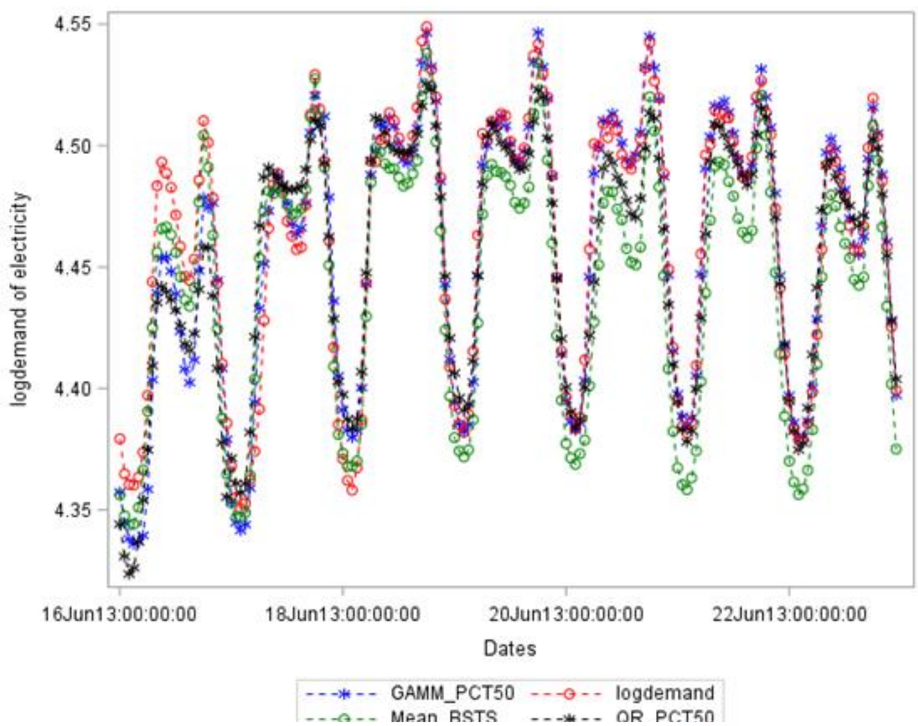


Figure 6.3: Prediction intervals BSTS vs GAMMQV 2015

BSTS model in Figures 6.4 to 6.6 that the model could under-predict demand in one year and over-predict in other years. The GAMMQV model captures the daily profile of electricity demand better than the QR and BSTS models as shown in Figures 6.4 to 6.6. The electricity demand forecasts at the 50 th percentile of the demand distribution from GAMMQV represented by the blue line is closest to the actual demand (red line) in Figures 6.4 to 6.6 indicating that it predicted the hourly electricity demand over this important high-demand period in South Africa better compared to BSTS and QR models. Hence, based on the closeness of the point forecasts to the actual electricity demand and the narrow prediction intervals, the GAMMQV model would be more suitable for forecasting demand, particularly in the short to medium-term horizon in which there is less uncertainty and therefore narrow prediction intervals would be acceptable.

The critical question is, of the three models considered in this study, which one would be recommended for forecasting the long-term electricity demand in South Africa? It has been noted that QR and GAMMQV predicted demand well in 2013, 2014 and 2015 with narrow prediction intervals. The BSTS credible intervals are wider than those of the QR and GAMMQV and they become even wider the further the forecasts are into the future, which is to be expected. It is also noted that one model would accurately predict demand for a certain year where other models would comparably predict poorly and vice versa. The averaging of the models forecasts at the 50 th percentile of the demand distribution using



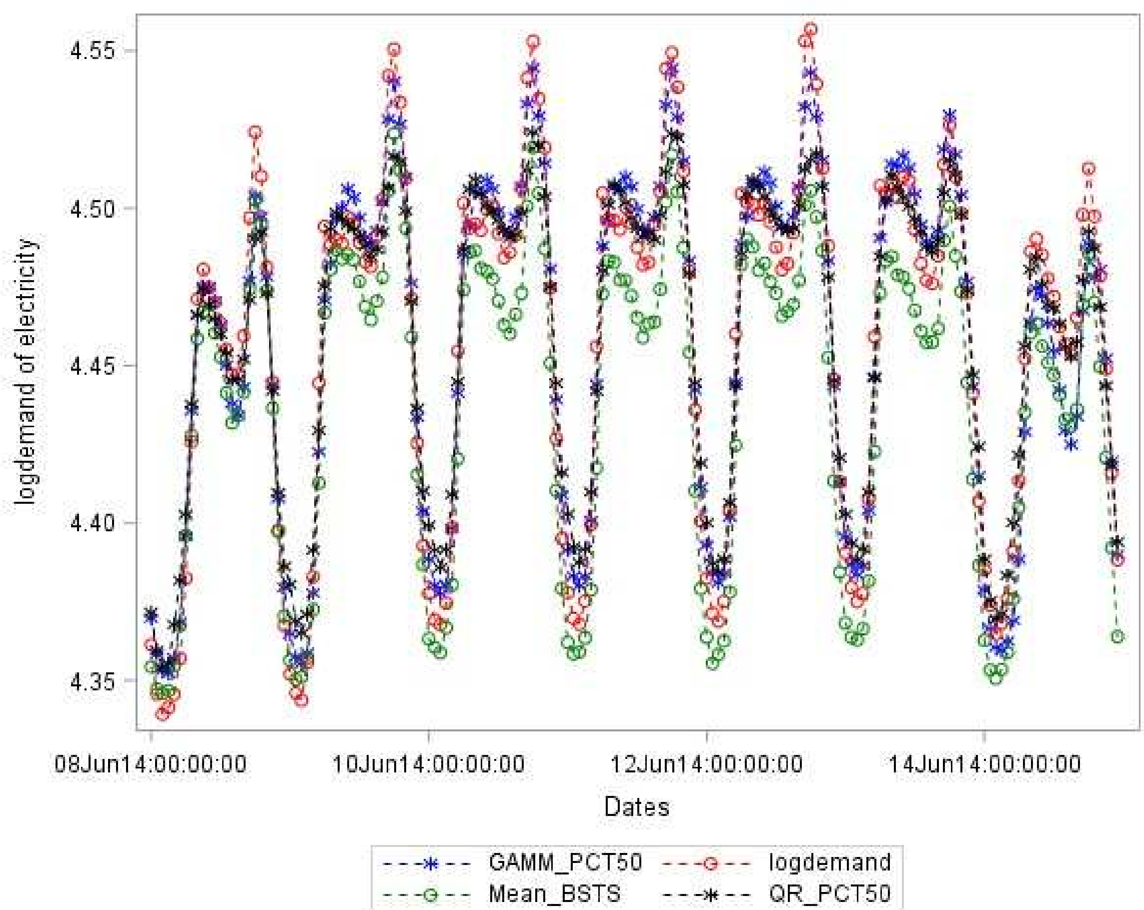


Figure 6.5: Models comparison of point forecasts to actual demand in 2014

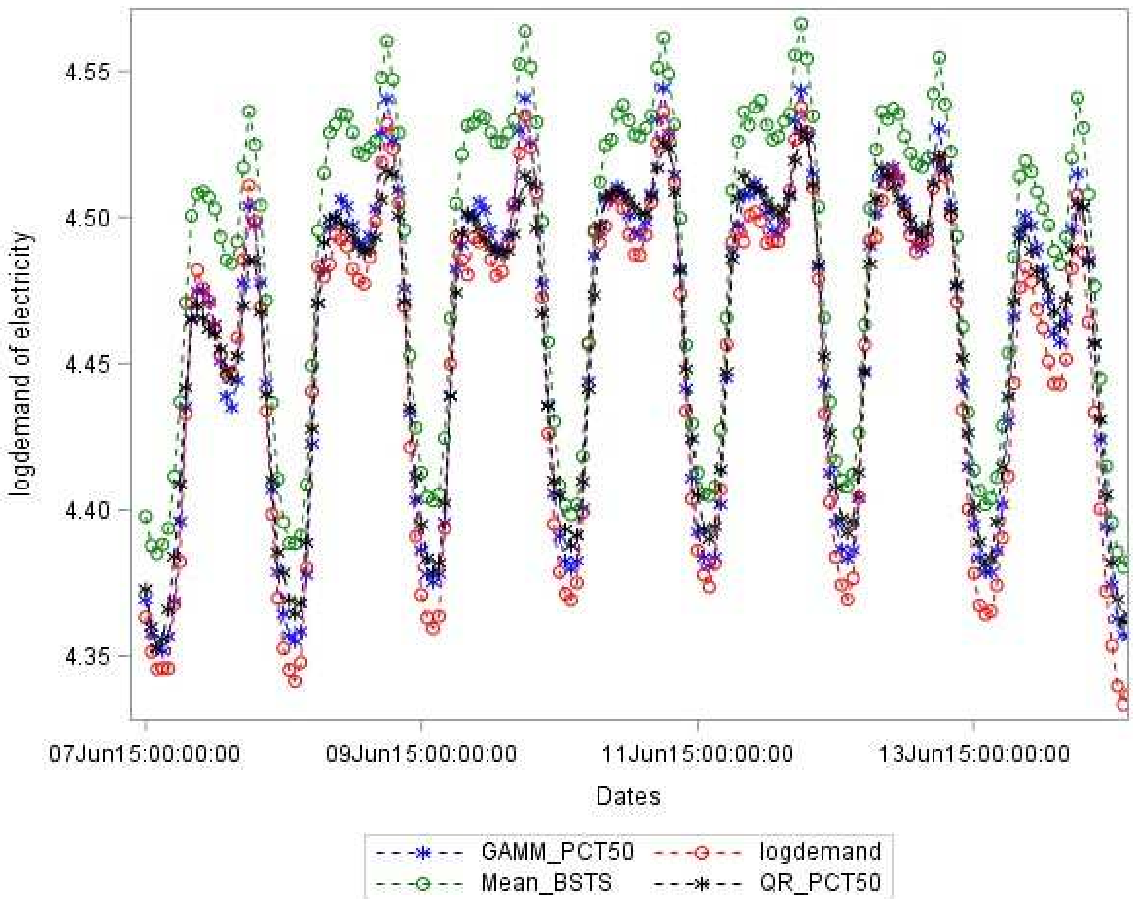


Figure 6.6: Models comparison of point forecasts to actual demand in 2015

QR is explored to avoid making a difficult choice between the models which all predicted well over some periods and poorly over other periods.

6.2 Model averaging using quantile regression

Based on the forecasted electricity demand from 2013 to 2015, the three different models considered in this study gave robust forecasts, but in the long-term horizon, it is difficult to choose one model over others because one model would predict better in one year where others would comparably predict poorly and vise versa, as indicated earlier. The averaging of point forecasts of sister models has been used in forecasting in order to avoid choosing one forecasting model over some possibly equally good models and according to Liu et al. (2017), averaging of the sister models eliminates the need to rely on high quality expert forecasts (point forecasts from a single model), which are rarely achievable in load forecasting practices. The averaging approach helps to develop probabilistic forecasts from points forecasts generated by a family of models. Liu et al. (2017) used the point forecasts generated by the family of regression models which were averaged using quantile regression averaging technique. Ravazzolo (2007) forecasted the financial time series using the averaging of sister point forecasts. The averaging of the point forecasts generated by the sister models approaches have been applied and discussed in literature by Hong and Fan (2016) and Goude et al. (2014). Wang et al. (2019) used a combined probability density model for medium-term load forecasting based on QR, where random forest regression, boosting decision tree and support vector regression models were combined. Our approach to quantile regression averaging is not done from the point forecasts generated by a family of models because the point forecasts are generated by three distinct model types, namely the GAMMQV, BSTS and QR models. Generally, as discussed in Chapter 3, the quantile regression model is given in equation (6.4),

$$Q_Y(\tau|X_t) = {X_t}^T \beta_\tau \tag{6.4}$$

where $Q_y(\tau|.)$ is the conditional τ th quantile of the electricity demand distribution (y_t), X_t are the drivers of electricity demand and β_τ is a vector of parameters for quantile τ. Parameters in QR are estimated by minimising equation (6.5) which is the function for a particular τ th quantile. A QR model is discussed in detail in Chapter 3.

$$Q_Y \beta_\tau = \min_{\beta_\tau \in R^p} \sum_{t:y_t \geq {x_t}^T \beta_\tau}^{n} \tau|y_t - {x_t}^T \beta_\tau| + \sum_{t:y_t < {x_t}^T \beta_\tau}^{n} (1-\tau)|y_t - x_t^T \beta_\tau| \tag{6.5}$$

where y_t is the actual electricity demand and $X_t = [1,\ \hat{y}_{1,t},\ \hat{y}_{2,t}, \ldots,\ \hat{y}_{m,t}]$ is a vector of point forecasts of m individual models where $m = 3$ in this study.

6.2.1 The recommended model

The recommended model is depicted in Figure 6.7 and it involves four stages of modelling. The calendar values, harmonic terms and lagged demand variables are used in the three models considered in this study. The variables used are dichotomous, polytomous or continuous. For the GAMM model, the trend is extracted from the electricity demand data and the future trend values are forecasted at stage one of modelling. At stage two, the GAMM model consisting of parametric, error and non-parametric components is fitted. Two sets of point forecasts are generated from the GAMM with trend and from the GAMM with no trend models. The point forecasts from the two GAMM models are averaged at modelling stage three in Figure 6.7 and a new set of demand forecasts are generated.

The QR model is fitted at modelling stage one and then the QR forecasts at the 50 th percentile are generated. The considered BSTS model has a regression component, therefore the spike and slab prior is used for variable selection at stage one. The BSTS model with time series and regression components is fitted at stage three and the BSTS forecasts at the 50 th percentile of the demand distribution are generated. The point forecasts from the three models are averaged at modelling stage four and the final forecasts at $0.01, 0.02, 0.03, \ldots, 0.99$ quantiles of the demand distribution using QR are generated.

In this proposed model, the forecasts at the 50 th percentile of the demand distributions from QR, GAMMQV and BSTS models are used as the predictor variables in the QR model while the electricity demand is the dependent variable. The long-term electricity demand is then forecasted using QR. This approach to forecasting is referred to as an averaging of point forecasts of sister models using QR and Liu et al. (2017) referred this model to as quantile regression averaging (QRA).

The purple lines in Figures 6.8 to 6.10 represent the forecasted hourly demand at the 50 th percentile of the electricity demand generated by averaging the BSTS, QR and GAMMQV forecasts (the forecasts generated from the model in Figure 6.7). The 2013, 2014 and 2015 demand forecasts from the recommended model in Figure 6.7 are compared to the actual demand and to the forecasted demand from models considered in this study. The red circles represent the actual hourly demand, the blue line represents the forecasts of hourly demand at the 50 th percentile generated by the GAMMQV model. The black line represents the forecasts of hourly demand at the 50 th percentile of the demand distribution generated by the QR model and the green line represents the forecasts at the 50 th percentile of the demand distribution generated by the BSTS model.

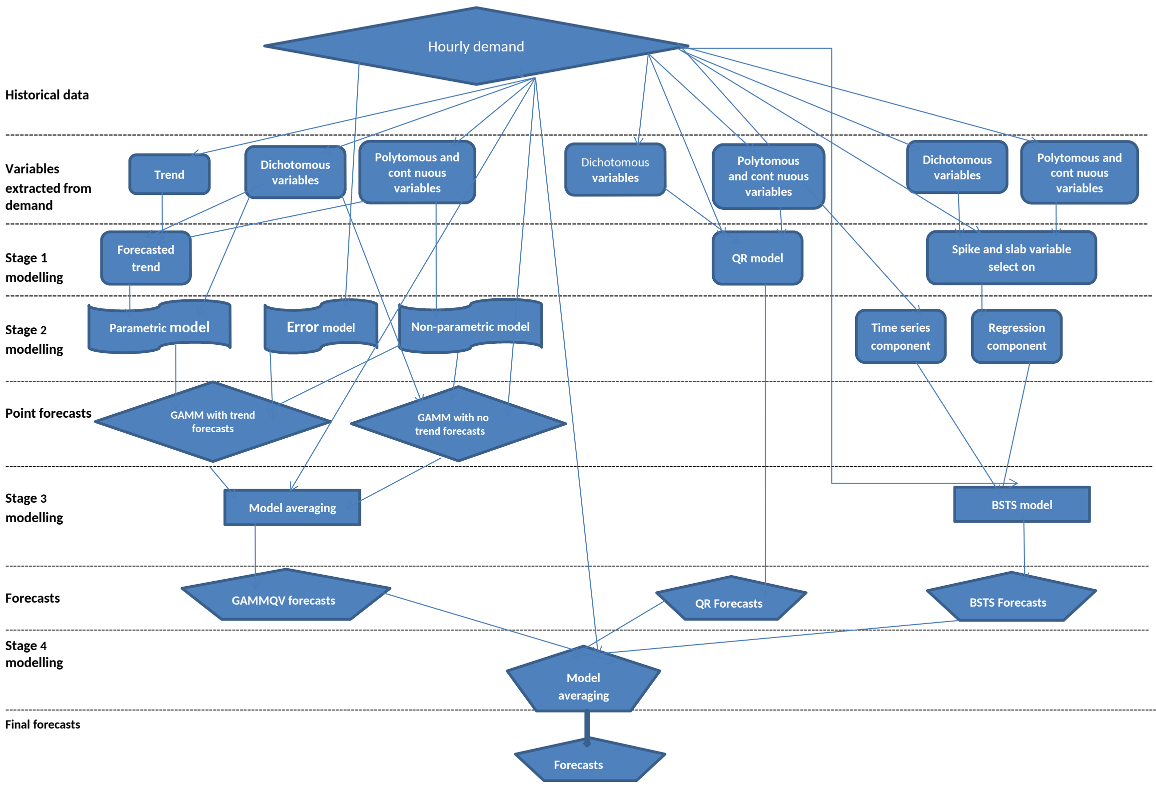


Figure 6.7: Modeling framework of the recommended model

6.2.2 Distributions of demand from 2013 to 2023

The distribution of the forecasted electricity demand obtained from the quantile averaging model for the period investigated suggest that the hourly electricity demand from Eskom is likely to shift towards lower demand over the years until 2023 (Figure 6.11). The distribution of the 2023 electricity demand in Figure 6.11 has shifted more towards lower demand compared to the distributions of demand in other years, hence the probability of demand exceeding 36 294 kW (logdemand = 4.560) is less likely in 2023 than it is in any previous years while 2015 had the highest probability of demand exceeding 36 294 kW (logdemand = 4.560) in the period between 2013 and 2023 as shown in Figure 6.11. The distribution shifts towards lower demand over the years indicating that if demand were met in 2015 it will more likely be met in the later years until 2023 and that the probabilities of not meeting future demand with the current generation infrastructure are slim.

Figures 6.12 and 6.13 are used to investigate the expected future peak demand and their probability of exceedance over the years from 2013 to 2023. The distribution of annual peak demand is very important for planning purposes, as these represent the distribution of the maximum that would need to be supplied in an hour over the years and if the power generating company could meet the daily peak hourly demand, it could meet any hourly demand. Figures 6.12 and 6.13 suggest that the distributions of the morning and afternoon peak demand are more likely to shift towards lower demand over the years until 2023.

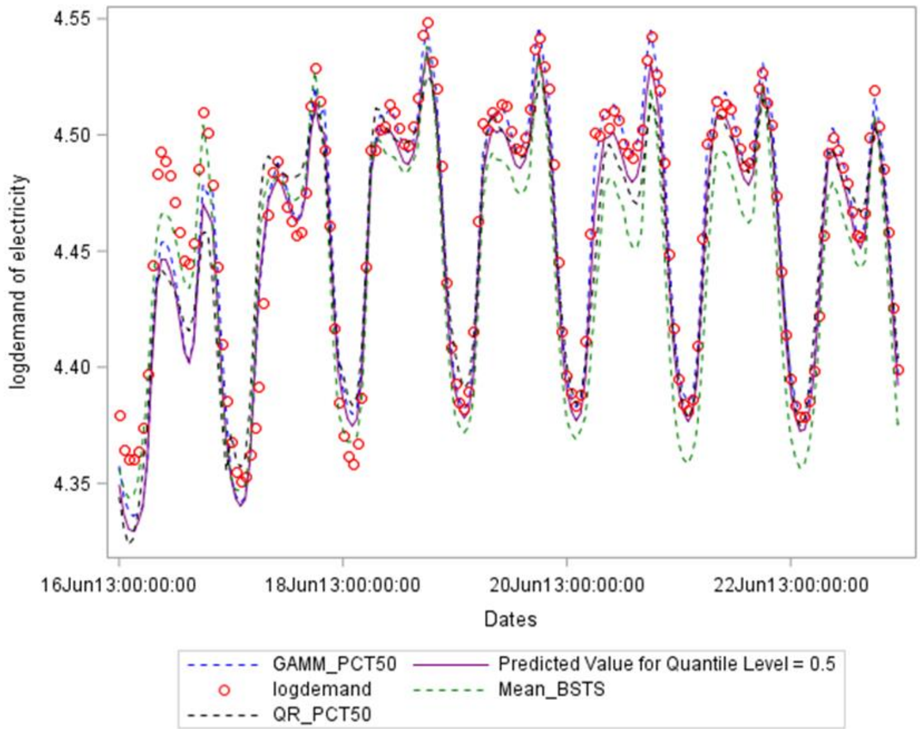


Figure 6.8: Comparison of forecasts from considered models to their averaged forecasts and actual demand in 2013

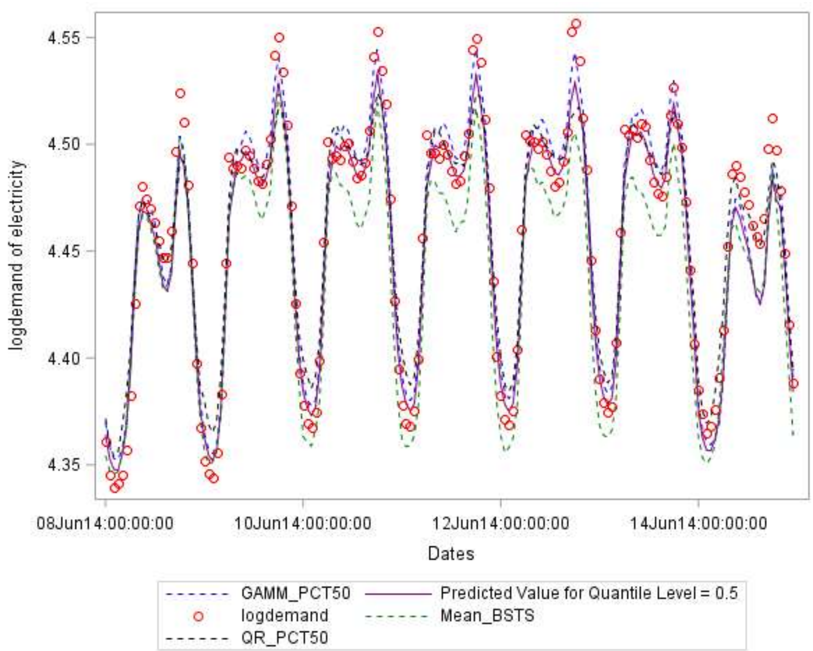


Figure 6.9: Comparison of forecasts from considered models to their averaged forecasts and actual demand in 2014

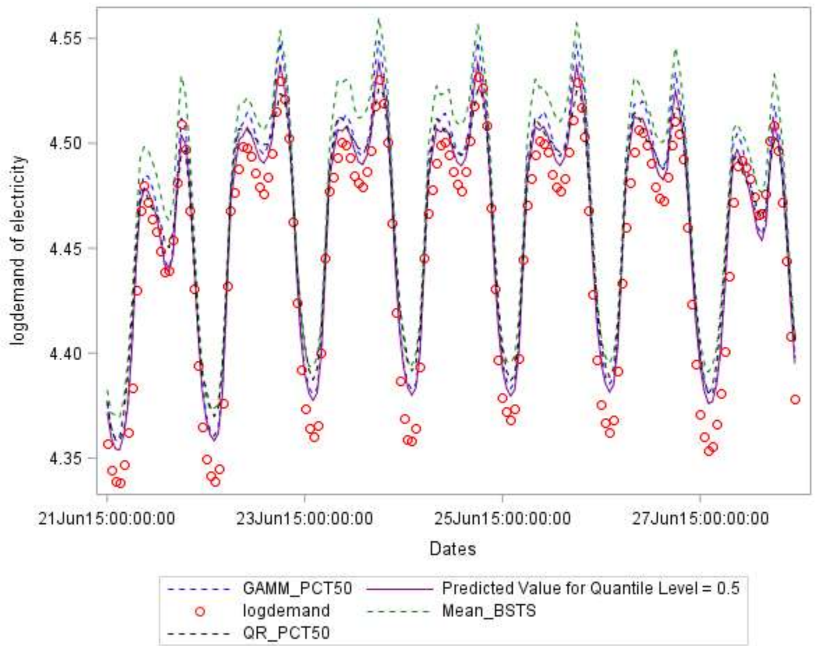


Figure 6.10: Comparison of forecasts from considered models to their averaged forecasts and actual demand in 2015

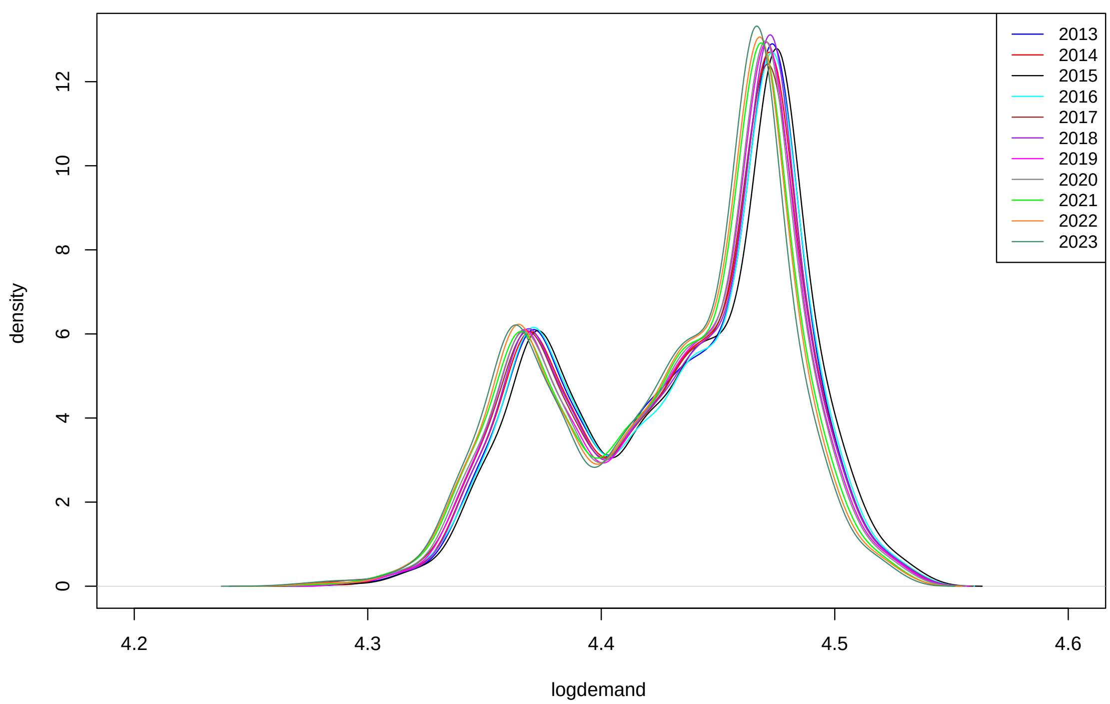


Figure 6.11: Distribution of forecasted hourly demand between 2013 and 2023

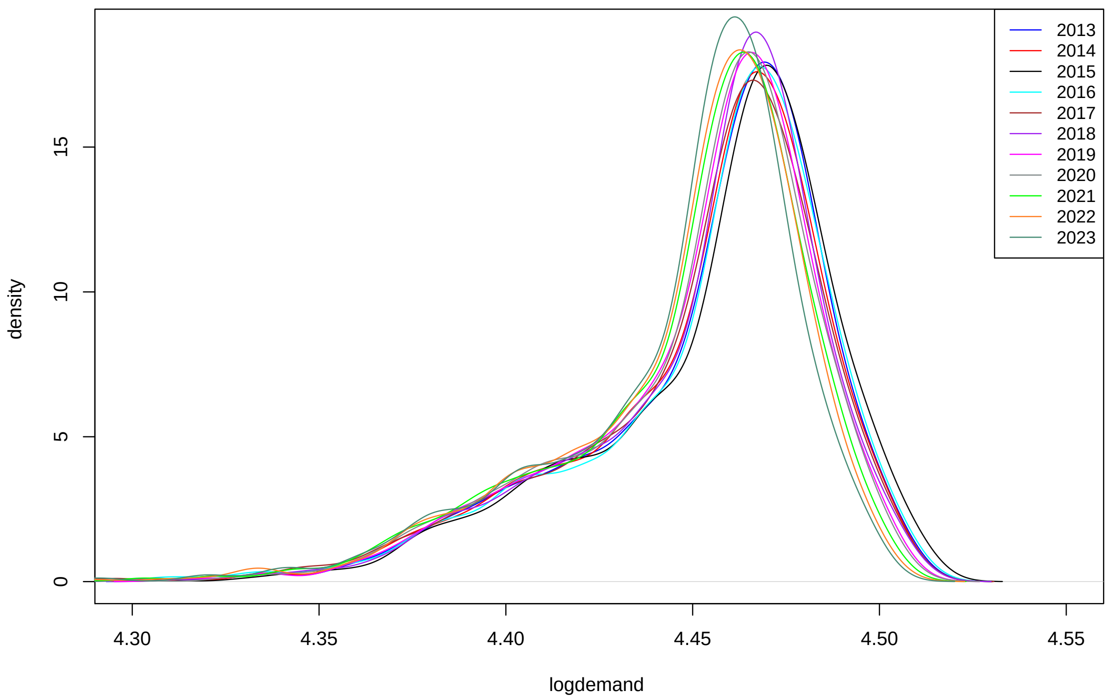


Figure 6.12: The morning peak demand distributions

Figures 6.12 and 6.13 suggest that if Eskom met peak electricity demand in 2015, it is more likely to continue meeting them until 2023 as the distribution of demand in 2015 is forecasted to shift more towards higher peak demand compared to the distribution of the 2023 peak demand. This finding is consistent with the results from QR model in Chapter 3, BSTS in Chapter 4 and GAMMQV in Chapter 5.

6.2.3 Historic and forecasted electricity demand in South Africa

Figure 6.14 represents the actual evolution of electricity demand in South Africa from 1997 to 2012 and the forecasted electricity demand from 2013 to 2023 at the 50 th percentile of the demand distribution from the recommended model. Electricity demand started dropping from 2012 and is unlikely to recover until 2023 according to our forecasts. Figure 6.14 shows that the highest hourly electricity demand reached in 2011 is unlikely in future years until 2023. The year to year drop in electricity demand is also evident in Figures 6.11, 6.12 and 6.13 where considering the distributions from 2013 to 2023, the 2023 demand distribution shifted more towards lower demand.

The sharp increase in demand between 1997 and 2007 in Figure 6.14 could partly be attributed to aggressive connection of more South African households to the electricity grid during this period indicated by the increased electricity consumption by the residential sector from 16% in 1993 to 23% in 2013 (Deloitte, 2017). As discussed in Chapter 1, only

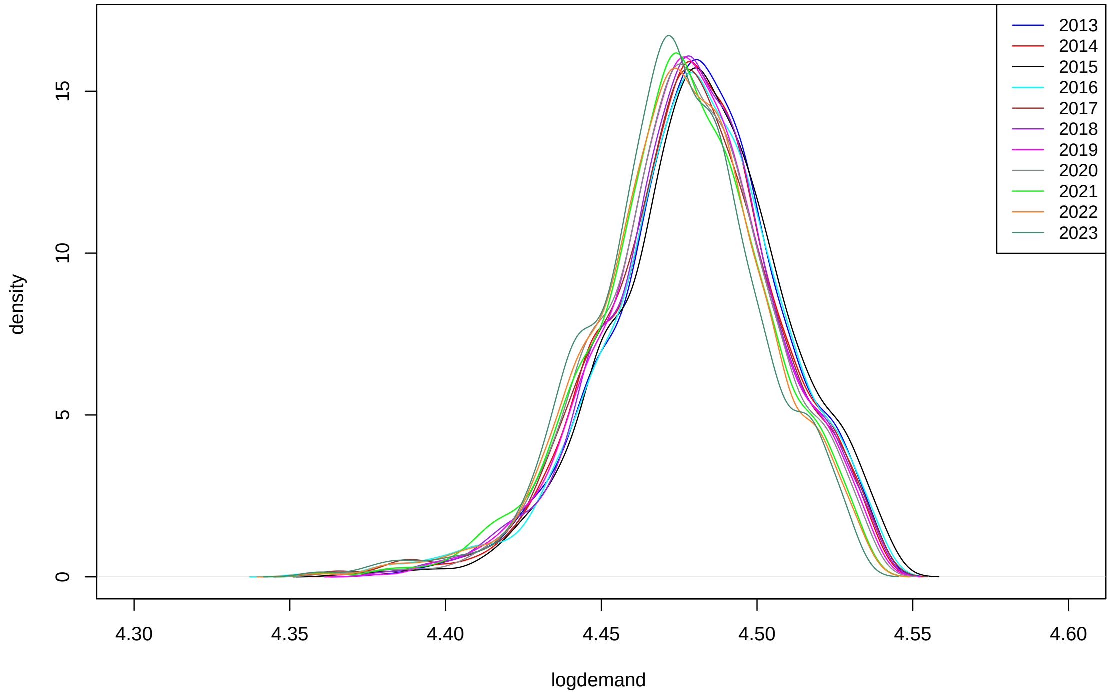


Figure 6.13: The Afternoon peak demand distributions

57.6% of the South African households had access to electricity in 1996 and this percentage went up to 84.7% in 2011. The economy was also performing well in this period as on average the South African economic growth rate between 2004 and 2007 was approximately 5% in real terms.

The decline in electricity demand in the latest years until 2015 could be attributed to the sluggish economic growth since the economy continued to performed poorly and slipped into a technical recession in the second quarter of 2018 showing decrease in (GDP) for April, May and June, as well as the possible growth in renewable sources of electricity and the slow-down in the rate at which new households were being connected to the electricity grid as the majority (84.7%) was already connected to the electricity grid in 2011. The decline in electricity demand could also be linked to the transition of the South African economy away from the reliance on energy-intensive mining and manufacturing sectors towards a more diverse range of service-oriented activities, as Deloitte (2017) asserted.

6.2.4 Assessment of models, forecasts on additional data

Additional data on hourly electricity demand were received from Eskom in December 2018 after the results in Chapter 3, Chapter 4 and Chapter 5 were published in two Journals and in the proceedings of an international conference, respectively. The additional data comprised the hourly demand from 2016 to June 2018. The additional data are different from

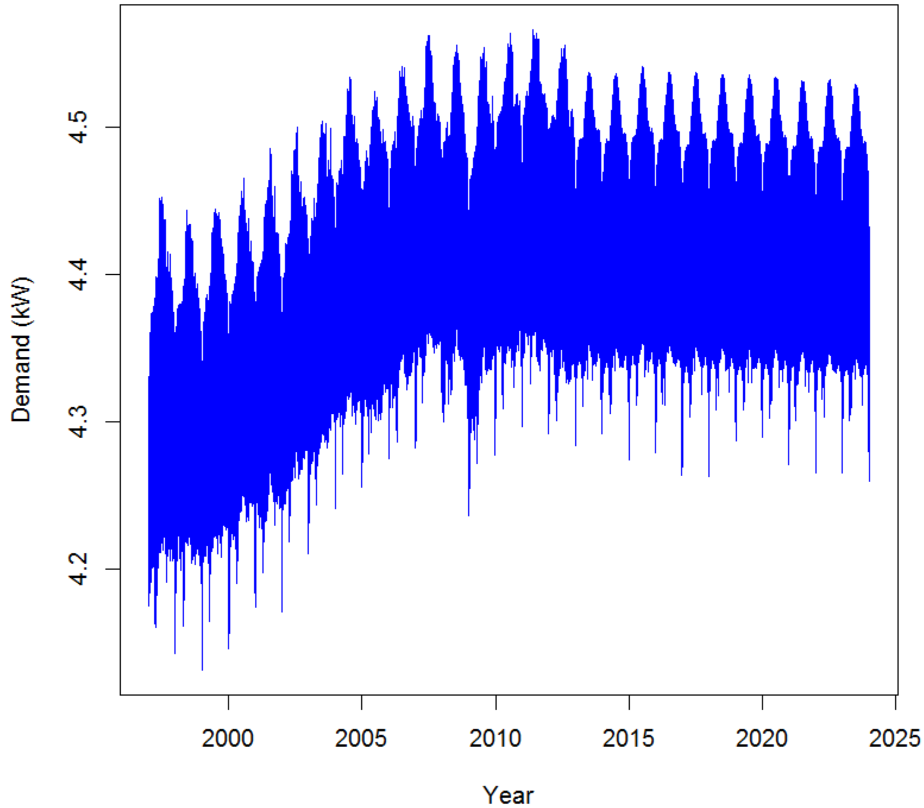


Figure 6.14: Evolution of demand for electricity in South Africa from 1997 to 2023

the validation data in that there was some load-shedding in 2014 and 2015 while there was none in 2016 and 2017. There was some load-shedding after June 2018. The additional data is different from the training data in that, there is a mix of years in terms of load-shedding in the training data, that is, there was some load-shedding in certain years and none in others. The data for more recent periods would of course also have included more recent patterns of change in electricity usage due to economic and demographic factors. Based on these additional data, the distribution of the actual electricity demand in 2016 to June 2018 are compared to the distribution of the forecasted demand of each considered model to assess how well the models forecasted the distributions of electricity demand represented by density functions in this very recent period. The hourly forecasts from each model are used to assess whether the forecasted daily profiles are close to the actual daily demand profiles and whether the forecasted daily demand profiles maintained the shapes of the actual daily profile over the years. MAPE values are used to assess how close our best guesses of demand are to the actual demand (where best guesses of demand refer to the forecasts at the 50 th percentiles of the demand distribution).

Before calculating the MAPE for each hour, the logdemand was transformed from the logarithmic scale back to the original demand scale. The MAPE of individual hours in 2016, 2017 and hours up to June 2018 are given in Figures 6.15, 6.16 and 6.17. There are 8 760 forecasted hourly demand in each year (the year has 8 760 hours). Most of the 8 760 forecasts of hourly demand in 2016, 2017 and up to June 2018 do not deviate from the actual hourly demand by more than 5% on average as shown in Figures 6.15, 6.16 and 6.17. This

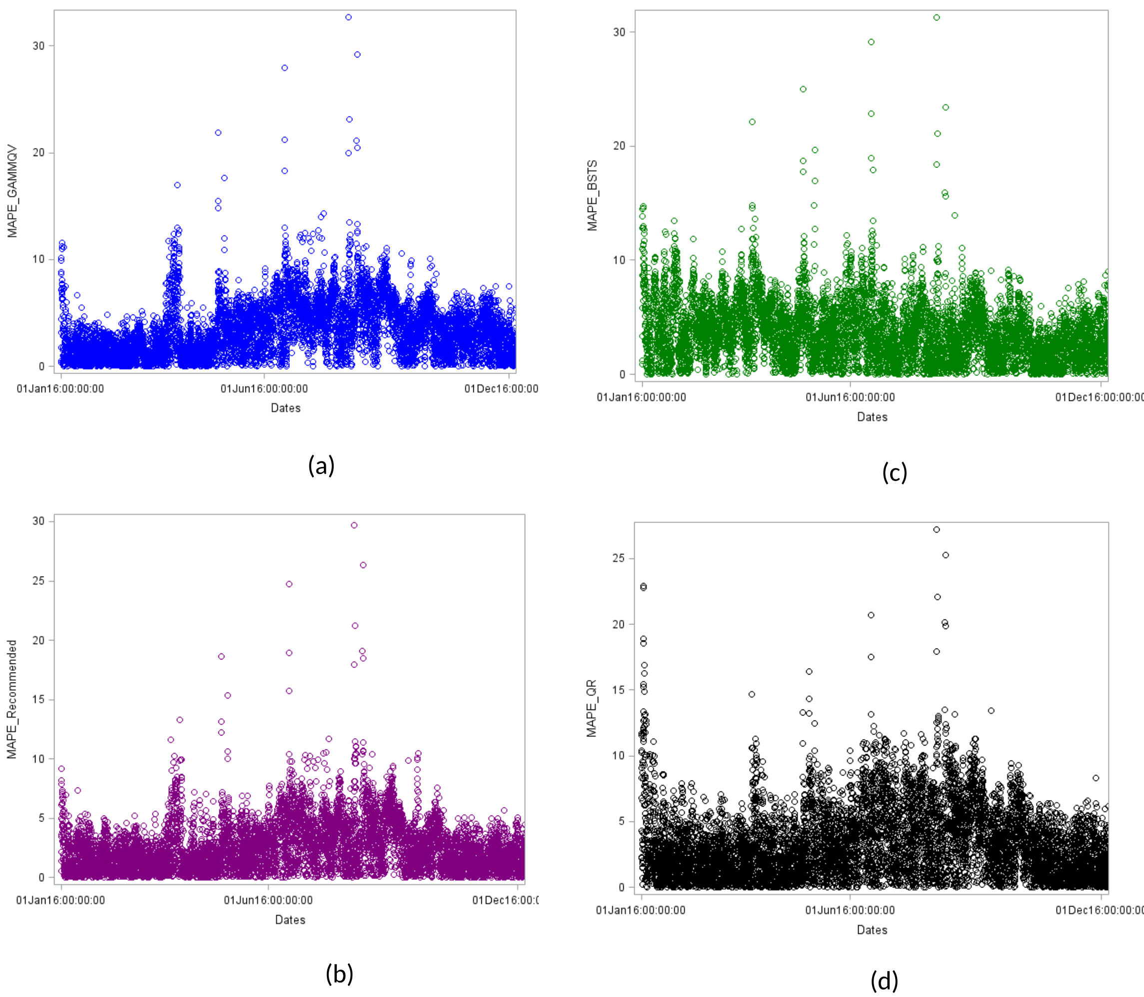


Figure 6.15: Mean absolute percentage errors 2016

implies that the prediction intervals (intervals between the 1 st and 99 th percentile) of the QR and GAMMQV are likely to include almost all forecasts while the credible intervals of the BSTS are more likely to include all the forecasts in the long-term. The models are doing well in terms of forecasting electricity demand for the additional data and there is no reason to doubt that they will continue giving robust results beyond the end of June 2018.

The density functions of the actual and predicted electricity demand between 2016 until June 2018 in Figure 6.18 were used to assess whether the densities of the forecasted demand represented the actual density functions of electricity demand in 2016, 2017 and up to June 2018. The 2016, 2017 and 2018 actual demand densities are well forecasted by the models. The forecasted density functions in 2016, 2017 and 2018 closely tracked the actual density functions as shown in Figure 6.18 (a), (b) and (c). The densities of the actual electricity demand for 2016, 2017 and 2018 are represented by the red lines in Figure 6.18 (a), (b) and (c).

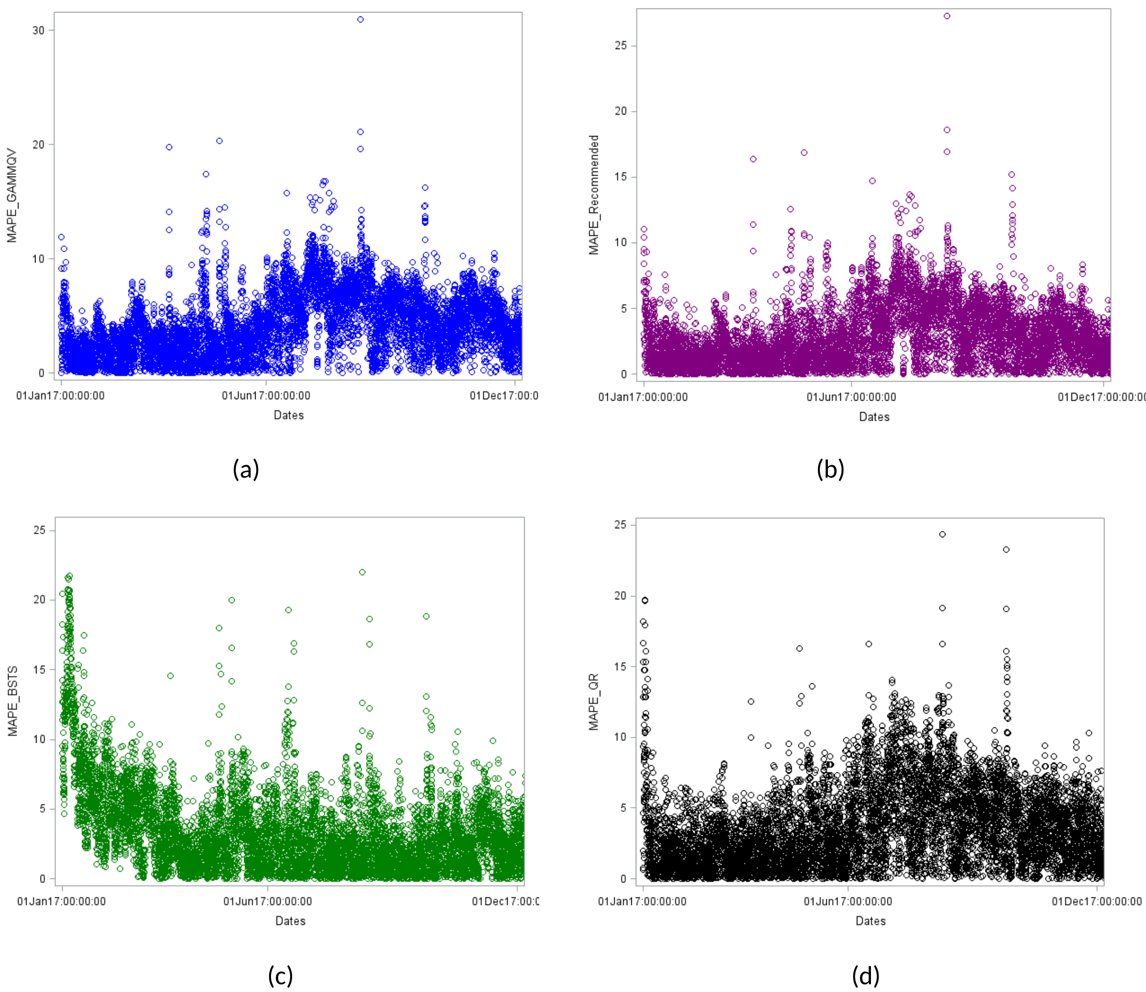


Figure 6.16: Mean absolute percentage errors 2017

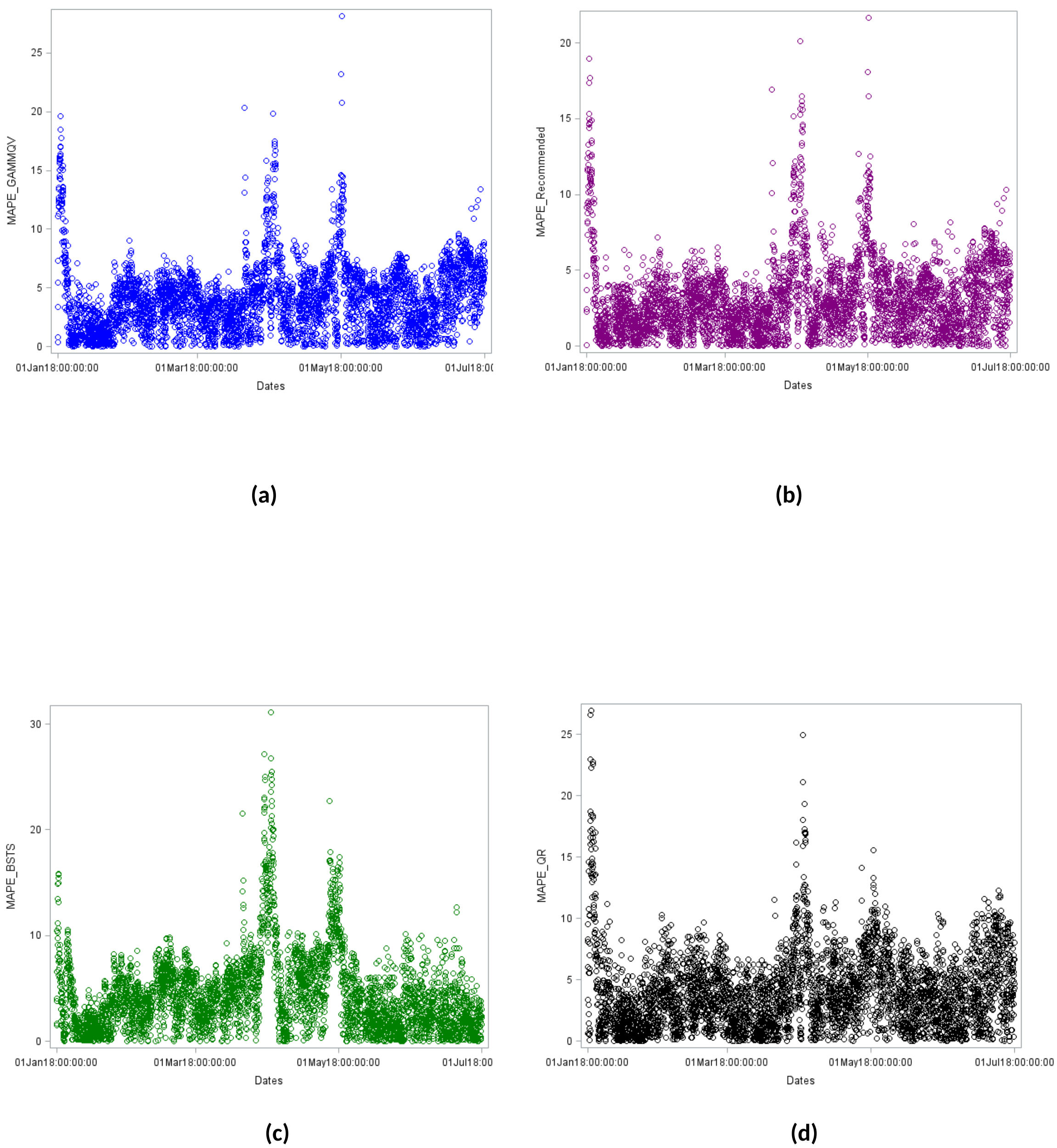


Figure 6.17: Mean absolute percentage errors 2018

Figures 6.19, 6.20 and 6.21 give the comparisons between the actual (red line) and the forecasted demand at the 50 th percentile of the electricity demand distribution from the models considered in this study for the 16 days in June 2016, 2017 and 2018. The 16 days are chosen in such a way that the weeks that had the highest hourly demand in 2013, 2014 and 2015 are included to maintain consistency in comparisons of forecasts between the different years. The additional data shows that the forecasted daily profiles maintained the shapes of the actual daily profiles of electricity demand and forecasted profiles are close to the actual profiles. Therefore, there is no reason to doubt that the forecasted daily profiles will continue to accurately estimate the actual daily profiles beyond June 2018. Figures 6.19, 6.20 and 6.21 show that any model considered in this study could be used to forecast the long-term electricity demand in South Africa since they all give robust forecasts. The demand forecasted from the quantile regression averaging model represented by the purple solid line in Figures 6.19, 6.20 and 6.21 could also be used to avoid choosing one model over others. The forecasts from the quantile regression averaging model do not deviate much from the actual demand and the forecasts of the BSTS, GAMMQV and QR.

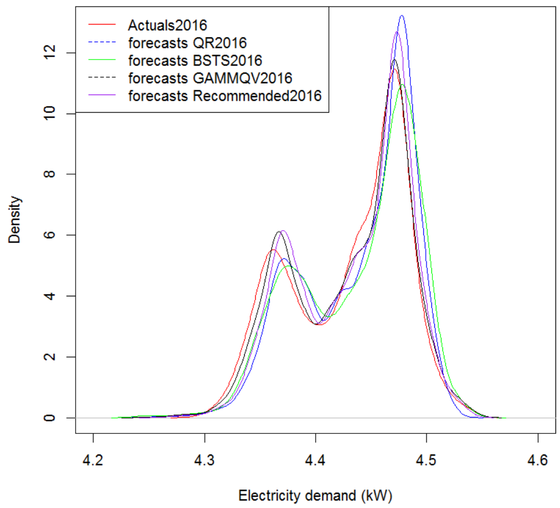
The contributions of Chapter 6 are:

1. the averaging of the point forecasts generated from the distinct models (BSTS, QR and GAMMQV) using quantile regression averaging for the long-term forecasting of electricity demand in South Africa. In this study the averaging of the point forecasts is done using forecasts at the 50 th percentile of the demand distributions generated by BSTS, QR and GAMMQV. The models are not coming from the same family,

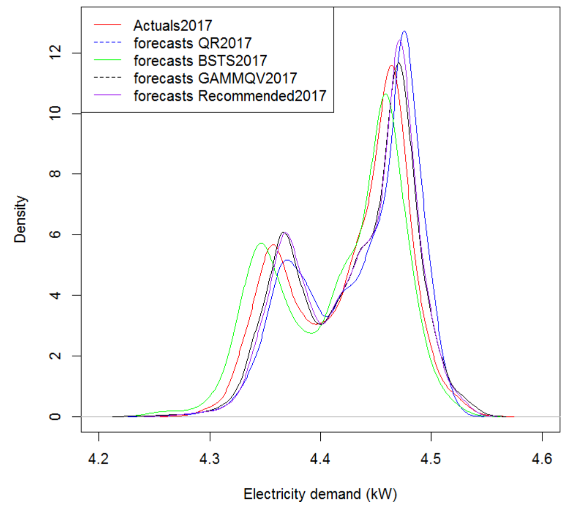
2. the averaging of point forecasts is useful for avoiding to rely solely on point forecasts generated by a single model, and

3. the averaging of point forecasts using QR is another way of generating probabilistic forecasts from point forecasts generated by different models.

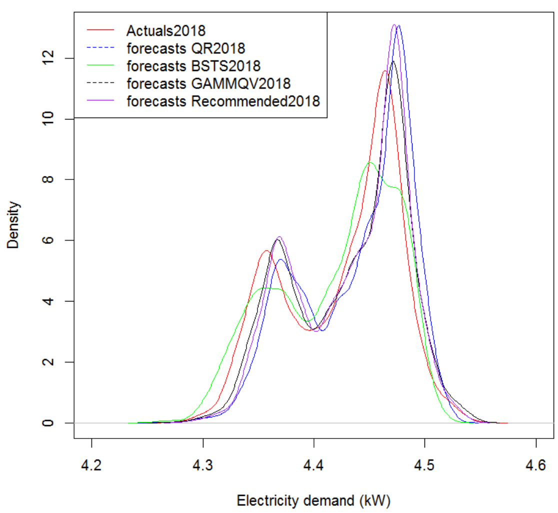
Chapter 7 presents the overall conclusions based on literature on electricity demand forecasting, the three models considered in this study and the summarised results. The contributions of the study are summarised in Chapter 7. Furthermore, the recommendations are made and the limitations of the study are discussed.



(a)



(b)



(c)

Figure 6.18: Comparison of actual and forecasted demand distributions 2016 to 2018

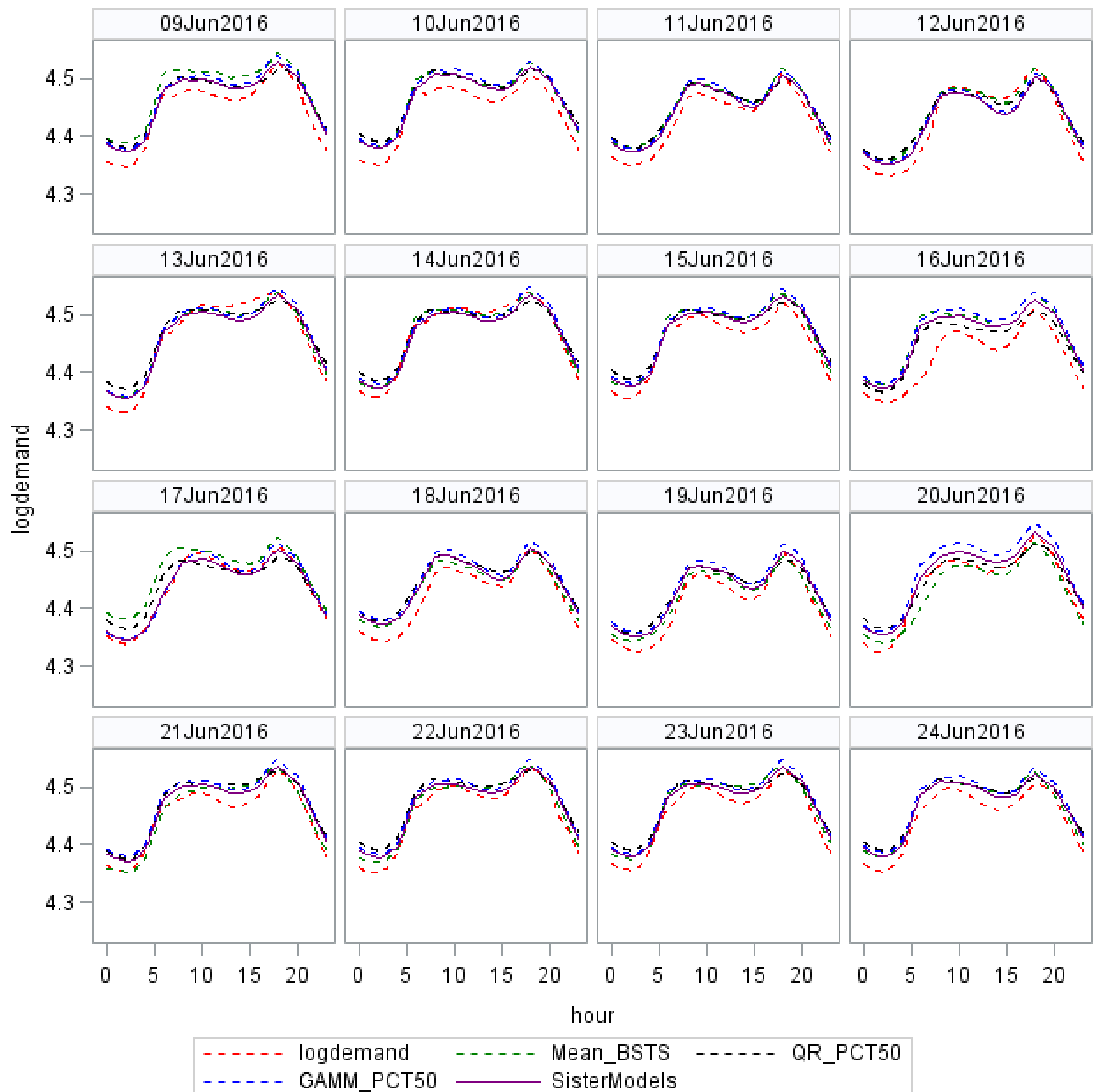


Figure 6.19: Comparison of actual and forecasted hourly demand at 50 th percentile 2016

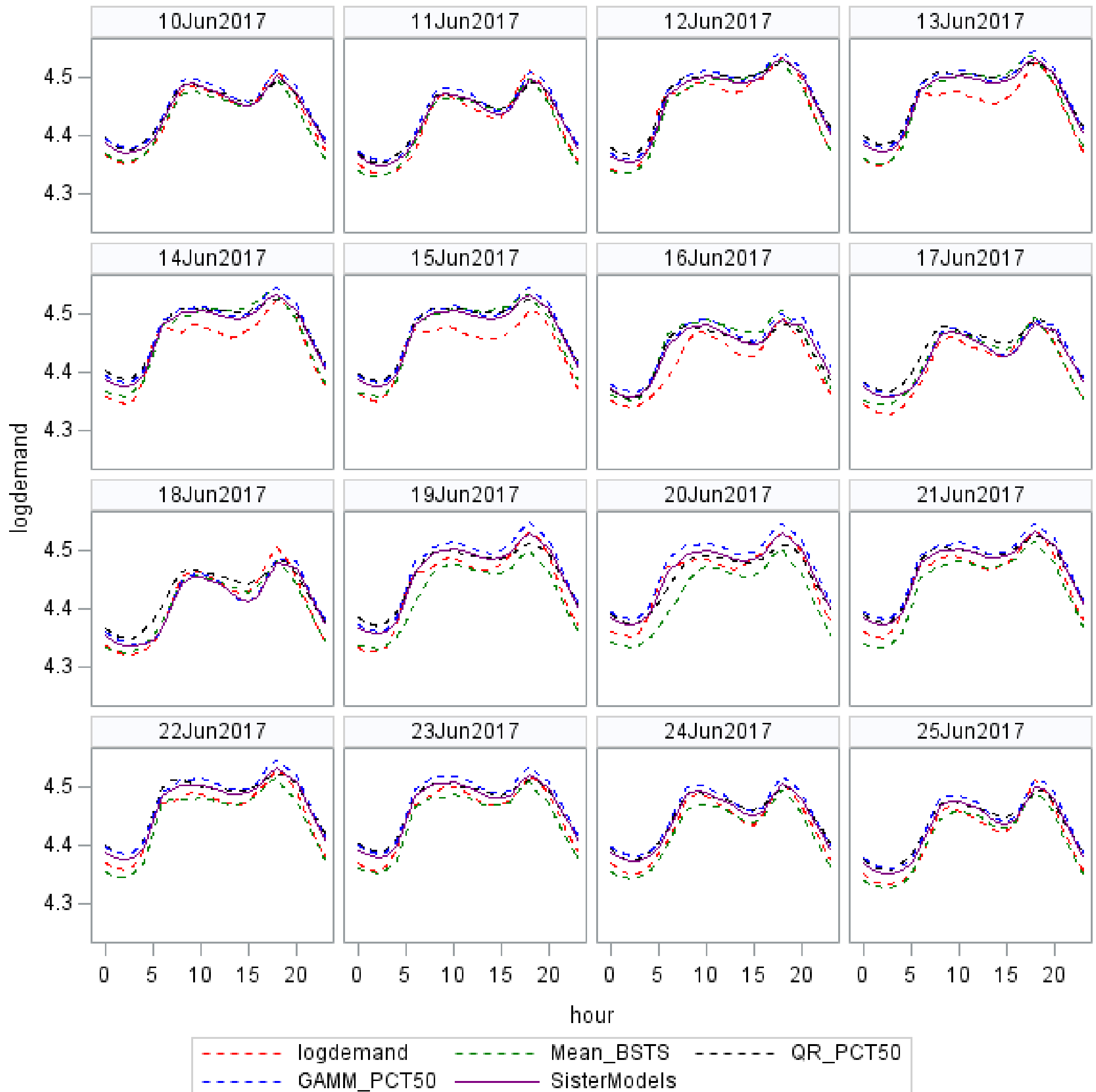


Figure 6.20: Comparison of actual and forecasted hourly demand at 50 th percentile 2017

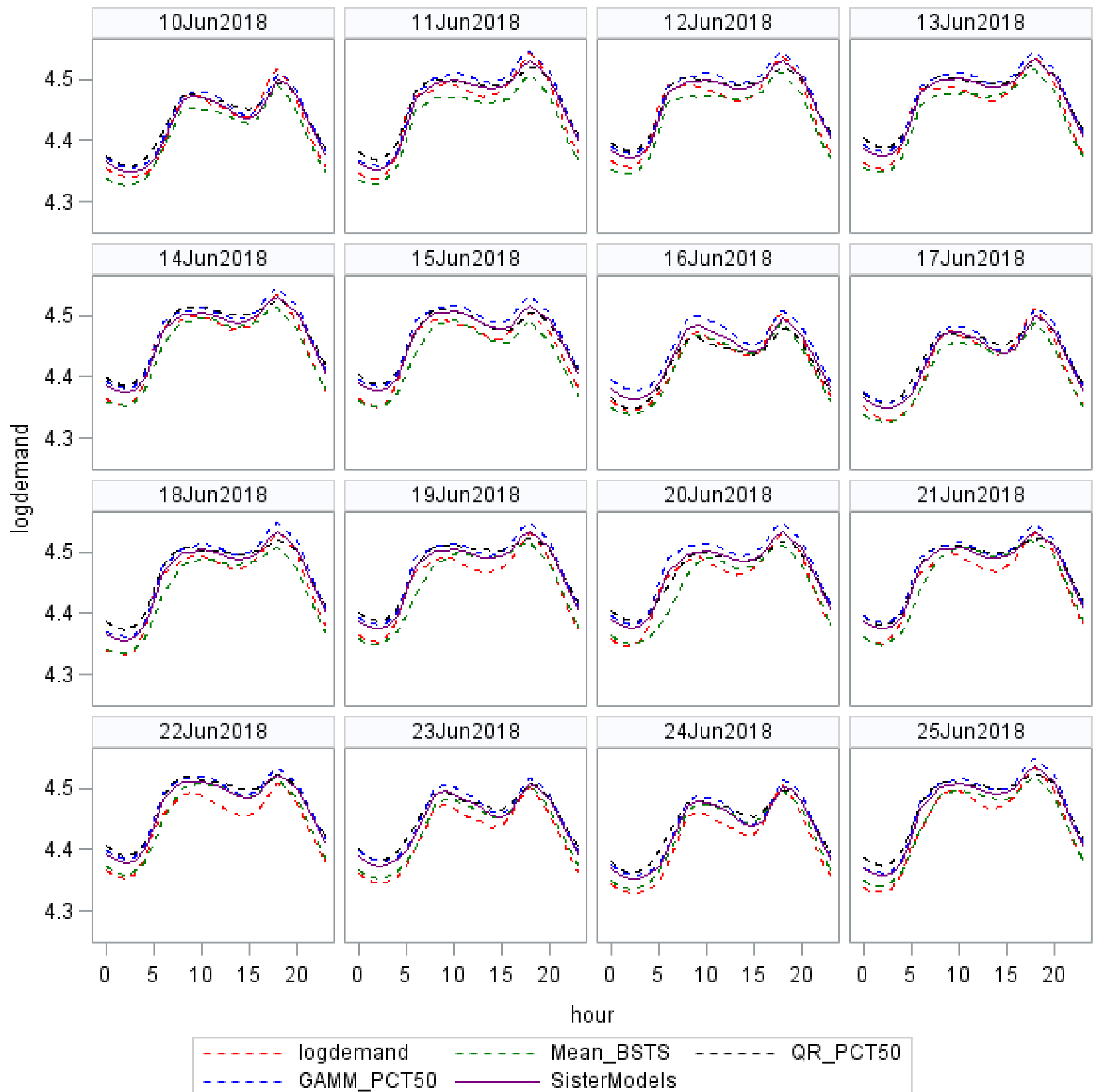


Figure 6.21: Comparison of actual and forecasted hourly demand at 50 th percentile 2018

Chapter 7

Conclusions and recommendations

7.1 Background

The long-term forecasts of electricity demand are important to those who manage the power utility companies, including political principals making decisions pertaining to the countries energy security. The decisions include generation capacity, maintenance planning, decisions on the optimal energy mix and plans for future infrastructure expansions. These decisions are made under uncertainties. Probabilistic modelling accounts for uncertainties in forecasting, enabling decision makers to make better decisions in the presence of such uncertainties. The decisions could have far reaching consequences because the decision for infrastructure expansion could result in the construction of unnecessary power generating facilities while the decision not to expand on the current infrastructure could result in failure to meet future electricity demand. For planning purposes, probabilistic forecasts of the long-term electricity demand can help managers to investigate a range of possible future values for the demand for electricity, so that the more informed decisions regarding the need for more (or fewer) power stations could be taken with minimum risk.

7.2 Findings

The conclusions from the results of the three models considered in this study are generally consistent. The main difference in the results is that, the prediction intervals of the Bayesian modeling approach (BSTS model) and the frequentist modeling approach (QR model) vastly differ in terms of their widths. The BSTS prediction interval widths are wider than those of the GAMMQV model (prediction interval of the GAMMQV are built from QR). The semi-parametric model (GAMMQV) is the best model for predicting electricity demand at the 50 th percentile of the demand distribution. The other findings are:

- the literature has shown that data on electricity demand attract different statistical models due to its complexity. The choice of a model to forecast electricity demand must be data-driven, objective-driven and must be country specific. A good scientific approach to forecasting electricity demand will produce good forecasts,

- the MAPE values from BSTS, QR and GAMMQV are all below 5% in 2013, 2014 and 2015 indicating that the forecasted demand at the 50 th percentile of the demand distribution deviated from the actual demand by less than 5% on average. The GAMMQV model gave much better predictions at the 50 th percentile of the demand distribution compared to QR and BSTS,

- the distribution of electricity demand in South Africa is expected to shift towards lower demand over the years until 2023.

- the distribution of peak electricity demand in South Africa is likely to shift towards lower demand over the years,

- the developed GAMMQV model indicates that the electricity demand in South Africa is unlikely to exceed the highest (36 826 kW) hourly demand until 2023. According to GAMMQV model, the probability of electricity demand exceeding 36 826 kW was below 0.15 from 2013 to 2023,

- the BSTS model shows that electricity demand in South Africa is likely to shift towards lower demand over the years until 2023 and it is estimated to lie between 15 849 kW (logdemand = 4.2) and 39 810 kW (logdemand = 4.6) during the period between 2013 and 2023 with a 90% probability. The utility company might want to assess whether with the current generation infrastructure, it would be able to meet electricity demand of 39 810 kW during the period $2013 - 2023$,

- the prediction interval widths for both QR and GAMMQV models are narrow compared to those of the BSTS model. There is a risk that some extreme electricity demand forecasts in QR and GAMMQV could fall outside the prediction intervals especially in the long-term,

- the additional data received in December 2018 support the published results on this study. The additional electricity demand data for 2016, 2017 and up to June 2018 received in December 2018 showed that the models considered in this study continue forecasting demand accurately. These models are therefore, expected to continue forecasting electricity demand accurately beyond June 2018.

7.3 Conclusions

Any stochastic time series model like SARIMA would perform poorly in forecasting the long-term electricity demand in South Africa because the forecasting model is built by decomposing the time series into trend, cyclic, seasonal and irregular components and if, for example, the trend in the training data follows an upward trajectory, then the extrapolated trend will continue the upward trend. Unfortunately, if the demand trend changes direction at the certain point, the predicted trend will continue the historical trajectory even after the point (year) where the demand trend started changing direction. This would result into the actual demand taking a downward trend while the predicted demand follow an upward trend and hence the demand forecasts in the long-term would vastly deviate from the actual demand. In addition, exogenous variables that are found in literature to be significant drivers of electricity demand such as temperature, calendar and economic variables cannot be used in the stochastic time series model because only previous values of the same time series are used to predict future demand values.

Simple regression models would perform poorly in forecasting the long-term electricity demand because the regression model assumes that individual data points are determined solely by the underlying parameters and some random noise. so, all information in a given data point can be used towards estimating the underlying parameters, as the regression models attempt to make educated guesses about the underlying parameters based on the observable data and this is only true if the data points are independent. The simple regression model ignores the inherent grouping, temporal or spatial structures. Therefore, a regression model with random effects or appropriately modelled error would be appropriate for modelling the South African electricity demand data.

A time series with a regression component such as USTS and SARIMAx would give better forecasts of the electricity demand compared to the stochastic time series and simple regression model because such a model has one component that treats the data as a time series which take the dependence of data points into consideration, as well as a regression that uses exogenous drivers of electricity demand. The forecaster should also consider the future trend if the model is used for forecasting demand in the long-term because this model will suffer the same fate as SARIMA in terms of the issues pertaining to trend. In this modelling framework, a linear relationship between electricity demand and its drivers is assumed, but unfortunately the historical electricity demand shows a nonlinear relationship with some of its drivers.

Regression based models like GAMMs which take the grouping, temporal or spatial structures into consideration by adding random effects, or appropriately modelled errors, or both, in the modelling and which also relaxes the assumption of linearity between predictors and demand would give better forecasts of the demand.

The GAMMQV model approach gives better forecasts but with narrow prediction intervals compared to the credible intervals of the BSTS model. Some of the long-term extreme electricity demand could fall outside the GAMMQV prediction intervals, therefore the model would be more suitable for forecasting demand in the medium-term.

Intuitively, the further the forecasts are in the forecasting horizon, the more uncertain one becomes about the forecasts and this expectation is captured well in the BSTS model, hence the BSTS model best captures uncertainties for long-term forecasting horizons.

The future evolution of the South African electricity demand trend is uncertain, therefore, in long-term forecasting, the trend could adversely affect the forecasts if it is inaccurately extrapolated, hence it must be taken into consideration when the model is fitted. The local level trend which assumes that the trend follows the random walk is used in BSTS, while in the GAMMQV, a nonlinear trend is extracted from the historical demand data and its future evolution is forecasted.

Considering the three economic drivers of electricity demand (*activity*, *structural* and *intensity* effects), it makes sense that the electricity demand in South Africa is declining and it will continue the downward trajectory until 2023. The economic *activity* has been declining in South Africa and this lately drove the economy into a technical recession in 2018. Due to the cost of electricity after a couple of steep tariff increases in South Africa, consumers of electricity became efficient (*intensity* effect) in their usage. Due to lack of capacity in the generation of electricity by Eskom in 2007, which resulted in a series of highly disruptive power outages in early 2008 through a programme of load-shedding further load-shedding in early 2014 and regularly between November 2014 and September 2015 as well as in 2018, companies and households had to find other sources of electricity. This could have led to the apparent decline in electricity demand from Eskom customers. The apparent decline in electricity demand could also be linked to the transition of the South African economy away from the reliance on energy-intensive mining and manufacturing sectors towards a more diverse range of service-oriented activities.

7.4 Recommendations

The recommended model for the long-term electricity demand forecasting should accurately predict electricity demand and should at the same time capture uncertainties since the long-term electricity demand in South Africa is forecasted with a lot of uncertainties as already highlighted in previous chapters. The BSTS model predicts the actual electricity demand well with wide credible intervals. On the other hand, GAMMQV predicts the actual electricity demand well (better than BSTS) but with narrow prediction intervals.

- The three models considered in this study could be used to forecast the long-term electricity demand in South Africa using probabilistic approach. However, in the long-term, some extreme electricity demand could fall outside the prediction intervals for QR and GAMMQV models because of their narrow prediction intervals.
- The GAMMQV model would be more suitable for medium-term probabilistic forecasting because it accurately forecast the actual electricity demand but with narrow prediction intervals. The GAMMQV is therefore recommended for medium-term forecasting of electricity demand in South Africa.

- The evolution of a long-term electricity demand trend is important because if it is not accurately captured, the forecasts will vastly deviate from the actual demand in the long-term. This is especially important for electricity demand in South Africa where the trend has breakpoints. At the breakpoint, the forecasts could continue the historical trend deviating from the actual trend and consequently the forecasts would vastly deviate from the actual demand. A model with trend taken into consideration is recommended for forecasting the South African electricity demand.
- In situations where it is difficult to choose one forecasting model over others, the averaging of point forecasts from those models would be recommended. In our case the forecasts at the 50 th percentiles of the demand distributions of QR, BSTS and GAMMQV models were averaged using QR and the long-term electricity demand was forecasted. This approach gave robust forecasts.

7.5 Contributions

As discussed in Chapter 1, the contributions of this study are the re-contextualisation of the existing QR, BSTS and the GAMM modelling techniques which resulted in the application of these modelling techniques in the totally new context (the probabilistic forecasting of the long-term electricity demand). Due to the complexity of the demand data, the parametric and non-parametric approaches to statistical modelling were considered and under the parametric approach, the frequentist and Bayesian approaches were considered. The expansion of an existing GAMM model to GAMMQV as well as fitting a GAMM model with nonlinear trend was made. The long-term electricity demand in South Africa was forecasted using probabilistic approaches and the forecasts can be effectively communicated so that planners who might not be statisticians would understand and appreciate their benefits. Due to the impact that the trend could have in the long-term forecasts, a local level trend that are assumed to follow a random walk was used in BSTS. A nonlinear trend was extracted from the demand data and its future evolution was forecasted in GAMMQV. The use of QR, BSTS and GAMMQV has shown that the distribution of electricity demand in South Africa is likely to shift towards lower demand. The forecasts at the 50 th percentile of the distribution of demand for the three considered models were averaged using QR to obtain the final forecasts.

7.6 Limitations of the thesis and future research

From this research the limitations of the thesis and future research are as follows:

1. additional improvement in forecasts accuracy could be achieved when the intraday correlation structures are explored which translates into the hours being modelled together and all model parameters estimated simultaneously. It would be interesting to consider models that harness the intraday and interday correlation structures in future research,

2. for QR, a predictor variable could be significant at the certain quantile but insignificant at other quantiles of the distribution. This creates a challenge for variable selection into the model. Research towards finding a method that could be used for variable selection into a QR model would eliminate the cumbersome approach where each QR model would be fitted separately or an adoption of the strategy that if the variable is significant at any quantile of the distribution then it is included in the model,

3. some exogenous drivers of electricity demand that have been found in the literature such as the natural variables (like temperature and cloud cover) and economical variables (like gross domestic product *GDP*) are not considered in this study because they were either unavailable or were not given at the hourly level like the electricity demand data. It would be interesting to consider these variables in future research,

4. the electricity demand in South Africa differs from one economic sector to another. It would therefore be ideal to predict the South African electricity demand at the sectorial level, but unfortunately the electricity demand data are given as the aggregated hourly electricity demand for all sectors of the South African economy such as transport, mining, manufacturing, residential and agricultural sectors,

5. it would have been interesting to analyse the electricity demand by the South African provinces, but unfortunately the hourly electricity demand is not currently readily available in the required format at the provincial level.

Bibliography

Alfares, H. K. and Nazeeruddin, M. (2002). Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34.

Amusa, H., Amusa, K., and Mabugu, R. (2009). Aggregate demand for electricity in South Africa: An analysis using the bounds testing approach to cointegration. *Energy policy*, 37(10):4167–4175.

Barakat, E., Al-Qassim, J., and Al Rashed, S. (1992). New model for peak demand forecasting applied to highly complex load characteristics of a fast developing area. In *IEE Proceedings C (Generation, Transmission and Distribution)*, volume 139, pages 136–140. IET.

Bedi, J. and Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. *Applied Energy*, 238:1312–1326.

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London*, 22(1):23–36.

Boraine, H. and Yadavalli, V. S. S. (2003). Electricity load forecasting with artificial neural networks. *South African Journal of Industrial Engineering*, 14(2):23–35.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Box, G. E. and Jenkins, G. M. (1976). *Time series analysis: forecasting and control, revised ed*. Holden-Day.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

Carstens, H., Xia, X., and Yadavalli, V. S. S. (2018). Bayesian energy measurement and verification analysis. *Energies*, 11(2):380.

Chen, BoJuen and Ming-Wei, Chang (2004). Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE transactions on power systems*, 19(4):1821–1830.

Cornec, M. (2014). Constructing a conditional GDP fan chart with an application to french business survey data. *OECD Journal: Journal of Business Cycle Measurement and Analysis*, 2013(2):109–127.

Cottet, R. and Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849.

Deloitte (2017). An overview of electricity consumption and pricing in South Africa, An analysis of the historical trends and policies, key issues and outlook in 2017. Techreport, Deloitte.

Fan, S. and Chen, L. (2006). Short-term load forecasting based on an adaptive hybrid method. *IEEE Transactions on Power Systems*, 21(1):392–401.

Fan, S., Methaprayoon, K., and Lee, W. (2009). Multiregion load forecasting for system with large geographical area. *IEEE Transactions on Industry Applications*, 45(4):1452–1459.

Farland, J. (2014). Zonal and regional load forecasting in the new england wholesale electricity market: A semiparametric regression approach. Master's thesis, University of Massachusetts.

Fay, D., Ringwood, J. V., Condon, M., and Kelly, M. (2003). 24-h electrical load data—a sequential or partitioned time series? *Neurocomputing*, 55(3):469–498.

Gelper, S., Fried, R., and Croux, C. (2010). Robust forecasting with exponential and holt–winters smoothing. *Journal of forecasting*, 29(3):285–300.

Gibbons, C. and Faruqui, A. (2014). Quantile regression for peak demand forecasting.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Goude, Y., Nedellec, R., and Kong, N. (2014). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid*, 5(1):440–446.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.*, 1(3):297–310.

Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55.

Hong, T. (2008). Long-term spatial load forecasting using human-machine co-construct intelligence framework. Master's thesis, North Carolina State University.

Hong, T. and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938.

Hong, T. and Shahidehpour, M. (2015). Load forecasting case study. *EISPC, US Department of Energy*.

Hong, T., Wilson, J., and Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, 5(1):456–462.

Hong, T., Xie, J., and Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*.

Huang, S. (1997). Short-term load forecasting using threshold autoregressive models. *IEE Proceedings-Generation, Transmission and Distribution*, 144(5):477–481.

Hyndman, R. and Athanasopoulos, G. (2013). Forecasting: Principles and practice, section 7/7. *URL: http://otexts. com/fpp/7/7.*, 10(13):2015.

Hyndman, R., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Hyndman, R. J. and Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2):1142–1153.

Inglesi, R. and Pouris, A. (2010). Forecasting electricity demand in South Africa: A critique of Eskom's projections. *South African Journal of Science*, 106(1-2):50–53.

Inglesi-Lotz, R. (2011). The evolution of price elasticity of electricity demand in South Africa: A Kalman filter application. *Energy Policy*, 39(6):3690–3696.

Jebaraj, S. and Iniyan, S. (2006). A review of energy models. *Renewable and Sustainable Energy Reviews*, 10(4):281–311.

Juberias, G., Yunta, R., Moreno, J. G., and Mendivil, C. (1999). A new ARIMA model for hourly load forecasting. In *Transmission and Distribution Conference, 1999 IEEE*, volume 1, pages 314–319. IEEE.

Koen, R., Holloway, J., Magadla, T., Mokilane, P., and Makhanya, S. (2017). Forecasts for electricity demand in south africa (2017 – 2050) using the csir sectoral regression model. techreport, Council for Scientific and Industrial Research.

Koen, R., Magadla, T., and Mokilane, P. (2014). Developing long-term scenario forecasts to support electricity generation investment decisions. In *43rd Annual Conference of the Operations Research Society of South Africa*, page 9.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 43–61.

Lai, S.-H. and Hong, T. (2013). When one size no longer fits all: Electric load forecasting with a geographic hierarchy. *SAS White Paper*.

Lehohla, P. (2005). Census 2001: Stages in the life cycle of South Africans.

Lewis, C. D. (1982). Industrial and business forecasting methods: A radical guide to exponential smoothing and curve fitting.

Liu, B., Nowotarski, J., Hong, T., and Weron, R. (2017). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, 8(2):730–737.

Liu, K., Subbarayan, S., Shoults, R., Manry, M., Kwan, C., Lewis, F., and Naccarino, J. (1996). Comparison of very short-term load forecasting techniques. *IEEE Transactions on power systems*, 11(2):877–882.

Maciejowska, K. and Weron, R. (2015). Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. *Computational Statistics*, 30(3):805–819.

Maharaj, Y. and Yadavalli, V. S. S. (2017). Modelling electricity spot prices with a three-regime markov model. In *Industrial Engineering and Engineering Management (IEEM), 2017 IEEE International Conference*, pages 2153–2158. IEEE.

Maistre, S., Lavergne, P., and Patilea, V. (2017). Powerful nonparametric checks for quantile regression. *Journal of Statistical Planning and Inference*, 180:13–29.

McSharry, P. E., Bouwman, S., and Bloemhof, G. (2005). Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Transactions on Power Systems*, 20(2):1166–1172.

Mori, H. and Takahashi, A. (2011). Hybrid intelligent method of relevant vector machine and regression tree for probabilistic load forecasting. In *Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition*, pages 1–8. IEEE.

Muñoz, A., Sánchez-Úbeda, E. F., Cruz, A., and Marín, J. (2010). Short-term forecasting in power systems: a guided tour. In *Handbook of power systems II*, pages 129–160. Springer.

Papalexopoulos, A. D. and Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, 5(4):1535–1547.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011.

Ramanathan, R., Engle, R., Granger, C. W., Vahid-Araghi, F., and Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International journal of forecasting*, 13(2):161–174.

Ravazzolo, F. (2007). *Forecasting financial time series, using model averaging*. Number 415. Rozenberg Publishers.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Scott, S. L. and Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23.

Sigauke, C. (2014). *Modelling electricity demand in South Africa*. PhD thesis, University of the Free State.

Sigauke, C. (2017). Forecasting medium-term electricity demand in a South African electric power supply system. *Journal of Energy in Southern Africa*, 28(4):54–67.

Sigauke, C. and Chikobvu, D. (2011). Prediction of daily peak electricity demand in South Africa using volatility forecasting models. *Energy Economics*, 33(5):882–888.

Soares, L. J. and Medeiros, M. C. (2005). Modeling and forecasting short-term electric load demand: a two-step methodology. Technical report, Working Paper 495, Department of Economics, Pontifical Catholic University of Rio de Janeiro.

Song, K., Baek, Y., Hong, D. H., and Jang, G. (2005). Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE transactions on power systems*, 20(1):96–101.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

Statistics South Africa (1998). The People of South Africa, Population Census, 1996. *Pretoria, South Africa: Statistics South Africa*.

Statistics South Africa (2008). *Community Survey, 2007: Basic Results Municipalities*. Statistics South Africa.

Statistics South Africa (2011). Census 2011 methodology and highlights of key results.

Suganthi, L. and Samuel, A. A. (2012). Energy models for demand forecasting—a review. *Renewable and sustainable energy reviews*, 16(2):1223–1240.

Taieb, S. B. and Hyndman, R. J. (2014). A gradient boosting approach to the kaggle load forecasting competition. *International journal of forecasting*, 30(2):382–394.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: a survey. *Journal of forecasting*, 19(4):235.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204(1):139–152.

Taylor, J. W., De Menezes, L. M., and McSharry, P. E. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22(1):1–16.

Taylor, J. W. and McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, 22(4):2213–2219.

Wang, S., Wang, S., and Wang, D. (2019). Combined probability density model for medium term load forecasting based on quantile regression and kernel density estimation. *Energy Procedia*, 158:6446–6451.

Weron, R. and Misiorek, A. (2004). Modeling and forecasting electricity loads: A comparison. In *International Conference EEM-04 proceedings*, pages 135–142.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wood, S. N. (2017). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing*, 27(4):985–989.

Xiong, T., Bao, Y., and Hu, Z. (2014). Interval forecasting of electricity demand: a novel bivariate emd-based support vector regression modeling framework. *International Journal of Electrical Power & Energy Systems*, 63:353–362.

Yue, Y. R. and Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, 55(1):84–96.

Zhang, J., Wei, Y.-M., Li, D., Tan, Z., and Zhou, J. (2018). Short term electricity load forecasting using a hybrid model. *Energy*, 158:774–781.

Zhao, H., Ren, Z., and Huang, W. (1997). Short term load forecasting considering weekly period based on par. *proceeding-chinese society of electrical engineering*, 17:211–213.

Ziramba, E. (2008). The demand for residential electricity in South Africa. *Energy Policy*, 36(9):3460–3466.