ELECTRICITY DEMAND PREDICTION USING A HYBRID APPROCH

SYAMND MIRZA ABDULLAH

FACULTY OF ECONOMICS AND ADMINISTRATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

University of Malaya

ELECTRICITY DEMAND PREDICTION USING A HYBIRD APPROACH

SYAMND MIRZA ABDULLAH

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF ECONOMICS AND ADMINISTRATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

University of Malaya

*(Please delete this part) Depending on the medium of thesis/dissertation, pick either the English or Bahasa Malaysia version and delete the other.

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **SYAMND MIRZA ABDULLAH**

Matric No: **EHA120003**

Name of Degree: **DOCTOR OF PHILOSOPHY (PHD)**

Title of Thesis ("this Work"): **ELECTRICITY DEMAND PREDICTION USING HYBRID APPROACH**

Field of Study: APPLIED STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: **SYAMND MIRZA ABDULLAH**

No. Matrik: **EHA120003**

Nama Ijazah: **DOCTOR OF PHILOSOPHY (PHD)**

Tajuk Tesis ("Hasil Kerja ini"): **ELECTRICITY DEMAND PREDICTION USING HYBRID APPROACH**

Bidang Penyelidikan:

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon Tarikh:

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh:

Nama:

Jawatan:

University of Malaya

ABSTRACT

Electricity demand prediction is an important field of study that supports the government in developing a good economic and control plan for the future of electricity power generation. Various techniques and tools have been utilized throughout the history of such predictions, and different parameters have been analyzed. The main aims of studies in this field were to predict electricity demand and to minimize errors by analyzing various effects, such as that of the relation between the patterns of the data set and the utilized tools.

In particular, this study focuses on reducing the degree of multicollinearity among independent variables to increase accuracy rate. In addition, the study aims to employ a combination system that accepts both linear and nonlinear patterns of the input data set to minimize the residual errors in prediction rate. To realize this objective, this thesis proposes a system that uses a hybrid approach that combines principal component analysis as a tool for lowering degree of multicollinearity, multiple linear regression (MLR) and a time series artificial neural network (ANN) to minimize errors. The novel electricity demand prediction model proposed in this thesis is called the principal component regression with back-propagation artificial neural networks model (PCR-BPNN). The data set fed into this model is the quarterly electricity usage in Malaysia from 1995 to 2013 provided by the Department of Statistics Malaysia.

According to the performance indicators such as mean squared error, root mean squared error, and mean absolute percentage error, the PCR-BPNN model generates a more accurate predictions than previous methods such as principal component (PC)—MLR, PCNN, and PC-Support Vector regression models. The results indicate the expected electricity demand in Malaysia for 2020 will be 13702.91 Ktoe.

University of Malaya

ABSTRAK

Ramalan terhadap permintaan elektrik adalah satu bidang kajian yang penting di dalam menyokong kerajaan untuk membangunkan satu plan ekonomi serta kawalan penjanaan tenaga elektrik yang baik untuk masa depan. Pelbagai teknik dan alat telah digunakan di dalam kajian ramalan terdahulu, dan analisis juga telah dibuat ke atas parameter-paremeter yang berbeza. Tujuan utama kajian ini adalah untuk membuat ramalan permintaan tenaga eletrik dan meminimumkan ralatdengan menganalisis pelbagai kesan, seperti kesan hubungan di antara corak set data input dan kesan kaedah yang digunakan. Tesis ini membincangkan multikolinearan di antara pembolehubah bebas dan kelinearan serta ketidaklinearan data input iaitu dengan merujuk kepada ketepatan model ramalan permintaan elektrik. Secara khususnya, kajian ini telah memberi tumpuan di dalam mengurangkan tahap multikolinearan di antara pembolehubah bebas untuk meningkatkan kadar ketepatan tersebut. Di samping itu, kajian ini juga bertujuan untuk mengkaji sistem gabungan yang menerima kedua-dua corak set data input (kelinearan dan ketaklelurusan/ketidaklinearan) untuk mengurangkan ralat sisa di dalam ramalan ini. Untuk merealisasikan matlamat ini, tesis ini telah mencadangkan satu sistem yang menggunakan pendekatan hibrid dengan menggabungkan analisis komponen utama sebagai alat untuk mengurangkan tahap multikolinearan, regresi linear berganda (RLB) dan rangkaian neural tiruan (RNT) bagi siri masa. Oleh itu, satu model ramalan permintaan elektrik yang novel dicadangkan di dalam tesis ini dan dikenali sebagai model regresi komponen utama dengan rangkaian neural pembiakan kembali PCR-BPNN. Kajian ini telah menggunakan data suku tahunan dari tahun 1995 hingga tahun 2013. Set data input yang digunakan di dalam model ini adalah di dalam konteks Malaysia dan telah disahkan oleh Jabatan Perangkaan Malaysia. Menurut petunjuk prestasi seperti min ralat kuasa, asas min ralat kuasa, dan min ralat peratusan mutlak, model PCR-BPNN ini dapat menyumbang kadar ketepatan yang lebih tepat jika dibandingkan dengan komponen

University of Malaya

utama (PC)-MLR, PCNN, dan ‘PC-support vector regression models’. Menurut ramalan yang diperolehi dari kaedah ini, permintaan elektrik di Malaysia bagi tahun 2020 adalah 13.702,91 Ktoe.

.

University of Malaya

ACKNOWLEDGEMENTS

I would like to thank Allah and his mercy, for enabling me to go through my doctoral studies and finishing this thesis writing. Secondly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Noor Azina Ismail for her continuous support to this thesis, her patience in teaching me the research works, her motivation when I was down, her immense knowledge that makes me more knowledgeable than before, her helpful guidance at the time of doing research and at the time of writing this thesis. I have never expected a better advisor and mentor for my Ph.D study than her.

Besides my advisor, I would like to thank all friends in the Faculty of Economics and Administration, especially, Adeel Ahmed and Ayshah Shoukat for their motivation and discussions, for their sleepless nights working together, and for all the knowledge they shared with me in the last four years. I would like to appreciate the role of Koya Technical Institute and Ministry of Higher Education in Kurdistan that funded and encouraged me to complete this study.

Finally, I would like to thank all of my family members especially my beloved wife Wanawsha Ismail Tahir. Thanks to my kids (Ram and Honey) for the time I spent to work instead of being a father. Thanks for the mercy and endless supports of my parents. Thanks to my brothers and sisters for their physical and mental supports and encouragement since I started this study until the end. Thanks to all who contributed in achieving this work even how small they are.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

University of Malaya

LIST OF TABLES

University of Malaya

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

ANFIS :	Adaptive Neuro Fuzzy Inference System
AGDP :	Agricultural Gross Domestic Product
ANN :	Artificial Neural Network
ANOVA :	Analysis Of Variance
ARDL :	Auto Regressive Distributed Lag
ARIMA :	Autoregressive Integrated Moving Average
ARMA :	Autoregressive Moving Average
BPNN :	Back-Propagation Neural Network
E.C :	Electricity Consumption
E.D :	Electricity Demand
ECM :	Error Correction Model
ES :	Exponential Smoothing
FLR :	Fuzzy Linear Regression
FN :	Fuzzy Neural
FR :	Fuzzy Regression
FWTNN :	Fuzzy Wavelet Transform Neural Network
GA :	Genetic Algorithm
GDP :	Gross Domestic Product
GNP :	Gross National Product
GP :	Grey Prediction
GPRM :	Grey Prediction with Rolling Mechanism
MAE :	Mean Absolute Error
MAPE :	Mean Absolute Percentage Error
MBE :	Mean Bias Error

University of Malaya

MLR : Multiple Linear Regression

MSE : Mean Square Error

NGDP : Non-agricultural Gross Domestic Product

PA : Prediction Accuracy

PCA : Principal Component Analysis

PCR : Principal Component Regression

PCR-BPNN : Principal Component Regression with Back-Propagation Neural Networks

PEDM : Prediction Electricity Demand Model

PMES : Peninsular Malaysia Electricity Supply

PPEDM : Preliminary Prediction Electricity Demand Model

RA : Auto- Regressive

RMSE : Root Mean Square Error

SOM : Self-Organized Map

SVM : Support Vector Machine

SVR : Support Vector Regression

VAR : Vector Autoregressive

VECM : Vector Error Correction Model

WFNN : Wavelet Fuzzy Neural Network

WNN : Wavelet Neural Network

University of Malaya

LIST OF APPENDICES

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Background of Study

Electrical energy is vital to any developing country. Demands for such energy have been increasing rapidly in most countries because of economic and population growth. The demand rate of electricity energy for an area is the rate that is required by the consumers in that area. The prediction of the electricity consumption rate of an area could be obtained by making a prediction of this consumption in advance so as to be able to prepare for the rise in consumption. Such a prediction could be obtained or estimated by analyzing the historical records of electricity consumption rates of a particular area (Yoo et al.,2007). In the aim to prepare a country for making plans that can help it to foresee future problems related with electricity demand, demand prediction is vital. The information acquired from the demand prediction model can be used for building a cost effective risk management plan for any kind of electric utility, specifically, for long-term forecasting issues which are related to the planning of power generation, operations and real time. Therefore, to understand any increase or decrease in electricity consumption for future needs, a good model with high prediction accuracy is imperative. The rational for developing such a model lies in the fact that the wrong estimation of electricity demand rates could harm the economy negatively. For instance overestimation can lead to unnecessary idle capacity i.e. wastage of financial resources while underestimation can lead to potential outages which could be devastating for the economy (Kavaklioglu et al.,2011). In that regard, a good model that can predict electricity demand rate with accuracy is in place.

According to the latest census, there are a total of 30.5 million people in Malaysia in the year 2015. The Gross Domestic Product (GDP) grew at an average rate of 4.77% during the past 15 years. Figure1.1 illustrates a time based seasonal data of the long term electricity consumption pattern of Malaysia from 1995 to 2013, where the electricity

University of Malaya

consumption rate of the population had increased rapidly because of its economic and population growth. It was noted that among the industries of Malaysia, the main consumers of electricity were construction commercial industries and residential sectors.

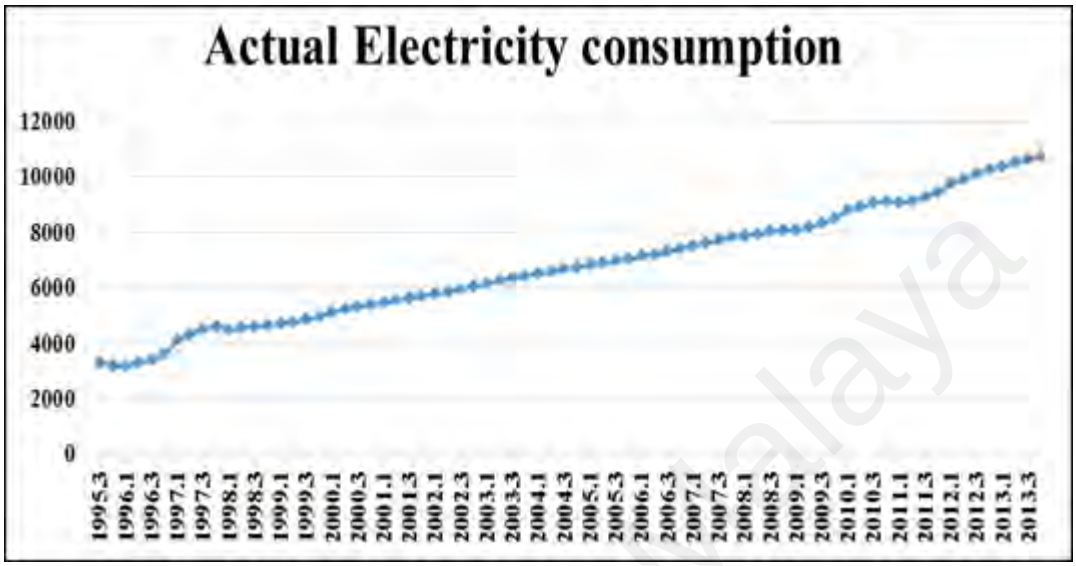


Figure 1.1: Actual electricity consumption in Malaysia

Based on the statistics shown in Figure 1.1, it appears that the electricity demand rate will continue to increase as Malaysia is fast approaching to become an industrialized country by 2020. This can be traced to some historcial facts. In 2014, Peninsular Malaysia Electricity Supply (PMES) provided the prediction (see Table1-1) of the electricity demand as well as electricity generation from 2014 to 2020 (PMES, 2014). The company depended on some predictor factors for this prediciton which were based on: GDP growth, Price of electricity, Population, Energy generation, Number of consumers, and Peak demands. The company proposed a long term load prediciton in its forecast by implementing industry-wide practice techniques. The company analyzed the historical data by using a top-down and bottom-up appraoch. The method are highly tied with time series analysis. Although the prediction of electricity generation by PMES was able to indicate future consumption demands, it was also affected by various problems. For instance, the depleting rate of the source of electricity generation had caused PMES to predict less accurately, thereby affecting the sustainability of the power sector in the

University of Malaya

country. This shows that the prediction made by PMES was not accurate (Mahlia et al., 2011).

Table 1.1: Prediction for electricity demand and generation (PMES)

Years	Generation / Ktoe	Demand / Ktoe	Shortage
2014	12440	11087.73	1352.27
2015	12800	11486.89	1313.11
2016	13236	11854.47	1381.53
2017	13672	12245.67	1426.33
2018	14124	12649.77	1474.23
2019	14590	13067.22	1522.78
2020	15027	13485.37	1541.63

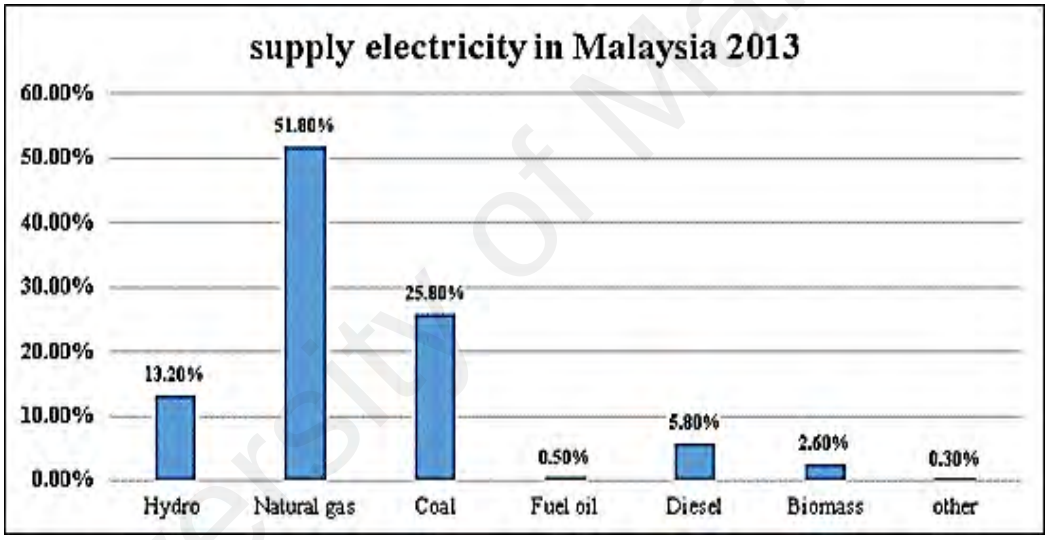


Figure 1.2: Sources of electricity supply in Malaysia, 2013

Figure 1.2, highlights the sources of electricity supply in Malaysia. As can be seen, the highest source of electricity supply came from natural gas, followed by coal and hydro. It is clear from the figure that most electricity power in Malaysia (around 76%) came from unrenewable energy which costs Malaysia billions of Ringgit. Even the 13% that comes from the hydro power systems costs Malaysia billion litters of rain water (clean for use), not counting the costs involved with the operation and maintenance of turbines. In both the two cases of unrenewable energy and hydro energy, it would appear

University of Malaya

that a wrong prediction and estimation of electricity demand could have financial and environmental implications on Malaysia.

Given that Malaysia is among the fastest growing country in Asia, more energy suppliers are needed to fuel its rapid pace of economic expansion. Therefore, it is crucial for Malaysia to have a reliable supply of electricity for meeting the social and development objectives of the nation. It is equally important to ensure that over supply would not happen as the generation of electricity may have an adverse environmental impact on the country. Thus, it is crucial that a good and accurate model(s) for predicting electricity demand be developed for use.

Researchers have continuously tested the accuracy of electricity prediction models by using different tools and techniques and this has in turn, resulted in the proposal of various tools for designing and building electricity demand prediction models (Akay et al.,2007); (Chen et al.,2007) and (Zhang et al., 2012)

There are two main steps in predicting electricity consumption models. The first step involves identifying the factors and parameters that are related to electricity consumption. These parameters which are known as the independent input variables are utilized differently from one study to another, given that the areas covered by the studies are also different. Therefore, the factors may be positively influenced by electricity demand in some areas whereas other factors may have a positive influence on electricity consumption in other regions (Zhang et al.,2012).

The second step involves searching for a suitable algorithm that can accurately compute and predict electricity consumption rates (Xin et al.,2010). Researchers have employed different tools and techniques to accomplish prediction and estimation processes (Aranda et al., 2012); (Bazmia et al.,2012) and (Kuo et al.,2012). They clamped

University of Malaya

different parameters into their proposed models and they used different parameters in their study of electricity demand. This is because both the types and numbers of such parameters can be influenced by the geographical area under study of electricity demand. This is because both the types and numbers of such parameters can be influenced by the geographical area under study (Abiyev et al.,2009).

Many studies involve factors that are related to weather as most countries throughout the world are influenced by different seasons, and inevitably, electricity demands definitely depend on weather changes. Some countries with tropical climates have temperatures that do not change drastically but remain constant throughout the year. In this regard, the weather in those countries is not as influential on electricity demand as it is elsewhere. With regard to the type of data, researchers have encountered different challenges such as finding algorithms that can address a mix of linear and nonlinear data. Some researchers Wang et al.,(2009) have recently proposed the hybrid system which considers both types of data.

The sought after statistical technique is deemed to be able to reduce errors that occur because of the change in the patterns of the data. Nonetheless, it faces a challenge which also deals with characteristics related to accuracy such as complexity and multicollinearity, both of which have not been actively discussed in previous studies (Pao,2006) ; (Dalvand et al.,2008) and (Kavaklioglu et al.,2009).

The complexity of any dataset consists of the change in the patterns of the data within a variable (Chia et al.,2011). Such complexity affects the type of tools that must be chosen when designing a prediction model. In pursuit of this goal, tools and techniques are classified into three groups for this study: linear, nonlinear, and hybrid systems. Depending on the complexity of the data, a method can be selected from any of these groups. A method that is grouped as linear or nonlinear can be employed for data that

University of Malaya

have only one pattern (either linearly or non-linearly) (Zhang et al.,2010). However, if the data change both linearly and nonlinearly, over a specific range, the hybrid system methodology becomes more suitable (Kavaklioglu,2011). The choice of prediction method clearly depends on the patterns of the data. Employing an inappropriate method for a dataset (e.g., linear methods for nonlinear datasets) can negatively affect the accuracy rate of the prediction model, hence, to reduce such negative effects on accuracy, researchers should choose a hybrid-based prediction model which can address data that have mixed patterns (linear and nonlinear patterns).

The hybrid prediction system was introduced by Bates et al.(1969) as an alternative to the individual methods. The idea of the hybrid system is to combine two or more individual prediction methods as one where each method has different features which can then be used to accurately predict those data that have different patterns or characteristics. Combining two different methods to form one prediction model can result in a better accuracy rate. It is certainly better than the individual prediction methods (Bates et al.,1969 ; Zheng et al.,2011). Nonetheless, minimizing and reducing the complexity and collinearity of the data can be another way to minimize errors, thereby improving the accuracy rate. Based on the intention mentioned earlier, the major part of this study aims to develop a model that is able to reduce the complexity and collinearity of the independent variables so that the output errors can be minimized, thereby, improving the accuracy rate of the prediction model.

1.2 Problem Statement

There are several studies (Wolde-Rufael, 2006) which investigate and analyze electricity demand via the utilization of many methods. In all of these studies, the accuracy rate of the prediction model is an important factor for explaining certain

University of Malaya

statements about the future rate of electricity demand. These statements are useful for decision-making in the energy sectors besides being necessary for managing the electricity power supply. In this regard, the main problem being addressed is improving the accuracy rate of the electricity demand prediction model. This study also addresses some sub-problems (as mentioned below) that are relevant to the accuracy rate of the electricity demand prediction model:

1. The inclusion of relevant factors which are related to electricity demands and taking into account the strong correlations between these factors.
2. Taking into account the complexity of the data with linear and non-linear patterns and to consider the effect of residual errors on the accuracy rate of prediction models, which has not been considered in linear, non-linear or hybrid based models.

1.3 Research Question

A major concern of this study aims to answer how the complexity of input data set can be reduced and how multicollinearity among independent variables can be removed so as to reduce errors and improve accuracy. Thus, the research questions formulated are:

1. Which characteristics of the input dataset affect the accuracy rate of the electricity demand prediction model?
2. Which statistical method can reduce the complexity of the input dataset?
3. How to develop a new electricity demand prediction model that takes into account the different patterns or characteristics of the data?
4. How can a hybrid approach help to increase accuracy of prediction in electricity demand?

University of Malaya

1.4 Research Objectives

The main objective of this study is to propose a new approach that can improve the accuracy rate of the electricity demand prediction model. To achieve the main objective, this study focuses on the following sub-objectives:

1. To investigate the relationship between different input dataset patterns and electricity demand.
2. To reduce dataset complexity and then improve accuracy of electricity demand prediction.
3. To assess the accuracy of the developed prediction model.

1.5 Significance of the Study

The prediction and modeling of electricity consumption plays a vital role in developing and developed countries. It is heavily linked to the accuracy of prediction rate. It also plays an important role for the related organizations and policy makers where underestimation of the consumption would lead to potential outages and overestimation would lead to unnecessary idle capacity i.e. wastage of financial resources. Decision makers in all countries are focusing on the accuracy of the demand predicting rate for many reasons. For example, in Malaysia, this accuracy rate of electricity consumption is considered for two reasons. The first reason, according to Ismaila,(2011) is that any single percentage of error rate may lead to losses or gains of millions of ringgit. The second reason is based on Razak et al.,(2009), where the prediction of electricity demand helps the relevant ministry to control the electricity consumption rate.

University of Malaya

1.6 Scope of the Study

Figure 1.3 explains the scope of this research. It also shows the important parts of the prediction of the electricity demand model. Details on the study scope are mentioned below:

1. A prediction model is a statement about how the events will occur in the future and it can simulate activity. The output of any prediction model is a quantitative estimation which can be used for any plans for possible development. This explains why the prediction model's applications can be found in many fields of study such as Energy demand, Medical application, Engineer's application, and Economic growth.
2. As mentioned above, one application that is more important in the prediction model is the energy demand. The energy models are developed to help a country sustain its economic progress. In this regard, the demand for an energy prediction model is vast as such energy models can be divided into energy demands of water, oil, gas, and electricity.
3. With regards to the energy demand for the prediction model envisaged, this study only focused on the electricity demand prediction model because the energy of electricity is the most significant driving force for economic growth. Therefore, the planning of electricity demand is one key success factor of development in any country. This key can only be achieved if the demand is accurately predicted by the right model.
4. For the input data characteristics, this study focuses more on multicollinearity because this property can negatively affect the accuracy of the prediction model.

University of Malaya

5. This study uses the Malaysia data set, as it is a developing country. The dataset contains nineteen independent variables that are related to electricity demand, which mentioned in chapter 3 from section 3.3.1.

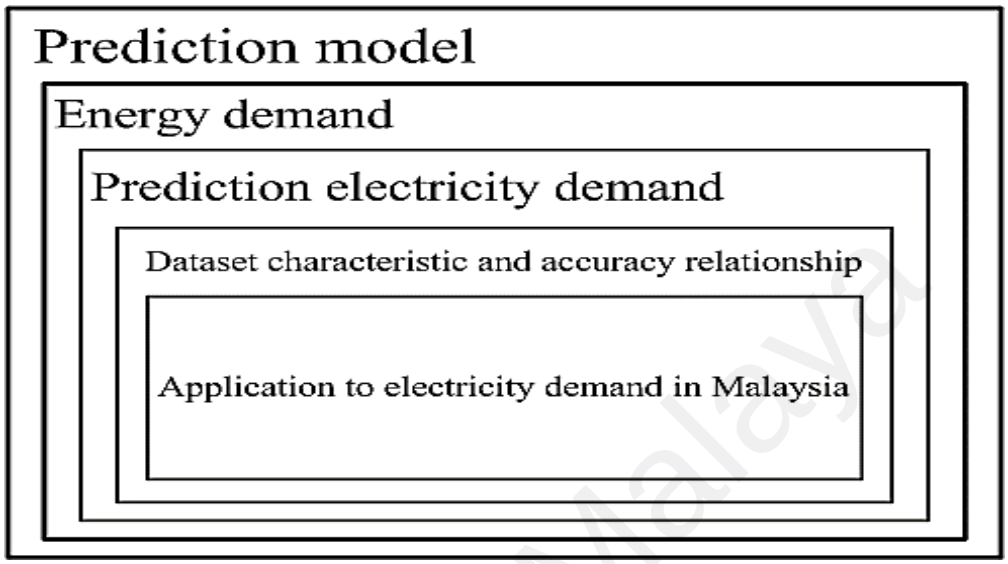


Figure 1.3: Scope of the study

1.7 Contributions of the Study

Based on the outcome of this study, a new long term approach and methodology for predicting electricity demand for Malaysia's consumption is provided. The study developed for this thesis illustrates how the accuracy problem of the electricity prediction model is due to analyzing an input dataset that has linearity and nonlinearity patterns. These can be solved with uncombined (pure-bred) approach models. The study also illustrates how a few percentages of errors in predicting the electricity demand rate can affect economy vitally, especially when viewed by decision makers. The main contribution of this thesis is for improving the accuracy of the electricity demand prediction model by designing and implementing a new prediction approach. The proposed approach depends on two things. First, by reducing the multicollinearity problem of the input data set, the proposed model becomes more reliable. Second, it solves the problem of the residual errors that occur as a result of the complexity pattern of the input dataset. Through this new approach, the study identifies a new effect on the

University of Malaya

accuracy of the electricity demand prediction model. The study thus, provides a new perspective in viewing the prediction model where errors can be recorded when a dataset that contains different patterns is applied to the linear or nonlinear based prediction model. Such errors could not be eliminated by just changing the process of building a prediction model from linear based to nonlinear based or vice-versa. It actually needs a special process that can minimize or eliminate changes in the dataset patterns whilst also keeping the information and reality of the prediction model intact.

The improvement process of the new approach comes in some sequenced layers where a type of problem is solved at each layer. However, the works of the overall layers are better at providing the accuracy of the prediction rate rather than other well-known predictor tools and methods.

In summary, this study has made the following contributions.

- This study has viewed the relevant works in a new taxonomy that group works into three main classes: linear, nonlinear, and hybrid models. The works in each group has been discussed from the view point of relations between the input dataset patterns and the property of the tools that were utilized as predictors. (Using a new taxonomy in Section 2.5)
- Selecting improper predictor tools with reference to the property of input dataset (linear data processed or analyzed by nonlinear tool, or vice versa) can lead to a high rate of residual errors. This work proposes a new method to deal with such residual errors by including them in the calculation process as a means of getting better accuracy. (Using BPNN to receive residual errors and process them for better accuracy obtaining)
- It was difficult for researchers in the past to take into account all the predictor variables that can impact on electricity demand in a single analysis. These

University of Malaya

researchers have also disregarded some predictor variables in the bid to minimize the dimensionality of the input dataset, which affects the accuracy rate. Such an action could result in an unreliable prediction result. This study found a tradeoff between minimizing the number of predictor variables (size dimension complexity) and the reliability of the obtained results. (Using PCA to extract as much information as possible in the input data set and then reduce the input variables).

- Propose a new long term based approach that can predict electricity demand rate for Malaysia.

1.8 Summary

Electricity demand prediction rate is very important process for developed and non-developed countries. There are many techniques used to estimate this rate based on historical (time based data) for an area.

Accuracy of prediction models are very important as it reflects huge amount economically. There are many things affecting the rate of accuracy and errors of a prediction model, among which are multicolinarity of input dataset and residual error of utilized models. These two problems could be overcome when a model can reduce the multicolinarity of input variables and involve the residual errors again in the prediction calculation.

University of Malaya

1.9 Organization of the Thesis

This thesis consists of 6 chapters followed by references and appendices.

Chapter 1 describes the background and introduction of the study, problem statement and research objectives of the study that are related to the research question, the significance of the study, the scope of the work, and finally, the organization of the thesis.

Chapter 2 provides a review of the literature. This involves a detailed exposition on the principal component in electricity demand and a discussion of the prediction model of electricity demand. This is followed by a discussion of each nonlinear- PCR, nonlinear-ANN, and hybrid system in electricity demand prediction.

Chapter 3 presents the methodology used in this study. The framework of the PCR-BPNN model is given a through explanation as it is the main component of the study. In addition, the chapter also provides detailed explanations about each part of the PCR-BPNN model. The theory and all the formula used for each part of the model that is employed to execute the study is also provided.

Chapter 4 illustrates the execution parameters and characteristics of each part of the PCR-BPNN model. It includes the results obtained through the model execution. The discussion is summarized at the end of the chapter.

Chapter 5 presents the testing of the main prediction models. Throughout this chapter, the study evaluates the results that have been obtained for testing these models. The comparison between the tested predictive models done previously and the PCR-BPNN model is also illustrated.

Chapter 6 consists of further discussion of the results and findings. It also explains the achievements of the current research and suggestions for future work.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter discusses an overview of the related literature and is divided into three major parts. The first section presents several approaches and techniques of the electricity demand prediction models, which mainly proposed to control overestimate and underestimate errors in the electricity consumption. In this study, these techniques are categorized into three groups: linear models, nonlinear models, and hybrid systems. The second section details some characteristics of input dataset that are related to accuracy of prediction models.

The third section details some performance indicators that used to evaluate the accuracy of prediction models. The chapter shows multiple methods that are utilized to determine factors that are significant in explaining the electricity demand. As identification of tools and factors related to electricity demand prediction models is important, this review discusses studies related to electricity demand prediction models conducted from 2000 to 2013. These methods are also evaluated to confirm the most accurate technique in predicting electricity rates. The chapter concludes with an overall summary and outputs for the literature review.

2.2 Electricity Demand

Electricity is a type of energy that becomes a basic requirement for economic development. It is also considered as an adequate standard of living. This type of energy is consumed in all fields of the human's live at every second, and day after day the rate of consumption increases with population and economic growth. Usually, policy maker can estimate the consumption rate of electricity the amount of the demand that is necessary for an area. According to Altinay et al.(2005) at a given point of time, electricity

University of Malaya

demand is represented by the maximum amount of electricity consumed and it is often represented in kilowatt or kilovolt amperes.

For any country, it is vital to expect the demand rate of an area for some years in advance. There are a lot of ways to do that. One of the most popular methods is using statistical approaches for analyzing factors that have been affected on electricity demand (Kheirkhah et al.2013). The results of these analyses give the rate of electricity demand. However, the predicted rate through these statistical approaches can be dependable only when their results could be validated against some accuracy or performance indicators. Therefore, validation process for electricity demand prediction models should be carried out in such studies (Akay et al.,2007).

Next section presents different factors which are related to the rate of electricity demand.

2.3 Factors Related to Electricity Demand

This section discusses important factors that have impacts on the rates of electricity consumption, and consequently, on the electricity demand. These factors , which also known as independent variables, varied from one country to another (Kucukali et al.,2010), as the characteristics of the countries involved vary in terms of temperature, environment, economic and population growth, and demands on electricity. For that, this work reviewed different studies in different countries to collect all factors and independent variables that influenced with electricity demand.

One of the important factors is economic growth. Dalvand et al.(2008) utilized the economic growth in designing an electricity prediction model because standard of living has been influenced by economic growth. They showed that commonly used indicators

University of Malaya

to represent economic growth are GDP, gross national product (GNP), GDP per capita, income, export, and import. Other studies included both population as well as economic growth in their study (Kavaklioglu et al., 2009). This is due to strong relationship between electricity demand and population. While several studies Nasr et al.,(2000) and Yuan et al.,(2007). Supported economic growth and population have great impact on electricity demand in developed countries, others for example, (Saravanan et al.,2012), also found that these factors are also affecting electricity demand in developing countries. .

Besides population and economic growth, other important factors are factors related to weather condition. A study investigated the impact of weather, such as climate, CO_2 emission, and humidity on the electricity prediction rate (Al-Ghandoor et al.,2008). Both heating and cooling a home or an office take large amount of energy, more than that what was used for any other appliance. Heating or cooling with natural gas produces carbon dioxide (CO_2). This means that in such countries demand on electricity will be changed with the weather. Another study found, weather, economic, and population factors, taken together are significant when they used simultaneously in predicting electricity demand (Ekonomou,2010). Moreover, studies were conducted in industrial and residential sectors to determine factor-independent variables for prediction models for electricity demand (Lai et al.,2008) and (Zhang et al.,2012).

Some studies included the rate of electricity or energy consumption into the model. In some studies these consumption rates (or load) has been sub-classified into residential, commercial, and industrial sectors. Even more, some studies depended on when and at what rate the maximum or minimum consumptions are occurred. Consumption rate is coming in another form in some studies, such as monthly, daily besides the annual load consumptions.

University of Malaya

Some factors related to factors discussed above were also included in other studies. For example urbanization development degree is related to the rate of population. There are many factors that are related to the economic impact such as gas price, oil price, price of electricity, agricultural-GDP (AGDP), and non-agricultural-GDP (NGDP). Still in the relation among factors, the weather factor has strong relation with many variables that employed in electricity demand studies, such as temperature and CO_2 emission.

As a summary, there are many factors that can be considered as significant independent variables, which demand rate of electricity in specific area will depend on them. The impact and the relation of these variables with electricity varies according to location of the study.

The next section presents a review of literature that utilized different techniques and algorithms for analyzing electricity demand related factors and getting an accurate prediction rate of electricity demand.

2.4 Electricity Demand Prediction Model

This section reveals the types of tools and algorithms utilized to build the prediction models. A review of previous works also allows us to trace the common steps followed by authors in developing an electricity prediction model.

Development of a prediction model includes preparation of input dataset, identifying suitable tools and algorithms, and testing the proposed models against errors and accuracy rate. All studies followed the same sequence of steps from conceptualization until validation of their models. Evaluation of factors include investigating where they they are positively or negatively related to electricity consumption. Although these factors affect and are related to the electricity demand or consumption, most researchers did not

University of Malaya

use all of them in their analyses because the analyses they used are not able to handle the complexity of the input dataset. The increase in complexity of the input dataset increases error rates and decreased accuracy. For instance, two articles estimated the electricity demand in Turkey,(Erdogdu, 2007) and (Akay et al., 2007). Although both studies were conducted in the same year, the types of factors reported and slightly differed, despite the presence of several constant factors. Nevertheless, these studies used different tools to build their respective prediction models. Generally, researchers tend to change the tools they are using from one study to another to demonstrate that a specific tool can provide higher accuracy than other tools. Table 2.1 shows that factors and a variety of tools utilized by different researchers from different countries. The detailed discuss can be found in Section 2.5.1 – 2.5.3.

University of Malaya

Table 2.1: Summary of the literature review of electricity demand 2000 – 2015

A - Linear model				
#	**Title**	**Method / applied**	**Factors**	**Names & Year**
1.	"Econometric modeling of electricity consumption in post-war Lebanon"	Co-integration	GDP	Nasr et al., 2000
2.	"The relationship between elasticity consumption, electricity prices and economics growth: Time series evidence from Asian developing countries"	Co-integration	GDP, income and Population	Asafu J., 2000
3.	"The relationship between energy consumption and economic growth in Pakistan"	Granger causality method	GDP	Aqeel & Butt, 2001
4.	" On the relationship between electrical energy consumption and climate factors in Lebanon: co-integration and error-correction models"	Co-integration	Humidity, and Temperature	Badr & Nasr, 2001
5.	"Economic variables and electricity consumption in Northern Cyprus"	MLR	electricity consumption and historical economic (GDP)	Egelioglu et al., 2001
6.	"Electricity consumption and economic growth in India"	Granger causality	GDP per capita	Ghosh, 2002
7.	"Forecasting the primary energy demand in Turkey and analysis of cyclic patterns"	exponential smoothing linear regression	Population, GNP, Industrial and Commercial	Ediger & Tatlıdil, 2002
8.	"Modeling and forecasting the demand for electricity in New Zealand: a comparison of alternative approaches"	ECM and ARDL	GDP and Previous Electricity Consumption	Fatai et al., 2003

University of Malaya

9.	"Electricity consumption and economic growth in China"	Co-integration with the Granger causality test.	GDP	Shiu & Lam, 2004
10.	"The impact of electricity supply on economic growth in Sri Lanka"	ordinary least squares regression models	GDP	Morimoto & Hope, 2004
11.	"Cointegration and causality between electricity consumption and GDP: empirical evidence from Malawi"	Co-integration and ECM	GDP,AGDP and NGDP	Jumbe, 2004
12.	"Residential electricity demand in Taiwan. Energy Economics"	ECM and Co-integration	income, population, price of electricity and degree of urbanization	Holtedahl & Joutz, 2004
13.	"Estimating residential demand for electricity in Greece. Energy Economics",	VECM	price of electricity, income and Temperature	Hondroyiannis, 2004
14.	"Electric energy demand of Turkey for the year 2050"	linear regression	Previous Electricity Consumption , income per capita and population	Yumurtaci & Asmaz, 2004
15.	"Energy consumption and GDP in developing countries: A cointegrated panel analysis"	Co-integration	GDP	Lee, 2005
16.	"Electricity consumption, employment and real income in Australia evidence from multivariate Granger causality tests"	multivariate Granger causality tests	Previous Electricity Consumption , employment and income	Narayan & Smyth, 2005a

University of Malaya

17.	"Forecasting electricity consumption in New Zealand using economic and demographic variables"	MLR	GDP, price of electricity and population	Mohamed & Bodger, 2005
18.	"Electricity consumption and economic growth: evidence from Korea"	Co-integration and ECM	GDP	Yoo et al., 2005
19.	"Electricity consumption and economic growth: evidence from Turkey"	VAR	GDP	Altinay & Karagol, 2005
20.	"The residential demand for electricity in Australia: an application of the bounds testing approach to cointegration"	bounds testing approach to Co-integration	income, temperature and price of electricity	Narayan & Smyth, 2005b
21.	"Electricity consumption and economic growth: a time series experience for 17 African countries"	Co-integration	GDP	Wolde-Rufael, 2006
22.	"The causal relationship between electricity consumption and economic growth in the ASEAN countries"	Granger causality test	GDP	Yoo et al., 2006
23.	"Comparing linear and nonlinear forecasts for Taiwan's electricity consumption"	ARMAX and ANN	Income, population ,GDP and consumer price index	Pao, 2006
24.	"Electricity consumption in G7 countries: A panel cointegration analysis of residential demand elasticities"	Panel Cointegration	Income and price of Electricity	Narayan et al., 2007
25.	"Electricity consumption and economic growth: Bounds and causality analyses of OPEC members"	Co-integration	GDP	Squalli, 2007

University of Malaya

26.	"Electricity demand analysis using cointegration and ARIMA modeling: A case study of Turkey"	ARIMA	Income, price of electricity and GDP	Erdogdu, 2007
27.	"The relationship between GDP and electricity consumption in 10 Asian countries"	Unit root and Co-integration	GDP	Chen et al., 2007
28.	"Electricity consumption and economic growth in China: Cointegration and co-feature analysis"	Cointegration	GDP	Yuan et al., 2007
29.	"An empirical analysis of electricity consumption in Cyprus"	time series techniques	Income, prices of Electricity and Temperature.	Zachariadis & Pashourtidou, 2007
30.	"Estimation of residential electricity demand function in Seoul by correction for sample selection bias"	Granger causality test	Income and price of electricity	Yoo et al., 2007
31.	"Causality relationship between electricity consumption and GDP in Bangladesh"	cointegration and vector error correlation	per capita electricity consumption and GDP per capita	Mozumder & Marathe, 2007
32.	"Electricity consumption and associated GHG emissions of the Jordanian industrial sector: Empirical analysis and future projection"	MLR	Emissions CO_2	Al-Ghandoor et al., 2008
33.	"The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach"	MLR	Temperature	Bessec & Fouquau, 2008

University of Malaya

34.	"Electricity demand loads modeling using Auto regressive Moving Average (ARMA) models"	ARMA	Previous Electricity Consumption	Pappas et al., 2008
35.	"Seasonal variations in residential and commercial sector electricity consumption in Hong Kong"	PCA with MLR	Residential and commercial sector electricity	Lam et al, 2008
36.	"Electricity consumption forecasting in Italy using linear regression models"	Linear regression	Population GDP, and GDP per capita	Bianco et al., 2009
37.	"The application of seasonal latent variable in forecasting electricity demand as an alternative method"	ARIMA, SARIMA and regression model	Previous Electricity Consumption	Sumer et al., 2009
38.	"Electricity consumption–growth nexus: The case of Malaysia"	ARDL	GDP	Chandran et al., 2010
39.	"An Improved Combined Forecasting Method for Electric Power Load Based on autoregressive Integrated Moving Average Model"	ARIMA	Previous Electricity Consumption	Xin et al., 2010
40.	"Short-term forecasting of power flows over major transmission interties: Using Box and Jenkins ARIMA methodology"	ARIMA	Previous Electricity Consumption	Paretkar et al., 2010

University of Malaya

41.	"The causal relationship between energy consumption and GDP in Albania, Bulgaria, Hungary and Romania: Evidence from ARDL bound testing approach"	ARDL and bounds test	GDP	Ozturk & Acaravci, 2010
42.	"Application of Principal Component Regression Analysis in power load forecasting for medium and long term"	PCR	GDP, primary industry output, secondary industry output, tertiary industry output per capita annual, disposable income of urban households, per capita annual net income of rural household, resident population and urbanization	Yingying & Dongxiao, 2010
43.	"Forecast of electricity consumption in Cyprus up to the year 2030: The potential impact of climate change"	ARDL	macroeconomic variables, prices of electricity and Temperature	Zachariadis, 2010
44.	"Modeling of energy consumption based on economic and demographic factors: The case of Turkey with projections"	MLR	Population and GDP	Aydin, 2014

University of Malaya

45.	"Modeling and forecasting demand for electricity in Bangladesh: econometrics mode"	Auto Regressive Econometric Modelling	Price of Electricity , GDP per capita	Shuvra et al., 2011
46.	"Principal Component Analysis of Electricity Consumption Factors in China"	PCA	GDP ,income, industrial output value, exports, imports, and added industry services	Zhang et al., 2012
47.	"Electricity consumption-GDP nexus in Pakistan: A structural time series analysis"	Regression	GDP, price of electricity and the underlying energy demand trend)	Zhang et al., 2012
48.	"The nexus between electricity consumption and economic growth in Bahrain"	ARDL and cointegration	real foreign direct investment per capita, and GDP per capita	Hamdi et al., 2014
49.	**Expectation from this model**	**PCR** **Can solve multicollinearity**	**More Variables**	
B - Nonlinear Model				
50.	"Forecasting the short-term demand for electricity: Do neural networks stand a better chance?"	ANN and ARIMA	Previous Electricity Consumption	Darbellay & Slama, 2000

University of Malaya

51.	"A fuzzy expert system for peak load forecasting application to the Greek power system"	Fuzzy Logic	Previous Electricity Consumption and Temperature.	Kiartzis et al., 2000
52.	"Support vector machines for short-term electrical load forecasting"	SVM	Previous Electricity Consumption	Mohandes, 2002
53.	"Energy demand estimation based on two-different genetic algorithm approaches"	GA	GDP, population, import and export	Ersel et al., 2004
54.	"Load forecasting using support vector machines: A study on EUNITE competition 2001"	SVM	Previous Electricity Consumption and Temperature.	Chen & Chang, 2004
55.	"Electricity estimation using genetic algorithm approach: a case study of Turkey"	GA	GNP, population, import and export	Ozturk et al., 2005
56.	"Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting"	Wavelet	Previous Electricity Consumption	Benaouda et al., 2006
57.	"Comparing linear and nonlinear forecasts for Taiwan's electricity consumption"	ANN	Income, Population, GDP and consumer price index	Pao, 2006
58.	"Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks"	ANN , decision Tree and Regression	Previous Electricity Consumption	Tso & Yau, 2007
59.	"Forecasting electrical consumption by integration of Neural Network, time series and ANOVA"	ANN	Previous Electricity Consumption	Azadeh et al. , 2007

University of Malaya

60.	"Electricity price forecasting in Iranian electricity market applying Artificial Neural Networks"	ANN	price of electricity	Zarezadeh et al., 2008
61.	"Long-term load forecasting of Iranian power grid using fuzzy and artificial neural networks"	ANN	GDP,GNP, Iranian oil price, value-added of manufacturing and mining group, oil income, population, consumer price index gas consumption electricity and supply exchange rate	Dalvand et al., 2008
62.	"Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors"	ANN , Regression model	Previous Electricity Consumption	Azadeh et al., 2008
63.	"Modeling and Forecasting of Short-Term Half-Hourly Electric Load at the University of Ibadan, Nigeria"	ANN	Previous Electricity Consumption	Fadare & Dahunsi, 2009
64.	"Computational intelligence approach to load forecasting-a practical application for the desert of Saudi Arabia"	ANN	Temperature, Previous Electricity Consumption and wind speed	Ahmmed et al., 2009
65.	"Clustering based short term load forecasting using artificial neural network"	ANN	Previous Electricity Consumption and temperature	Jain & Satish, 2009
66.	"Medium and long-term load forecasting based on PCA and BP neural network method"	PCNN	Previous Electricity Consumption	Zhang & Wang, 2009

University of Malaya

67.	"Load forecasting of a desert: A computational intelligence approach"	ANN	Previous Electricity Consumption	Saber & Al-Shareef, 2009
68.	"Short term load forecasting using a robust novel Wilcoxon Neural Network"	Wilcoxon neural network	Previous Electricity Consumption	Mishra & Patra, 2009
69.	"Artificial Neural Networks and regression approaches comparison for forecasting Iran's annual electricity load"	ANN	GDP and population	Ghanbari et al., 2009
70.	"Modeling and prediction of Turkey's electricity consumption using Artificial Neural Networks"	ANN	population, GNP, import and export	Kavaklioglu et al., 2009
71.	"Energy demand estimation of South Korea using artificial neural network"	ANN	GDP, population, and import and export	Geem & Roper, 2009
72.	"Short term load forecasting using an artificial neural network trained by artificial immune system learning algorithm"	ANN	temperature, holidays, and days in a week	Abdul Hamid & Abdul Rahman, 2010
73.	"Turkey's short-term gross annual electricity demand forecast by fuzzy logic approach"	Fuzzy	GDP	Kucukali & Baris, 2010
74.	"Research on short-term power load time series forecasting model based on BP neural network"	ANN	Previous Electricity Consumption	Niu et al., 2010
75.	"Greek long-term energy consumption prediction using artificial neural networks"	ANN, linear regression and SVM	Temperature, Previous Electricity Consumption ,GDP and Installed power capacity	Ekonomou, 2010

University of Malaya

76.	"Electricity demand forecasting of Electricite Du Lao (EDL) using neural networks"	ANN	GDP, population, price of electricity & Number of house	Sackdara, Premrudeep reechacharn, & Ngamsanroaj, 2010
77.	"Estimation of electricity demand of Iran using two heuristic algorithms"	GA	GDP, population, number of customers and price electricity	Amjadi et al., 2010
78.	"Short term load forecasting in Mauritius using Neural Network"	ANN	Previous Electricity Consumption	Bhurtun et al., 2011
79.	"One day-ahead load forecasting by artificial neural network"	ANN	hourly load electricity	Mosalman et al., 2011
80.	"Daily peak load forecasting using ANN"	MLP-ANN	maximum load (Lmax) and maximum temperature (Tmax)	Tasre et al., 2011
81.	"Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variables"	ANN & Regression model	GDP, population, import and export amounts, and employment	Kankal et al., 2011
82.	"Forecasting electricity demand in Thailand with an artificial neural network approach"	ANN	Population, GDP, export, Previous Electricity Consumption	Kandananond, 2011
83.	"Hourly load forecasting using Artificial Neural Network for a small area"	ANN	Previous Electricity Consumption and Temperature	Tasre et al., 2012
84.	"ANN application for the next day peak electricity load prediction"	ANN	Previous Electricity Consumption	Milojkovic et al. , 2012
85.	"India's Electricity Demand Forecast Using Regression Analysis And Artificial Neural Networks Based On Principal Components"	PCNN and PCR	Population, GDP per capita, Imports, Export and Electricity Consumption Per capita	Saravanan et al., 2012

University of Malaya

86.	"Area-Load Based Pricing in DSM Through ANN and Heuristic Scheduling"	ANN	Previous Electricity Consumption	Kunwar & Kumar, 2013
87.	"Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis"	PCNN with DEA	Previous Electricity Consumption	Kheirkhah et al., 2013
88.	"Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types"	Regression and ANN	Previous Electricity Consumption	Ardakani & Ardehali, 2014
89.	"Development of a Novel Approach for Electricity Forecasting"	NNs and decision tree	Previous Electricity Consumption	Moghaddam & Bahri, 2014
90.	"Comparative Study of Grey Forecasting Model and ARMA Model on Beijing Electricity Consumption Forecasting"	GP and ARIMA	Previous Electricity Consumption	Guo et al., 2014
91.	"Electricity Consumption Forecasting in Thailand Using an Artificial Neural Network and Multiple Linear Regression"	ANN and MLR	GDP, Population, Maximum ambient temperature and E D	Panklib et al., 2015
92.	"Up to year 2020 load forecasting using neural nets"	ANN	population, GDP, GNP, number of households, CO_2, index of industrial production, electricity consumption, oil price and number of air-conditioners	Kermanshahi & Iwamiya, 2002

University of Malaya

93.	"Predicting Annual Electricity Consumption In Iran Using Sing Artificial Neural Networks (Narx)"	Narx	population, GNP, import and export	(Kargarl et al,. 2014)
94.	**Expected from this work**	**ANN** **Can compute residual errors**	**More variables than others**	
C - Hybrid system				
95.	"A Short-term Load Forecasting Based on Support Vector Regression"	SVR	Daily electricity consumption	Lu Yu 2015
96.	"Wavelet transform and neural networks for short-term electrical load forecasting"	WNN	Previous Electricity Consumption and Temperature	Yao et al., 2000
97.	"Short-term load forecasting for the holidays using fuzzy linear regression method"	Fuzzy Linear Regression Method	Previous Electricity Consumption	Song et al. , 2005
98.	"Short-term load forecasting based on an adaptive hybrid method"	Hybrid network with (SOM) and (SVM).	daily Electricity consumption	Shu & Luonan, 2006
99.	"Grey prediction with rolling mechanism for electricity demand forecasting of Turkey"	GPRM	Previous Electricity Consumption	Akay & Atak, 2007

University of Malaya

100.	"Modeling of electricity consumption in the Asian gaming and tourism center—Macao SAR, People's Republic of China"	MLR, ANN and WANN	Population, the number of tourists, hotel room occupancy and days per month.	Lai et al., 2008
101.	"A hybrid simulation-adaptive network based fuzzy inference system for improvement of electricity consumption estimation"	FN	Monthly electricity consumption	Azadeh et al., 2009
102.	"A trend fixed on firstly and seasonal adjustment model combined with the ε-SVR for short-term forecasting of electricity demand"	SVR	Previous Electricity Consumption	Wang et al., 2009
103.	"Fuzzy wavelet neural network for prediction of electricity consumption"	FWNN	Previous Electricity Consumption	Rahib H Abiyev, 2009
104.	"An Efficient Hybrid Model to Load Forecasting"	NN-PSO	Previous Electricity Consumption	Hasan et al., 2010
105.	"Forecasting of short-term electric load using application of wavelets with feed-forward neural networks"	AWNN and MLPANN	Previous Electricity Consumption	Pindoriya et al.,2010
106.	"Combined modeling for electric load forecasting with adaptive particle swarm optimization"	S-ARIMA with S-ESM	Previous Electricity Consumption	Wang, Zhu et al., 2010

University of Malaya

107.	"Short-term load forecasting: Similar day-based wavelet neural networks"	WNN	Previous Electricity Consumption and Temp.	Chen et al., 2010
108.	"Modeling and forecasting electricity consumption of Malaysian large steel mills"	regression model with (MAED_EL)	population, consumption per capita, GDP and Price of Electricity	Aman et al., 2011
109.	Load forecasting using hybrid models.	WFNN	Previous Electricity Consumption and Temp.	Hanmandlu & Chauhan, 2011
110.	"Modeling and prediction of Turkey's electricity consumption using Support Vector Regression"	SVR	Population, GNP, import and export	Kavaklioglu, 2011
111.	"Electricity Demand Estimation Using an Adaptive Neuro-Fuzzy Network: A Case Study from the State of Johor, Malaysia"	ANFIS	GDP, GNP, employment, and unemployment	Bazmia et al., 2012
112.	"A review on short term load forecasting using hybrid neural network techniques"	SAPSO neural network	E.C (daily, weekly and monthly)	Raza & Baharudin, 2012
113.	"An integrated fuzzy mathematical model and principal component analysis algorithm for forecasting uncertain trends of electricity consumption"	PCA-FR	monthly electricity consumption	Azadeh et al., 2013

University of Malaya

The most suitable approach is starting with the oldest work on electricity demand prediction. To the best of our knowledge, the oldest article was written by Murray (1978), who investigated the electricity demand in Virginia in 1978.

In the current study, models are classified into three categories: linear, non-linear, and hybrid system. Method(s) and related factors used in each article were evaluated and classified based on these categories. Linear models employ co-integration, multiple linear regression, vector error correction model, ARDL, and autoregressive integrated moving average (ARIMA). Nonlinear models utilize support vector machine (SVM), fuzzy logic, genetic algorithm (GA), and ANN. Most commonly used tools in hybrid system are Wilcoxon neural network (WNN), GPRM, SVR, FN, adaptive neural fuzzy inference system (ANFIS), PCR, and PC–BPNN models. Application of these methods is dependent on the purpose or goal of the study.

2.5 Types of Models

In this section, the type of approaches that was utilized for building electricity demand prediction models are explained with reference to their impact on the accuracy. As mentioned in Table 2.1, the commonly used approaches are linear, nonlinear, and hybrid systems. Subsequent sub-sections discuss several approaches utilized by previous studies and explain the relationship between these proposed approaches and the patterns of the dataset fed to the models.

University of Malaya

2.5.1 Linear Models

This section focuses on the relationship between linear approaches and the types and patterns of the dataset used.

Linear approaches are used to identify the causal relationship between several independent variables and the rate of the electricity consumption. As an example for the linear approaches, Narayan et al.(2007), employed co-integration and equation-correction model (ECM) to determine causal relationship between consumption rate of electricity and income per capita. However, co-integration and ECM model were also employed by (Jumbe 2004) to investigate the causal relationship of GDP and employment rates to electricity consumption in Malawi.

Aqeel et al. (2001) used the Hisao's version of the Granger causality method for data collection in Pakistan. The results showed that GDP significantly affects the rate of electricity consumption. The Granger causality method was co-integrated with vector auto-regression model (VAR) by (Ghosh 2002) to evaluate the relationship between consumption of electricity and GDP in India. This study showed a strong relationship between these two factors. A similar relationship between consumption of electricity and GDP was also investigated by Shiu et al. (2004) in China. These researchers used co-integration with the Granger causality test. The causality between consumption of electricity and GDP was also investigated by Morimoto et al. (2004) in Sri Lanka by using the co-integration method. The relationship between parameters that used as independent variables and electricity consumption was further investigated by Holtedahl et al. (2004) in Taiwan. In this study, they evaluated different parameters, such as earned income, population, electricity price, and urbanization degree. The results showed varied influences of the parameters tested. The causality relationship was further expanded by Lee (2005) to include 18 developed countries by using two panels, namely, a panel unit

University of Malaya

and a panel based on error correction. The results showed that income changes in developed countries significantly affect electricity consumption rates. Similarly, a study in Australia assessed the causality between the given parameters and electricity demand (Narayan et al.,2005a). In this study, the relationship among consumption of electricity, employment factor, and real income factor was examined using co-integration and causality framework. This work was also expanded to analyze the causality of electricity demand on long and short terms run of the elasticity of residential electricity demand in Australia by using the bounds testing procedure co-integrated within an autoregressive–distributive lag framework (Narayan et al. 2005b). The G7 countries were further investigated by this group of researchers to expand their work by applying panel unit root and panel co-integration techniques; they estimated long- and short-term income factor and price elasticity for residential demand (Narayan et al.,2007). Similarly, long- and short-term causality issues in South Korea were investigated by Yoo (2005) by using co-integration and ECM model found that consumption of electricity is affected by growth of economic. More Asian countries were included by Chen et al. (2007) to investigate electricity consumption. In this study, single datasets and panel data procedures were applied in 10 newly industrialized countries. The results demonstrated that the type of data affects the directionality of the relationship between the economy and electricity. A one-way short-term causality running was further found within the single dataset and started from economic attributes to electricity consumption, whereas bidirectional long-term causality was observed in the panel data procedures. These findings were confirmed in the study of (Pappas et al., 2008). Moreover, a new approach was proposed by Altinay et al., (2005) to investigate the causality relationship between electricity rate consumption and real GDP in the Turkey. The Granger causality and non-causality were examined using two different tests in their framework, and the results showed that the supplied electricity rate should satisfy the growth of electricity consumption.

University of Malaya

The causality relationship between electricity demand and growth of economic could be bidirectional or unidirectional, and this relationship exhibits independent or dependent forms. In this regard, Yoo,(2006) investigated the direction of this relationship among four countries included in the Association of South East Asian Nations. The economic–electricity consumption relationship is unidirectional in Malaysia and Singapore and bidirectional in Indonesia and Thailand.Chandran et al., (2010) estimated the relationship between electricity consumption and GDP rate in Malaysia. This study showed that the short-term causality between the economy and the rate of electricity consumption is unidirectional (Squalli,2007) investigated the dependent and independent relationship between the economy and electricity consumption in members of the Organization of the Petroleum Exporting Countries (OPEC). In this study, the economic growth is dependent on electricity consumption in several countries but independent in other OPEC countries. Countries that newly entered into the European zone, such as Albania, Bulgaria, Romania, and Hungary, were further included in the study of (Acaravci et al.,2010). The results demonstrated that the economic growth in these countries increases the rate of electricity consumption, although the effect is unidirectional in Hungary and bidirectional in the three remaining countries. This evidence supported the study of (Fatai et al.,2003) who showed that factors influence consumption of electricity rate vary from one region to another region; thus, models developed for one region may differ from those developed for another region.

In addition to the economy and population, many other factors that influence electricity demand have been investigated through linear techniques. Badr et al., (2001) evaluated climate-based factors, such as temperature, humidity, and clearness-of-sky index. Egelioglu et al., (2001) studied the number of customers and tourists to predict annual electricity consumption. Several studies further assessed mixed parameters, such as economics, electricity price, and temperature to determine the parameter that influences

University of Malaya

electricity demand (Hondroyiannis,2004). Yumurtaci et al., (2004) utilized previous electricity consumption rate as a factor in computation of electricity demand rate. The results showed that previous electricity consumption rate provides an accurate electricity demand prediction model. Therefore, many papers have focused in predicting electricity rate (Mohamed et al., 2005).

Several studies have investigated the economy and price elasticity in certain regions for a long period of time to predict electricity consumption rate. On the basis of this concept, (Erdogdu,2007) conducted a study in Turkey by combining co-integration with ARIMA. Shuvra et al.,(2011) forecasted the demand rate of electricity consumption in Bangladesh. These works utilized the following parameters, namely, price of gas, GDP per capita, and income. The prediction rate estimated through prediction models should be evaluated and tested. Therefore, Al-Ghandoor et al.,(2008) employed ANOVA to check the significance of the results and multiple linear regression (MLR) to estimate the electricity consumption by the industrial sector in Jordan. Another linear regression-based forecasting model was designed by Bianco et al.,(2009) to estimate the demand rate in Italy. These authors assessed many factors, such as GDP, per capita GDP, population, and electricity price and found that inclusion of more factors in prediction models may result in more accurate prediction rates which employed the MLR approach and inputted with a dataset comprising population and GDP. Nevertheless, the GDP used in this study presents nonlinear patterns, which are difficult to be captured using MLR. The accuracy of the MLR model is also affected by the multicollinearity of the input dataset, which cannot be solved by the model. Consequently, the multicollinearity affects the accuracy and the reality of the results. Most linear studies in this sub-section did not determine the multicollinearity problem among independent variables. In cases with highly correlated independent variables, multiple regression analysis faces serious challenges (McAdams et al.,2000a). Fekedulegn et al.,(2002) reported that multicollinearity, which shows that

University of Malaya

the high correlation coefficient among independent variables in a regression model, negatively affects the ability of MLR to correctly identify the most important factors affecting the process. This result was confirmed by Maddala,(1992), who suggested that MLR cannot easily interpret the estimation of the individual coefficients if the variables are highly inter-correlated. Therefore, a method for removing such multicollinearity and redundant information must be developed and one of the proposed strategies is multivariate data analysis (MDA).

Previous studies employed these two linear approaches in setting accurate prediction rates when using input datasets with linear patterns. MLR is a widely used linear-based method to express the relationship between a response variable and several independent variables, whereas PCR model is a sequential process with the combined MLR and PCA techniques (Draper et al.1981).

This technique can be used identify the trends and relationships large environmental data (Saravanan et al.,2012); (Yingying et al.,2010) and (Zhang et al., 2012). MDA can reduce data dimensionality, thereby simplifying the possible models that can be used to describe the dataset. A well-known MDA method is PCA, which was proposed by Hotelling in 1933. PCA is a multivariate statistical technique that can be applied to quantitatively explain the degree of inter-dependency for a set of correlated variables (Von Storch et al.,2001).

In many regression analysis processes, PCA is used to moderate the multicollinearity problem. This approach explores the relationships among independent variables when the defined predictors are insignificant. PCA utilizes principal components as new independent variables, which are ideal predictors in regression equations (PCR) because they can optimize special patterns and problems caused by multicollinearity (Jolliffe,2005) and (Myers,1986).

University of Malaya

When a regression model becomes incompatible and complicated because of complex and nonlinear relationships among multiple variables (Comrie,1997), prediction models are expected to underperform if utilized to fit the relationships between electricity demand and other related independent variables (Pao,2006).

Several studies have applied PCA in load prediction to solve the multicollinearity problem. In this scenario, PCA is used to reduce the correlation between independent variables without losing any information from response variables (Azadeh et al., 2009). These components vary from those with high variances to those with low variances. For example, the principal component 1 (PC1) presents a higher variance degree than PC2, PC2 presents higher variance degree than PC3, and so on until up to PCn, which is usually the remainder.

Zhang et al., (2010). Utilized PCA and MLR to forecast electricity consumption. With the use of PCA, researchers can evaluate more parameters (i.e., GDP, income, industrial output value, exports, household numbers, population, price index, and added services industry value).

Ndiaye et al.,(2011) used PCA to generate regression models for electricity consumption of 221 households in Canada. The result showed that only nine factors among 59 factors are significant. Lam et al.(2008) proposed a multiple linear regression model based on two principal components to examine the electricity consumption for commercial and residential sectors in Hong Kong. The result showed that the commercial sector could be predicted more accurately than the residential sector as evidenced by the error rates measured based on normalized mean-bias error (NMBE). Similarly, Yingying et al.(2010) proposed a PCR analysis to predict medium- and long-term power loads and the result showed that the model is feasible and effective for load prediction. PCR-based load prediction also effectively retains most information of the original variables

University of Malaya

compared with other models. As the PCR model can remove the multicollinearity problem among independent variables, its performance indicators illustrate the improvement in the accuracy rate of such prediction models. In this regard, Zhang et al.,(2012) considered 10 major economical principal components to investigate their effect on electricity consumption in China. They depended on two principal components with eigenvalues of 8.28 and 1.04 and cumulative variances of 82.77% and 93.19%, respectively. The results showed that the regression model with two PCs can more accurately predict the actual electricity consumption.

Researchers have succeeded in removing multicollinearity problem when they treat datasets with linear patterns using PCA in the MLR-based prediction model for electricity demand. Given that not all input datasets change linearly, researchers must use a nonlinear approach to capture nonlinear dataset patterns and PCA to reduce or remove the multicollinearity among such nonlinear variables.

2.5.2 Nonlinear Model

This section explains a specific type of nonlinear approach, which receives and manipulates the nonlinearity pattern of input dataset. The present study focuses on artificial neural networks (ANN) because it demonstrated the optimal nonlinear approach in various prediction studies. These studies have established that accurate prediction can be obtained if more parameters are included in the model. However, increasing the number of parameters may shift the patterns of input dataset from linear to nonlinear, resulting in increased errors at the output stage of linear prediction (Pao, 2006). Therefore, researchers started to change the techniques they employed from linear to non-linear, as shown in the second part of Table 2.1.

The commonly used techniques and tools in nonlinear models include ANN, fuzzy systems, SVM, and GA (Abdulalla et al.,2010).

University of Malaya

The work mentioned that ANN is the most commonly used tool by researchers because it can deal perfectly with nonlinear patterns of a dataset, this argument has been confirmed by Darbellay et al.(2000) and Pao,(2006).

ANN is used to reduce the error rates recorded in most linear models, such as PCR. Previous studies showed that the ANN model provides better prediction rates than linear models (Azadeh et al.,2008). ANN is also identified as the best nonlinear approach for quantitative prediction models (Aggarwal et al.,2009), as confirmed by Kavaklioglu et al.(2009). This section shows that how to improve ANN when the input dataset involves linear and nonlinear patterns, this research gap has been implicitly explored in several studies. For example, Zarezadeh et al.,(2008) employed ANN to predict electricity price in Iran. Input datasets are classified into warm and cold day-based records, with each group sub-divided based on low load, normal load, and peak load hours. With this grouping, the architectures were modeled for six ANN to effectively cover the scenario. The results showed that the MAPE values change between 0.58 percent and 3.09 percent, and comparison results indicated that ANN provide more accurate prediction than the MLR model. However, the MAPE value reveals a large distance between the actual and predicted ANN outputs. This large distance represents that some parts of the input dataset could not be captured by ANN during the training phase or ANN could not effectively learn from the input dataset. This scenario may revert to the linear patterns existing in some parts of the input dataset.

This ANN capability was confirmed through ANOVA by Azadeh et al.(2007), who forecasted electricity demand in Iran. In this regard, researchers from far-eastern, middle-eastern, and western countries have used an ANN to forecast electricity demand. A historical dataset of electricity consumption in Czech Republic was inputted to an ANN to predict short-term electricity demand (Darbellay et al.,2000). The researchers obtained

University of Malaya

a more accurate electricity demand rate when a nonlinear tool was used to analyze the nonlinear dataset. If the dataset contains linear patterns, nonlinear tools can generate some errors. However, the error percentage is lower than that when a linear tool is used to analyze a nonlinear dataset. Errors caused by evolving mixed (i.e., linear and nonlinear) dataset patterns have been evaluated using a nonlinear tool. Kandananond,(2011) employed an ANN to forecast electricity demand in Thailand by using three parameters: population, GDP, and consumer price index (CPI). Although two of these three factors present nonlinear patterns, the population factor was treated as a linear pattern.

Kermanshahi et al. (2002) used a back-propagation neural network and Jordan recurrent network to predict electricity demand for Japan by employing 9 factors such as population, GDP, GNP, number of households, CO2, index of industrial production, energy consumption, oil price and number of air-conditioners.

Ghanbari et al.(2009) showed that ANN can predict long-, medium-, and short-term electricity demands as it exhibits high root mean square error (RMSE) and mean absolute percentage error (MAPE); GDP and population are also considered the most significant factors on electricity demand rate in Iran at that time ANN was then applied to predict each independent variable and electricity demand. The MAPE results indicated that the ANN approach is more accurate for prediction model than another models.

Lu Yu (2015) presented the Support Vector Regression (SVR) for Short-term Load Forecasting (STLF) to predict electricity composition. The results indicate that the linear regression model with SVM is suitable combined model to predict electricity consumption. Furthermore, Geem et al.,(2009) identified two additional parameters, namely, export and import cost indicators, as significant in electricity demand rate in South Korea.

University of Malaya

Sackdara et al., (2010) employed the number of households and electricity price as nonlinear patterns to predict electricity demand rate in Thailand. They showed that ANN outperforms other nonlinear regression models, such as MLR. Another work performed by Kandananond (2011) in Thailand recognized other significant parameters, such as stock index and revenue from exporting industrial products.

An ANN comprising input factors of GDP, GNP, Iranian oil price, oil income, and CPI was proposed by Dalvand et al.(2008) to predict electricity loads in Iran. The network was trained using feed-forward back-propagation algorithm, and percentage error was used to evaluate the model. Although this work obtained a good accuracy, the multicollinearity among the input variables still need to be removed.

Ardakani et al.(2014) compared linear (MLR) and nonlinear approaches (ANN) at the level of input dataset. In this work, two different data sets (electrical energy consumption and socio-economic data) for two different countries were inputted to the two proposed approaches. The result showed that the use of socio-economic data set leads to more accurate electricity consumption prediction than that when the other data set was used. Moreover, changing the patterns of the input data set provides different accuracy rates. The effect of historical input data or the range of the time span of data was investigated by Fadare et al.(2009). This work developed a short term load prediction model for the Ibadan University by using five years dataset of peak load. This study verified the influence of short data and the stopping criterion proposed by Demuth et al.(2008) on the over-fitting status of the prediction model by using the coefficient technique (R2). The obtained R2 is 0.846, which indicates that a very good relation exists between the size of the dataset and the over-fitting status in ANN, the result showed that both techniques are affected by the size of the data set. A short time span of dataset was also adopted by Saber et al.(2009) to propose a model integrated with ANN and PSO to predict short-term load

University of Malaya

in Saudi Arabia. This model can predict the load by utilizing the data obtained for a utility company, but the results showed that the prediction rate is lower than the desired level of accuracy. Hence, PSO was adopted to improve predictions. This work depended on MAPE performance evaluation to check the result of the comparison. Another short-term load prediction by using ANN approach was proposed by Mosalman et al.(2011) for 1 day prediction. The prediction rate of the developed model was evaluated using the dataset that obtained from the power system of Yazd, and the MAPE obtained is 1.78%.

Jain et al.(2009) proposed a novel clustering-based ANN model, which was designed for a short-term load prediction by using 48 half-hourly loads. This model, which can predict loads for the next day, was trained with historical load and temperature data. The model performance was evaluated using average and maximum peak loads, and cluster and cluster-less ANN were compared. The results showed that the error percentage of ANN with clusters is better.

The preseason pattern of the target in clustering training does not change as much as that in prediction training. Mishra et al.(2009) also compared and analyzed WNN with Wilcoxon norm cost function and multi-layer perceptron neural network (MLPNN) with least mean square cost function. The compression confirmed that the short-term load-based BPNN and WNN for prediction are affected by the size of the data set.

A comparison work was conducted by Kankal et al.(2011) to predict electricity consumption in Turkey by using ANN and linear regression model. This work used demographic factor and socio-economic rates as independent variables, such as GDP rate, population factor, import and export factors, and employment rate. The models in this work were validated using relative errors and RMSE. The result showed that the proposed model more accurately predicted electricity consumption than regression models. Another work verified the excellent performance of nonlinear approaches compared with

University of Malaya

linear approaches (Azadeh et al.,2008) by employing ANN to predict annual consumption of electricity in high-energy consuming industries in Iran. The proposed ANN approach was based on the multilayer perception structure. This study showed that the ANN approach presents higher accuracy than regression models in predicting electricity consumption, as evaluated through ANOVA.

Another Multi-layer perceptron neural network (MLPNN) was applied to the Maharashtra State data (Tasre et al.,2011), with the maximum temperature factor and maximum load rate as independent variables inputted to the network. This work utilized MAPE to evaluate the performance indicator of the developed model. As the maximum temperature is significantly correlated with the maximum loads, the accuracy of the results is regarded unsatisfactory.

The patterns and sizes of input dataset are not the only issues that should be considered in investigating the accuracy of electricity demand prediction models. Removing the multicollinearity by reducing the size of the data by using MDA techniques should also be considered. The following studies showed the advantage of using PCA in improving the accuracy of electricity demand prediction models.

Zhang et al.(2009) used PCNN to predict long- and medium-term load electricity demands. However, this study showed that PCA not only reduces duplicated information, but also extracted the leading factors. This work also computed errors to evaluate model performance, and the result showed that the PCNN model is an effective algorithm to predict electricity demand.

Another work conducted by Kheirkhah et al.(2013) presented an approach using ANN, PCA, DEA, and ANOVA methods to evaluate and predict demand of electricity for monthly change and seasonal change in electricity consumption.

University of Malaya

Researchers have succeeded in overcoming the shortages of uncovering the nonlinear part that exists in input dataset by utilizing nonlinear approaches instead of linear approaches. The multicollinearity existing among the independent variables on input dataset has been successfully removed using MDA techniques, such as PCA. Nevertheless, the accuracy of prediction models must still be improved because nonlinear-based prediction approaches cannot capture the linear patterns of the input dataset. These un-captured patterns may cause few residual errors in the system model to accumulate, thus obtaining a high rate value. In the next sub-section, residual errors induced by the un-captured linear patterns of the input data set are addressed.

So far, studies mentioned in this section demonstrated how researchers have attempted to determine significant parameters with respect to prediction of electricity demand rate. New or modified parameters (e.g., the overall export rate of a country vs. the industrial export rate of a country) have been proposed in each work. These studies aimed to achieve an accurate prediction rate by using significant independent parameters. Nevertheless, involving several parameters increases the complexity of the input dataset (Popovic. 2013), thereby increasing error rates. Therefore, many studies have evaluated different prediction models as an alternative to ANN and developed several testing and comparison processes among different nonlinear prediction tools and models.

Yau et al.(2007) used three different tools such as regression analysis, decision tree, and ANN to predict electricity demand in Hong Kong. ANN, linear regression, and log linear regression were also employed by Ghanbari et al.(2009), whereas SVM was utilized as a comparatively strong nonlinear tool by Ekonomou (2010).

As an alternative to ANN, additional common tools, such as fuzzy systems, SVM, and GA, have been used by researchers as prediction models. Researchers believed that these alternative tools can be used to decrease the complexity of input datasets, which increases

University of Malaya

with increasing number of significant parameters included. Mohandes,(2002) employed SVM to build a short-term electricity prediction model and compared this model with an auto-regressive-based model. Chen et al.(2004) used SVM to build an electricity demand prediction model for Taiwan. Although these researchers concluded that a time-series concept, which can be easily found in the time-series neural network structure, can yield accurate results, they did not found any similarity between SVM and ANN.

Various studies have proposed GA as a nonlinear tool to construct an electricity prediction model (Amjadi et al.,2010) and (Ersel Canyurt et al.,2004). Parameters used in GA include GDP, population, gross national product (GNP), customer number, average electricity price, and export and import incomes. The fuzzy system is also another technique used in electricity prediction models of different studies (Azadeh et al.,2008) and (Kucukali et al.,2010). Most researchers used GA and fuzzy systems to enable prediction models to mimic human thinking and reasoning.

Lewis (1982) they employed artificial neural network (ANN) and a regression model to predict electricity consumption for long term in Thailand. This study employed GDP, Population, Maximum ambient temperature and electricity power demand as input for both models. The results prove that the ANN model can give more accurate predictions than multiply linear regression model.

Researchers on non-linear models believed that partitioning a dataset to decrease the number of factors that have to be simultaneously dealt with can decrease the complexity of that dataset, thereby enhancing the accuracy of the model. Therefore, many researchers used PCA on their dataset to simplify or reduce input data (Kheirkhah et al., 2013) and (Saravanan et al.,2012). In these studies, PCA can improve the accuracy if the model is combined with an ANN or any other non-linear prediction tools. However, Saravanan et al., (2012) reported that the PCA and ANN prediction models are more accurate than the

University of Malaya

PCA and MLR models and proposed the PCR and PCNN methods for prediction of long-term electricity demand. The paper used 11 factors that affect electricity demand in India and applied PCA to decrease multicollinearity among independent variables. The PCA variables were then used as a new dataset of independent variables in the MLR and ANN methods to predict electricity demand. The PCNN model is more effective than the PCR model based on RMSE, MAPE, and mean bias error (MBE) as performance indicators.

Other works utilized additional tools, such as data envelopment analysis (DEA) and ANOVA (Kheirkhah et al.,2013). Hamid et al.(2010) proposed an ANN-based prediction model that utilized an artificial immune system (AIS) as the learning algorithm. The developed model contains an input layer, a hidden layer, and an output layer. Historical dataset from Malaysia and North Carolina, USA were inputted to the model. A MAPE indicator and another training algorithm called back propagation (BP) were utilized to evaluate the performance of the developed model. As an improvement of 1.347 was obtained, thus, AIS can be replaced by the ANN algorithm for electric load forecasting models.

Niu et al.(2010) developed a BP neural network by using MATLAB and applied the model to data obtained from an undisclosed city power company. Data were fed from February 1 to April 30. The results showed that ANN-based prediction models present satisfactory predictive and generalization ability as evidenced by small errors detected in comparison of the forecasted and actual values of prediction rates.

Another ANN-based prediction model was proposed by Bhurtun et al.(2011) to predict electricity load in Mauritius by utilizing a feed-forward back-propagation training algorithm. This study evaluated errors by using performance indicators, namely, MBE, RMSE, and MAPE. The results implied that the ANN model is suitable for load forecasting. Tasre et al.(2012) presented an ANN incorporated with the BP algorithm to

University of Malaya

be applied in a small area network in India. The independent variables were historical load dataset and temperatures, and the latter was assessed as it fluctuates during the four seasons of the year. MAPE was also employed to estimate the performance of the developed model. The performance results obtained on monthly and annual bases are 1.987% and 4.291%, respectively. This study concluded that errors can be less if hourly load prediction can be estimated for demand-side management and security analysis through company utilities.

Ekonomou (2010) determined the long-term energy consumption of Greece by using the proposed ANN multilayer perception model. The selected independent variables for this model are annual ambient temperature, installed power capacity, yearly per resident electricity, consumption, and GDP. The work tested the proposed ANN, linear regression method, and SVM by using real forecasting records available from 2005 to 2008. The results indicated that ANN is superior to the other tested methods.

Milojkovic et al.(2012) proposed a peak load prediction method based on feed-forward ANN. This network contains a hidden layer in addition to the input and output layers. The records inputted to the proposed model were obtained from the UNITE 1997 file. The results showed that the average of prediction error was 0.14%. Overall, the study concluded that the proposed method is the only known tool that can accurately predict time-based peak load.

Another back-propagation ANN with a hidden layer was proposed by Kunwar et al.(2013). The developed model was tested using data obtained from the New Hampshire Electricity Corporation. Data were classified into three parts: 65% for training dataset, 20% for validation dataset, and 15% for testing dataset. The results showed that the proposed model improved compared with models based on fuzzy systems, SVM, and GA.

University of Malaya

Benaouda et al.(2006) used a wavelet-based nonlinear multi-scale decomposition model to forecast electricity load.

Two different nonlinear models of NN and decision tree were proposed by Moghaddam et al.(2014) to predict electricity demand for 1 and 7 days in the interconnected system of Southwest Australia. These researchers considered the maximum and minimum prediction of temperature and relative humidity as available future inputs. The result showed that the two different nonlinear models of NNs and decision tree properly fit to the models based on the modified MAPE. The general literature on ANN and other models demonstrate that the nonlinear model cannot capture the linear component of datasets (Zheng et al.,2011). However, Darbellay et al., (2000) reported that the accuracy of a nonlinear model with PCA, such as PCNN, is higher than that of linear, PCA-linear, and nonlinear models without PCA. Therefore, the next section discusses the use of the combination of two models (hybrid approach) to capture different dataset patterns.

Guo et al.(2014) proposed grey prediction (GP) and ARIMA to predict electricity consumption in Beijing, and the results indicated that the GP model has better accuracy prediction model than ARIMA model for electricity consumption. Wavelets can also be used for short-term load forecasting. A recent work by Aqeel and Butt,(2001) used Narx NN to predict annual electricity consumption in Iran by using population, GNP, import, and export as independent variables. The results were compared with those of the ARIMA model and Perceptron NN (PNN). Narx NN exhibits higher accuracy than ARIMA and PNN.

All these studies were conducted because researchers believed that the accuracy rate of electricity prediction models must be improved. Thus, the use of hybrid system has been introduced in electricity prediction models.

University of Malaya

2.5.3 Hybrid Models

Hybrid approach is a combination of two different sub-models or more than two sub-models. This combination is used in designing and building prediction models to improve the accuracy rate of predictions. To the best of our knowledge, Batesr et al., (1969) were the first to introduce a combination approach as an alternative to single prediction. The concept of combining predictors is to use the unique features of each model to capture different patterns or features in the dataset. Therefore, the rate of prediction accuracy can be improved by building a prediction model with more than a single sub-model compared with using an individual predictor (Clemen,1989); (Makridakis et al.,1982); (Makridakis et al.,1993) and (Ismaila,2011). The current literature reveals that a hybrid approach has been rarely employed in prediction models for electricity demand.

Hybrid systems are used to capture the linearity and nonlinearity patterns of input dataset. To the best of our knowledge, accuracy can be improved through the use of combination models because they can capture the linearity and nonlinearity patterns in the input dataset. Studies that proposed combination approaches are presented in this sub-section

Electricity demand prediction model based on a hybrid system comprises two or more techniques and tools to cover continuous (linear) and discrete (nonlinear) dataset patterns (Schaft et al.,2000). Studies shown in Table 2.1 differ from studies utilizing nonlinear models in two aspects: tools for a hybrid-based system are combined and work together as one unit and ability to cover linear and non-linear patterns of the input dataset. Both differences can be easily detected in these studies.

Yao et al.(2000) combined the wavelet transform (WT) analysis and ANN to forecast short-term electrical load in Brunel. These researchers used wavelet analysis to decompose the input data and transfer them to two different frequency types. The

University of Malaya

frequencies were then clamped to ANN to predict short-term electrical load. Another combination of WT and ANN was performed by Chen et al.(2010), who showed that a combination of tools (WTANN) provides a more accurate prediction rate than individual models, namely, ANN and wavelet transformation.

Nonlinear tools can be combined to produce hybrid systems. Song et al.,(2005) utilized hybrid model by using a fuzzy system and regression to forecast electricity load for South Korea and reported that the combination presents higher accuracy than the fuzzy–neural combination. Rahib H Abiyev (2009) proposed another fuzzy combination by using WT and NN to form a hybrid prediction system (FWTNN). The model was clamped with a complex time-series dataset, and this combination exhibits better performance than the FWT and FNN combination. Another hybrid combination (ANFIS) was proposed by Bazmia et al.(2012) to predict long-term electricity load for Malaysia. A new combination (SVM and regression) was proposed by Wang et al.(2009) for prediction of long-term electricity load in China. The same dataset group was used by these researchers to propose another model (ARIMA). The accuracy rates of SVM and regression are higher than the ARIMA model, which used adaptive particle swarm optimization (PSO).

Azadeh et al.(2008) proposed a combination method comprising regression sub-model and another sub-model to analyze electric demand/electric load for predicting the daily maximum electricity demand of large steel mills in Malaysia. The result showed that this combined model can accurately predict the daily maximum electricity demand of large steel mills.

Hanmandlu and Chauhan (2011) combined fuzzy and neural network to develop a wavelet fuzzy neural network (WFNN) and fuzzy neural network with Choquet integral. The records obtained from the utility company of Indian were used to the developed model. MAPE was used and compared with the performance of ANFIS to evaluate the

University of Malaya

developed model. The results showed that the developed model is more accurate than few conventional models.

Hasan et al.(2010) also proposed a new hybrid model called NN-PSO to resolve short-term load prediction. This model was designed for load prediction for weekdays, weekends, and holidays.

Choy et al.(2008) propose the combined MLR, ANN, and wavelet ANN to determine electricity consumption in the Asian gaming and tourism center in Macau SAR, China. Temperature, population, number of tourists, number of hotel room occupancy, and number of occupant days per month were used to characterize monthly electricity consumption in Macau. The performance indicators utilized to evaluate the accuracy of the model were MSE, MSPE, and MAPE. The wavelet ANN provided accurate results compared with the other two models. A new hybrid ANN predictor model, namely, SAPSO, was proposed by Raza & Baharudin (2012). The proposed ANN contains three layers, and the SAPSO training algorithm was used instead of BP. The proposed training algorithm is a combination of simulated annealing (SA) and practical swarm optimization (PSO). This paper concluded that the proposed model SAPSO neural network exhibits accurate load prediction and can solve convergence problems of conventional techniques.

Yao et al.(2000), which combined WT and ANN to obtain a model for short-term electrical load prediction in Brunel. WT analysis was used to decompose input data and transfer them to two different types of frequencies. These frequencies were then clamped to ANN to obtain a short-term electrical load prediction. Another work conducted by Azadeh et al.(2013) utilized PCA as the input variables for fuzzy regression model and time-series models, such as ARMA, to predict electricity consumption in Iran. This paper established a good model to provide less error in electricity demand prediction. ANOVA was further employed to compare the fuzzy regression and time-series models. The results

University of Malaya

indicated that the (FR) provides better prediction than the time-series model (ARIAM). Many nonlinear tools can be combined to create hybrid systems. Song et al.(2005) proposed a hybrid model by using a fuzzy system and regression to predict electricity load demand in South Korea. The combination of fuzzy and regression provides higher accuracy than the fuzzy and neural combination.

Rahib H Abiyev,(2009) proposed another fuzzy combination with wavelet transform and NN to form the FWTNN hybrid prediction system. The model was clamped with a complex time-series data set. The combination showed better performance than the FWT and FNN combination. Another hybrid combination (adaptive neural fuzzy system ANFIS) was proposed by Bazmia et al.(2012) to predict long-term electricity in Malaysia. A new combination (SVM and regression) was also proposed by Wang et al.(2009) to predict long-term electricity in China. These researchers used similar group of datasets and clamped these data in another proposed model (ARIMA). The accuracy of SVM and regression is higher than that of the ARIMA model, which used adaptive PSO. Pindoriya et al.(2010) combined adaptive wavelet neural network and feed-forward neurons to predict short-term loads. The result showed that the AWNN model exhibit higher accuracy than MLPNN.

2.6 Accuracy Related Components

Accuracy is the most important feature of prediction models. Many works have investigated and analyzed the rate of the accuracy with reference to the factors that have impact on. This study presents the influence of two important features on accuracy. The first feature is multicollinearity, which is related to the characteristic of the input dataset. The second feature is the errors that recorded due to the type of approach that proposed by researchers in designing and building their predictive models. The three commonly used approaches include linear, nonlinear, and hybrid models.

University of Malaya

2.6.1 Multicollinearity of Dataset

A dataset characteristic that affects the accuracy rate of electricity demand prediction models is the multicollinearity among independent variables. Multicollinearity problem presents a high degree of correlation (linear dependency) among several independent variables. In 1934, the term multicollinearity was first introduced by Zhou et al.,(2006). Multicollinearity occurs when a linear relationship exists among one or more of the independent variables (Bakheit et al.,2008). This problem also occurs when the original dataset is directly used for prediction or estimation (Saberi et al., 2007). The problem regularly happens when large number of independent variables are incorporated in a prediction model because an independent variable may measure the same concept or effect on the output rate of the model (Kheirkhah et al.,2013). Therefore, it causes generally unreliable predicted results (Asteriou & Hall,2011)

In general, the occurrence of multicollinearity during inclusion of independent variables in a model may result in complex estimation and prediction. This phenomenon could be due to insufficient independent information provided in the model by estimation factors for independent variables; hence, coefficients are inaccurately estimated and contain much uncertainties (Asteriou & Hall,2011). As a result, the standard error of estimation is very high and an accurate prediction model cannot be obtained (Azadeh et al.,2007). The occurrence of multicollinearity when explanatory variables exhibit high variation and correlation may results in a low accuracy of prediction (Azadeh et al.,2007).

In cases with highly correlated independent variables, the release of the separate effects of each explanatory variable on the explained variable becomes complicated (Azadeh, et al.,2007). Under this phenomenon, developing an accurate prediction model is difficult and dataset cannot be directly used in the regression model. Therefore,

University of Malaya

multicollinearity negatively affects the stability of the regression model (Gabriel et al., 2011).

Shalamu,(2009) reported that when MLR is used on several factors affecting electricity demand, the model may sufficiently fit with the dataset but may produce worse predictions on the new dataset. Figure 2.1 shows the different degrees of multicollinearity for independent variables (x1, x2, and x3). No collinearity degree means that independent variables exhibit a collinearity coefficient within the tolerant range (between 0–0.3). Low collinearity denotes that independent variables exhibit a collinearity coefficient over a certain range, without causing significant effects on the accuracy and reliability of the model. High or very high collinearity coefficients (over 0.7) represent a significant problem related to accuracy.

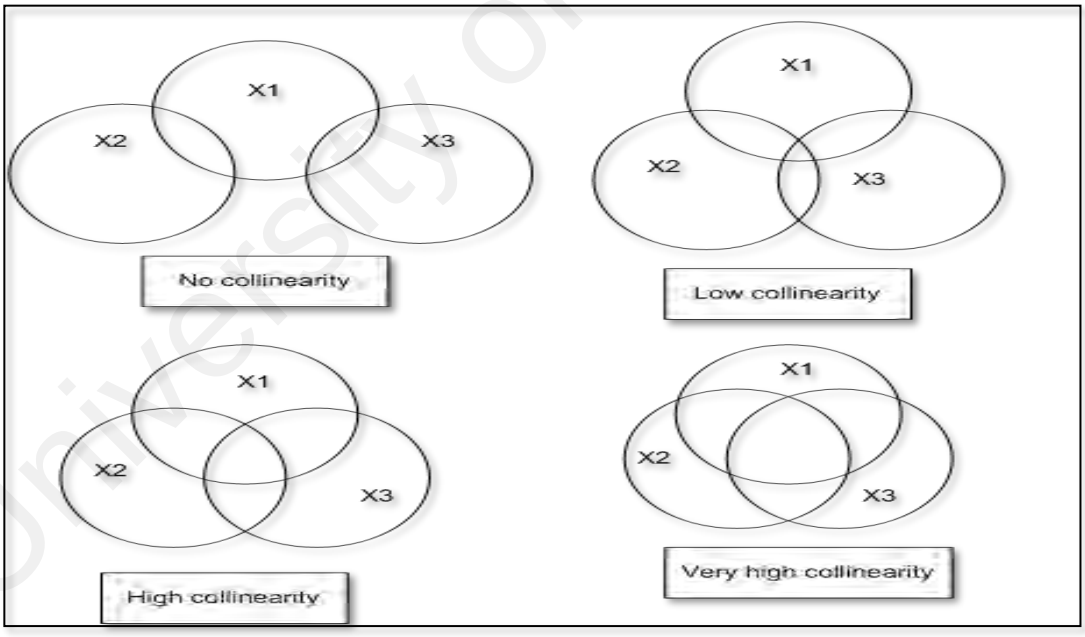


Figure 2.1: Collinearity in different degree for independent variables

Different methods could be used to determine whether collinearity exists among variables in a dataset or not. The present work used the value of the coefficient R2. Although coefficients are jointly significant and the R2 for the regression is quite high,

University of Malaya

coefficients exhibit very high standard errors and low significance levels when collinearity exists

The main question addressed by this work is reduction and minimization of collinearity coefficient. As the multicollinearity problem is related to high collinearity coefficient among few independent variables, it could be reduced or minimized using the following methods: 1) removing predictor factors related to irrational coefficients, 2) removing predictor factors by using stepwise regression model, 3) constructing composite indices as predictor factors, and 4) orthogonally transforming the correlation matrix to provide an equal number of uncorrelated dataset and inter-correlated dataset (Garen,1992) and (McCuen,1985).

Elimination of predictor variables may not be an efficient solution as the information obtained from very few sites may not spatially represent the variable information of the basin. Although constructing composite indices can remove the major source of inter-correlation of predictor variables, these indices are usually determined without considering regression and thus may not be statistically optimal for prediction (Garen,1992). The current Z-score method used in electricity consumption may be appropriate to minimize the collinearity problem. The weightings used in this method (Z-score) are computed based on correlations with dependent variables. However, in this approach, inter-correlations among independent variables remain unknown (Azadeh et al.,2008) and (Azadeh et al.,2009). Limited studies have used orthogonal transformation on original independent variables (highly correlated) to provide uncorrelated independent variables by removing multicollinearity. The reduced rank regression (RR) and PCR are adopted in different studies to reduce multicollinearity.

In RR, which was first introduced by Yona et al.(2010), a constant τ is added to the variances of explanatory variables before solving normal equations. Although this

University of Malaya

method is used to explain factors as much response factor variation as possible, it does not provide an accurate prediction model (Tobias,1995). By contrast, PCR is used to explain as much prediction factor variation as possible but may not be related to the response of variation factor (Tobias,1995). PCA is a well-known method with inherent ability to identify multicollinearity (Hocking,1976). To improve the accuracy of prediction models, researchers have used PCR to determine load demand. The result demonstrated that the PCR model is more accurate than other linear models when utilized for a prediction model because it can remove multicollinearity (Sousa et al., 2007).

2.6.2 Errors in the Models

This section presents the relation between the group of the utilized model and the type of errors that possibility a model will fall in. This section also shows the relation between the characteristics of input factors (dataset) and the accuracy of electricity prediction models. It also presents the argument of errors, which probably occurred while processing the nonlinear patterns of the dataset by using a linear method.

To achieve that, three different datasets are considered; linear, nonlinear and mixed patterns. Figure 2.2 shows the relation between an independent variable (X) with a dependent variable (Y), a linear fitting line, and an error case. When the linear fitting line cannot pass through all existing cases, an error (E1) will occur, which is similar to the other cases. The accuracy of this model could be evaluated using an error evaluation method (such RMSE). In an ideal case, the linear fitting line or equation should be drawn as close as possible to all cases. In this scenario, errors are minimized and thus the accuracy is set at a high level.

The next type of error is expected when linear data are inputted to a nonlinear method. Figure 2.2 shows this error situation. The figure shows the scattering of points that are almost linearly distributed and an error case. A nonlinear fitting equation is used to

University of Malaya

express the regression. The error discussed in Figure 2.2 is further explained, but using a different situation. In this case, the curve could not pass through all cases perfectly. The distance between the curve and the scattering case is the weight of the error (E1).

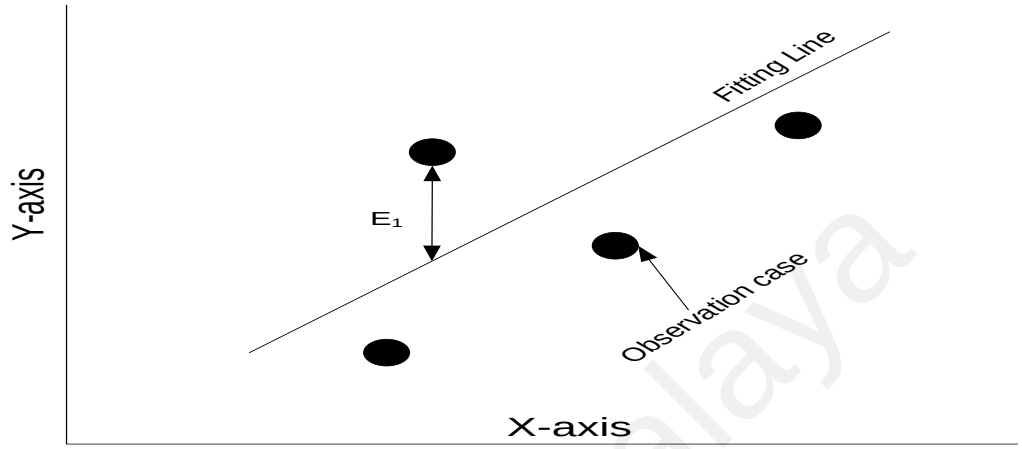


Figure 2.2: Case (A) of error due to nonlinear patters of scattering and linear fitting line

In Figure 2.3, instead of the curve, a linear fitting line (dash line) is used as a fitting equation. The value or the weight of error will be almost less than the weight that probably occurred with the curve. As a result, a high accuracy could be obtained.

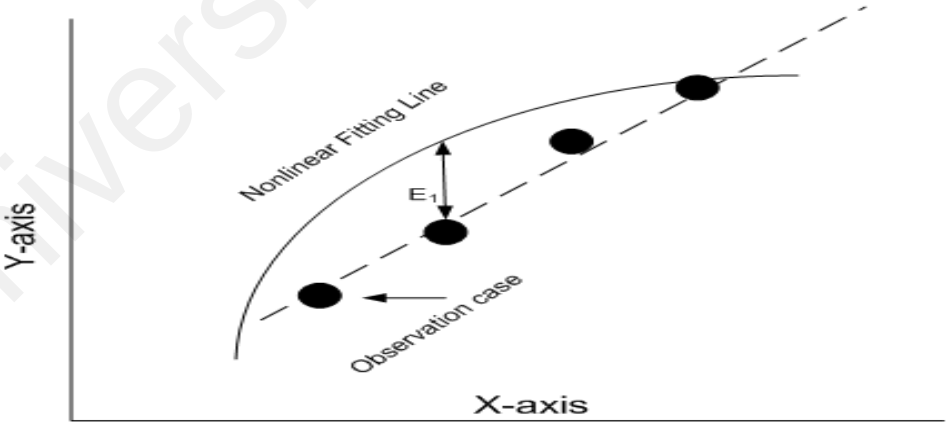


Figure 2.3: Case (B) of error due to linear patters of scattering and nonlinear fitting line

Most electricity demand prediction models use a dataset with variables of two different patterns. In such cases, capturing all data by using linear methods or nonlinear methods alone is difficult. The proposed solution to address this problem is to use hybrid systems

University of Malaya

as prediction models (Hanmandlu et al., 2011). Figure 2.4 illustrates the schematic of such method. The figure shows a few scattering cases of the relationship between an independent variable (X) and a dependent variable (Y). A case is considered to explain how linear and nonlinear fitting line will result in error while obtaining the value of this relation (E1 and E2). A similar case is considered for the hybrid system line. The figure clearly shows how the hybrid fitting line can pass through the selected case perfectly. In the first phase, the hybrid system uses the linear fitting line to determine the relation between independent and dependent variables. Errors are then computed. The function of the nonlinear fitting is used in the next phase, which is calculation of residual errors for all cases. Mathematically, a residual error is the summation of all errors occurring over all cases. Finally, the summation of linear fitting prediction with the value of the residual errors will present the output of the hybrid prediction model.

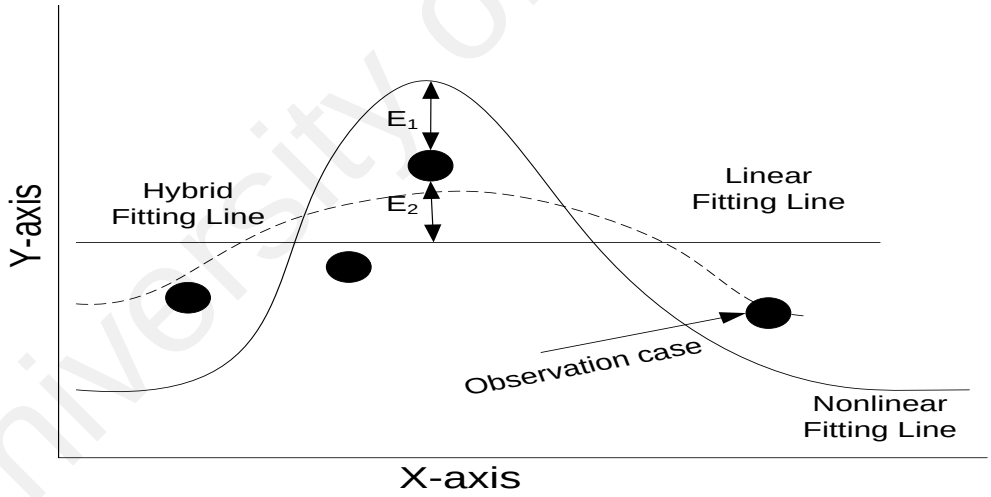


Figure 2.4: Case (C) Hybrid fitting line VS linear and nonlinear fitting line error due to linear patters of scattering and nonlinear fitting line

2.6.3 Performance Indicators

The most important part of designing an electricity prediction model is error rate evaluation to measure the accuracy of the proposed model. Researchers always used something called performance indictors to measure the accuracy of their proposed prediction models. The literature review (section 2.5) shows that root mean square error

University of Malaya

(RMSE), mean absolute percentage error (MAPE), mean bias error (MBE), mean absolute deviation (MAD), and prediction accuracy (PA) are the most popular statistical techniques used as performance indicators. Each tool is used to evaluate errors in a different direction.

RMSE and MAPE are employed to estimate the difference between actual and predicted values and explain the divergence or distance between the fitting equation and each tested case. MBE is then used to determine whether the proposed model performs underestimation or overestimation, which illustrate whether the results fit well, generate more waste, or are damaging for the public and economy (Saravanan et al.,2012). MAD and MSE performance indicators are generally used to measure the average magnitude of prediction errors (Zhou et al.,2006). Finally, PA is used to evaluate each model (Ramli et al.,2011). Several tests, such as ANOVA, MANOVA, paired t-test, and Wilcoxson signed-rank test, are used to assess the significance of the proposed model and determine the optimal structure for electricity demand prediction (Sohrabkhani et al.,2007).

The current study used a historical dataset to determine the appropriate prediction model by using MSE (Zhou et al.,2006), RMSE (Saravanan et al.,2012), and MAPE (Kandananond,2011) as measures to justify the suitability of the model. The formula for these indicators can be found in Section 3.6.

2.7 Summary

This chapter presented briefly the contents of more than hundred works that were done previously for predicting the electricity demand rate of some areas. The chapter discussed the works in the direction of utilized tools, the type of factors or independent variables that involved, and the performance indicators that used to validate their works. Another important direction that followed by this work is discussing the previous works in viewpoint of linearity and nonlinearity of input dataset and the ability of the tools and

University of Malaya

techniques for processing such patterns. Characteristics of input dataset, such as multicollinearity, were another point that addressed in Chapter Two. Hence the selection of the method in this study is based on the review of the literature done in this chapter, that takes care of linear and nonlinear pattern of the input data. At the same time, the method selected is able to take into account all relevent input data and used them for prediction. Previously researchers had to exclude some relevant parameters due to multicollinearity problem and the method used cannot handle large input data. The present review also noted that researchers avoid the inclusion of many independent parameters to reduce the complexity of input dataset as increasing the number of independent parameters negatively affects the performance indicators.

The present study also aims to improve the accuracy of predicting electricity demand rate. Previously, researchers minimized errors recorded at the output stage by testing different prediction tools and algorithms. In this study, the complexity of input dataset is reduced by measuring and improving multicollinearity between independent variables and this procedure positively affects all electricity prediction models.

The outcomes of this chapter are:

1. According to best of our knowledge, no study has been conducted using the principal components regression with back-propagation artificial neural networks (PCR-BPNN) combination to improve the accuracy and reliability of prediction models for electricity demand.
2. To the best of our knowledge, no study has determined the taxonomy of prediction models for electricity demand depending on patterns of input dataset.
3. To the best of our knowledge, no study has explored the relation between the characteristics of input dataset and the accuracy rate of prediction models for electricity demand.

University of Malaya

To address all these outcomes, chapter three presents the techniques and tools that proposed by this work as a work methodology, to explain how the above mentioned gaps could be solved.

University of Malaya

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter presents the orientation of the study to design, build, and implement the required prediction model, which namely PCR-BPNN approach. It begins with presenting the main components of the study's methodology in a sequential form. For each component, the chapter defines the input variables and the required output. The study is going to narrowing down each main component to some actions and activities that occur throughout the study's process.

Through giving details and mathematical expressions, the chapter explains the functionality of each main component, which thoroughly, inputs to a block will be mapped to the required output of the same block. To achieve these functionalities, the study depends on some linear and nonlinear techniques such as multi-regression and BPNN. The techniques have been fed with an input data set that collected and prepared by this study. The data set preparation covers many process starts with defining the expecting features of the model, then selecting the most significant ones through selecting feature and reducing collinearity processes.

Finally, the chapter presents the process of testing and validating the proposed model.

3.2 Framework of Methodology

The approach of this study relates to the accuracy of electricity demand prediction model that effects by the linearity and nonlinearity characteristics of the input data set when analyzed. As shown in the Figure 3.1, the methodology of this approach comes in two main parts; data collection part and techniques combination part. The first step is gathering information for identifying independent variables that have strong relations with PEDM model. Then using one statistical technique to remove multicollinearity

University of Malaya

problem among independent variables. A clean data set means a new independent uncorrelated dataset. The part two is finding out the tools and algorithms that could be utilized for building PEDM models. Later step is to clamping input dataset and testing whether the designed model can give better result or not. Testing PEDM models in terms of result can be done by computing the rate of errors then accuracy rate, and then gives the best accuracy. As shown in the Figure 3.1 , each main step has many sequential activities.

In most cases of building a prediction model, the proposed techniques and algorithms cannot fit the linearity and nonlinearity of the input dataset, and cannot reduce problems that affects accuracy rate due to existing multicollinearity among the selected features. Therefore, this study is proposing the PCR-BPNN model that has ability to answer the main questions and objectives targeted by this work. The questions that this study can answer is related to building a combination model that can minimiz the linearity/nonlinearity and the multicollinearity characteristics of input data sets from a side, and can improve the accuracy rate of the electricity demand prediction models in another side. The reliability proposed model has been tested and validated in three different environments; Malaysia, Turkish, and Sweden. The next section starts with data collection part. Later chapter will explain and show the ability of this model.

University of Malaya

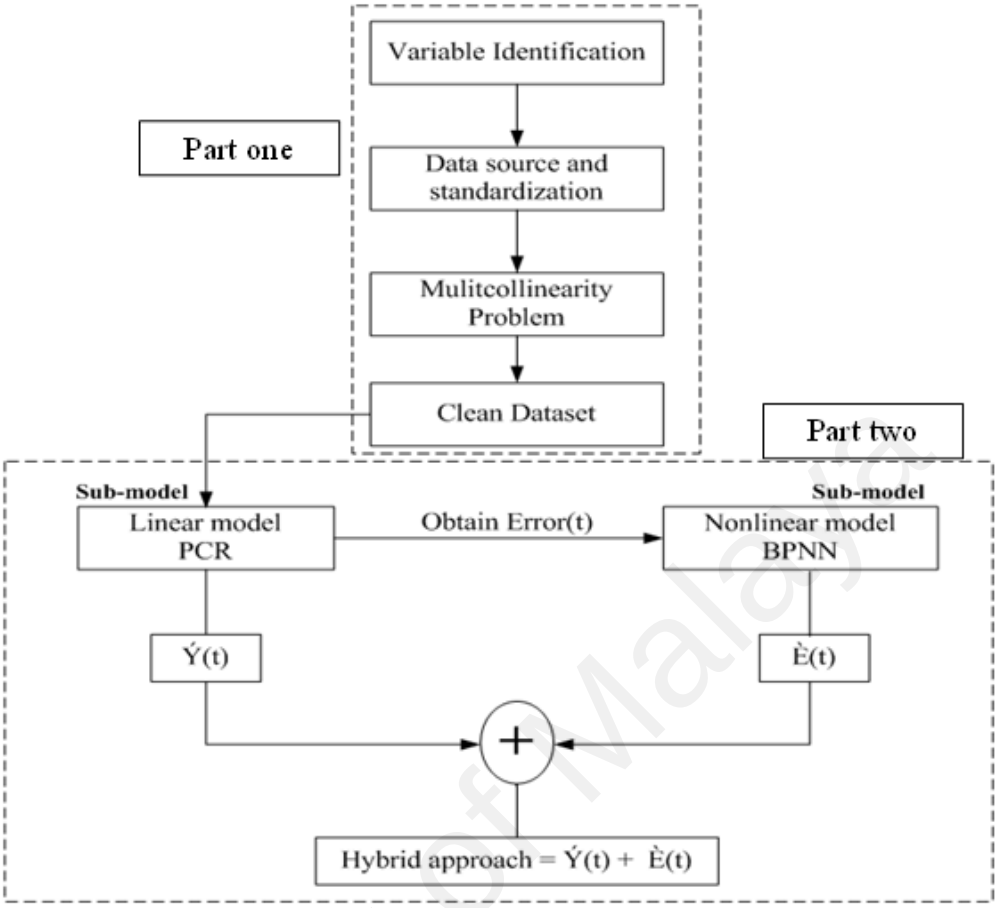


Figure 3.1: Structure of methodology

$\widehat{Y}\,(t)$: *PPEDM model which provided from linear model* $-$ *PCR model*

$\widehat{E}\,(t)$: *the output of nonlinear model which BPNN model*

3.3 Data Preparation

This section focuses on sources of data, variable selection and standardization of the data. To collect data, in general, previous works have collected their data from two different sources; either questionnaires (primary data) or historical (secondary data). Because most reviewed studies of electricity demand prediction models depended on historical data, therefore this study collected the required data through the same source. Historical data for electricity consumption always shows the patterns of consuming electricity for a specific country or a specific area. Although it needs less time and efforts,

University of Malaya

collecting historical data still has some disadvantages such as missing data and necessity to add new features to the model. Another reason for using historical data in this study is the range of the prediction; long-term, medium-term and short-term. Short term data is defined as data collected between 6 months to 2 years, medium term is between 3 to 5 years and long term is more than 5 years (Zhang et al.2009).

In most cases, long-term prediction models should depend on historical data with at least ten years of data. This main part of the model contains a sequence of sections that starts with variable or feature identification for the model and ends by providing without clean dataset (no multicollinearity) set to the study.

3.3.1 Variable Identification

Based on the works that were done previously in the field of electricity demand estimation or prediction model (Section 2.3 and Table 2.1 in Section 2.4), 19 variables were identified as commonly used and are significant in predicting electricity demand. Table 3.1 presents the summary of the selected 19 variables.

Table 3.1: Significant variables that obtained through previous works

#	**Variable Name**	**Authors**	**Area**
1	Population	Asafy J., 2000	Asian Developed countries
2	GDP	Nasr et al., 2000	Lebanon
3	Income per capita	Asafy J. 2000	Asian Developed countries
4	Humidity	Badar & Nasr, 2000	Lebanon
5	Temperature	Badar & Nasr, 2000	Lebanon
6	GNP	Ghosh, 2002	Indian
7	Industrial sector	Ediger & Talidth, 2002	Turkey
8	AGDP	Jumb, 2004	Malawi
9	NGDP	Jumb, 2004	Malawi
10	Price of electricity	Holtedahl &joutz, 2004	Taiwan
11	Degree of urbanization	Holtedahl &joutz, 2004	Taiwan
12	Number of Employers	Narayan&Smyth,2005a	Australia
13	Consumer Price Index	Pao,2006	ASEAN
14	Emission CO2	Al-Ghardoor et al. 2008	Jordanian

University of Malaya

Continuous Table

15	Residential sector Electricity demand	Lam et al. 2008	Hong Kong
16	Amount of export for a country	Zhang et al. , 2012	China
17	Amount of import for a country	Zhang et al. , 2012	China
18	Number of Tourists per year	T. Lai et al., 2008	China
19	Number of unemployed	Kavaklioglu, 2011	Turkey

In the next section, the source identification for getting data is explained.

3.3.2 Data Source

To collect required data on the electricity demand rates and records about significant independent variables, in general, and to get records on Malaysia specifically, this work consulted the Department of Statistic in Malaysia. This study also collected data for Sweden and Turkey to validate the model that proposed for Malaysia. Data for Sweden and Turkey have been collected from Statistics Sweden at (Sweden. 2015) and (Turkish, 2015).

The period of the records obtained from the mentioned departments covers the years 1995 to 2013. The records obtained for each factor (i.e., variable) are not in the same format. Some factors are derived on a yearly base and others on a quarterly base. For this study, all records and data should be in the quarterly format. To do that, this work uses Chow Lin method as discussed in section 3.3.4. Next section is about the process of selecting significant variables for the proposed model.

3.3.3 Selection of Variables

In this section, selection of variables to be included into further analyses is carried out by investigating the correlation coefficient of each variable with the electricity demand rate

University of Malaya

The correlation coefficient of the independent variables is very important because it indicates how strong the linear correlation among the said variables is. Several types of correlation coefficients exist; a common type of correlation coefficient is the Pearson product moment correlation used in linear regression, which is given by

$$R_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} - S_{YY}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

$$S_{XY} = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

$$S_{xx} = \sqrt{(X_i - \bar{X})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.3)$$

$$S_{YY} = \sqrt{(Y_i - \bar{Y})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.4)$$

Where

S_{XY} is the standard deviation of the covariance X and Y

S_{XX} is the standard deviation of the variable X

S_{YY} is the standard deviation of the variableY

X_i *and* Y_i are the pairs of measurements, in this case the independent variables

$\bar{X}$ *and* $\bar{Y}$ *are the sample means*

The range of the values of the correlation coefficients is expressed from the +1 to –1. The value of +1 shows a perfect positive correlation, which means the related variables have the same direction of changes. By contrast, the coefficient of –1 shows a perfect

University of Malaya

negative correlation, which means that the related variables have an opposite direction of changes. Degrees of correlation that are far from –1 and +1 are expressed as non-zero decimals. A coefficient of zero shows no discernible relationship between the variables of the fluctuations.

3.3.4 Chow Lin Method

After selecting the significant independent variables, time span format of these variables are not coming similarly. The dissimilarity of the data set comes when some records based on annual format and other variables have collected based on quarter format will be found in a set. Therefore, it is necessary to put them all in the same format through some statistical methods, such as Chow Lin.

The Chow Lin method is a basic method of estimating a quarterly dataset from an annual series of dataset, considering that some datasets are collected on a quarterly basis and others on an annual basis.

Many studies have used the Chow Lin method in converting annual datasets into quarterly datasets. Lahari, Haug, and Garces-Ozanne (2011) used the Chow Lin method to convert an annual GDP dataset to a quarterly GDP dataset. The process involved in 1971 Chow Lin method is described below.

Let *m* be the observations in the quarterly series of independent variables, where Y_q is related to the *m* observations on the independent variables-related variables, Xq, is based on a regression in the form of equation 3.5.

$$\boldsymbol{Y_q = X_q \beta_q + \breve{u}_q} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.5)$$

University of Malaya

Where Y_q is (m × 1) and X_q is (m × k). The error term follows a stationary first-order auto-regression $u_{q,t} = p_q u_{q,t-1} + e_{q,t}$ for t = 1....m, with $e_{q,t}$ having zero mean and a covariance matrix of $\sigma^2 I_m$.

The Chow and Lin (1971) equation disaggregates n annual independent variables estimates to 4n = m. Quarterly estimates are expressed as Equation 3.6.

$$\widehat{Y}_q = X_q \hat{\beta}_a + V_q C'(CV_q C')^{-1} \hat{u}_a \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.6)$$

where $\hat{\beta}_a$ is estimated using Equation 3.7:

$$\hat{\beta}_a = [X_q' C' \, (CV_q C')^{-1} C X_q]^{-1} X_q' C' (CV_q C')^{-1} Y_a \dots\dots\dots\dots\dots (3.7)$$

The (4n x 1) vector of the disaggregated quarterly independent variables estimates is represented by $\widehat{Y}_q \; and \; X_q$ is a (4n x k) matrix of k predictors, excluding the constant term. $\hat{\beta}_a$ is a (k x 1) vector of generalized least squares (GLS) estimates derived from the annual data. V_q is the covariance matrix (4n x 4n) of the quarterly error; $u_{q,t}$ and $\hat{u}_a = Y_a - X_a \hat{\beta}_a$ is (n x 1) vector of residuals from an annual regression of independent on predictor variables ($X_a = CX_q$). C is the (n x 4n) averaging matrix if multiplied by 0.25 or an aggregation matrix as presented in matrix below and Y_a represents the n x 1 vector of annual independents figures.

$$C = \begin{bmatrix} 1\;1\;1\;1 & 0\;0\;0\;0 \dots\dots\dots\dots\dots\dots\dots & 0\;0\;0\;0 \\ 0\;0\;0\;0 & 1\;1\;1\;1 \dots\dots\dots\dots\dots\dots\dots & 0\;0\;0\;0 \\ 0\;0\;0\;0 & 0\;0\;0\;0 \dots\dots\dots\dots\dots\dots\dots & 1\;1\;1\;1 \end{bmatrix}$$

The next section discusses the standardization of the independent and dependent variables.

University of Malaya

3.3.5 Data Standardization

The aim of standardizing a dataset is to achieve maximum compatibility and an optimum degree of parameters through data pre-processing. Standardizing datasets is important when dealing with variables of different scales and units. Many models, especially nonlinear models, require data standardization. For example Mashudi,(2001) standardized all datasets prior to using input variables in the ANN model, given that the best range of dataset for the said model is from –1 to 1.

The present study standardizes two datasets, which consists of the dependent and independent variables. Function (3.8) is more commonly used in standardizing independent variables, following, (Zhang et al.,2012), while Tobias,(1995) used Function (3.9) in standardizing common dependent variables.

$$\acute{x} = \frac{x_i - \bar{x}}{\sigma^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

$$\check{\mathrm{Y}} = \frac{\mathrm{Y_i} - \overline{\mathrm{Y}}}{\sigma^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

In equation (3.8), $\acute{x}$ is the result of standardizing all the independent variables, $\mathrm{x_i}$ is the mean of the original independent variables, $\bar{\mathrm{x}}$ is the mean of the each independent variable, and σ^2 is a standard deviation of the original dataset.

In equation (3.9), $\check{Y}$ is the result of the standardized dependent variable, Y_i is an original dataset, $\bar{Y}$ is the mean of the dependent variables, and σ^2 is the standard deviation of the independent variables. The correlation coefficient of the independent variables must also be determined to determine the multicollinearity problem from the independent variables.

University of Malaya

3.3.6 Dataset Characteristic Problems

The two main problems that relevant to the input dataset characteristics are linearity-nonlinearity patterns and multicollinearity problems. These two problems are addressed by this work in reference to the accuracy of the electricity demand prediction models and the solutions for these problems are presented in Section 3.5.

3.4 Linearity – Nonlinearity Problem

Based on previous research, the availability of dissimilar patterns inside input datasets directly affects the performance indicators and the accuracy rates of electricity demand prediction models (Saravanan et al., 2012) and (Zuhaimy, 2011). The reason behind this effect is also going back to the dimensionality of the input dataset. Increasing the dimension size of an input dataset increases these dissimilarity (linearity and nonlinearity) patterns, which in turn negatively affects the accuracy rate of the prediction models. Increasing this dimensionality increases the possibility of an input dataset to present more than two different patterns (i.e., linear and nonlinear), which makes it difficult for the electricity demand prediction models to capture both. Figure 3.2 shows an example for the change in the patterns of data and records within a variable.

University of Malaya

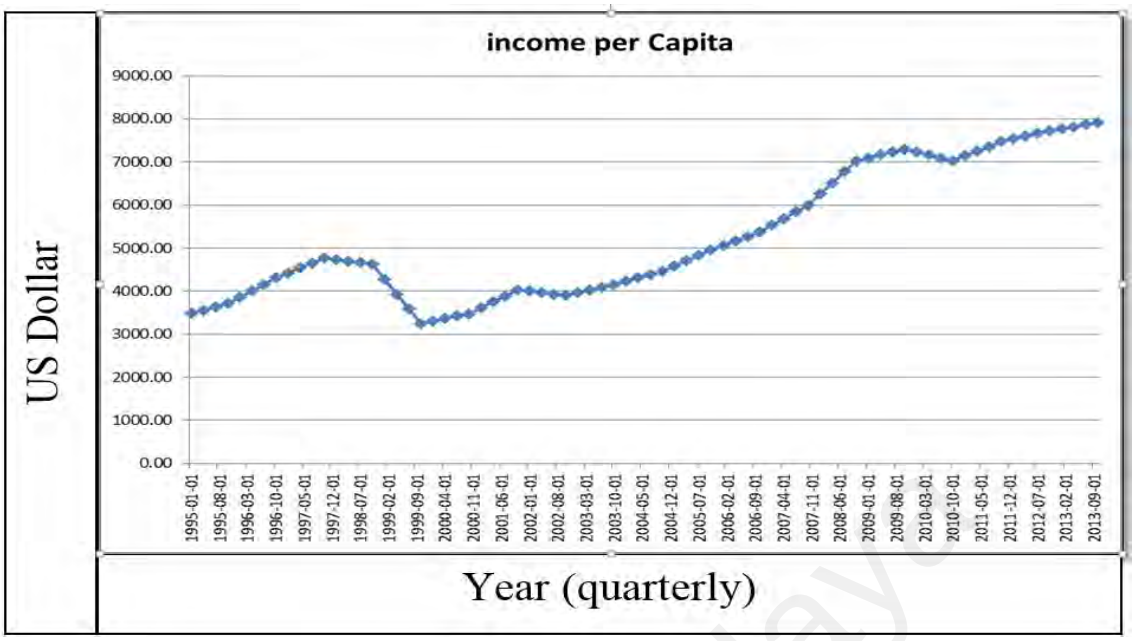


Figure 3.2: Change in the pattern of a variable over the time

The figure shows the pattern change in income per capita variable over the time span of 1995 to 2013, which the indicated quarterly dataset for income per capita. The patterns of the data have not been changed purely linearly or nonlinearly. The patterns have been changed differently. For example, data from 1996Q1 to 2000Q3 records nonlinear changes. However, from 2003Q2 to 2007Q4 the change becomes linear. Nonlinearity has been found again between the years of 2009Q1 to 2011Q4. Therefore, linear or nonlinear approaches cannot cover such mix patterns. A hybrid approach should be used so that the impact of both patterns on electricity demand can be showed.

To test this argument, this work tested three different prediction models (i.e., linear, nonlinear, and hybrid models) with three different sizes of input datasets. Appendix F shows the results of these argument tests. To overcome this problem, this work proposes a novel hybrid approach.

University of Malaya

3.4.1 Multicollinearity Problem

Multicollinearity causes unreliable predicted results. Therefore, it should be minimized as much as possible. The process is to minimize the correlation among independent variables, if exists. To check the availability of the multicollinearity inside the collected data set, this work followed the same process that mentioned in the section (3.3.3) through using the equation 3.1. The equation that used by this study depends on Pearson correlation coefficient, which can determine the correlation coefficient among the independent variables. The summarized of correlation coefficient of the independent variables shows in the Table 4.2.

The high correlation among the independent variables was caused by the multicollinearity problem for the independent variables. In a perfect multicollinearity problem, no methods can provide unique estimates for any independent variable, which also often results in correctible mistakes (Asteriou & Hall,2011), meaning that one can be linearly predicted from the others, with a non-trivial degree of accuracy. Therefore, a variety of techniques of removing/reducing the multicollinearity problem of the independent variables, such as factor analysis, PCA, and Ridge Regression (RR), must be used. This study uses the PCA statistic technique to reduce the multicollinearity problem of the independent variables.

3.5 Combination of Models

The present study combines Principal Component Regression (PCR) and Back Propagation Neural Network (BPNN) as a novel hybrid system for predicting electricity demand. PCR is a combination between PCA and MLR model. The results from PCA are used as input to MLR. The (PCR) part represents the linear sub model while the BPNN part represents the nonlinear sub model. The combination of these two sub models formed PCR-BPNN. Each sub model has its own output, and the idea of PCR-BPNN is

University of Malaya

combining these two outputs together in order to cover both problems; linear and nonlinear patterns with multicollinearity of input dataset in the prediction process.

Figure 3.3, which is taken from Figure 3.1, shows the architecture of this sub models combination.

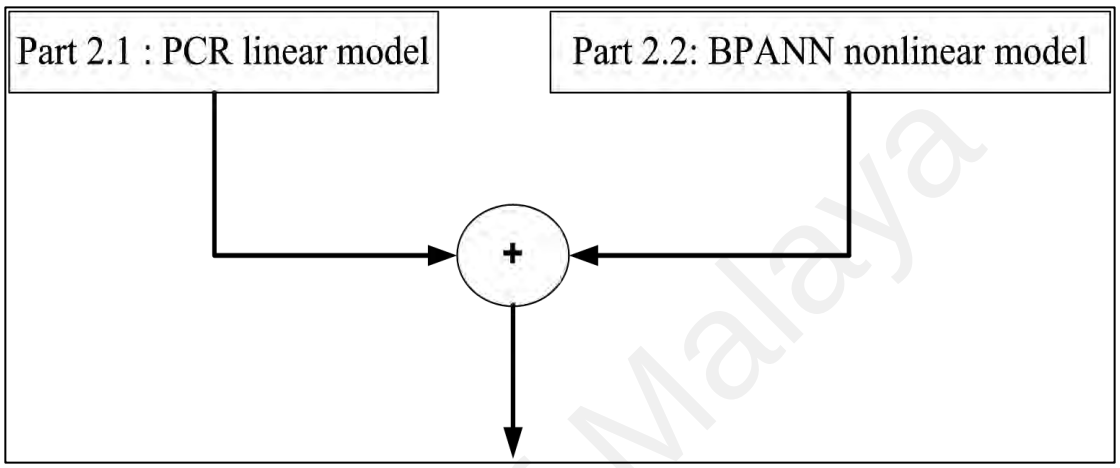


Figure 3.3: Processes the hybrid system PCR-BPNN

As mentioned before, the idea of combining PCR and BPNN is to capture differences in the pattern dataset. This is because, the PCR sub model (Part 2.1 in the Figure 3.3), can be used to analyze the linear pattern and solve multicollinearity problem of a dataset. However, the BPNN sub model (Part 2.2 in the Figure of 3.3 is developed to model the residual errors from the PCR model. The PCR model cannot capture the pattern of the nonlinear dataset, and the residual errors of the linear model contain information on the pattern of a nonlinear dataset. The results from the BPNN model can be employed as a prediction model of the residual error terms of the PCR model. The contribution of the hybrid approach can be determined for datasets with different patterns, and the overall combined predictions can improve the modeling and performance measurement indicators.

The process of building the PCR-BPNN as hybrid approach has been explained in detail through sections 3.5.1 to 3.5.3.

University of Malaya

3.5.1 PCR Linear Sub-Model

The basic idea behind PCR is to calculate the principle component (PC) and then use some of these components as predictors in linear regression to form PCR. The aim of applying PCA in the MLR model is to reduce the multicollinearity problem among the independent variables. However, the PCA technique can help determine the relevant independent variables for the preliminary prediction electricity demand model (PPEDM). In addition, to discriminate significance independent variables, this study have been using correlation coefficient between independent variables and dependent variable. Figure 3.4 shows the architecture of the PCA application based on the MLR model and details on each steps and blocks of the figure are explained in sections 3.5.1.1 to 3.5.1.4.

In short, only independent variables (input data) are included in applying the PCA. The results from PCA are then used as independent variables in determining the electricity demand (dependent variable) using MLR (sub-model).

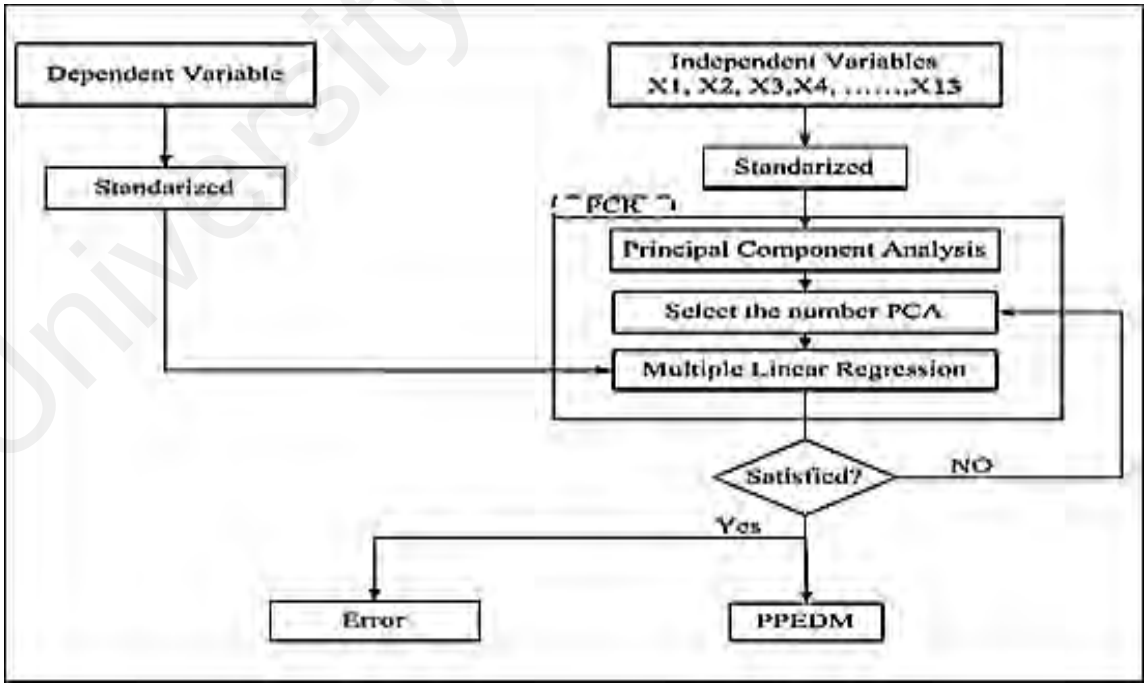


Figure 3.4: PCA-MLR combination (Multicollinearity reduction)

University of Malaya

3.5.1.1 Principal Component Analysis

This study uses the PCA statistical technique to reduce the multicollinearity problem of the independent variables. PCA is a multivariate statistical technique that linearly transforms the original dataset of ***n*** variables, which are correlated to the independent variables, into a new small dataset with ***n*** number of uncorrelated Principal Components (PCs). It also transforms the number of variables, particularly when the number is too large such that the size of the problem itself becomes quite unmanageable in many realistic situations. However, selecting the optimal number of uncorrelated variables should represent most of the information in the original dataset of the independent variables (Valle, 1999) and (Dray,2008). The goal of PCA is to reduce the multiple dimensions associated with MLRs, which create new parameters called PC, which are orthogonal and uncorrelated to one another. Analyzing a set with a small uncorrelated variable size is easier than analyzing a large set of correlated variables (Garen, 1992). According to Azadeh et al (2009) the advantage of Principal component analysis is reducing the number of dimensions without losing a much information of independent variables.

PCs are sequenced from the highest variance to the lowest variance. The first PC provides the highest amount of variance in the dataset. The second PC is larger than the third PC, and so on (Al-Alawi et al.,2005),(Wang et al., 2004) and (Sousa et al.,2007). Meanwhile, to determine the number of PCs that are relevant, the percentage of cumulative variance should be provided because it contains the most information from the original dataset (Saravanan et al.,2012). However, the quantitative based PCA describes the degree of interdependency of correlated parameters in a dataset (Zwiers et al., 1999)..

University of Malaya

PC components, which are the new variables from the PCA, can be considered perfect predictors in a regression model because they optimize space patterns and remove possible complexities resulting from multicollinearity (Mashudi,2001).

Eigen analysis is the important mathematical technique in PCA to find Eigen values and eigenvector. PCA uses eigenvalues and eigenvectors to solve a square symmetric matrix, sums of squares, and cross products. Eigenvectors associated with the largest eigenvalues have the same orientation as the first PC. Eigenvectors associated with the second largest eigenvalues define the direction of the second PC. The sum of the eigenvalues is equal to the trace of a square matrix, while the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

We let A be a square matrix if δ is a scalar and X is a non-zero; as such,

$$AX = \delta X \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

Where:

$X : is\ an\ eigenvector\ of\ A$

$\delta : is\ an\ eigenvalue\ of\ A$

Therefore, eigenvectors are possible only for square matrices; δ is an eigenvalue of a $n * n$ matrix, and A with corresponding eigenvector X.

$$(|A - \delta|X = o\ , with\ X \neq 0\ \ lead\ to\ |A - \ \delta I| = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.11)$$

At the most, *n* distinct eigenvalues of *A* exist.

University of Malaya

The PCA transforms the ("P") original correlated variables into ("P") uncorrelated components. These components are linear functions of the original variables. The transformation is written as

$$\boldsymbol{Z = XA} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.12)$$

Where

$X\ is\ \ n * P\ matrix\ of\ n\ observation\ on\ P\ variables$

$Z\ is\ \ n * P\ matrix\ of\ n\ values\ for\ each\ of\ P\ components$

$A\ is\ \ P * P\ matrix\ of\ coefficients\ defining\ the\ linear\ transformation$

All *X* are assumed to be deviations from their respective means, hence *X* is a matrix of deviations from the mean.

3.5.1.2 Optimum Number of PCs

The fundamental core of selecting the optimum number of PCs depends on the eigenvalues and eigenvectors that can be found through the following steps:

1. The process begins with the data for n observations on P variables.
2. A matrix of size $n * P$ with deviations from the mean for each of the variables is formed.
3. The covariance matrix $(P * P)$ is calculated.
4. The eigenvalues and eigenvectors of the covariance matrix are calculated.
5. PCs are chosen and form a feature vector.

According to Valle et al.(1999), ten different approaches have been tested in the past to select the number of PCs. According to their conclusion, the most reliable method is

University of Malaya

the cumulative percent variance (CPV) because it is capable of selecting the optimal number of PCs, as confirmed by Zhang et al.(2011).

The idea of the CPV is illustrated in Figure 3.5. It shows the process of defining the optimal number of PCs (Zhang et al.,2010) . The figure indicates the relation between the number of PCs and accumulative variance percent, which has a curved shape. The location at which this curve becomes a straight line defines the optimal number of PCs. For example, the optimal number of PCs is 13, as shown in the figure 3.5.

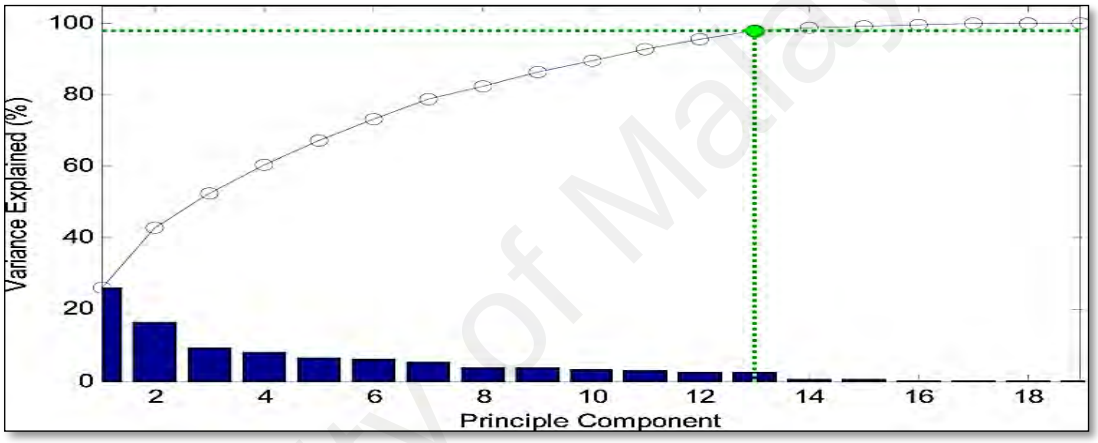


Figure 3.5: PC and CPV to select optimal numbers of PCs (Zhang et al., 2010)

Finally, the percentage of each PC component can be obtained through the equation (3.13)

$$P_i = \frac{\delta i}{Trace(S)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.13)$$

Where:

$P_i : Percentage\ of\ principal\ components$

$\delta i : is\ each\ eigenvector$

$Trace\ (s)$: $sum\ of\ the\ eigenvector\ \sum \delta_i$

University of Malaya

3.5.1.3 Multiple Linear Regression Model

MLR is one of the most widely employed statistical techniques in prediction applications. Typically, the purpose of MLR is to obtain the relationship between several independent variables and a dependent variable by fitting a linear equation to the observed data (Shalamu, 2009). The general equation of an MLR model illustrates in equation 3.14 (Montgomery et al,2012):

$$Y_t = \beta_0 + \boldsymbol{\beta_1 X_{1,t}} + \boldsymbol{\beta_2 X_{2,t}} + \boldsymbol{\beta_3 X_{3,t}} + \ldots\ldots\ldots + \boldsymbol{\beta_n X_{n,t}} + \boldsymbol{e_t} \ldots\ldots\ldots (3.14)$$

Where: $Y = dependet\ variable$

$\beta_0 = intercept\ constant$

$\beta_1 = slope\ coefficient\ correlation$

$X = independent\ Variables\ (explanatory)$

$t = the\ time\ period$

$n = the\ number\ of\ independent\ variables\ (explanatory)$

$e = the\ residual\ (error)$

This study posits that MLRs are unrelated when the explanatory variables are correlated. With such methods, predictor applications are usually challenged by certain disadvantages because high correlations among predictor variables cannot be easily and correctly analyzed (Pires et al.,2008). The problems related to the MLR model become more difficult when the input variables are high correlated coefficient with each other (McAdams et al.,2000b). Therefore, to remove correlation coefficient among independent variables, this study is conducted with PCA. According to Sousa et al.(2007), the combined PCA and MLR is called PCR.

University of Malaya

3.5.1.4 Combination Concepts between MLR and PCA

PCR is a type of regression analysis that confirms PCs as independent variables instead of adopting the original variables (Pires et al.,2008). PCR analysis is a combination of MLRs with PCA and establishes the relationship between the independent variable and the selected PCs of the input variable.

The PCs obtained from PCA are taken as independent variables in the MLR equation in performing PCR analysis. Equation 3.15 illustrates the PCR model:

$$\boldsymbol{Y = \beta_1 * PC_1 + \beta_2 * PC_2 + \beta_3 * PC_3 + \cdots + \beta_n * PC_n} \ldots\ldots\ldots\ldots\ldots\ldots (3.15)$$

As mentioned in Section 3.3.5, the standardization of the independent variables is based on equation (3.8) and that of the dependent variables on equation (3.9). Equation (3.10) is used to find the transform matrix dataset. The equation (3.16) is used to estimate β, and equation (3.17) is used for the regression model of the PCs. Then,

$$\widehat{\boldsymbol{\beta}} = (\acute{\boldsymbol{Z}}\boldsymbol{Z})^{-1}\acute{\boldsymbol{Z}}\boldsymbol{Y} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.16)$$

$$\boldsymbol{Y = Z * \beta} \quad \boldsymbol{or}\ y_i = \sum_{j=1}^{p} \beta_j Z_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.17)$$

Where

$Y: is\ \ (n * 1) vector\ of\ \ n\ \ obsevations\ of\ the\ centered\ dependent\ variable.$

$Z: is\ \ (n * p)\ matrix\ of\ \ n\ values\ for\ transformed\ data\ of\ \ (p)\ \ variables$

$\beta: is\ a\ (p * 1) vector\ of\ \ n\ \ unknown\ parameters.$

$\acute{Z}\ : is\ a\ transormation\ of\ value\ vectros\ matrix$

University of Malaya

As a summary of the PCR process:

1. The numerical value of the $\beta's$ retained in the regression is not altered by reducing the size.
2. The interpretation of $\beta's$ in terms of the independent variables is simplified.
3. The resulting regression coefficient is more stable when applied to a new dataset.
4. The disadvantage is that all the original variables must still be measured even though some PCs are eliminated.

The output of the PCR model consists of two dimensions: the first dimension is called the preliminary prediction of the electricity demand model, and the second is called the residual of dataset (error), which is used in Section 3.5.2 to improve the prediction model. However, the Figure 3.3 shows that the first processes from the input dataset to the linear model are based on PCA.

3.5.2 Nonlinear Model Back Propagation Neural Networks

Artificial Neural Network (ANN) shows better performance in dealing with nonlinear patterns found in input dataset. Kazemi et al.,(2009) and is also the best model for analyzing nonlinear datasets for the prediction model (Kavaklioglu et al.,2009). One of the most widely applied algorithms in neural network models is the back propagation (BP) algorithm, which is more accurate than other algorithms in obtaining minimum errors in the ANN model (Li et al.,2012). Williams et al.(1988) first proposed the BPNN and Figure 3.6 shows a typical processing element of an ANN. It is a supervised learning algorithm with feed-forward network.

University of Malaya

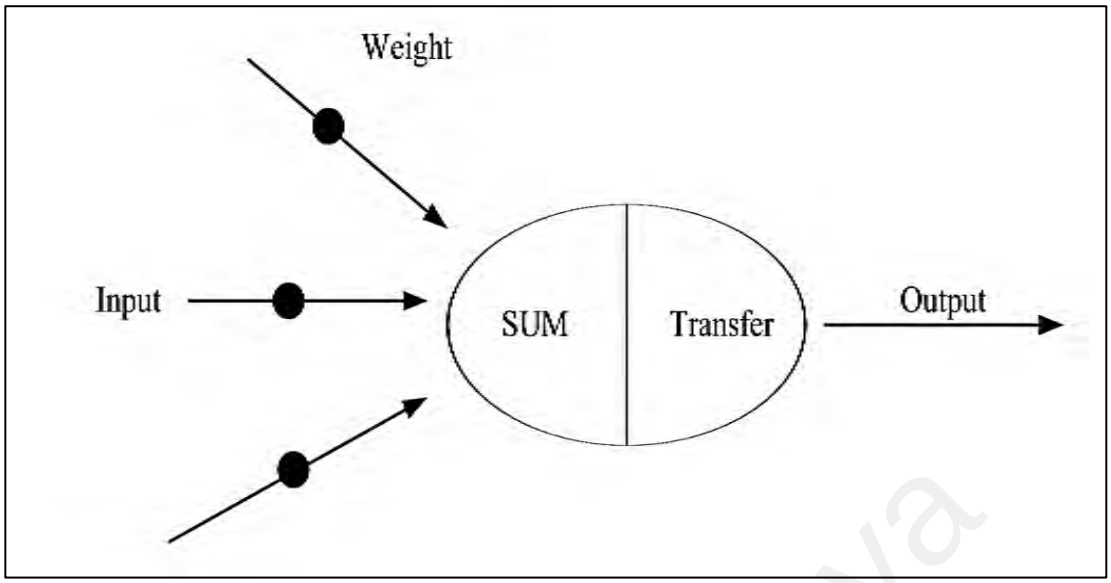


Figure 3.6: Typical processing element of an ANN

The term BP algorithm is derived from the way error rates in hidden layers are obtained after back propagating the error rates in the output layer. It utilizes the mean square error and gradient descent in modifying the connection weight of the network.

The architecture of the BPNN model is usually grouped into three types of layers, the input, hidden, and output layers. Data are fed into the nodes in the input layer after being transferred to the subsequent layer. Figure 3.7 presents a three-layer network with (*n*) input and propagation of signals throughout the output layer.

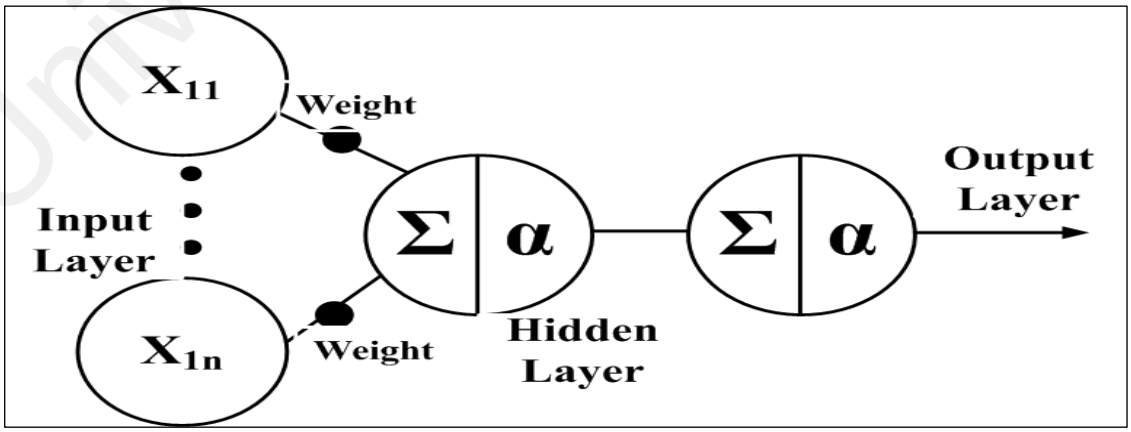


Figure 3.7: Three layers network with (n) input

University of Malaya

According to Figure 3.8, the variables $e_{t1}, e_{t2}, e_{t3} \ldots e_{tn}$ mean that the input layer (i.e., independent variables) normally have scales between (–1 to 1). The hidden layer is a medium layer that connects the input layers, and the output layer in the hidden layer combines the entire input layer with weights that are related to each input layer. After the calculation, the results are processed and sent to the next layer. The final layer is called the output layer, or target (y), and this layer normally presents the feature of the prediction (Tobias, 1995). The application of error rate in the BPNN model purposes to reduce the error of the dataset, which is provided in the first part from the PCR error.

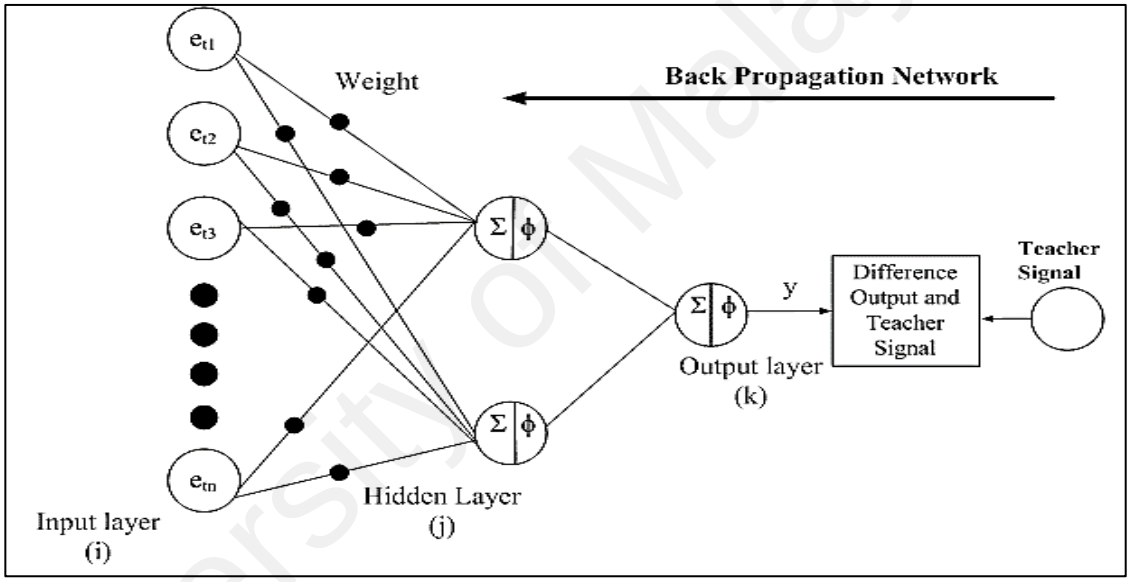


Figure 3.8: Structure of back propagation neural network

The next step takes the output signal of the ANN and compares it with the desired output value (i.e., the target), and the difference between these two values is called the error signal of the output layer neuron (Err_j).

The most common and best way to learn the neural network for predicting or classifying datasets is the back propagation algorithm (Dost et al.,1996). Figure 3.8 shows the structure of the BPNN model. This section also consists of the algorithm of BPNN. However, according to Kumarm et al., (2013) the initial weight for the neural network

University of Malaya

normally uses a small random number, such as (–1 to 1 or –0.5 to 0.5), where each unit has a bias associated with it.

For example, when the terminating condition is not convinced, the first part of the theory of BPNN and the second part are applied as an example.

First part: Mathematical theory of BPNN

I: input layers

O: output layers

T: Target unit

W: weight of neural network

i, j, k: the number of layer unit

$I_j = \sum_{i=0}^{n} W_{ij} O_i + \theta_j \ldots\ldots\ldots\ldots\ldots\ldots (3.18)$

Compute the net input of unit j with respect to the previous layer$i$.

$O_j = \frac{1}{1+e^{-I_j}} \ldots\ldots\ldots\ldots\ldots (3.19)$ Calculate the output of each unit j

Now, starting BPNN based on error of each unit (j) the output layer according to below the question;

$Err_j = O_j(1 - O_j)(T_j - O_j) \ldots\ldots\ldots\ldots\ldots\ldots (3.20)$ Compute the error

To each unit j the hidden layer from the last to the first hidden layer.

$Err_j = O_j(1 - O_j)\sum_k Err_k W_{jk} \ldots\ldots\ldots\ldots\ldots\ldots (3.21)$

Through equation 3.20, error rate is calculating with respect to the next higher layer$k$.

University of Malaya

In each weight W_{ij} in neural network, equation 3.22 is used to find the difference in the weight.

$$\Delta W_{ij} = (i) Err_j O_j \ldots\ldots\ldots\ldots (3.22)$$

Equation 3.23 is used to update the weights.

$$W_{ij} = W_{ij} + \Delta W_{ij} \ldots\ldots\ldots\ldots\ldots\ldots (3.23)$$

Finally, to update the θ_j from neural network

$\Delta\theta = (i) Err_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.24)$ Bias increment

$\theta_j = \theta_j + \Delta\theta_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.25)$ To update bias

Section 3.5.1.4 shows that one output from the PCR model is the error rates. Residual errors are most important in the diagnosis of linear model sufficiency. Figure 3.9 presents the architecture of the residual errors, which are used in the BPNN nonlinear model.

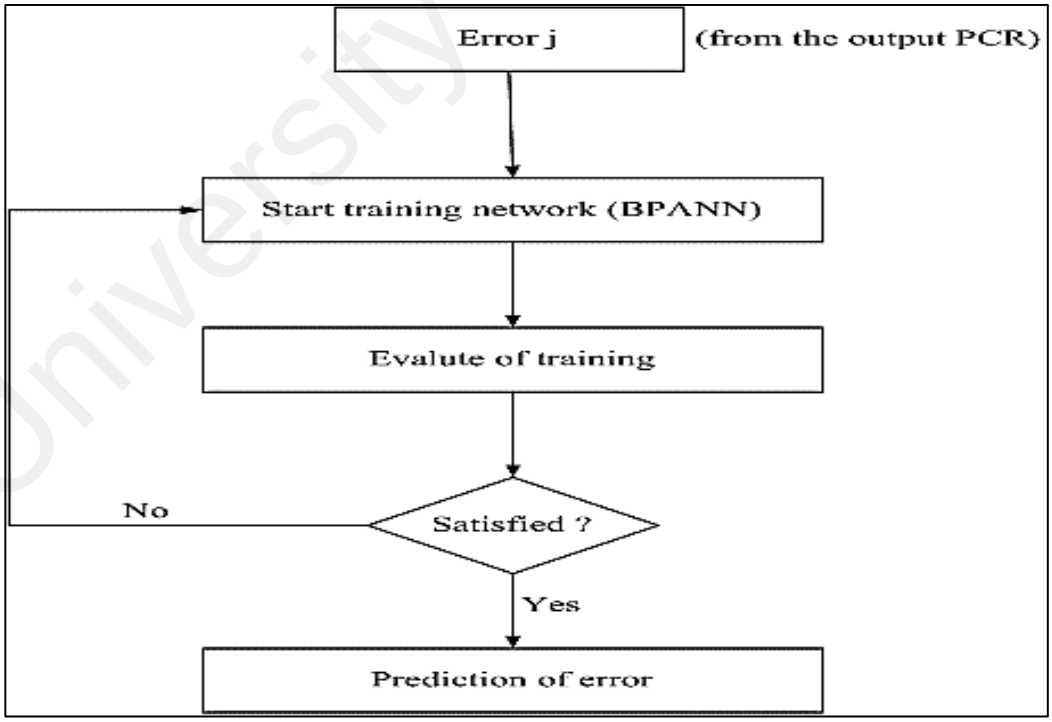


Figure 3.9: Nonlinear part of BPNN as residual errors processing

No nonlinear dataset patterns can be detected through residual analysis, even though the model passes the diagnostic checking. The adequacy of the model still results in

University of Malaya

doubly when nonlinear relationships have not been appropriately modeled. Therefore, this study used the BPNN model to reduce error rates.

3.5.3 Combination Process - Hybrid Approach

Figure 3.3 in Section 3.5 showed the architecture of this sub models combination process of the linear and nonlinear sub-models. Each sub-model has its own output, and the idea of PCR-BPNN is combining these two outputs together in order to cover both linear and nonlinear patterns of input dataset in the prediction process. Mathematically, Equation (3.26) represents this output combination, which obtaining better accurate in electricity demand predication rate is expected.

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{G}}_{\boldsymbol{t}} + \hat{\boldsymbol{J}}_{\boldsymbol{t}} \quad \text{..... (3.26)}$$

Where

$\hat{y}$: is hybrid approach

$\hat{G}_t$: *is inear estiamted by PCR*

$\hat{J}_t$: *is nonlinear estimated by BPNN*

Next sections show the results and analyze the accuracy of the proposed combination model in terms of some selected performance indicators.

3.6 Measures of Accuracy

When a model or a number of models fits a particular dataset, the models are often compared based on how well the models fit the historical data and how well they estimate future demand values. The first is generally referred to in this thesis as goodness of fit or

University of Malaya

fit, and the second is referred to as prediction accuracy. In estimating the model fit and goodness of fit, all the available data are used. However, in measuring prediction accuracy for a particular period, some actual data are discarded while developing the models. Forecasts provided by the developed models are then compared with the actual data that are used to measure the prediction accuracy. Therefore, prediction accuracy provides a better measure of model performance than goodness of fit because a better model fit does not necessarily imply good predictions (Garen, 1992).

Three indicators were employed to measure the goodness of fit and prediction accuracy throughout this thesis. Statistical measures were employed to evaluate the performance of the prediction models. This study depends on the following three statistical measures to evaluate the accuracy of prediction models:

a) Mean Square Error (MSE)

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=0}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}})^{\mathbf{2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.27)$$

b) Root Mean Square Error (RMSE)

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{MSE}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.28)$$

c) Mean Absolute Percentage Error (MAPE)

$$\boldsymbol{MAPE} = \frac{\mathbf{100}}{\boldsymbol{n}}\sum_{\boldsymbol{i=0}}^{\boldsymbol{n}}\left|\frac{\boldsymbol{y_i} - \boldsymbol{\hat{y}}}{\boldsymbol{y_i}}\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.29)$$

$Y_t : is\ \ the\ actual\ dataset$

$\hat{Y}_t : is\ \ the\ prediction\ of\ \ dataset$

$n : is\ \ the\ number\ of\ \ dataset\ points\ used$

University of Malaya

Table 3.2, adopted from (Lewis,1982), provides the criterion for judging the prediction accuracy obtained from MAPE.

Table 3.2: Criterion for referring prediction accuracy

MAPE	The accuracy of prediction
Less than 10 percent	High
From (11 to 20) percent	Good
From (21 to 50) percent	Reasonable
More than 50 percent	Worse (undesirable)

All three question (3.27), (3.28) and (3.29) used to measure the errors between actual value and estimate value, which predicted values by a model such as regression model.

3.7 Validation of Model

An important part of this study is model validation. Through this part or process, the results that will be obtained by PCR-BPNN prediction model will be evaluated against results of other prediction models that built by other theories and techniques than PCR-BPNN, such as SVR, PCR and PCNN. The process of validation depends on tests to confirm that result of the PCR-BPNN model should be superior to those of other models for an accurate prediction model.

The evaluation process will be done between PCR-BPNN and models that depend on three kinds of prediction based techniques; linear technique based model, nonlinear technique based model, or hybrid systems based model.

For the linear technique based model, this study will utilize the PCR model, which is more commonly used as linear technique for electricity demand prediction. For the nonlinear techniques based model this study will utilize PCNN. Finally, PC-SVR will be used as hybrid approach based model.

University of Malaya

3.8 Summary

This chapter explains the steps of the PCR- BPNN model in the form of a flow chart:

1. The PCR-BPNN model is proposed by this work to predict electricity demand.
2. The PCR -BPNN model is made up of two main parts, each consisting of different steps.
3. Parts of the PCR-BPNN model are mainly related to the following:
    a. Data collection and standardization with correlation among the independent variables. The PCA technique is then used to provide the number of input datasets for MLRs. The output in this section (first part) is two rates; the first rate is PPEDM, while the second is the residual (error).
    b. To improve the prediction model using the error rate mentioned in step (a) in the nonlinear models, such as BPNN, the output of the stage is called the prediction of error.
    c. The combination of the two steps above, (a + b), is called a hybrid approach (PCR-BPNN).

Finally, the results are compressed in the hybrid system (PCR-BPNN) with the other methods mentioned in the next chapter.

University of Malaya

CHAPTER 4: DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter presents the execution of the methodology and discusses the results of the data analysis. In Chapter 3, this study presented a new hybrid approach that combines the PCR and BPNN techniques. The new combination can improve the accuracy of the prediction model for the electricity demand.

This chapter illustrates the application of some statistical techniques on the performance measurement indicators to demonstrate the efficiency of the proposed model and presents the results at the end. The chapter also compares the new proposed hybrid approach and other models, such as the linear PCR and nonlinear PCNN models. Moreover, the comparison includes other hybrid approaches that have been previously studied in the field of prediction models for electricity demand, such as the PC-SVR model. All parts of the new proposed approach are coded using the MATLAB 2013 software.

4.2 Variables Selection

After standardizing the dataset, the data covariance is analyzed. The covariance shows the measure of the strength of the correlation coefficients between the dependent variable electricity demand and independent variables such economic factors for long term load prediction. The covariance can show the significant impact of the variables on the electricity demand. This process is used to find out the significant independent variables in a predefined population. Equation 3.6 is used to determine the covariance measure between the independent and dependent variables. Table 4.1 estimates the correlation degree of each independent variable with the dependent variable for Malaysian Dataset (i.e., actual demand).

University of Malaya

A p-value of 0.05 is used to indicate that an input variable is significantly correlated to electricity demand. Based on that, only the first 13 independent variables (among 19) have significant correlation with electricity demand rate of Malaysia, as shown in the Table 4.1. Humidity, industrial electricity demand, residential electricity demand, AGDP, NGDP, price of electricity, degree of urbanization, and unemployment are excluded from further analysis.

Table 4.1: Correlation coefficient between input variables and output

		r	**p-value**
No.	**Factors**	**Electricity demand**	
1	Population	0.987	<0.05
2	GDP	0.972	<0.05
3	GNP	0.978	<0.05
4	Income per capita	0.887	<0.05
5	Number of Employers	0.989	<0.05
6	Amount of export for a country	0.967	<0.05
7	Amount of import for a country	0.968	<0.05
8	Number of Tourists per year	0.975	<0.05
9	Consumer price index	0.605	<0.05
10	CO2 Emission	0.610	<0.05
11	Climate	0.611	<0.05
12	Industrial electricity	0.931	<0.05
13	Residential sector Electricity demand	0.985	<0.05
14	Humidity	0.35	>0.05
15	AGDP	0.47	>0.05
16	NGDP	0.41	>0.05
17	Price of electricity	0.42	>0.05
18	Degree of urbanization	0.34	>0.05
19	Unemployment	0.40	>0.05

In such case, variables with weak correlations will not be considered in future analysis and only 13 independent variables are selected as significant to the rate of electricity demand in Malaysia.

University of Malaya

4.3 Standardization

To code and implement the above transformation, MATLAB 2013a is used to standardize the dataset from the original dataset. The following MATLAB code can be used to achieve standardization.

```
Zx=zscore(x);
Zy=zscore(y);
```

The zscore(x) code is used to standardize the independent variables, whereas the zscore(y) is used to standardize the dependent variables. In the above two standardization codes, x is the independent variable of the dataset, and y is the dependent variable of the dataset. The data are imported from MS- Excel to the MATLAB tool using the two codes below:

```
[x]=xlsread('QX.xlsx'); %% the independent
[y]=xlsread('QY.xlsx'); %% the independent
```

Where:-

QX: Independent variables (input the variables)

QY: Dependent variable (output variable)

Table 4.2 shows a part of the dataset that has been transformed into the same range. The formula of Z-score is given by equations 3.8 and 3.9.

University of Malaya

Table 4.2: Transformation of the data set

opulatio n	GDP ('000)	GNP ('000)	income per Capita	Employeme nt ('000)	Export	Import	Tourist arrivals (million)	CO2 emissions (kt)	Consumer price index	Mean Climate	Industrial electricity	Residential electricity	E. D (ktoe)
-1.7854	-1.4716	-1.2930	-1.1611	-1.7691	-1.6642	-1.5440	-1.2487	-1.7419	-2.1989	-1.6985	-1.4219	-1.3169	-1.285
-1.7377	-1.4369	-1.2650	-1.1129	-1.7367	-1.6194	-1.4783	-1.1829	-1.5098	-2.1042	-0.1713	-1.3711	-1.3071	-1.238
-1.6900	-1.4022	-1.2371	-1.0647	-1.7043	-1.5746	-1.4126	-1.1171	-1.2777	-2.0095	-0.1365	-1.3202	-1.2974	-1.192
-1.6423	-1.3676	-1.2092	-1.0166	-1.6719	-1.5297	-1.3470	-1.0513	-1.0457	-1.9147	-0.4865	-1.2694	-1.2876	-1.145
-1.5945	-1.3275	-1.1772	-0.9172	-1.5356	-1.5125	-1.3158	-1.0617	-1.0094	-1.8167	-0.3420	-1.2295	-1.2773	-1.102
-1.5467	-1.2874	-1.1452	-0.8178	-1.3994	-1.4952	-1.2847	-1.0722	-0.9731	-1.7187	0.8682	-1.1897	-1.2670	-1.060
-1.4990	-1.2472	-1.1132	-0.7184	-1.2631	-1.4780	-1.2536	-1.0826	-0.9367	-1.6207	-0.2427	-1.1498	-1.2567	-1.017
-1.4512	-1.2071	-1.0812	-0.6190	-1.1268	-1.4608	-1.2225	-1.0930	-0.9004	-1.5227	-0.6519	-1.1100	-1.2464	-0.975
-1.4029	-1.1711	-1.0546	-0.5432	-1.0961	-1.4266	-1.2236	-1.1222	-0.9052	-1.4508	-1.2684	-1.0328	-1.2355	-0.989
-1.3546	-1.1351	-1.0281	-0.4675	-1.0654	-1.3924	-1.2248	-1.1514	-0.9099	-1.3790	-0.5057	-0.9557	-1.2247	-1.003
-1.3063	-1.0991	-1.0015	-0.3917	-1.0347	-1.3582	-1.2259	-1.1806	-0.9147	-1.3071	-0.9932	-0.8786	-1.2138	-1.018
-1.2581	-1.0631	-0.9749	-0.3159	-1.0040	-1.3240	-1.2271	-1.2098	-0.9194	-1.2352	-0.7843	-0.8014	-1.2030	-1.032
-1.2082	-1.0612	-0.9736	-0.3408	-0.9986	-1.2300	-1.1472	-1.2303	-1.0104	-1.0817	-1.6549	-0.8036	-1.1465	-1.013
-1.1584	-1.0594	-0.9723	-0.3657	-0.9931	-1.1359	-1.0673	-1.2507	-1.1014	-0.9281	-0.4186	-0.8057	-1.0901	-0.993
-1.1086	-1.0575	-0.9710	-0.3906	-0.9876	-1.0419	-0.9875	-1.2712	-1.1924	-0.7746	-0.6276	-0.8079	-1.0337	-0.974
-1.0587	-1.0557	-0.9697	-0.4155	-0.9821	-0.9478	-0.9076	-1.2916	-1.2834	-0.6210	-0.4360	-0.8101	-0.9773	-0.954
-1.0078	-1.0332	-0.9569	-0.6416	-0.9390	-0.8977	-0.9382	-1.2171	-1.3369	-0.5328	-1.2022	-0.7747	-0.9713	-0.930
-0.9570	-1.0107	-0.9441	-0.8677	-0.8960	-0.8476	-0.9688	-1.1425	-1.3904	-0.4446	-0.1400	-0.7394	-0.9653	-0.906
-0.9061	-0.9882	-0.9312	-1.0938	-0.8530	-0.7975	-0.9994	-1.0679	-1.4439	-0.3564	-0.0181	-0.7041	-0.9594	-0.882
-0.8552	-0.9657	-0.9184	-1.3198	-0.8099	-0.7473	-1.0300	-0.9933	-1.4974	-0.2682	-0.7146	-0.6687	-0.9534	-0.858
-0.8035	-0.8943	-0.8674	-1.2820	-0.7320	-0.6733	-0.9150	-0.9212	-1.3367	-0.2192	-1.2370	-0.6267	-0.9046	-0.813
-0.7518	-0.8230	-0.8163	-1.2442	-0.6541	-0.5992	-0.8001	-0.8491	-1.1761	-0.1702	0.2431	-0.5847	-0.8557	-0.768
-0.7001	-0.7516	-0.7652	-1.2064	-0.5761	-0.5252	-0.6852	-0.7771	-1.0155	-0.1212	-0.3664	-0.5427	-0.8069	-0.723
-0.6484	-0.6802	-0.7142	-1.1686	-0.4982	-0.4511	-0.5703	-0.7050	-0.8549	-0.0722	-0.0007	-0.5007	-0.7581	-0.678

4.4 Multicollinearity Problem

Equation (3.1) in Section 3.3.3 is used to estimate the correlation coefficient of the independent variables. The following code used in MATLAB 2013a environment to build the equation (3.1) and determine the correlation coefficient among independent variables.

```
Correlation=corrcoef(x); % correlation
```

The code `corrcoef(x)` is used to determine the correlation coefficient values of (x) independent variables. The results of the given code are shown in the Table 4.3.

University of Malaya

Table 4.3: Summary of correlation coefficient among independent variables

		economic factors							environment factors		eclectics sector		other factors	
	Variables	**GDP**	**GNP**	**income per Capita**	**No. Employers**	**Export**	**Import**	**Consumer price**	**Climate**	**CO2 emissions**	**Industrial electricity**	**Residential electricity**	**Tourist arrivals**	**Population**
economic factors	**GDP**	1.00												
	GNP	0.98	1.00											
	income per Capita	0.93	0.95	1.00										
	No. Employers	0.95	0.98	0.88	1.00									
	Export	0.98	0.95	0.86	0.94	1.00								
	Import	0.97	0.97	0.88	0.96	0.98	1.00							
	Consumer price index	0.87	0.79	0.73	0.79	0.86	0.80	1.00						
environment factors	**Climate**	0.60	0.62	0.61	0.58	0.54	0.56	0.53	1.00					
	CO2 emissions	0.79	0.69	0.65	0.65	0.81	0.75	0.71	0.39	1.00				
eclectics sector	**Industrial electricity**	0.85	0.94	0.84	0.96	0.85	0.91	0.64	0.56	0.48	1.00			
	Residential electricity	0.96	0.99	0.90	0.99	0.95	0.97	0.78	0.60	0.69	0.95	1.00		
other factors	**Tourist arrivals**	0.97	0.99	0.92	0.97	0.95	0.97	0.81	0.62	0.70	0.92	0.98	1.00	
	Population	0.97	0.96	0.86	0.97	0.98	0.97	0.86	0.55	0.78	0.88	0.97	0.96	1.00

To discuss the values shown in Table 4.3, the population variable is taken as an example. Population is considerably correlated with most of the other independent variables within a range of 0.85 and 0.98. The only medium correlation coefficient of the

University of Malaya

population exists with a climate variable by 0.55. The economic factors (i.e., GDP, GNP, income per capita, employment, export, and import, consumer price index, and tourist arrivals) are also highly correlated with one another within a range of 0.60 to 0.9. However, the correlation coefficients of the export and import variables with the climate are quite medium, and they are 0.54 and 0.56, respectively. The factors industrial electricity and residential electricity are highly correlated with other factors, within a range of 0.6 to 0.95. However, the correlation coefficient between industrial electricity and CO_2 emissions is as 0.48 as weakness. Two factors related to the weather affect electricity demand: a CO_2 emission, which is highly correlated with other variables within a range of 0.65 to 0.85, and climate, which has a median correlation coefficient with other independent variables.

4.5 Hybrid Approach

The new hybrid approach is consisting of two main parts. The first part is PCR. This part is utilizing the technique of PCA to solve multicollinearity issue of input dataset (as shown in Section 4.4) before feeding it to the MLR. The output in the form of the residual errors from the first part is processing by the second part, which is BPNN. The work of the second part is to eliminate the impact of the residual error on the prediction results.

4.5.1 Selection of Optimal Principal Components

This study utilizes the PCA statistic technique to reduce the multicollinearity problem of the independent variables. The most important part of the PCA method is the determination of the eigenvalue and eigenvectors in selecting the number of PCs based on the accumulative variances. Equation 3.10 provides the eigenvalue and eigenvectors of the independent variables. The code of the MATLAB 2013 program that is used to determine the eigenvalue and eigenvectors is

University of Malaya

Eigenvalue=eig(correlation)*; %eigne value*

[eigenVector, Eigenvalues]=eig(correlation); *% find eign value and eign vector*

The first equation in MATLAB 2013 is used to determine the eigenvalue, and the second equation is used to determine the eigenvectors. The eigenvalue can be found based on the percentage of the variance for each PCs. Equation 3.13 is used to apply the percentage of the variances. However, selecting the number of the PCs to be used requires the accumulation of the variance of the percentage. For example, the first variance of the PC is constant with the same value of the variance, but the second variance is accumulated from the first variance of the PC with the second variance of the PC, the third variance is accumulated from the second variance of the PC with the third variance of the PC, and so on. Table 4.4 explains the percentage of the variation and the selection of the number of PCs based on the accumulated variances, and Figure 4.2 shows the accumulation of variances and eigenvalues.

Table 4.4: Total variance explained by the PCs

	Eigenvalue	Variability (%)	Cumulative %
PC1	**10.02**	**77.07**	**10.02**
PC2	**1.07**	**85.34**	**1.07**
PC3	**0.87**	**92.04**	**0.87**
PC4	**0.71**	**97.52**	**0.71**
PC5	0.20	99.02	0.20
PC6	0.05	99.40	0.05
PC7	0.04	99.70	0.04
PC8	0.02	99.86	0.02
PC9	0.01	99.93	0.01
PC10	0.01	99.96	0.01
PC11	0.00	99.99	0.00
PC12	0.00	100.00	0.00
PC13	0.00	100.00	0.00

University of Malaya

Table 4.4 summarizes the results of the PCA on these 13 factors with the amount of variance that explains each PC component and the total variance of the original variables. As the table makes clear, the first four PCs have the highest eigenvalues among the thirteen factors, with 10.02, 1.07, 0.87, and 0.71 respectively. Moreover, the cumulative contribution of the total explained variance for these selected four components is 97.52%. Therefore, the first four PCs can provide the most information on the original dataset and the extracted dataset.

Another confirmation for selecting the optimum number of PCs is through explaining the relation between PC number and accumulative of variance. Figure 4.2 simultaneously shows the level of the eigenvalue and cumulative variance. It shows how the optimal number of PCs is selected based on the Cumulative Percent Variance (CPV). As shown in the figure and according to the CPV method, the optimal number of PCs is four. As the curve of CPV becomes a straight line, it is no longer affected by the PCs at 97.5%.

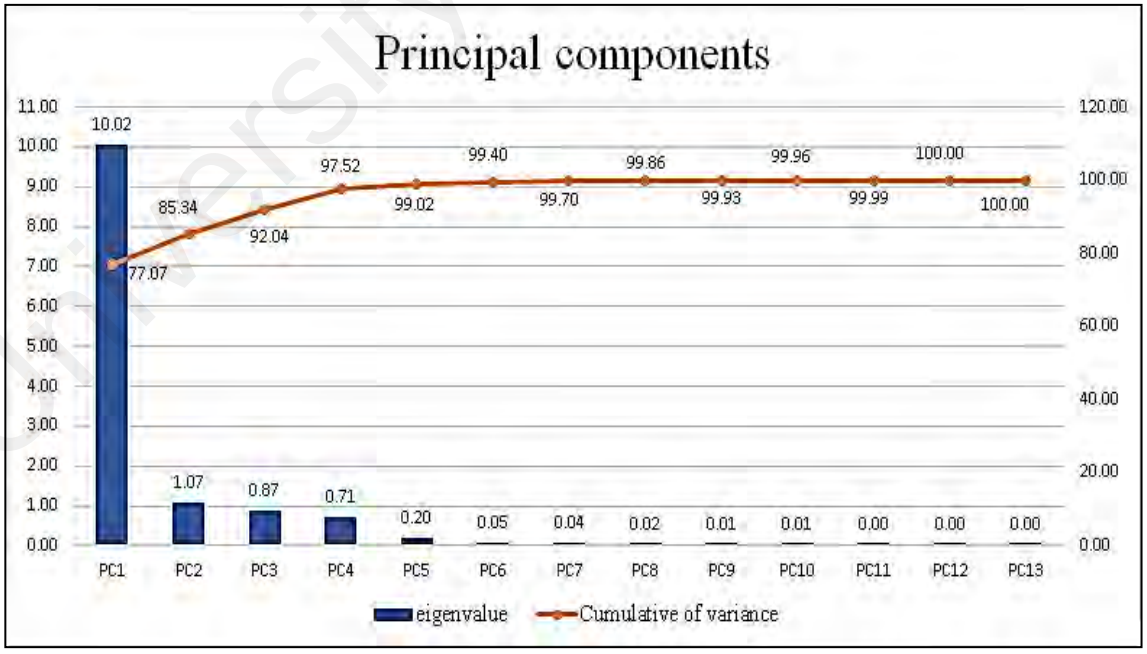


Figure 4.1: Principal components and accumulative variance

University of Malaya

The main predictor variables as selected by the PC technique (according to the Table 4.4 and Figure 4.1) are population, GDP, GNP, exports, imports, income per capita, and industrial electricity. The PC_1 provides the highest amount of variance in the dataset, and PC_2 is larger than PC_3, and so on. To see the accumulative variance of all PCs (fours) the following Matlab 2013 code is implemented. The output of this code is summarized in table 4.5.

PCs = eigenvector (1:13, 1:4); *%to find from PC_1 to PC_4*

Table 4.5: First four principal components

	PC1	PC2	PC3	PC4
Population	**0.985**	-0.076	0.008	0.033
GDP	**0.988**	-0.006	0.032	0.087
GNP	**0.991**	0.086	-0.031	-0.070
income per Capita	**0.916**	0.141	-0.034	-0.031
Number of Employers	**0.979**	0.038	-0.027	-0.148
Amount of Export	**0.981**	-0.093	0.039	0.097
Amount of Import	**0.986**	-0.006	-0.033	0.013
Number of Tourist arrivals per years	**0.986**	0.087	-0.032	-0.032
CO2 emissions	0.661	-0.215	-0.075	**0.708**
Consumer price index	0.365	-0.432	**0.817**	-0.106
Climate	-0.087	**0.873**	0.424	0.217
Industrial sector for electricity demand	**0.915**	.144	-.093	-.316
Residential sector for electricity demand	**0.986**	.085	-.051	-.076

Determining the number the PCs through PCA techniques is converting the input dataset with multicollinearity into a new dataset without multicollinearity. Through this process, new dataset from PCA is obtained by multiplying the original standardized dataset with the number of the principal components chosen, as mentioned in the Matlab 2013a code below:

dataset=Zx * PCs;

University of Malaya

Where dataset refers to a new dataset, Zx refers to the original standardized dataset, and PCs refers to the selected number of PCs. According to the selected number of PCs, Table 4.6 shows only 20 observations out of (76) as parts of the new dataset that has been obtained through the above Matlab code.

Table 4.6: New dataset obtained from PCA

Observation	PC-1	PC-2	PC-3	PC-4
Obs1	-1.54	0.79	-2.37	0.35
Obs2	-1.51	0.98	-2.11	0.47
Obs3	-1.41	-0.15	-0.79	-0.29
Obs4	-1.41	2.37	0.39	0.64
Obs5	-1.35	-0.22	-0.93	-0.30
Obs6	-1.32	1.85	0.53	0.46
Obs7	-1.26	0.71	0.06	0.04
Obs8	-1.25	2.67	0.91	0.75
Obs9	-1.23	1.65	-0.58	0.32
Obs10	-1.20	1.03	-0.90	0.05
Obs11	-1.14	0.54	-0.47	-0.18
Obs12	-1.12	-0.17	-1.50	-0.49
Obs13	-1.10	-0.63	-1.61	-0.74
Obs14	-1.07	-0.06	-1.11	-0.60
Obs15	-1.00	-0.01	1.48	-0.66
Obs16	-0.94	-1.30	1.67	-1.22
Obs17	-0.93	-1.70	1.54	-1.43
Obs18	-0.94	-0.72	2.23	-1.12
Obs19	-0.92	-1.27	2.14	-1.38
Obs20	-0.98	0.92	2.10	-0.62

The new data set that obtained throughout the PCA process should be uncorrelated, which means correlation among independent variables should be as less as possible. Minimizing the correlation among independent variables means the multicollinearity problem has been eliminated. To test this, the Karl Pearson product moment method is used to determine the correlation coefficient of the new dataset (PCs). Table 4.7 shows the correlation coefficient of the new dataset (PCs). The following MATLAB 2013 code can achieve the above-mentioned processes:

Universiti Malaya

correlationPCs=corrcoef(PCs); % correlation PCs

Table 4.7: Correlation coefficient for new dataset

	PC-1	**PC-2**	**PC-3**	**PC-4**
PC1	1			
PC2	-1.5E-07	1		
PC3	-2.1E-07	3.5E-07	1	
PC4	-8.8E-08	2.78E-07	-8.2E-07	1

According to the same threshold value that used in section (4.4) for Table 4.2, independent variables with correlations of less than 0.5 are uncorrelated (insignificant), all obtained PCs shown in Table 4.6 are uncorrelated because their *P* values are less than 0.5.

4.5.2 Principal Components Regression

The combination of the PC technique to solve multicollinearity problem and MLR as a prediction tool forms a kind of linear combination called PCR. This section uses PCR to show the Preliminary Prediction Electricity Demand Model (PPEDM). Equation 3.15 is used to determine the PCR, and the MATLAB 2013 code is used to complete the PCR:

Yhat_PCR= new_dataset*beta;

Where

Yhat_PCR: predict model

new_dataset: PCs data multiple independent variables when applied standardization.

Beta: coefficients efficient

The first step to get the output of the PCR is to have the values of the *Beta*. To get that Equation (3.16) in Chapter 3 is used to estimate the general parameters of β_i (i = 1, 2, 3… n).

University of Malaya

Table 4.8: Result of regression analysis

Parameters	Coefficients β_i	Standard Error	t Stat	P-value	R2 (%)
PC1	0.9892	0.01501	65.4627	0.0200	0.9838
PC2	-0.0077	0.01511	-0.5104	3.23E-65	
PC3	-0.0110	0.01493	-0.7292	0.03113	
PC4	-0.0716	0.014660	-4.7371	0.0468	

As Table 4.8 shows, the four PCs (i.e., PC_1, PC_2, PC_3 and PC_4) can explain 98.38% of the variation ($\boldsymbol{R^2}$) in electricity demand. According to the *p-value*, the PCs are the most significant independent variables in the regression analysis, given that each of the *P-value* is less than *0.05 (*i.e., *P-value < 0.05).* The PC1 has a positive effect on the electricity demand model, whereas the rest of the PCs (i.e., PC2, PC3, and PC4) have a negative effect on the electricity demand. Table 4.9 shows all of the estimated coefficients (β_i) that are used in the PCR model

Table 4.9: Estimated parameters for all Betas

Betas	Coefficients
β_1	0.9891
β_2	-0.0077
β_3	-0.0110
β_4	-0.0716

Therefore, the equation for the developed PCR linear model could be written as: -

$$PPEDM = 0.9892\,PC_1 - 0.0077\,PC_2 - 0.0110PC_3 - 0.0716PC_4 \ldots\ldots\ldots\ldots\ldots (4.1)$$

Although Equation 4.1 shows linearity, theβ_i 's value that shown in the Table 4.8 can't be considered for estimating their direct impact on PPEDM. Because, the value of PC_i has been associated with its correspondingβ_i, which comes with negative or positive signs (a part of PC_i values have been shown in the Table 4.6).

University of Malaya

Finally, according to the Equation 4.1 and the Matlab 2013a code shown below,

error= Yhat_PCR - Zy;

Error: error rates and the *Zy*: actual electricity demand.

Table 4-10 provides the part of result preliminary prediction model and residuals of error for the electricity demand in Malaysia based on the Principal Components Regression.

Table 4.10: Prediction results based on PCR

Observation	Predicted E.D	Error
1	-1.531545789	0.058164
2	-1.512937529	-0.04998
3	-1.361717913	-0.26084
4	-1.458500156	-0.19383
5	-1.297565274	-0.35464
6	-1.356111161	-0.26609
7	-1.256258323	-0.30606
8	-1.319677958	-0.15287
9	-1.248985882	0.03126
10	-1.184187513	0.061929
11	-1.11558672	0.064609
12	-1.056894208	0.053011
13	-1.01277873	-0.032
14	-1.007081506	-0.01345

University of Malaya

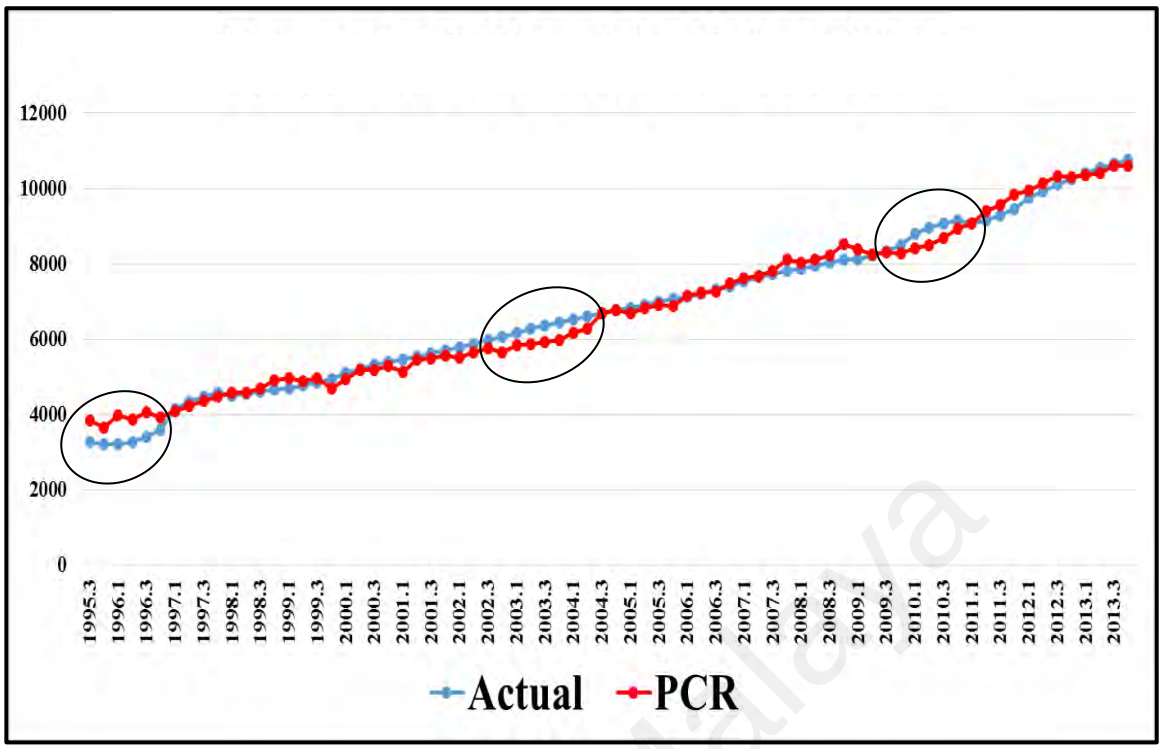


Figure 4.2: PCR model with actual data of electricity demand

Figure 4.2 and the results in Table 4.9 show the output of the PCR model (Yhat). From 1995Q1 to 1997Q1, based on the figure4.2 they are not accurate and have high a gap between them. However, which is almost the same as the actual data on the electricity demand from $1997Q_2$ to $2002Q_3$. Again, 2003Q1 – 2005Q1 this similarity means that errors from that time were high and they are not accurate. However, noticeable differences exist between the actual output values of the PCR from $2005Q_2$ to 2009Q4 the desired output of the electricity demand for the same duration. However, from 2010Q1 to 2011Q3 also inaccurate rate the error rate meaning high and rest they are close to actual data. These differences mean that errors from that time were high and inaccurate. To measure that, Section 3.6 discusses the measurement of the accuracy prediction, and Table 4.11 gives the RMSE and MAPE of the PCR model. The following MATLAB 2013a code is used to find out the mentioned performance indicators:

RMSE= sqrt(mean((error).^2));

MAPE = (mae(Zy - Yhat_PCR))*100;

Table 4.11: Measure of performance indicators

No.	Model	MSE	RMSE	MAPE%
1	Principal component regression (PCR)	0.015998	0.126484	27

The overall result in Table 4.11 shows that the performance of the PCR model is significant based on Table 3.2, which explains the performance indicators. However, for the prediction model, especially of electricity demand, this accuracy still needs improvement. Therefore, BPNN is employed to minimize the error rates and improve the accuracy of such prediction model. The next part shows the process of BPNN when applied to the error rates.

4.5.3 BPNN Based Residual Error Processing

As mentioned in Section 4.5.2, the PCR model has two kinds of output: the PPEDM and residual error rates. However, for the second output related to the residual error rate, this study applies the BPNN model on the time series to improve the accuracy of the predictions by minimizing the error rates. In such time series-based BPNN model, the network has one input unit and one output unit, as shown in Section 3.5.1.5. This study uses the MATLAB 2013 code in building the desired BPNN model. Figure 4.3 shows the popped message, which gives the availability of different types of BPNN, one of which is the time series type (*ntstool*). This study uses the neural time series tool, BPNN (*ntstool*), as a structure to improve the accuracy prediction rates.

University of Malaya

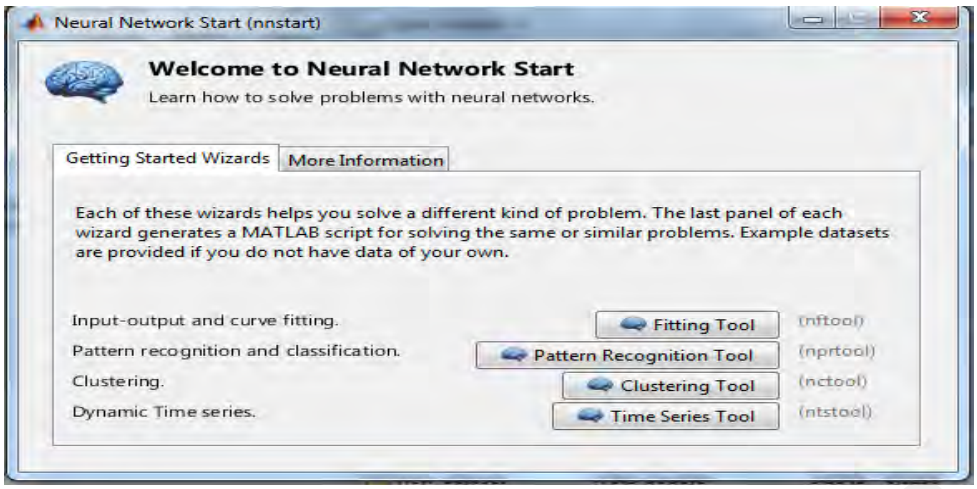


Figure 4.3: Time series tool (ntstool)

The standard structure of a BPNN has an input layer, hidden layer and an output layer. These layers are interconnected with nodes. Number of nodes at each layer depends on type and the complexity of the problem. For the present study, the input node is one as the network receives one error as a time. The output node is also one because the network provided one number (rate of residual errors) each time. The complex one is estimating the number of hidden layer and the nodes at each hidden layer. This could be done through different ways. However, the most optimum one is using the Cascade-Correlation algorithm. According to this method, number of hidden layer and the nodes at each layer will be added sequentially. When a layer is added, number of node at this added layer will be increased one by one. The method initialized at zero hidden layers and zero node numbers at each hidden layer. After each adding, the training performance of the network will be checked for 10 times, and then the average performance will be obtained and considered. The process of adding nodes and checking the performance of training will be continued until no better performances will be obtained. Table (4.12) shows the results of the Cascade-Correlation process for this study, and it shows that the optimum neural network's architecture for this study is one hidden layer with 10 nodes. To choose an optimum architecture, researchers should always select the simplest architecture that

University of Malaya

provides the better performance because adding more hidden layers and nodes means increasing the time and the space complexity of a model. However, without adding hidden layers the accuracy and the performance of the model will be very weak. Therefore, a tradeoff between complexity and the accuracy should be considered. As shown in the Table (4.12), the performance of the model is improving from architecture (0, 0) until the (1, 10). After that, no significant improvements have been noticed. Even, another hidden layer is added the performance have not improved notably.

According to the performances shown in the Table (4.12), the best architecture for the present model is (1, 10), which means having one hidden layer and 10 nodes at that layer.

Table 4.12: Performance of the model is improving from architecture (0, 0) until the (1, 10)

NN Architecture **(hidden number, node's number)**	**Average Performance (MSE)** **Training Process**
(0, 0)	1.563
(1, 2)	0.235
(1, 4)	0.128
(1, 6)	0.069
(1, 8)	0.017
(1, 10)	0.002
(1, 12)	0.00199
(1, 14)	0.00185
(1, 16)	0.00179
(2, 2)	0.00196
(2, 4)	0.00193
(2, 6)	0.00191
(2, 8)	0.00189
(2, 10)	0.00185

The format of the input dataset is time based; therefore, the proposed structure of the BPNN is the time series-based NN that can be learned according to a nonlinear autoregressive (NAR) equation. This learning is used to accumulate the residual error rates and to improve the accuracy of the electricity prediction models. For the NAR

University of Malaya

equations, the number of the input and output nodes are similar. The relationship between the output $e_{(t)}$ and the inputs $(e_{(t-1),}e_{(t-2)}, \dots, e_{(t-d)}$ is represented by equation (4.2):

$$\boldsymbol{e_{(t)} = f(e_{(t-1),}e_{(t-2)}, \dots \dots \dots \dots \dots \dots \dots \dots e_{(t-d)})} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (4.2)$$

Where $e_{(t)}$ represents the residual errors at time t from the PCR model (i.e., linear model), and f is a function conducted by the NN structure and connection weights.

After choosing the equation type of BPNN, the target should be selected from the dataset, and based on Kavaklioglu et al.,(2009), is divided into three types: 70% training of the dataset, 15% validation of the dataset, and 15% testing of the dataset. Table 4.13 explains the classification of the dataset.

Table 4.13: Classification of the dataset

Training of the dataset	Validation of the dataset	Testing of the dataset	Total
70%	15%	15%	100%
54 target	11 target	11 target	76 target

Particularly, the train dataset used for learning and in the back-propagation used to find the optimal weight. The validation data has been used to get the optimal dataset of the hidden layer. Moreover, it is used to determine a stopping point in the neural network such as back-propagation. The last one is testing dataset, usually used to estimate the residual error rate after selecting the final optimal model.

This study clamps the $e_{(t)}$ dataset to the BPNN and is also based on Equation 4.1, selecting the number of hidden neurons (layers) and feedback delays. The architecture of the neural network has ten nodes at the hidden layer, and the value of the delays (d) is equal to 1:2 that shows from the Figure 4.4. The following written code in MATLAB 2013 is used to create an NAR neural

University of Malaya

network:*feedbackDelays = 1:2;*

hiddenLayerSize = 10;

net = narnet(feedbackDelays,hiddenLayerSize);

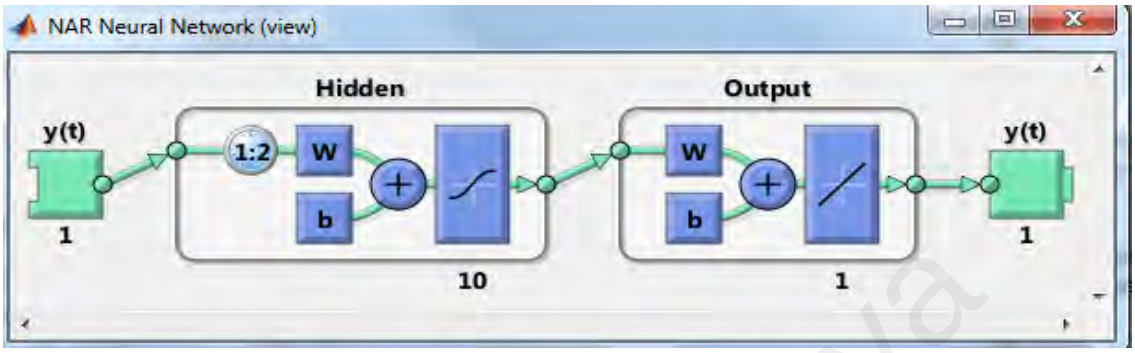


Figure 4.4: Architecture of the ANN as tested predicted model

To achieve the division mentioned in Table 4.2, the following MATLAB 2013 code is used:

net.divideFcn = 'dividerand';

net.divideMode = 'time';

net.divideParam.trainRatio = 70/100;

net.divideParam.valRatio = 15/100;

net.divideParam.testRatio = 15/100;

The *net.divideFcn* code is used to randomly divide the dataset, and the *net.divideMode* is used to divide every values based on time. According to Levenberg-Marquardt optimization, the *trianlim* function is often faster than the other functions in the BPNN model for the training dataset. The *trianlim* is highly recommended as the first choice supervised algorithm. The code *trianlim* in MATLAB 2013 is as follows:

net.trainFcn = 'trainlm'

After preparing the train, test, and validation datasets and fixing the training function, the MSE is employed as a performance function for all the classified dataset in the NN model. The MSE code in MATLAB 2013a is written as

University of Malaya

net.performFcn = 'mse'; % Mean squared error

The training stage of any BPNN based model usually uses the k-fold to estimate the generalization and input dataset validation. At the first stage, the training process will randomly select 15% of the records in the error dataset. Then, they will be passed the BPNN to check the testing. A sample of the selected records is shown in the table 4.13. Each sample like this will be prepared for training and validation too. However, for training, 70% of the dataset will be selected and for validation only 15%. This process will be repeated to K-folds. At each fold, performance indicators are calculated. In this study, the process repeated up to 10-folds. The process of selecting samples will be done randomly, as shown in the table 4.14.

Table 4.14: Randomly select of testing dataset

#/ test	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
1	0.063	0.063	0.000	0.075	-0.070	0.011	0.094	0.020	0.020	0.072
2	-0.067	0.020	-0.025	0.000	0.007	0.062	-0.047	0.094	-0.077	-0.037
3	-0.044	-0.025	-0.047	-0.070	0.045	0.020	0.059	0.013	-0.047	0.007
4	-0.035	-0.067	-0.035	-0.055	0.020	0.094	-0.004	0.041	-0.044	0.035
5	0.035	0.007	0.045	0.015	0.052	-0.081	0.033	-0.025	0.007	-0.025
6	0.000	0.052	0.020	0.058	-0.025	0.007	-0.012	-0.007	0.058	-0.055
7	-0.015	-0.079	-0.007	0.059	-0.036	-0.035	-0.015	-0.015	0.065	-0.079
8	-0.079	-0.024	0.015	-0.004	-0.024	0.045	-0.024	0.015	0.059	0.094
9	0.001	0.013	0.054	-0.079	0.094	0.153	0.102	-0.024	-0.025	0.153
10	0.094	-0.313	-0.313	0.094	0.153	-0.284	0.054	0.094	-0.047	0.129
11	-0.224	0.238	-0.284	0.153	-0.284	0.162	0.068	0.102	0.102	0.162

Same sets of input dataset will be selected randomly for testing and validating. In all cases, the sets will be passed to time based BPNN to check the performance. As shown in the table 4.15, ten MSE indicators have been obtained for each selected sets.

University of Malaya

Table 4.15: Performance indicators-MSE

No	training Performance	Validation Performance	Testing Performance
1	0.0026	0.0049	0.0043
2	0.0011	0.0025	0.0142
3	0.0014	0.0067	0.0154
4	0.0043	0.0032	0.0032
5	0.0014	0.0057	0.0217
6	0.0040	0.0021	0.0022
7	0.0033	0.0017	8.0609e-04
8	0.0034	0.0091	0.0071
9	0.0013	0.0021	0.0019
10	0.0029	0.0130	0.0056

In the Table 4.15, the performances of the proposed model somehow for all folds and for of training, validating, and testing are similar. This means that input dataset for the time based BPNN model are validated. Among all performances in the Table 4.15, stage seven is 0.00080609, and it is the best result based on the MSE. The output shows that the value of the testing performance in the stage seven indicators is smaller than the values of other testing performances. Table 4.14 shows the random selection for the testing dataset. Table 4.14 is selecting the values of the testing dataset. This study uses 15% dataset to test the proposed model. Table 4.16 illustrates the performance indicators for each testing dataset. As previously mentioned, this study uses the K-fold validations for the testing dataset. The idea of the K-fold involves using the average of RMSE and MSE for all the testing datasets tin set in the BPNN.

Table 4.16: Performance indicators for testing of dataset

#	MSE	RMSE
Test1	0.004	0.066
Test2	0.014	0.119
Teset3	0.015	0.124
Test4	0.003	0.057
Test5	0.022	0.147
Test6	0.002	0.047
Test7	0.001	0.028

University of Malaya

Continuous Table

Test8	0.004	0.06
Test9	0.002	0.044
Test10	0.006	0.075
Mean	0.007	0.077

Table 4.17 is related to section 3.5.1.5, which explains the weights and biases used in this neural network from the Equation 3.22 and 3.23. The values that indicated in this table are optimum values, which means at these weight's values the time based BPNN gives the best performance indicator.

Table 4.17: Weight input, layer and Bias

Weight /input		**Weight Layer**	**bias**
3.081757	3.110265	-0.29937	-4.44612
-0.53802	-4.34395	-0.49445	3.52379
-3.58235	-2.50557	-0.00352	2.501008
-3.70975	2.119171	-0.35918	1.590331
1.403156	4.190317	0.158071	-0.4959
3.203768	3.05637	-0.15816	0.462392
3.03723	-3.21035	0.187196	1.476869
-1.79293	-4.01561	-0.07951	-2.5098
3.471012	2.670818	0.215155	3.494909
0.263317	4.2469	0.313435	4.581473
			-0.0248

Figure 4.5 explains the training, validating, and testing process with best validation performance. The figure shows that time based BPNN for this fold has got best trained, validated and tested at epoch number two.

University of Malaya

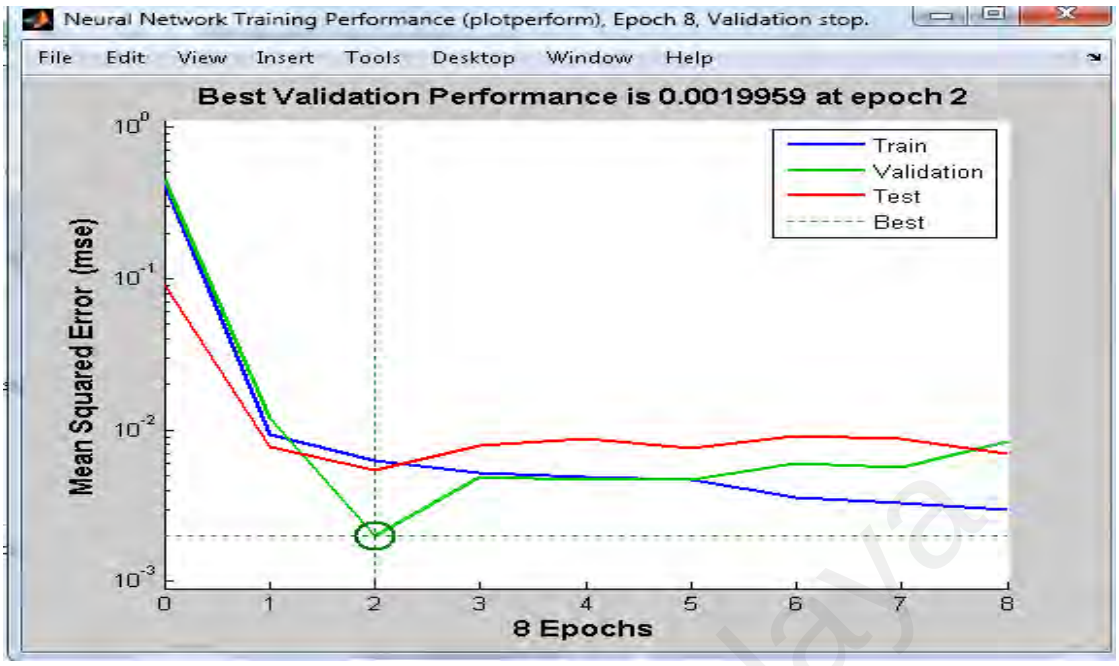


Figure 4.5: Performance of MSE

The training, validating, and testing processes for the time based BPNN at this fold completed with only eight iterations (i.e., epochs). The best performance validation at epoch two is 0.0019959. Figure 4.5 shows the process of finding the best validation performance indicator using the MSE function.

Figure 4.6 shows the error points between the output and target for each training, testing, and validation dataset. The figure shows that the distance between actual output and the desired output are acceptable

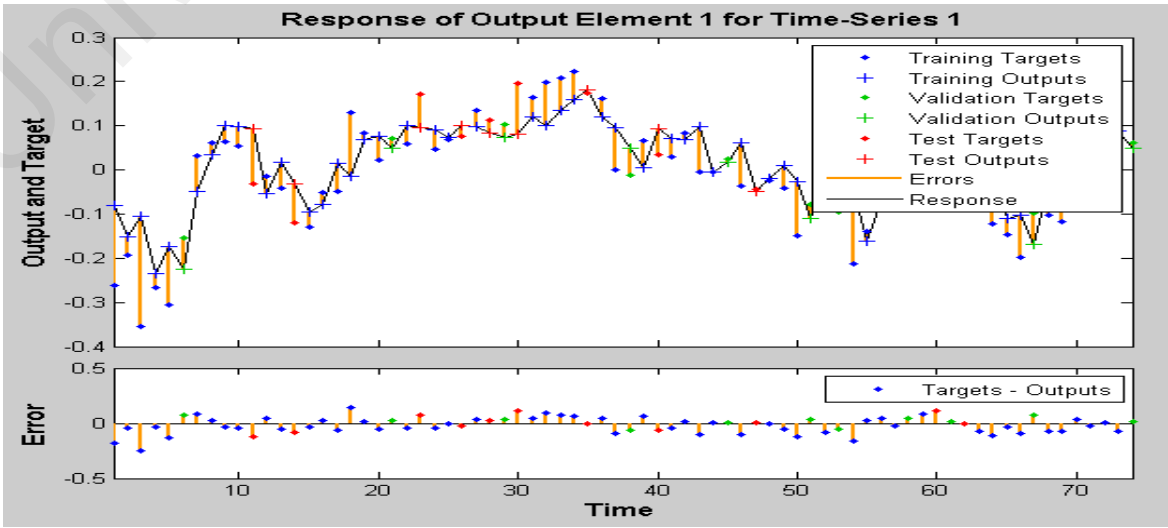


Figure 4.6: Error (target – output)

Figure 4.6 explains the error of the point by the formula target–output. The blue points indicate the training target, the red points the testing target, and the green points the validation targets.

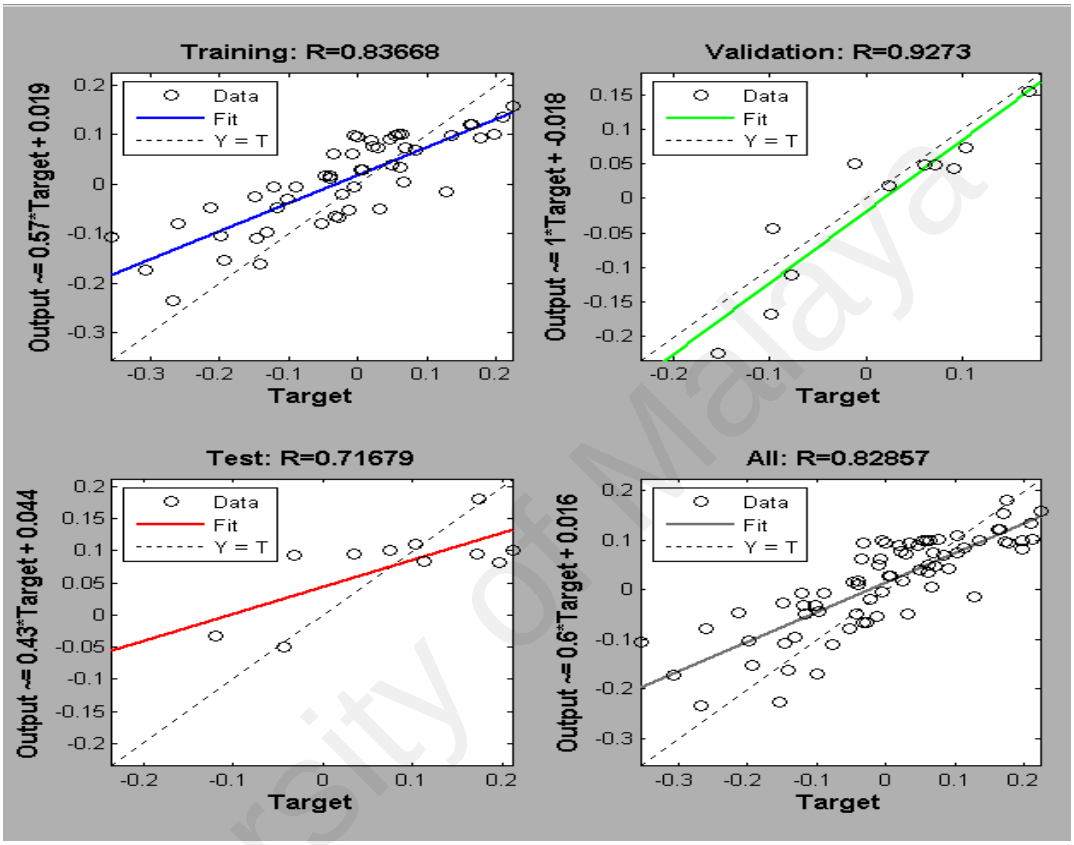


Figure 4.7: Regression training, testing and validation dataset

Figure 4.7 shows the regression for the training, validation and testing dataset. The figure shows a good fit with an actual dataset of 0.83, 0.92, and 0.71, respectively. Moreover, for all the datasets, the regression model for fitting a dataset is also good, such as 0.82. This study focuses on the improvement of the entire prediction accuracy. Therefore, in the next section, this study employs a hybrid approach to improve accuracy prediction model by decreasing the error rates and the combination with the preliminary prediction model.

University of Malaya

4.5.4 Combination PCR – BPNN Approach

As mentioned in (3.5.1.6), this study uses a combination of the PCR linear and BPNN nonlinear models. Equation (3.26) explains this hybrid approach, and the code for this hybrid approach in MATLAB 2013a is as follows:

Hybrid = error BPNN+ Yhat PCR;

Where *hybrid* represents the output of the hybrid estimation approach, *error BPNN* represents the output of the estimated nonlinear part, and *Yhat PCR* is the preliminary prediction model. Table 4.18 shows a part of the output of the hybrid approach.

Table 4.18: Result of hybrid approach

Time	Error hat ($\widehat{er}$)	Yhat_PCR	hybrid = Yhat_PCR+ Error hat ($\widehat{er}$)
obs.1	-0.181349593	-1.361717913	-1.543067506
obs.2	-4.13E-02	-1.458500156	-1.499843063
obs.3	-0.248980783	-1.297565274	-1.546546057
obs.4	-0.03164426	-1.356111161	-1.387755421
obs.5	-0.13329473	-1.256258323	-1.389553053
obs.6	0.072550315	-1.319677958	-1.247127644
obs.7	0.080289531	-1.248985882	-1.16869635
obs.8	0.026855785	-1.184187513	-1.157331728
obs.9	-0.034847538	-1.11558672	-1.150434257
obs.10	-0.045831447	-1.056894208	-1.102725655
obs.11	-0.125330549	-1.01277873	-1.138109279
obs.12	0.040011767	-1.007081506	-0.967069738

The weight and bias used for the hybrid approach in the BPNN model uses the following formula:

weights = getwb(net); % to find the neural network

getwb is used to find the neural network weights and bias. However, to separate them, the following formula is used:

weights = getwb(net); % to find the neural network

[b,IW,LW] = separatewb(net,weights); %to spreate bias and weight

Where, *sepratewb* is used to separate the bias and weight of the neural network. b, IW, and LW are used to determine the cell array of the bias input weight matrices and layer matrices, respectively.

In this study, Equations 3.27, 3.28 and 3.29 determine the performance indicators for the hybrid system. Table 4.19 and Figure 4.8 illustrate the performance indicators for the PCR-BPNN model with the actual independent variables that have been tested throughout this study.

Table 4.19: Performance indicators for PCR-BPNN

No.	Model	MSE	RMSE	MAPE%
1	Hybrid approach PCR – BPNN	0.008865	0.094155	13

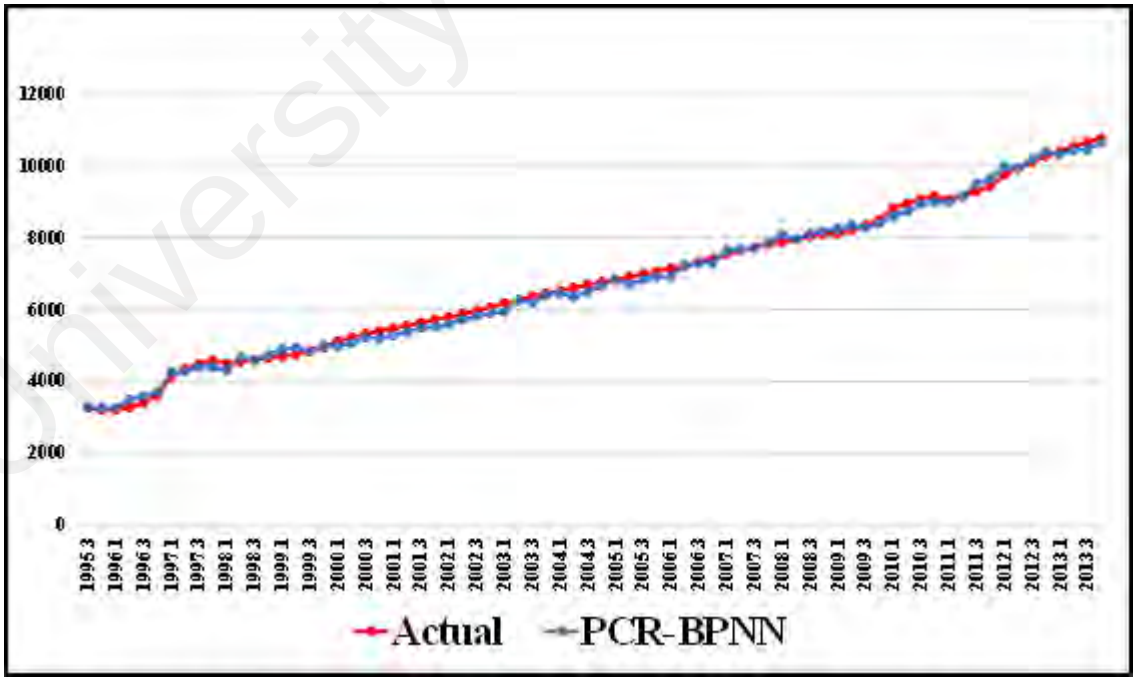


Figure 4.8: Comparison actual output with hybrid approach PCR-BPNN

University of Malaya

According to the results, RMSE and MAPE, which are the performance indicators of the hybrid system, the accuracy of PCR-BPNN is better than that of the linear model (PCR). This is because the proposed hybrid approach can capture both patterns of the input dataset. The results of the performance indicators for the linear model (PCR) are explained in Table 4.10.

4.6 Summary

The chapter presents the methodology of building the PCR-BPNN prediction model. The methodology of the study has been achieved in some steps starts with identifying variables and ends with building the proposed model. Through this methodology, 19 independent variables have been collected, but only 13 variables have been selected as significant as their correlations with the demand of electricity have passed the threshold value (0.5).

The study addressed the problems of linearity-nonlinearity patterns in dataset and multicollinearity among independent variables in the viewpoint of the accuracy. To improve the accuracy, PCA as solution for multicollinearity problem and residual error analysis for errors due to linearity-nonlinearity patterns have been discussed. The implementation of the proposed model (PCR-BPNN) has been done step by step and results for each step has been presented.

University of Malaya

CHAPTER 5: VALIDATION AND GENERALIZATION

5.1 Introduction

The new approach of the electricity demand in Malaysia has been tested in the chapter four. The model needs to be validated and generalized too. To achieve the validation, this study utilizes three types of techniques as prediction based model; PCR as linear technique, PCNN as nonlinear technique, and SVR as a hybrid system technique. Results of the PCR-BPNN based model show the outperformance compared with the other mentioned techniques (PCR, PCNN, and SVR).

The second part of this chapter presents generalization. The model tests datasets collected from two different countries (Turkey and Sweden). The countries are different in environment, weather, economic, and other significant variables. Results of generalization show that the proposed approach is valid for using in different environments.

5.2 Principal Component Neural Network Model

To test the nonlinear prediction model, this study develops the PCNN model using the principal components as inputs for the ANN model. The purpose of this part is to determine the accuracy of the prediction model and to compare it with three other models, the linear model PCR, the hybrid system SVR, and our proposed model PCR-BPNN.

The PCNN uses the same number of principal components used in the linear model PCR, as previously mentioned. In the chapter four, Table 4.3 presents the number of PCs, and Figure 4.2 illustrates the input of the dataset used for the nonlinear model PCNN to realize the PCNN model using five inputs and one output.

University of Malaya

Table 5.1 shows several testing datasets that are randomly being chosen to select the best performance indicators. Table 5.2 shows the best result of the PC-neural network based on the MSE, RMSE, and MAPE. The structure of PCNN used in this study is shown in Figure 5.1.

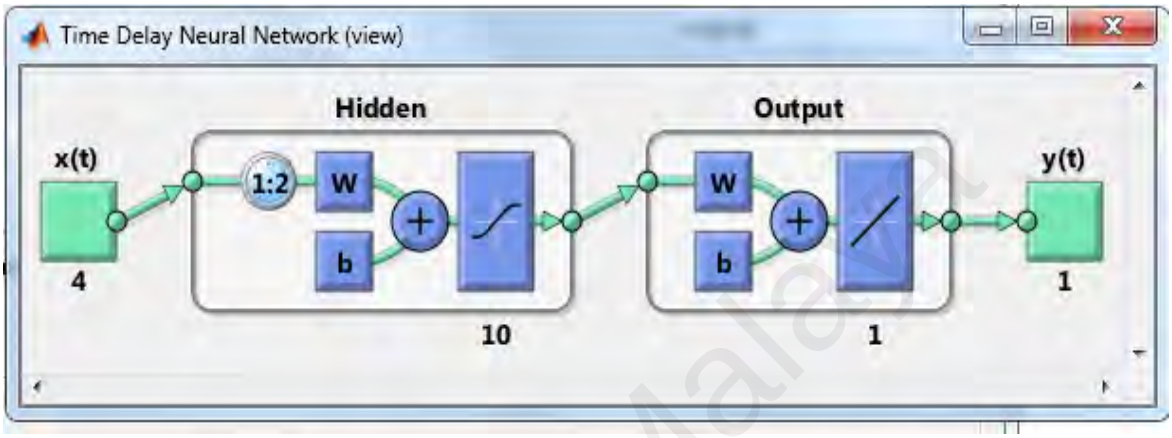


Figure 5.1: Structure of PCNN model

Table 5.1: Testing for PCs dataset

#/ test	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
1	-0.98	-1.10	-1.15	-0.72	-0.99	-1.24	-1.24	-1.02	-1.19	-1.19
2	-0.99	-0.99	-1.10	-0.68	-1.01	-1.10	-1.19	-0.88	-1.01	-0.98
3	-1.02	-1.01	-0.99	-0.54	-0.97	-1.06	-1.15	-0.72	-0.97	-0.95
4	-0.97	-0.95	-0.68	-0.51	-0.61	-1.02	-0.97	-0.61	-0.93	-0.93
5	-0.77	-0.86	-0.33	-0.25	-0.51	-0.99	-0.72	-0.58	-0.54	-0.86
6	-0.51	-0.72	0.33	-0.15	-0.19	-1.01	-0.54	-0.48	-0.48	-0.81
7	-0.15	-0.09	0.46	0.49	-0.09	-0.64	-0.41	-0.22	-0.22	-0.51
8	0.52	0.00	0.62	0.62	0.37	-0.06	0.30	-0.12	0.14	-0.37
9	0.62	0.14	0.96	1.57	0.62	0.43	2.04	0.14	0.37	0.10
10	1.81	1.34	2.04	2.49	1.21	0.49	2.27	0.26	0.69	0.49
11	2.27	2.42	2.56	2.56	2.27	0.96	2.35	0.40	0.76	1.21

Table 5.2: Measure performance indicators of PCNN model

No.	Model	MSE	RMSE	MAPE%
1	PC – neural network (PCNN)	0.011144	0.105565	16

University of Malawi

In Table 5.2, MSE, RMSE, and MAPE are used as the measures of performance. The result shows that the nonlinear model neural network (NN) based on the PCs as the input dataset is better than the linear model (PCR), which means that the ANN model can capture part of the linear dataset.

5.3 Support Vector Regression Model

This section discusses SVR-based models for the prediction of electricity demand, which have been considered in some studies (Section 2-3). The SVR model is a combination of linear regression and nonlinear support vector machine (SVM) models.

The aim of using the SVR model is to compare its accuracy with the accuracy of the other models and our proposed model; as such, the same dataset fed into other models is also fed into SVR. As mentioned from chapter four, the independent variables fed into the models are explained in Figure 4.1. However, they are then applied into the regression model, as mentioned in Table 4.10, which results in the output of PCR has two outputs: the PPEDM and the error rates. The error rates used in the SVM improve the accuracy of the prediction model. The total of the prediction model combines PPEDM and the output of the prediction error rates used in the SVM model.

Table 5.3 shows the performance indicators for SVR based on MSE, RMSE, and MAPE. To realize the prediction model in the SVR model using MATLAB 2013a in SVM requires the installation of *libsvm* from www.csie.ntu.edu.tu, which then allows the application of the SVR model in the prediction model.

Table 5.3: Measure performance indicators of SVR model

No.	Model	MSE	RMSE	MAPE%
1	Support Vector Regression (SVR)	0.00952	0.09757	14.2

University of Malaya

In Table 5.3, the results of the performance indicators, such as MSE, RMSE, and MAPE, show that the SVR is better than both the linear and nonlinear models, such as the PCR and PCNN models in analyzing complex datasets, such as linear and nonlinear dataset that proved in chapter four, tables 4.10 and current chapter 5.3 by the performance indicators of the two other approaches. However, the result of the SVR model is not better than that of our proposed model (PCR-BPNN) based on the measure of the performance indicators.

5.4 PCR-BPNN Validation

As mentioned before, the validation step is a compression process between the performance indicators of PCR-BPNN and the performance indicators of four other prediction models (PCR, PCA-BPNN, and PC-SVR). All models are tested against three types of performance indicator MSE, RMSE, and MAPE.

The first step is to find out the values of the actual and predicted electricity demand for the years of 1995 to 2013 for each above-mentioned techniques. The format of the year is taken on quarter base. The result of this step is shown in the Table 5.4. The table shows the results of each model. Here direct comparisons between values obtained from each method and the actual values can be made. From this table, it could be easy to find out the most accurate prediction model so that predictions from 2015 to 2020 could be made.

Table 5.4: Comparison of the actual and predicted electricity demand for all models

Q.years	Actual	PCR	PCNN	PC-SVR	PCR-BPNN
1995.3	3280.72	3849.01	3468.16	3150.69	3269.59
1995.4	3218.03	3646.52	3396.76	2908.54	3260.03
1996.1	3218.28	3983.23	3338.28	2936.88	3262.31
1996.2	3281.47	3860.74	3432.18	2813.51	3494.54
1996.3	3407.59	4069.65	3566.86	3011.67	3590.77
1996.4	3596.66	3936.97	3736.77	3101.17	3688.76
1997.1	4133.34	4084.87	4162.61	4011.43	4252.85

University of Malaya

Continuous table

1997.2	4334.41	4220.44	4246.51	4448.81	4276.63
1997.3	4484.53	4363.97	4451.29	4696.43	4391.06
1997.4	4583.72	4486.76	4333.53	4793.60	4390.88
1998.1	4497.59	4579.06	4632.55	4718.29	4316.85
1998.2	4548.66	4590.98	4709.02	4432.89	4674.69
1998.3	4602.53	4700.64	4791.44	4658.31	4580.00
1998.4	4659.22	4921.32	4753.73	4605.29	4740.06
1999.1	4693.09	4978.63	4795.13	4638.95	4907.64
1999.2	4765.66	4886.59	4829.72	4711.03	4943.44
1999.3	4851.28	4964.30	4954.04	4888.82	4830.61
1999.4	4949.97	4692.17	5309.88	4924.49	4992.04
2000.1	5112.50	4948.06	5380.21	5251.76	4978.17
2000.2	5217.00	5182.81	3914.05	5401.06	5065.54
2000.3	5314.25	5175.01	5417.79	5444.05	5222.48
2000.4	5404.25	5289.14	5501.07	5622.41	5203.17
2001.1	5469.19	5119.30	5574.18	5666.06	5277.27
2001.2	5551.81	5460.13	5597.09	5748.83	5369.93
2001.3	5634.31	5497.68	5136.00	5809.35	5486.89
2001.4	5716.69	5565.35	5831.19	5933.93	5512.40
2002.1	5788.94	5514.40	5864.97	5989.66	5589.28
2002.2	5875.06	5645.64	6082.09	6054.72	5706.27
2002.3	5965.06	5753.04	6053.68	6150.70	5815.20
2002.4	6058.94	5651.10	6107.26	6204.85	5893.36
2003.1	6175.91	5837.04	6278.60	6431.88	5926.82
2003.2	6269.84	5859.81	6314.46	6469.89	6264.65
2003.3	6359.97	5926.56	6398.80	6644.47	6179.93
2003.4	6446.28	5982.16	6579.70	6778.74	6417.14
2004.1	6523.94	6160.69	6620.56	6840.33	6446.73
2004.2	6604.56	6267.33	6615.65	6933.77	6351.81
2004.3	6683.31	6684.34	6716.20	6892.37	6485.20
2004.4	6760.19	6783.72	6714.84	6864.73	6654.82
2005.1	6826.12	6686.47	6829.73	6836.32	6814.23
2005.2	6902.87	6830.52	7207.65	7087.26	6704.35
2005.3	6981.37	6916.86	6964.90	7144.55	6828.30
2005.4	7061.62	6885.28	7047.54	7231.43	6914.65
2006.1	7134.56	7143.07	7110.04	7316.25	6924.75
2006.2	7221.94	7229.48	7169.84	7206.89	7230.07
2006.3	7314.69	7259.32	7304.50	7347.55	7273.68
2006.4	7412.81	7483.45	7382.92	7538.38	7277.81
2007.1	7547.25	7631.57	7550.55	7437.93	7645.36
2007.2	7643.75	7685.86	7601.07	7596.93	7678.08
2007.3	7733.25	7811.87	7698.87	7763.60	7703.90
2007.4	7815.75	8119.98	7860.67	7762.78	7862.72

Universiti Malaya

Continuous table

2008.1	7874.06	8029.31	7784.44	7634.85	8097.75
2008.2	7949.44	8125.79	7452.95	7927.01	7955.35
2008.3	8024.69	8217.56	7960.29	7924.32	8107.54
2008.4	8099.81	8535.93	8042.62	7992.63	8188.49
2009.1	8111.06	8394.76	8020.73	7762.39	8241.08
2009.2	8211.44	8258.53	8662.46	8060.87	8342.03
2009.3	8337.19	8310.12	8275.80	8386.21	8266.52
2009.4	8488.31	8284.96	8066.19	8605.76	8387.30
2010.1	8800.75	8418.91	8688.91	8960.31	8589.80
2010.2	8948.25	8492.21	8885.84	9144.35	8722.44
2010.3	9066.75	8696.79	8771.64	9374.15	8928.08
2010.4	9156.25	8923.64	9078.12	9370.14	8989.91
2011.1	9060.81	9061.16	9424.61	9174.04	8986.39
2011.2	9154.69	9391.85	9104.40	9125.60	9151.34
2011.3	9281.94	9573.27	9460.67	9037.02	9492.74
2011.4	9442.56	9837.55	9412.77	9208.47	9640.42
2012.1	9750.78	9937.30	9718.37	9378.30	9982.96
2012.2	9932.47	10125.08	9909.65	9846.77	9935.46
2012.3	10101.84	10323.94	10063.56	9941.79	10180.24
2012.4	10258.91	10301.54	10219.17	10135.23	10372.47
2013.1	10403.66	10368.99	10385.71	10438.74	10319.65
2013.2	10536.09	10410.33	10491.31	10775.35	10427.09
2013.3	10656.22	10590.08	10611.61	10816.02	10444.16
2013.4	10764.03	10609.19	10720.40	10839.33	10635.05

The second step is to find out the performance indicators for all models. The result of this step is shown in the Table 5.5. The value of the performance indicators showed that the performance of the PCR-BPNN is better than the other types of utilized prediction models (PCR, PCNN, and PC-SVR). The formula for each performance indicators which related to section 3.6 by equation (3.27), (3.28) and (3.29).

Table 5.5: Comparison of residual error for four models

Statistical parameters	**PCR**	**PCNN**	**PC-SVR**	**PCR-BPNN**
MSE	0.015998	0.011144	0.009339	0.008865
RMSE	0.126484	0.105565	0.09664	0.094155
MAPE	27	16	14.2	13

Universiti Malaya

For making visualized compression between the actual and predicted output electricity demand for all models, Figure 5-2 shows all this cases. The figure shows the strangeness of the all three models in prediction demands. Predicted output for all model somehow is close to each other and to the actual results, however, the prediction of the PCR-BPNN is shown better performance than all. Although the difference is small, it costs for billions of US and Ringgits for decision maker on electricity demand.

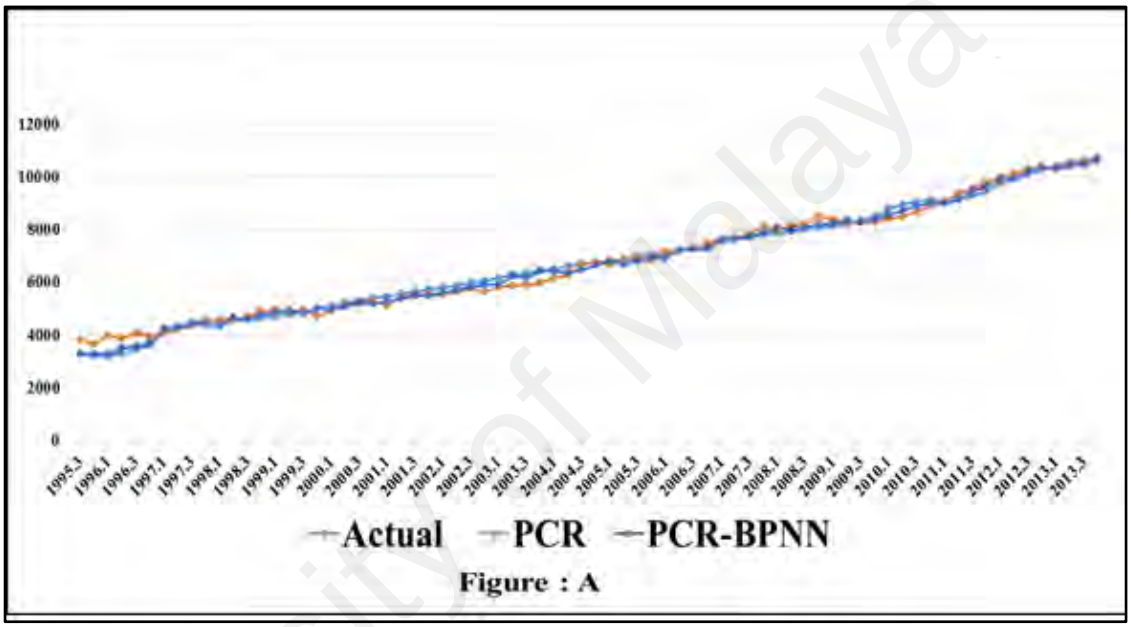


Figure : A

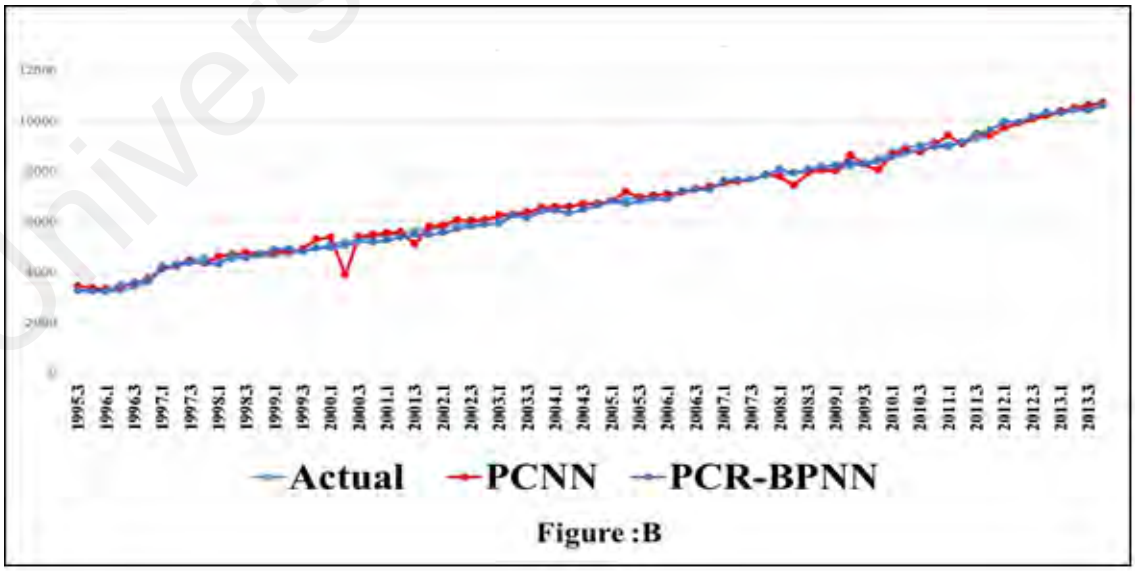


Figure :B

University of Malaya

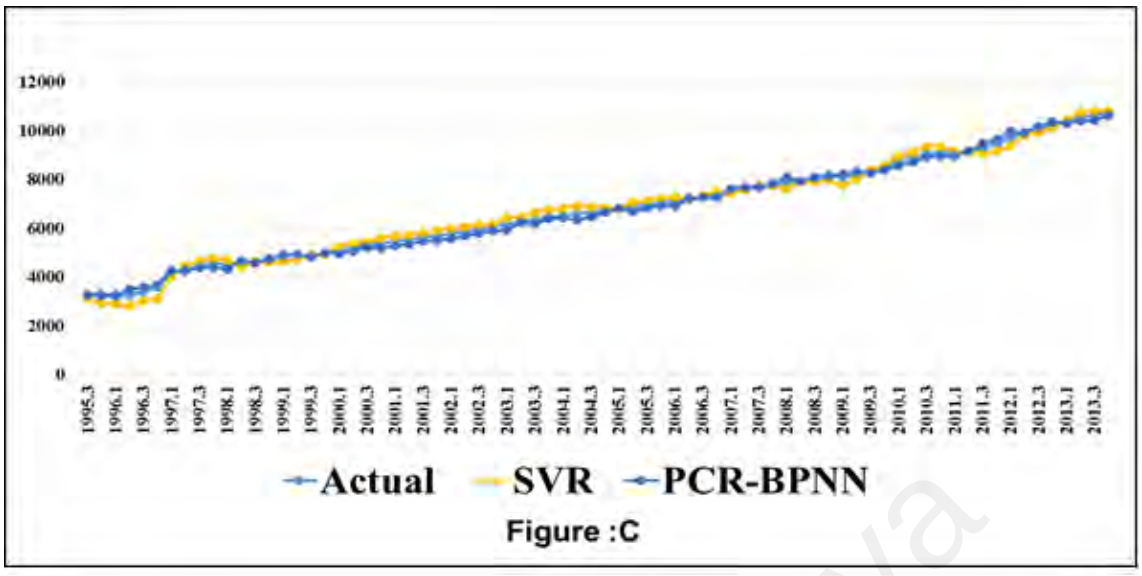


Figure 5.2: Comparison of real dataset with predicted dataset (Mtoe) in (A,B and C)

Figure 5.2 explains the accuracy comparison for three nonlinear prediction models (including the proposed PCR-BPNN) with the actual demand of electricity in Malaysia. It seems that models of PCR, PCNN, and PC-SVR are doing some fluctuations around the line of actual demand. At most points on the lines of the three graphs if the difference between the two lines of predicted demand (PCR-BPNN and other nonlinear model) with actual demand be measured, the less difference will be found between the actual demand and the line of the PCR-BPNN predictor. This means that the overall accuracy of the PCR-BPNN model is better than the accuracy of the other nonlinear predictor models. This also means that the target of the present work, which is improving the accuracy of the electricity demand prediction models, has been achieved.

To return the standardized dataset to the original dataset, the conversion of the Z-score formula in the MATLAB 2013a software is given by

mean_y= mean(y);

st_y=std(y);

*yi_predict= (Yhat_PCR*st_y)+mean_y;*

University of Malaya

Where

Mean(y) is the average of the actual output electricity demand.

Stud(y) is the standard deviation of the actual output electricity demand

yi_predict is the estimated output of each model.

According to the comparison of the aforementioned four models, the PCR-BPNN model yields more accurate predictions. The future electricity demand model for 2014–2020 is calculated for the electricity demand prediction with the estimated input variables.

The PCR-BPNN model is employed to predict future electricity demand based on the input variables, such as population, GDP, GNP, income per capita, employment, export, import, tourist arrivals, CO_2 emissions, consumer price index, climate, industrial electricity, and residential electricity. The predictions of the future electricity demand are evaluated using PCR-BPNNs from 2014–2020 are the rates of the maximum of the quarterly electricity demand. The results are given in Table 5.7 and Figure 5.3. However, the Table 5.6 which predicted for 13 independent variables by using time series for 2014 – 2020.

Table 5.6: Predicted independent variables 2014 – 2020

years	Population ('000)	GDP ('000)	GNP ('000)	income per Capita	Employement ('000)	Export	Import	Tourist arrivals (million)	CO2 emissions (kt)	Consumer price index	Mean Climate	Industrial electricity	Residential electricity
2014	29907.47	8013.85	8644.79	743.554	12167.56816	60880.93	46861.14	28456.28	182166.6	109.67814	27.7872	5314.972	2058.27585
2015	30921.43	8708.2	9468.13	791.743	12657.86146	66031.81	50696.26	31261.45	189059.4	110.92095	27.7242	5730.446	2220.01405
2016	31428.41	9055.38	9879.8	815.837	12903.00811	68607.25	52613.82	32664.03	192505.8	111.54235	27.6926	5938.182	2300.88315
2017	31935.39	9402.56	10291.5	839.932	13148.15476	71182.69	54531.38	34066.61	195952.2	112.16375	27.6611	6145.919	2381.75226
2018	32442.37	9749.74	10703.1	864.026	13393.30142	73758.13	56448.94	35469.19	199398.6	112.78515	27.6296	6353.656	2462.62136
2019	32949.35	10096.9	11114.8	888.121	13638.44807	76333.57	58366.5	36871.77	202845	113.40656	27.598	6561.392	2543.49046
2020	33456.33	10444.1	11526.5	912.215	13883.59472	78909.01	60284.06	38274.35	206291.4	114.02796	27.5665	6769.129	2624.35956

University of Malaya

Table 5.7: Predicted electricity demand for best model (PCR-BPNN)

Year	Predicted Output (Ktoe)
2014	11641.12
2015	11911.34
2016	12242.08
2017	12510.95
2018	13021.11
2019	13432.92
2020	13702.91

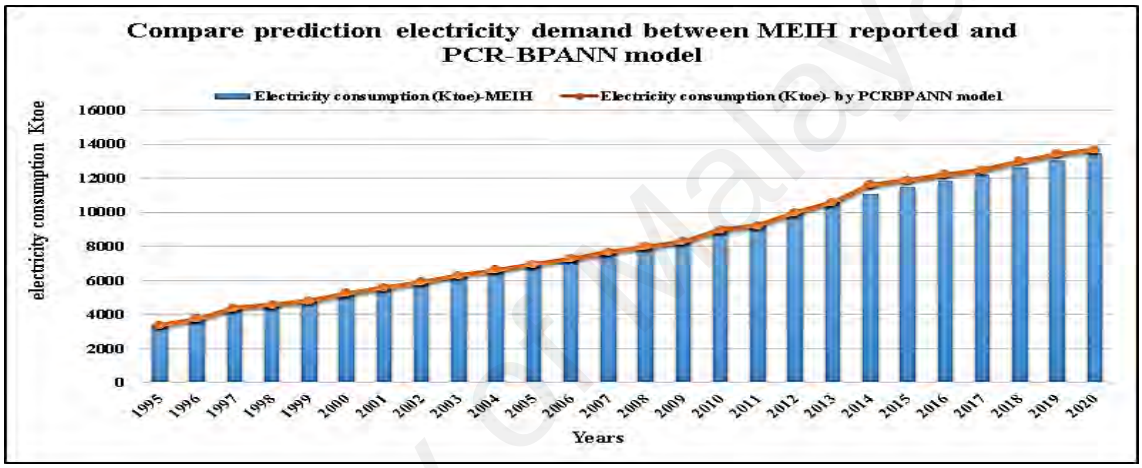


Figure 5.3: Future electricity demand prediction

The predicted model boosts the accuracy of the prediction as a result of the application of PCR-BPNN. The reports on the 18th of Malaysia Energy Information Hub (MEIH) predicted the electricity demand of 2018 to be 12649.77 Ktoe, but the predicted electricity demand calculated using PCR-BPNN is 13021.11 Ktoe.

The methods in this chapter are compared using 3 indicators and as shown in Table 5-5 and from the results obtained it can be concluded that PCR-BPANN gives better prediction than others. This is supported by Figure 5.2. Although the difference in performance (MSE) between PC-SVR and PCR-BPNN is small, predicting electricity demand accurately is of great importance in order to ensure enough electricity supply for Malaysians so that their lives will not be disrupted due to shortage of electricity. On the other hand, overestimating the electricity demand will increase the need for additional

University of Malaya

investment in the power plants. These unnecessary expenses as a result of inaccurate due diligence, translate to higher tariffs for consumers. Therefore a good prediction method will save billions of Ringgit and can avoid air pollution.

5.5 Model Generalization

Artificial Intelligent (AI) techniques usually have a generalization phase through checking the validity of the proposed AI model with different population of samples. This process supports the learning phase of an AI-based model to be more generalizable across a verity span of data. To this end, data from different sources should be collected and fed into the proposed model.

In this study, data set have been collected from three different countries: Malaysia, Turkey, and Sweden. These countries have been selected because of their differences in seasons. Malaysia has only one season, Turkey four, and Sweden three. Moreover, Malaysia and Turkey are coming in the list of the developing countries, while Sweden is considered as a developed country.

In the previous chapter, we presented the PCR-BPNN model as a new approach to build an electricity demand prediction model. The new approach has been trained and tested using data set collected from Malaysia. To validate the generalization of the proposed model, this study fed the proposed approach with different data sets that collected from two different countries; Sweden and Turkey. The following sections explain the process of checking the generalization of the PCR-BPNN model in predicting electricity demand for long term through using data sets that specialized to Sweden and Turkey countries.

Although, the collected data sets for Turkey and Sweden come with different records and patterns of data inside, the independent variables that considered as predictors in these

University of Malaya

two data sets are the same as used for Malaysia data set. Only through this process, the generalization of the new approach can be tested against the variety of data sets and expanded population samples. Therefore, the same predictors that proposed in the Section 5.6 for Malaysia prediction model are selected as predictors for the Turkey and Sweden data sets and fed to the PCR-BPNN model without applying any data preparation processes.

The first step in this work is removing the correlation and minimizing the dimensionality of the input data set through using PC method. After passing both input data sets (for Sweden and Turkey) under the process of PC, results as shown in Table 5.8 and Table 5.9 are obtained, and Figure 5.4 and Figure 5.5 are shown that relation between the eigenvalues and the component numbers of PC for both Sweden and Turkey data set. From the results and figures, the adequate PC number for each data set could be obtained. For the Sweden data set, the suitable number of PCs is six, while for the data set that belongs to Turkey this number is becoming four.

Table 5.8: Total variance - Sweden

Total Variance Explained						
#	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**		
	Total	**Variance %**	**Cumulative %**	**Total**	**Variance %**	**Cumulative %**
PC1	6.947	53.441	53.441	6.947	53.441	53.441
PC2	1.524	11.723	65.164	1.524	11.723	65.164
PC3	1.175	9.038	74.202	1.175	9.038	74.202
PC4	1.033	7.948	82.151	1.033	7.948	82.151
PC5	.931	7.164	89.314	.931	7.164	89.314
PC6	.788	6.060	95.375	.788	6.060	95.375
PC7	.249	1.915	97.289	.249	1.915	97.289
PC8	.175	1.345	98.634	.175	1.345	98.634
PC9	.085	.657	99.291	.085	.657	99.291
PC10	.052	.397	99.688	.052	.397	99.688
PC11	.022	.168	99.856	.022	.168	99.856
PC12	.010	.079	99.935	.010	.079	99.935
PC13	.008	.065	100.000	.008	.065	100.000

University of Malaya

Table 5.9: Total variance - Turkey

Total Variance Explained						
#	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	Variance%	Cumulative %	Total	Variance%	Cumulative %
PC1	9.460	72.767	72.767	9.460	72.767	72.767
PC2	1.218	9.366	82.133	1.218	9.366	82.133
PC3	.983	7.564	89.697	.983	7.564	89.697
PC4	.849	6.531	96.228	.849	6.531	96.228
PC5	.269	2.071	98.299	.269	2.071	98.299
PC6	.095	.734	99.033	.095	.734	99.033
PC7	.056	.430	99.463	.056	.430	99.463
PC8	.032	.245	99.708	.032	.245	99.708
PC9	.016	.124	99.833	.016	.124	99.833
PC10	.011	.087	99.919	.011	.087	99.919
PC11	.006	.043	99.962	.006	.043	99.962
PC12	.003	.024	99.986	.003	.024	99.986
PC13	.002	.014	100.000	.002	.014	100.000

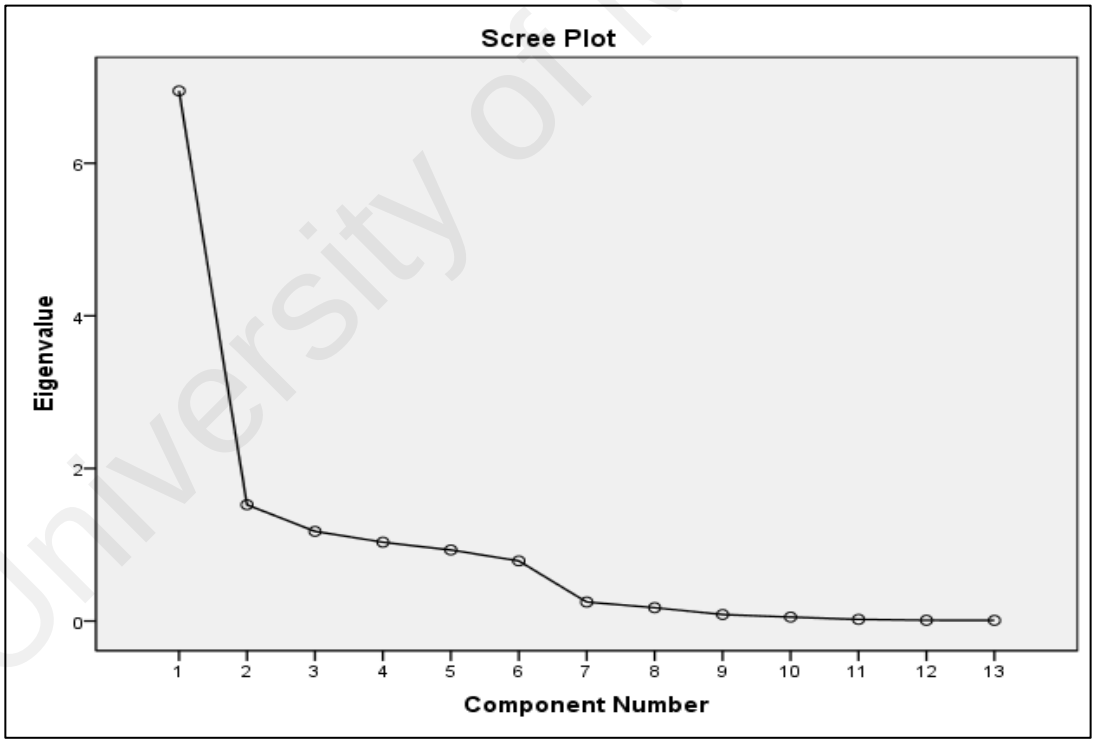


Figure 5.4: Number of components and eigenvalue – Sweden

University of Malaya

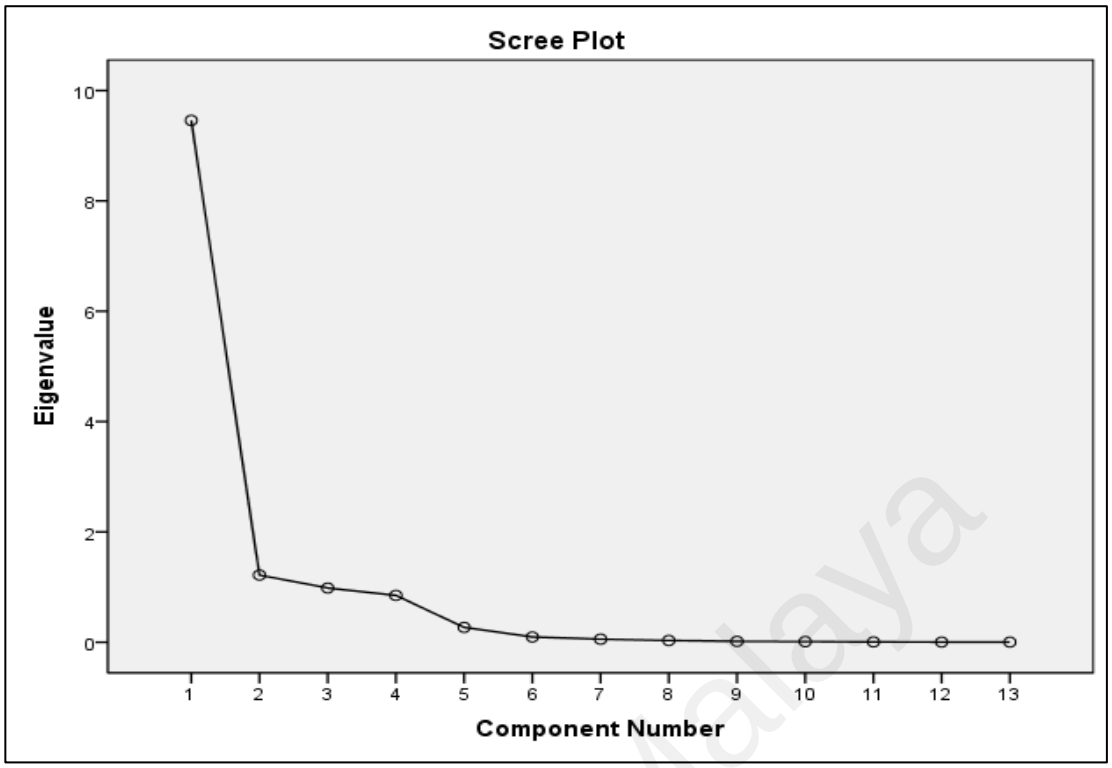


Figure 5.5: Number of components and eigenvalue – Turkey

The decision that made on the number of included PCs is going back to the cumulative percentage and eigenvalues for each obtained PC. The cumulative percentage that considered for both data sets are started from 53% and 72% for Sweden and Turkey respectively and increased up to around 96% for both.

Table 5.10 represents the component score coefficient matrix reveals the system information of principal components with the original independent variables. The scores demonstrate the relative importance of each standardized predictors in the PC calculations in Sweden and Table 5.11 shows the same relative information impotence for Turkey dataset.

University of Malaya

Table 5.10: Component matrix - Sweden

Component Matrix						
	Component					
	PC1	PC2	PC3	PC4	PC5	PC6
Population	0.886	0.416	-.006	-0.020	0.016	0.012
GDP	-0.293	-0.205	-.302	0.383	0.762	0.222
GNP	0.104	-0.128	.739	0.080	0.370	-0.053
Export	0.942	-0.135	-.053	0.088	0.067	0.123
Import	0.961	-0.020	-.089	0.087	0.063	0.055
Employment	0.955	0.012	.073	-0.059	0.031	0.025
CO2_ emission	0.448	-0.826	-.011	0.102	-0.163	0.012
Climate	-0.205	0.043	.072	0.154	-0.073	0.640
Tourism	-0.038	0.156	-.039	0.907	-0.341	-0.172
Income per capita	0.985	-0.109	.008	0.011	-0.036	0.014
Industrial electricity	-0.828	0.450	-.037	0.048	0.018	-0.019
Residential electricity	0.718	0.561	.012	0.036	0.225	0.019
Consumer Price Index	0.964	0.187	.009	0.035	-0.051	-0.008
Extraction Method: Principal Component Analysis.						

Table 5.11: Component matrix - Turkey

Component Matrix				
	Component			
	PC1	PC2	PC3	PC4
Population	0.942	0.007	-0.012	0.119
GDP	0.083	0.726	-0.276	0.616
GNP	0.060	0.721	-0.133	0.764
Export	0.968	0.025	0.055	-0.088
Import	0.980	0.047	-0.024	-0.079
Employment	0.905	-0.007	-0.047	0.145
CO2_ emission	0.984	-0.042	0.010	-0.038
Climate	0.050	0.321	0.940	0.106
Tourism	0.986	-0.002	-0.008	-0.035
Income per capita	0.977	0.012	0.005	-0.113
Industrial electricity	0.994	-0.005	-0.002	-0.047
Residential electricity	0.992	-0.060	0.011	0.010
CPI	0.986	0.001	-0.025	0.040
Extraction Method: Principal Component Analysis.				

According to the component scores of variables, information of all 13 independent variables is accumulated in six PCs for Sweden's data set while for the Turkey data set

University of Malaya

the information are accumulated in four PCs, as shown in the Table 5.8 and Table 5.9. All PCs are not loaded with the same information rate. For Sweden's dataset as an example, all thirteen independent variables were included in the six selected PCs. However, purely certain variables showed high loadings within each PC; the first PC is weightily loaded on population, export, import, employment, income per capita, industrial electricity, residential electricity and consumer price index. The second PC is heavily loaded with CO_2 emission only. The third PC is heavily loaded on the GDP factor, while the forth PC is heavily loaded with tourism factor; the fifth PC is heavily loaded with GDP factor. The last PC is heavily loaded with climate factor. For weighting the PCs against the rate of information loading the same scenario that followed for Sweden dataset is used for Turkey dataset too. The population, export, import, employment, CO_2-emission, Tourism, Income per capita, industrial electricity, residential electricity and consumer price index CPI that more related with the first Principal component. Second PC is heavily loaded with GDP. The third PC is heavily loaded with the Climate factor. Finally, the forth PC is heavily loaded with GNP factor.

The next step is attaching the PCs (less multicollinearity components) to linear regression model to form principle component regression (PCR). In this study PCR can receive PCs to obtain the electricity prediction rate. For the Sweden data set, PCR receives six components, while for Turkey PCR receives four components. The output of the PCR is considered as a preliminary prediction of electricity demand rate. Mathematically, the PCR equation is formulated based on the included PCs number. Accordingly, the PCR equation for Sweden data set involves six independent variables equation 5.1, while the same equation for the Turkey data set involves four independent variables equation 5.2.

$$PCR = 0.67\,3 \times PC1 + 0.410 \times PC2 + (-0.019) \times PC3 + 0.095 \times PC4 + (-0.164) \times PC5 + (-0.083) \times PC6 \ldots\ldots\ldots\ldots (5.1)\ for\ Sweden$$

University of Malaya

$$PCR = 1.004 \times PC1 + (-0.0308) \times PC2 + (-0.0024) \times PC3$$
$$+ (-0.0147) \times PC4 \ldots\ldots\ldots (5.2)\ for\ Turkey$$

For the equation 5.1, the components of 0.673, 0.410, -0.019, 0.095, -0.164 and -0.083 are called regression coefficients of the PCR model, and they are the value of the PC1, PC2, PC3, PC4, PC5 and PC6 for Sweden data set. However, the regression coefficients for Turkey data set are 1.004,-0.0308, -0.0024 and -0.0147 which they are assigning the value of PC1, PC2, PC3, PC4. Using the equation 5.1 and 5.2, preliminary prediction model and residual errors for the electricity demand in Sweden and Turkey are calculated and a part of these results are presented in Table 5.12 All results are presented in the Appendix F.

Table 5.12: Result of preliminary prediction model for both countries

Sweden			Turkey	
Obs.	Predicted E.D	Residuals	Predicted E.D	Residuals
1995-01-01	-1.187	1.313	-1.54491	0.099485
1995-04-01	-1.168	1.312	-1.48802	0.051485
1995-07-01	0.134	0.026	-1.46635	0.048583
1995-10-01	0.296	-0.126	-1.42524	0.023173
1996-01-01	-0.924	1.121	-1.39313	0.027256
1996-04-01	-0.922	1.110	-1.35714	0.022749
1996-07-01	-0.870	1.035	-1.31718	0.045037
1996-10-01	-0.873	1.026	-1.2759	0.036374
1997-01-01	-0.740	0.877	-1.20534	0.015633
1997-04-01	-0.742	0.430	-1.12936	-0.01877
1997-07-01	-0.926	0.429	-1.16827	0.054745
1997-10-01	-0.870	0.134	-1.06597	0.012172
1998-01-01	-0.863	-0.090	-1.0222	-0.01571
1998-04-01	-0.869	-0.107	-1.00748	0.020298
1998-07-01	-0.954	-0.314	-0.97475	-0.00349
1998-10-01	-0.812	-0.588	-0.92562	0.011693
1999-01-01	-0.708	-0.780	-0.87171	-0.02892
1999-04-01	-0.849	-0.709	-0.91169	0.015817

University of Malaya

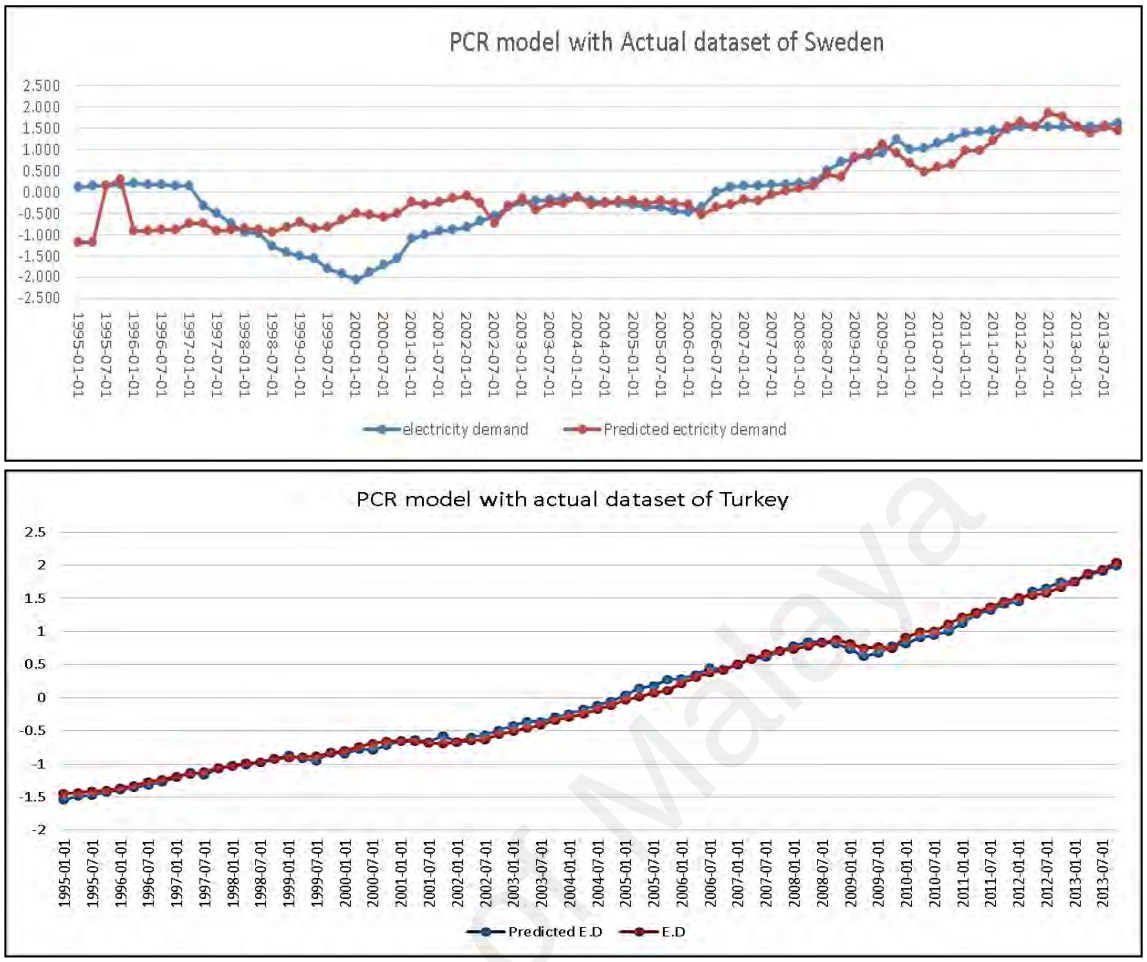


Figure 5.6: Actual and predicted PCR model – Sweden and Turkey

Figure 5.6 provides all the rates of the electricity demand for the period of 1995 to 2013 for the Sweden and Turkey dataset. Based on the figure 5.6, the actual consumption and predicted electricity demand result some errors, which named as residuals errors. Residual errors for Sweden are very high compared to the rate the obtained for the dataset of Turkey still.

To improve the accuracy, the residual errors will be more analyzed through Back-propagation artificial neural network, which is a nonlinear approach. The network works based of the equation 5.3

$$\boldsymbol{e_{(t)} = f(e_{(t-1)}, e_{(t-2)}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots e_{(t-d)})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.3)$$

University of Malaya

Figure 5.7 shows the performance indictor for training dataset. The best validation performance indictor is 0.034711 at epoch 13 for the Sweden dataset, while for the Turkey dataset best validation performance indictor is 0.0015929 at epoch 2

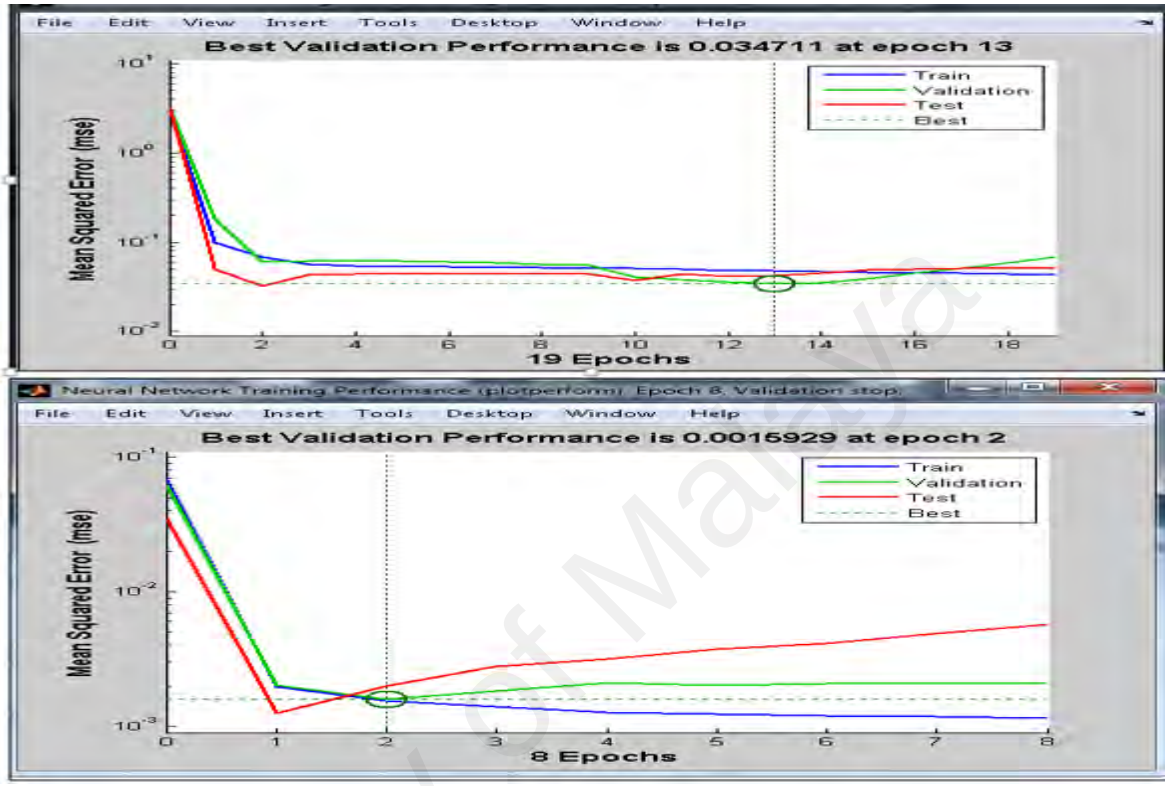


Figure 5.7: Best validation of performance for Sweden and Turkey

Figure 5.8 shows the regression status for training, validation and testing phases. The figure shows a good fit with an actual dataset of 0.91887, 0.97709 and 0.89099 respectively for Sweden. While, for Turkey a good fit is 0.76, 0.83 and 0.73 respectively.

University of Malaya

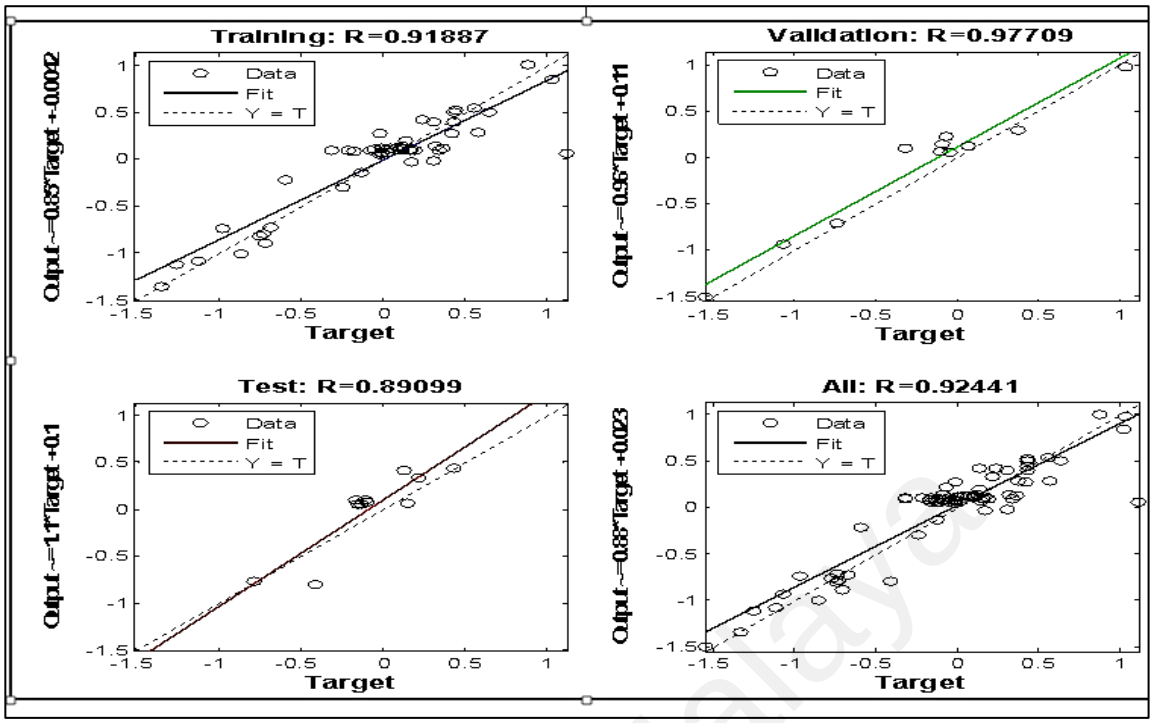


Figure 5.8: Regression for both Sweden and Turkey

Table 5.13 provides the overall prediction results of the PCR-BPNN model. The PCR subpart of the model is a linear function and results the preliminarily prediction rate. Later, analyzing the residual error is done through BPNN, which has nonlinear functioning. The combination of both subparts improves the accuracy of the prediction model. The Appendix F shows all the result of PCR-BPNN model.

Table 5.13: Hybrid approach PCR-BPNN model for Sweden and Turkey

Sweden country			Turkey country		
PCR model	**error in BPNN**	**PCR+BPNN**	**PCR model**	**error in BPNN**	**PCR+BPNN**
-1.187	0.069	-1.118	-1.488	0.094	-1.394
-1.168	-0.136	-1.304	-1.466	0.082	-1.406
0.134	0.060	0.194	-1.425	0.075	-1.391
0.296	1.124	1.420	-1.393	0.094	-1.331
-0.924	0.980	0.056	-1.357	0.093	-1.300
-0.922	0.848	-0.074	-1.317	0.100	-1.258
-0.870	1.004	0.134	-1.276	0.098	-1.219
-0.873	0.503	-0.369	-1.205	0.093	-1.183
-0.740	0.402	-0.337	-1.129	-0.001	-1.207
-0.742	0.417	-0.325	-1.168	0.014	-1.116
-0.926	0.140	-0.786	-1.066	0.094	-1.074

University of Malaya

Finally, Table 5.14 shows the most important performance indicator for the PCR-BPNN model to predict electricity demand for long term on the Sweden and Turkey

Table 5.14: Accuracy of PCR-BPNN model different countries

#	Country	MSE	RMSE	MAPE
1	Turkey	0.022	0.044	1.57
2	Sweden	0.113	0.335	1.137

5.6 Summary

This chapter presented a new approach to predict electricity demand, which combines the PCR and BPNN techniques and other models based on the PCA input dataset. The technique used PCA to remove the multicollinearity problem in the independent variables. PCA computes the number of PCs that can be found based on the accumulation of the variance test.

A prediction model has been tested with PCA techniques, which is MLR, to build a preliminary model of prediction electricity demand. Through this step a PCR prediction technique has been built. The residual error rates from the PCR were used in BPNN to improve the prediction model. The overall prediction model was combined to develop the PCR-BPNN model presented in the chapter. Using the three indicators (MSE, RMSE and MAPE) and graphs, it can be concluded that PCR-BPNN gives better prediction of electricity demand compared to other methods. Furthermore, the method can also be used to predict electricity demand in other countries with different characteristics of input dataset.

University of Malaya

CHAPTER 6: FUTURE WORKS AND CONCLUSIONS

6.1 Introduction

This chapter summarizes the study on the prediction of electricity demand and discusses its major contributions and suggestions to the field. Section 6.2 details the achievements of this study by showing all the defined objectives and their respective validations. The section explains and illustrates the way the objectives have been realized. Section 6.3 provides some future work and recommendations for the field of electricity demand prediction. This study has found that most researchers predict electricity demand via three approaches: linear, nonlinear, and hybrid. These approaches are proposed based on the original independent variables. However, the hybrid approach has been proposed to overcome both the linearity and non-linearity problems. Furthermore, PCA is used to reduce multicollinearity problem in the data. As such, more significant input data can be included in the analysis.

The approaches that have been proposed by previous researchers focus on determining techniques that can provide high-accuracy prediction. Researchers have sought to improve the accuracy of the models using alternative linear and nonlinear tools without accounting for the patterns of the dataset. Therefore, the selection of the method used in this study is based on the following:

- The patterns of the dataset, both linear and nonlinear input data;
- Ability to include all relevant input data and reduce multicollinearity problem;
- Ability to provide the most accurate prediction of electricity demand.

6.2 Achievement of Research Objective

The objectives of this study are given in Section 1.4. Each objective focuses on the process of improving the accuracy of the prediction models for electricity demand. Listed

University of Malaya

below are the illustrations of the objectives that are relevant to their effect on minimizing the errors of models that predict electricity demand rates.

Objective 1: To investigate the relationship between different input dataset patterns and electricity demand

The first objective is to collect information on the relations between the performance indicators and the size of the input dataset or the dissimilarity patterns of the input dataset (i.e., linear and nonlinear). The study initiated by this objective determines how the types of models proposed by researchers and the tools used in their methods are affected by the patterns and the dimensional size of the dataset. This objective also seeks to identify the relationship between the pattern input dataset and proposed prediction model to satisfy the accuracy of the prediction. The results shown in Appendix E can validate this argument.

Objective 2: To reduce dataset complexity and then improves accuracy of electricity demand prediction

This is for a complex problem with both linear and nonlinear correlation structures. Therefore, one of the techniques for avoiding the complexity problem is the PCA. This technique reduces the number of the input dataset, eliminates the collinearity problem in the independent variables, and decreases the complexity of the model. The performance indices are higher, using the PCA technique as input dataset for the model as compared with the original dataset.

Objective 3: **To assess the accuracy of the developed prediction model.**

Both objectives concerning the design and implementation of the validation of the PCR-BPNN model are targeted here. The objective is realized in the context of residual error rates and the performance of the methods are investigated using three indicators.

University of Malaya

The use of PCA allows all relevant input data to be included in the analysis and reduce the multicollinearity problem. The results of PCA are used as input data and render the independent variables to be uncorrelated. The combination of PCR-BPNN gives better prediction of electricity demand as shown in Chapter 4. Although the results obtained from other methods such as PC-SVR approach give almost similar results, it is noted that an accurate prediction is desirable as a little variation in the predicted values give serious implication.

6.3 Suggestions for Future Studies

Several studies have been conducted on prediction models for electricity demand, PCR, BPNN, and SVR. The present thesis extends these studies a little further, but for unavoidable reasons, not all aspects could be covered. This, we believe, forms the basis for future research in this area. The suggestions given below may be classified into two main categories: selection of independent variables and model.

6.3.1 Involving more Independent Variable

As mentioned in Section 3.4.1, this study employs 13 factors that directly affect electricity demand in the prediction model, and these factors affect the increase in electricity demand. In future, some factors that decrease future electricity demand can be determined. One factor that is more significant in decreasing future electricity demand is energy house. It is a new technology that is being installed for any type of house or building that uses less energy. However, this variable is still unpopular.

We suggest that the energy house is one important factor that should be used as an independent variable in computing electricity demand because it directly affects the decrease in electricity consumption and electricity bills. This may result in an accurate and stable picture for electricity demand.

University of Malaya

6.3.2 Modeling for Different Applications

First, this study uses the sigmoid function with k-fold cross validation from the nonlinear model side (BPNN) to improve the accuracy of the prediction model for electricity demand. Previous studies on electricity demand using BPNN suggest utilizing the tangential sigmoid called *tansig* function, but in this study, the logistic sigmoid is called *logsig* function, which performs better than the *tansig* function. Therefore, studies should focus on other activation functions for prediction models for electricity demand.

Second, the PCR-BPNN model can be suggested and recommended to be used as prediction models in several areas, such as water, oil, and gas demands. This is because PCR-BPNN is able to capture both linear and nonlinear datasets and remove the multicollinearity in the independent variables. Therefore, in the area of electricity demand, PCR-BPNN based on the sigmoid function presents opportunities for further research in demand optimization.

This research is conducted to provide the current demand model with a more accurate input, but further research should be aimed at enabling far better capturing of the demand model.

6.4 Conclusion

This work proposed a novel prediction model known as the PCR-BPNN. This model is able to predict electricity demand based on certain factors, such as population, GDP, GNP, income per capita, employment, export, import, tourist arrivals, CO_2 emissions, consumer price index, climate, industrial electricity, and residential electricity. These predictors were selected via the calculation of the correlation coefficient. This study also utilizes four different approaches that regard PCs as independent variables, such as PCR, PCNN, PC-SVR, and PCR-BPNN.

University of Malaya

This study uses five PCs as uncorrelated independent variables, and the variables of the PCs are associated with all of the original variables. The output of the PCA technique for the prediction model can be regarded as the independent variables, and it takes into account all relevant input data.

This study demonstrates that a most accuracy prediction model for electricity demand could be obtained from using PCR as a variable selection approach in identifying the most appropriate explanatory variable subset data for the regression model of electricity demand, followed by applying a BPNN technique on the resulting residual errors. The BPNN component of the combined model is able to fit more accurately the remaining non-linearity in the residuals, which the PCR analysis fails to capture. The combined PCR and BPNN significantly improve the predicted accuracy of the electricity demand models in the long term. Table 4-17 illustrates the results for the four methods in terms of MSE, RMSE, and MAPE to determine the best prediction model. Based on the independent variables of the PCs, we conclude that the hybrid approach provides better results than the linear and nonlinear models in predicting the electricity demand.

Obtained from this study can be used by policy makers to better predict electricity demand in Malaysia. Based on the report provided by PMES, it was concluded that actual demand of electricity never matched the demand that predicted by PMES. Since 1995 every year a percentage of error between the predicted and actual demand of electricity in Malaysia is recorded. These errors occurred due to the weak ability of the approach that was used by PMES to estimate the demand prediction. Of course errors in prediction costs Malaysia billions of Ringgits and minimizing these errors saves Malaysia from these losses. The proposed PCR-BPNN model for prediction of electricity demand proved its ability to minimize the residual errors. Therefore, PMES should use the results obtained in this study to predict their future electricity demand.

University of Malaya

REFERENCES

Abdul-Wahab, S. A., Bakheit, C. S. & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10), 1263-1271.

Abdul Hamid, M. B., & Abdul Rahman, T. K. (2010, 24-26 March 2010). Short Term Load Forecasting Using an Artificial Neural Network Trained by Artificial Immune System Learning Algorithm. Paper presented at the Computer Modelling and Simulation (UKSim), 2010 *12th International Conference on.*

Abdulalla, S. M., Kiah, Laiha Mat, & Zakaria, Omar. (2010). A biological model to improve PE malware detection: Review. *International Journal of Physical Sciences*, 5(15), 2236-2247.

Abiyev, R. H. (2009). Fuzzy wavelet neural network for prediction of electricity consumption. *Energy Procedia*, 17(Part B), 1332–1338.

Abiyev, R. H., & Altunkaya, K. (2009). Neural Network based Biometric Personal Identification with fast iris segmentation. *International Journal of Control, Automation and Systems*, 7(1), 17-23.

Aggarwal, S. K., Saini, L. M., & Kumar, A. (2009). Electricity price forecasting in deregulated markets: A review and evaluation. International Journal of Electrical Power & *Energy Systems*, 31(1), 13-22.

Ahmmed, S., Rahman, D. M. F., Hasan, M. K., Saber, A. Y., & Rahman, M. Z. (2009). Computational intelligence approach to load forecasting-a practical application for the desert of Saudi Arabia. Paper presented at the Computers and Information Technology, 2009. ICCIT'09. *12th International Conference on*.

Akay, D., & Atak, M. (2007). Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. *Energy*, 32(9), 1670-1675.

Al-Alawi, S. M., Abdul-Wahab, S. A., & Bakheit, Charles S. (2008). Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, 23(4), 396-403.

Al-Ghandoor, A., Al-Hinti, I., Jaber, J. O., & Sawalha, S. A. (2008). Electricity consumption and associated GHG emissions of the Jordanian industrial sector: Empirical analysis and future projection. *Energy Policy*, 36(1), 258-267.

Alamaniotis, M., Ikonomopoulos, A., & Tsoukalas, L. H. (2011). A pareto optimization approach of a Gaussian process ensemble for short-term load forecasting. Paper presented at the Intelligent System Application to Power Systems (ISAP), 2011 *16th International Conference on*.

Altinay, G., & Karagol, E. (2005). Electricity consumption and economic growth: evidence from Turkey. *Energy Economics*, 27(6), 849-856.

University of Malaya

Aman, S., Ping, H. W., & Mubin, M. (2011). Modelling and forecasting electricity consumption of Malaysian large steel mills. *Scientific Research and Essays*, 6(8), 1817-1830.

Amjadi, M. H., Nezamabadi-pour, H., & Farsangi, M. M. (2010). Estimation of electricity demand of Iran using two heuristic algorithms. *Energy Conversion and Management,* 51(3), 493-497.

Aqeel, A., & Butt, M. S. (2001). The relationship between energy consumption and economic growth in Pakistan. *Asia-Pacific Development Journal*, 8(2), 101-110.

Aranda, A., Ferreira, G., Mainar-Toledo, M. D., Scarpellini, S., & Llera Sastresa, E. (2012). Multiple regression models to predict the annual energy consumption in the Spanish banking sector. *Energy and Buildings*, 49, 380-387.

Ardakani, F. J., & Ardehali, M. M. (2014). Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types. *Energy*, 65(0), 452-461.

Areekul, P., Senjyu, T., Toyama, H., & Yona, A. (2010). Notice of Violation of IEEE Publication Principles
A Hybrid ARIMA and Neural Network Model for Short-Term Price Forecasting in Deregulated Market. Power Systems, *IEEE Transactions on*, 25(1), 524-530.

Asafu-Adjaye, J. (2000). The relationship between elasticity consumption, electricity prices and economics growth: Time series evidence from Asian developing countries. *Energy Economics*, 22, 615–625.

Asteriou, D., & Hall, S.G. (2011). *Applied Econometrics*: Palgrave Macmillan.

Aydin, G. (2014). Modeling of energy consumption based on economic and demographic factors: The case of Turkey with projections. *Renewable and Sustainable Energy Reviews*, 35(0), 382-389.

Azadeh, A., Saberi, M., Ghaderi, S. F., Gitiforouz, A., & Ebrahimipour, V. (2008). Improved estimation of electricity demand function by integration of fuzzy system and data mining approach. *Energy Conversion and Management*, 49(8), 2165-2177.

Azadeh, A., Ghaderi, S. F., & Sohrabkhani, S. (2007). Forecasting electrical consumption by integration of Neural Network, time series and ANOVA. *Applied Mathematics and Computation*, 186(2), 1753-1761.

Azadeh, A., Ghaderi, S. F., & Sohrabkhani, S. (2008). Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. *Energy Conversion and Management*, 49(8), 2272-2278.

Azadeh, A., Ghaderi, S. F., Tarverdian, S., & Saberi, M. (2007). Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. *Applied Mathematics and Computation*, 186(2), 1731-1741.

University of Malaya

Azadeh, A., Saberi, M., & Gitiforouz, A. (2013). An integrated fuzzy mathematical model and principal component analysis algorithm for forecasting uncertain trends of electricity consumption. *Quality & Quantity*, 47(4), 2163-2176.

Azadeh, A., Saberi, M., Gitiforouz, A, & Saberi, Z. (2009). A hybrid simulation-adaptive network based fuzzy inference system for improvement of electricity consumption estimation. *Expert Systems with Applications*, 36(8), 11108-11117.

Badr, E. A., & Nasr, G. E. (2001). On the relationship between electrical energy consumption and climate factors in Lebanon: co-integration and error-correction models. *International journal of energy research*, 25(12), 1033-1042.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *OR*, 451-468.

Bazmia, A. A., Davoodya, M., & Zahedia, G. (2012). Electricity Demand Estimation Using an Adaptive Neuro-Fuzzy Network: A Case Study from the State of Johor, *Malaysia. International Journal*, 3(4), p284.

Benaouda, D., Murtagh, F., Starck, J. L., & Renaud, O. (2006). Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting. *Neurocomputing*, 70(1–3), 139-154.

Bessec, M., & Fouquau, J. (2008). The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach. *Energy Economics*, 30(5), 2705-2721.

Bhurtun, C., Jahmeerbacus, I., & Jeewooth, C. (2011). Short term load forecasting in Mauritius using Neural Network. Paper presented at the Industrial and Commercial Use of Energy (ICUE), 2011 *Proceedings of the 8th on the Conference*

Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. *Energy*, 34(9), 1413-1421.

Chandran, V. G. R., Sharma, S., & Madhavan, K. (2010). Electricity consumption–growth nexus: The case of Malaysia. *Energy Policy*, 38(1), 606-612.

Chen, B. J., & Chang, M. We. (2004). Load forecasting using support vector machines: A study on EUNITE competition 2001. Power Systems, *IEEE Transactions on*, 19(4), 1821-1830.

Chen, S. T., Kuo, H. I., & Chen, C. C. (2007). The relationship between GDP and electricity consumption in 10 Asian countries. *Energy Policy*, 35(4), 2611-2621.

Chen, Y. L., Peter, B., Guan, C. Z., Yige, M., Laurent, D., Coolbeth, M. A., . . . Rourke, S. J. (2010). Short-term load forecasting: Similar day-based wavelet neural networks. Power Systems, *IEEE Transactions* on, 25(1), 322-330.

Chia, K. S., Rahim, H. A., & Rahim, R. A. (2011). A comparison of Principal Component Regression and Artificial Neural Network in fruits quality prediction. Paper presented at the Signal Processing and its Applications (CSPA), 2011 *IEEE 7th International Colloquium* on.

University of Malaya

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4), 559-583.

Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association*, 47(6), 653-663.

Dalvand, M. M., Azami, S., & Tarimoradi, H. (2008). Long-term load forecasting of Iranian power grid using fuzzy and artificial neural networks. Paper presented at the Universities *Power Engineering Conference*, 2008. UPEC 2008. 43rd International.

Darbellay, G. A., & Slama, M. (2000). Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting,* 16(1), 71-83.

Demuth, H., Beale, M., & Hagan, M. (2008). *Neural network toolbox™* 6. User's guide.

Dincer, I., & Dost, S. (1996). Energy intensities for Canada. *Applied Energy*, 53(3), 283-298.

Draper, N. R., & Smith, Harry. (1981). *Applied regression analysis 2nd ed*.

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics & Data Analysis*, 52(4), 2228-2237.

Ediger, V. Ş, & Tatlıdil, H. (2002). Forecasting the primary energy demand in Turkey and analysis of cyclic patterns. *Energy Conversion and Management*, 43(4), 473-487.

Egelioglu, F., Mohamad, A. A., & Guven, H. (2001). Economic variables and electricity consumption in Northern Cyprus. *Energy*, 26(4), 355-362.

Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. *Energy*, 35(2), 512-517.

Erdogdu, E. (2007). Electricity demand analysis using cointegration and ARIMA modelling: A case study of Turkey. *Energy Policy*, 35(2), 1129-1146.

Ersel C., Olcay, C., Halim, K. O., Harun, & Hepbasli, A. (2004). Energy demand estimation based on two-different genetic algorithm approaches. *Energy Sources*, 26(14), 1313-1320.

Fadare, D. A., & Dahunsi, O. A. (2009). Modeling and Forecasting of Short-Term Half-Hourly Electric Load at the University of Ibadan, Nigeria. *Pacific Journal of Science and Technology*, 10(2), 471-478.

Fatai, K., Oxley, L., & Scrimgeour, F. G. (2003). Modeling and forecasting the demand for electricity in New Zealand: a comparison of alternative approaches. *The Energy Journal*, 24(1), 75-102.

University of Malaya

Fekedulegn, B. D., Colbert, J. J., Hicks Jr, R. R., & Schuckers, M. E. (2002). Coping with multicollinearity: An example on application of principal components regression in dendroecology. *Research & Development Treesearch*, 43p.

Ferguson, R., Wilkinson, W., & Hill, R. (2000). Electricity use and economic development. *Energy Policy*, 28(13), 923-934.

Garen, D. C. (1992). Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management*, 118(6), 654-670.

Geem, Z. W., & Roper, W. E. (2009). Energy demand estimation of South Korea using artificial neural network. *Energy Policy*, 37(10), 4049-4054.

Ghanbari, A., Naghavi, A., Ghaderi, S. F., & Sabaghian, M. (2009). Artificial Neural Networks and regression approaches comparison for forecasting Iran's annual electricity load. Paper presented at the Power Engineering, *Energy and Electrical Drives*, 2009. POWERENG'09. International Conference on.

Ghods, L., & Kalantar, M. (2008, 21-24 April 2008). Methods for long-term electric load demand forecasting; a comprehensive investigation. Paper presented at the Industrial Technology, 2008. ICIT 2008. *IEEE International Conference on*.

Ghods, L., & Kalantar, M. (2011). Different Methods of Long-Term Electric Load Demand Forecasting; A Comprehensive Review. *Iranian Journal of Electrical & Electronic Engineering*, 7(4), 249.

Ghosh, S. (2002). Electricity consumption and economic growth in India. *Energy Policy*, 30(2), 125-129.

Guo, W., Shen, X., Ma, X., Ma, L., & Cao, T.. (2014). Comparative Study of Grey Forecasting Model and ARMA Model on Beijing Electricity Consumption Forecasting. *Mechatronics and Automatic Control Systems* (pp. 501-508).

Hamdi, H., Sbia, R., & Shahbaz, M. (2014). The nexus between electricity consumption and economic growth in Bahrain. *Economic Modelling*, 38(0), 227-237.

Hamid, A., & Abdul Rahman, T. K. (2010). Short term load forecasting using an artificial neural network trained by artificial immune system learning algorithm. Paper presented at the Computer Modelling and Simulation (UKSim), 2010 12th *International Conference on*.

Hanmandlu, M., & Chauhan, B. K. (2011). Load forecasting using hybrid models. Power Systems, *IEEE Transactions on*, 26(1), 20-29.

Hasan, M. K., Khan, M. A. A, & Saber, S. A. A. Y. (2010). An Efficient Hybrid Model to Load Forecasting". IJCSNS International *Journal of Computer Science and Network Security*, 10(8), 61-68.

Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1), 1-49.

University of Malaya

Holtedahl, P., & Joutz, F. L. (2004). Residential electricity demand in Taiwan. *Energy Economics*, 26(2), 201-224.

Hondroyiannis, G. (2004). Estimating residential demand for electricity in Greece. *Energy Economics*, 26(3), 319-334.

Hv, S., & Zwiers, F. W. (1999). Statistical analysis in climate research: *Cambridge University Press, UK.*

Jain, A., & Satish, B. (2009). Clustering based short term load forecasting using artificial neural network. Paper presented at the Power *Systems Conference and Exposition*, 2009. PSCE'09. *IEEE*/PES.

Javid, M., & Qayyum, A. (2014). Electricity consumption-GDP nexus in Pakistan: A structural time series analysis. *Energy*, 64(0), 811-817.

Jolliffe, I. (2005). Principal component analysis: Wiley Online Library.

Jumbe, C. B. L. (2004). Cointegration and causality between electricity consumption and GDP: empirical evidence from Malawi. *Energy Economics*, 26(1), 61-68.

Kandananond, K. (2011). Forecasting electricity demand in Thailand with an artificial neural network approach. *Energies*, 4(8), 1246-1257.

Kankal, M., Akpınar, A., Kömürcü, M. İ., & Özşahin, T. Ş. (2011). Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variables. *Applied Energy*, 88(5), 1927-1939.

Kareem, Y. H., & Majeed, A. R. (2006). Monthly Peak-load Demand Forecasting for Sulaimany Governorate Using SARIMA. Paper presented at the Transmission & Distribution Conference and Exposition: *Latin America, 2006. TDC'06. IEEE/PES.*

Kargar, M. J., & Charsoghi, K. (2014). Predicting annual electricity consumption in iran using artificial neural networks (narx). Indian J. *Sci. Res*, 5(1), 231-242.

Kavaklioglu, K. (2011). Modeling and prediction of Turkey's electricity consumption using Support Vector Regression. *Applied Energy*, 88(1), 368-375.

Kavaklioglu, K., Ceylan, H., Ozturk, H. K., & Canyurt, O. E. (2009). Modeling and prediction of Turkey's electricity consumption using Artificial Neural Networks. *Energy Conversion and Management*, 50(11), 2719-2727.

Kaytez, F, T, M. Cengiz, C, E. & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems,* 67(0), 431-438.

Kazemi, A., Shakouri, H., Mehregan, M. Taghizadeh, M., Menhaj, M. B., & Foroughi, A. A. (2009, 28-30 Dec. 2009). A Multi-level Artificial Neural Network for Gasoline Demand Forecasting of Iran. Paper presented at the Computer and Electrical Engineering, 2009. ICCEE '09. *Second International Conference on*.

University of Malaya

Kermanshahi, B., & Iwamiya, H. (2002). Up to year 2020 load forecasting using neural nets. *International Journal of Electrical Power & Energy Systems*, 24(9), 789-797.

Kheirkhah, A., Azadeh, A., Saberi, M., Azaron, A., & Shakouri, H. (2013). Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis. *Computers & Industrial Engineering*, 64(1), 425-441.

Kiartzis, S, B., Theocharis, J., & Tsagas, G. (2000). A fuzzy expert system for peak load forecasting. Application to the Greek power system. Paper presented at the Electrotechnical Conference, 2000. *MELECON 2000. 10th Mediterranean.*

Kucukali, S, & Baris, K. (2010). Turkey's short-term gross annual electricity demand forecast by fuzzy logic approach. *Energy Policy*, 38(5), 2438-2445.

Kunwar, N, & Kumar, R. (2013). Area-Load Based Pricing in DSM Through ANN and *Heuristic Scheduling*. 4(3), 1275 - 1281.

Kuo, R., Wang, C., & Chen, Z. (2012). Integration of growing self-organizing map and continuous genetic algorithm for grading lithium-ion battery cells. *Applied Soft Computing*, 12(8), 2012–2022.

Lahari, W., Haug, A., & Garces-O, A. (2011). Estimating quarterly GDP Data for the South Pacific Island Nations. *The Singapore Economic Review,* 56(01), 97-112.

Lai, T. M., To, W. M., Lo, W. C., & Choy, Y. S. (2008). Modeling of electricity consumption in the Asian gaming and tourism center—Macao SAR, People's Republic of China. *Energy*, 33(5), 679-688.

Lam, J, C., Tang, H. L., & Li, D. H. W. (2008). Seasonal variations in residential and commercial sector electricity consumption in Hong Kong. *Energy*, 33(3), 513-523.

Lee, C. (2005). Energy consumption and GDP in developing countries: A cointegrated panel analysis. *Energy Economics*, 27(3), 415-427.

Lewis, C. (1982). Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting: *Business & Economics Butterworth Scientific London.*

Li, J, C, Ji-hang, S, J, & Huang, F. (2012). Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement Advances in *Computer Science and Information Engineering* (pp. 553-558).

Maddala, G. S. (1992). Introduction to Econometrics New York.

Makridakis, S, Andersen, A, Carbone, R, Robert, H, Michele, L, R, . . . Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting*, 1(2), 111-153.

University of Malaya

Makridakis, S, Chatfield, C, Hibon, M, Lawrence, M, Mills, T, Ord, K, & Simmons, L, F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9(1), 5-22.

Mashudi, M R. (2001). Forecasting water demand using neural networks in the operation of reservoirs in Citarum Cascade, West Java, Indonesia. The *George Washington University.*

McAdams, HT, Crawford, RW, & Hadder, GR. (2000a). A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions. *Society of Automotive Engineers*, Paper, 2000-2001-1961.

McAdams, HT, Crawford, RW, & Hadder, GR. (2000b). A vector approach to regression analysis and its application to heavy-duty diesel emissions. Society of Automotive Engineers, Inc, Contract with the *Energy Division of Oak Ridge National Laboratory (ORNL)*. Contract No. DE-AC05-00OR22725.

McCuen, R,H. (1985). Statistical methods for engineers (Vol. 439). Duxbury Press: *Prentice-Hall Englewood Cliffs*, NJ.

Milojkovic, J, Litovski, I, & Litovski, V. (2012). ANN application for the next day peak electricity load prediction. *Paper presented at the Neural Network Applications in Electrical Engineering (NEUREL)*, 2012 11th Symposium on.

Mishra, S, & Patra, S. (2009). Short term load forecasting using a robust novel Wilcoxon Neural Network. Paper presented at the Nonlinear Dynamics and Synchronization, 2009. INDS'09. *2nd International Workshop on*.

Moghaddam, M, & Bahri, P. (2014). Development of a Novel Approach for Electricity Forecasting. In H. K. Kim, S.-I. Ao, M. A. Amouzegar & B. B. Rieger (Eds.), *IAENG Transactions on Engineering Technologies* (Vol. 247, pp. 635-649).

Mohamed, Z, & Bodger, P. (2005). Forecasting electricity consumption in New Zealand using economic and demographic variables. *Energy*, 30(10), 1833-1843.

Mohandes, M. (2002). *Support vector machines for short-term electrical load forecasting.* International journal of energy research, 26(4), 335-345.

Montgomery, D C, Peck, E, & Vining, G.G. (2012). Introduction to linear regression analysis (Vol. 821): *John Wiley & Sons*.

Morimoto, R, & Hope, C. (2004). The impact of electricity supply on economic growth in Sri Lanka. *Energy Economics*, 26(1), 77-85.

Mosalman, F, Mosalman, A, Yazdi, H M & Yazdi, M M. (2011). One day-ahead load forecasting by artificial neural network. *Scientific Research and Essays,* 6(13), 2795-2799.

Mozumder, P, & Marathe, A. (2007). Causality relationship between electricity consumption and GDP in Bangladesh. *Energy Policy*, 35(1), 395-402.

University of Malaya

Murray, M P. (1978). The demand for electricity in Virginia. *The Review of Economics and Statistics*, 60(4), 585-600.

Myers, RH. (1986). Classical and Modern Regression with Applications: *Prindle. Weber & Schmidt (PWS).*

Narayan, Paresh K, & Smyth, R. (2005a). Electricity consumption, employment and real income in Australia evidence from multivariate Granger causality tests. *Energy Policy*, 33(9), 1109-1116.

Narayan, Paresh K, & Smyth, R. (2005b). The residential demand for electricity in Australia: an application of the bounds testing approach to cointegration. *Energy Policy*, 33(4), 467-474.

Narayan, Paresh K, Smyth, R, & Prasad, A. (2007). Electricity consumption in G7 countries: A panel cointegration analysis of residential demand elasticities. *Energy Policy*, 35(9), 4485-4494.

Nasr, GE, Badr, EA, & Dibeh, G. (2000). Econometric modeling of electricity consumption in post-war Lebanon. *Energy Economics*, 22(6), 627-640.

Ndiaye, D, & Gabriel, K. (2011). Principal component analysis of the electricity consumption in residential dwellings. *Energy and Buildings*, 43(2–3), 446-453.

Niu, D, Shi, H, Li, J, & Wei, Y. (2010). Research on short-term power load time series forecasting model based on BP neural network. *Paper presented at the Advanced Computer Control (ICACC), 2010 2nd International Conference on.*

Niu, D. X., Wang, Q., Li, J. C., et al.: Short Term Load Forecasting Model Based on Support Vector Machine, in: Advances in Machine Learning and Cybernetics, vol. 3930, *Springer- Verlag, Berlin*, 880–888, 2006.

Ong, H., Mahlia, T., & Masjuki, H. (2011). A review on energy scenario and sustainable energy in Malaysia. *Renewable and Sustainable Energy Reviews*, 15(1), 639-647.

Ozturk, Harun K, Ceylan, H Canyurt, O E, & Hepbasli, A. (2005). Electricity estimation using genetic algorithm approach: a case study of Turkey. *Energy*, 30(7), 1003-1012.

Ozturk, I, & Acaravci, A. (2010). The causal relationship between energy consumption and GDP in Albania, Bulgaria, Hungary and Romania: Evidence from ARDL bound testing approach. *Applied Energy*, 87(6), 1938-1943.

Panklib, K., C. Prakasvudhisarn and Khummongkol (2015). Electricity Consumption Forecasting in Thailand Using an Artificial Neural Network and Multiple Linear Regression. *Energy Sources, Part B: Economics, Planning, and Policy* 10(4): 427-434.

Pao, Hsiao-T. (2006). Comparing linear and nonlinear forecasts for Taiwan's electricity consumption. *Energy,* 31(12), 2129-2141.

University of Malaya

Pappas, S. Sp, Ekonomou, L., Karamousantas, D. Ch, Chatzarakis, G. E., Katsikas, S. K., & Liatsis, P. (2008). Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models. *Energy*, 33(9), 1353-1360.

Paretkar, P. S., Mili, L., Centeno, V., Kaiyan, Jin, & Miller, C. (2010, 25-29 July 2010). Short-term forecasting of power flows over major transmission interties: Using Box and Jenkins ARIMA methodology. Paper presented at the *Power and Energy Society General Meeting, 2010 IEEE.*

Pindoriya, NM, Singh, SN, & Singh, SK. (2010). Forecasting of short-term electric load using application of wavelets with feed-forward neural networks. *International Journal of Emerging Electric Power Systems*, 11(1).

Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Alvim, M. C. M., & Pereira, M. C. (2008). Selection and validation of parameters in multiple linear and principal component regressions. *Environmental Modelling & Software,* 23(1), 50-55.

Popovic, J J. (2013). Three Universal methods of reducing complexity. Retrieved 2013, 2013, from *www.computing.dcu.ie/~renaat/ca2/ca214/ca214vii.*

Raza, MQ, & Baharudin, Z. (2012). A review on short term load forecasting using hybrid neural network techniques. Paper presented at the Power and Energy (PECon), 2012 *IEEE International Conference on*.

Razak, F Abd, S, Mahendran, H, Amir H, & Abidin, I, Z. (2009). Load forecasting using time series models. *Jurnal Kejuruteraan (Journal of Engineering),* 21, 53-62.

Rumelhart, David E, H, Geoffrey E, & Williams, R, J. (1988). *Learning representations by back-propagating errors*. Cognitive modeling, 323(Nature), 533-536.

Saber, A, Y, & Al-Shareef, A. (2009). *Load forecasting of a desert: A computational intelligence approach*. Paper presented at the Intelligent System Applications to Power Systems, 2009. ISAP'09. 15th International Conference on.

Sackdara, V, Premrudeepreechacharn, S, & Ngamsanroaj, K. (2010*). Electricity demand forecasting of Electricite Du Lao (EDL) using neural networks*. Paper presented at the TENCON 2010-2010 IEEE Region 10 Conference.

Saravanan, S, Kannan, S, & Thangaraj, C. (2012). *India's electricity demand forecast using regression analysis and artificial neural networks based on principal components*. 02(04), 365 - 370.

Shalamu, A. (2009). *Monthly and seasonal streamflow forecasting in the Rio Grande Basin*. New Mexico State University.

Shiu, A, & Lam, P. (2004). *Electricity consumption and economic growth in China*. Energy policy, 32(1), 47-54.

Shu, F, & Luonan, C. (2006). *Short-term load forecasting based on an adaptive hybrid method*. Power Systems, IEEE Transactions on, 21(1), 392-401. doi: 10.1109/TPWRS.2005.860944.

University of Malaya

Shuvra, M Ali, R, Md M Ali, A, & Khan, S, I. (2011). *Modeling and forecasting demand for electricity in Bangladesh: econometrics model*. Paper presented at the International Conference on Economics, Trade and Development.

Song, K, Baek, Y, Hong, D, & Jang, G. (2005). *Short-term load forecasting for the holidays using fuzzy linear regression method*. Power Systems, IEEE Transactions on, 20(1), 96-101.

Sousa, SIV, Martins, FG, Alvim-Ferraz, MCM, & Pereira, MC. (2007). *Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations*. Environmental Modelling & Software, 22(1), 97-103.

Squalli, J. (2007). *Electricity consumption and economic growth: Bounds and causality analyses of OPEC members*. Energy Economics, 29(6), 1192-1205.

Sumer, K, Kagan, G & Hepsag, A. (2009). The application of seasonal latent variable in forecasting electricity demand as an alternative method. *Energy Policy*, 37(4), 1317-1322.

Tadayoshi F. Estimation of prediction error by using K-fold cross-validation. *Springer Science+Business Media*, 10 October 2009

Tasre, M B, Bedekar, P, P, & Ghate, V, N. (2011). Daily peak load forecasting using ANN. Paper presented at the Engineering (NUiCONE), 2011 *Nirma University International Conference on*.

Tasre, M B, Ghate, V, N, & Bedekar, P, P. (2012). Hourly load forecasting using Artificial Neural Network for a small area. Paper presented at the Advances in Engineering, Science and Management (ICAESM), 2012 *International Conference on*.

Tobias, Randall D. (1995). An introduction to partial least squares regression. Paper presented at the Proc. Ann. SAS Users Group Int. Conf., 20th, *Orlando*, FL.

Tso, Geoffrey K. F., & Yau, Kelvin K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9), 1761-1768.

Ul-Saufie, AZ, Yahya, AS, & Ramli, NA. (2011). Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang. *International Journal of Environmental Sciences*, 2(2), 403-409.

Valle, S, Li, W, & Qin, S . (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*, 38(11), 4389-4401.

Van ,S, Abraham J, & Schumacher, J. (2000). An introduction to hybrid dynamical systems (Vol. 251): *Springer London.*

Von S, H, & Zwiers, F, W. (2001). Statistical analysis in climate research (Vol. ISBN 0 511 01018 4 virtua): *Cambridge University Press.*

University of Malaya

Wang, J, Zhu, W, Zhang, W, & Sun, D. (2009). A trend fixed on firstly and seasonal adjustment model combined with the ε-SVR for short-term forecasting of electricity demand. *Energy Policy*, 37(11), 4901-4909.

Wang, S, & Xiao, F. (2004). AHU sensor fault diagnosis using principal component analysis method. *Energy and Buildings*, 36(2), 147-160.

Wolde-Rufael, Y. (2006). Electricity consumption and economic growth: a time series experience for 17 African countries. *Energy Policy*, 34(10), 1106-1114.

Xin, J, Yao, D, Jie, W, & J, W. (2010, 7-8 Aug. 2010). An Improved Combined Forecasting Method for Electric Power Load Based on Autoregressive Integrated Moving Average Model. Paper presented at the Information Science and Management Engineering (ISME), 2010 *International Conference of*.

Yao, SJ, S, YH, Z, LZ, & C, XY. (2000). Wavelet transform and neural networks for short-term electrical load forecasting. *Energy Conversion and Management*, 41(18), 1975-1988.

Yingying, Li, & Dongxiao, Niu. (2010). Application of Principal Component Regression Analysis in power load forecasting for medium and long term. Paper presented at the Advanced Computer Theory and Engineering (ICACTE), 2010 *3rd International Conference on*.

Yoo, S. H. (2006). The causal relationship between electricity consumption and economic growth in the ASEAN countries. *Energy Policy*, 34(18), 3573-3582.

Yoo, S. (2005). Electricity consumption and economic growth: evidence from Korea. *Energy Policy*, 33(12), 1627-1632.

Yoo, S, Lee, J, & Kwak, S. (2007). Estimation of residential electricity demand function in Seoul by correction for sample selection bias. *Energy Policy*, 35(11), 5702-5707.

Yuan, J, Zhao, C, Yu, S & Hu, Z. (2007). Electricity consumption and economic growth in China: Cointegration and co-feature analysis. *Energy Economics*, 29(6), 1179-1191.

Yumurtaci, Z & Asmaz, E. (2004). Electric energy demand of Turkey for the year 2050. *Energy Sources*, 26(12), 1157-1164.

Zachariadis, T. (2010). Forecast of electricity consumption in Cyprus up to the year 2030: The potential impact of climate change. *Energy Policy*, 38(2), 744-750.

Zachariadis, T, & Pashourtidou, N. (2007). An empirical analysis of electricity consumption in Cyprus. *Energy Economics*, 29(2), 183-198.

Zarezadeh, M, Naghavi, A, & Ghaderi, SF. (2008). Electricity price forecasting in Iranian electricity market applying Artificial Neural Networks. Paper presented at the Electric Power Conference, 2008. EPEC 2008. *IEEE Canada*.

University of Malaya

Zhang, J, Y, Shen, F, Li, Y, Xiao, H Qi, H, Deng, S. (2012). Principal Component Analysis of Electricity Consumption Factors in China. *Energy Procedia*, 16, 1913-1918.

Zhang, S, & Wang, Q. (2009). Medium and long-term load forecasting based on PCA and BP neural network method. Paper presented at the Energy and Environment Technology, 2009. ICEET'09. *International Conference on*.

Zhang, X, Wang, Q, Yu, M, & Wu, J. (2010). Combining principal component regression and artificial neural network to predict chlorophyll-a concentration of Yuqiao Reservoir's outflow. *Transactions of Tianjin University*, 16, 467-472.

Zheng, F & Zhong, S. (2011). Timeseries Forecasting Using a Hybrid RBF Neural Network and AR Model Based on Binomial Smooting. *World Academy of Science and Technology*, 75.

Zhou, P., Ang, B. W., & Poh, K. L. (2006). A trigonometric grey prediction approach to forecasting electricity demand. *Energy*, 31(14), 2839-2847.

Zuhaimy I.Member, *IEEE and Rosnalini Mansorb*. (2011). Fuzzy Logic Approach for Forecasting Half-hourly Malaysia Electricity Load Demand.

www.st.gov.my/index.php. (2014). Peninsular Malaysia Electricity Supply Industry Outlook 2014 *Retrieved from*

http://www.st.gov.my/index.php/component/k2/item/606-peninsular-malaysia-electricity-supply-industry-outlook-2014.html

Sweden, D. 2015. Available: http://www.statistikdatabasen.scb.se/pxweb/en/ssd/?rxid=0adc9826-6f78-4830-84ab-a24b0197ffa9.

Institute, T. S. 2015. Turkish Statistical Institute [Online]. Available: http://www.turkstat.gov.tr/Start.do;jsessionid=gt42Y1kRxLXyxx0m5xqwsphd1k8l9Kdn8nM1vgJGVqzk8rXGn5Pt!979482609

University of Malaya

University of Malaya

LIST OF PUBLICATIONS AND PAPERS PRESENTED

ISI Journal

Noor Azina Ismail and Syamnd Mirza Abdullah (2016), Principal Component Regression with Artificial Neural Network to Improve Prediction of Electricity Demand. International Arab Journal of Information Technology (IAJIT) . 2016, Vol. 13 Issue 1A, p196-202. 7p. http://ccis2k.org/iajit/?option=com_content&task=blogcategory&id=109&Itemid=395

Conference

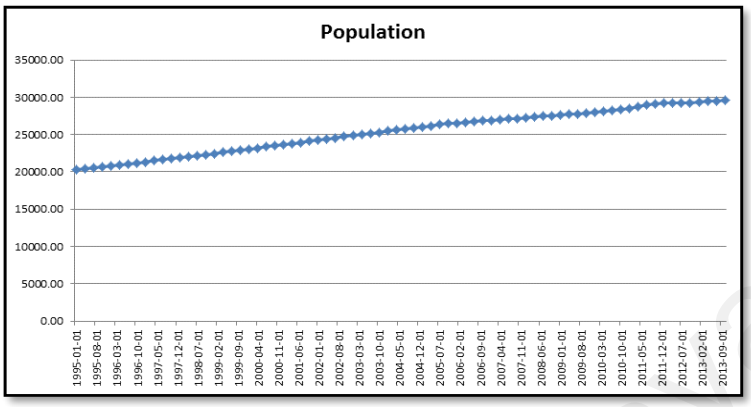
Noor Azina Ismail and Syamnd Mirza Abdullah (2016), A Two-Layer Approach Model for Industry Electricity Demand in Malaysia. The International Arab Conference on Information Technology (ACIT). http://www.acit2k.org/ACIT/index.php/component/content/category/41-acit

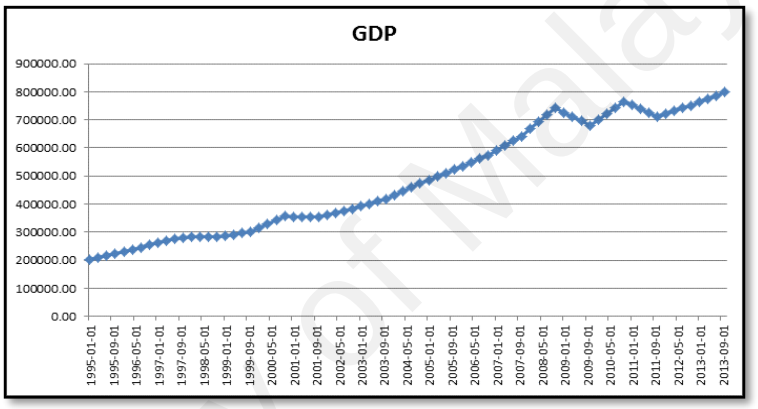
Noor Azina Ismail and Syamnd Mirza Abdullah (2013), Past, Present and Future for Electricity Consumption Studies A Review . International Conference on Innovation Challenges in Multidisciplinary Research & Practice (ICMRP). http://www.globalilluminators.org/wp-content/uploads/2013/08/LIST.pdf

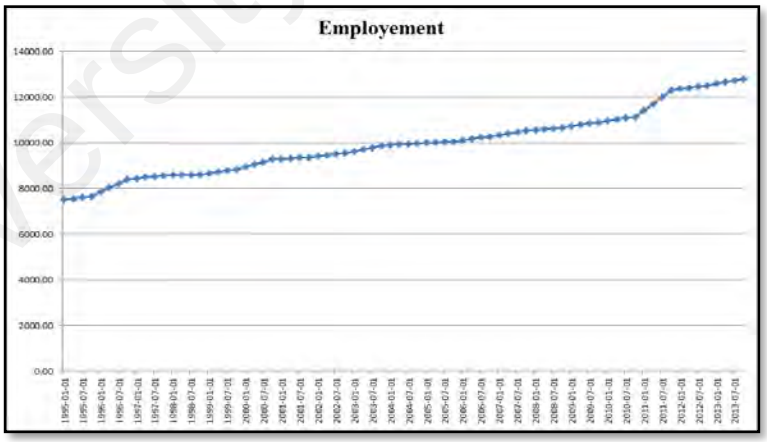
University of Malaya

APPENDIX

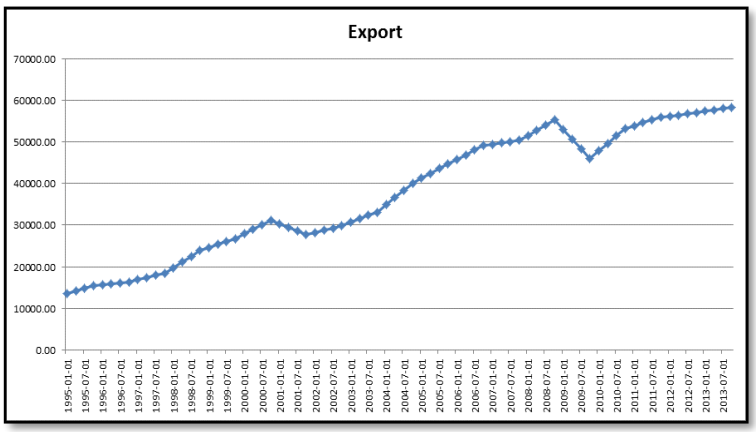
Appendix A: Linear and Nonlinear Dataset

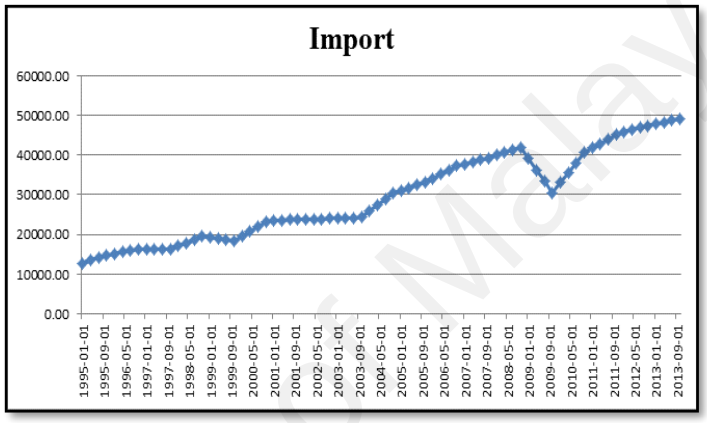


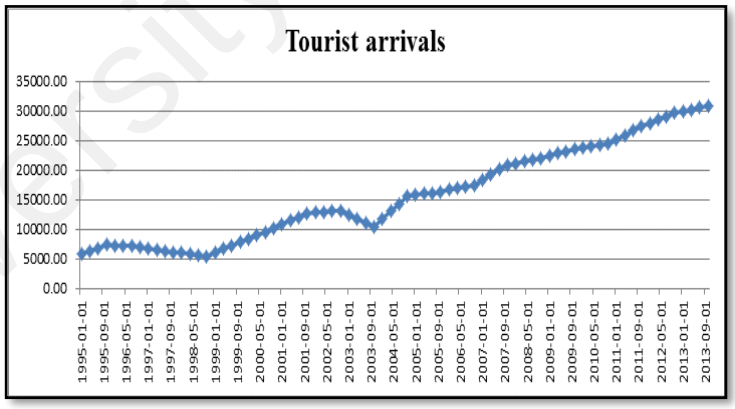


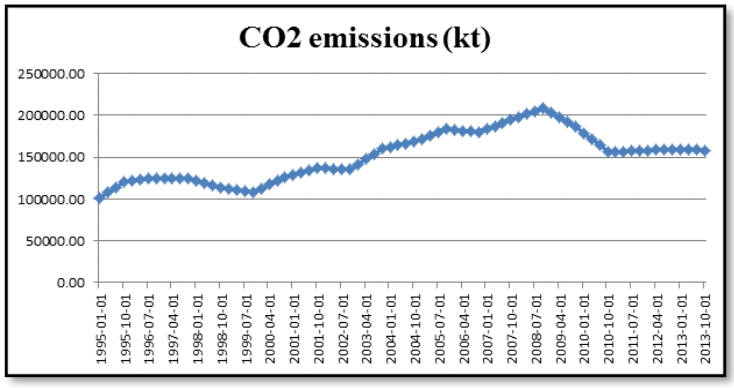


University of Malaya

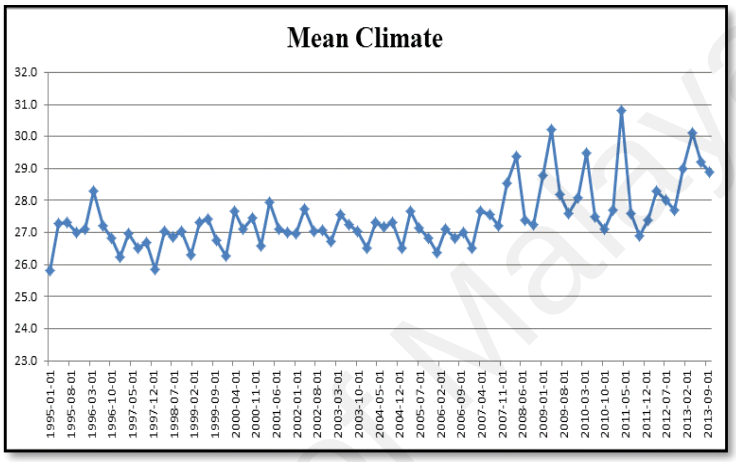


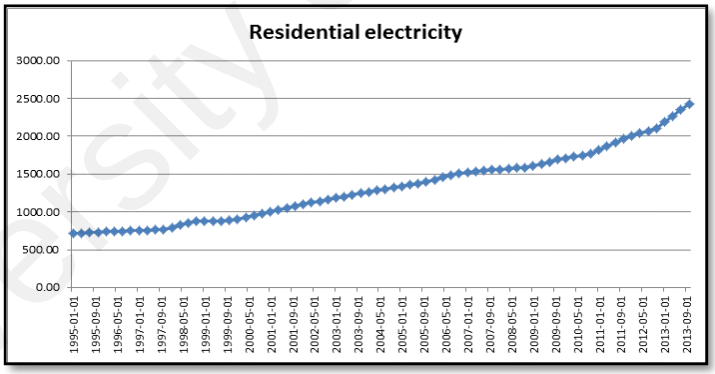


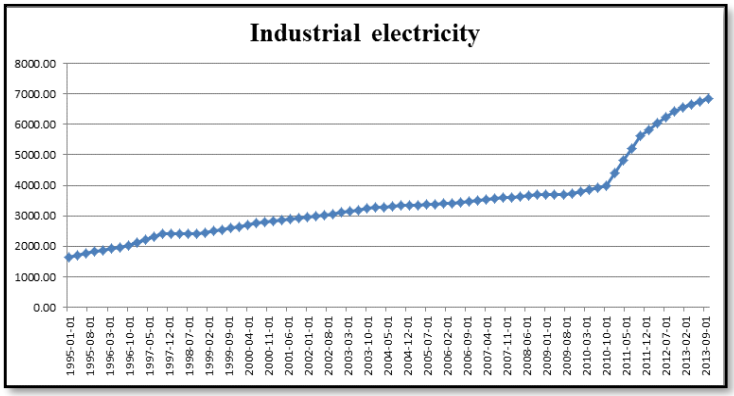




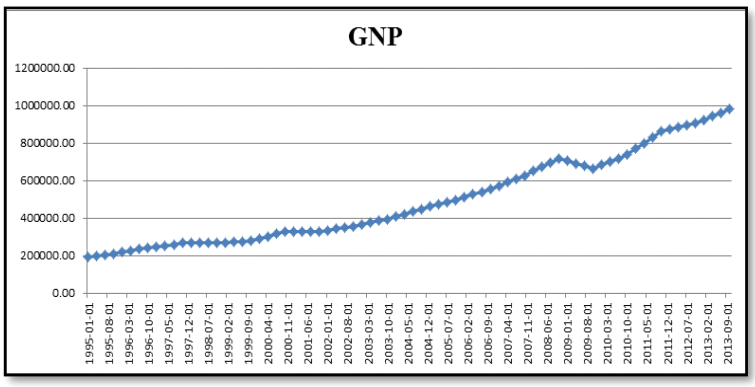
University of Malaya

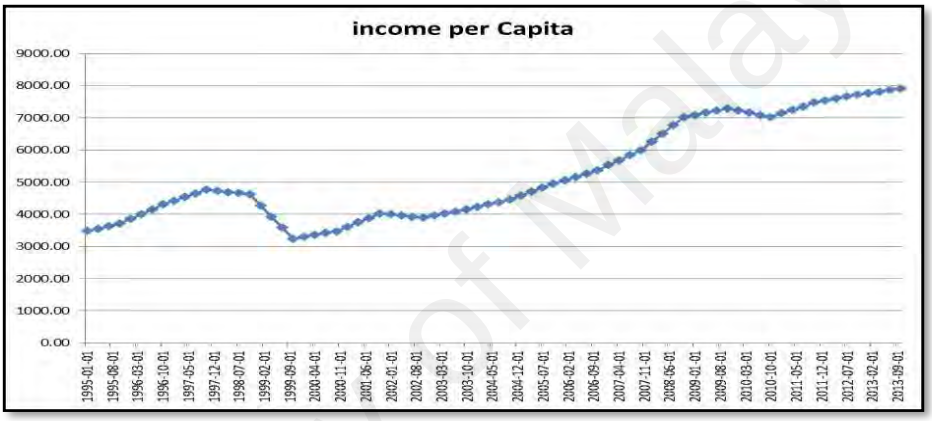






University of Malaya





Appendix B: Codes Used Throughout Building PCR- BPNN

Code-1; the prepare dataset from starting till using PCR

```
clc;
clear all;
[x]=xlsread('QX.xlsx'); %% the independent
[y]=xlsread('QY.xlsx'); %% the independent
Zx=zscore(x);
Zy=zscore(y);
correlation=corrcoef(x); % correlation
eigenvalue=eig(correlation); %eigne value
[eigenVector, Eigenvalues]=eig(correlation); % find eign value and eign vector
```

University of Malaya

```
PCsall = eigenVector(1:13,1:13); %to find from PC1 to PC5
PCs = eigenVector(1:13,1:5); %to find from PC1 to PC5
correlationPCs=corrcoef(PCs); % correlation PCs
datasetall=Zx * PCsall;
dataset=Zx * PCs;
new_dataset=dataset * -1;
PCA_data=xlswrite('PCAdata',[new_dataset]); % to write the number PC
beta= ((new_dataset'*new_dataset)^-1)*(new_dataset'*Zy);
Yhat_PCR= new_dataset*beta;
error=  Zy - Yhat_PCR;
RMSE= sqrt(mean((error).^2));
MSE= RMSE *RMSE;
MAPE = (mae(Zy - Yhat_PCR))*100;
```

Code 2, applying BPNN on the resulting residuals error from PCR model

```
% Solve an Autoregression Time-Series Problem with a NAR Neural Network
% Script generated by NTSTOOL
% Created Sat Jul 12 17:09:46 SGT 2014
% This script assumes this variable is defined:
%   error - feedback time series.
targetSeries = tonndata(error,false,false);
% Create a Nonlinear Autoregressive Network
feedbackDelays = 1:2;
hiddenLayerSize = 10;
net = narnet(feedbackDelays,hiddenLayerSize);
% Choose Feedback Pre/Post-Processing Functions
% Settings for feedback input are automatically applied to feedback output
```

University of Malaya

```
% For a list of all processing functions type: help nnprocess
net.inputs{1}.processFcns = {'removeconstantrows','mapminmax'};
% Prepare the Data for Training and Simulation
% The function PREPARETS prepares timeseries data for a particular network,
% shifting time by the minimum amount to fill input states and layer states.
% Using PREPARETS allows you to keep your original time series data unchanged, while
% easily customizing it for networks with differing numbers of delays, with
% open loop or closed loop feedback modes.
[inputs,inputStates,layerStates,targets] = preparets(net,{},{},targetSeries);
% Setup Division of Data for Training, Validation, Testing
% For a list of all data division functions type: help nndivide
net.divideFcn = 'dividerand';  % Divide data randomly
net.divideMode = 'time';  % Divide up every value
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
 % Choose a Training Function
% For a list of all training functions type: help nntrain
net.trainFcn = 'trainlm';  % Levenberg-Marquardt
 % Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse';  % Mean squared error
 % Choose Plot Functions
% For a list of all plot functions type: help nnplot
net.plotFcns = {'plotperform','plottrainstate','plotresponse', ...
  'ploterrcorr',  'plotregression','plotinerrcorr'};
```

University of Malaya

```
% Train the Network
[net,tr] = train(net,inputs,targets,inputStates,layerStates);
 % Test the Network
outputs = net(inputs,inputStates,layerStates);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)
 % Recalculate Training, Validation and Test Performance
trainTargets = gmultiply(targets,tr.trainMask);
valTargets = gmultiply(targets,tr.valMask);
testTargets = gmultiply(targets,tr.testMask);
trainPerformance = perform(net,trainTargets,outputs)
valPerformance = perform(net,valTargets,outputs)
testPerformance = perform(net,testTargets,outputs)


 % View the Network
view(net)
 % Plots
% Uncomment these lines to enable various plots.
%figure, plotperform(tr)
%figure, plottrainstate(tr)
%figure, plotresponse(targets,outputs)
%figure, ploterrcorr(errors)
%figure, plotinerrcorr(inputs,errors)
 % Closed Loop Network
% Use this network to do multi-step prediction.
% The function CLOSELOOP replaces the feedback input with a direct
```

University of Malaya

```
% connection from the outout layer.
netc = closeloop(net);
[xc,xic,aic,tc] = preparets(netc,{},{},targetSeries);
yc = netc(xc,xic,aic);
perfc = perform(net,tc,yc)
% Early Prediction Network
nets = removedelay(net);
[xs,xis,ais,ts] = preparets(nets,{},{},targetSeries);
ys = nets(xs,xis,ais);
closedLoopPerformance = perform(net,tc,yc)
errorBPNN= errors';
errorBPNN= cell2mat(errorBPNN);
YhatPCR=Yhat_PCR(3:76);
Zyy=Zy(3:76);
hybrid=errorBPNN+YhatPCR;
errorhybrid= Zyy - hybrid;
RMSE_hybrid= sqrt(mean((errorhybrid).^2)); % for current process
MAPE_hybrid = (mae(errorhybrid))*100; % for current process
weights = getwb(net); % to find the neural network
[b,IW,LW] = separatewb(net,weights); %to spreate bias and weight
%%%b  : Cell array of bias vectors
%%%IW : Cell array of input weight matrices
%%%LW : Cell array of layer weight matrices
```

Code 3 using PC- Back propagation neural networks

```
clc;
clear all;
```

University of Malaya

```
[x]=xlsread('PCAdata.xls'); %% the independent
[Y]=xlsread('QY.xlsx'); %% the independent
y=zscore(Y);
% Solve an Input-Output Fitting problem with a Neural Network
% Script generated by NFTOOL
% Created Thu Jul 17 17:35:32 SGT 2014
% This script assumes these variables are defined:
%   x - input data.
%   y - target data.
inputs = x';
targets = y';
 % Create a Fitting Network
hiddenLayerSize = 10;
net = fitnet(hiddenLayerSize);
% Choose Input and Output Pre/Post-Processing Functions
% For a list of all processing functions type: help nnprocess
net.inputs{1}.processFcns = {'removeconstantrows','mapminmax'};
net.outputs{2}.processFcns = {'removeconstantrows','mapminmax'};
 % Setup Division of Data for Training, Validation, Testing
% For a list of all data division functions type: help nndivide
net.divideFcn = 'dividerand';  % Divide data randomly
net.divideMode = 'sample';  % Divide up every sample
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
 % For help on training function 'trainlm' type: help trainlm
```

University of Malaya

```
% For a list of all training functions type: help nntrain
net.trainFcn = 'trainlm';  % Levenberg-Marquardt
 % Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse';  % Mean squared error
 % Choose Plot Functions
% For a list of all plot functions type: help nnplot
net.plotFcns = {'plotperform','plottrainstate','ploterrhist', ...
  'plotregression',  'plotregression', 'plotfit'};
% Train the Network
[net,tr] = train(net,inputs,targets);
 % Test the Network
outputs = net(inputs);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)
 % Recalculate Training, Validation and Test Performance
trainTargets = targets .* tr.trainMask{1};
valTargets = targets  .* tr.valMask{1};
testTargets = targets  .* tr.testMask{1};
trainPerformance = perform(net,trainTargets,outputs)
valPerformance = perform(net,valTargets,outputs)
testPerformance = perform(net,testTargets,outputs)
 % View the Network
view(net)
 % Plots
% Uncomment these lines to enable various plots.
```

University of Malaya

```
%figure, plotperform(tr)

%figure, plottrainstate(tr)

%figure, plotfit(net,inputs,targets)

%figure, plotregression(targets,outputs)

%figure, ploterrhist(errors)

ErNN= errors';

outputNN=outputs';

RMSE_NN= sqrt(mean((ErNN).^2)); % for current process

MAPE_NN = (mae(ErNN))*100; % for current process

 weights = getwb(net); % to find the neural network

[b,IW,LW] = separatewb(net,weights); %to spreate bias and weigh

Trantest=testTargets';

%%%%%%to convert Yhat to original %%%%%

mean_y= mean(y);

st_y=std(y);

yi_predict= (outputs*st_y)+mean_y;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Code 4; coding for PC- SVR model

```
clc;

clear all;

[x]=xlsread('QX.xlsx'); %% the independent

[y]=xlsread('QY.xlsx'); %% the independent

Zx=zscore(x);

Zy=zscore(y);

correlation=corrcoef(x); % correlation

eigenvalue=eig(correlation); %eigne value
```

University of Malaya

```
[eigenVector, Eigenvalues]=eig(correlation); % find eign value and eign vector
PCs = eigenVector(1:13,1:5); %to find from PC1 to PC5
correlationPCs=corrcoef(PCs); % correlation PCs
dataset=Zx * PCs;
new_dataset=dataset * -1;
beta= ((new_dataset'*new_dataset)^-1)*(new_dataset'*Zy);
Yhat_PCR= new_dataset*beta;
error=  Zy - Yhat_PCR;
%%%%%%SVM %%%%%%%
[err]=xlsread('er.xlsx'); %% the independent
in =err(1:76,1:1);
tr=err(1:76,2:2);
testin=error(66:76,1:1);
svm=svmtrain(in,tr)
predict=svmprdict(svm,testin);
RMSE= sqrt(mean((Zy - predict).^2));
MAPE = (mae(Zy - preidct))*100;
%%%%%%%to convert Yhat to original %%%%%%
mean_y= mean(y);
st_y=std(y);
yi_predict= (Yhat_PCR*st_y)+mean_y;
```

Appendix C: All PCs Transfer From Original Dataset.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
-1.54	0.79	-2.37	0.35	0.48	0.30	-0.45	0.11	-0.13	0.04	-0.01	0.00	-0.01
-1.51	0.98	-2.11	0.47	0.17	0.28	-0.37	0.08	-0.09	0.01	-0.02	-0.01	0.02

University of Malaya

-1.41	-0.15	-0.79	-0.29	0.16	0.22	-0.34	0.00	-0.07	0.01	-0.02	-0.03	0.04
-1.41	2.37	0.39	0.64	0.21	0.16	-0.33	-0.11	-0.06	0.02	-0.02	-0.04	0.06
-1.35	-0.22	-0.93	-0.30	0.23	0.10	-0.24	-0.09	-0.01	0.01	-0.01	-0.01	0.02
-1.32	1.85	0.53	0.46	0.05	0.06	-0.11	-0.03	0.06	-0.02	0.00	0.02	-0.01
-1.26	0.71	0.06	0.04	0.31	-0.03	-0.07	-0.07	0.08	-0.01	0.01	0.05	-0.05
-1.25	2.67	0.91	0.75	0.44	-0.10	0.00	-0.07	0.12	-0.01	0.02	0.08	-0.09
-1.23	1.65	-0.58	0.32	0.59	-0.15	0.05	-0.08	0.09	0.02	0.02	0.05	-0.08
-1.20	1.03	-0.90	0.05	0.49	-0.17	0.15	-0.03	0.09	0.01	0.02	0.01	-0.05
-1.14	0.54	-0.47	-0.18	0.62	-0.21	0.20	-0.04	0.06	0.04	0.02	-0.02	-0.04
-1.12	-0.17	-1.50	-0.49	0.62	-0.24	0.28	-0.01	0.05	0.05	0.01	-0.06	-0.02
-1.10	-0.63	-1.61	-0.74	0.75	-0.17	0.28	-0.02	-0.01	0.01	0.01	-0.04	0.00
-1.07	-0.06	-1.11	-0.60	0.48	-0.05	0.36	0.05	-0.03	-0.07	0.01	-0.02	0.02
-1.00	-0.01	1.48	-0.66	0.49	0.03	0.38	0.06	-0.08	-0.12	0.00	-0.01	0.04
-0.94	-1.30	1.67	-1.22	0.42	0.13	0.42	0.10	-0.12	-0.18	0.00	0.01	0.06
-0.93	-1.70	1.54	-1.43	0.38	0.12	0.29	0.07	-0.11	-0.08	-0.01	0.01	0.06
-0.94	-0.72	2.23	-1.12	0.00	0.16	0.23	0.11	-0.07	-0.01	-0.01	0.00	0.06
-0.92	-1.27	2.14	-1.38	-0.21	0.18	0.13	0.12	-0.04	0.07	-0.02	0.00	0.07
-0.98	0.92	2.10	-0.62	-0.26	0.17	0.00	0.09	-0.03	0.16	-0.02	0.00	0.07
-0.90	-0.30	1.13	-0.94	-0.19	0.17	-0.01	0.02	-0.02	0.17	-0.01	0.01	0.04
-0.80	-0.96	1.70	-1.04	-0.50	0.21	0.06	0.04	0.03	0.14	0.01	0.02	0.03
-0.80	-0.43	-0.48	-0.71	-0.42	0.20	0.05	-0.03	0.04	0.15	0.02	0.03	0.00
-0.72	-0.26	-0.35	-0.51	-0.51	0.21	0.07	-0.06	0.07	0.13	0.03	0.04	-0.02
-0.71	1.83	0.92	0.33	-0.29	0.10	-0.01	-0.14	0.05	0.08	0.03	0.03	-0.02
-0.63	-0.43	0.79	-0.46	-0.51	0.03	0.01	-0.13	0.08	-0.02	0.02	0.02	0.00
-0.62	-0.48	-0.43	-0.44	-0.29	-0.08	-0.07	-0.21	0.06	-0.07	0.01	0.00	0.00
-0.59	-0.54	-0.45	-0.41	-0.21	-0.18	-0.11	-0.26	0.06	-0.14	0.00	-0.01	0.00
-0.57	0.47	0.21	-0.06	-0.23	-0.19	-0.13	-0.24	0.04	-0.12	-0.01	-0.01	0.00
-0.53	-0.22	-0.15	-0.31	-0.41	-0.17	-0.11	-0.17	0.05	-0.12	-0.02	-0.02	-0.01
-0.50	-0.64	-0.21	-0.47	-0.30	-0.19	-0.16	-0.17	0.01	-0.09	-0.03	-0.02	-0.02
-0.49	0.89	0.89	0.01	-0.34	-0.20	-0.17	-0.14	0.00	-0.08	-0.03	-0.02	-0.02
-0.45	-0.57	-0.90	-0.33	-0.34	-0.27	-0.11	-0.05	-0.03	-0.03	-0.03	-0.02	-0.05
-0.41	-0.20	-0.53	0.02	-0.57	-0.30	0.01	0.08	-0.03	-0.01	-0.03	-0.02	-0.07
-0.37	-0.36	-0.87	0.17	-0.58	-0.37	0.07	0.17	-0.05	0.03	-0.03	-0.01	-0.09
-0.34	-0.42	-1.15	0.36	-0.61	-0.43	0.14	0.27	-0.08	0.07	-0.03	0.00	-0.12
-0.26	-0.69	-1.29	0.32	-0.50	-0.35	0.05	0.20	-0.05	0.06	-0.01	-0.01	-0.08
-0.18	-0.29	-1.02	0.54	-0.66	-0.22	0.03	0.19	0.01	0.03	0.01	-0.01	-0.04
-0.01	-1.87	0.99	0.01	-0.63	-0.13	-0.04	0.14	0.05	0.01	0.03	-0.02	-0.01
0.07	-1.37	1.45	0.27	-0.65	-0.02	-0.10	0.10	0.09	-0.01	0.04	-0.02	0.03
0.06	-0.17	0.40	0.85	-0.45	-0.04	-0.11	0.06	0.05	0.00	0.03	-0.02	0.04
0.12	-0.62	0.21	0.82	-0.64	0.00	-0.03	0.10	0.05	-0.03	0.02	-0.03	0.06
0.16	-0.75	-0.44	0.92	-0.49	-0.01	-0.02	0.06	0.02	-0.02	0.01	-0.03	0.07
0.19	0.25	-0.29	1.43	-0.39	-0.01	0.00	0.04	-0.02	-0.02	0.00	-0.03	0.08
0.26	-0.99	-0.93	0.91	-0.25	0.05	-0.01	0.02	-0.05	-0.03	0.00	-0.01	0.06
0.32	-0.65	-0.63	0.98	-0.35	0.14	0.04	0.05	-0.05	-0.06	0.01	0.01	0.05

Universiti Malaya

0.37	0.16	-0.04	1.41	-0.25	0.20	0.05	0.04	-0.08	-0.08	0.01	0.03	0.03
0.45	-0.79	-0.51	1.07	-0.23	0.27	0.07	0.05	-0.10	-0.10	0.01	0.04	0.02
0.51	-1.06	-0.63	1.04	-0.05	0.23	0.02	-0.02	-0.08	-0.07	0.01	0.04	0.00
0.57	-0.45	-0.22	1.31	-0.22	0.23	0.04	-0.02	-0.03	-0.07	0.00	0.03	-0.01
0.63	-0.52	-0.22	1.35	-0.12	0.20	0.00	-0.07	0.00	-0.04	0.00	0.03	-0.02
0.71	-2.27	-1.23	0.81	0.03	0.16	-0.04	-0.14	0.02	-0.01	0.00	0.03	-0.04
0.76	-0.12	0.23	1.61	-0.09	0.20	0.07	-0.11	0.03	0.00	-0.01	0.01	-0.03
0.83	0.28	0.63	1.87	-0.11	0.24	0.17	-0.10	0.03	0.02	-0.02	-0.01	-0.03
0.90	0.75	1.08	2.05	0.44	0.20	0.14	-0.21	-0.03	0.10	-0.03	-0.02	-0.05
0.99	-0.87	0.27	1.44	0.61	0.22	0.19	-0.25	-0.05	0.14	-0.03	-0.04	-0.05
0.95	-0.08	0.75	1.66	0.40	0.03	0.14	-0.11	-0.02	0.09	-0.02	-0.02	0.00
0.91	0.64	1.18	1.92	0.20	-0.16	0.09	0.03	0.01	0.04	-0.01	0.00	0.05
0.90	-0.38	0.57	1.57	0.70	-0.44	-0.12	0.02	-0.03	0.06	0.00	0.03	0.09
0.87	-0.67	0.40	1.39	0.91	-0.68	-0.26	0.07	-0.04	0.05	0.01	0.06	0.13
0.82	-0.10	0.74	-0.20	0.76	-0.40	-0.23	0.12	-0.03	0.02	0.02	0.03	0.07
0.88	0.63	1.23	-0.14	0.43	-0.11	-0.15	0.21	0.01	-0.03	0.02	-0.01	0.00
0.95	0.09	0.96	-0.33	0.78	0.11	-0.23	0.15	-0.02	-0.01	0.03	-0.03	-0.07
1.03	-0.64	0.57	-0.63	0.81	0.36	-0.24	0.16	-0.03	-0.03	0.03	-0.06	-0.14
1.11	0.39	0.72	-0.54	0.54	0.30	-0.15	0.17	0.05	-0.04	0.01	-0.03	-0.10
1.22	-0.38	-0.21	-1.07	-0.23	0.30	0.04	0.29	0.17	-0.10	-0.01	-0.01	-0.04
1.31	0.23	-0.32	-1.08	0.26	0.14	-0.04	0.14	0.16	-0.03	-0.03	0.03	-0.01
1.41	0.07	-0.88	-1.38	0.25	0.05	-0.01	0.10	0.21	-0.02	-0.06	0.07	0.03
1.46	0.35	-0.71	-1.42	0.14	0.03	0.01	0.04	0.17	0.00	-0.04	0.03	0.04
1.52	-0.19	-1.03	-1.71	-0.05	0.02	0.06	0.00	0.15	0.00	-0.02	-0.01	0.05
1.58	-0.84	-1.41	-2.04	0.00	-0.02	0.05	-0.09	0.10	0.03	0.00	-0.04	0.06
1.62	0.56	-0.57	-1.60	0.06	-0.06	0.04	-0.18	0.05	0.06	0.02	-0.08	0.06
1.68	1.22	-0.14	-1.36	-0.21	-0.05	0.06	-0.12	-0.03	0.03	0.02	-0.04	0.03
1.72	1.98	-0.20	-1.13	-0.43	-0.05	0.08	-0.08	-0.11	0.00	0.02	-0.01	0.00
1.79	1.56	-0.56	-1.30	-0.25	-0.09	0.00	-0.12	-0.24	0.01	0.01	0.03	-0.04
1.85	2.96	0.29	-0.85	-0.20	-0.12	-0.05	-0.13	-0.36	0.01	0.01	0.07	-0.08

- The PC1, PC2, PC3 and PC4 selected to compute PPED base on the accumulative percentage variance

PC1	PC2	PC3	PC4
-1.54	0.79	-2.37	0.35
-1.51	0.98	-2.11	0.47
-1.41	-0.15	-0.79	-0.29
-1.41	2.37	0.39	0.64
-1.35	-0.22	-0.93	-0.30
-1.32	1.85	0.53	0.46
-1.26	0.71	0.06	0.04
-1.25	2.67	0.91	0.75

Universiti Malaya

-1.23	1.65	-0.58	0.32
-1.20	1.03	-0.90	0.05
-1.14	0.54	-0.47	-0.18
-1.12	-0.17	-1.50	-0.49
-1.10	-0.63	-1.61	-0.74
-1.07	-0.06	-1.11	-0.60
-1.00	-0.01	1.48	-0.66
-0.94	-1.30	1.67	-1.22
-0.93	-1.70	1.54	-1.43
-0.94	-0.72	2.23	-1.12
-0.92	-1.27	2.14	-1.38
-0.98	0.92	2.10	-0.62
-0.90	-0.30	1.13	-0.94
-0.80	-0.96	1.70	-1.04
-0.80	-0.43	-0.48	-0.71
-0.72	-0.26	-0.35	-0.51
-0.71	1.83	0.92	0.33
-0.63	-0.43	0.79	-0.46
-0.62	-0.48	-0.43	-0.44
-0.59	-0.54	-0.45	-0.41
-0.57	0.47	0.21	-0.06
-0.53	-0.22	-0.15	-0.31
-0.50	-0.64	-0.21	-0.47
-0.49	0.89	0.89	0.01
-0.45	-0.57	-0.90	-0.33
-0.41	-0.20	-0.53	0.02
-0.37	-0.36	-0.87	0.17
-0.34	-0.42	-1.15	0.36
-0.26	-0.69	-1.29	0.32
-0.18	-0.29	-1.02	0.54
-0.01	-1.87	0.99	0.01
0.07	-1.37	1.45	0.27
0.06	-0.17	0.40	0.85
0.12	-0.62	0.21	0.82
0.16	-0.75	-0.44	0.92
0.19	0.25	-0.29	1.43
0.26	-0.99	-0.93	0.91
0.32	-0.65	-0.63	0.98
0.37	0.16	-0.04	1.41
0.45	-0.79	-0.51	1.07
0.51	-1.06	-0.63	1.04
0.57	-0.45	-0.22	1.31
0.63	-0.52	-0.22	1.35
0.71	-2.27	-1.23	0.81

Universiti Malaya

0.76	-0.12	0.23	1.61
0.83	0.28	0.63	1.87
0.90	0.75	1.08	2.05
0.99	-0.87	0.27	1.44
0.95	-0.08	0.75	1.66
0.91	0.64	1.18	1.92
0.90	-0.38	0.57	1.57
0.87	-0.67	0.40	1.39
0.82	-0.10	0.74	-0.20
0.88	0.63	1.23	-0.14
0.95	0.09	0.96	-0.33
1.03	-0.64	0.57	-0.63
1.11	0.39	0.72	-0.54
1.22	-0.38	-0.21	-1.07
1.31	0.23	-0.32	-1.08
1.41	0.07	-0.88	-1.38
1.46	0.35	-0.71	-1.42
1.52	-0.19	-1.03	-1.71
1.58	-0.84	-1.41	-2.04
1.62	0.56	-0.57	-1.60
1.68	1.22	-0.14	-1.36
1.72	1.98	-0.20	-1.13
1.79	1.56	-0.56	-1.30
1.85	2.96	0.29	-0.85

Appendix D: Dataset of Predicted ED (PCR) and Error

Obse.	predict E.D (PCR)	error
1	-1.53155	0.058164
2	-1.51294	-0.04998
3	-1.36172	-0.26084
4	-1.4585	-0.19383
5	-1.29757	-0.35464
6	-1.35611	-0.26609
7	-1.25626	-0.30606
8	-1.31968	-0.15287
9	-1.24899	0.03126
10	-1.18419	0.061929

University of Malaya

11	-1.11559	0.064609
12	-1.05689	0.053011
13	-1.01278	-0.032
14	-1.00708	-0.01345
15	-0.95467	-0.04028
16	-0.84919	-0.11884
17	-0.8218	-0.13015
18	-0.86579	-0.0517
19	-0.82865	-0.04819
20	-0.95872	0.128732
21	-0.83641	0.0836
22	-0.72421	0.021016
23	-0.72794	0.070918
24	-0.67339	0.0591
25	-0.75456	0.17111
26	-0.59166	0.047436
27	-0.57371	0.06866
28	-0.54137	0.075432
29	-0.56572	0.134089
30	-0.50299	0.112253
31	-0.45166	0.103652
32	-0.50038	0.19695
33	-0.41151	0.163614
34	-0.40062	0.197331
35	-0.36872	0.208219
36	-0.34215	0.222629
37	-0.25682	0.174171
38	-0.20584	0.161479
39	-0.00653	-0.00045
40	0.040972	-0.01145
41	-0.00551	0.066346
42	0.063341	0.033936
43	0.104609	0.02994
44	0.089512	0.083141

University of Malaya

45	0.212729	-0.00544
46	0.254027	-0.00526
47	0.268293	0.024516
48	0.375419	-0.03602
49	0.446213	-0.04298
50	0.472165	-0.02311
51	0.532392	-0.04084
52	0.679656	-0.14894
53	0.636321	-0.07791
54	0.682432	-0.08824
55	0.726295	-0.09637
56	0.878468	-0.21287
57	0.810991	-0.14005
58	0.745879	-0.02728
59	0.770537	0.007767
60	0.75851	0.091549
61	0.822536	0.175871
62	0.857572	0.21087
63	0.955354	0.169353
64	1.063779	0.103424
65	1.129508	-0.00762
66	1.287566	-0.12111
67	1.374277	-0.1474
68	1.500594	-0.19745
69	1.548275	-0.09878
70	1.638028	-0.10227
71	1.733074	-0.11689
72	1.722367	-0.03161
73	1.754606	0.004879
74	1.774363	0.048004
75	1.86028	0.019124
76	1.869415	0.061179

University of Malaya

Appendix E: Applied PCR-BPNN on Sweden and Turkish

Sweden			Turkey		
PCR model	error in BPNN	PCR+BPNN	PCR model	error in BPNN	PCR+BPNN
-1.187	0.069	-1.118	-1.488	0.094	-1.394
-1.168	-0.136	-1.304	-1.466	0.082	-1.406
0.134	0.060	0.194	-1.425	0.075	-1.391
0.296	1.124	1.420	-1.393	0.094	-1.331
-0.924	0.980	0.056	-1.357	0.093	-1.300
-0.922	0.848	-0.074	-1.317	0.100	-1.258
-0.870	1.004	0.134	-1.276	0.098	-1.219
-0.873	0.503	-0.369	-1.205	0.093	-1.183
-0.740	0.402	-0.337	-1.129	-0.001	-1.207
-0.742	0.417	-0.325	-1.168	0.014	-1.116
-0.926	0.140	-0.786	-1.066	0.094	-1.074
-0.870	0.061	-0.809	-1.022	-0.004	-1.070
-0.863	0.091	-0.771	-1.007	0.019	-1.003
-0.869	-0.215	-1.083	-0.975	0.013	-0.995

University of Malaya

-0.954	-0.767	-1.721	-0.926	0.001	-0.973
-0.812	-0.892	-1.704	-0.872	0.000	-0.926
-0.708	-0.737	-1.446	-0.912	0.001	-0.870
-0.849	-1.119	-1.969	-0.949	0.041	-0.871
-0.817	-1.506	-2.323	-0.828	0.094	-0.855
-0.649	-1.351	-2.000	-0.851	0.008	-0.820
-0.511	-1.087	-1.599	-0.771	0.096	-0.755
-0.540	-0.936	-1.476	-0.791	0.066	-0.705
-0.601	-1.009	-1.609	-0.722	0.097	-0.694
-0.489	-0.787	-1.276	-0.647	0.022	-0.701
-0.239	-0.726	-0.965	-0.635	-0.021	-0.668
-0.282	-0.713	-0.995	-0.667	-0.014	-0.649
-0.226	-0.814	-1.040	-0.580	-0.041	-0.708
-0.145	-0.793	-0.938	-0.673	-0.015	-0.595
-0.088	-0.032	-0.119	-0.598	-0.031	-0.703
-0.273	0.067	-0.207	-0.566	-0.052	-0.649
-0.725	0.094	-0.631	-0.495	-0.060	-0.626
-0.330	0.093	-0.238	-0.425	-0.066	-0.560
-0.154	0.122	-0.032	-0.361	-0.024	-0.449
-0.404	0.115	-0.289	-0.368	-0.016	-0.377
-0.251	0.118	-0.133	-0.293	-0.023	-0.391
-0.259	0.100	-0.159	-0.241	-0.028	-0.321
-0.121	0.112	-0.010	-0.174	-0.052	-0.294
-0.307	0.106	-0.201	-0.119	-0.060	-0.234
-0.271	0.096	-0.175	-0.055	-0.046	-0.165
-0.215	0.084	-0.131	0.033	-0.054	-0.109
-0.193	0.097	-0.096	0.141	-0.092	-0.059
-0.274	0.060	-0.213	0.181	-0.092	0.049
-0.203	0.081	-0.122	0.272	-0.024	0.157
-0.273	0.070	-0.204	0.283	-0.091	0.181
-0.307	0.097	-0.210	0.339	-0.035	0.248
-0.518	0.271	-0.247	0.446	-0.032	0.307
-0.339	0.397	0.058	0.422	0.003	0.449
-0.303	0.289	-0.014	0.495	0.017	0.439

University of Malaya

-0.161	0.336	0.175	0.588	0.011	0.506
-0.217	0.192	-0.025	0.619	0.030	0.618
-0.047	0.137	0.090	0.699	0.091	0.711
0.039	0.127	0.166	0.776	-0.027	0.673
0.087	0.120	0.207	0.833	0.030	0.806
0.148	0.121	0.269	0.841	-0.002	0.831
0.401	0.272	0.674	0.812	0.026	0.867
0.360	0.057	0.417	0.735	0.028	0.840
0.823	0.098	0.922	0.629	0.004	0.740
0.907	-0.018	0.889	0.672	0.096	0.725
1.123	0.140	1.264	0.776	0.094	0.766
0.919	0.285	1.205	0.810	0.094	0.870
0.691	0.540	1.230	0.905	0.193	1.003
0.467	0.505	0.972	0.946	0.084	0.989
0.589	0.522	1.111	1.002	0.100	1.046
0.641	0.443	1.084	1.123	0.085	1.088
0.960	0.421	1.381	1.265	0.092	1.215
0.983	0.221	1.204	1.323	0.092	1.358
1.203	0.052	1.255	1.420	0.097	1.420
1.532	0.077	1.609	1.456	0.094	1.514
1.651	0.100	1.751	1.604	-0.001	1.455
1.532	-0.297	1.235	1.651	-0.099	1.505
1.844	0.045	1.889	1.746	-0.009	1.642
1.774	0.096	1.870	1.764	-0.020	1.725
1.547	0.122	1.668	1.851	0.097	1.861
1.385	0.110	1.496	1.909	0.092	1.942

University of Malaya

Appendix F: Dimension Size of Input Dataset, Multicollinearity, and RMSE

The relation between the dimension size of input data set, multicollinearity, and RMSE rates have been tested using two prediction techniques (MLR and ANN) that most commonly employed by many researchers. To test this relation argument, this study tested two different prediction models with three different sizes of the input data set. This table illustrates the results of those tests. The results show the negative effect of the dimensionality size of input dataset on the performance accuracy indicator (RMSE). These results support the argument made by this work and different works as well.

#	Dimensionality size of input data set	Linear model MLR		nonlinear model ANN	
		RMSE	MAPE	RMSE	MAPE
	22 X 13	0.02	5.7341	0.036	4.205
	46 X 13	0.03	6.1561	0.051	4.8231
	74 X13	0.097	6.52713	0.083	5.5137

University of Malaya