Forecasting electricity demand in South Africa using artificial intelligence

UNIVERSITY OF JOHANNESBURG

Lufuno R. Marwala

Supervised by: Professor Bheki Twala
16 July 2015

This Doctor of Philosophy is submitted to the
Department of Engineering Science, University of Johannesburg

Abstract

This thesis introduces a novel artificial intelligence technique called extreme learning machines (ELM) and structural causal models (SCM) for forecasting electricity consumption using time series and causality approaches. Time series data is used to construct univariate models for forecasting a one step ahead electricity consumption on a monthly basis. For causal analysis, the study is novel in that it mathematically models the relationship between electricity consumption and production levels in the manufacturing sector and mining sector in South Africa.

This work contributes to knowledge, firstly by conducting an empirical comparison of existing forecasting techniques which include a linear forecasting method, namely autoregressive moving average (ARMA) and artificial intelligence techniques namely artificial neural networks (ANN), neuro-fuzzy network (ANFIS), support vector regression (SVR). Total electricity consumption data series, sampled on a monthly basis, from 1985 to 2011 was used for experimentation. ARMA performance was hindered by nonlinear dynamics in the data. ANN can handle nonlinearities but uses empirical risk minimization (ERM) which has a local minima problem and over fitting and the model is difficult to read which makes it a black box. ANFIS uses fuzzy clustering which brings about readability of the model and improved performance compared to ANN but uses ERM to learn which is vulnerable to the local minima problem. SVR uses structural risk minimization (SRM) which overcomes the local minima problem and gives better performance than ANN and ANFIS however, SVR models are computationally costly.

Secondly, this work proposes the use of two novel techniques namely, basic ELM and optimally-pruned ELM (OP-ELM) to forecast the electricity consumption data series. The two techniques use a learning technique that converts single layer feed-forward network learning problem into a linear problem which can find the universal minima and requires very small processing time. OP-ELM prunes the hidden layer to eliminate unnecessary hidden units and avoid over-fitting. ELM and OP-ELM were found to be significantly better in accuracy and computationally faster than the ANN, ANFIS, and SVR. It was also found the ELM differs significantly in duration of computation compared to OP-ELM, with ELM computing faster.

The third and novel contribution in this work is the proposal to use SCM and graphical

causal model for time series causal analysis in electricity consumption forecasting to identify the causal variables. Unlike Granger causality which focuses on accurate modeling of the systems, SCM provides a framework for reasoning about the causal relationship between variables. SCM was successfully used to identify the causal variable.

The fourth and also novel contribution is the use ELM and OP-ELM in conducting granger causality testing using the causal variable identified using SCM. Experiments were performed using data series of electricity consumption and the manufacturing production index in the manufacturing sector and consumption and mining production index in the mining sector. Using OP-ELM, empirical results showed that a granger causal relationship exists between manufacturing production index and electricity consumption in the manufacturing sector which was not the case when using ELM. No causal relationship was found between the mining production index and electricity consumption using both ELM and OP-ELM.

The experiments performed with ELM and OP-ELM also explored the question of whether a lagged dependent variable should be included on the right-hand side of the regression equation. In both manufacturing sector and the mining sector it was found that there was no justification for excluding the dependent variable on right hand side of the regression model.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof Bhekisipho Twala for his guidance and support throughout this study. He provided me with invaluable inputs and insights that made the success of the study possible.

I would like to also convey my sincere gratitude to the database administators at Quantec, NERSA, Statssa and Eskom. Without their support and dedication to data collection this study would not have been possible. I also benefitted from the funding provided by NRF and for that I am very grateful.

Lastly, I would like to thank my family, my mother Tshinakaho Marwala, my late father Thifulufhelwi Marwala, my grandmother Takalani Marwala and my siblings Lutendo, Pfunzo and Andani. To my wife Busisiwe and my son Phathutshedzo you are my inspiration.

Contents

List of Figures

List of Tables

Nomenclature

- AI: Artificial Intelligence
- ANN: Artificial Neural Network
- AR:Autoregressive
- ARMA: Autoregressive Moving Average
- ARIMA: Autoregressive Integrated Moving Average
- ELM: Extreme Learning Machines
- ERM: Emperical Risk Minimisation
- DSM: Demand Side Management
- FCM: Fuzzy C-means
- GDP: Gross Domestic Product
- GK: Gustafson-Kessel HCM: Hard c-means
- IRLS: Iteratively Reweighted Least Squares
- IRP: Integrated Resource Planning
- LOO: Leave One Out
- MRSR: Multiresponse Sparse Regression
- MLP: Multilayer perceptron
- OP-ELM: Optimally-Pruned Extreme Learning Machines

- PSO: Particle Swarm Optimisation
- RBF: Radial Basis Function
- RNN: Recurrent Neural Network
- SEM: Structural Equation models
- SVM Support Vector Machines
- SRM: Structural Risk Minimisation
- SCM: Structural Causal Model
- TS: Takagi-Sugeno
- VAR: Vector Autoregression
- VEC: Vector Error Correction

Chapter 1

Introduction

The main objective of electricity consumption forecasting is to ensure that the supply of electricity meets the demand at all times. Supply/demand balance can only be ensured through a thorough electricity-supply planning exercise which requires efficient management of existing power systems and optimization of the decisions concerning additional capacity. Any electricity planning model relies on demand forecasting. Through the forecasting exercise decision makers are assisted in making decisions about electricity generation investments that will maintain acceptable supply levels to avoid catastrophic power shortages. Yet forecasting is not an exact science thus making the decision making process highly uncertain. According to Herbert Simon every decision is made within the constraints of bounded rationality. Bounded rationality asserts that rational decision making is bounded by failures of knowing all the alternatives, uncertainty about relevant exogenous events, and inability to calculate consequences [1]. Planning for energy supply includes but not limited to decisions about the optimal energy source mix, when to invest in new capacity, maintenance schedule, pricing, and energy market structure. Forecasting in this regard, serves to reduce the uncertainty or to expand the space within the boundaries for decision makers by attempting to predict the consequences of decisions taken. Marwala has termed this phenomenon of expanding the boundaries Flexibly-bounded rationality [2].

Electricity demand forecasting is divided into short-term forecasting which covers hourly

to weekly forecasting, medium-term forecasting which covers from monthly to quarterly forecasting and lastly, long-term which covers years. Short-term forecasts are used for control and scheduling of the power system and also as inputs to the load flow study or contingency analysis. Medium to long-term demand forecasting are useful in determining the capacity of generation required in the future to meet the forecasted demand and also to plan for transmission or distribution system additions and the type of facilities that are required in transmission planning and maintenance planning. This also assists in keeping an acceptable supply reserve margin especially to accommodate the peak demand. There are two types of reserves namely the spinning reserve and the cold reserve. Spinning reserve is defined as the unused capacity which can be activated on decision of the system operator and which is provided by devices which are synchronized to the network and able to affect the active power [3]. Cold reserve is defined as the extra generation capacity that is not already connected to the system [4]. When the future demand is overestimated it results in unused spinning reserve which is resource wastage and when the demand is underestimated the cold reserve has to be brought to life which is costly. For certain level of reserve needed to be maintained the future demand has to be determined through forecasting.

Demand forecasting relies on functional analysis of the electricity demand system which requires the identification of variables that control the behaviour of the electricity demand. The analysis has to specify how these variables interact to determine a particular response. The objective of the analysis is to create a model that better approximate the structural functioning of the electricity supply and demand system. The modeling process limits itself to observable input-output relation which is specified in terms of the history of the inputs. In the end, modeling process strives to discover a lawful relationship between the input and the output.

1.1 Electricity consumption data series

Electricity consumption is characterized by cyclicality, seasonality and randomness [5]. There are three seasonality types namely, daily, weekly and yearly cycles. The weekly cycle is caused

by the weeks work cycle. The yearly cycle is as a result of climatic conditions. And day cycles are dependent on weather conditions. All these characteristics are reflected on the data series depending on the frequency of sampling which could be hourly, daily, monthly or yearly. A modelling technique that has the ability to comprehend the complexity of the system that exhibit these dynamics is required.

This study implements two strategies that of autoregression and that of causality. Autoregression uses a single variable time series, electricity consumption to construct a model for forecasting. Unlike explanatory forecasting, time series forecasting treats the system as a black box and endeavours to discover the factors affecting the behaviour. There are two reasons for wanting to treat a system as a black box [6]. First, the system may not be understood, and even if it were understood it may be extremely difficult to measure the relationships assumed to govern its behaviour. Second, the main concern may be only to predict what will happen and not why it happens. For these reasons, time series forecasting is one of the least understood areas which has been under scrutiny for some time.

Explanatory forecasting assumes a cause and effect relationship between the inputs and output. According to explanatory forecasting, changing the inputs will affect the output of the system in a predictable way, assuming the cause and effect relationship is constant. This approach involves finding variables such as the GDP, weather, etc. that can be used as control variables to forecast the electricity consumption. The study of the causal relationship between GDP and electricity consumption has dominated previous studies [7]. The relationship between electricity demand and the economy can also be studied under regression analysis but the causal analysis has been the main focus.

In the causal framework, the determination of the causal direction is the main objective. The causal direction can be unidirectional, meaning that it is running from the economy to the electricity demand or vice versa; or it could be bidirectional, both variables causing each other; or there could simply be no causal relation which means neutrality. Determining these relationships is critical for decision makers and policy makers. No causality implies that each variable evolves on its own without affecting the other.

Causality in this study differs with other previous studies in that the consumption of electricity is decomposed into different sectors. In this study the focus is on the manufacturing sector and the mining sector. The point of the decomposition is to establish a relationship between levels of production and electricity consumption in these different sectors. The manufacturing production index measures the total output of industrial/manufacturing sector of the economy. Its fluctuation reflects the performance of the manufacturing sector on month to month basis. A data series was collected, sampled on a monthly basis from 1985 to 2011, from statistics South Africa for the study. In the mining sector, the volume of mining production, also known as the production index, is a statistical measure of the change in the volume of production. Similarly a data series sampled on a monthly basis from 1985 to 2011 from statistics South Africa.

The data sources used for this work include statistics South Africa, national energy regulator of South Africa, Eskom, and Quantec database. The database administators were contacted for confirmation of the integrity of the data for each of these institutions.

1.2 Methodologies and literature survey

1.2.1 Regression tools

The tools used in load/consumption forecasting range from regression-based approaches over time-series approaches towards artificial intelligence and expert systems. The objective of the linear regression model is to describe the output variable through a linear combination of one or more input variables [8]. Regression methods are relatively easy to implement and make it easy for researchers to understand the relationship between input and output variables. In regression modeling, it is mainly linear methods that are used and as a result they are limited when dealing with non-linear systems. Kyriakides and Polycarpou reported the inherent problems in regression models in identifying the correct model, due to the complex non-linear relationship between the load and the influencing factors [8]. There are many other studies that appear in the literature on regression models, for example regression models based on local polynomial regression for Short-Term Load Forecasting [9], non-parametric regression [10],

or robust regression methods [11]. Regression analysis also assumes that observations are independent of each other and, therefore, the errors are uncorrelated. In most cases this is not case and as a result regression modeling produces limited models that poorly approximate the underlying system.

1.2.2 Time series methods

Time series analysis enables researchers to describe variation in variables of interest over time, to gain a better understanding (or explanation) of the data-generating mechanism, to be able to forecast future values of a time series, and to allow for the optimal monitoring and control of a systems performance over time [12]. Time series analysis is advantageous because of the ability to model both linear and nonlinear relationships between variables over time.

The most popular time series methods are the Box-Jenkins methods. They are both univariate and multivariate. Univariate approach is usually applied to short-term load forecasting and multivariate is used for all time horizons [13]. The Box and Jenkins methods have also been used widely in forecasting electricity demand and these includes, autoregressive models (AR), autoregressive moving avarage (ARMA) and autoregressive integrated moving average (ARIMA) modelling. These tools have used by many, within a univariate framework, see, Abraham and Nath, [14]; Darbellay and Slama, [15]; Laing and Smith, [16]. Taylor [17] applied three univariate models, namely, the autoregressive, the autoregressive integrated moving average (ARIMA) and a novel configuration combining an AR(1) with a highpass filter. They assessed the forecasting performance of each model and found that the AR(1)/highpass filter model yields the best forecast.

Wang et al. used residual modification models to improve the precision of seasonal ARIMA for electricity demand forecasting [18]. In this study, PSO optimal Fourier method, seasonal ARIMA model and combined models of PSO optimal Fourier method with seasonal ARIMA are applied in the Northwest electricity grid of China to correct the forecasting results of seasonal ARIMA. They found that prediction accuracy of the three residual modification models is higher than the single seasonal ARIMA model and that the combined model was the better

of the three models.

The box-Jenkins methodologies involve an iterative model-building procedure through which any deterministic component of the time series is identified and removed in order to make the data stationary before standard data analysis methods could be used with time series data. The drawback with these approaches is that sometimes the assumption of stationarity may not be true for the lifetime of the time series which can lead to misspecification of the parameters and therefore, erroneous predictions. Furthermore, these approaches are optimised to model linear relationships and are not optimal for nonlinear relationships between variables.

There are alternatives to Box-Jenkins approaches that have emerged these include two of which are distributed lag models (ARDL) [19] and differential equation models [20][21]. Distributed lag analysis is a specialized technique for examining the relationships between variables that involve some delay. This technique relies on a simple structural equation model that can be estimated by an ordinary least squares (OLS) regression. ARDL is mostly used in causality studies.

1.2.3 Causalilty studies

Most of the studies in causality are two variable causality studies relating growth and electricity consumption. These studies have significantly increased our understanding of the relationship between the two variables in different economic contexts. The drawback, however, is that the two variable causality test is vulnerable to specification bias by ignoring other variables. This is because the causality model is sensitive to specification and number of lags [22]. The most pupular methodology for causal analysis in time series is the Granger causality test. A simple formulation of a Granger Causality is that if the prediction of one time series is improved by incorporating the knowledge of the second time series, then the latter has a causal influence on the former [23].

The standard granger causality test use vector auto-regression models (VAR) and vector error correction model (VEC). VAR assumes that the underlying variables are stationary, or integrated of order zero in nature. This approach is based upon the Granger Representation

Theorem, in which he showed that if a pair of I(1) series are cointegrated there must be a unidirectional causation in either way [24]. Thus, testing for a unit root and cointegration become the starting point in testing for causality between two time series. The Dickey-Fuller type tests [25] are used for unit root test(s), and the Engle-Granger or the Johansen test [26], are used for the cointegration test . If cointegration exists, the causality test may be conducted in two ways. First, the integrated data may be used in levels in a bivariate autoregressive model, due to the consistency properties of estimation [24]. Secondly, a bivariate model containing error correction mechanism terms due to the Granger representation theorem may be used in causality testing [24]. If the data are integrated but not cointegrated, then causality tests can be conducted by using the first differenced data to achieve stationarity. Dickey-Fuller test and the Phillips-Perron test [27], among others, have been developed to test for staionarity in the time-series econometrics literature, however, these tests suffer from very low power in distinguishing between a unit-root and a near- unit-root process.

The two variable causality test methodologies are the most widely used methods and they have weaknesses. Most of the time series is non-stationary and according to Gujarati when the series are integrated the F-test is no longer valid [22]. Non-stationarity could be an indication of spurious regression problem. This renders the usage of VAR impossible or inappropriate. If the variables are co-integrated the VAR model is transformed into VEC model. VEC came out of the work done by Granger, Engle and Granger and Johansen [24] [28] [26]. VEC captures the stochastic characteristic of the data by separating the short-term relations and long-term relations of the variables under consideration. The VEC model requires a co-integration test which is nontrivial task, before the estimating the parameters of the model. However, the interest of the researchers in these studies is to estimate a causal model or the significance of the coefficients of the VAR parameters, hence the introduction of the Toda-Yamamoto model [29].

Toda-Yamamoto (TY) model is a statistical inference model that makes parameter estimation possible without the co-integration of the VAR system. Toda-Yamamoto procedure makes Granger-causality much easier and researchers do not have to test for integration or transform

VAR to VEC. Toda-Yamamoto uses a modified Wald test for restriction on the parameters of the VAR (k) with k being the lag length of the VAR system. With this approach the correct order of the system (k) is augmented by the maximal order of integration (dmax) then the VAR(k + dmax) is estimated with the coefficients of the last lagged dmax vector being ignored [29]. Toda and Yamamoto can confirm that with this approach the Wald statistic converges in distribution to a chi-square random variable with degrees of freedom equal to the number of the excluded lagged variables regardless of whether the process is stationary, possibly around a linear trend or whether it is cointegrated. The Toda-Yamamoto procedure circumvents the bias associated with unit roots and cointegration tests as it does not require pre-testing for cointegration properties of the system [30] [31]. However, Toda-Yamamoto has weaknesses, as compared to methods in which cointegration is considered TY is considered inefficient; TY cannot distinguish between short run and long run causality; and lastly, TY cannot test for hypothesis on the long run equilibrium [32].

Using these methodologies, most studies have found a causal relationship between GDP and energy in industrialised countries. These studies are limited to a two variable analysis. The relationship between energy and GDP was particularly significant in the case of Japan for the period 1950-1982 [33]. However, when the period was altered and restricted to 1950-1973 the relationship was no longer significant. The study by Stern, advanced beyond the Just using GDP and energy in testing for Granger causality in a multivariate setting using a vector autoregression (VAR) model and he included GDP, energy use, capital, and labor inputs [7]. He also included a quality-adjusted index of energy input in place of gross energy use in the model. The inclusion of more variables reduces the bias on the model. In this study he starts by investigating the co-integration between variables which makes the study complex and very uncertain. Other studies included, studying the impact of the energy source composition change on the causal relationship between GDP and energy. Such changes as the substitution of higher quality energy sources such as electricity for lower quality energy sources such as coal [34] [35]. It was found that, after the innovations, energy granger causes the GDP. A summary of the studies on economic growth and electricity demand is presented in Table 1.1 and Table 1.2

A study by Yu and Jin, which was the first of its kind, tested whether energy and output cointegrate [62]. The study found that there is no relationship that exists between energy use and either employment or an index of industrial production. However, the study by Stern, contradicted these findings [7]. Stern found that single equation static cointegration analysis and the multivariate dynamic cointegration analysis shows that energy is significant in explaining GDP or output.

A new area of study in causality has emerged namely the structural causal models (SCM) [63]. SCM is a structural theory developed in [63][64] which combines features of the structural equation models (SEM) used in economics and social science [65][66], the potential-outcome framework of Neyman [67] and Rubin [68], and the graphical models developed for probabilistic reasoning and causal analysis [63][69][70].

Wold in a paper published in 1954 proposed using a recursive model structure to analyze causal relations among economic time series [71]. Developed with this thinking, SCM is an attempt to provide a framework for modeling causal relations and can be extended to time series causal relations. The objective of the model is to gain the ability to infer cause-effect relations that are implied by the observed time series data. This is done by identifying the complex set of causal processes that have generated the observed data. The task of identification is called structural analysis and is intimately connected with the possibility of guiding policy direction and answering counterfactual questions (for instance, asking what would happen to the economy if particular variable is changed) [72].

Granger causality method focuses on providing accurate modeling of the systems as opposed to providing insight on how to events are linked. SCM, however, provides an analytical framework for causal effect estimation that brings theoretical understanding of the problem particularly the nature of the dependencies. A study on the causality and graphical models in time series analysis was conducted by Eichler and Dahlhaus [73]. SCM is more particularly used for the identification of the causal variables such that the causal effect can be estimated. Similar studies have been conducted by [74] [72]. This explores SCM to analyse the electricity consumption system.

1.2.4 Artificial Intelligence Methods

Artificial intelligence (AI) methods are data-driven non-parametric learning techniques that are used to model relationships that are considered to be nonlinear and complex between variables. These techniques rely on the data to derive mathematical relationships between the variables and they include popularly known technques such as artificial neural networks (ANN), support vector machines (SVM), neurofuzzy systems (ANFIS) and most recently extreme learning machines (ELM). Of all AI methods, the mostly widely used technique is arguably ANN and there are reasons for this.

Firstly, NN is a data driven self-adaptive method and for that reason very few a priori assumptions are required to construct a model. Given enough input and output data ANN is able to extract underlying relationships. Hence, they are described as multivariate nonlinear nonparametric statistical method [75][76] [77]. In this sense the modeling approach has the ability to learn from experience that is captured in the measured data. However, the drawback is that the underlying rules are hard to describe and the data used for modeling can sometimes be distorted by noise.

Secondly, ANN learn from data presented (in-sample-data) and can often correctly infer the unseen part of the data regardless of the noise in the data. This ability to generalize makes ANN most ideal for the forecasting future behavior based on past behavior. Thirdly, ANNs are described as universal approximators in that they have been shown to have the ability to approximate any continuous function with a reasonable accuracy [78].

Finally, ANN is a nonlinear modelling technique which makes it more suitable for modeling nonlinear systems. The real world systems are thought to be nonlinear [79]. What differentiates ANN from other nonlinear modeling tools such as autoregressive conditional heteroskedasticity (ARCH) is that ANN has the ability to perform nonlinear modeling without prior knowledge of the relationship between the inputs and outputs.

The characteristics of ANN that have just been outlined also also extend to other AI techiniques. All the techiques are nonparametric which allows the data to speak for themselves in the sense of determining the form of mathematical relationships between time series vari-

ables A number of studies have been conducted comparing neural networks and Box-Jenkins forecasting ability [80] [81] [82]. The superiority of ANN was also demonstrated in a competition organized through the Santa Fe institute of which winners of each set of data used ANN models [83]. Furthemore, there has been lots of work that has been done on using neural networks in forecasting yielding mixed results. The reason for this is that there is a lack of systematic approaches to neural network model building which is probably the primary cause of inconsistencies in reported findings [84].

The learning method used by neural networks is called error risk minimization (ERM). Neural networks such as Multilayer perceptron (MLP) uses gradient descent method learn the pattern in the data and adjust the network parameters through backpropagation to minimize the error [78]. The learning method, particularly gradient descent method, has a problem of getting stuck in the local minima of the error solution space and is therefore unable to find the universal minimum point. Another limitation with ANN is overfitting. Overfitting happens when the ANN model has more parameters than is required to model the input-output data such that it cannot generalize but has merely memorised the dynamics of the data. During training regularisation parameters are introduced which penalize the complexity of the network and thus avoid overfitting [78].

Support vector machines or regressions models are formulated using the Structural Risk Minimisation (SRM) principle, which has been shown to be superior to the traditional Empirical Risk Minimisation (ERM) principle employed by the conventional neural networks [85]. SRM minimises the upper bound on the expected risk, as opposed to ERM which minimises the empirical error on the training data. SRM equips SVM with a greater ability to generalise which is the goal of statistical learning. This is because SRM enables SVM to overcome the problem of the local minima. SVR avoids underfitting and overfitting of the training data by minimizing the regularization term as well as the training error.

SVR has been applied to short-term load forecasting medium-term forecasting. Chen et al. applied support vector regression for Mid-term Load Forecasting [86]. The study was performed for the European Network on Intelligent TEchnologies for Smart Adaptive Systems

(UNITE). The forecasting period considered for the competition was daily peak loads in January 1999. The SVR model designed by by Chen at al won the competition. Ceperic et al conducted a study short-term load forecasting (STLF) based on the support vector regression machines (SVR) [87]. They applied feature selection algorithms for automatic model input selection and used of the particle swarm global optimization based technique for the optimization of SVR hyper-parameters. They found that SVR yielded results with better accuracy than nonlinear autoregressive (NARX) model. Hong found that an SVR model with immune algorithm (IA) had better forecasting performance than the other methods, namely SVMG, regression model, and ANN model [88].

The drawback with SVR is that determining the proper learning parameters such as C, which defines cost of constraint violation, and ε, the loss function, is still a heuristic process and almost surely suboptimal. In addition, the response speed of trained SVM to external new unknown observations is much slower than feedforward neural networks since SVM algorithms normally generate much larger number of support vectors (computation units) while feedforward neural networks require very few hidden nodes (Computation units) for same applications [89]. Hence, it is not ideal to use SVMs to make real-time prediction since several hours may be spent for such prediction (testing) set.

Neuro-fuzzy systems or adaptive neuro-fuzzy inference (ANFIS) is a combination of neural networks and fuzzy inference system. It involves a procedure where fuzzy sets and rules are adjusted using neural networks tuning techniques in an iterative way with data vectors (input and output system data). Unlike neural networks which is regarded as a black box because its models are difficult to read, ANFIS is regarded as semi-transparent or a grey box. This is because the rule based system used by fuzzy inference system is readable. Research has shown that the neural-fuzzy network has a good performance in time series prediction [90], [91] and for that reason it has also been used for load forecasting. It was found to perform better, in terms of accuracy, than backpropagation neural network when they were both applied in hourly load forecasting for 24 h ahead [92]. Neuro-fuzzy uses clustering methods to organize the data into fuzzy clusters before it fed into a neural networks. In Neuro-fuzzy, the fuzzy inference

brings the advantage of interpretation capability and ease of encoding a priori knowledge [93]. The limitation is that fuzzy inference lacks the learning capability and combined with neural networks, to form neuro-fuzzy, it does not overcome learning weaknesses of ANN.

1.3 Why artificial intelligence

This study seeks to contribute to the forecasting studies in general and electricity consumption forecasting in particular. There has been a shift from simple linear forecasting tools to more complex nonlinear forecasting in the endeavor to find a more suitable tool. Accordingly, novel Artificial intelligence tools are proposed in this work. The use of artificial intelligence techniques such as neural networks falls within the logic of introducing complex methods that are able to deal with the non-stationary data sets. Marvin minsky divided the task of creating an intelligent machine into five main areas: Search, Pattern-Recognition, Learning, Planning, and Induction [94]. A machine searches for solutions in a solution space but the search is often inefficient because of the vast solution space especially now with big data. It is therefore, important to introduce pattern-recognition to make the search efficient by restricting the machine to use its methods only on the kind of attempts for which they are appropriate. The efficiency of the search is further improved through learning which direct Search in accord with earlier experiences. Planning helps in dividing the problem into smaller chunks that make it relatively easy for the machine to search for the solution. Induction is creating model for the machine to generalize on unseen data. Artificial intelligence has come under heavy criticism for the usage of statistical analysis. Noam Chomsky, the world renowned linguist, criticized the definition of success in AI which is defined as getting a fair approximation to a mass of chaotic unanalyzed data [95]. This way of studying AI, he stated, does get the kind of understanding that the sciences have always been aimed at which is to understand the underlying principles of the system. He, however, acknowledged that the statistical analysis gives much better prediction of phenomena than the physics models will ever give.

Artificial intelligence techniques do not require a priori assumptions about the statistical

characteristics of the data or the problem space. Unlike other modeling methods that have to assume that the data is normally distributed and is stationary, AI techniques require no such assumptions. The AI techniques perform the necessary analytical work during training, which ordinarily would require non-trivial effort when using other methods. The proposed AI techniques in this study, which are the extreme learning machines (ELM), will have to overcome the limitations of the AI techniques that are currently widely used which include:

- Finding or coming closer to the universal minima during learning
- Overfitting
- Reduction of computation cost by reducing the computation time
- Improvement of the accuracy
- Overcome the curse of dimensionality

1.4 Forecasting method selection

The selection of methods for forecasting is influenced by multiple factors. These factors serve as a criteria for selecting the appropriate forecasting methods. There are several ways in which the criteria for selecting and comparing forecasting methods are ranked. They could be ranked in order of importance, and accuracy is often given the top priority. Other criteria used is the pattern of the data to be forecast, the type of series, the time horizon to be covered in forecasting and the ease of application.

- Accuracy: Accuracy in forecasting experiments is used to measure the deviation of the forecasted values from the actual values. The lack of accuracy of a forecast reflect other factors, for example, insufficient data or use of a technique that does not fit the pattern of the data [96]. Forecasters generally agree that forecasting methods should be assessed for accuracy using out-of-sample tests rather than test for goodness of fit to past data (in-sample tests) [97]. Fildes and Makridakis [98] concluded that the performance of a model

on data sample outside that used in its training or construction remains the touchstone for its utility in all applications.

- Pattern of the data: The forecasting method has to be able to distinguish between randomness and the underlying pattern of the data. Time series analysis has also revealed that a pattern itself can be thought of as consisting of sub-patterns or components, namely trend, seasonality and cycle [96]. Understanding the three sub-patterns helps in selecting the appropriate forecasting model, since different methods vary in their ability to cope with different kinds of patterns.

- Time horizon: In certain data series sub-patterns change with the length of the time horizon. In the short term, randomness is usually the most important element. Then, in the medium term the cyclical element becomes important and finally in the long term, the trend element dominates. There is generally a greater uncertainty as the time horizon lengthens. The method taken to be able to take these factors into consideration for example, long term horizon would require a method that is adaptive over time.

- Ease of Application: Included under this heading are such things as complexity of the methods, the timeliness of the forecasts it provides, the level of knowledge required for application, and the conceptual basics and the ease with which it can be conveyed to the final user of the forecast [96].

Upon selection these methods, they are used to predict or forecast future values of a data series in this future electricity load.

1.5 Prediction vs. Forecasting

A forecast is merely a prediction about the future values of data. However, it is not a forecast if it does not involve time. Prediction is part of statistical inference. Prediction includes both regression and classification, for example a prediction model can be used to predict a class that a certain object belongs. A forecasting model is used to forecast the magnitude of a parameter

and at a certain point in time. Because this study uses time series data it is therefore referred to as a forecasting study.

1.6 Classification vs. Regression

A lot of work has been done on the topic prediction and forecasting. According to [99], the classification problem can be formally stated as estimating a function $f : R_N(-1, 1))$ based on an input-output training data generated from an independently, identically distributed unknown probability distribution $P(x, y)$ such that f will be able to classify previously unseen (x, y) pairs. The classification approach differs from the regression approach.

Regression involves forecasting raw price values. Regression analysis looks for a relationship between the X variable (sometimes called the "independent" or "explanatory" variable) and the Y variable (the "dependent" variable). This work focuses on regression, using electricity consumption as a dependent variable and economic variables as explanatory variables.

1.7 Methods comparison and statistical tests

Forecasting models, created from the different forecasting tools, yield forecasting results with different levels accuracy when tested on unseen data. These accuracy differences are calculated by using accuracy measure tools such the mean square error (MSE). The differences in residuals may some reflect differences of the particular sample under consideration not necessarily that of the populations from which the data was sampled. It is, therefore important to assess the statistical significance of the differences in accuracy. The statistical significance test, which assesses whether two samples are from the same population through the mean or the variance, is used to perform this assessment. The null hypothesis of the test is that residuals of the accuracy results from different forecasting tools are from the same population. If the null hypothesis is not rejected it may mean that the two forecasting methods are equally capable of approximating or modeling the underlying system under consideration. However, if it is rejected it may mean

that the method that has shown better accuracy, is a better modeling tool. Using the estimate of the population variability and the known sample size, the mean differences of a particular size can be mathematically calculated. For example, for a t test the calculations provides a p value, which is called the significance level, such as $p = 0.05$. P value is a number that illustrates the proportion of the number of times the mean differences can be expected to be as large as or larger than a particular sized difference obtained when sampling from the same population assumed under the null hypothesis [100]. If $p = 0.02$, it means that 2% of the time when sampling a pair of means from the same population, it will have the expected difference. In academic studies a $p = 0.05$ has been adopted as the cut off so that any value larger than 0.05 the null hypothesis can be rejected.

1.8 Applications of demand forecasting and contribution

It is an established fact that forecasting electricity demand is important for capacity planning and more importantly for scheduling. In South Africa, Eskom, which is the power utility, maintains a monopoly over the electricity supply. In 2007, electricity demand in South Africa outstripped the demand and as a result experienced rolling power blackouts. Inglesi and Pouris argue that part of the crisis was exacerbated by the inadequacy of the demand forecasting models used by Eskom [101]. Because of this shortage, Eskom has to continuously implement load shedding programmes until the power plants under construction are connected into the grid. For this reason, planning for maintenance and coal supply scheduling has become very critical. A monthly demand forecasting is a vital element of this planning. The first part of this study provides a one step ahead monthly forecast for the total consumption of electricity in South Africa.

The second part of this study decomposes the consumption forecasting into different economic sectors. Electricity is a commodity that drives all sectors of the South African economy. These sectors include manufacturing, household, agriculture and mining. Each of these sectors is an independent driver of electricity consumption and therefore, each induces its own par-

ticular effect on the total consumption. The end-use of electricity in South Africa is currently (year 2014) divided between domestic (17.2%), agriculture (2.6%), mining (15%), industrial (37.7%), commercial (12.6%), transport (2.6%) and general (12.3%) [102]. Since the beginning of the electricity supply crisis in 2007 in South Africa, the system has been operating at a tight reserve margin. The reserve margin decreased from 15% in 2001 to 7% in 2007 [102]. Medium term forecasts in such circumstances is critical in that maintenance of the generation plants need to be planned such that the peak demand can be met by the supply. Understanding the medium term future demand in South Africa will ensure that there are no rolling blackouts which can have a serious negative impact on the economy. This study is only limited to the manufacturing sector and mining sector. The reason for choosing these two sectors is that they are the most critical sectors of the South African economy and they are the largest consumers of electricity. By conducting these multivariable study, policy makers will be able to understand how the growth of these industries will affect the electricity demand and respond accordingly. Electricity customers in the manufacturing industry have a special requirement. They require un-interrupted supply of electricity so that a whole cycle of production is completed otherwise manufactured goods are wasted. The mining is very prone to accidents that can be very costly in terms of lost lives and therefore, it is important that there is certainty in terms of electricity supply.

This study proposes the use of structural causality model (SCM) to identify a causal variable that can be used on the conditioning set for determining the dependent variable (electricity consumption in the sector under consideration). The SCM framework allows researchers to reason about the problem unlike other causality methodologies that focus on providing accurate modeling of the systems as opposed to providing insight on how to events are linked. The application SCM to electricity demand modelling is the first of its kind and it is a contribution that will be a great addition to the forecasting literature. Under SCM, this work proposes the use graphical causal models to identify the variables to use on the conditioning set.

To perform the estimation of the causal effect, the study proposes the use of extreme learning machines which are a new artificial intelligence tools optimized to make the task of model

estimation quicker and more accurate. ELM is distinct from traditional function approximation approaches, which require the adjustment of input weights and hidden layer biases, in that input weights and hidden layer biases are randomly assigned provided an activation function that is infinitely differentiable is used for training the model. The only free parameters that are learned are the weights between the neurons in the hidden layer and the output layer. Hence, ELM is formulated as linear-in-the-parameter model which means solving a linear problem. In this way, ELM is remarkably efficient and tends to reach the global minimum. OP-ELM is an optimized ELM that introduces pruning techniques to select the optimal number of hidden layers. These tools use the variables identified through SCM to estimate the causal effect. In addition, ELM is also used to construct autoregressive models with the total energy consumption in South Africa.

The use of artificial intelligence to model the South African electricity consumption is almost non-existent. Furthermore, the approach of decomposing the economic sectors and predicting the consumption for each sector is the first of its kind.The work introduced in this thesis is a great contribution to academia and industry.

1.9 Objectives of the thesis

During the past several decades, researchers have developed and applied widely forecasting techniques which enabled them, to a considerable extent, to forecast time series data. It is of great interest to build on this work and explore more forecasting techniques The main objective can be divided in four particular objectives, being:

- Particular objective 1: To propose tools that are able to forecast a one step ahead monthly electricity consumption.
- Particular objective 2: To propose forecasting tools that outperform tools commonly used in the literature.
- Particular objective 3: To propose structural causal model for the identification of the

control parameters for the causal models.

- Particular objective 4: To propose artificial intelligence tools with low computational cost to process the data.

1.10 Overview of the thesis

The remainder of this thesis describes the development of methodologies to model and predict electricty consumption or load . The main objective is to create a model that predicts future electricty load as accurately as possible. Chapter 2 outlines the background in load forecasting and the presents the relavant literature survey.

Forecasting is a topic that has attracted researchers for a long time and there has been a number of studies in this area which including electricty load forecasting. These studies have always sought to find the most suitable mathematical and statistical tools to model sytems. The suitability of each method is function firstly, type of method e.g. parametric or non-parametric, secondly, the type of system e.g stationary or non-stationary, and lastly, modelling aproach e.g. data-driven or expert system. Chapter 3 outlines forecasting techniques from linear techniques to non-linear artificial intelligence (AI) techniques that were used for modelling in this work. The experiments conducted with this techniques is presented in chapter 4.

The load forecasting literature shows that Granger causality has been used widely used in causal studies. However, Granger causal modelling seeks to find the most accurate model without a strong focus on the problem analyis. This work proposes the use of structural causal modelling (SCM) which introduces a way of reasoning about the problem which helps identify the causal variables to be used for causal estimation. The SCM methodology are presented in chapter 5. The variables identified in chapter 5 are used for estimation in chapter 6. The techniques used in chapter 6 are the extreme learning machine and optimally-pruned extreme learning machines. This chapter presents the techniques and the experimets conducted. Finally, Chapter 7 summarizes the findings of this thesis and identifies avenues for further research.

Table 1.1: Causality studies

Authors	Methodology	Hypothesis/Period
Kraft and Kraft [36]	Standard Granger causality	Growth-led energy U.S.A,1947-1974,
Akarca and Long [37]	Standard Granger causality	Growth-led energy, South Africa,1973-1974,
Yu and Hwang [38]	Standard Granger causality	Growth-led energy U.S.A,1973-1981,
Soyatas and Sari [39]	Vector error correction model granger causality	Growth-led-energy, Italy, Japan, South Korea, 1950-1992,
Akinlo [40]	ARDL Bounds test Neutrality.	Nigeria, Cameroon, Ivory Coast, Kenya, Togo, 1980-2003
Wolde-Rufael [41]	Toda and Yomamoto granger causality test	Growth-led-energy, Algeria, Congo, Egypt, Ghana, Ivory coast,1971-2001
Akinlo [42]	Full Modified OLS	Energy-led-growth-led-Energy, Ghana, Senegal, Gambia, 1980-2003
Lee [43]	Vector error correction model granger causality	Growth-led-energy, Ghana, 1975-2001
Twerefo et al [44]	Vector error correction model granger causality	Growth-led-energy, Ghana, 1975-2006
Fatai et al [45]	Toda and Yomamoto	Energy-led-growth-led-Energy, Philippines, 1960-1999
Stern [7]	Cointegration, Granger causality	Energy-led-growth, U.S.A, 1948-1994
Ghali and El-Sakka [46]	Cointegration, VEC Granger causality	Energy-led-growth-led-Energy, Canada, 1961-1997
Ho and Siu [47]	VEC Granger Causality	Energy-led-growth, Hong Kong, 1966-2002

Table 1.2: Causality studies

Soytas and Sari [48]	Toda and Yomamoto causality test	Neutrality, 1960-2000
Payne [49]	Toda and Yomamoto causality test	Neutrality, 1949-2006
Masih [50]	VEC Granger Causality	Energy-led-growth-led-Energy, Taiwan and Energy-led-growth, South Korea, 1952-1992
Haciciglou [51]	Granger causality, Bounds testing	Growth-led-electricity, Turkey, 1968-2005
Tang [52]	ECM based F-test, ARDL	Growth-led-electricity-led-growth, Malaysia, 1972-2003
Morimoto and Hope [53]	Standard granger causality	Electricity-led-growth, Sri Lanka, 1960-1998
Shiu and Lam [54]	Cointegration, ECM	Growth-led-electricity-led-growth,China,1971-2000
Odhiambo [55]	ARDL Bounds test	Growth-led-electricity, Tanzania, 1971-2006
Odhiambo [56]	Standard granger causality	Growth-led-electricity-led-growth, South Africa, 1971-2006
Ghosh [57]	ARDL test	Growth-led-electricity, India, 1970-2006
Ghosh [58]	Standard granger causality	Growth-led-electricity, India, 1950-1997
Narayan and Smyth [59]	Multivariate Granger causality	Growth-led-electricity, Australia,1966-1999
Solarin Sakiru Adebola [60]	Granger causality test	Electricity-led-Growth, 1980-2008
E. Ziramba [61]	Granger causality test Neutral,	Egypt; Growth-led-hydroelectricity, South Africa; Growth-led-hydroelectricity-led-growth, Algeria (1980-2009)

Chapter 2

Artificial Intelligence and Other Modelling Techniques

According to Alfares and Nazeeruddin [103] load forecasting techniques are classified into nine categories. The techniques can be classified in this order:

- Multiple regression;
- Exponential smoothing;
- Iterative reweighted least-squares;
- Adaptive load forecasting;
- Stochastic time series;
- Fuzzy logic;
- Knowledge-based expert systems; and
- Artificial Intelligence

2.1 Multiple regression

Multiple regression applies weighted least-squares to estimate the coefficient of the independent variable. Using this analysis, the statistical analysis between a dependent variable e.g. total electricity load and the independent variables e.g. weather. This has been used widely [104][105][106][107]. The main weakness of these multiple linear regression models is that transformations include a priori or parametric assumptions about the type and consistency of the relation between 2 parameters which may not be met completely. In addition, if the system under consideration exhibits non-linear dynamics these regression models may suffer serious gaps in their representations of the system leading to poor prediction ability.

2.2 Exponential smoothing

Exponential smoothing schemes weight past observations using exponentially decreasing weights. The equation for exponential smoothing is expressed as:

$$y(x) = \beta(t)^T f(t) + \varepsilon(t) \tag{2.1}$$

where $f(t)$ is a fitting function vector of the process, $\beta(t)$ is a coefficient vector,(t) is white noise and T is a transpose. The advantage with exponential smoothing models is their simplicity so that they can be applied to a large number of series quickly. In addition, exponential smoothing are always attractive due to the small number of parameters involved, which make them easy to implement [108]. The disadvantage is that the exponential smoothing models may be too narrow for some data series [109].

2.3 Iterative reweighted least-squares

This algorithm uses an operator that controls one variable at a time. The operator is used to determine an optimal starting point. The algorithm makes use of the autocorrelation func-

tion and partial autocorrelation function of the resulting differenced past data in identifying a sub-optimal model. Iteratively reweighted least squares (IRLS) is one of the most effective methods to minimize the regularized linear inverse problem. Unfortunately, the regularizer is non-smooth and non-convex when $0 < p < 1$. In spite of its properties and mainly due to its high computation cost, IRLS is not widely used in forecasting [110].

2.4 Adaptive load forecasting

The model is adaptive because the model parameters are automatically corrected to keep track of the changing conditions of the system. Kalman filter theory for regression analysis is used under this methodology. The filter is designed in a two-step way:

- A prediction step where a priori estimation determines the optimal one-step-ahead prediction of the former estimate.
- A correction step where the prediction is updated according to a new observation resulting then in the optimal a posteriori estimation of the state vector.

The drawback with the Kalman filter is that it assumes that the moments of the noises are known, which is often untrue. As the noises are usually centered, only variances are considered [111].This approach becomes inaccurate as soon as there is significant change in the time series. An adaptive Kalman filter has been introduced to overcome these limitations. However, finding the right balance between adaptivity, reliability and forecast range may require tedious efforts when complex systems are considered [112].

2.5 Stochastic time series

There are various time series methods that are used for forecasting. These includes autoregressive model (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), etc [113].

2.5.1 Autoregressive model

This method used with the assumption that the current value or the forecast is the linear combination of the previous values of the same parameter. In case of electricity load, the current will be assumed to the linear combination of the previous period load values and can be represented in an equation as follows:

$$\hat{L_t} = -\sum_{j=1}^{k} \phi_t j L_{t-} + \varepsilon_t \tag{2.2}$$

where $\hat{L_t}$ is the predicted load at time t, ε_t is a random load disturbance and $\phi_j, j = 1, \ldots, k$ are unknown coefficients of order k.

2.5.2 Autoregressive moving average model

In the method the current value of a parameter is expressed linearly in terms of its values at a previous periods and in terms of the previous values of white noise. An equation for ARMA can be written as follows:

$$y(t) = \phi_1 y(t-1) + \cdots + \phi_k y(t-k) + \alpha(t) - \theta_1 \alpha(t-1) + \cdots + \theta_n \alpha(t-n) \tag{2.3}$$

An ARMA model with a lag p for the variable value and lag q for white noise is written as ARMA(p, q).

2.5.3 Autoregressive integrated moving average

If the data under consideration exhibits nonstationarity over time then it has to be differenced to transform the series into a stationary series. A time series that is differenced d times has an ARIMA model written as ARIMA(p, d, q). The three stochastic models considered above are parametric which means that the models parameters are pre-specified. As a result the models lack flexibility and are unable to predict turning points in the data series.

2.6 Fuzzy logic

Fuzzy Logic was initiated in 1965 [114], [115], [116], by Lotfi A. Zadeh , professor for computer science at the University of California in Berkeley. Basically, Fuzzy Logic (FL) is a multivalued logic,that allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low, etc.

The drawback with fuzzy logic is that, they rely on fuzzy rules that are extracted from experts' and operators' experience, which can be inconsistent and thus unreliable [117]. Overcome this problem neural networks learning techniques have introduced to construct models such as neuro-fuzzy models or adaptive neuro-fuzzy inference system which will be covered later in this chapter.

2.7 Expert systems

Expert systems, combines rules and procedures used by human experts in the field of interest to create a software algorithm that is then able to automatically make forecasts without human assistance [5]. The process of developing the software is dependent on the availability of the human expert to work with software developers for a considerable amount of time in imparting the expert's knowledge to the expert system software. In addition, it is helpful if expert's knowledge is appropriate for codification into software rules. The drawback of this approach is the over-reliance on knowledge of the expert.

Ho et al. [118] proposed the use of the knowledge-based expert system for the short-term load forecasting of the Taiwan power system. They developed an algorithm that performed better compared to the conventional Box-Jenkins method.

2.8 Neural Networks

The theory of neural network computation provides interesting techniques that mimic the human brain and nervous system. A neural network is characterized by the pattern of connections

among the various network layers, the numbers of neurons in each layer, the learning algorithm, and the neuron activation functions. In general, a neural network is a set of connected input and output units where each connection has a weight associated with it. Neural networks can be used for classification or regression. For this study neural networks were used as a regression tool for predicting the future price of a stock market index.

Neural networks gained interest after McCulloch and Pitts introduced a simple version of a neuron in 1943 [78]. This model of a neuron was inspired by the biological neuron in the human brain and it was presented as a simple mathematical model. Neural network is a network consisting of neurons and paths connecting the neurons. They are interconnected assemblies of simple processing nodes whose functionality is loosely based on the animal neuron. NN can also be defined as generalizations of classical pattern-oriented techniques in statistics and engineering areas of signal processing, system identification and control. Figure 5.8 shows a neural network model with the major components of the network.

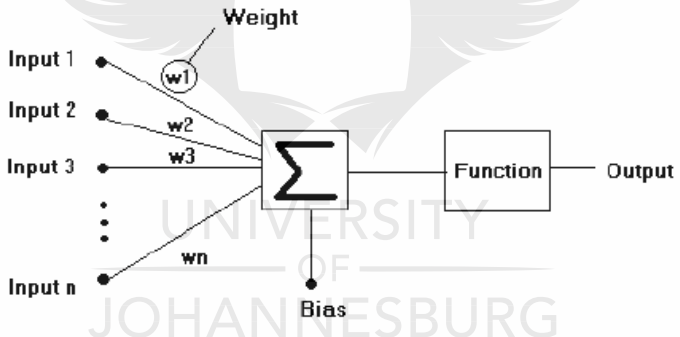

Figure 2.1: Architecture of a neuron

Each input is multiplied by weights along its path and the weighted inputs are then summed and biased. This weighted input is then biased by adding a value unto the weighted input. The output of the summation is sent into a function which is called an activation function which the user specifies (linear, logistic). The output of the function block is fed to the output neuron.

Rosenbatt introduced the concept of a perceptron in the late fifties. A perceptron is considered as a more sophisticated model of the neuron. The perceptron as pattern classifier can solve classification problems with various number of data classes depending on the number of neurons incorporated. Minsky and Papert showed in 1969 that for the correct classification, the

data classes have to be linearly separable which was a major setback [78]. They then further suggested that a two layer feed-forward network can overcome many restrictions, but they did not present a solution of how to adjust the weights from the input to the hidden units.

Studies of neural networks were revived again in the eighties when Kohonen introduced self organising maps (SOM) [78]. SOMs use an unsupervised learning algorithm for applications such as data mining. At about the same time Hopfield was building a bridge between neural computing and physics. Hopfield networks, which are initialised with random weights continously computes until it reaches a final state of stability. For physicists, a Hopfield network resembles a dynamical system falling into a state of minimal energy.

The neural networks research was gained a tremendous momemtum by the discovery of the backpropagation algorithm in 1986. This learning algorithm has gone unchallenged as the most popular learning algorithm for training multilayer perceptrons. The central idea of the error backpropagation algorithm is to determine the errors of the hidden layers of the multilayer perceptron. These errors are determined by back-propagating the errors of the units of the output layer through the network. Then there was the discovery of radial basis functions (RBF) in 1988 [78]. RBF came as an alternative to multilayer perceptron for finding a solution to the multivariable interpolation problem.

2.8.1 Network Topologies

This section presents the pattern of connections between the units and the propagation of data. The pattern of connections can be distinguished as follows:

- $Feedforward\ Networks$, where the data flow from the input to the output units is strictly feedforward as shown in fig. 2.2. This type of connection has no feedback connection.

- $Recurrent\ networks$, that contain feedback connections as shown in fig. 2.3. Contrary to feedfoward networks, the dynamic properties of the network are important.

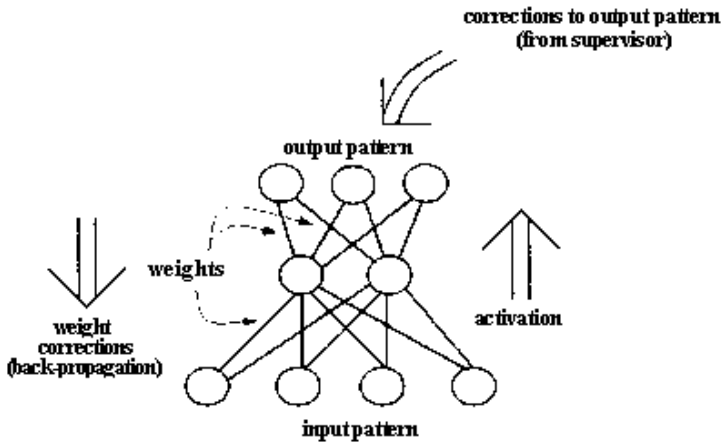

Figure 2.2: A diagram of the feedfoward neural networks

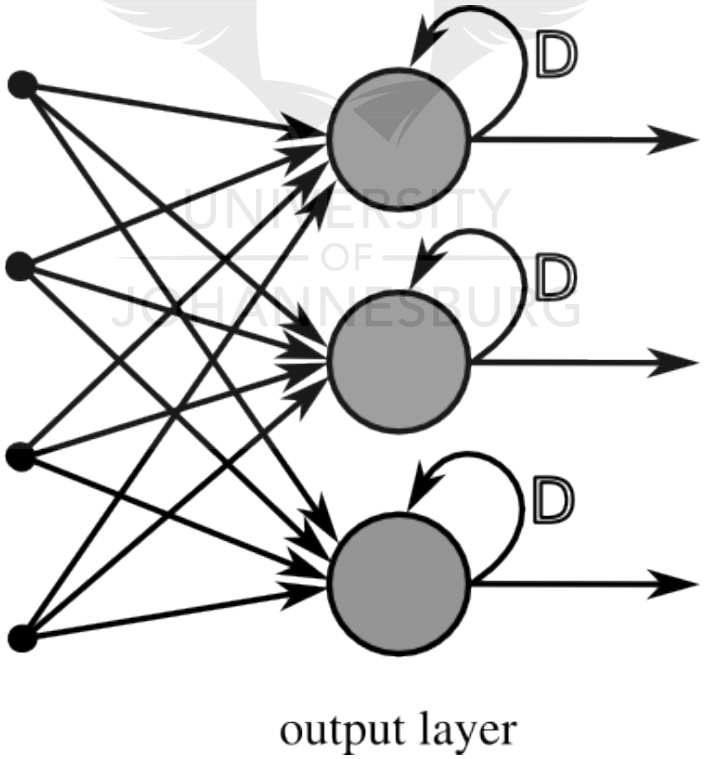

Figure 2.3: A diagram of the recurrent neural networks

It has been widely accepted that a three-layer feed forward network (i.e., one type of multilayer perceptrons models). According to Hornik et al. [119], it is widely accepted that a three-layer feedforward network with an identify transfer function in the output unit and logistic functions in the middle-layer units can approximate any continuous functions arbitrarily well, given suficiently many middle-layer units. This research also uses a three-layer feedback network with a backprogation learning algorithm which is the focus of the next section.

2.8.2 Multi-layer Perceptron

Neural network in it's simplest form has a single layer with directed inputs, and it is only limited to linearly separable classes as a classifier. In order for the network to deal with more complex non-linear problems, hidden non-linear layers are added to form a multilayer perceptron. MLP is structured in a feedfoward topology whereby each unit gets it's input from the previous one. A diagram of a generalised multilayer neural network model is shown in Fig. 2.4.

Neural networks are most commonly used as function approximators which map the inputs of a process to the outputs. The reason for their wide spread use is that, assuming no restriction on the architecture, neural networks are able to approximate any continuous function of arbitrary complexity [120].

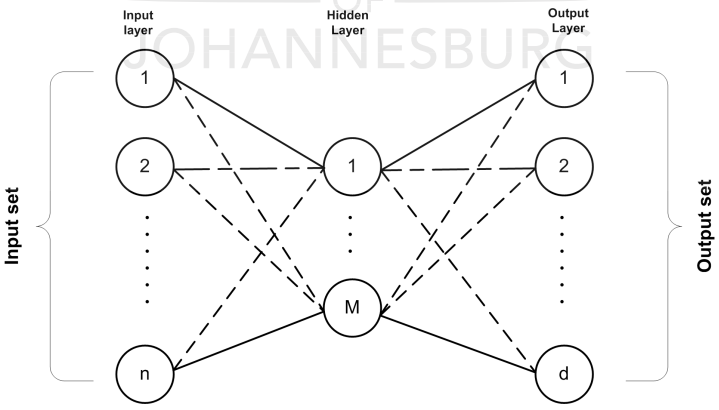

Figure 2.4: A diagram of a generalised neural network model

The mapping of the inputs to the outputs using an MLP neural network can be expressed as

follows:

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{2.4}$$

In Eq. 2.4, $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$ indicate the weights in the first and second layers, respectively, going from input i to hidden unit j, M is the number of hidden units, d is the number of output units while $w_{j0}^{(1)}$ indicates the bias for the hidden unit j and $w_{k0}^{(2)}$ indicates the bias for the output unit k. For simplicity the biases have been omitted from the diagram.

The inputlayer consists of just the inputs to the network.The input-layer neurons do not perform any computations, they merely distribute the inputs to the weights of the hidden layer. Then, it follows a hidden layer, which consists of any number of neurons, or hidden units placed in parallel. Each neuron performs a weighted summation of the inputs, which then passes a nonlinear activation function, also called the neuron function.

Activation Function

The activation function can also be called the transfer function of a neural networks system. This function mathematically defines the relationship between the inputs and the output of a node and a network. The activation function introduces non-linearity that is important to the neural networks applications. It is the non-linearity or the ability to model a non-linear function that makes MLP so powerful. A study was conducted by Chen and Chen to identify general conditions for a continuous function to qualify as an activation function. In practice they found that only a small number on bounded, monotonically increasing and differentiable activation functions are used. These includes the sigmoidal function, the hyperbolic tanget function, the sine or cosine function and the linear function.

Zhang et al concludes that it is not clear whether different activation functions have major effects on the perfomance of the networks. A network may have different activation functions for different nodes in the same or different layers (Schoneburg 1990 and wong 1991). Functions such as tanh or arctan that produce both the positive and the negative values tend to yield faster

training than functions that produce only positive values such as logistic function because of better numerical conditioning. For continuous-valued targets with a bounded range, the logistic and tanh functions can be used, provided that either the outputs are scaled to the range of the targets or the targets are scaled to the range of the output activation. The latter option has been chosen for the purpose of this research. Furthermore, the tanh represeted in Eq. 2.6 and the logistic functions represeted in Eq. 2.5 are used as activation functions of the hidden layer and the output layer respectively. Tanh function shown in fig 2.6 is chosen for the hidden because it converges faster to a solution and therefore reduces the cost of computation for the hidden layer with multiple nodes. The financial time series under consideration is highly non-linear and as a result it requires a sufficiently non-linear to represent all the properties of this series. Hence, non-linear logistic function shown in fig 2.5 is chosen to output layer.

$$g(v) = \frac{e^v}{1 + e^v} \tag{2.5}$$

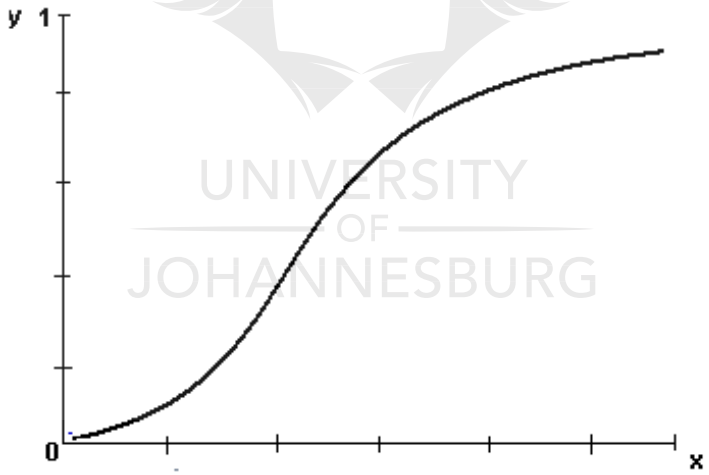

Figure 2.5: A diagram of the sigmoid activation function

where v is the is the result of the weighted summation of each neuron. The neurons in the output layer are linear and they only compute the weighted sum of the inputs.

$$g(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} \tag{2.6}$$

The MLP architecture is a feedforward structure whereby each unit receives inputs only

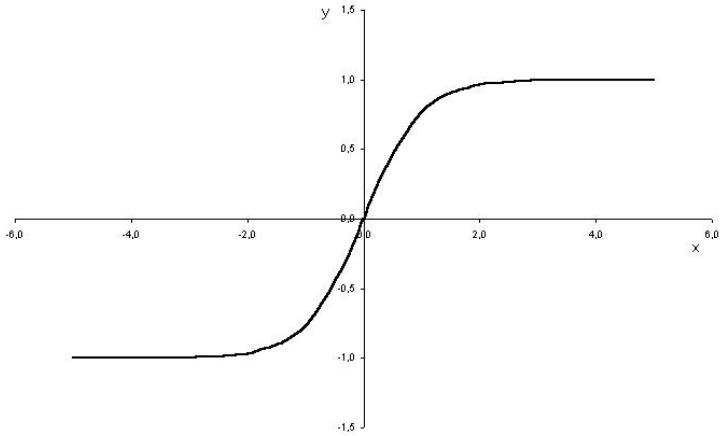

Figure 2.6: A diagram of a tanh activation function

from the lower layers units. Gradient methods are used to find the sets of weights that work accurately for the practical cases. Backpropagation is also used to compute derivatives, with respect to each weight in the network, of the error function. The error function generally used in the neural network computation is the squared difference between the actual and desired outputs. The activities for each unit are computed by forward propagation through the network, for the various training cases. Starting with the output units, backward propagation through the network is used to compute the derivatives of the error function with respect to the input received by each unit.

2.8.3 Network training

Given a training set comprising a set of input X_n, where $n = 1, ...N$, together with a corresponding set of target vectors t_n, the objective is to minimise the error function.

$$E(w) = \frac{1}{2} \sum_{n}^{N} ||y(x_n, w) - t_n)||^2 \tag{2.7}$$

where E is the total error all patterns, the index n ranges over the set of input patterns. The variable t_n is the desired output for the nth output neuron when the nth pattern is presented, and $y_{n,w}$ is the actual outputof the nth output neuron when pattern h is presented.

This type of learning is called supervised learning, where every input has an associated

target output. After the computation of the error the weight vector is then updated as follows:

$$w_{new} = w_{old} - \nabla E(w) \tag{2.8}$$

where $\nabla E(w)$ is the gradient.

$$\nabla E(w) = [\frac{\partial}{\partial w_0}, \frac{\partial}{\partial w_1}, ..., \frac{\partial}{\partial w_n}] \tag{2.9}$$

so each k, w can be updated by:

$$w_k = w_k + \Delta w_k \tag{2.10}$$

where

$$\Delta w_k = -\eta \frac{\partial E}{\partial w_k} \tag{2.11}$$

where η is the learning rate.

The weights of the neural network are optimised via backpropagation training using, most commonly, scaled conjugate gradient method [78]. The cost function representing the objective of the training of the neural network can be defined. The objective of the problem is to obtain the optimal weights which accurately map the inputs of a process to the outputs. The gradient descent method suffers the problems of slow convergence, inefficiency and of robustness. Thus, it is very sensitive to the choice of the learning rate. Smaller learning rates tends to slow the learning process and larger learning rates creates oscillations. One way to achieve a faster learning without experiencing oscillations is to introduce a momentum term which essentially minimise the tendency to oscillate (reinhart). By introducing the momentum term Eq. 5.6 changes to the following format:

$$\Delta w_k = -\eta \frac{\partial E}{\partial w_k} + \alpha \Delta w_k \tag{2.12}$$

The standard backpropagation techniques with a momentum term has been adopted by most

researchers. There are few known ways of selecting the learnign rate parameters, and as a result the parameters are usually chosen through experimentation. Learning rate and momentum take any value between o and 1. Starting with a higher learning rate and decreasing as training proceeds is common practice. McClelland and Rumelmert have indicated that the momentum term is especially useful in error spaces containing long ravines that are characterised by steep, high walls and a gently sloping floor. By using the momentum term the use of very small learning rate is avoided which requires excessive training time.

The learning algorithm and number of iterations determines how good the error on the training data set is minimized meanwhile the number of learning samples determines how good the training samples represent the actual function. The perceptron learning rule is a method for finding the weights in a network. The perceptron has the property that if there exist a set of weights that solve the problem, then the perceptron will find these weights. This rule follows a linear regression approach, that is, given a set of inputs and output values, the network finds the best mapping from inputs to outputs. Given an input value which was not in the set, the trained network can predict the most likely output value. This ability to determine the output for an input the network was not trained with is known as generalization.

MLP with a sigmoid transfer function in the hidden layer and linear transfer functions in the output layer can approximate any function provided a sufficient number of hidden units are available [78]. These hidden units make use of non-linear activation functions. The sigmoid activation function was used in this work.

2.8.4 Radial Basis Function Network

A radial basis function (RBF) is a two layer neural network with a radially activated function on each hidden unit [78]. RBF has an architecture dipicted on Fig. 2.7. Given a data set (y_i, t_i), $i \epsilon N$ of input vectors y_i and associated targets t_i,measured in the presence of noise. The input vector is $Y = y_1, y_2, y_3, ...y_n$ is a collection of inputs in n dimensional space.In the case of regression model, the output is a scalar t and represents the target value of a single function $t = f(y_1, y_2, ..., y_n)$. For a classification problem the output is a vector $T = (t_1, t_2, ..., t_n)$

and represents p functions, like posterior probabilities of different classes. RBF Networks are universal approximators of any continuous functions in regression and classification.

Each input y_i is passed to each node of a hidden layer. Nodes of a hidden layer are RBF functions which perform nonlinear mapping of inputs to a new feature space. Then outputs are fitted in a nonlinear transformed space using Least Squares approximation or relative technique.

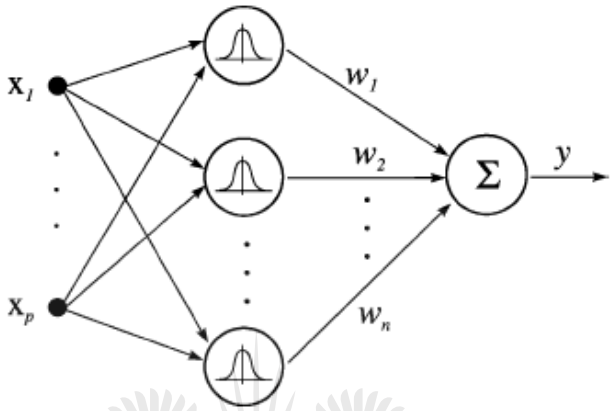

Figure 2.7: Radial basis function network.

This network can be represented in an equation as follows:

$$y = \sum_{i=1}^{m} w_i \phi_i(x) \tag{2.13}$$

where the basis function $\phi_i(x)$ is chosen to be the gausian function. The gausian function for RBF is given by:

$$\phi(x)_i = exp(\frac{||x - c_i||^2}{2\sigma^2}) \tag{2.14}$$

The architecture of RBF shows that the weights that are adjusted during training are found in the output layer only. The connections from the input layer to the hidden layer are fixed to unit weights. The free parameters of RBF networks are the output weights w_i and the parameters of the basis functions (centers c_i and radii σ_i). Since the networks output is linear in the weights w_i,these weights can be estimated by least-squares methods. The output of the first layer for an

input data point d_i is computed as follows:

$$o_{ki} = \phi_i(x)d_i \tag{2.15}$$

these outputs are then put into a matrix O =[o_{ki}]. The output of the first stage is then taken to the next stage which involves the introduction of weights vector $w = [w_1, w_2, ...w_i]$. The output of the network can be written in the following matrix:

$$y^* = Vw \tag{2.16}$$

the objective is to find a weight vector that will minimise the error between the actual output and the output of the network, $e = y^* - y$. To solve for w from Eq.2.16 a method of pseudo-inverse is introduced because V may not be a sqaure matrix and the equation is written as follows:

$$w = [V^T V]^{-1} V^T y^* \tag{2.17}$$

The adaptation of the RBF parameters c_i and σ_i is a nonlinear optimization problem that can be solved by the gradient-descent method.

Training RBF

A training set is an m labelled pair (y_i, t_i) that represents associations of a given mapping or samples of a continuous multivariate function. The sum of squared error criterion function can be considered as an error function E to be minimized over the given training set. That is, to develop a training method that minimizes E by adaptively updating the free parameters of the RBF network. These parameters are the receptive field centres mj of the hidden layer Gaussian units, the receptive field widths sj, and the output layer weights (wij). Because of the differentiable nature of the RBF network transfer characteristics, one of the training methods considered here was a fully supervized gradient-descent method over E.

In particular, μ_j, σ_j and w_ij are updated as follows:

$$\Delta\mu_j = -\rho_\mu \nabla_{\mu_j} E \tag{2.18}$$

$$\Delta\sigma_j = -\rho_\sigma \frac{\partial E}{\partial \sigma_j} \tag{2.19}$$

$$\Delta w_j = -\rho_w \frac{\partial E}{\partial w_l j} \tag{2.20}$$

where ρ_μ,ρ_σ , and ρ_w are small positive constants. This method is capable of matching or exceeding the performance of neural networks with back-propagation algorithm, but gives training comparable with those of sigmoidal type of neural networks [14].

The training of the RBF network is radically different from the classical training of standard NNs. In this case, there is no changing of weights with the use of the gradient method aimed at function minimization. In RBF networks with the chosen type of radial basis function, training resolves itself into selecting the centres and dimensions of the functions and calculating the weights of the output neuron. The centre, distance scale and precise shape of the radial function are parameters of the model, all fixed if it is linear. Selection of the centres can be understood as defining the optimal number of basis functions and choosing the elements of the training set used in the solution. It was done according to the method of forward selection15. Heuristic operation on a given defined training set starts from an empty subset of the basis functions. Then the empty subset is filled with succeeding basis functions with their centres marked by the location of elements of the training set; which generally decreases the sum-squared error or the cost function. In this way, a model of the network constructed each time is being completed by the best element. Construction of the network is continued till the criterion demonstrating the quality of the model is fulfilled. The most commonly used method for estimating generalization error is the cross-validation error.

For the purpose of this work RBF is used to create a neurofuzzy model in combination with fuzzy logic. The combination allows for the optimization methods such as gradient descent

methods from the area of neural networks to be used to optimize parameters in a fuzzy system. A detailed explanation can be found in section 3.3 of this chapter.

2.9 Support Vector Machine

Traditional neural network approaches have challenges of generalisation, and occasionally produce models that can overfit the data. This is a consequence of the optimisation algorithms used to select the model parameters and the statistical measures used to select the model that approximates the relationship between the input and the output. The foundations of support vector machines (SVM) have been developed by Vapnik [85] and are gaining popularity due to multiple attractive features and promising emperical perfomance. SVM models are formulated using the Stuctural Risk Minimisation (SRM) principle, which has been shown to be superior [89] to traditional Emperical Risk Minimization principle, employed by the convectional neural networks. SRM minimises the upper bound on the expected risk, as opposed to ERM which minimises the empirical error on the training data. SRM equips SVM with a greater ability to generalise which is the goal of statistical learning. SVMs were developed for solving classification problems, but recently they have been extended to the domain of regression problems [121].

2.9.1 Support Vector Machine Classification

SVM uses linear models to implement nonlinear class boundaries through some nonlinear relationship of mapping the input vectors x into the high-dimensional feature space. A linear model constructed in the new space can represent a nonlinear decision boundary in the original space. In the new space, an optimal separating hyperplane is constructed. Thus, SVM is known as the algorithm that finds a special kind of linear model, the maximum margin hyperplane. The maximum margin hyperplane gives the maximum separation between the decision classes. The training examples that are closest to the maximum margin hyperplane are called support vectors. All other training examples are irrelevant for defining the binary class boundaries. For the

linearly separable case, a hyperplane separating the binary decision classes in the three-attribute case can be represented as [121]:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \tag{2.21}$$

where y is the outcome, x_i are the attribute values, and there are four weights w_i to be learned by the learning algorithm. In Eq. 2.21, the weights w_i are parameters that determine the hyperplane. The maximum margin hyperplane can be represented by the following equation in terms of the support vectors [121]:

$$y = b + \sum \alpha_i y_i x\left(i\right) \cdot x \tag{2.22}$$

where y_i is the class value of training example $x(i)$, $\cdot$ represents the dot product. The vector x represents a test example and the vectors $x(i)$ are the support vectors. In this equation, b and α_i are parameters that determine the hyperplane. From the implementation point of view, finding the support vectors and determining the parameters b and α_i are equivalent to solving a linearly constrained quadratic programming (QP). As mentioned above, SVM constructs a linear model to implement nonlinear class boundaries by transforming the inputs into high-dimensional feature space. For the nonlinearly separable case, a high-dimensional version of Eq. 2.22 is simply represented as follows [121]:

$$y = b + \sum \alpha_i y_i K\left(x\left(i\right), x\right) \tag{2.23}$$

The function $K\left(x\left(i\right), x\right)$ is defined as the kernel function. There are some different kernels for generating the inner products to construct machines with different types of nonlinear decision surfaces in the input space. Choosing among different kernels the model that minimises the estimate, one chooses the best model. Common examples of the kernel function are the polynomial kernel $K\left(x, y\right) = \left(xy + 1\right)^d$ and the Gaussian radial basis function $K\left(x, y\right) = exp\left(-1/\delta^2 \left(x - y\right)^2\right)^d$ where d is the degree of the polynomial kernel and δ^2 is the

bandwidth of the Gaussian radial basis function kernel. For the separable case, there is a lower bound 0 on the coefficient α_i in Eq. 2.23. For the non-separable case, SVM can be generalised by placing an upper bound C on the coefficients α_i in addition to the lower bound [121].

2.9.2 Support Vector Regression

Given a training data set $D = (y_i, t_i)|i = 1, 2, ...n$,of input vectors y_i and associated targets t_i,the objective is to fit a function $g(y)$which approximates the relation inherited between the data set points and this function can then be used to infer the output t for a new input point y. The deviation of the estimated function from the true one is measured by a loss function $L(t, g(y))$. There are different types of loss functions namely linear, quadratic,exponential, etc. For the purpose of this study Vapnik's loss function is used, which is also known as ϵ -sensitive loss function and defined as:

$$L(t, g(y)) = 0 \quad if \quad \|t - g(y)| \leq \varepsilon \tag{2.24}$$

$$|t - g(y)| - \varepsilon \quad otherwise \tag{2.25}$$

where $\varepsilon > 0$ is a predefined constant that controls the noise tolerance. With the ε -insensitive loss function, the goal is to find $g(y)$ that has at most ϵ deviation from the actual target t_i for all training data, and at the same time is as flat as possible. The regression algorithm does not care about errors as long as they are less than ϵ, but will not accept any deviation larger than ε. The ε-tube is illustrated in fig 2.8.

The estimated function is first given in a linear form taking such as:

$$g(y) = w.y + b \tag{2.26}$$

The goal of a regression algorithm is to fit a flat function to the data points. Flatness in the case of Eq. 2.26 means that one seeks a small w. One way to ensure this flatness is to minimise

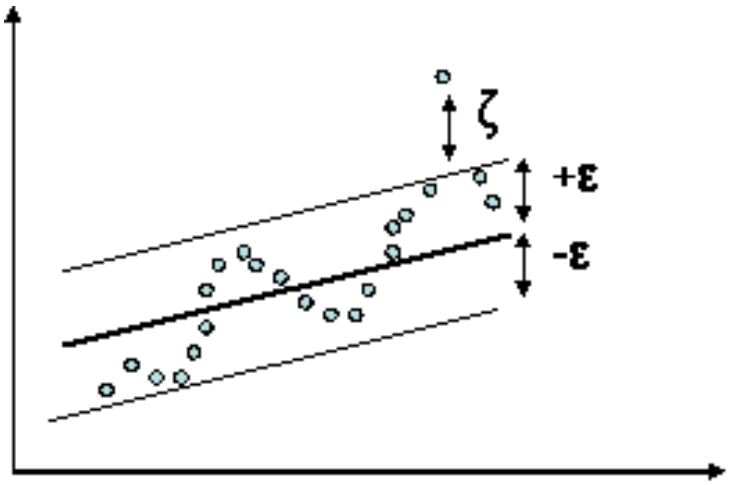

Figure 2.8: Support vector regression to fit a tube with radius ε to the data and positive slack variables ζ_i measuring the points lying outside of the tube

the norm, i.e. $||w||^2$. Thus, the regression problem can be written as a convex optimization problem:

$$minimise \ \frac{1}{2}||w||^2 \tag{2.27}$$

$$subject\ to \ t_i - (w.y + b) \leq \varepsilon \quad and \tag{2.28}$$

$$(w.y + b) - t_i \leq \varepsilon \tag{2.29}$$

The implied assumption in Eq. 2.28 is that such a function g actually exists that approximates all pairs (y_i, t_i) with ϵ precision, or in other words that the convex optimization is feasible. Sometimes, however, this may not be the case or we may want to allow some errors. Analogously to the "soft margin" loss function [122]which was adapted to SVM machines by Vapnik [85], slack variables ζ_i, ζ_i^* can be introduced to cope with otherwise infeasible constraints of the optimization problem in Eq. 2.28 Hence the formulation stated in [85]is attained:

$$minimise \ \frac{1}{2}||w||^2 + C\sum(\zeta_i + \zeta_i^*) \tag{2.30}$$

$$subject\ to \quad t_i - (w.y + b) \leq \varepsilon + \zeta_i \tag{2.31}$$

$$(w.y + b) - t_i \leq \varepsilon + \zeta_i^* \tag{2.32}$$

$$\zeta_i, \zeta_i^* \geq 0 \tag{2.33}$$

The constant $C > 0$ determines the trade-off between flatness of g and amount up to which deviations larger than ϵ are tolerated. This corresponds to dealing with the so-called ϵ-sensitive loss function which was described before.

It turns out that in most cases the optimization problem Eq. 2.31 can be solved more easily in its dual formulation. Moreover,the dual formulation provides the key for extending SVM to non-linear functions.

The minimization problem in Eq. 2.31 is called the primal objective function. The basic idea of the dual problem is to construct a Lagrange function from the primal objective function and the corresponding constraints, by introducing a dual set of variables. It can be shown that the Lagrange function has a saddle point with respect to the primal and dual variables at the solution see ([123, 124]). The primal objective function with its constraints are transformed to the Lagrange function. Hence the dual variables (lagrange multipliers) from the lagrange function have to satisfy positivity constraints. This means that they have to be greater or equal to zero.

It follows from the saddle point condition that the partial derivatives of the lagrange function with respect to the primal variables $(w, b, \zeta_i, \zeta_i^*)$ have to vanish for optimality. The outcome of the partial derivatives are then substituted into Eq. 2.28 to yield the dual optimozation problem:

$$maximise -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(y_i, y_(j)) - \varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{2.34}$$

$$subject\ to \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) = 0\ and\ \alpha_i^*, \alpha_i \in [0, C] \tag{2.35}$$

Eq. 2.26 can be rewritten as follows:

$$w = \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*), \quad thus \tag{2.36}$$

$$g(y) = \sum_{i=1}^{n} (y_i, y)(\alpha_i - \alpha_i^*) + b \tag{2.37}$$

This is called the Support Vector Regression Expansion, i.e. w can be completely described as a linear combination of the training patterns y_i.

In a sense, the complexity of a function's representation by SVs is independent of the dimensionality of the input space and depends only on the number of SVs.

Moreover, the complete algorithm can be described in terms of dot products between the data. Even when evaluating $g(y)$, the value of w does not need to be computed explicitly. These observations will come in handy for the formulation of the nonlinear extension.

The Karush-Kuhn-Tucker (KKT)conditions [125][126] are the basics for the lagrangian solution. These conditions state that at the solution point, the product between dual variables and constraints has to finish.

Several useful conclusions can be drawn from these conditions. Firstly only samples (y_i, t_i) with corresponding $\alpha_i^* = C$ lie outside the ϵ insensitive tube. Secondly, $\alpha_i, \alpha_i^* = 0$ which are both simultaneoulsy nonzero. this allows us to conclude:

$$\varepsilon + -t_i + w.y_i + b \geq 0\ and\ \zeta_i = 0 \qquad if\ \alpha_i \leq C \tag{2.38}$$

$$\varepsilon - t_i + w.y_i + b \leq 0 \qquad if\ \alpha_i > 0 \tag{2.39}$$

It follows that only for $|g(y)| \geq \varepsilon$ the lagrange multipliers may be nonzero or in other words for all samples inside ϵ-tube the α_i, α_i^* vanish: for $g(y) < \varepsilon$ the second factor is nonzero, hence

α_i, α_i^* has to be zero such that the KKT conditions are satisfied. therefore there is a sparse expansion of w in terms of y_i. The training samples that come with nonvanishing coefficients are called support vectors.

There are many ways to compute the value of b in Eq. 2.37. One of the ways can be found in [89]:

$$b = \frac{1}{2}(w.(y_r + y_s)) \tag{2.40}$$

where y_r and y_s are the support vectors

The next step is to make the SVM algorithm nonlinear. As noted in the previous section, the SVM algorithm only depends on dot products between patterns y_i. Hence it suffices to know $K(y, y') = \psi(y, y')$ rather than ψ explicitly which allows us to restate the SVM optimisation problem as:

$$maximise - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(y, y') - \varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{2.41}$$

$$subject\ to \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) = 0\ and\ \alpha_i^*, \alpha_i \in [0, C] \tag{2.42}$$

Likewise the expansion of g in Eq. 2.37 maybe written as follows:

$$w = \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*))K(y_i),\ and \tag{2.43}$$

$$g(y) = \sum_{i=1}^{n}(y_i, y)(\alpha_i - \alpha_i^*)K(y_i, y) + b \tag{2.44}$$

Even in the nonlinear setting, the optimization problem corresponds to finding the flattest function feature space, not in input space. The SVM kernel functions that can be used can be found in [127]. Roughly speaking, any positive semi-definite reproducing kernel hilbert space

(RKHS)is an admissable kernel.

2.9.3 SVM in a nutshell

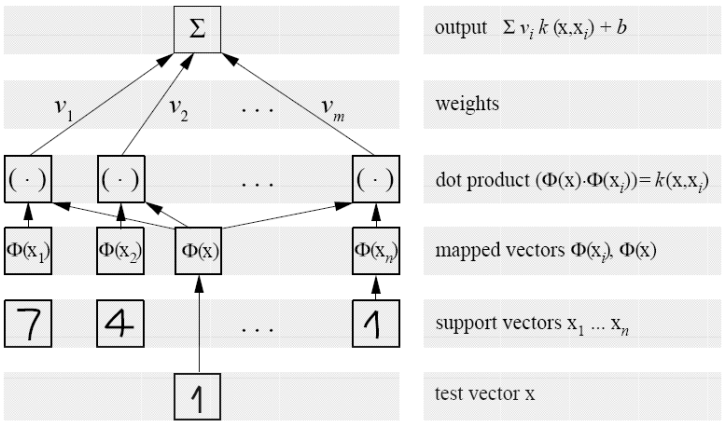

Figure 2.9: Architecture of a regression machine constructed by the SV algorithm.

Figure 2.9 illustrates graphically the different steps in the regression stage. The input pattern (for which a prediction should be made) is mapped into feature space by a map Φ. Then dot products are computed with the images of the training patterns under the map Φ. This corresponds to evaluating kernel k functions at locations $k(x_i, x)$. Finally the dot products are added up using the weights $\alpha_i - \alpha_i^*$. This, plus the constant term b yields the final prediction output. The process described here is very similar to regression in a three layered neural network, the difference is that in the SV case the weights in the input layer are predetermined by the training patterns.

Figure 2.10 shows how the SV algorithm chooses the flattest function among those approximating the original data with a given precision. Although requiring flatness only in feature space, one can observe that the functions also are very flat in input space.

Finally, Fig 2.11 shows the relation between approximation quality and sparsity of representation in the SV case. The lower the precision required for approximating the original data, the fewer SVs are needed to encode that. The non-SVs are redundant i.e. even without these patterns in the training set, the SV machine would have constructed exactly the same function f. One might think that this could be an efficient way of data compression, namely by

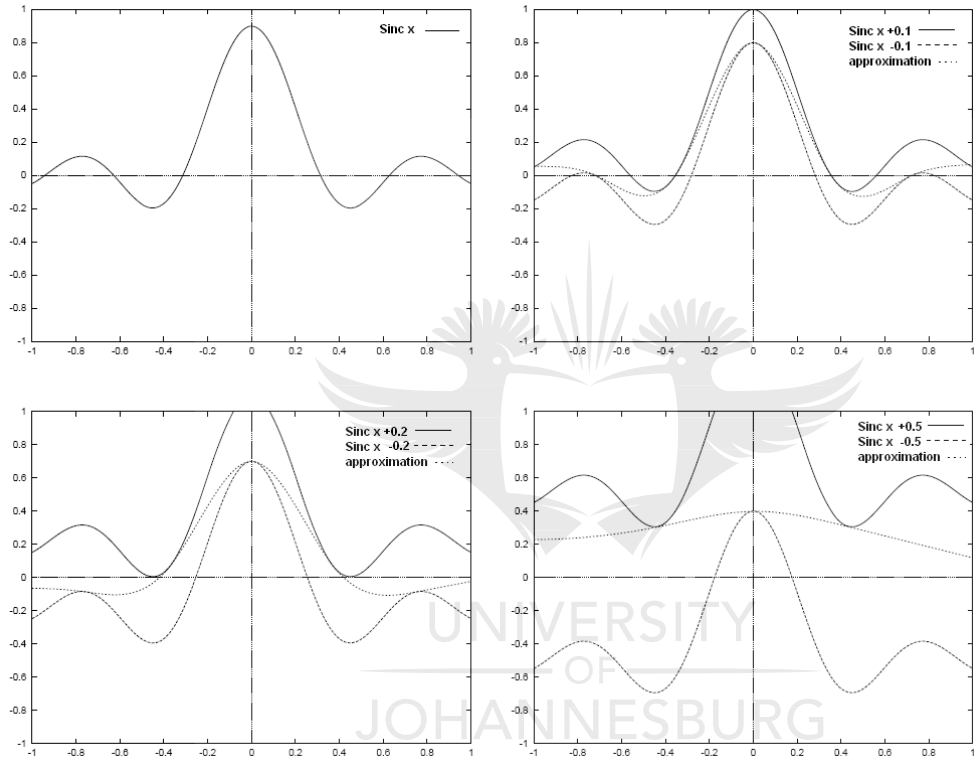

Figure 2.10: Upper left: original function $sincx$ upper right: approximation with $\varepsilon = 0.1$ precision (the solid top and the bottom lines indicate the size of the $\varepsilon - tube$ the dotted line in between is the regression)lower left:$\varepsilon = 0.2$, lower right: $\varepsilon = 0.5$

storing only the support patterns, from which the estimate can be reconstructed completely. However, this simple analogy turns out to fail in the case of high-dimensional data, and even more drastically in the presence of noise. In (Vapnik et al) one can see that even for moderate approximation quality, the number of SVs is considerably high yielding rates worse than the Nyquist sampling (NyquistShannon)rate.

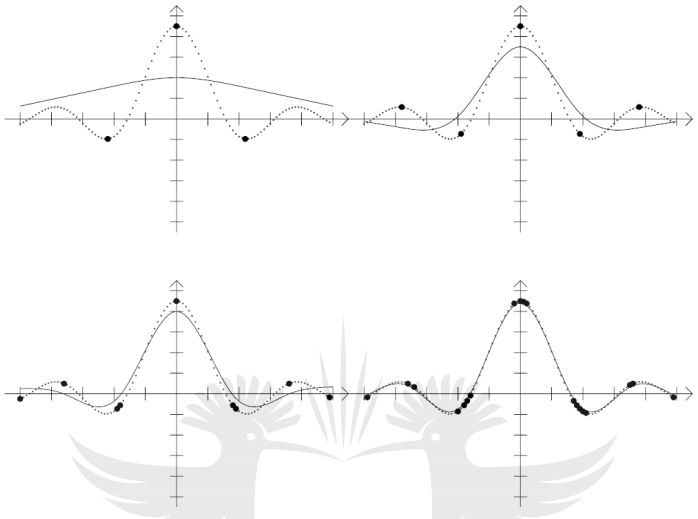

Figure 2.11: Upper left: regression (solid line) datapoints, (small dots) and SVs (big dots) for an approximation with $\varepsilon = 0.5$, upper right $\varepsilon = 0.2$, lower left $\varepsilon = 0.1$, lower right $\varepsilon = 0.1$. Note the increase in the number of SVs.

In Fig 2.12 one can observe the action of Lagrange multipliers acting as forces (α_i, α_i^*) pulling and pushing the regression inside the $\varepsilon - tube$, These forces, however, can only be applied at the samples where the regression touches or even exceeds the predetermined tube. This is a direct illustration of the KKT-conditions, either the regression lies inside the tube (hence the conditions are satisfied with a margin) and consequently the Lagrange multipliers are 0 or the condition is exactly met and forces have to applied $\alpha_i \neq 0$ or $\alpha_i^* \neq 0$ to keep the constraints satisfied. This observation will prove useful when deriving algorithms to solve the optimization problems (Osuna et al, Saunders et al)

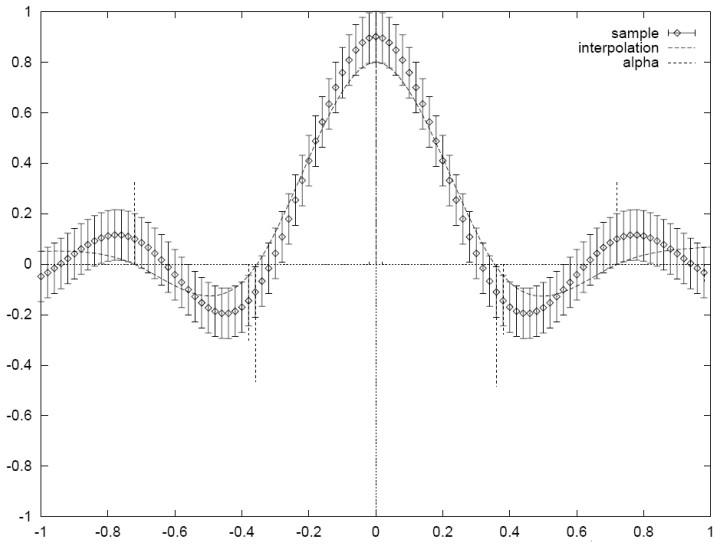

Figure 2.12: Forces (dashdotted line) exerted by the ε-tube (solid interval) lines on the approximation (dotted line)

2.10 Neuro-fuzzy Models

The concepts of fuzzy models and neural network models can be combined in various ways. This section covers the theory of fuzzy models and shows how a combination with neural network concepts gives what is called the neuro-fuzzy model. The most popular neuro-fuzzy model is the Takagi-Sugeno model which is widely used in data driven modeling [128]. The model which is used in this work is described in the subsections that follow.

2.10.1 Fuzzy Systems

Fuzzy logic concepts provide a method of modeling imprecise models of reasoning, such as common sense reasoning, for uncertain and complex processes [129]. Fuzzy set theory resembles human reasoning in its use of approximate information and uncertainty to generate decisions. The ability of fuzzy logic to approximate human reasoning is a motivation for considering fuzzy systems in this work. In fuzzy systems, the evaluation of the output is performed by a computing framework called the *fuzzy inference* system. The fuzzy inference system maps fuzzy or crisp inputs to the output - which is usually a fuzzy set [130]. The fuzzy inference system performs a composition of the inputs using fuzzy set theory, fuzzy *if-then* rules and fuzzy

reasoning to arrive at the output. More specifically, the fuzzy inference involves the fuzzification of the input variables (i.e. partitioning of the input data into fuzzy sets), evaluation of rules, aggregation of the rule outputs and finally the defuzzification (i.e. extraction of a crisp value which best represents a fuzzy set) of the result. There are two popular fuzzy models: the Mamdani model and the takagi-sugeno (TS) model. The TS model is more popular when it comes to data-driven identification and has been proven to be a universal approximator [130].

A fuzzy model is a mathematical model that in some way uses fuzzy sets. The models are premised on rule based system identification. The if-then rules, with imprecise predicates, are the means used to represent relationships between variables such as:

If the percentage increase of an asset price today is $high$ then tomorrow's percentage increase will be $medium$

For the representation to be operational, terms 'high' and 'medium' need to be defined more precisely. The definition of these terms can be represented with fuzzy sets that have membership that gradually changes. Fuzzy sets are defined through their membership functions (denoted by μ) which map the elements of the considered space to the unit interval [0, 1]. The extreme values 0 and 1 denote complete membership and non-membership, respectively, while a degree between 0 and 1 means partial membership in the fuzzy set. Depending on the structure of the if-then rules, two main types of fuzzy models can be distinguished: the Mamdani (or linguistic) model and the Takagi-Sugeno model.

2.10.2 Mamdani Models

The Mamdani fuzzy model is a modeling technique that is typically used in knowledge-based or expert systems. The reason for this is that the model is a linguistic fuzzy model that is very useful in representing qualitative knowledge, illustrated in the following example: Consider a qualitative description of the relationship between the oxygen supply to a gas burner (x) and its

heating power (y).

$$\text{If } O_2 \text{ flow rate is } Low \text{ then power is } Low \tag{2.45}$$

$$\text{If } O_2 \text{ flow rate is } OK \text{ then power is } High \tag{2.46}$$

$$\text{If } O_2 \text{ flow rate is } High \text{ then power is } Low \tag{2.47}$$

The linguistic terms that are used for representation are defined by membership functions such as the ones shown in Fig 2.10.2. The complexity of the definition of the linguistic terms arises because the terms are not universally defined and can mean different things in different context.

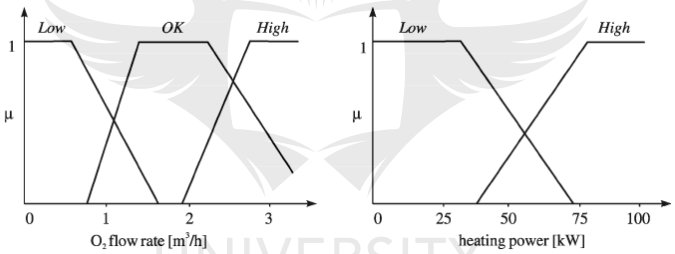

Figure 2.13: Membership functions for the Mamdani model of Example 1.

2.10.3 Takagi-Sugeno Models

The Takagi-Sugeno (TS) model is used in data driven system identification. This model defines the antecedent in the same manner as the mamdani model while the consequent is an affine linear function of the input variables:

$$R_i : \text{If } x \text{ is } A_i \text{ then } y_i = a_i^T x + b_i \tag{2.48}$$

Where a_i is the consequent parameter vector, b_i is a scalar bias value and $i = 1, ..., K$.

What differentiates the Takagi-Sugeno model from the mamdani model is that the former combines the linguistic description with standard functional regression while the latter only uses the linguistic description. The antecedents describe the fuzzy regions in the input space in which the consequent functions are valid. The y output is computed by taking the weighted average of the individual rules' contributions:

$$y = \frac{\sum_{i=1}^{K} \beta_i(x) y_i}{\sum_{i=1}^{K} \beta_i(x)} = \frac{\sum_{i=1}^{K} \beta_i(x)(a_i^T x + b_i)}{\sum_{i=1}^{K} \beta_i} \tag{2.49}$$

Where β_i is the degree of fulfillment of the i rule. The antecedent's fuzzy sets are usually defined to describe distinct, partly overlapping regions in the input space. In the TS model the parameters a are local linear models of the considered nonlinear system. The TS model can thus be considered as a smooth piece-wise linear approximation of a nonlinear function.

Consider a static characteristic of an actuator with a dead-zone and a non-symmetrical response for positive and negative inputs. Such a system can conveniently be represented by a TS model with three rules each covering a subset of the operating domain that can be approximated by a local linear model, see Fig. 2.14. A Takagi-Sugeno fuzzy model as a piece-wise linear approximation of a nonlinear system and the input-output equation are:

$$R_1 : \text{If } \ u \ \text{ is Negative then } \ y_1 = a_1 u + b_1 \tag{2.50}$$

$$R_2 : \text{If } \ u \ \text{ is Zero then } \ y_2 = a_2 u + b_2 \tag{2.51}$$

$$R_3 : \text{If } \ u \ \text{ is Positive then } \ y_3 = a_3 u + b_3 \tag{2.52}$$

and

$$y = \frac{\mu Neg(u) y_1 + \mu Zero(u) y_2 + \mu pos(u) y_3}{\mu Neg(u) + \mu Zero(u) + \mu Pos(u)} \tag{2.53}$$

As the consequent parameters are first-order polynomials in the input variables, model in

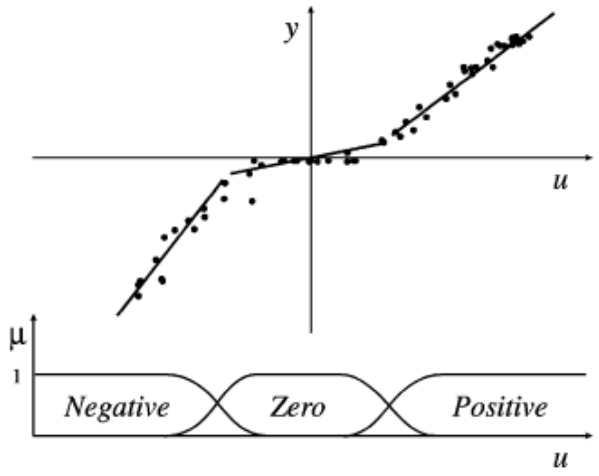

Figure 2.14: A Takagi Sugeno fuzzy model as a piece-wise linear approximation of a nonlinear system.

the literature are also called the first-order TS models. This order is different from the zero-order TS model whose consequents are simply constants (zero-order polynomials):

A TS model of a zero order is a special case of a Mamdani model in which the consequent fuzzy sets degenerate to singletons (real numbers):

$$R_i = \text{If } \ x \ \ A_i \ \text{ then } \ y_i = b, \qquad i = 1, 2, 3....., K \tag{2.54}$$

For this model, the input-output Eq. 2.54

$$y = \frac{\sum_K^{i=1} \beta_i(x) b_i}{\sum_K^{i=1} \beta_i(x)} \tag{2.55}$$

The TS model has been proven to have the ability to approximate any nonlinear function abitrarily well given that the number of rules is not limited. It is for these reasons that it is used in this study. The most common form of the TS model is the first order one. A diagram of a two-input and single output TS fuzzy model is shown in Fig. 2.15:

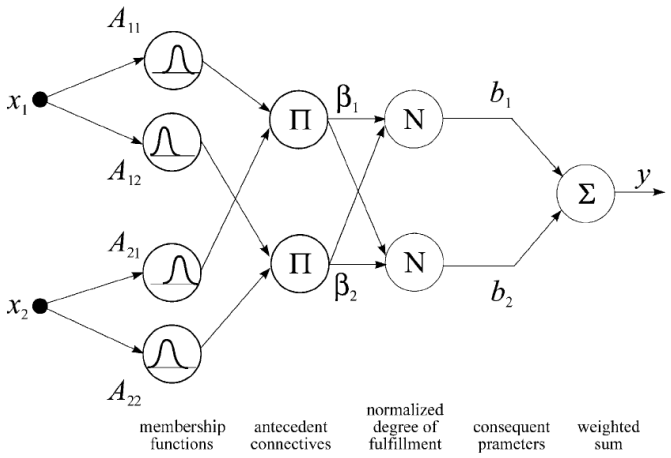

Figure 2.15: A two-input first order Takagi-Sugeno fuzzy model

2.10.4 Fuzzy logic operators

In dealing with systems with more than one inputs, the antecedent proposition is usually defined as a amalgamation of terms with univariate membership functions by using logic operators 'and' (conjuction), 'or' (disjunction), and 'not' (complement). As an example, consider the conjunctive form of the antecedent,which is given by:

$$R_i = \text{If } x_1 \text{ is } A_{i1} \text{ and...} and \ x_p \text{ is } A_{ip} \text{ then } y_i = a_i^T x + b_i \tag{2.56}$$

with a degree of fulfillment

$$\beta_i(x) = \min(\mu A_{i1}(x_i), ..., \mu A_{ip}(x_p)) \tag{2.57}$$

2.10.5 Fuzzy to Neuro-Fuzzy

Fuzzy logic limits the membership functions and the consequent models to a priori knowledge of the model designer. The adaptive neuro-fuzzy inference system (ANFIS) is an architecture which is functionally equivalent to a Sugeno type fuzzy rule base [131]. Under certain minor constraints the ANFIS architecture is also equivalent to a radial basis function network [130]. In a loose kind of way ANFIS is a method for tuning an existing rule base with a learning algorithm based on a collection of training data.What differentiates neuro-fuzzy modeling from

fuzzy logic is that in the absense of the knowledge and the availability of the input-output data observed from the process the components of the fuzzy system membership and consequent models can be represented in a parametric form and the parameters tuned by a learning procedure as shown in its architectecture in Fig. 2.16. In this case the fuzzy system turns into a neuro fuzzy approximator. Neuro-fuzzy systems combine human-like representation and the fast learning methods used in neural networks and therefore has a tradeoff in terms of readability and efficiency. However, what mainly distinguishes neuro-fuzzy estimators from other kinds of non linear approximators is their potentiality for combining available a priori first principle models with data driven modeling techniques.

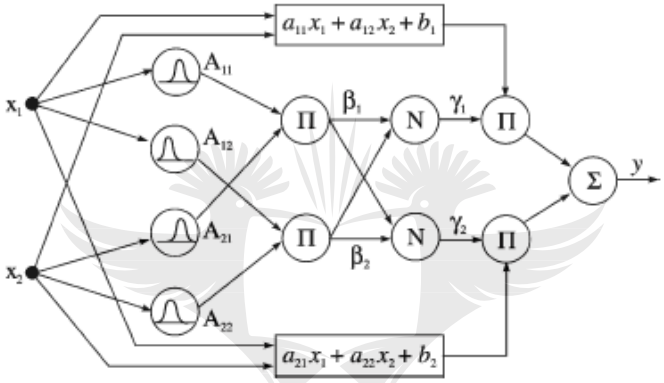

Figure 2.16: An example of a first-order TS fuzzy model with two rules represented as a neuro-fuzzy network called ANFIS.

2.10.6 Neuro-fuzzy Learning Procedure

In a neuro-fuzzy system, there are two types of model tuning which are required, namely structural and parametric tuning.

Structural tuning is a procedure that finds the suitable number of rules and the proper partitioning of the input space. Upon finding the suitable structure, the parametric tuning searches for the optimal membership functions together with the optimal parameters of the consequent models. The problem can be formulated as that of finding the structure complexity which will give the best performance in generalization.The structure selection involves the following two

choices:

1. $Selection\ of\ input\ variables$. In stock market forecasting the inputs are selected from the past data. The past prices give insight in the process behavior and may contain some information about what will happen in the future.

2. $Number\ and\ type\ of\ membership\ functions,\ number\ of\ rules$. The number of membership functions or fuzzy sets are determined by the number of rules defined and which then determine the level of detail of the model, called the granularity of the model. Automated, methods can be used to add or remove membership functions and rules.

In the TS model, the antecedent part of the rule is a fuzzy proposition and the consequent function is an affine linear function of the input variables as shown in Eq. 2.48. Too many rules may lead to an overly complex model with redundant fuzzy rules which compromises the integrity of the model [132]. The antecedents in the model describe the fuzzy regions in the input space in which the consequent functions are valid.

The first step in any inference procedure is the partitioning of the input space in order to form the antecedents of the fuzzy rules. The shapes of the membership functions of the antecedents can be chosen to be Gaussian or triangular. The Gaussian membership function of the form shown in Eq. 2.58 is most commonly used [130].

$$\mu^i(x) = \prod_{j=1}^{n} e^{-\frac{(x_j - c_j^i)^2}{(b_j^i)^2}} \tag{2.58}$$

In Eq. 2.58, μ^i is the combined antecedent value for the ith rule, n is the number of antecedents belonging to the ith rule, c is the center of the Gaussian function and b describes the variance of the Gaussian membership function.

The output y of the entire inference system is computed by taking a weighted average of the individual rules' contributions as shown in Eq. 2.49 [130]. The parameters obtained from the training are then used to approximate models of the non-linear system under consideration [133].

2.10.7 Neuro-fuzzy Modeling

When setting up a fuzzy rule-based system we are required to optimise parameters such as membership functions and consequent parameters. At the computational level, a fuzzy system can be seen as a layered structure (network), similar to artificial neural networks of the RBF-type [130]. In order to optimise these parameters, the fuzzy system relies on training algorithms inherited from artificial neural networks such as gradient descent-based learning. It is for this reason that they are referred to as neuro-fuzzy models. There are two approaches to constructing neuro-fuzzy models [128]:

1. Fuzzy rules may be extracted from expert knowledge and used to create an initial model. The parameters of the model can then be fine tuned using data collected from the operational system being modeled.

2. The number of rules can be determined from collected numerical data using a model selection technique. The parameters of the model are also optimised using the existing data. The Takagi-Sugeno model is most popular when it comes to this identification technique.

2.10.8 Neuro-fuzzy Learning Algorithm

Fixing the premise parameters of the Neuro-fuzzy ANFIS model makes the overall output combination linear . A hybrid algorithm adjusts the consequent parameters of the model in a forward pass and the premise parameters of the model in a backward pass [130]. In the forward pass the network inputs propagates forward until output layer of the network, where the consequent parameters are identified by the least-squares method. In the backward pass, the error signals propagate backwards and the premise parameters are updated by gradient descent.

Because the update rules for the premise and consequent parameters are decoupled in the hybrid learning rule, a computational speedup may be possible by using variants of the gradient method or other optimisation techniques on the premise parameters. Since ANFIS and radial

basis function networks (RBFNs) are functionally equivalent under some minor conditions, a variety of learning methods can be used for both of them.

2.10.9 Clustering of Data

In order set up a Takagi-Sugeno model of a system, it is important to have an automatic method that finds clusters in the data that would form the linear regions. The objective of such a cluster analysis is to partition the data set into a number of natural and homogeneous subsets, where the elements of each subset are as similar to each other as possible, and at the same time as different from those of the other sets as possible. Clusters are such a grouping of objects, that is, it consists of all points close, in some sense, to the cluster centre.

(a) Hard clusters

Hard c-means algorithm attempts to locate clusters of data with a multidimensional feature space . The basic approach of the algorithm is as follows [134]:

1. The algorithm is manually seeded with with c cluster centres, one for each cluster. This turns the algorith into supervised learning because information about the number of different clusters is already known.

2. Each point is assigned to the cluster centre nearest to it.

3. A new cluster centre is computed for each class by taking the mean values of the coordinates of the points assigned to it.

4. If not finished according to some stopping criterion, go to step 2.

(b) Fuzzy Clusters

The fuzzified c-means algorithm [135] allows each data point to belong to a cluster to a degree specified by amembership grade, and thus each point may belong to several clusters. The fuzzy c-means (FCM) algorithm partitions a collection of N data points specified by p-dimensional

vectors $u_k(k = 1, 2, ..., K)$, into f fuzzy clusters, and finds a cluster centre. Fuzzy c-means is different from hard c-means (HFC), mainly because it employs fuzzy partitioning where a point can belong to several clusters with degrees of membership. To accommodate the fuzzy partitioning, the membership matrix M is allowed to have elements in the range $[0, 1]$. The total membership of all clusters, however, must always be equal to unity to maintain the properties of the M matrix.

(c) GK algorithm

The GK algorithm is an extension of the standard fuzzy c-means algorithm by employing an adaptive distance norm in order to detect clusters of different geometric shapes in one data set. The GK algorithm basically contains four steps. Step 1 is computation of cluster prototypes or means. Step 2 then calculates the cluster covariance matrices. Step 3 then calculates the cluster distances. Step 4 then updates the partition matrix. The algorithm then iterates through these steps until the change in membership degrees is less than a given tolerance. For a more detailed explanation of the algorithm refer to appendix A.

The advantages of using the GK algorithm are listed below:

- The resulting fuzzy sets induced by the partition matrix are compact and are therefore easy to interprete.
- In comparison to other clustering algorithms, the GK algorithm is relatively insensitive to the initialisations of the partition matrix.
- The algorithm is based on an adaptive distance measure and is therefore less sensitive to the normalisation of the data.
- The GK algorithm can detect clusters of different shapes, i.e., not only linear subspaces.

Chapter 3

Time series forecasting and Artificial Intelligence

In 2007, the peak demand in South Africa went up to 32000 MW against the generating capacity of 30800 MW. Thus, the electricity supply crisis in South Africa began. Since the beginning of the electricity supply crisis in 2007 in South Africa, the system has been operating at a tight reserve margin. It is, therefore important to conduct medium term forecasts so that maintenance of the generation plants can be planned properly. Understanding the medium term future demand in South Africa will ensure that there are no rolling blackouts which can have a serious negative impact on the economy.

3.1 Time series

A time series is defined as a sequence of observations on a variable measured at successive points in time or over successive periods of time. This means that a set of observations x_t are observed and recorded at a specific time t. A time series model for the observed data x_t can be defined as a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables X_t of which x_t is postulated to be a realization [136].The measurements may be taken every hour, day, week, month, or year, or at any other regular

interval. The pattern of the data is an important factor in understanding how the time series has behaved in the past. If such behavior can be expected to continue in the future, we can use the past pattern to guide us in selecting an appropriate forecasting method.

A time series can either be stationary or non-stationary. A stationary series is a time series whose statistical properties such as the mean, and the variance are independent of time. In a non-stationary series these properties may vary with time. A time series can also exhibit important patterns namely, trend and seasonal patterns. Trend pattern illustrates long-term factors in a series and seasonal illustrates repeating patterns over successive periods of time. In addition to these pattern, an irregular component can be observed in a time series and it arises from random shocks either in the system generating the data or in the data recording instruments. These are some of the characteristics that are uncovered during data analysis.

The underlying pattern and characteristics in the time series are important determining factors in selecting a forecasting method or model. There are various techniques that are used to analyze the data. Time series decomposition can be used to separate or decompose a time series into seasonal, trend, and irregular components. While time series decomposition method can be used for forecasting, it is primarily applied in getting a better understanding of the time series.

Mathematical modeling methods have been developed to model time series. These time series methods are based on the assumption that the data have an internal structure, such as autocorrelation, trend, or seasonal variation as described before. Time series forecasting methods detect and explore such a structure. Time series have been used for decades in such fields as economics, digital signal processing, as well as electric load forecasting.

3.2 Data collection

The data used for the study includes the total electricity consumption was obtained from statistics South Africa (Stats SA). The data was sampled on a monthly basis and stats SA published under P4141 Electricity generated and available for distribution (201406) [137]. The total con-

sumption data used is called the total electricity available for distribution in South Africa which includes losses.

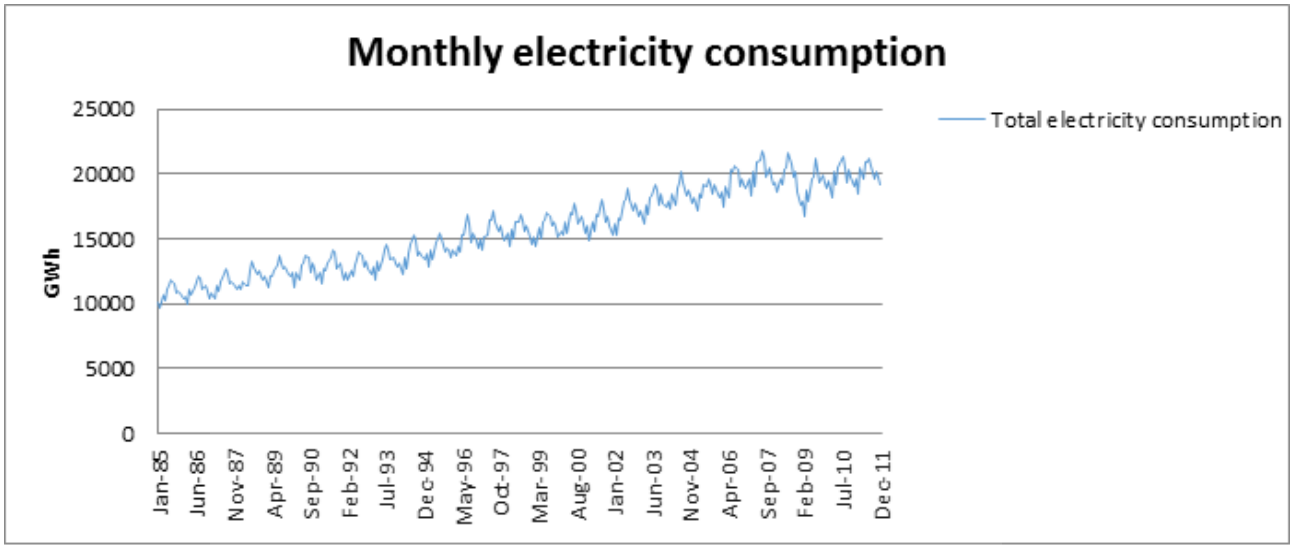

Figure 3.1: Illustration of the monthly total energy consumption

The data used in this experiment was sampled on a monthly basis from January 1985 to December 2011 in South Africa. The demand for electricity has been on the increase over the years in South Africa as illustrated in Figure 3.1. There are trends in the data; these are seasonal trends and monthly trends. Seasonally, the consumption for electricity is the high in winter (May through august) reaching a peak in the month of July and it is the low in summer (November through March) reaching the lowest point in the month of February. This type of seasonality illustrates the relationship between the electricity demand and the weather conditions in different seasons. The peak consumption percentage increase between 1985, which is the lowest peak (11835 GWh) in the data sample, and 2007, which highest peak (21780 GWh) thus far in South Africa, is about 84%. The increase is quite significant and is important to recognize this level of increase which might even be higher going forward.

3.3 Consumption by Sector

The study looks at the consumption of electricity in the mining and the manufacturing sector. The manufacturing sector in South Africa is the largest consumer of electricity and it currently consumes 37% of the total energy to supply. The mining industry is the second largest consumer of electricity. Therefore, in compiling the data for the study the following assumptions were made:

- The percentage of the total annual consumption for each sector was consistent through the months of the year
- Where the variation for the consumption of each sector was more than 2% the average of the two is reasonable to use

The data for this two sectors were obtained from stats SA publications. The trends of the different sectors are illustrated in the graph in Figure 3.2 which covers the consumption from 1957 to 1996. From Figure 3.2, notably consumption in the mining sector has been on the decline as opposed to all the other sectors.

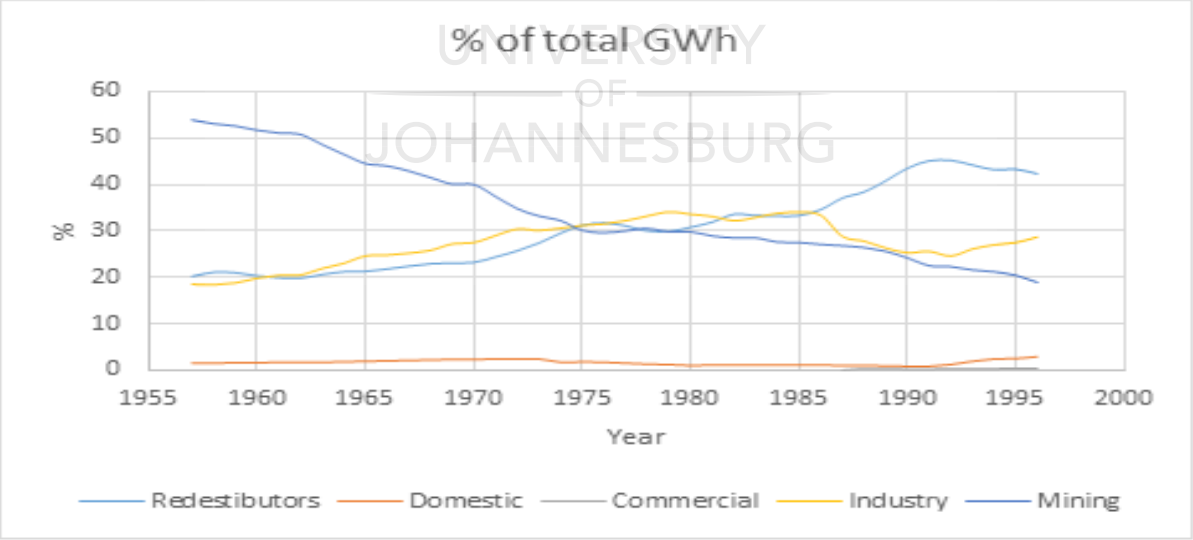

Figure 3.2: The change in electricity consumption in different sector

3.4 Univariate modelling of electricity consumption in South Africa

3.4.1 Data preprocessing

Data preprocessing converts raw data and signals into data representation suitable for training forecasting models through a sequence of operations. The objectives of data preprocessing include size reduction or dimensions of the input space, smoother relationships, data normalization, noise reduction, and feature extraction.

3.4.2 Data normalisation

The data was normalised to lie between 0 and 1 which allows the different algorithm and the activation functions used to comprehend the data more intelligently and make valid deductions from the input series. The data retains its inherent characteristics. The algorithm that is used for normalization is as follows:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

where x_{norm} is the normalised value of x, x_{min} is the minimum value of the series on the data set and x_{max} is the maximum value of the series from the data set.

3.4.3 Feature extraction

As noted by Kaastra et al. and Kolarik popular method to extract features for univariate forecasting is the sliding window approach [138][139]. The data is partitioned into windows of sizes $[n]$ from the time series data. The window is then evaluated from month $[t - (t - n)]$ to month $[t]$. The training example consisted of a sequence of window size $[n] + 1$ demand values. The earlier window size $[n]$ demand values make up the attributes for that example, and the latest demand is the realised target example. To obtain n examples, we have to slide

the window n steps, extracting an example at each 4,5,6,7,8,9,10,11,12,13 and 14 months were used for the extraction of the features that were used to train and test the models.

3.4.4 Accuracy measure

After training the models, out-of-sample data is used test the models on how good they are able to predict on unseen data. The outputs of these tests are compared to the target outputs to measure the accuracy of the prediction. The methodology used in this work to measure the accuracy of the out-of-sample data is the mean square error (MSE), which can represented as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.2}$$

where n is the number of elements, Y_i is the target output and

3.4.5 Experimental techniques and tools

All the simulations for the algorithms are carried out in MATLAB 7 environment running in an Intel CORE i7, CPU. A neural network toolbox developed by Nabney MATLAB was used to conduct the experiments for neural networks [140]. Fuzzy network toolbox developed by Gianluca Bontempi and Mauro Bitattari based on MATLAB was used to conduct the experiments training examples were used for training and one hundred instances were used as out-of-sample data for testing the models [141]. For support vector regressions a toolbox developed by Steven Gunn [142]. The three artificial intelligence techniques are similar in that they are all nonlinear and therefore to give the experiments a perspective on the suitability of these techniques for modelling a linear technique was used to model this system. The linear technique selected was a stochastic technique namely, autoregressive moving average (ARMA).

3.5 ARMA

The concept of AutoRegressive (AR) modelling was first introduced by Yule in 1926 [143]. A further development in this work came about when Slutsky presented Moving Average (MA) schemes in 1937 [144]. However, it was Wold (1938), who thought to combine AR and MA schemes and showed that ARMA processes can be used to model all stationary time series as long as the appropriate order of p, the number of AR terms, and q, the number of MA terms, was appropriately specified [145]. It was Box and Jenkins who popularized the use of ARMA models through the following:

- Providing guidelines for making the series stationary in both its mean and variance
- Suggesting the use of autocorrelations and partial autocorrelation coefficients for determining appropriate values of p and q
- Providing a set of computer programs to help users identify appropriate values for$p$ and q, and
- Once the parameters of the model were estimated, a diagnostic check was proposed to determine whether or not the residuals e_t were white noise, in which case the order of the model was considered final. If the diagnostic check showed random residuals then the model developed was used for forecasting or control purposes assuming of course constancy that is that the order of the model and its non-stationary behavior, if any, would remain the same during the forecasting, or control, phase.

Autoregressive Moving Average (ARMA) model is used to model time series with the purpose of using the model for forecasting. ARMA model is constructed such that the current value of a time series is estimated using prior values of the same time series. The model is a linear combination of the prior values and the coefficients of the linear combination are the parameters which are computed during the modelling process. This is an autoregressive (AR)

model and an AR(1) model represented in an equation as follows:

$$x_t = \phi x_{(t-1)} + e_t \tag{3.3}$$

where e_t $WN(0, \sigma_e^2)$ and similarly AR(p) can be represented as follows:

$$x_t = \phi_1 x_{(t-1)} + \phi_2 x_{(t-2)} \ldots\ldots\ldots \phi_p x_{(t-p)} + e_t. \tag{3.4}$$

The objective of the modeling process is to find a mechanism of converting time series observations to a series of uncorrelated white noise values. The modeling is done such that the observations of a random variable at time t are not only affected by the shock at time t, but also the shocks that have taken place before time t. This is represented as follows:

$$x_t = e_t + \delta e_{(t-1)} \tag{3.5}$$

which is called a moving average (MA), and the above equation is denoted MA(1). And similarly MA(q) is defined as:

$$x_t = e_t + \delta_1 e_{(t-1)} +_2 e_{(t-2)} + \ldots + \delta_q e_{(t-q)} \tag{3.6}$$

By transforming observations into white noise the modeling is broken into two distinct linear filters namely, the autoregressive model and the moving average. The autoregressive (AR) model includes lagged terms on the time series itself, and that the moving average (MA) model includes lagged terms on the noise or residuals [113]. By including both types of lagged terms, we arrive at what are called autoregressive-moving-average, or ARMA, models.

The order of the ARMA model is included in parentheses as ARMA(p,q), where p is the autoregressive order and q the moving-average order. The ARMA model can be represented as follows:

$$X_t = \theta + \sum_{i=1}^{p} \phi_i X_(t-i) + \sum_{i=1}^{p} \delta_i e_{t-i} \tag{3.7}$$

where X_t is the estimated value, θ is a constant, ϕ_i and δ_i are ARMA model parameters for AR and MA respectively, e_{t-i} is white noise sequence. If ϵ_t is a random variable with mean zero and variance σ^2 then for every $t, \tau \geq 0$ with $t \neq \tau, e_t$ and e_τ are uncorrelated. This can be represented formally as:

$$E(\epsilon_t) = 0, E(e_t^2) = \sigma^2, E(e_t e_\tau) = 0 \tag{3.8}$$

Least squares minimization is used to estimate the parameters of the ARMA models. Residuals of the model have to be random, and the estimated parameters have to be statistically significant. Usually the fitting process is guided by the principle of parsimony, by which the best model is the simplest possible model, the model with the fewest parameters that adequately describe the data.

3.6 Experimental setup

In Box-Jenkins modeling process the time series observations are generated by an underlying stochastic process [113]. This underlying stochastic process is assumed to be a stationary process if not then steps are taken to ensure it is stationary. ARMA modelling of time series requires four steps. First the original series x_t must be transformed to become stationary around its mean and its variance. Second, the appropriate order of p and q must be specified. Third, the value of the parameters θ_i and δ_i must be estimated using some non-linear optimization procedure that minimizes the sum of square errors or some other appropriate loss function. Finally, practical ways of modelling seasonal series must be envisioned and the appropriate order of such models specified.

The data was examined for stationarity by plotting autocorrelogram. The data was found to be non-stationary because the plot was decaying slowly as shown in Figure 3.3 and only after performing first order differencing was the data found to be stationary because the plot decayed rapidly and decays to zero after 4 lags as shown in Figure 3.4. The partial autocorrelation function, illustrated in Figure 3.5, was used to determine the AR lag for the model and it was

found to decay to close zero at a lag of 5 which means that the model is given as ARMA(5,4).

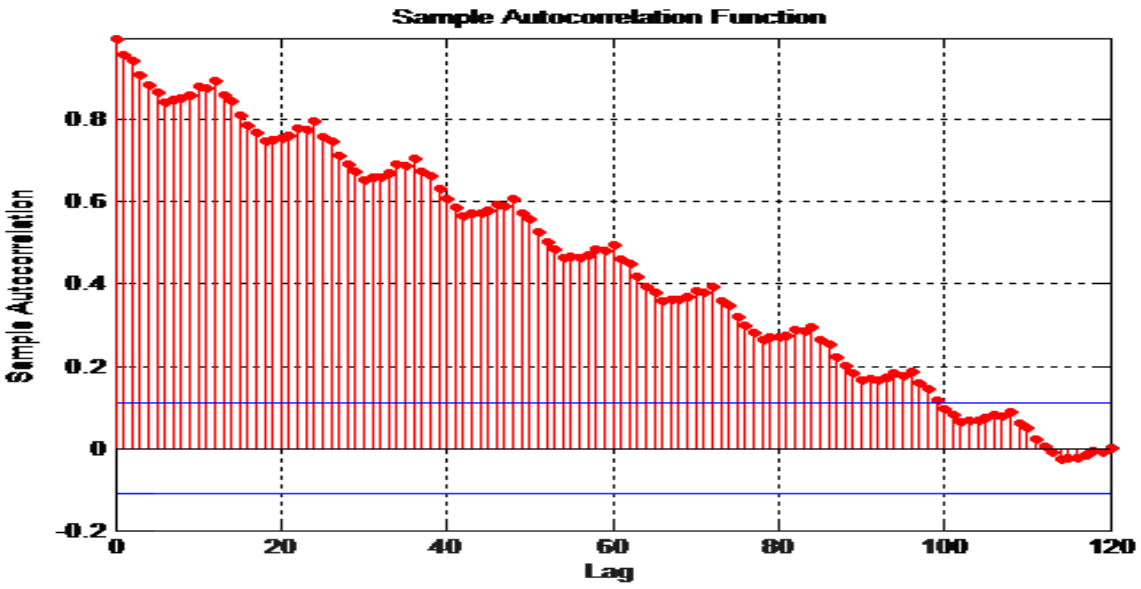

Figure 3.3: Autocorrelation graph before first differencing

3.7 Experimental results

An autocorrelogram plot of the residuals was used to check for model suitability. The autocorrelations of the residuals were found to be insignificant which means the model is suitable. Table 3.1 and 3.6 present the out-of the sample results of ARMA model.

Table 3.1: ARMA accuracy results

No of lags	MSE
5[y(t-5)]	0.011

The autocorrelation function (ACF) computed of the residuals of ARMA model and found that the autocorrelation was very small for all non-zero lags, thus there the error terms are independent and normally distributed.

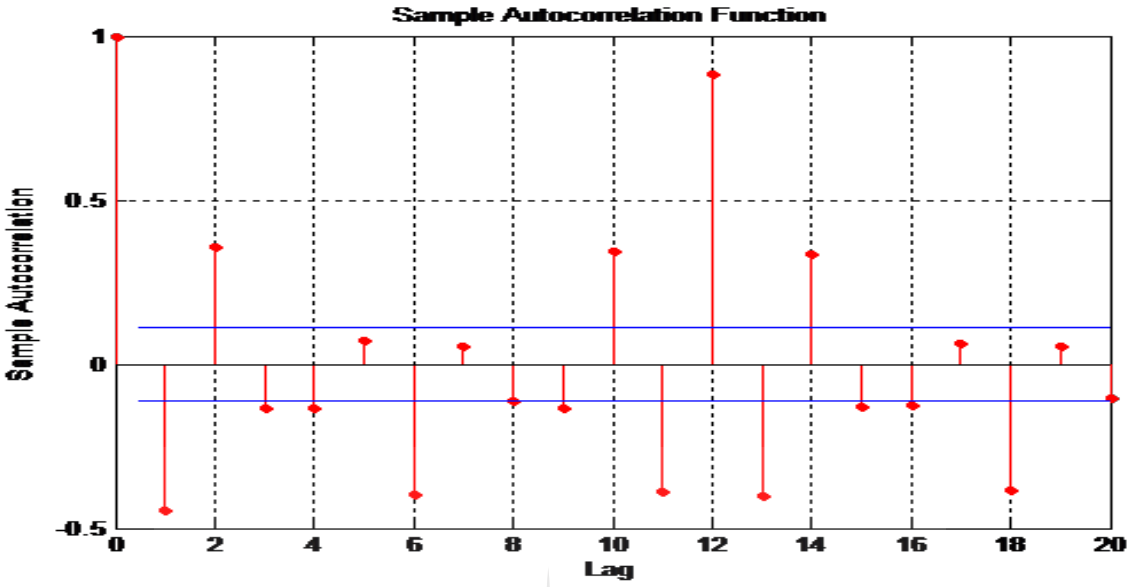

Figure 3.4: Autocorrelation graph after first differencing

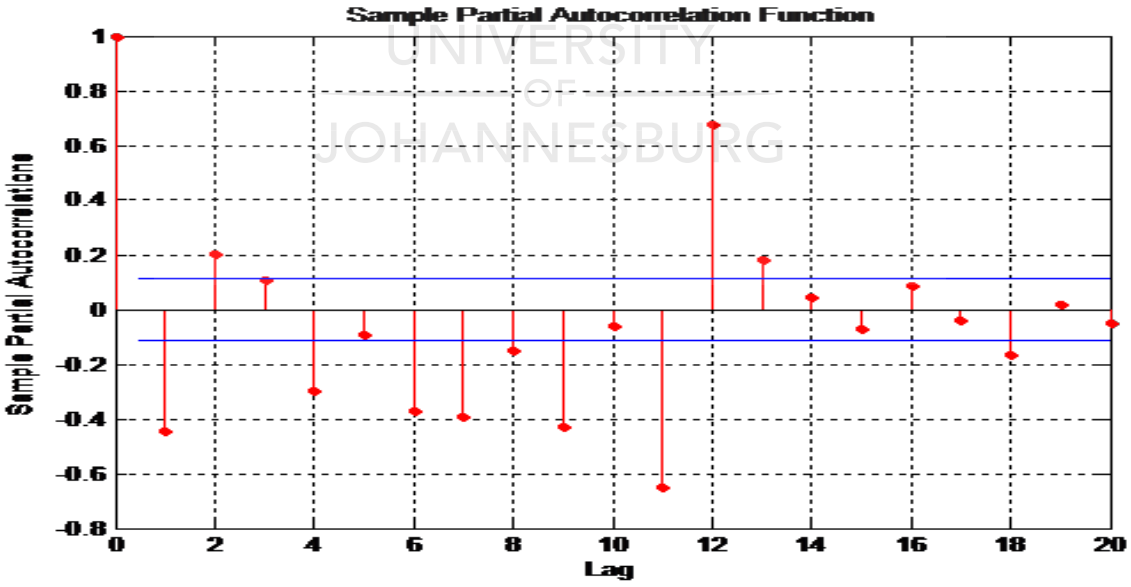

Figure 3.5: Partial autocorrelation plot

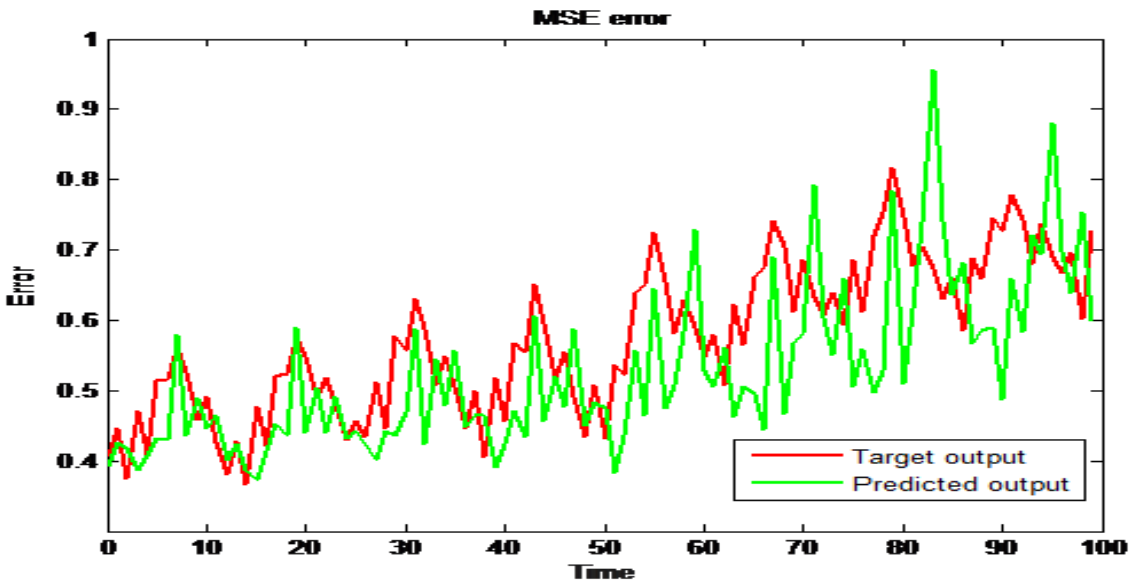

Figure 3.6: Comparison of target output and predicted output for ARMA model)

3.7.1 Neural networks

To train a neural network involves presenting it with different input patterns as contained in the data so that the network can reduce its empirical error and improve its performance. The algorithm that is used to train the network may vary depending on the network architecture, but the most common training algorithm used for neural networks is the backpropagation algorithm. Backpropagation is the process of backpropagating errors through the network from the output layer towards the input layer during training. Backpropagation is necessary because hidden units have no training target value that can be used, so they must be trained based on errors from previous layers.

The output layer is the only layer which has a target value for which to compare. As the errors are backpropagated through the nodes, the connection weights are changed. During the process of backpropagation of errors through the network when different input patterns are presented to the network, the error is gradually reduced to a minimum. Training continues until the errors in the weights are sufficiently small to be accepted. The procedures for developing the neural network BP model are as follows:

- Normalize the learning set;

- Decide the architecture and parameters: i.e., learning rate, momentum, and architecture. There are no criteria in deciding the parameters except on a trial-and-error basis;
- Initialize all weights randomly;
- Training, where the stopping criterion is either the number of iterations reached or when the total sum of squares of error is lower than a pre-determined value;
- Choose the network with the minimum error;
- Forecast future outcome.

MLP results: Neural networks models were trained using back-propagation algorithm (3000 training epochs). The architectures of the neural networks models were as shown in Table 3.2. The Model with 12 inputs for MLP produced the most accurate out-of-sample test results as shown in Table D.1 in Appendix D.

Table 3.2: MLP architectures

MLP architectures used for the experiment
4-6-1
5-6-1
6-7-1
7-8-1
9-10-1
10-11-1
11-12-1
12-12-1
13-13-1
14-14-1

3.7.2 Adaptive neuro-fuzzy inference system

To build and train a neuro-fuzzy model takes two types of tuning, namely structural and parametric tuning. The algorithm is summarised in Figure 3.8 Structural tuning has two objectives, firstly to find a suitable number of rules and secondly, properly partition of the input space into clusters. Once a satisfactory structure is attained, the parametric tuning searches for the optimal

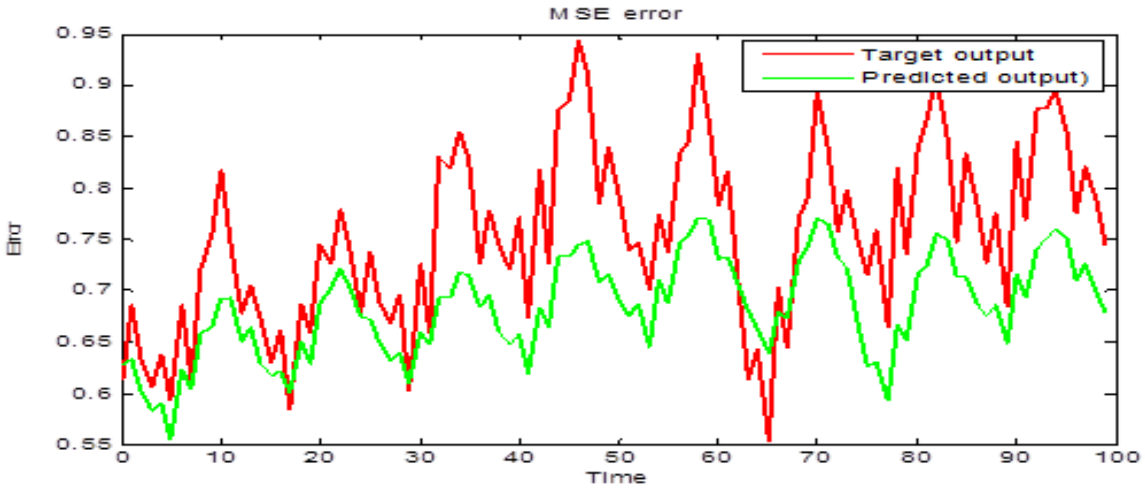

Figure 3.7: Comparison of target output and predicted output for the MLP model with the best accuracy (12 inputs))

membership functions together with the optimal parameters for consequent models. In some cases there may be multiple structure and parameter combinations which make the fuzzy model perform in a satisfactory way. The problem can be formulated as that of finding the structural complexity that is able to give the best generalisation performance. The approach taken in this process chooses the number of rules as the measure of complexity to be properly tuned on the basis of available data. An incremental approach where different architectures having different complexity (i.e number of rules) are first assessed in cross validation and then compared in order to select the best one. To train the neuro-fuzzy systems the Gaussian membership function and hard cluster means (HCM) were used for clustering and structure determination. To train the neuro-fuzzy models the error rates were propagated backwards and the parameters were updated by gradient descent method. The model with 12 inputs for neuro-fuzzy had the most accurate out-of-sample test results as shown in Table D.2 in Appendix D and Figure 3.9.

3.7.3 Support vector regressions

SVR results: There are three methods for controlling the regression model, the loss function, the kernel, and additional capacity control. The experiments for the regression model used for

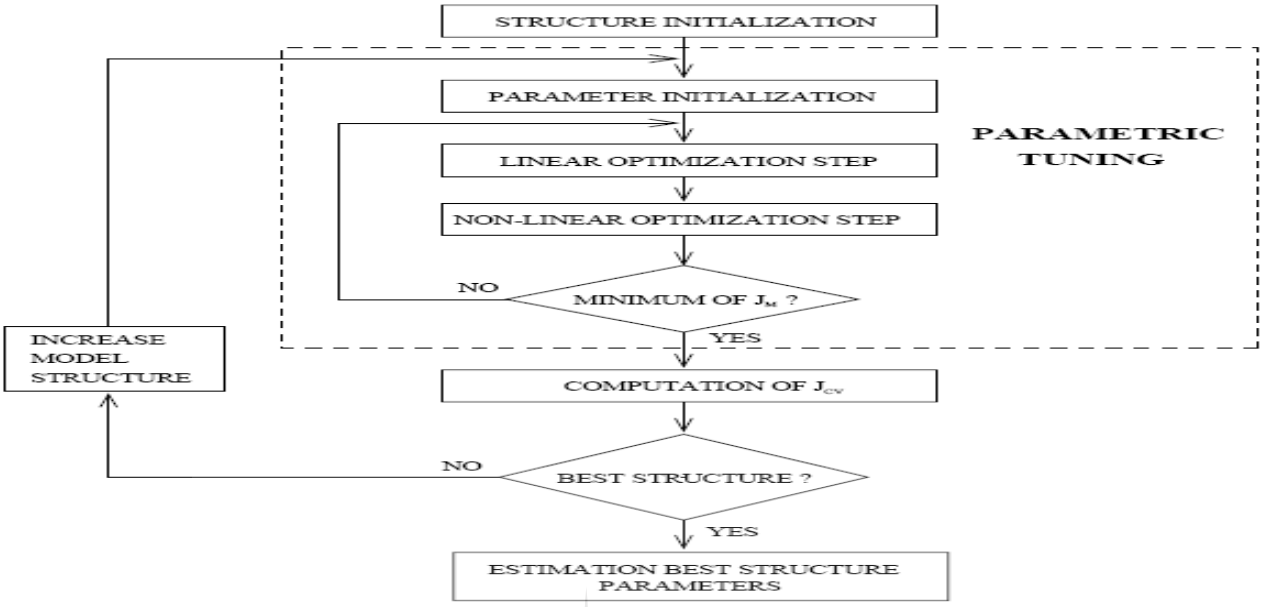

Figure 3.8: ANFIS learning algorithm

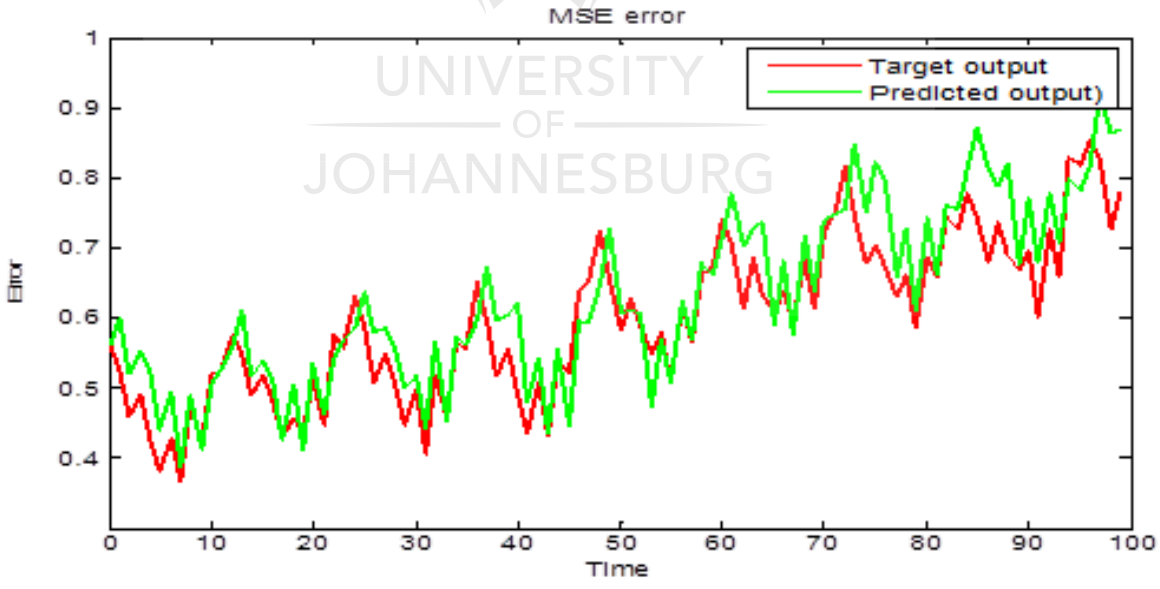

Figure 3.9: Comparison of target output and predicted output for the ANFIS model with the best accuracy (12 inputs)

the prediction of the electricity consumption were performed with different ϵ-insensitive loss functions, with different kernels and different degrees of capacity control. Several combinations were tried, and finally a radial basis function was chosen as the kernel with $\epsilon = 0.01$ and C = 100. The model is trained with the training data set with 150 instances, and then tested with the test data set with another 100 instances. The model with 13 and 14 inputs have the most accurate out-of-sample test results for svr as shown in Table D.3 in Appendix D and Figure 3.10.

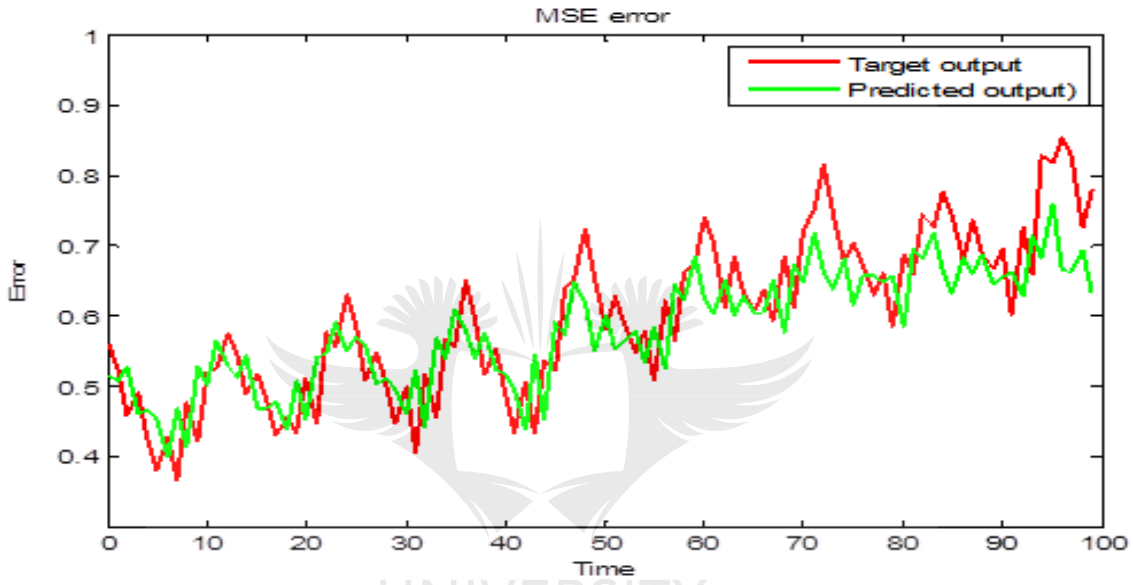

Figure 3.10: Comparison of target output and predicted output for the SVR model with the best accuracy (13 inputs)

3.8 Conclusion and Discussion of results

For neuro-fuzzy and neural networks models, the model with 12 inputs produces the most accurate results. For support vector regression the model with 13 inputs produces the most accurate results as measured by the mean square error. This seems to indicate that 12 and more inputs are a better representation of the dynamics of the consumption in a year which makes monthly modeling much more efficient. The results in TableD.1, D.2 and D.3 also show that the magnitude of the errors decline as the number of inputs increase. For MLP and ANFIS

the errors decline until the input size reaches 12 and 13 inputs for SVR and beyond that the errors begin to increase. It is further shown in the three tables that the support vector regression model outperforms both the neural network model and neuro-fuzzy model. The results also show that a ranking of methods that puts across SVR at the top, followed by ANFIS and then MLP which performed better than ARMA on the considered load series. SVR gives lower mean square error than both MLP, ANFIS and ARMA. The statistical significance tests indicate that SVR results are significantly better than all the other methods. ANFIS results were found not be statistically significantly different from MLP results. ANFIS and MLP significance results indicated that they were both better than that of ARMA.

Lastly, the results show that the error for each method, on the test data set, gets larger as the time series time moves further away from the training data set. This could be an indication that as the time moves the dynamics of the data modeled during training are changing over time which renders the model less accurate in its predictions.

Chapter 4

Causality Approach to Electricity forecasting

This chapter outlines the application of explanatory forecasting to electricity consumption. Explanatory forecasting assumes a cause and effect relationship between the inputs and output. According to explanatory forecasting, changing the inputs will affect the output of the system in a predictable way, assuming the cause and effect relationship is constant. Correlation in explanatory forecasting is a necessary but not a sufficient condition. Causality concepts that will be covered under this section include Granger causality and structural causal models.

4.1 Cauasality

According to David Hume ”We have no other notion of cause and effect, but that of certain objects, which have been always conjoined together, and which in all past instances have been found inseparable. We cannot penetrate into the reason of the conjunction. We only observe the thing itself, and always find that from the constant conjunction the objects acquire a union in the imagination [146].” This was one of the many attempts to understand causal relationships between events that occurs in reality. However, the problem with this definition is that it relied on observation to detect causation is that it could not account for spurious correlation (a

correlation that does not imply causality e.g. the crow of the rooster and sunrise have no causal relationship). Tshilidzi Marwala defines causality as a flow of information from the cause to the effect [147]. The problem with Marwala's definition is that it relies on the ability to measure the information flow which is not possible.

Causality is used to define a relationship between two events such that one event called the cause is believed to have caused another event called the effect. In a causal relationship the cause is necessary for the occurrence of effect. The likelihood of the cause is non zero, and the likelihood of occurrence of the effect given that the cause has occurred is bigger than the likelihood of the effect occurring alone

Sellitz et al. further outlined three conditions for the existence of causality [148]:

1. There must be a concomitant co-variation between the cause and the effect.
2. There should be a temporal asymmetry or time ordering between the two observed sequences (the cause should happen before the effect).
3. The covariance between the cause and the effect should not disappear when the effects of any confounding variables are removed.

Covariation implies that there should an association between the cause and the effect statistically called correlation. However, correlation does not imply causality.

Before quantum mechanics causality had largely dealt with deterministic variables or events. However, there was a realization that certain variables cannot be described using deterministic mathematical representation and therefore, probability was used for representing these types of events. This led to a formulation of an area of study called probabilistic causality. Causation should be differentiated from correlation. It is common to obtain between quantities varying with the time (such as time series) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be significant

1. The cause occurs before the effect; and

2. The cause contains information about the effect that that is unique, and is in no other variable.

From these two statements Granger concludes that causal variable can help forecast the effect variable. This is expressed in a linear regression model:

$$Y_t = a_0 + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \sum_{k=1}^{L} b_{2k} Y_{t-k} + \epsilon_t \tag{4.1}$$

where ϵ_t is an uncorrelated random variable with zero mean and variance, L is the specified number of time lags, and $L = L + 1, \ldots, N$. The null hypothesis that X_t does not Granger cause Y_t is supported when $b_{2k} = 0$ for $k = 1, \ldots, N$ which reduces equation 4.1 to:

$$Y_t = a_0 + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \epsilon_t \tag{4.2}$$

The granger causality modeling has been applied in many areas including but not limited economics, finance, and relevant to this study in electricity consumption forecasting (see [97][98][149] [150] [41] [40] The limitation of the Granger causality method is that it focuses on providing accurate modeling of the systems as opposed to providing insight on how to events are linked. And as can be seen from the literature survey in outlined Chapter 2 Granger causality is used widely and its results are mixed in terms of causal direction. Practitioners need to be careful not misinterpret the results when using this method.

4.2 Structural Causal Model (SCM) and time series analysis

Developed by Judea Pearl, structural causal models (SCM) is an attempt to answer the question: Do we rely on powerful computing and statistical approaches to tease apart signal from noise, and find the causal relation or do we look for the more basic principles that underlie the system and explain its essence? Furthermore, SCM provides a graphical criterion for choosing the "right hand side" variables to include in a forecasting model [151]. In this work, SCMs are used identify variables that can be used to model electricity consumption in the different economic

sectors in South Africa.

The origins of SCM are very closely linked to the Neyman-Rubin model.

4.2.1 Neyman-Rubin model

The Neyman-Rubin model was preceded by Neyman's non-parametric model where each unit has two potential outcomes, one if the unit is treated and the other if untreated [67]. In this model a causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed at an instance[152]. The model was further developedl into a general framework for causal inference with implications for observational research by Rubin [68] [153], among others and including most notably Cochran [154][155].

As already mentioned the Neyman-Rubin model is defined in terms of potential outcomes which can be denoted as follows: $Y(x, u)$ the potential outcome in unit u if X is set equal to x [156][157][158]. The potential outcomes are used to define the unit-specific causal effects. This can be illustrated by assuming that X can only take on the values zero and one. The unit-specific causal effect of $X = 1$ on Y relative to the effect of $X = 0$ in unit u is calculated by comparing $Y(1, u)$ to $Y(0, u)$. The difference between the two effects is used as a comparison:

$$Y(1, u) - Y(0, u) \tag{4.3}$$

If both could be observed, the objective of the comparison would be to observe the unit-specific cause effect of causal variable at different levels.

By assuming consistency of the observed outcomes, it may be possible to observe one of these two outcomes for each individual. Operationally, the assumption makes it a necessary condition that the observed outcome for each unit $Y(u)$ matches the potential outcome for unit u for the observed value of X. This is defines as:

$$X(u) = x = Y(u) = Y(x, u) \tag{4.4}$$

which if it is the case, then the observed Y can be written as:

$$Y^{obs}(u) = X(u).Y(1,u) + (1 - X(u).Y(0.u) \tag{4.5}$$

Unit u only gets one of either $X = 0$ or $X = 1$ but not both and as a result makes it impossible for the unit specific causal effect to be observed. This is called the fundamental problem of causal inference.

Because of this limitation, the causal inference is usually confined to characteristics of populations as opposed to specific individual units. Causal models have been developed with the shift from unit specific effect to population causal effects. This means that population causal model is created assuming a distribution over U.

4.2.2 Structural Causal Model

Structural Causal Model (SCM) is a structural theory developed in [63][64] which combines features of the structural equation models (SEM) used in economics and social science [65][66], the potential-outcome framework of Neyman [67] and Rubin [68], and the graphical models developed for probabilistic reasoning and causal analysis [63][69][70]. SEM was developed for linear analysis and therefore, could not be extended to nonlinear problems.

According to Judea pearl [63] the SCM as general theory of causation is able to:

1. Represent causal questions in some mathematical language,

2. Provide a precise language for communicating assumptions under which the questions need to be answered,

3. Provide a systematic way of answering at least some of these questions and labeling others ”unanswerable,” and

4. Provide a method of determining what assumptions or new measurements would be needed to answer the ”unanswerable” questions.

5. Subsume any other theory or methods that scientists have found useful in exploring the various aspects of causation.

The benefits of using SCM includes:

1. Proving that a particular conditioning set identifies a casual effect of interest
2. Helps researcher to make the assumptions about causal dependencies explicit

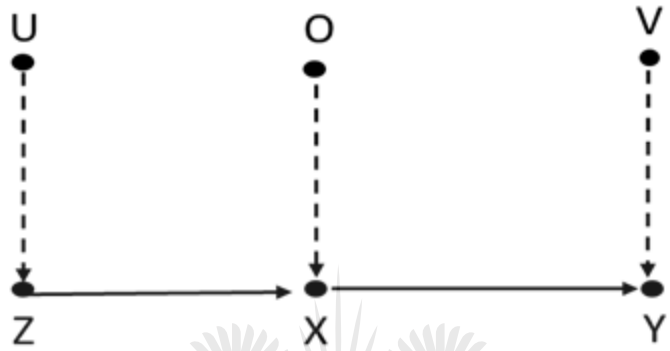

Figure 4.1: Illustration of the structural model

.

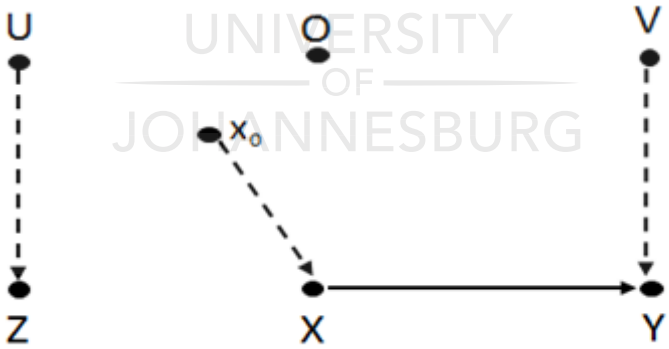

Figure 4.2: Illustration of the modified structural model representing the intervetion $do(X = x_0$

.

The SCM can be illustrated through a structural model M, consisting of two sets of functions that determine, or simulate how values are assigned to each variable

$$z = f_z(u_z) \tag{4.6}$$

$$x = f_X(z, u_X) \tag{4.7}$$

$$y = f_Y(x, u_Y) \tag{4.8}$$

where U_z, U_X and U_Y are assumed to be jointly independent but, arbitrarily distributed. Each of the functions represents a causal process that determines the desired output on the left hand side from the inputs in the right hand side. The absence of the variable Z from the arguments of the function f_Y means that variation in Z leaves Y unchanged if U_Y and x remain constant. These equations can be defined as structural if they are assumed to be autonomous, that is, each function is invariant to possible changes in the form of the other functions [64][159]. The invariance provides a framework for using structural equations as a basis for modeling causal effects and counterfactuals. A mathematical operator $do(x)$ is used to simulate a physical intervention by deleting certain functions from the model, replacing them by a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in model M of Figure 4.1, we replace the equation for x with $x = x_0$, and obtain a new model, $M_(x_0)$,

$$y = f_Y(x_0, u_Y) \tag{4.9}$$

$$x = x_0 \tag{4.10}$$

$$z = f_z(u_z) \tag{4.11}$$

the graphical description of which is shown in Figure 4.2

$P(z, y|do(x_0))$ is the joint distribution associated with the modified model, and it describes the post-intervention distribution of variables Y and Z (also called "controlled" or "experimental" distribution), to be distinguished from the pre-intervention distribution, $P(x, y, z)$, as-

sociated with the original model shown in Figure 4.2. In general, we can formally define the post-intervention distribution by the equation:

$$P_M(y|do(x)) \triangleq P_(M_x)(y) \tag{4.12}$$

In the framework of model M, the post-intervention distribution of outcome Y is defined as the probability that model M_x assigns to each outcome level $Y = y$. However, using directed graphical analysis (DGA) circumvents the derivation of the structural analysis algebraically because the graphs used in DGA encode all the information that non-parametric structural equations need.

Identification

The central question in the analysis of causal effects is the question of identification: Can the controlled (post-intervention) distribution, $P(Y = y|do(x))$, be estimated from data governed by the pre-intervention distribution, $P(z, x, y)$? Unlike in linear parametric with a model parameter with a unique solution, in nonparametric formulation the identification is more involved since the notion of unique solution does not directly apply.

The following definition overcomes these difficulties:

Definition 2 (Identifiability [159]). A quantity $Q(M)$ is identifiable, given a set of assumptions A, if for any two models M_1 and M_2 that satisfy A, we have

$$P(M_1) = P(M_2)Q(M_1) = Q(M_2) \tag{4.13}$$

The details of $M1$ and $M2$ do not matter; what matters is that the assumptions in A (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of $P's$ would entail equality of $Q's$. When this happens, Q depends on P only, and should therefore be expressible in terms of the parameters of P. The next subsections exemplify and operationalize this notion. The condition for identifying the causal is captured in Theorem 1 in Appendix B.

Estimating the effect of the interventions

Estimation deals with the process of estimating hypothetical entities such as $P(y|do(x))$. Through derivation, Judea pearl in [159] , show that $P(y|do(x))$ can be derived from $P(y|x)$.

4.2.3 Directed Acyclic Graph

Directed acyclic graph (DAG) models are popular tools developed for describing causal relationships and for guiding attempts to learn them from data. They appear to supply a means of extracting causal conclusions from probabilistic conditional independence properties inferred from purely observational data. Simply put, causal analysis in the graphical models starts off with the realization that all causal effects are identifiable whenever the model is markovian, which means that the graphs are acyclic and all the error terms are jointly independent. The causal Markov Condition is explained in Appendix B in Theorem 2 and corollary 1. If the markovian condition is met the ability of the conditioning to determine the causal effect can be determined from the DAG using the backdoor criterion.

The following criterion, named "back-door" in [70], provides a graphical method of selecting such a set of factors for adjustment. It states that a set S of covariates is appropriate for adjustment or for identifying the effect of the causal variable on the effect if two conditions hold:

1. No element of S is a descendant of X

2. The elements of S "block" all "back-door" paths from X to Y , namely all paths that end with an arrow pointing to X also called d-separation.

Definition 1 (d-separation . A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks all paths from X to Y , it is said to d-separate X and Y, and then, X and Y are independent given S [70].

A directed path is a path of edges will all arrows pointing on the same direction along the path. A path is blocked by a conditioning set Z when either:

1. It contains a chain structure $a \rightarrow b \rightarrow c$ or a fork structure a$\leftarrow b \rightarrow c$ or

2. It contains a collider structure $a \rightarrow b \leftarrow c$,

where b is not in the set Z nor is any descendent of b in Z (a descendent of b would be a variable on a directed path out of b) The thinking behind the back-door criterion can be summarized as follows: The back-door paths in the diagram imply spurious associations from X to , while the paths directed along the arrows from X to Y imply causative associations. Blocking the former paths (by conditioning on S) ensures that the measured association between X and Y is purely causative, namely, it correctly represents the target quantity: the causal effect of X on . The reason for excluding descendants of X (e.g., $W3$ or any of its descendants) is given in [63]. The joint distribution for the DAG in figure 4.1 can be written as follows:

$$P(z, x, y) = P(z)P(x|z)P(y|x) \tag{4.14}$$

Where each marginal or conditional probability on the right hand side is directly estimate from the data. Given an intervention where variable X is set to x_0, the post-intervention distribution can be written as:

$$P(z, y|do(x_0)) = P(z)P(y|x_0) \tag{4.15}$$

where $P(z)$ and $P(y|x_0)$ are identical to those associated with the pre-intervention distribution of equation 4.18. Figure 4.2 represents the modified model. The causal effect of X on Y can be obtained immediately by marginalizing over the Z because the distribution of Z is not affected by the intervention, since

$$P(y|do(x_0)) = \sum_y P(z)P(y|x_0) = P(z) \tag{4.16}$$

while that of Y is sensitive to x_0 and is given by

$$P(y|do(x_0)) = \sum_z P(z, y|do(x_0)) = \sum_z P(z)P(y|x_0) = P(y|x_0) \tag{4.17}$$

This example demonstrates how the (causal) assumptions embedded in the in the model M that allows for the calculation of the post-intervention distribution from pre-intervention distribution and also allows for the estimation of the causal effect of X on Y.

4.3 Framework for SCM time series analysis

This model was adopted from the paper written by Glynn and Quinn called "Structural Causal models and the specification of Time-Series-Cross-Section model" [151]. The framework of the problem can be described by regression equations:

$$y_{i,t} = \alpha y_{t-1} + \beta x_t + u_t \tag{4.18}$$

$$x_t = \rho x_t + e_t \tag{4.19}$$

$$u_t = \phi u_t + e_t;\ t = 1, \ldots, T, \tag{4.20}$$

where y_t is the dependent variable , x_t is the causal variable, u_t represents unobserved errors, and all variables have been standardized in order to avoid the need for intercepts. To construct a regression model y_t is regressed on x_t and perhaps lagged y and x variables to determine the causal parameter β and other parameters of the model.

The lagged right-hand-side variables are referred to as the conditioning set because they are used to determine β and their inclusion in the model is based on whether they improve the estimation of β . There three conditioning sets that can be used to estimate β:

- Empty set: Lagged dependent variable only or auto-regression (y_{t-1})

- A lagged explanatory variable (x_{t-1})
- A lagged dependent and explanatory variable and lagged dependent variable (y_{t-1}, x_{t-1})

4.3.1 The model without a Causal Effect for the lagged Dependent variable

Using the DGA and the backdoor condition causal effect or the lack thereof, of the causal effect of the lagged dependent variable can be determined. From Figure 4.3 [151] it can be observed that there is no conditioning variable on a directed path from X, which means the backdoor criterion is satisfied. There is only one other path from X to Y which is $x_t \leftarrow x_{t-1} \rightarrow y_{t-1} \leftarrow u_{t-1} \rightarrow u_t \rightarrow y_t$ and it is blocked by $\leftarrow y_{t-1} \rightarrow$ collider structure. It is clear from this that regressing y_t on x_t will identify the causal effect β. Similarly, the conditioning sets x_{t-1}, y_{t-1} and x_{t-1} will also identify β because these block the backdoor at the $\leftarrow x_{t-1} \rightarrow$ fork structure.

SCM analysis ensures that by simply knowing that y_{t-1} is not the cause y_t does not allow one to simply conclude that all the regression specifications that identify the causal effect of x_t on y_t from the right hand side.

4.3.2 The model with serial Correlation and causal dependent lagged variables

DGA analysis is also useful in analyzing causal modeling wherein the dependent variable has a causal effect on the effect or outcome. Figure 4.4 illustrates this model [151]. Using y_{t-1} leaves open a backdoor from x_t to y_t, through the path $x_t \leftarrow x_{t-1} \rightarrow y_{t-1} \leftarrow u_{t-1} \rightarrow u_t \rightarrow y_t$ because conditioning on a collider at y_{t-1} opens up a closed path. As a result this conditioning set cannot be used for the estimation of the causal of x_t on y_t. The conditioning sets x_{t-1} and x_{t-1}, y_{t-1} block all backdoor paths from x_t to y_t. These two conditioning sets can be used to identify the causal effect.

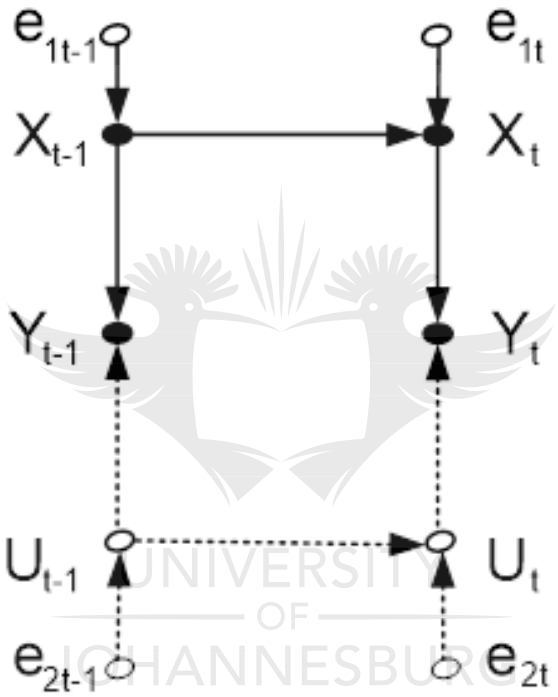

Figure 4.3: Illustration of model without dependent variables

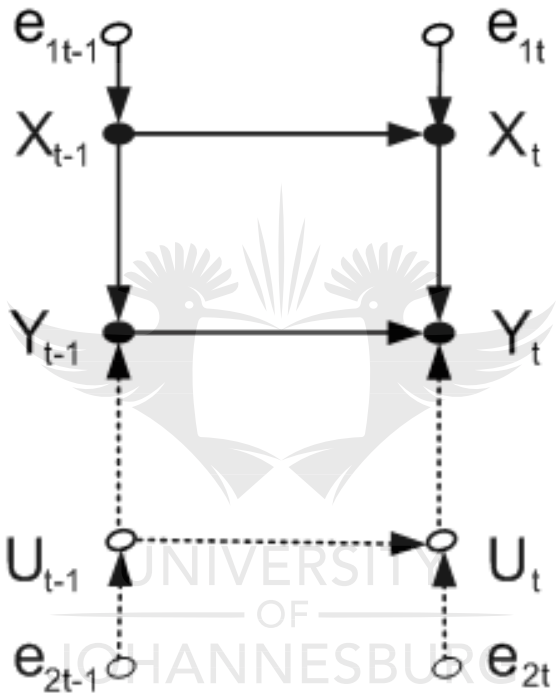

Figure 4.4: Illustration of model with dependent variables

4.4 SCM and Load forecasting

In this DAG will be used to analyse electricity demand for the identification of the causal variables which are subsequently used for estimation of the causal effect. The structural approach follows four very important steps that should be part of every exercise in causal inference:

1. Define: Express the target quantity$Q$ or causal effect as a function $Q(M)$ that can be computed from any model M.

2. Assume: Formulate causal assumptions using ordinary scientific language and represent their structural part in graphical form.

3. Identify: Determine if the target quantity is identifiable.

4. Estimate: Estimate the target quantity if it is identifiable, or approximate it, if it is not.

The South African energy demand is determined by consumption in the different sectors of the economy namely, mining sector $(MinC)$, Manufacturing/industrial sector $(ManC)$, Agricultural sector $(AgriC)$, transport sector $(TransC)$ and the domestic sector $(DomC)$. The change in the electricity load in this sectors causes a change in the total electricity demand or consumption. The relationship can be represented as follows:

$$Total\ \ elec\ \ consumption = f(ManC, MinC, DomC, AgriC, TransC) \tag{4.21}$$

It can be assumed that there is causal relation between the load variation in the various sectors and the total demand. The trick however is to find the quantifiable parameters in the various sectors that are sensitive to the causal assumptions. In this case the manufacturing index, mining index and agricultural production index as quantifiable parameters representative of the dynamics of electrical load variation in these sectors.

The data collected is a time series. These data is used to extract causal relations between the electricity consumption as a dependent variable and sectoral data. The SCM is used to create a

theoretical analysis and understanding of the problem regarding of the causal relations between these variables. What will follow from the analysis is the estimation of the causal effect using time series regression models.

The CSIR developed models for electricity consumption forecasting in South Africa in 2003/4 which was revised in 2010 with updated data [160]. Information on both total consumption and consumption per sector was used to create the models. The sectors considered for the study included mining; domestic; manufacturing and commerce; transport and Agriculture. Multiple regression modelling was used for forecasting the annual consumption within the individual electricity sectors by relating the demographic and economic conditions to the demand in each sector.

Drivers for each sector were identified as follows:

1. Mining: platinum production index, coal production index, and gold ore treated
2. Commerce and Manufaturing: Population, and manufacturing index
3. Transport: Mining Index
4. Domestic: Population and Final Consumption Expenditure by Households (FCEH) also called private consumption expenditure (PCE)
5. Agriculture: Final Consumption Expenditure by Households (FCEH)

This study has restricted itself to two major sectors of the economy which manufacturing or industry and mining which collectively make up the majority of the total consumption of electricity. The main reason for the restriction is the limited data available to make the study much more rigorous in the sectors that have excluded. The data is important not only as a measure of the dynamics of the sector but also because the current study uses artificial intelligence techniques which are mainly data driven for modeling.

4.4.1 Manufacturing Industry and Electricity demand

Bell and Madula assert that manufacturing performed poorly in the 1980's, a continuation of the deterioration that started in the 1970's [161]. There was a rapid growth in manufacturing between 1992 and 1996 which according to bell and madula was due to the recovery of the OECD economies. South Africa at this point had a well-established manufacturing export capacity. The industry was picking up from a rapid decline between 1989 and 1992. Manufacturing sector is currently largest contributor to the South African economy and it contributes about 15.1% of the South African GDP [137]. Manufacturing refers to industries belonging to International Standard Industrial Classification (ISIC) of All Economic Activities, Rev.3.

With regards to electricity demand, there was a sharp increase in electricity sales to the manufacturing sector from 1980 to 2008. The increase was largely driven by economic growth and increased consumption by the non-ferrous metals and iron and steel sub-sectors. Figure 4.5 shows the consumption of electricity in the manufacturing sector from 1985 to 2011. As illustrated in the figure the data reflects the seasonal fluctuations and the response of the electricity demand to the economic performance. Factors determining consumption of electricity in the Manufacturing sector include:

1. Income/Gross Domestic product
2. Export
3. Production levels
4. Foreign direct investment

As the economy grows the demand for manufactured goods also grows which means the manufacturing industry requires more inputs including electricity to produce more goods. Similarly when there is demand for more or exports, depending on the type of products produced or the exchange rate, etc., to other markets the local production levels also increase. Investment in the industry, especially foreign direct investment, leads to an increase in the use of capital versus labour which leads to increase in the use of electricity [162]. Price has been excluded because

studies have shown that electricity prices in South Africa have been so low that prices increases have little or no effect on the electricity demand in the manufacturing sector in the period that is being considered in this study [163][164][165].

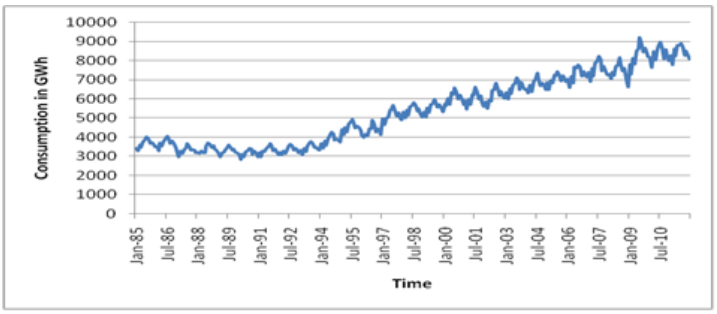

Figure 4.5: Electricity consumption in the manufacturing sector

.

Manufacturing production index

In South Africa, manufacturing production index measures the total output of industrial/manufacturing sector of the economy. Its fluctuation reflects the performance of the manufacturing sector on month to month basis. Figure 4.6 illustrates how the index has changed from January 1985 to December 2011. The general trend of the index went on an increase from the early 1990s up until 2008 when the global economic crisis started. From 2008 the production index went on a steep decline and its recovery started in mid-2009.

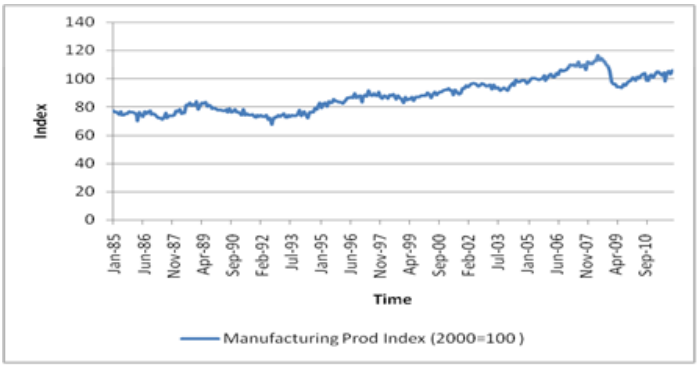

Figure 4.6: The manufacturing production index

.

SCM and manufacturing industry

Figure 4.7 illustrates the foreign direct investment has a direct impact on the increase or decrease of exports, gross domestic product and manufacturing production index. GDP and PI in turn have a causal relation with the consumption of electricity in the manufacturing sector. The selection of these variables is not exhaustive in both the observed and the unobserved confounding variables but these are regarded as the major variables. The relations reflected in Figure 4.7 are drawn from other studies that evaluated these relations in other countries [166][167][168][169].

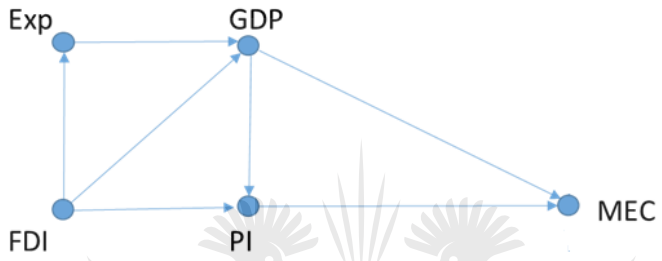

Figure 4.7: Graphical model for electricity consumption in the manufacturing sector

.

The directed path between the manufacturing production index (PI) and the manufacturing electricity consumption (MEC) does not have a conditioning variable as shown in Figure 4.7. And there are two backdoor paths from PI to MEC and they are both blocked by the collider structures. The path $PI \leftarrow FDI \rightarrow Exp \rightarrow GDP \rightarrow MEC$ is blocked by the for structure $\leftarrow FDI \rightarrow$ and the path $PI \leftarrow GDP \rightarrow MEC$ is blocked by the fork $\leftarrow GDP \rightarrow$. This means that PI can be used to estimate the causal effect between the production index and electricity consumption. Similarly, price or tariff (P) and technological efficiency (TE) can be used to estimate the causal effect. Sets such as Exp, GDP and FDI unblocks backdoor paths. For example FDI has directed path $FDI \rightarrow PI \rightarrow MEC$ but has a backdoor path $FDI \rightarrow GDP \rightarrow MEC$. This means that these conditioning sets cannot be used to estimate the causal effect. The fluctuations of the PI reflects the dynamics of other economic variables such as the GDP, FDI, Exp and other unobserved economic variables. This makes it possible to regress of PI and get the causal effect between the PI and the MEC.

4.4.2 Electricity and the mining industry

The discovery of mineral deposits South Africa led to the emergence of one of the largest mining industry in the world. It was the discovery of diamonds and gold in the 1850's that changed South Africa's economic path from an agrarian economy to a modern industrial economy [170]. The birth of the electricity industry and later the formation of Eskom was due to the mining industry. This means that when the electricity industry began the main user was the mining industry. Over the years since the discovery of mineral resources the contribution of the mining industry to the total GDP of South has declined. This is due to the lowering demand for minerals which gained a boost in the early 2000 by the rapid growth in china, the growth of the services sector and lastly, by the growth of the manufacturing industry. The decline in the economic contribution is also reflected in the consumption of electricity by the mining sector which has been on the decline. For example the consumption of the mining sector in 1950 was 58.8% of the total energy consumption in the country and by 2000 the consumption had declined to 18% [102]. Currently the South African mining industry consumes 15% of the total electricity supplied to the South African consumers. Within the mining industry, Gold mining sector is the largest consumer, consuming 47% of the electricity [171]. Gold mining is followed by the platinum sector which consumes 33% and the remainder of the mining activity consumes the remaining 20% [171]. The process of mining requires a large amount of electricity supply. The electricity is used to energize areas in mining such as Materials handling, processing, compressed air, pumping, fans, cooling, lighting and other associated activities that require electricity.

South Africa has the highest known reserves of platinum in the world boosting about 80% of the world reserves. These reserves are concentrated on the bushveld complex. Platinum mines tend to progress deeper as shallower areas of the reefs constituting the Bushveld Complex are being depleted, resulting in increasing energy demands and labour efforts. South Africa also has some of the deepest gold mines in the world. TauTona mine or Western deep No. 3 is the deepest gold mine in the world going as deep as 3.9 Kilometres (Km) underground [172]. The impact of deep mining is that of higher electricity demand to keep the operations going. In

comparison, the mining industry relies much more on electrical energy than on liquid fuels for its operations. The main reason for this is that the mineral and metal processing consumes large amounts of electricity. The energy demand in the mining industry is influenced by production level, energy efficiency and commodity price. Energy requirements in mining vary with the type of mineral being mined, whether it is underground or on the surface and also the extent to which it must be beneficiated or processed. Energy requirements for underground gold mining are significantly higher on a per ton basis than underground coal mining where the resource can be obtained in larger amounts. This is because the energy consumption in the mining industry is determined by the quantity of material that must be handled for every ton of useful resource. If large quantities of material must be extracted, transported, processed and disposed, then large amounts of energy will be required per unit of production of that particular resource. Furthermore, underground mining operations require much higher amounts of electricity than surface operations. liquid fuel is extensively used in surface mining for haulage whereas underground mining use electricity for activities such as hoisting to the surface and mine ventilation. A deep mine requires more pumps or bigger pumps for water distribution throughout the mine.

Consumption of electricity in the mining industry

Power on the mines is primarily used for the following activities: Refrigeration and ventilation; Compressed air generation; Pumping; Vertical transportation / hoisting; Conveying of materials; Milling; Processing; Arc furnaces; Hostels - lighting / heating and cooling; and Administration offices. Pumping, cooling and ventilation consume about 50% of the total power used by mining.

Mining production index

The index of the volume of mining production, also known as the production index, is a statistical measure of the change in the volume of production. The production index of a mineral group is the ratio between the volume of production of a mineral group in a given period and

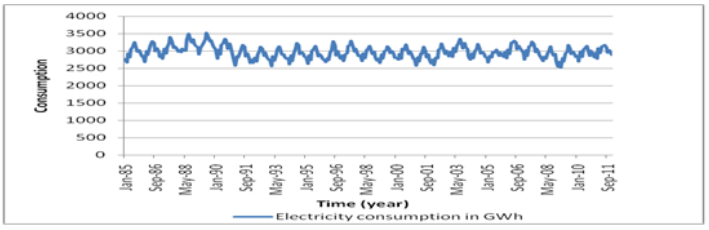

Figure 4.8: Electricity consumption in the manufacturing sector

.

the volume of production of the same mineral group in the base period. The base period in the data used is 2000. The production in the base period is set at 100.

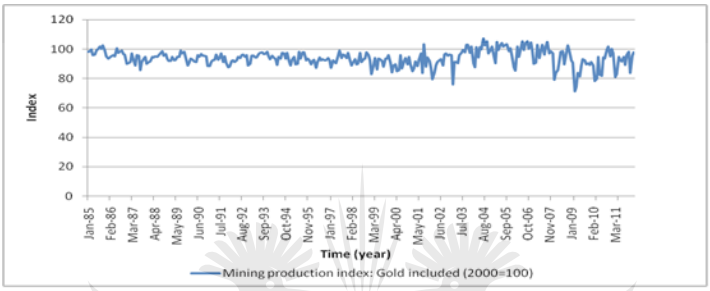

Figure 4.9: Mining production Index

.

SCM and mining industry

South Africas mineral industry can be broken down into five broad categories gold, PGM, diamonds, coal and vanadium [173]. South Africa's mineral industry is export-oriented, due to the small domestic market for most these commodities. Figure 4.10 illustrates that foreign direct investment has a direct impact on the increase or decrease of exports, gross domestic product and mining production index. The mineral prices determine whether a deeper mine will be profitable or not, which has an impact on the production index. Prices also determine if mining companies should continue to export. All these activities have an impact on the GDP of the country. Mining depth and production index are linked to electricity consumption. The effect of price of electricity conumption in the mining industry was the same as in the manufacturing sector, neglegible. In summary factors determining electricity consumption in the mining sector include:

1. Income/Gross Domestic product (GDP)
2. Export (Exp)
3. mining Production levels (MPI)
4. Foreign direct investment (FDI)
5. Mineral prices (MP)
6. Mine depth (MD)

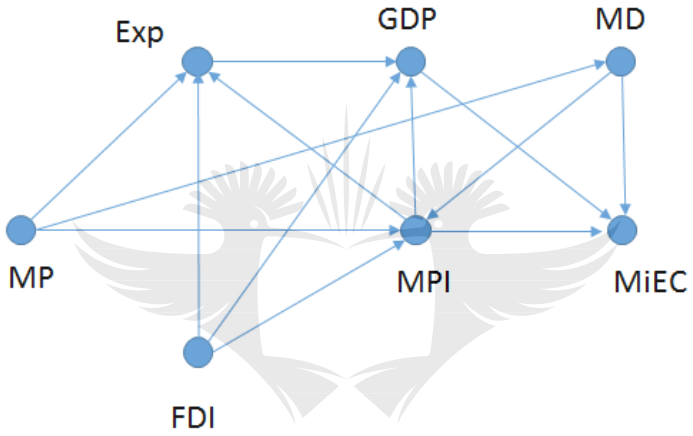

Figure 4.10: Graphical model for electricity consumption in the mining sector

From Figure 4.10, it is clear each of the variables cannot be used for the estimation of the causal effect because each path has a back-door path that is not blocked. For the purpose of this study MPI is the independent variable of interest to determine electricity consumption to test whether SCM results can be falsified.

4.5 Conclusion

Causal analysis is a useful concept in forecasting. Standard statistical analysis, typified by regression and other estimation techniques, is to infer parameters of a distribution from samples drawn of that population so long as experimental conditions remain the same. Causal analysis

goes one step further; its aim is to infer aspects of the data generation process which makes it possible to deduce not only the likelihood of events under static conditions, but also the dynamics of events under changing conditions [63].

Causal analysis relies on unproven assumption that there may exist a causal relationship between two or more variables. The introduction of the SCM enhances causal analysis in that allows researchers to move beyond just data analysis and reason about causal effect identification. SCM provides an analytical framework for causal effect estimation that brings theoretical understanding of the problem particularly the nature of the dependencies to the fore. The estimation of the causal analysis is conducted in the in Chapter 6 using the analytical framework covered in Section 4.3.

Chapter 5

Extreme Learning Machines and Forecasting

Model estimation methods are always striving to improve the accuracy of the approximation and reducing the computation time. Improving the accuracy requires the methodology used to better capture the dynamics of the system under consideration and the ability to reach or approach the universal minima during training. Methods that have long computation time are computationally costly and therefore, it is better to have methods that use less computational time. It is within this context that extreme learning machines have been introduced as an improvement in accuracy and computation time of the single hidden layer feedfoward neural networks.

5.1 Backpropagation networks

A widely used method for training a single-hidden layer feedforward neural networks (SLFNs) is the gradient descent algorithm. The mathematical representation for SLFNs is given by:

$$y_k = f\left(\sum_{j=1}^{M} w_{kj}^{(2)}\left(\sum_{i=1}^{n} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{5.1}$$

5.1.1 Network training

Given a training set comprising a set of inputs x_n, where $n = 1, \ldots\ldots\ldots, N$, together with a corresponding set of target vectors t_n, the objective is to minimise the error function.

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} ||y(x_n, w) - t_n)||^2 \tag{5.2}$$

where E is the total error of all patterns and the index n ranges over the set of input patterns. The variable t_n is the desired output for the nth output neuron when the nth pattern is presented, and $y_{n,w}$ is the actual output of the nth output neuron when pattern n is presented.

This type of learning is called supervised learning, where every input has an associated target output. After the computation of the error the weight vector is then updated as follows:

$$w_{new} = w_{old} - \nabla_w E(w) \tag{5.3}$$

where $\nabla E(w)$ is the gradient.

$$\nabla_w = [\frac{\partial}{\partial w_0}, \frac{\partial}{\partial w_1}, \ldots\ldots\ldots, \frac{\partial}{\partial w_n}] \tag{5.4}$$

so each k, w can be updated by:

$$w_k = w_k + \Delta w_k \tag{5.5}$$

where

$$\Delta w_k = -\eta \frac{\partial E}{\partial w_k} \tag{5.6}$$

where η is the learning rate.The weights of the neural network are optimized via back propagation training using, most commonly, scaled conjugate gradient method [174]. The cost function representing the objective of the training of the neural network can be defined. The objective of the problem is to obtain the optimal weights which accurately map the inputs of a

process to the outputs.

This method of learning has several drawbacks:

- When the learning rate η is too small, the learning algorithm converges very slowly. However, when η is too large, the algorithm becomes unstable and diverges.
- Another peculiarity of the error surface that impacts the performance of the BP learning algorithm is the presence of local minima [174].
- It is undesirable that the learning algorithm stops at a local minima if it is located far above a global minima.
- Neural network may be over-trained by using BP algorithms and obtain worse generalization performance. Thus, validation and suitable stopping methods are required in the cost function minimization procedure.
- Gradient-based learning is very time-consuming in most applications. Extreme earning machine is an algorithm that aims to overcome these drawbacks.

5.1.2 Extreme Learning Machines

Extreme learning machines were proposed by Huang et al [175] for SLFN architecture. With further development ELM is now a learning technique that provides efficient unified solutions to generalized feed-forward networks including but not limited to (both single- and multi-hidden-layer) neural networks, radial basis function (RBF) networks, and kernel learning [175].

The standard SLFNs in equation 5.1 with N hidden nodes with activation function f can approximate these N samples with zero error means that $\sum_{j=1}^{N} \|o_j - t_j\| = 0$ there exist w_i, b_i and β_j such that

$$\sum_{j=1}^{N} \beta_j g(\sum_{i=n}{}_k w_{ij} x_i + b_i) = o_j \tag{5.7}$$

The above N equations can be written compactly as

$$H\beta = T \tag{5.8}$$

where

$$H = (w_1, ...w_{\tilde{N}}, b_1, ..., b_{\tilde{N}}, x_1, ..., x_N = \begin{bmatrix} g(w_1 x_1 + b_1), ..., g(w_M x_1 + b_{\tilde{N}}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ g(w_1 x_N + b_1), ..., g(w_{\tilde{N}} x_N + b_{\tilde{N}}) \end{bmatrix} \tag{5.9}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix} \tag{5.10}$$

and

$$T = \begin{bmatrix} x_1^T \\ \vdots \\ \vdots \\ x_{\tilde{N}}^T \end{bmatrix} \tag{5.11}$$

H is called the the hidden layer output matrix of the neural network; the column ith of H is the ith hidden node output with respect to inputs $x_1, ..., x_N$ [176][177].

For fixed input weights w_i and the hidden layer biases b_i, seen from Eq. 5.9, to train an

SLFN is simply equivalent to finding a least squares solution β of the linear system $H\beta = T$

$$\|H(w_1, \ldots\ldots\ldots, w_{\tilde{N}}, b_1, \ldots\ldots\ldots, b_{\tilde{N}})\tilde{\beta} - T\| = \min_{\beta} \|H(w_1, \ldots\ldots\ldots, w_{\tilde{N}}, b_1, \ldots\ldots\ldots, b_{\tilde{N}})\beta - T\| \tag{5.12}$$

If the number M of hidden nodes is equal to the number N of distinct training samples, $M = N$, matrix H is square and invertible when the input weight vectors w_i and the hidden biases b_i are randomly chosen, and SLFNs can approximate these training samples with zero error.

However, in most cases the number of hidden nodes is much less than the number of distinct training samples, $\tilde{N} << N$, H is a non-square matrix and there may not exist $w_i, b_i; \beta_i (i = 1, \ldots, \tilde{N})$ such that $H\beta = T$. According to Theorem 5.1 in the Appendix C, the smallest norm least squares solution of the above linear system is

$$\beta = H^* T \tag{5.13}$$

where H^* is the Moore-Penrose generalized inverse of matrix H [178][179]. A regularization term is added to improve generalization performance and make the solution more robust, as shown:

$$\beta = (\frac{1}{C} + HH^T)^{-1} HH^T \tag{5.14}$$

Proposed learning algorithm for SLFNs

Thus, a simple learning method for SLFNs called extreme learning machine (ELM) can be summarized as follows:

Algorithm ELM: Given a training set $Q = (x_i, t_i) | x_i \epsilon R^n, t_i \epsilon R^m, i = 1, \ldots\ldots\ldots, \tilde{N}$ activation function $g(x)$, and hidden node number$\tilde{N}$,

Step 1: Randomly assign input weight w_i and bias b_i, $(i = 1, \ldots\ldots, \tilde{N})$.

Step 2: Calculate the hidden layer output matrix H.

Step 3: Calculate the output weight β from the equation: $\beta = H^*T$

5.2 Training ELM

During training, hidden layer performs a mapping of the original d-dimensional space into an L-dimensional space through the random matrix R, which is set independently from the distribution of the training data. In principle, the feature mapping phase may either involve a reduction in dimensionality $(L < d)$ or conversely, map the input space into in an expanded space $(L > d)$ [176].

With ELM algorithm, the input weights (of the connections linking the input neurons to hidden neurons) and the bias of the hidden neurons are randomly generated based on continuous distribution probabilities and kept fixed [176].

5.3 Optimally Pruned Extreme Learning Machine

The Optimaly-Pruned Extreme Learning Machine (OP-ELM) is regarded as an improvement of the original Extreme Learning Machine (ELM) proposed in [180] by Y. Miche et al. ELM has certain shortcomings that need to be overcome: ELM models tend to have problems when irrelevant or correlated variables are present in the training data set [181]. A method called optimally pruned extreme machine learning has been introduced to over the ELM drawback. This method involves the pruning of irrelevant variables by pruning of the related neurons of the SLFN built by the ELM. OP-ELM methodology has three main steps:

Step 1: The construction of an MLP model using ELM algorithm

Step 2: Ranking of neurons through Multiresponse Sparse Regression (MRSR) method

Step 3: Selection of the optimal number of neurons using Leave-One-Out (LOO) validation method

The OP-ELM algorithm brings together a combination of three different types of kernels, for robustness and more generality, whereas the original ELM uses only sigmoid kernels cite-journal:miche2. The kernel types used in OP-ELM include linear, sigmoid, and Gaussian kernels. Linear kernels included in the network help to model linear relationships between the input and the output.

5.3.1 Multiresponse Sparse Regression

MRSR was introduced by by Simila and Tikka [182] , for eliminating irrelevant neurons in the hidden layer of the MLP. Suppose that the targets are denoted by a matrix $n \times m, T = [t_1, \ldots\ldots, t_n]$ and the matrix $n \times p, X = [x_1, \ldots\ldots, x_n]$ denotes regressors. The MRSR algorithm adds sequentially active regressors to the mode

$$Y^k = WX^k \tag{5.15}$$

where $Y^k = [y_1^k, \ldots\ldots\ldots\ldots, y_n^k]$ is the target approximation of the model and W^k is the weight matrix with k nonzero rows at the kth step of the MRSR. Every step introduces a new nonzero row and thus, a new regressor to the model. In the case $m = 1$ MRSR is similar the least angle regression (LARS) algorithm [182]. This makes MRSR rather an extension than an improvement of LARS. An important detail shared by the MRSR and the LARS is that the ranking obtained is exact, if the problem is linear. In fact, this is the case with the OP-ELM, since the neural network built in the previous step is linear between the hidden layer and the output. Therefore, the MRSR provides an exact ranking of the neurons for this problem. Because of the exact ranking provided by the MRSR, it is used to rank the kernels of the model. The target is the actual output y_i, while the "variables" considered by the MRSR are the outputs of the kernels $h_i = ker(x_i^T)$, the columns of K.

5.3.2 Leave one out

After the MRSR provides a ranking of the kernels, the optimal number of neurons is determined for the model using an LOO validation method. The drawback with the LOO error is that it can be very time consuming, if the data sample is large. To cofound the time consuming process, the Prediction Sum of Squares (PRESS) statistics is used to provide a direct and exact formula for the calculation of the LOO error for linear models [183][184].

$$\varepsilon^{PRESS} = \frac{y_i - h_i b_i}{1 - h_i P h_i^T} \tag{5.16}$$

where$P$ is defined as $P = (H^T H)^{-1}$ and H is the hidden layer of the output matrix.

The final decision over the appropriate number of neurons for the model can then be taken by evaluating the LOO error versus the number of neurons used. Here, the neurons are already ranked by the MRSR. The convergence is faster, because the LOO error gets to the minimum faster when the MRSR is used than when it is not. Also, the number of neurons is far fewer in the LOO error minimum point when using the MRSR ranking, thus leading to sparser network with the same performance. In the end, an SLFN possibly using a mix of linear, sigmoid, and Gaussian kernels is obtained, with a highly reduced number of neurons, all within a small computational time.

5.4 Forecasting with Extreme learning machines

ELM is relatively new compared to other learning algorithms in artificial intelligence such MLP and as a result applications of the method are very few. There has been an increasing interest in ELM and this interest is demonstrated by the ELM conference that is held annually since 2012. Because of the training efficiency and speed, ELM is also gaining popularity amongst practitioners in dealing with big data problem (big data paper). Demand forecasting with ELM is still in its infancy and therefore, there are very few studies in this area. The current study is a novel study by using ELM in forecasting medium term electricity demand.

Mateo et al [185] used ELM for short-term Electric Power Demand Prediction and they found that did not improve the errors of the other machine learning techniques such LS-SVM, MLP etc. but their computational time is much lower while enabling reasonable estimations. Zhang et al [186] uses an ensemble model of a promising novel learning technology called extreme learning machine (ELM) for high-quality Short-Term Load Forecasting of Australian National Electricity Market (NEM). The model consists of a series of single ELMs. During the training, the ensemble model generalizes the randomness of single ELMs by selecting not only random input parameters but also random hidden nodes within a pre-defined range. The forecast result is taken as the median value the single ELM outputs. The approach takes advantage of the very fast training/tuning speed of ELM, to ensure that the model is efficiently updated to track, on-line, the variation trend of the electricity load and maintain the accuracy. The results show that the training efficiency and the forecasting accuracy of the ensemble model are superior over the competitive algorithms.

Shrivastava and Panigrahi [187] used ELM in the price forecasting problem. They coupled ELM with the Wavelet technique to develop a hybrid model termed as WELM (wavelet based ELM) to improve the forecasting accuracy as well as reliability. The experimental results demonstrated that the proposed method is one of the most suitable price forecasting techniques.

5.5 Granger causality with ELM and OPLM

The experiments conducted were performed to test whether the there is a unidirectional causal relationship from the manufacturing production index and mining production index to electricity consumption in the manufacturing sector and the mining sector respectively. These experiments are performed with electricity consumption as a univariate and secondly, with the indexes included on the right hand side of their respective regression equations.

5.5.1 Experimental setup

In this experiments, all the inputs (attributes) have been normalized into the range [0.1] as well as the outputs (targets). For univariate experiments the data features were structured as described in Section 3.4.1 of Chapter 4. For multivariate experiments, the index data was input window sizes equal in sizes as consumption. These data windows were combined with the consumption of the same to form one input vector with a target vector as the electricity consumption. This is represented as:

$$y_t = f(x_{t-1}, x_{t-2}, \ldots\ldots, x_{t-N}, z_{t-1}, z_{t-2}, \ldots\ldots z_{t-N}) \tag{5.17}$$

where x is the electricity consumption, z is the production index, N is the lag and y is the desired output. The electricity consumption time series was sampled on a monthly basis and has 324 data points. 150 instances where used as the input for training purposes and 100 instances were used for testing the trained model.

5.6 ELM results

In this experiments ELM is used to forecast electricity consumption using electricity consumption parameter. Past values of this parameter are used to predict future. The experiment is performed in a MATLAB environment. The matlab tool used for ELM experiments was written by Qin-Yu Zhu and Guang-Bin Huang from the University of Singapore.

5.6.1 Univariate results

The forecasting results for autoregressive models for manufacturing and the mining sector are presented in Table D.4 and Table D.5 in Appendix D . These are experimental results obtained from forecasting electricity consumption in the manufacturing sector and the mining sector using the past values of the same parameter, consumption. All the models were created with a sigmoid activation function.

5.6.2 Multivariate (with both consumption and index on the conditioning set) results

The results for consumption forecasting in the manufacturing and the mining sector using the manufacturing production index and the mining production index respectively, as an independent variable and excluding the lagged dependent variable, consumption, as part of the conditioning set. The results are presented in Table D.6 and Table D.7 in Appendix D .

5.6.3 Multivariate (with only index on the conditioning set) results

The results presented in Table D.8 and Table D.9 present experimental results for consumption in the manufacturing sector and the mining sector with a conditioning set that has only the lagged values of the manufacturing production index and the mining production index respectively.

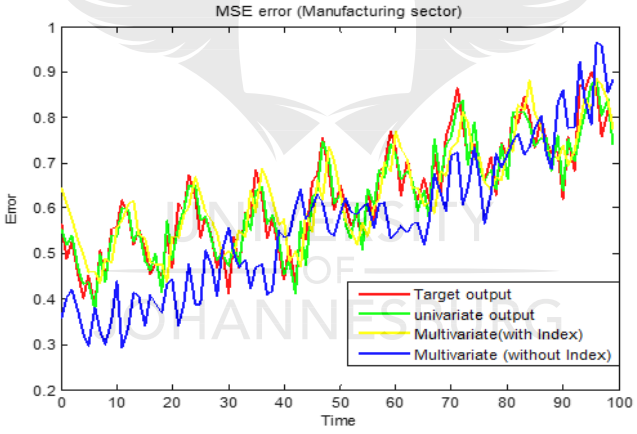

Figure 5.1: A comparison of the best predicted outputs of ELM models with different conditioning variables with the target output in the manufacturing sector

5.7 OP-ELM results

The Matlab toolbox used for OP-ELM experiments was developed by Lendasse et al.. The data was setup as described in Section 5.5.1.

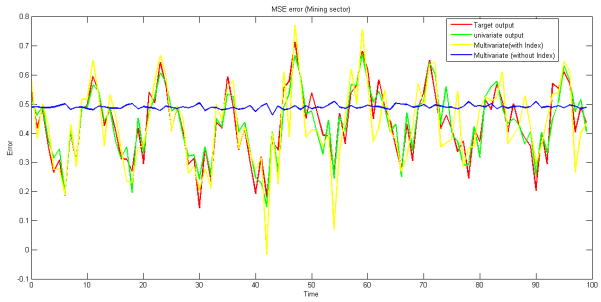

Figure 5.2: A comparison of the best predicted outputs of ELM models with different conditioning variables with the target output in the mining sector

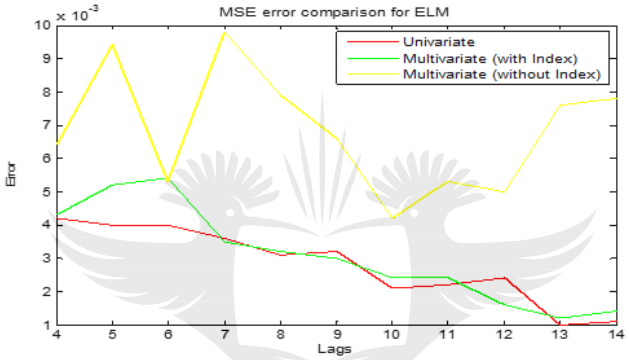

Figure 5.3: MSE error comparison ELM forecasting models in the manufacturing sector

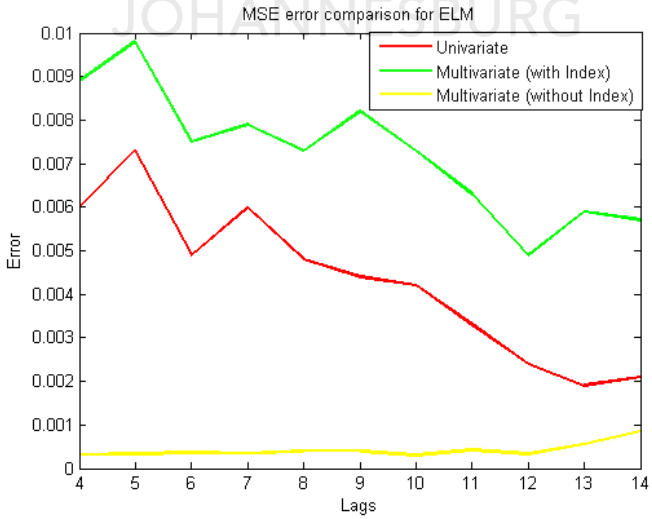

Figure 5.4: MSE error comparison ELM forecasting models in the mining sector

5.7.1 Univariate results

The results for autoregressive models are presented in Table D.10 and Table D.11 in Appendix D for the manufacturing sector and mining sector respectively. These are experimental results obtained from forecasting electricity consumption in the manufacturing sector using the past values of the same parameter, consumption.

5.7.2 Multivariate (with both consumption and index on the conditioning set) results

The results for consumption forecasting in the manufacturing sector and mining sector using the manufacturing production index and the mining production index respectively as an independent variable and excluding the lagged dependent variable, consumption, as part of the conditioning set. The results are presented in Table D.12and TableD.13 in Appendix D for the manufacturing sector and mining sector respectively.

5.7.3 Multivariate (with only index on the conditioning set) results

The results presented in Table D.14 and Table D.15 in Appendix D, present experimental results for the manufacturing sector and mining sector respectively. The conditioning set includes the production indexes only.

5.8 Conclusions

ELM and OP-ELM were used to perform causality experiments with electricity consumption data in the mining sector and the manufacturing sector. The objective of these experiments was to assess whether there is causal relation between the manufacturing production index and the mining production index and the electricity consumption in each of these sectors. The results show that in autoregressive experiments, OP-ELM performs significantly better than ELM. In the manufacturing sector, it was found that the ELM results with the model conditioned on

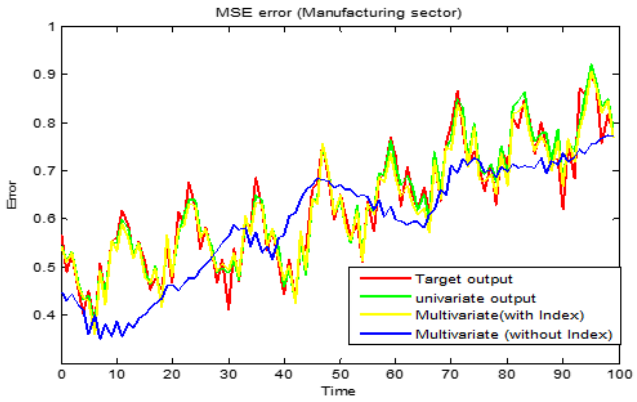

Figure 5.5: A comparison of the best predicted outputs of OP-ELM models with different conditioning variables with the target output in the manufacturing sector

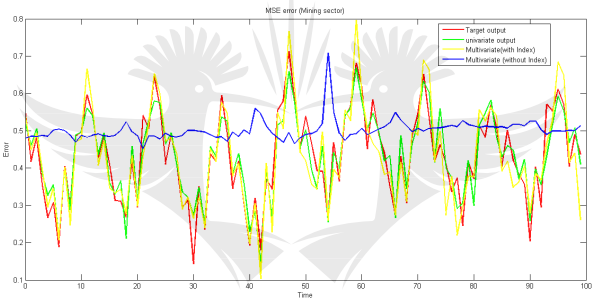

Figure 5.6: A comparison of the best predicted outputs of OP-ELM models with different conditioning variables with the target output in the mining sector

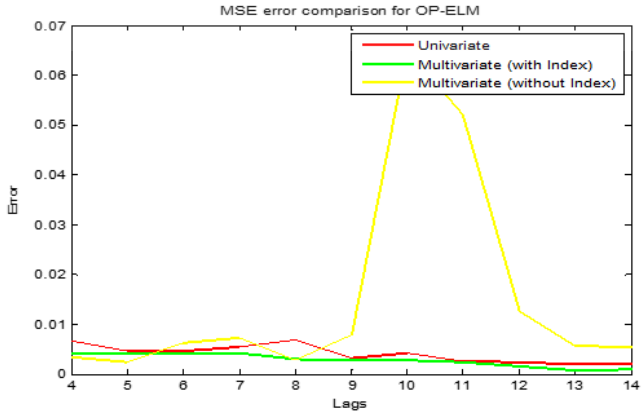

Figure 5.7: MSE error comparison for OP-ELM in the Manufacturing sector

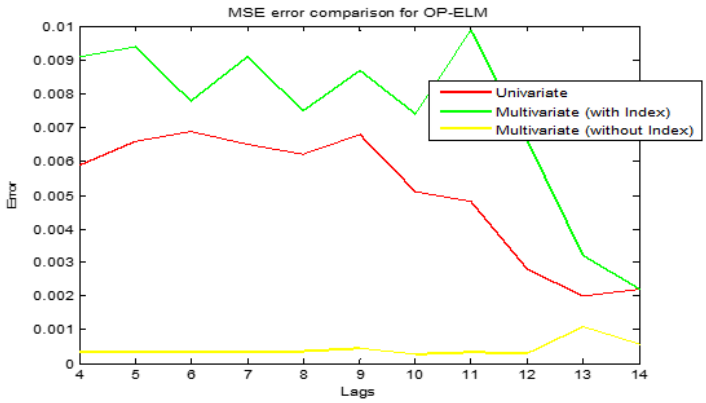

Figure 5.8: MSE error comparison for OP-ELM in the mining sector

both the production index and the consumption data the results were better than that of autoregression but they were not statistically significantly better. With OP-ELM the results were statistically significantly better than that of autoregression models. The results in the mining sector showed no statistically significant improvement when the production index was included in the conditioning set for both ELM and OP-ELM. The experiments with the production index only in conditioning set showed that the models were unable to track the fluctuations of electricity consumption in both sectors with both OP-ELM and ELM. It was found, therefore, that using OP-ELM models, there is a granger causal relationship running from the manufacturing index to electricity consumption in the manufacturing sector. No causal relation was found running from electricity consumption to manufacturing index.

Chapter 6

Conclusions

This study looked at the one step ahead forecasting of electricity consumption. The time series and the explanatory (causality approach) forecasting approaches were both explored to model this system and models were constructed accordingly. The mathematical tools chosen for modelling are the artificial intelligence tools. Neural networks, Neuro-fuzzy systems and support vector machines were used to develop nonlinear models in the dataset under consideration. These three techniques have been widely used but they provide very valuable insights in terms of modelling. As a contribution to the forecasting studies, this work introduces the use of structural causal models and the use of extreme learning machines to model the electricity consumption system. The modelling approaches adopted are firstly, the univariate which uses past values electricity consumption series to predict values and multivariate which uses economic variables such as manufacturing production index and the mining production index to predict the electricity consumption. The causal studies were conducted under the multivariate studies, for example the lagged values manufacturing production index (independent variable) were used to predict the future values of the electricity consumption (dependent variable). From the experiments performed the following conclusions were drawn:

- Non-linear methods, namely Artificial intelligence techniques performed significantly better than the linear method, ARMA.
- OP-ELM gave better forecasting results than all the other artificial intelligence techniques

- There is a casual relationship between manufacturing production index and electricity consumption in the manufacturing sector
- There was not causal relationship found between mining production index and electricity consumption in the mining sector
- Structural causal models were able to identify the causal relationships through graphical methods
- There is no justification for not including the dependent variable on the right-hand side of the regression equation
- The optimal number of lags for modeling was found to be thirteen.

6.1 Results Comparison and discussion

Several artificial intelligence techniques including a linear stochastic method have been used in this work to forecast electricity consumption. As it is usually the case, there are several plausible methods to forecast a time series. ANN, SVR, ANFIS and ARMA have been widely used for forecasting and can therefore be used as a benchmark for other techniques. ELM and OP-ELM are relatively new machine learning and are an improvement of conventional neural networks techniques. All the methods used for forecasting in this work have demonstrated the ability to forecast a time series. Beyond looking at the accuracy measure to see which method gives the best performance, it is also important to conduct statistical significance tests to test whether the differences in accuracy are significantly different. The best results of each method, meaning the model in each method with the most accurate results, were used to make comparison of all methods and between the new techniques and the ones that have been widely.

Considering the univariate total consumption forecasting results in Table 6.1 the OP-ELM model with thirteen lagged inputs had the most accurate perfomance. This is followed by ELM model twelve lagged inputs. ANN, ANFIS and SVR all have their best models at the twelve

Table 6.1: MSE errors for artificial intelligence techniques

Lags	MLP	ANFIS	SVR	ELM	OP-ELM
4[y(t-4)]	0.0212	0.0159	0.0090	0.0044	0.0052
5[y(t-5)]	0.0099	0.0113	0.0093	0.0033	0.0044
6[y(t-6)]	0.0126	0.0099	0.0078	0.0071	0.0033
7[y(t-7)]	0.0163	0.0094	0.0072	0.0038	0.0039
8[y(t-8)]	0.0186	0.0087	0.0061	0.0044	0.0038
9[y(t-9)]	0.0133	0.0064	0.0060	0.0024	0.0038
10[y(t-10)]	0.0131	0.0062	0.0050	0.0039	0.0022
11[y(t-11)]	0.0116	0.0060	0.0033	0.0020	0.0018
12[y(t-12)]	0.0075	0.0033	0.0027	0.0011	0.00098
13[y(t-13)]	0.0082	0.0035	0.0026	0.0013	0,00064
14[y(t-14)]	0.0092	0.0044	0.0026	0.0021	0.0012

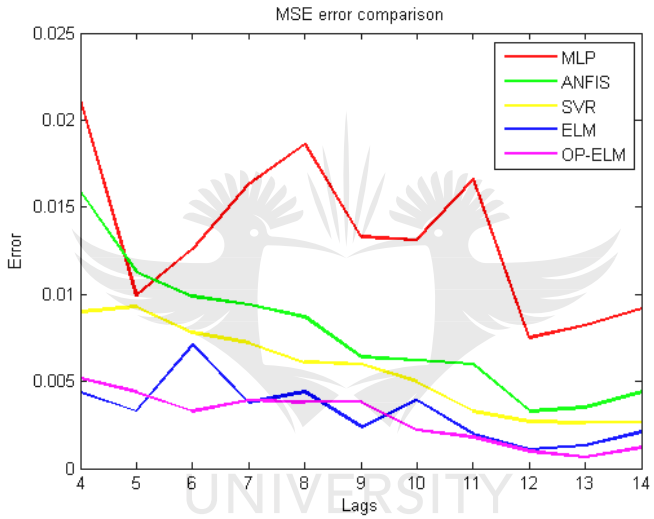

Figure 6.1: A comparison of the MSE errors for different AI techniques at different lags

lagged inputs with SVR in accuracy followed by ANFIS and lastly MLP has the poorest performance in terms of accuracy of all the AI methods as illustrated in Figure 6.1. MLP performs better than ARMA which has a lag of five and stationarity in the first difference. Neural networks and support vector machines are the most widely used artificial intelligence techniques. However, it has now become well known that both of them including the neurofuzzy technique suffer from, slow learning speed and poor computational scalability. ELM together with its improvement, OP-ELM have the ability to overcome these challenges because the hidden layer parameters need not be tuned. In addition, these techniques provide better computation general-

ization results as demonstrated by the experimental results presented in Table 6.1. Using MSE as an accuracy measure, the performance of the OP-ELM prediction model conditioned on the lagged values of both the manufacturing production index and the electricity consumption that is better than that of a model conditioned on electricity consumption alone. A statistical significance test is conducted for the model with thirteen inputs for both and it is found the performance of the former is significantly better than that of the latter. By the Granger causality definition there is a causal relationship between the manufacturing production index. The results are in line with the findings of the structural causal analysis in chapter 5 that identified manufacturing production index as a causal variable. In the mining the sector, the results for this two models are not very different. Conditioning the regression equation of the lagged values of both production index and electricity consumption does not yield more accuracy that of the univariate model. The statistical significance test confirms this finding and the null hypothesis that the two means are the same is not rejected at a tolerance level of 1% and 5%. The experiments performed on both ELM and OP-ELM in both the mining sector and the manufacturing with the conditioning set with only the production index yielded disappointing results. In these experiments the question of whether to exclude the dependent variable on the right hand side of the regression equation is answered in the negative. The figures 20, 21, 24 and 25 illustrate the failure of the models, with the index only in the conditioning set, to predict or track the fluctuations of the electricity consumption in the respective sectors. The computation time was reviewed for all five AI methodologies. It can be noted that for all five methodologies, the computational times for the test phase are negligible compared to the training times; this is especially clear for large training times, like the SVM, ANFIS or MLP ones. SVM is the slowest of all the AI techniques. ELM is the fastest algorithm by several orders of magnitude compared, for example, to the SVM. This is in line with the claims of the ELM authors. The proposed OP-ELM is between one and three orders of magnitude slower than the original ELM, but still much faster than the rest of the compared methods in this data set.

The study met all the objectives that were set out in the beginning of the study. Firstly, the study explored various forecasting tools for forecasting one step ahead monthly electricity con-

sumption with varying success. ELM techniques were found to yield the most accurate results especially the optimally pruned ELM. However, it is important to mention that ELM does not overcome the challenge of overfitting, hence the intoduction of the pruning technique which helps to eliminate unnecesary hidden units and therefore reducing the number of parameters. The random selection of input weights is vulnerable to the selection of suboptimal weights which may have an impact on the modeling ability. Research on ELM has been focusing on developing methodologies for improving the selection of the input weisghts. Furthermore, ELM relies on emperical risk minimisation which has its weaknesses.

Secondly, structural causal model was introduced and used successfully to identify control variables for causal models. The manufacturing and mining production indices were used as causal variables for the electricity consumption models in the manufacturing the mining sectors respectively. SCM provides a powerful framework for reasoning about the electricity consumption problem. However, the weakness of SCM is that it relies on unproven assumed causal directions betweeen variables based on the reasoning of the researcher.

Thirdly and lastly, ELM tools were found to have a very low computation time compared to other forecasting techniques. However, the pruning technique used to optimize the number of hidden neurons introduced a slight delay increasing the ELM computation time. The lower computation time is as a result of the elimination of backpropagation in ELM learning. The computation time can be reduced in that ELM generates its hidden layers randomly and it usually requires more hidden neurons than that of a convectional neural networks to achieve matched perfomance. This may result in a larger than desired network size which requires longer running time in the testing phase of the ELM.

6.2 Further work

Extreme learning machines are relatively new tools which are still undergoing research to futher improve their stability and generalisation perfomance and therefore, need further exploration. ELM which uses the structural risk minimisation approach to learning can also be explored as

opposed to emperical risk minimisation.

South Africa has just recently introduced renewable energy which include wind, solar, hydrogen fuel cell, etc. into the energy mix. These energy sources make demand forecasting even more important because of the way they operate. Solar works during day light and wind works when there is wind blowing and therefore, cannot be used for base load.

Appendix A

Hard C-means

The classical 'hard' c-means clustering algorithm has the following characteristics,

- Each data point is a member of ONE and only ONE cluster;
- The number of clusters, c, necessary to correctly cluster the data is known a priori.

  Moreover,

$$2 \leq C < P \tag{A.1}$$

  where, P is the number of data points. More formally we need to identify the characteristic function, F, relating each data point, x_k, to one of a family of sets $\{A_i, i = 1, 2, \ldots\ldots, C\}$.

  Thus

$$F_{(A_i)}(x_k) = \begin{cases} 0 & \text{if } x_k \in A_i \\ 1 & \text{if } x_k \notin A_i \end{cases}$$

  Some properties of the (hard c-means) clustering process Property 1 - the union of all

cluster (sets), A_i, spans the set of data points, X

$$\bigcup_{i-}^{\cup} F_{A_i}(x_k) = i \forall k \tag{A.2}$$

Property 2 - there is no overlap between clusters

$$F_{A_i}(x_k) \cap F_{A_i}(x_k) = i \forall k \tag{A.3}$$

Property 3 - clusters cannot be empty, and cannot contain all data points.

$$0 < \sum F_{A_i}(x_k) < P \forall k \tag{A.4}$$

section Defining a Cluster

So far we have just set the scene - nothing has been said about how data points are related to a specific cluster.

Let matrix, U, be a $c \times P$ matrix of assignments between data points and clusters.

That is to say, if $F_{ij} = F_{A_i}(x_j)$ represents the membership (0 or 1) between the jth data point and ith cluster, then,

- $\cup$ is a matrix of $F_i j (i = 1, 2, \ldots\ldots, c_j; j = 1, 2, \ldots, P)$

Let $M_c p$ be the universe of hard $'c'$ partitions, or

- The allocation of memberships (0 or 1) such that each data point us associated with one class;

$$M_{cp} = \{U \mid F_{ij} \in [0, 1]; \sum_{i=1}^{c} F_{ik} = 1; 0 < \sum_{i=1}^{P} F_i k < P\} \tag{A.5}$$

A.1 Ranking Clusters

- What represents a good cluster and what represents a bad cluster?
- Objective function differentiates between quality of different cluster allocations.
- C-means algorithm uses a sum of distances between,
  1. Proposed cluster and,
  2. Associated data points belonging to this cluster.
- Objective is to find the best centroid and allocation of data points such that the distance is minimized.

That is we wish to minimize,

$$J(U,V) = \sum_{k=1}^{P} \sum_{i=1}^{c} F_{ik}(d_{ik})^2 \tag{A.6}$$

where, d_{ik} is a suitable distance metric, say an Euclidean norm, between the kth data sample, x_k, and ith cluster center v_i,

$$d_{ik} = d(x_k - v_i) = \|x_k - v_i\| = [\sum_{j=1}^{m}(x_{ik} - v_{ik})^2]^{1/2} \tag{A.7}$$

i.e. each data point lies in an 'm' dimensional space. We need the optimal combination, (U^*, V^*) , minimizing $J(U,V)$

Appendix B

Causality

Theorem 1 (The Causal Markov Condition). Any distribution generated by a Markovian model M can be factorized as:

$$P = \prod_i P(v_i \mid pa_i) \tag{B.1}$$

where $V_1, V_2, \ldots\ldots, V_n$ are the endogenous variables in M, and pa_i are (values of) the endogenous "parents" of V_i the causal diagram associated with M.

Corollary 1 (Truncated factorization). For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the truncated factorization

$$P(v_1, v_2, \ldots\ldots, v_n|do(x_0)) = \prod_{(} i \mid V \notin X) P(v_i \mid pa_i) \mid_{x=x_0} \tag{B.2}$$

where $P(v_i \mid pa_i)$ are the pre-intervention conditional probabilities.

Theorem 2 . A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that every path between X and any of its children traces at least one arrow emanating from a measured variable.

Appendix C

Moore-Penrose

The resolution of a general linear system $Ax = y$, where A may be singular and may even not be square, can be made very simple by the use of the Moore-Penrose generalized inverse

Definition 5.1 . A matrix G of order $n \times m$ is the Moore-Penrose generalized inverse of matrix A of order $m \times n$, if

$$AGA = A; GAG = G; (AG)^T = AG; (GA)^T = GA. \tag{C.1}$$

For the sake of convenience, the Moore-Penrose generalized inverse of matrix A will be denoted by A^*.

C.1 Minimum norm least-squares solution of general linear system

For a general linear system $Ax = y$, we say that $\tilde{x}$ is a least-squares solution (l.s.s) if

$$\|A\tilde{x} - y\| = min_x\|Ax - y\|, \tag{C.2}$$

where $\|.\|$ is a norm in Euclidean space.

Definition 5.2 . $x_0 \in R^n$ is said to be a minimum norm least squares solution of a linear system $Ax = y$ if for any $y \in R^m$

$$\|x_0\| \leq \|x\|, \forall x \in x : \|A\tilde{x} - y\| \leq \|Az - y\|, \forall z \in R^n \tag{C.3}$$

That means, a solution x_0 is said to be a minimum norm least-squares solution of a linear system $Ax = y$ if it has the smallest norm among all the least-squares solutions.

Theorem 5.1 . Let there exist a matrix G such that Gy is a minimum norm least-squares solution of a linear system $Ax = y$. Then it is necessary and sufficient that $G = A^*$, the Moore-Penrose generalized inverse of matrix A.

Remark . Seen from Theorem 5.1, we can have the following properties key to our proposed ELM learning algorithm:

1. The special solution $x_0 = A^*y$ is one of the least squares solutions of a general linear system $Ax = y$:

$$\|Ax_0 - y\| = \|AA^* - y\| = \min_x \|Ax - y\| \leq \|Az - y\| \tag{C.4}$$

2. In further, the special solution $x_0 = A^*y$ has the smallest norm among all the least-squares solutions of $Ax = y$:

$$\|x_0\| = \|A^* \leq \|x\|,$$

$$\forall x \in x : \|A\tilde{x} - y\| \leq \|Az - y\|, \forall z \in R^n \tag{C.5}$$

3. The minimum norm least-squares solution of $Ax = y$ is unique, which is $x_0 = A^*y$.

Appendix D

Results

Table D.1: MLP results

No of inputs	MSE
4[y(t-4)]	0.0212
5[y(t-5)]	0.0099
6[y(t-6)]	0.0126
7[y(t-7)]	0.0163
8[y(t-8)]	0.0186
9[y(t-9)]	0.0133
10[y(t-10)]	0.0131
11[y(t-11)]	0.0116
12[y(t-12)]	0.0075
13[y(t-13)]	0.0082
14[y(t-14)]	0.0092

Table D.2: ANFIS results

No of inputs	MSE
4[y(t-4)]	0.0159
5[y(t-5)]	0.0113
6[y(t-6)]	0.0099
7[y(t-7)]	0.0094
8[y(t-8)]	0.0087
9[y(t-9)]	0.0070
10[y(t-10)]	0.0062
11[y(t-11)]	0.0058
12[y(t-12)]	0.0033
13[y(t-13)]	0.0039
14[y(t-14)]	0.0044

Table D.3: SVR results

No of inputs	MSE
4[y(t-4)]	0.0090
5[y(t-5)]	0.0093
6[y(t-6)]	0.0078
7[y(t-7)]	0.0072
8[y(t-8)]	0.0061
9[y(t-9)]	0.0060
10[y(t-10)]	0.0050
11[y(t-11)]	0.0033
12[y(t-12)]	0.0027
13[y(t-13)]	0.0026
14[y(t-14)]	0.0026

Table D.4: ELM univariate results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0042
5[y(t-5)]	0.0040
6[y(t-6)]	0.0040
7[y(t-7)]	0.0036
8[y(t-8)]	0.0031
9[y(t-9)]	0.0032
10[y(t-10)]	0.0021
11[y(t-11)]	0.0022
12[y(t-12)]	0.0024
13[y(t-13)]	0.00085
14[y(t-14)]	0.0092

Table D.5: OP-ELM univariate results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.0060
5[y(t-5)]	0.0073
6[y(t-6)]	0.0049
7[y(t-7)]	0.0060
8[y(t-8)]	0.0048
9[y(t-9)]	0.0044
10[y(t-10)]	0.0042
11[y(t-11)]	0.0033
12[y(t-12)]	0.0024
13[y(t-13)]	0.0019
14[y(t-14)]	0.0021

Table D.6: ELM multivariate results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0043
5[y(t-5)]	0.0052
6[y(t-6)]	0.0054
7[y(t-7)]	0.0035
8[y(t-8)]	0.0032
9[y(t-9)]	0.0030
10[y(t-10)]	0.0024
11[y(t-11)]	0.0024
12[y(t-12)]	0.0016
13[y(t-13)]	0.0012
14[y(t-14)]	0.0014

Table D.7: ELM multivariate results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.0089
5[y(t-5)]	0.0098
6[y(t-6)]	0.0075
7[y(t-7)]	0.0079
8[y(t-8)]	0.0073
9[y(t-9)]	0.0082
10[y(t-10)]	0.0073
11[y(t-11)]	0.0063
12[y(t-12)]	0.0049
13[y(t-13)]	0.0059
14[y(t-14)]	0.0057

Table D.8: ELM multivariate (with index only) results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0089
5[y(t-5)]	0.0098
6[y(t-6)]	0.0075
7[y(t-7)]	0.0079
8[y(t-8)]	0.0073
9[y(t-9)]	0.0082
10[y(t-10)]	0.0073
11[y(t-11)]	0.0063
12[y(t-12)]	0.0049
13[y(t-13)]	0.0059
14[y(t-14)]	0.0057

Table D.9: ELM multivariate (with index only) results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.032
5[y(t-5)]	0.034
6[y(t-6)]	0.037
7[y(t-7)]	0.035
8[y(t-8)]	0.04
9[y(t-9)]	0.04
10[y(t-10)]	0.031
11[y(t-11)]	0.042
12[y(t-12)]	0.034
13[y(t-13)]	0.055
14[y(t-14)]	0.085

Table D.10: OP-ELM univariate results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0066
5[y(t-5)]	0.0047
6[y(t-6)]	0.0046
7[y(t-7)]	0.0054
8[y(t-8)]	0.0068
9[y(t-9)]	0.0032
10[y(t-10)]	0.0042
11[y(t-11)]	0.0025
12[y(t-12)]	0.0024
13[y(t-13)]	0.0019
14[y(t-14)]	0.0021

Table D.11: OP-ELM univariate results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.0059
5[y(t-5)]	0.0066
6[y(t-6)]	0.0069
7[y(t-7)]	0.0065
8[y(t-8)]	0.0062
9[y(t-9)]	0.0068
10[y(t-10)]	0.0051
11[y(t-11)]	0.0048
12[y(t-12)]	0.0028
13[y(t-13)]	0.0020
14[y(t-14)]	0.0022

Table D.12: OP-ELM multivariate (with index only) results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0039
5[y(t-5)]	0.0042
6[y(t-6)]	0.0042
7[y(t-7)]	0.0041
8[y(t-8)]	0.0029
9[y(t-9)]	0.0028
10[y(t-10)]	0.0027
11[y(t-11)]	0.0024
12[y(t-12)]	0.0015
13[y(t-13)]	0.0007
14[y(t-14)]	0.0009

Table D.13: OP-ELM multivariate results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.0091
5[y(t-5)]	0.0094
6[y(t-6)]	0.0078
7[y(t-7)]	0.0091
8[y(t-8)]	0.0075
9[y(t-9)]	0.0087
10[y(t-10)]	0.0074
11[y(t-11)]	0.0099
12[y(t-12)]	0.0068
13[y(t-13)]	0.0033
14[y(t-14)]	0.0025

Table D.14: OP-ELM multivariate (with index only) results for the Manufacturing sector

No of inputs	MSE
4[y(t-4)]	0.0033
5[y(t-5)]	0.0023
6[y(t-6)]	0.0062
7[y(t-7)]	0.0072
8[y(t-8)]	0.0028
9[y(t-9)]	0.0079
10[y(t-10)]	0.0651
11[y(t-11)]	0.0521
12[y(t-12)]	0.0125
13[y(t-13)]	0.0056
14[y(t-14)]	0.0054

Table D.15: OP-ELM multivariate (with index only) results for the Mining sector

No of inputs	MSE
4[y(t-4)]	0.033
5[y(t-5)]	0.033
6[y(t-6)]	0.037
7[y(t-7)]	0.036
8[y(t-8)]	0.037
9[y(t-9)]	0.044
10[y(t-10)]	0.029
11[y(t-11)]	0.034
12[y(t-12)]	0.030
13[y(t-13)]	0.011
14[y(t-14)]	0.058

Bibliography

[1] H. A. Simon, "Rational decision-making in business organizations," *Nobel Memorial Lecture*, 1978.

[2] T. Marwala, *Causality, Correlation and artificial Intelligence for rational Decision Making*. Singapore: World Scientific, 2015.

[3] Y. Rebours and D. Kirschen, "What is spinning reserve," *University of Manchester*, no. release 1, 2005.

[4] A. Pereira-Neto, O. Saavedra, C. Unsihuay, and J. Pessanha, "Profit based unit commitment considering the cold reserve under competitive environment," *Power Tech, IEEE Russia*, no. release 1, pp. 1–4, 2005.

[5] E. A. Feinberg and D. Genethliou, "Load forecasting, applied mathematics for restructured electric power systems," *Power Electronics and Power Systems*, pp. 269–285, 2005.

[6] Makridakis, Wheelwright, and McGee, *Forecasting: Methods and Applications*. 605 Third Avenue, New York 10158: John Wiley and Sons, Inc, 2nd ed., 1997.

[7] D. Stern, "A multivariate cointegration analysis of the role of energy in the us macroeconomy.," *Energy Economics*, vol. 22, pp. 267–283, 2000.

[8] E. Kyriakides and M. Polycarpou, "Short term electric load forecasting: A tutorial," *In: Chen, K., Wang, L. (Eds.), Trends in Neural Computation, Studies in Computational Intelligence, Springer*, vol. 35, pp. 391–418, 2007.

[9] R. Zivanovic, "Local regression-based short-term load forecasting," *Journal of Intelligent and Robotic Systems*, vol. 31, pp. 115–127, 2001.

[10] W. Charytoniuk, M. Chen, and P. V. Olinda, "Nonparametric regression based short-term load forecasting.," *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 725–730, 1998.

[11] L. Jin, Y. Lai, and T. Long, "Peak load forecasting based on robust regression," *In:Eighth International Conference on Probabilistic Methods Applied to Power Systems*, vol. 13, no. 3, pp. 123–128, 2004.

[12] C. Chatfield, *The analysis of time series: An introduction*. London: New York: Chapman and Hall, 4th ed. ed., 1989.

[13] H. Hahn, S. Meyer-Nieberg, and S. Pickl, "Electric load forecasting methods: Tools for decision making," *European Journal of Operational Research*, vol. 199, pp. 902–907, 2009.

[14] A. Abraham and B. Nath, "A neuro-fuzzy approach for modelling electricity demand in victoria," *Applied Soft Computing*, vol. 1, no. 2, pp. 127–138, 2001.

[15] G. A. Darbellay and M. Slama, "Forecasting the short-term demand for electricity - do neural networks stand a better chance?," *International Journal of Forecasting*, vol. 16, pp. 71–83, 2000.

[16] W. D. Laing and D. G. C. Smith, "A comparison of time series forecasting methods for predicting the cegb demand.," *Proceedings of the ninth power systems computation conference*, 1987.

[17] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing.," *Journal of the Operational Research Society*, vol. 54, pp. 23–49, 2003.

[18] Y. Wang, J. Wang, G. Zhao., and Y. Dong, "Application of residual modification approach in seasonal arima for electricity demand forecasting: A case study of china.," *Energy Policy*, vol. 48, pp. 284–294, 2012.

[19] C. W. Ostrom, *Time series analysis: Regression techniques*. Newbury Park, CA: Sage, 2nd ed. ed., 1990.

[20] D. P. Fan, *Prediction of public opinion from the mass media: Computer content analysis and mathematical modeling*. New York: Greenwood Press, 1988.

[21] J. Zhu, "Issue competition and attention distraction: A zero-sum theory of agenda-setting," *Journalism and Mass Communication Quarterly*, vol. 68, pp. 825–836, 1992.

[22] D. N. Gujarati, *Basic econometrics*. New York: McGraw-Hill Press, 3rd ed ed., 1995.

[23] C. Granger, "Strategies for modelling non linear time series relationship," *Economic Record*, no. 60, 1993.

[24] C. Granger, "Some recent developments in a concept of causality," *Journalof Econometrics*, vol. 39, no. 1, pp. 199–211, 1988.

[25] D. Dickey and W. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *J. Am. Stat. Ass.*, vol. 74, pp. 427–431, 1979.

[26] S. Johansen, "Estimation and hypothesis testing of cointegrating vectors in gaussian vector autoregressive models," *Econometrica*, vol. 59, pp. 1551–1580, 1991.

[27] P. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika*, vol. 75, pp. 335–346, 1988.

[28] R. Engle and C. Granger, "Cointegration and error correction representation: Estimation and testing," *Econometrica*, vol. 55, pp. 251–276, 1987.

[29] H. Toda and Yamamoto, "Statistical inference in vector autoregressions with possibly integrated processes," *Journal of Econometrics*, vol. 66, pp. 225–250, 1995.

[30] H. Zapata and A. Rambaldi, "Monto carlo evidence on cointegration and causation," *Oxford Bull. Econ. Stat.*, vol. 59, pp. 285–298, 1997.

[31] J. Clarke and S. Mirza, "Comparison of some common methods of detecting granger noncausality," *Journal of Statistical Computation and Simulation*, vol. 76, pp. 207–231, 2006.

[32] J. Lin, "Notes on causality," *faculty.ndhu.edu.tw/jlin/files/causality.pdfl*, 2008.

[33] U. Erol and E. Yu, "Time series analysis of the causal relationships between us energy and employment," *Resources Energy*, vol. 9, pp. 75–89, 1987.

[34] D. Jorgenson, "The role of energy in productivity growth," *Energy J.*, vol. 5, no. 3, pp. 11–26, 1984.

[35] C. Hall, C. Cleveland, and R. Kaufmann, *Energy and Resource Quality: The Ecology of the Economic Process*. New York: Wiley Interscience, 1986.

[36] J. Kraft and A. Kraft, "On the relationship between energy and gnp," *Journal of Energy Development*, vol. 3, pp. 401–403, 1978.

[37] A. Akarca and T. Long, "On the relationship between energy and gnp: a reexamination," *J. Energy Dev*, vol. 5, pp. 326–331, 1980.

[38] E. Yu and B. Hwang, "The relationship between energy and gnp: further results.," *Energy Econ.*, vol. 6, pp. 186–190, 1984.

[39] U. Soytasa and R. Sari, "Energy consumption and gdp: causality relationship in g-7 countries and emerging markets," *Energy Economics*, vol. 25, pp. 33–37, 2003.

[40] A. Akinlo, "Electricity consumption and economic growth in negeria: Evidence from cointegration and co-feature analysis.," *J. Policy Model.*, vol. 31, pp. 681–693, 2009.

[41] Y. Wolde-Rufael, "Electricity consumption and economic growth: A time series experience for 17 african countries.," *Energy Policy*, vol. 34, pp. 1106–1114, 2006.

[42] A. Akinlo, "Energy consumption and economic growth: Evidence from 11 subsahara african countries.," *Energy Economics*, vol. 30, pp. 2391–2400, 2008.

[43] C. Lee, "Energy consumption and gdp in developing countries: a cointegrated panel analysis," *Energy Economics*, vol. 27, pp. 415–27, 2005.

[44] D. Twerefo, S. Akoena, F. Egyir-Tettey, and G. Mawutor, "Energy consumption and economic growth: evidence from ghana.," *Department of Economics, University of Ghana, Ghana*, 2008.

[45] L. O. K. Fatai and F. Scrimgeour, "Modelling the causal relationship between energy consumption and gdp in new zealand, australia, india, indonesia, the philippines and thailand.," *Mathematics and Computers in Simulation*, vol. 64, pp. 431–45, 2004.

[46] K. Ghali and M. E. Sakka, "Energy use and output growth in canada: a multivariate cointegration analysis," *Energy Economics*, vol. 26, pp. 225–238, 2004.

[47] C. Ho and K.W.Siu, "A dynamic equilibrium of electricity consumption and gdp in hong kong: an empirical investigation.," *Energy Policy*, vol. 35, no. 4, pp. 2507–2513, 2007.

[48] U. Soytasa and R. Sari, "Energy consumption, economic growth, and carbon emissions: Challenges faced by an eu candidate member.," *Ecological Economics, Elsevier*, vol. 68, no. 6, pp. 1667–1675, 2009.

[49] J. Payne, "On the dynamics of energy consumption and output in the us.," *Applied Energy*, vol. 86, no. 4, pp. 575–577, 2009.

[50] A. Masih and R. Masih, "On temporal causal relationship between energy consumption, real income and prices; some new evidence from asian energy dependent nics based on a multivariate cointegration/vector error correction approach.," *Journal of Policy Modeling*, vol. 19, no. 4, pp. 417–440, 1997.

[51] F. Halicioglu, "Residential electricity demand dynamics in turkey.," *Energy Economics*, vol. 29, no. 2, pp. 199–210, 2007.

[52] C. Tang, "A re-examination of the relationship between electricity consumption and economic growth in malaysia.," *Energy Policy*, vol. 36, no. 8, pp. 3077–3085, 2008.

[53] R. Morimoto and C. Hope, "The impact of electricity supply on economic growth in sri lanka.," *Energy Economics*, vol. 26, pp. 77–85, 2004.

[54] A. Shiu and P. Lam, "Electricity consumption and economic growth in china.," *Energy Policy*, vol. 32, pp. 47–54, 2004.

[55] N. Odhiambo, "Energy consumption and economic growth nexus in tanzania: an ardl bounds testing approach.," *Energy Policy*, vol. 37, no. 2, pp. 617–622, 2009.

[56] N. Odhiambo, "Electricity consumption and economic growth in south africa: A trivariate causality test," *Energy Economics*, vol. 16, no. 31, pp. 635–640, 2009.

[57] S. Ghosh, "Electricity supply, employment and real gdp in india: evidence from cointegration and granger-causality tests.," *Energy Policy*, vol. 37, no. 8, pp. 2926–2929, 2009.

[58] S. Ghosh, "Electricity consumption and economic growth in india," *Energy Policy*, vol. 30, no. 8, pp. 125–129, 2002.

[59] P. Narayan and R. Smyth, "Electricity consumption, employment and real income in australia evidence from multivariate granger causality tests.," *Energy Policy*, vol. 1, no. 33, pp. 1109–1116, 2005.

[60] S. S. Adebola, "Electricity consumption and economic growth: Trivariate investigation in botswana with capital formation.," *International Journal of Energy Economics and Policy*, vol. 1, no. 2, pp. 32–46, 2011.

[61] E. Ziramba, "Hydroelectricity consumption and economic growth nexus: time series experience of three african countries.," *European Scientific Journal*, vol. 9, no. 1, 2013.

[62] Y. E. S. H. and J. C. Jin, "Cointegration tests of energy consumption, income, and employment," *Resources and Energy*, vol. 14, pp. 259–266, 1992.

[63] J. Pearl, "Causal inference in statistics: A gentle introduction," *TECHNICAL REPORT, Computing Science and Statistics, Proceedings of Interface '01*, vol. 33, 2001.

[64] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York: Cambridge, University Press, 2nd ed., 1992.

[65] H. Simon, "Causal ordering and identifiability," *In Studies in Econometric Method (W. C. Hood and T. Koopmans, eds.Wiley and Sons, Inc., New York, NY*, pp. 49–74, 1953.

[66] T. Koopmans, "Causal ordering and identifiability," *In Studies in Econometric Method (W. C. Hood and T. Koopmans, eds.Wiley and Sons, Inc., New York, NY*, pp. 27–48, 1953.

[67] J. Neyman, "On the application of probability theory to agricultural experiments. essay on principles.," *Statistical Science, Section 9*, vol. 5, no. 4, pp. 465–472, 1923(1990).

[68] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[69] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 1, pp. 669–710, 1995.

[70] J. Pearl, "Comment: Graphical models, causality, and intervention.," *Statistical Science*, vol. 33, no. 8, pp. 266–269, 1993.

[71] H. Wold, "Causality and ecoometrics," *Econometrika*, vol. 22, pp. 162–177, 1954.

[72] A. Moneta and P. Spirtes, "Graphical models for the identification of causal structures in multivariate time series models," *Joint Conference on Information Sciences*, 2006.

[73] R. Dahlhaus and M. Eichler, "Causality and graphical models in time series analysis," *In P. Green, N. Hjort and S. Richardson (eds), Highly structured stochastic systems, Oxford University Press*, 2003.

[74] M. Eichler and V. Didelez, "Causal reasoning with graphical time series models," *UAI*, 2007.

[75] H. White, "Some asymptotic results for learning in single hidden layer feedforward network models," *Journal of the American Statistical Association*, vol. 84, pp. 1003–1013, 1989.

[76] B. Ripley, "Statistical aspects of neural networks," *In: Barndorff-Nielsen, O.E., Jensen, J.L., Kendall, W.S. Networks and Chaos-Statistical and Probabilistic Aspects*, vol. 2, pp. 491–494, 1993.

[77] B. Cheng and D. Titterington, "Neural networks: A review from a statistical perspective," *Statistical Science*, vol. 9, no. 1, pp. 2–54, 1994.

[78] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1996.

[79] T. Tersvirta, C.-F. Lin, and C. W. J. Granger, "Power of the neural network linearity test," *Journal of Time Series Analysis*, vol. 14, pp. 309–323, 1993.

[80] Z. Tang, C. Almeida, and P. Fishwick, "Time series forecasting using neural networks vs box-jenkins methodology," *Simulation*, vol. 57, no. 5, pp. 303–310, 1991.

[81] R. Sharda and R. Patil, "Connectionist approach to time series prediction: An empirical test," *Journal of Intelligent and Manufacturing*, vol. 3, pp. 317–323, 1992.

[82] Z. Tang and P. Fishwick, "Feedforward neural nets as models for time series forecasting," *ORSA Journal on Computing*, vol. 5, no. 4, pp. 374–385, 1993.

[83] A. Weigend and N. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past," *Addison-Wesley, Reading, MA*, 1993.

[84] G. Zhang, B. Patuwo, and M. Hu, "A simulation study of artificial neural networks for nonlinear time-series forecasting,"

[85] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[86] B. Chen, M. Chang, and C. Lin, "Load forecasting using support vector machines: A study on eunite competition 2001.," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 1821–1830, 2004.

[87] E. Ceperic, V. Ceperic, and A. Baric, "A strategy for short-term load forecasting by support vector regression machines," *IEEE Transactions On Power Systems*, 2013.

[88] W. Hong, "Electric load forecasting by support vector model," *Applied Mathematical Modelling*, vol. 33, no. 5, pp. 2444–2454, 2009.

[89] S. R. Gunn, "Support vetor machines for classication and regression," *Technical Report, University of Suthoampton, School of Electronics and Computer Science*, 1998.

[90] M. Figueiredo, R. Ballini, S. Soares, M. Andrade, and F. Gomide, "Learning algorithms for a class of neurofuzzy network and application," *IEEE Trans. Syst., Man, Cyber. C, Appl. Rev.*, vol. 34, no. 4, pp. 381–396, 2004.

[91] C. J. Lin, C. H. Chen, and C. T. Lin, "A hybrid of cooperative particle swarm optimization and cultural algorithm for neural fuzzy networks and its prediction applications," *IEEE Trans. Syst., Man, Cyber. C, Appl. Rev.*, vol. 39, no. 1, pp. 55–68, 2009.

[92] S. P. Y. Bodyanskiy and T. Rybalchenko, "Multilayer neuro-fuzzy network for short term electric load forecasting," *Lect. Notes Comput. Sci.*, vol. 5010, pp. 339–348, 2008.

[93] V. Cherkassky, "Fuzzy inference systems: A critical review," *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications NATO ASI Series*, vol. 162, pp. 177–197, 1998.

[94] M. Minsky, "Steps toward artificial intelligence," *Proceedings of the IRE*, pp. 8–30, 1961.

[95] N. Chomsky, "On where artificial intelligence went wrong," *The Atlantic, http://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/?single$_page = true$*, 2012.

[96] S. Makridakis and M. Hibbon, "Accuracy of forecasting: An emperical investigation," *J. Roy Statist. Soc.*, 1979.

[97] L. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting*, vol. 7, no. 4, pp. 437–450, 2000.

[98] R. Fildes and S. Makridakis, "The impact of emperical accuracy studies on time series analysis forecasting," *International Statistical Review*, 1995.

[99] H. Kim and K. Shin, "A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets.," *Applied Soft Computing*, vol. 7, pp. 569–576, 2007.

[100] R. Carver, "The case against statistical significance testing," *Harvard educational review*, vol. 48, no. 3, 1978.

[101] R. Inglesi-Lotz and A. Pouris, "The influence of scientific research output of academics on economic growth in south africa: an autoregressive distributed lag (adrl) application," *Scientometrics*, vol. 95, no. 1, pp. 129–139, 2013.

[102] Eskom, "Eskom annual report," *www.eskom.co.za*, 2011.

[103] H. Alfares and M. Nazeeruddin, "Electric load forecasting: literature survey and classification of methods," *International Journal of Systems Science*, vol. 33, no. 1, pp. 23–34, 2002.

[104] G. A. Mbamalu and M. E. El-Hawary, "Load forecasting via suboptimal autoregressive models and iteratively recursive least squares estimation.," *IEEE Transactions on Power Systems*, vol. 33, no. 8, pp. 343–348, 1993.

[105] I. Mogghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques.," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 1484–1491, 1989.

[106] E. H. Barakat, M. A. Qayyum, M. N. Hamed, and S. A. Al-Rashed, "Short-term peak demand forecasting in fast developing utility with inherent dynamic load characteristics.," *IEEE Transactions on Power Systems*, vol. 5, pp. 813–824, 1990.

[107] A. D. Papalexopulos and T. C. Hesterberg, "A regression based approach to short-term load forecasting.," *IEEE Transactions on Power Systems*, vol. 5, pp. 1214–1221, 1990.

[108] W. R. Christiaanse, "Short-term load forecasting using general exponential smoothing," *IEEE Transactions on Power Apparatus and Systems, PAS-90*, pp. 900– 902, 1971.

[109] K. Krycha, "A comparison of standard methods with a neural network approach to forecast a univariate time series.," *Research memorandum no. 310, Institute of Advanced Studies, Vienna, https://www.ihs.ac.at/publications/ihsfo/fo310.pdf*, 1992.

[110] X. Zhou, R. Molina, F. Zhou, and A. Katsaggelos, "Fast iteratively reweighted least squares for lp regularized image deconvolution and reconstruction.," *www.decsai.ugr.es/vip/files/conferences/1569899677.pdf*, 2013.

[111] P. Chemouil and B. Garnier, "An adaptive short-term traffic forecasting.," *Elsevier Science Publishers B. V. (North-Holland)*, 1985.

[112] P. Crochet, "Adaptive kalman filtering of 2-metre temperature and 10-metre wind-speed forecasts in iceland, meteorol.," *A Appl.*, vol. 11, pp. 173–187, 2004.

[113] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis*. New York: John Wiley and Sons, 4th edition ed., 2008.

[114] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[115] L. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, vol. SMC-3, no. 1, pp. 28–44, 1973.

[116] L. Zadeh, "Fuzzy algorithms," *Information and Control*, vol. 12, pp. 94–102, 1968.

[117] G. Tso and K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks.," *Energy*, vol. 32, pp. 1761–1768, 2007.

[118] K. Ho, Y. Hsu, F. Chen, T. Lee, C. Liang, T. Lai, and K. Chen, "Short-term load forecasting of taiwan power system using a knowledge based expert system," *IEEE Transactions on Power Systems*, vol. 5, pp. 1214–1221, 1990.

[119] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators.," *Neural Networks*, no. 2, pp. 359–366, 1989.

[120] D. Tikk, L. T. Kóczy, and T. D. Gedeon, "A survey of universal approximation and its applications in soft computing," *Research Working Paper RWP-IT-2001. School of Information Technology, Murdoch University, Perth*, 2001.

[121] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer, second ed., 1999.

[122] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, vol. 1, pp. 23–34, 1992.

[123] O. L. Mangasarian, *Nonlinear Programming.* New York: McGraw-Hill, 1969.

[124] R. J. Vanderbei, "Loqo users manualversion 3.10," *Statistics and Operations Research, Technical Report SOR-97-08*, 1997.

[125] W. Karush, "Minima of functions of several variables with inequalities as side constraints," *Masters thesis,Dept. of Mathematics, University of Chicago*, 1939.

[126] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," *2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pp. 481–492, 1951.

[127] B. Scholkopf, C. Burges, and A. Smola, "Advances in kernel methods-support vector learning," *MIT Press, Cambridge, MA*, 1999.

[128] R. Babuska and H. Verbruggen, "Neuro-fuzzy methods for nonlinear sytem identification," *Annual Reviews in Control*, vol. 27, pp. 73–85, 2003.

[129] C. Harris, C. Moore, and M. Brown, *Intelligent control: Aspects of Fuzzy Logic and Neural Nets*. Singapore: World Scientific Publishing, first ed., 1993.

[130] J. Jang, C. Sun, and E. Mizutani, *Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence*. New Jersey: Prentice Hall, first ed., 1997.

[131] J. Jang, C. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence," *Upper Saddle River, NJ: Prentice-Hall.*, 1997.

[132] M. Sentes, R. Babuska, U. Kaymak, and H. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, vol. 28, no. 3, pp. 376–386, 1998.

[133] R. Babuska, *Fuzzy modeling and Identification*. PhD thesis, Technical University of Delft, Delft, Holland, 1991.

[134] R. Lewis, "Practical digital image processing," *Ellis Horwood Series in Digital and Signal Processing, Ellis Horwood Ltd, New York, etc.*, 1990.

[135] J. Bezdek and S. K. Pal, *Fuzzy models for Pattern Recognition*. New York.: IEEE Press, 1992.

[136] J. Grandell, "Time series analysis," *http://www.math.kth.se/matstat/gru/sf2943/ts.pdf*.

[137] S. S. Africa, "P414, electricity generated and available for distribution," *www.statssa.gov.za*, 2013.

[138] I. Kaastra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, pp. 215–236, 1996.

[139] T. Kolarik and G. Rudorfer, "Time series forecasting using neural networks, department of applied computer science," *Vienna University of Economics and Business Administration*, no. 1090, pp. 2–6, 1997.

[140] I. Nabney, *Netlab:Algorithms for Pattern Recognition*. Berlin: Springer, 2002.

[141] "http://www.ulb.ac.be/di/map/gbonte/software/local/fis.html,"

[142] "http://www.isis.ecs.soton.ac.uk/isystems/kernel/,"

[143] G. Yule, "Why do we sometimes get nonsense correlations between time series? a study of sampling and the nature of time series (with discussion)," *Journal of the Royal Statistical Society*, vol. 89, pp. 1–64, 1926.

[144] E. Slutsky, "The summation of random causes as the source of cyclic processes," *Econometrica*, vol. 5, pp. 105–146, 1931.

[145] H. Wold, "A study in the analysis of stationary time series," *Almgrist Wiksell,Stockholm*, 1938.

[146] D. Hume, *A Treatise of Human Nature*. Oxford: Oxford University Press, 1740,(1967).

[147] T. Marwala., *Artificial Intelligence Techniques for Rational Decision Making"*. Switzerland: Springer, 2014.

[148] C. Selltiz, L. Wrightsman, and S. Cook, *Research Methods in Social Relations*. New York: Holt, Rinehart and Winston, 1959.

[149] H. kim and K. Shin, "A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets," *Applied Soft Computing*, vol. 7, pp. 569–576, 2007.

[150] J. W. Taylor and R. Buizza, "Using weather ensemble predictions in electricity demand forecasting," *International Journal of Forecasting*, vol. 19, pp. 57–70, 2003.

[151] A. Glynn and K. Quinn, "Structural causal models and the specification of time-series-cross-section models.," *http://scholar.harvard.edu/files/aglynn/files/tscs-scm.pdf*, vol. 37, no. 3, pp. 424–438, 2013.

[152] J. Sekhon, "The neyman-rubin model of causal inference and estimation via matching methods," *The Oxford Handbook of Political Methodology, Janet Box-Steffensmeier, Henry Brady, and David Collier, eds. 2007. http://sekhon.berkeley.edu/papers/SekhonOxfordHandbook.pdf*, 2007.

[153] D. Rubin, "Matched sampling for causal effects.," *Cambridge, England: Cambridge, University Press*, 2006.

[154] W. Cochran, "Matching in analytical studies.," *American Journal of Public Health*, vol. 43, no. 5, pp. 684–691, 1953.

[155] W. Cochran, "The planning of observational studies of human populations (with discussion).," *Journal of the Royal Statistical Society, Series A*, vol. 128, pp. 234–255, 1965.

[156] B. D. Rubin, "Bayesian inference for causal effects: The role of randomization.," *The Annals of Statistics*, vol. 6, no. 1, pp. 34–58, 1978.

[157] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects.," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[158] P. Holland, "Statistics and causal inference.," *Journal of the American Statistical Association*, vol. 81, no. 1, pp. 945–960, 1986.

[159] P. Judea, "Causal diagrams for empirical research," *Biometrika*, vol. 33, no. 82, pp. 669–710, 1995.

[160] P. Debba, R. Koen, J. Holloway, T. magadla, M. Rasuba, S. Khuluse, and C. Elphinstone, "Forecasts for electricty demand in south africa (2010-2035) using the csir sectoral regression model," 2010.

[161] T. Bell and N. Madula, "Where has all the growth gone? south african manufacturing industry 1970-2000," *TIPS, 2001 Annual Forum, Misty Hills, Muldersdrift*, vol. 33, 2001.

[162] J. Blignaut and T. de Wet, "Some recommendations towards reducing electricity consumption in the south african manufacturing sector," *South African Journal of Economic and Management Sciences*, vol. 4, no. 2, pp. 359–379, 2001.

[163] R. Inglesi-Lotz, "The sensitivity of the south african industrial sector's electricity consumption to electricity price fluctuations," *University of Pretoria Working Paper*, no. 25, 2012.

[164] R. Inglesi-Lotz and J. Blignaut, "Estimating the price elasticity of demand for electricty by sector south africa," *South African Journal of Economic and Management Sciences*, vol. 14, pp. 449–465, 2011.

[165] R. Inglesi-Lotz and J. Blignaut, "South africa's electricty consumption:a sectoral decomposition analysis," *Applied Energy*, vol. 88, pp. 4779–4784, 2011.

[166] F. Hsiao and M. W. Hsiao, "Fdi, exports, and growth in east and southeast asia–evidence from time-series and panel data causality analyses," *2006 International Conference on Korea and the World Economy V*, 2006.

[167] M. Holden and A. Gouws, "Determinants of exports in south africa," *Trade and Industrial Policy Secretariat Annual Forum*, 1994.

[168] N. Shombe, "Causality relationships between total exports with agricultural and manufacturing gdp in tanzania," *Institute of Developing Economies Discussion paper No. 136*, no. 136, 2003.

[169] A. A. Alici and M. S. Ucal, "Foreign direct investment, exports and output growth of turkey: causality analysis," *European Trade Study Group (ETSG) Fifth Annual Conference, Madrid*, 2003.

[170] P. Janisch, "Gold in south africa," *Journal of the South African Institute of Mining and Metallurgy*, 1985.

[171] Deloitte, "The economic impact of electricty price increases on various sectors of the south african economy," 2013.

[172] A. Tan, T. Zhang, and S. Wu, "Pressure and density of air in mines," *Indian Journal of Radio and Space Physics*, vol. 37, pp. 64–67, 2008.

[173] S. S. Africa, "Mineral accounts south africa: 1980-2009," *www.statssa.gov.za*, vol. D0405.2, 2012.

[174] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice-Hall, 2nd ed., 1998.

[175] Q.-Y. Z. Guang-Bin Huang and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks, international joint conference on neural networks," *Journal of Educational Psychology*, vol. 2, no. 1, pp. 985–990, 2004.

[176] G. Huang, "Learning capability and storage capacity of two hidden-layer feed-forward networks," *IEEE Trans. Neural Networks*, vol. 14, no. 2, pp. 274–281, 2003.

[177] G. Huang and H. Babri, "Upper bounds on the number of hidden neurons in feed-forward networks with arbitrary bounded nonlinear activation functions," *IEEE Trans. Neural Networks*, vol. 9, no. 1, pp. 224–229, 1998.

[178] D. Serre, *Matrices: Theory and Applications.* New York: Springer, 2002.

[179] C. Rao and S. Mitra, *Generalized Inverse of Matrices and its Applications*. New York: Wiley, 1971.

[180] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "Op-elm: Optimally pruned extreme learning machine," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II: EXPRESS BRIEFS*, vol. 21, no. 1, pp. 18–20, 2010.

[181] Y. Miche, A. Sorjamaa, and A. Lendasse, "Op-elm: Theory, experiments and a toolbox," *Lecture Notes in Computer Science*, vol. 5163, pp. 145–154, 2008.

[182] T. Simila and J. Tikka, "Multiresponse sparse regression with application to multidimensional scaling.," *in Proc. Int. Conf. Artif. Neural Netw.,*, vol. 3697, no. 1, pp. 97–102, 2005.

[183] R. Myers, *Classical and Modern Regression with Applications*. Duxbury: Pacific Grove, 2nd ed. ed., 1990.

[184] G. Bontempi, M. Birattari, and H. Bersini, "Recursive lazy learning for modeling and control," *in Proc. Eur. Conf. Mach. Learn.,*, vol. 3697, no. 1, pp. 292–303, 1998.

[185] F. Mateo, J. J. Carrasco, M. Millan-Giraldo, A. Sellami, P. Escandell-Montero, J. M. Martinez-Martinez, and E. Soria-Olivas, "Machine learning techniques for short-term electric power demand prediction," *SANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium)*, 2013.

[186] R. Zhang, Z. Y. Dong, Y. Xu, K. Meng, and K. P. Wong, "Short-term load forecasting of australian national electricity market by an ensemble model of extreme learning machine," *ISSN : 1751-8687, DOI: 10.1049/iet-gtd.2012.0541*, pp. 391 – 397, 2013.

[187] N. A. Shrivastava and B. K. Panigrahi, "A hybrid wavelet-elm based short term price forecasting for electricity markets," *Electrical Power and Energy Systems*, vol. 55, pp. 41–50, 2014.