ANALYSIS OF SHORT TERM LOAD FORECASTING TECHNIQUES

TAN VY LUOH

SUBMITTED TO THE
GRADUATE SCHOOL OF FACULTY OF ENGINEERING
UNIVERSITY OF MALAYA, IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF POWER
SYSTEM ENGINEERING

2019

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: TAN VY LUOH

Matric No: KQI170008

Name of Degree: Master of Power System Engineering

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

Analysis of Short Term Load Forecasting Techniques

Field of Study: Load Forecasting

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:

ABSTRACT

Nowadays, the implementation of advanced technology load and the introduction of multiple renewable energy sources to the grid have created major impacts to the electricity utilities provider with problems of power fluctuation, over generation and conventional power interruption. Therefore, short term load forecasting (STLF) is widely implemented as a necessary technique in power system planning and operation to ensure the power system is functioning in reliable and secure condition. In this report, three common numerical STLF techniques including Multiple Linear Regression (MLR), Curve Fitting and Bagged Tree Regression are proposed to forecast one-day ahead load profile with a yearly historical load data. The algorithms for each respective techniques are modelled in MATLAB Toolbox for simulation purpose. Forecasted curve of three techniques are obtained for evaluation with the diagnosis statistics including mean absolute percentage error (MAPE), mean absolute error (MAE), standard deviation absolute percentage error (StdAPE) and standard deviation absolute error (StdAE). The relative error between actual load and forecasted load is computed and used to compare the performance among three STLF techniques. As a result, bagged tree regression has lower relative error in MAPE and StdAPE which can be used to indicate it is more accurate STLF technique compare to the other two STLF techniques studied in this paper.

ABSTRAK

Kini, kebanyakan pengunaan alat teknologi modern dan pengunaan pelbagai jenis tenaga boleh baharu telak memainkan peranan penting dalam kehidupan kita serta membawa banyak implikasi kepada pihak penjanaan tenaga elektrik tentang masalah fluktuasi tenaga, kemungkinan untuk menjana sumber tenaga elektrik yang lebih dan gangguan bekalan elektrik. Untuk mengelakkan situasi atas berlaku, model prediksi beban listrik jangka pendek (short term load forecasting) telah dikembangkan dan dilaksanakan dalam proses operasi dan perancangan sistem tenaga elektrik. Dalam projek ini, tiga jenis teknik prediksi iaitu Multiple linear regression, curve fitting dan bagged tree regression telah dikembangkan dengan program komputer MATLAB Toolbox untuk simulasi proses bagi sehari masa tempoh. Keluk yang telah dihasilkan akan dinilai dengan statistik MAPE, MAE, StdAPE dan StdAE. Pembezaan antara keluk yang telah dihasilkan serta keluk sebenar juga dibangkitkan sebagai ralat yang dapat dibandingkan untuk menilai performasi teknik prediksi. Dapat dibuktikan dari sumber hasil bahawa teknik prediksi bagged tree regression mempunyai ralat yang paling rendah dalam statistik MAPE dan StdAPE serta merujukkan teknik ini lebih akurat berbanding dengan teknik yang dibangkitkan dalam projek ini..

ACKNOWLEDGEMENT

First and foremost, I would like to sincerely appreciate to my supervisor Dr Tan Chia Kwang for his great support and guidance throughout my research project and master study, offering invaluable advice. With his careful supervision and contribution to the direction, this project has done smoothly.

Besides, I would also like to acknowledge the lecturers and friends who contributed their assistance during the period oy my project. Their kindness on sharing of information and knowledge have ease my project’s progress.

Finally, I wish to express my deepest thanks to my family for encouraging me all the time especially when I was tough. They have motivated me to continue my project work in good performance and always stay far from indolence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neutral Network
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
CART	Classification And Regression Trees
DS	Decision Stump
LTLF	Long Term Load Forecasting
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLR	Multiple Linear Regression
RE	Renewable Energy
RF	Random Forest
RAE	Relative Absolute Error
RES	Renewable Energy Sources
RMSE	Root Mean Square Error
RRSE	Root Relative Squared Error
Reptree	Reduced Error Pruning Tree
SCR	Selective Catalytic Reduction
SSE	Sum Of Square Error
SVM	Support Vector Machine

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

Energy crisis is a rising issue all over the world due to the increasing demand of energy. To cope with this challenge for future development, load forecasting acts as the prominent process for the operation of power system and prepare schedule maintenance plan. Load forecasting basically can be classified into two main categories based on the variety of time horizon: very short term load forecasting (VSTLF), short term load forecasting (STLF), medium term load forecasting (MTLF) and long term load forecasting (LTLF). Short term load forecasting usually deals with the load ranging from a couple of hours to one day for load scheduling and generation planning (W. Lee, Jung, & Lee, 2017). Long term load forecasting analyses the load pattern throughout a year or a decade with seasonal weather factor and human behaviour.

The day-of-day operation of electricity utility system is the main focus for load forecast thus more approaches are developed for STLF. Every STLF forecasting algorithm has own advantages and drawbacks on data modelling process and assessment of performance (Debnath & Mourshed, 2018). The common challenges of load forecasting are the unexpected weather condition, usage behaviour varies between consumer and misunderstanding of data contributing factors (Kuster, Rezgui, & Mourshed, 2017).

1.2 Problem Statement

Load forecasting is one of the necessary process for economic and effective generation of power. Energy suppliers use load forecasting to make optimal unit commitment decision to reduce the spinning reserve capacity and determine the least cost decision which is mainly affected by the load demand and energy purchase. Furthermore, load forecasts is widely used by load serving entitles for system security and schedule generator operation with well-planned order dispatch. The unexpected variation of the load without accurate load forecasting leads to undesired increase in the reserve capacity. In other word, load forecasting acts as an initial step for electric supplier to schedule an operation decision with economic dispatch consideration.

Besides that, the utility system nowadays is further compounded with unpredicted renewable energy sources (RES) such as solar, wind, biomass and hydro. Integrating RES to the grid requires sufficient conventional power generation to serve as a backup capacity. Meanwhile, multiple RES injections might impact the cycling of conventional power plants. Cycling can be defined as starting-up, ramping up, ramping down and shutting down of the power plant generation. In order to cover variable residual demand, the electricity supplier has to control the generation output by ramping or switching on/off the cycling. The changes of operation mode requires long response time and it might cause mismatch between power generation and load demand within the time period.

The accurate STLF optimize the operation scheduling, reduce the cost and supply secure and economical electric energy to the customer. STLF provides the key information for power system scheduling, load flow analysis and day-to day operation for power delivery and planning. For instance, the residual load demand can be pre-determined by the load forecasting technique thus minimize the impact of cycling. During the event of power supply shortage due to intermittence nature of RE, the mismatched load can be covered by the conventional power plant without over-estimating or under-estimating and time delay.

Based on present researches, there are up to fifteen forecast approaches applying in energy planning model for short term load forecasting (STLF). Each approaches has its own strength and limitation to generate forecast load with different input parameter. The selection of these various forecast approach will directly affect the forecast result and they can be compared by determine the error between forecasted load and actual load.

In this paper, linear regression, curve fitting and bagged tree regression are used for the STLF of a conventional commercial building to analyse and compare the performance of these three forecast approaches. Real time load data of that particular building is extracted to build a pre-trained data pool for STLF techniques' training and validation then forecast upcoming 1 day load profile.

1.3 Objective

(i) To develop multiple linear regression, curve fitting and bagged tree regression algorithms for short time load forecasting (STLF).

(ii) To analyse the error rate of the developed STLF algorithms using MAPE, MAE STDAPE and STDAE.

(iii) To compare and discuss the performance of the different STLF algorithms in load prediction.

1.4 Scope and Limitation

(i) Among various STLF techniques, this project only discusses three conventional techniques such as linear regression, curve fitting and bagged tree regression for analysis and evaluation.

(ii) The pre-trained load data is purely based on the past load profile from IEEE & Kaggle - Global Energy Forecasting Competition 2012 website.

(iii) All coding and simulation in this report are conducted using MATLAB.

1.5 Outline of Research Report

This report is organized as follows: Chapter 2 presents the literature studies by other researches which is relevant to this research topic. Chapter 3 develop and study the process of constructing model for simulation and evaluation. Chapter 4 describes the results and discussions for the output achieved from previous chapter . Finally, Chapter 5 summarizes the findings, contribution of research and recommendations for this project.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In present, the global electricity sector is facing multiple power system planning and operation challenges such as power system resilience issue, the security of supply due to rising of demand and global trend toward urbanization. As bulk electrical energy is not easy to be stored and it aims to be generated wherever there is a demand for it, the power industry is required to estimate the load usage in advance. The process of load estimation in advance is likely to be called as load forecasting.

According to the time zone of planning strategy and the power system structure, load forecasting can be categorized into short term load forecasting, medium term load forecasting and long term load forecasting. Each forecast process is executed to predict the load demand for different time horizon with multiple consideration factors taking in account. With these, the power utility company could generate a comprehensive operation plan to encounter any contingency condition, to make decision on generating and purchasing power and to mitigate the risk related to energy crisis and market economic. As such, a reliable and precise load prediction could ease the participants of market to drive and expand their electricity business.

Short term load forecasting is to serve requisite information for the power system management of daily operation, fuel resources utilization and unit commitment for the period ranging from hour to week. For different industry practice and location region, there are different load behavior due to the weather condition, nature of business and authority policy. Therefore, multiple variety of short-term load forecasting techniques are developed and refined for precise demand prediction (Singh, Ibraheem, & Muazzam, 2013). Some research studies found that the appropriate mathematical tool will lead to more accurate forecasting technique when dealing with the randomness on the probabilistic load flow (García-Ascanio & Maté, 2010).

2.2 Short Term Load Forecasting (STLF) Techniques For Existing Power Planning.

Generally, STLF techniques are divided into four categories including statistical, deterministic, artificial intelligence as well as the hybrid technique. For statistical method, it can be further summarized into regression technique, stochastic time series technique and exponential smoothing technique. The regression-based technique is to predict the relative change of variable by the common relationships derived among the affected variables. Most of the regression algorithm is developed based on weather-dependent factor and socio-economic factor.

Stochastic time series technique can be classified into univariate and multivariate based on either one variable or multiple variables is varying over time. The most common time series technique are curve fitting, autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) (Fengxia & Shouming, 2011). These techniques predict future value based on previous observed values. Exponential smoothing technique is the advance method of stochastic time series with the use of window function to smooth data instead of moving average.

Aside from statistical techniques, neural network as a part of the artificial intelligent machine learning, it has been employed as the main load forecasting technique which is used extensively in current power system study. The properties of less complicated linear equation and rapid optimization process are the main advantages to support the utility company in optimizing their real-time operation and saving the investment for any additional facilities construction required especially with the well-development of artificial intelligence (AI) in recent years (Tealab, Hefny, & Badr, 2017). The common topology of neutral network is known as multilayer perceptron which comprises inner layer, multiple hidden layer and output layer (K. Y. Lee, Cha, & Park, 1992). Each layer consists numbers of neuron to represent the set of weights, input variable and the output signal. There are few researches stated that Artificial Neutral Network (ANN) model can be derived with feed forward and back propagation network (Widodo & Fitriatien, 2016) through using the historical electricity data as the input layer nodes for network training and machine learning in pre-processing stage of load forecasting (Papadopoulos, Delerue, Ryckeghem, & Desmet, 2017).

Likewise, the network architecture of neutral network can also be done in several classes such as recurrent network, radial basis function network and self-organizing maps for clustering (Han, Chen, & Qiao, 2010). The similarity of all neutral network machine learning methods for load forecasting is the repetitive of learning process through feeding the input-output patterns to the network until the convergence of the sum of squared error reach the minimum value. The differences among neutral network is the selection of activation function and the weights of neuron. Instead of neutral network, the most promising alternative of learning machine algorithm is support vector machine (SVM) which considering statistical learning in pattern classification problems.

Several studies proposed SVM method as the most practical short-term load forecasting due to its capability to solve linear constrained quadratic programming problem (Abbas & Arif, 2006). Moreover, SVM has few significant characteristics such as optimal generalization performance, no limitation due to local minima and sparse approximation of solution. There are two main approaches of SVM applied in load forecasting such as non-linear kernel-based method and selective catalytic reduction (SCR) method. Non-linear kernel-based SVM perform classification to take the load data as the input and outputs a line to separate class which is defined as hyperplane (Cocianu, 2013).

All individual STLF technique has its own limitation to handle the challenges of forecasting such as the randomness and non-linear system load, variation of socio-economic environment as well as the weather condition. Conventional load forecasting techniques has their advantages to predict linear load with the historical load data but might lack of accuracy when predict in some particular time zone which is strongly influenced by temperature sensitivities and real time control (Mat Daut et al., 2017). Therefore, to increase the system reliability, some studies propose the use of hybrid model comprises multiple load-forecasting approaches (Lajevardy, Parand, Rashidi, & Rahimi, 2015).

The combination of time series method and support vector regression is proposed (Nie, Liu, Liu, & Wang, 2012) while some research combined expert system with neutral network to predict faster and more accurate compare either one of two approaches alone (Wen, Li, Tan, Cao, & Tian, 2015). For instance, this hybrid model could accommodate the performance error occurred from time series method during initiate peak load by using support vector regression while time series method still contributing the seasonal load features and behavior of the load (José Montaño Moreno, Palmer, & Muñoz Gracia, 2011).

Even though artificial intelligent computing method such as genetic algorithm, neutral network, support vector machine and fuzzy logic are gaining more advantage for their effective use, most of the conventional load forecasting method such as multiple regression, exponential smoothing and stochastic time series are still commonly applying in modern power system especially when the load forecasting trend move toward hybrid model. In fact, the research has been changing and superseding the old approaches with better performance one.

2.3 Multiple Linear Regression Technique

2.3.1 Description of Multiple Linear Regression Technique

Multiple linear regression is a statistical attempt to model the linear relationship between dependent variable and multiple independent variables and it is broadly used in many research fields of economic, natural science and computer science (Amral, Ozveren, & King, 2007). The purpose of multiple linear regression is to evaluate and analyze both linear effect and integrated linear effect of each dependent variable or independent variables based on the observed data. Once the dependent variable with relative linear impact and independent variables have been chosen, the optimal multiple linear regression algorithm can be created to obtain the deviation degree (Suganthi & Samuel, 2012). As such, one of the major advantages of multiple linear regression is that it capable to resolve multivariate non-linear regression problem through translating the polynomial non-linear regression condition into multiple linear regression (Wang, He, & Nie, 2017).

In the multiple linear regression, the general model with random variable y and the explanatory variables $x_1, x_{2,} \dots, x_k$ which affected the load is expressed in the form as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (1)$$

Where y is the load (the dependent variable), and $x_1, x_{2,} \dots, x_k$ are the affecting factor such as weather, time of observation and etc. (k number of independent variables). The independent variables can be controlled and measured through the pre-processing data (Kumar, Mishra, & Gupta, 2016). When k = 1, equation 1 turns into linear regression model; when k is equal or greater 2, equation 1 serves as multiple linear regression model. An error term ε is defined in the equation 1 which refer to the undesired noise of the dependent variable y and ε usually has a mean value equal to zero and constant variance in the case of the probability distributions for the dependent variable y at the various level of the independent variable are in normal distribution as shown in bell shaped. Among equation 1, $\beta_0, \beta_1, \beta_{2,} \dots, \beta_k$ is k + 1 unknown parameters with β_0 as regression constant while $\beta_1, \beta_{2,} \dots, \beta_k$ is partial regression coefficient to be estimated from observation of y and x_k.

Practically, for n sets of data $x_{i1}, x_{i2,} \dots, x_{ik}; y_i$ in term of i = 1, 2, …, n, then the linear regression model can be expressed (Ferrera, Hu, Tomasi, & Pastrone, 2014) and written in matrix form,

$$\overline{y} = \overline{X}\overline{\beta} + \overline{\varepsilon} \tag{2}$$

Where $\overline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \overline{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \vdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \overline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \overline{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$

Matrix $\overline{X}$ is known as regression data matrix which comprises multiple basic linear regression to facilitate model parameter estimation with n sets of actual observed data. The predicted value of $\hat{y}$ is a dependent value with respect to the computed partial regression coefficient parameters $b_0, b_1, b_{2,} \dots, b_k$ are derived from observation of $\overline{y}$ and $\overline{X}$. The predicted value of $\hat{y}$ is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{3}$$

The procedure flow chart of regression model that indicate the methods to obtain a linear correlation coefficient formula is shown in Figure 2.1

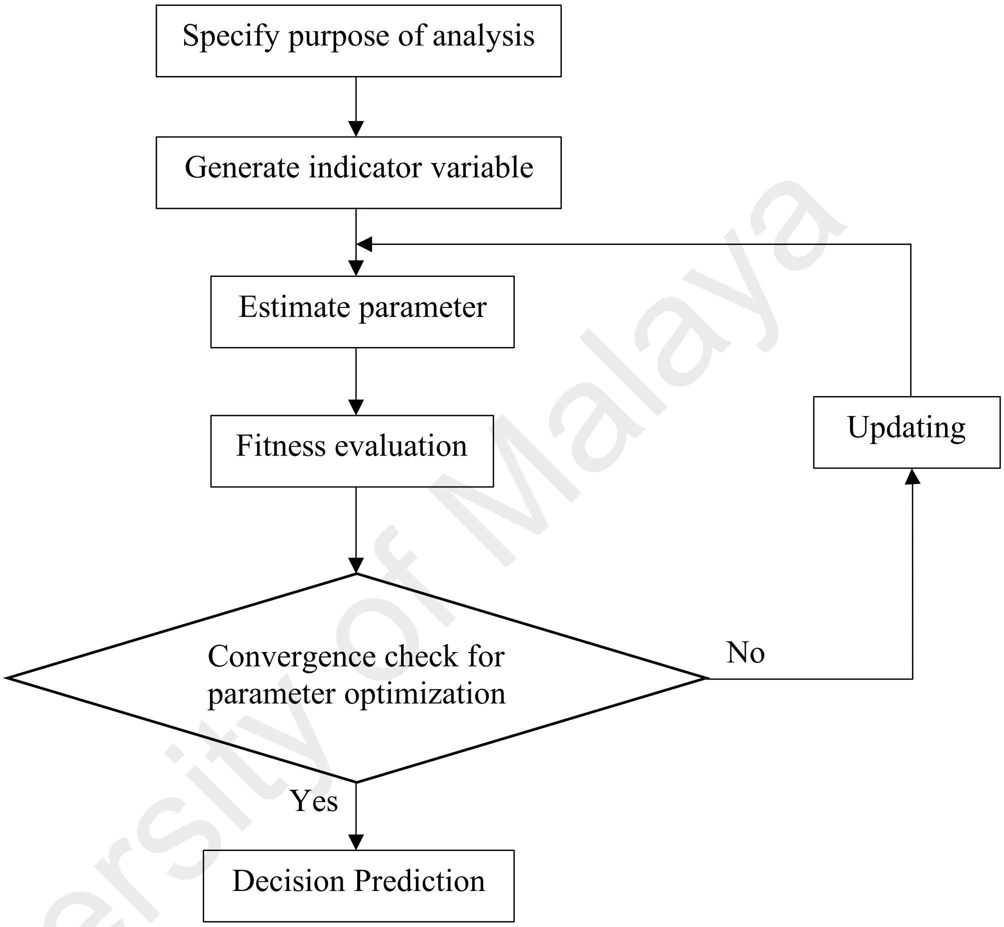
Specify purpose of analysis
Generate indicator variable
Estimate parameter
Fitness evaluation
Convergence check for parameter optimization
No
Updating
Yes
Decision Prediction

Figure 2.1: Procedure flow chart of regression modelling

2.3.2 Estimation of Regression Model Parameter

The most general used method in estimation for unknown regression parameter after the determination of multiple linear regression model is the ordinary least square method (Patel, Pandya, & Aware, 2015). This method aims to minimize the sum of squared error to get the parameters $\boldsymbol{b_i}$, (i = 1, 2, … , k):

$$[\boldsymbol{b_0}\ \boldsymbol{b_1}\ \boldsymbol{b_2}\ \dots\ \boldsymbol{b_k}]^T = (X^T X)^{-1} X^T Y \qquad (4)$$

Once the regression model parameters are computed, this model can be used for decision prediction, factor analysis and result optimization. For load prediction, the discrepancy between the predicted load value $\hat{\boldsymbol{y}}$ and the actual load value of y is defined as the residual sum of square (RSS) and the chi-square goodness of fit (R^2).

$$\text{RSS} = \sum(\boldsymbol{y_i(t)} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{(t)})^{\boldsymbol{2}} \tag{5}$$

$$R^2 = 1 - \frac{\sum(\boldsymbol{y_i(t)} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{(t)})^{\boldsymbol{2}}}{\sum(\boldsymbol{y_i(t)} - \bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{(t)})^{\boldsymbol{2}}} \tag{6}$$

$\boldsymbol{y_i(t)}$: Actual load from observation

$\hat{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{(t)}$: Estimation load derived from regression method

$\bar{\boldsymbol{y}}_{\boldsymbol{i}}\boldsymbol{(t)}$: Mean value of actual load

Assuming the independent variables have been observed comprehensively and therefore the standard error is minimized. Goodness-of-fit is used to determine the significance of regression parameters where comparing the observed sample distribution with the expected probability distribution (Hong, Gui, Baran, & Willis, 2010). The closer R^2 to 1 indicates that the optimal regression parameters are used to establish the best-fit model for prediction. Hence, the convergence checking for goodness-of-fit value is continuously updating to achieve model optimization. Then, the well-established multiple linear regression model can be utilized to make load prediction (Hahn, Meyer-Nieberg, & Pickl, 2009).

2.4 Curve Fitting Technique

2.4.1 Description of Curve Fitting Technique

Curve fitting is known as a regression analysis technique which aims to find the best-fit curve for a series of data sample (A Farahat & Talaat, 2012; Jain, Nigam, & Tiwari, 2012). Curve fitting is to establish a parametric equation that able to smooth the data and improve the behavior of plotted graph. Curve fitting technique can be divided into three common categories such as nonlinear curve fitting, smoothing curve fitting and least squares curve fitting (Molugaram & Rao, 2017). In nonlinear curve fit, researchers are required to specified and customize an equation to be fitted to the data. Generally, nonlinear curve fit is based on the Levenberg-Marquardt algorithm and is simulated using an iterative procedure. During the iteration process, the sum of squared error is calculated and re-evaluated with previous value until it reaches the best fit.

Least squares curve fitting is a technique that minimizes the square of the error between observed data and the predicted value by the pre-defined equation such as linear, polynomial, power, exponential and logarithmic. Besides, smoothing curve fits do not require to generate equation for the resulting curve. It uses variety of technique to arrive at the final curve with either incorporating a geometric weight or combining a series of cubic polynomials (Jingfei & Stenzel, 2005). Among three major curve fitting categories, least squares curve fit is the most popular use while nonlinear curve fit is more flexible to fit equation with multiple independent variables.

In most of the technical computing system, curve fitting technique come along with data pre-processing capabilities, pre-defined parametric and nonparametric models and also allow the users to customize their own model for specified data analysis (Willis, Powell, & Wall, 1984). Moreover, the integrated curve fitting technique is capable to differential, extrapolate and interpolate the fit during the post processing stage. After the curve fitting equation is built and trained, the curve fit result can be interpreted through correlation coefficients, parameter errors and the sum of the squared error (Chi Square). The typical flow chart of curve fitting prediction method (Reddy Cheepati & Nageswara Prasad, 2016) which is extracted from MATLAB toolbox is shown in Figure 2.2

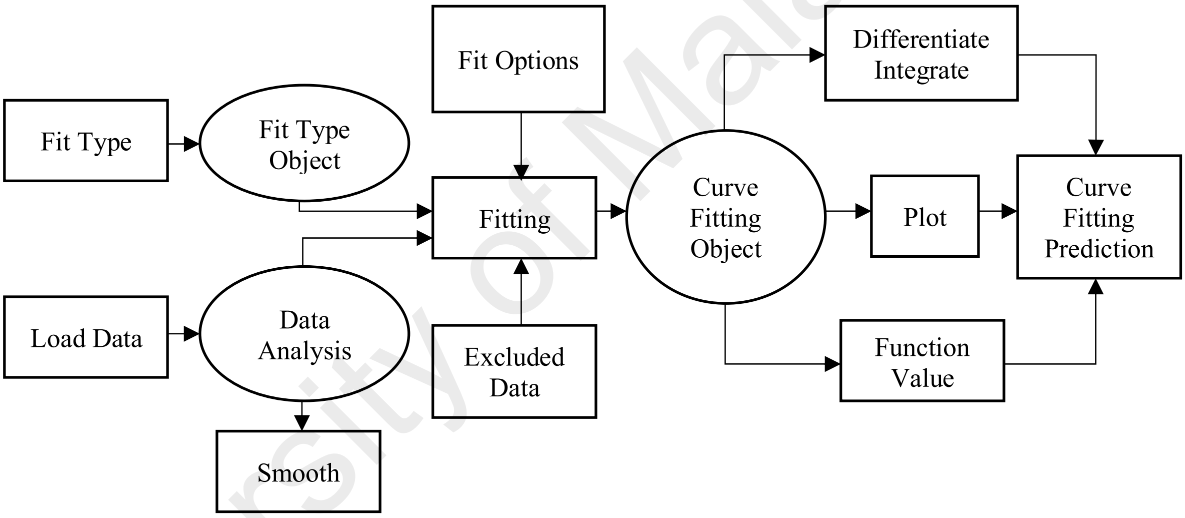

Figure 2.2: A typical flow chart of curve fitting prediction method

2.4.2 Approach of Curve Fitting to Determine the Best-Fit

Several studies have proposed two types of fit results such as goodness of fit and confidence intervals on the fitted coefficients to identify the accuracy of curve model and the performance for the curve fits the data. In certain cases, in order to reduce error from basic lease square method, orthogonal polynomial curve fitting is proposed (Aman, Simmhan, & Prasanna, 2015; Kafazi, Bannari, Abouabdellah, Aboutafail, & Guerrero, 2017) to make sure all vertical distance square sum between every points and fitting curve is minimum. Besides, genetic algorithms which widely used for optimization is proposed to improve the curve model fitting performance with analyzing the Sum of Square Error (SSE) and Root Mean Square Error (RMSE).

For instances, multiple pre-define functions are used as the training model to fit the data and these proposed models are evaluated by RMSE and SSE method. During every iteration, genetic algorithms or support vector regression are implemented to optimize the coefficients of initial equation then reduce the RMSE. When the RMSE reaches the convergence limit, the best-fit curve is obtained and is used to forecast one day ahead load.

2.5 Bagged Tree Regression Technique

2.5.1 Classification of Decision Tree

A decision tree is a graphical representation in form of tree structure with internal node that denotes a test on an attribute, branch expresses an outcome of test and leaf node (defined as terminal node) represents the class label (Dudek, 2015). The tree is built in the first stage by recursively splitting the training data into multiple subset based on the assigned optimal criteria. Basically, the construction of tree is completed when all subsets have been assigned the same class label. Then, some algorithms include pruning phase to minimize the outliers and noise that might cause overfitting problem to the data in the decision tree.

Generally, there are several kinds of decision tree classification algorithms including Classification and Regression Trees (CART), Reduced Error Pruning Tree (REPTree), Decision Stump (DS) and Random Forest (RF) proposed for the data mining process of load forecasting (Lahouar & Slama, 2015). Many studies employed and compared the efficiency of the algorithms to determine the best pattern in load forecasting model for the interpretation stage. Among them, CART is the earliest introduced non-parametric decision tree learning method that works in either classification or regression trees based on the nature of variables. The decision tree of CART is generated based on the valuable value to achieve best split at every node (Hambali, Akinyemi, Oladunjoye, & N, 2017). REPtree algorithm grew a decision tree with least complexity as it reduces error pruning as well as the error arising from variance. Decision stump is a one-layer decision tree model which purely depending on the single feature value. It is usually used as components in machine learning technique like boosting. A random forest is comprised of a set of decision trees which grow the tree in the same way as CART initially. Then, random selection of subset for best split at each node is done iteratively for each of the decision trees.

2.5.2 Description of Bagged Tree Regression

Bagged tree regression is one of the two ensemble techniques that can be used on the Classification and Regression Trees (CART) to improve the accuracy of decision tree. Ensemble methods including bagging and boosting are defined as a technique which combines several estimates to construct better predictive performance decision tree (del Carmen Ruiz-Abellón, Gabaldón, & Guillamón, 2018). For bagged tree regression, it is applied with bootstrap aggregation to reduce the variance of estimates. At first, bootstrap aggregation feeds multiple random observations from training set to train multiple models through bootstrap sampling with replacement. Then, all outputs from different fully-grown trees are aggregated into one large model then the average of predicted outcome is used for cross validation.

2.6 Assessment Method for STLF Method

For load forecast evaluation, there are multiple measures are taken to interpret and compare the prediction model's performance from various aspects. Some studies focus on the measures of cost, time spend, sensitivity and relationship for decision making through load prediction. There are several comprehensive measures have been proposed to assess load forecasting as a multi-dimensional problem such as scale-independent errors, reliability of model, development cost of model and application independent.

Scale-independent error is the most common evaluation measurement which is observed through the difference between the forecasted value and observed value (Zhang et al., 2012). Reliability of model indicates the consistency of model performance to produce similar unseen data relative to the testing data in future. Meanwhile, the development cost of model is expressed in term of time and effort for data pre-processing, model training and analysing. Application independent measures are based on the comparison across different load forecasting models (Aman et al., 2015).

Most load forecasting studies evaluate model based on the statistical measures which can be divided into two main categories: standalone accuracy measures and relative measures. In general, the most used evaluation indicators are Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and Standard Deviation Absolute Percentage Error (StdAPE)

2.6.1 Mean Absolute Percentage Error

The MAPE is scale independent statistical measure that is determined using the absolute error in each period normalized by the observed value in same period then present in percentage form (Khair, Fahmi, Hakim, & Rahim, 2017; W. Lee et al., 2017). This measure is commonly used to determine the error in predicting compare with the actual value.

$$MAPE = \frac{1}{n}\sum\left|\frac{y_1 - y'_t}{y_1}\right| x\ 100\%$$

2.6.2 Mean Absolute Error

The MAE is scale dependent statistical measure that is computed using the average of absolute error between forecasted load and observed load.

$$MAE = \frac{1}{n}\sum|y_1 - y'_t|$$

2.6.3 Root Mean Squared Error

The RMSE is also scale dependent statistical measures which is calculated by taking the squared root of the average of the squared differences between each forecasted value and its respective actual value (Vazquez et al., 2017).

$$RMSE = \sqrt{\frac{\sum(y_1 - y'_t)^2}{n}}$$

2.6.4 Relative Absolute Error

The RAE is the error that related to the average of the actual values. It is measured by taking the sum of absolute error and normalizes it by dividing by the total absolute error of the average of the actual values.

$$RAE = \frac{\sum|y_1 - y'_t|}{\sum|y'_t - \bar{y}|}\,,\ \bar{y} = \frac{\sum y'_t}{n}$$

2.6.5 Root Relative Squared Error

The RRSE is the advanced measures of relative squared error through square root function.

$$RRSE = \sqrt{\frac{\sum(y_1 - y'_t)^2}{\sum(y'_t - \bar{y})^2}}\,,\ \bar{y} = \frac{\sum y'_t}{n}$$

2.6.6 Standard Deviation of Absolute Percentage Error

The StdAPE is a measure of volatility to study the spread of forecasted load at each period of time by comparing the relative error with the average relative error, $\hat{y}$

$$StdAPE = \sqrt{\frac{\sum[\left(\frac{y_1 - y'_t}{y_1}\right) - \hat{y}]^2}{n-1}}\,,\hat{y} = \frac{\sum\frac{y_1 - y'_t}{y_1}}{n}$$

All above assessment metrics are commonly used to assess STLF techniques from different perspective. The advantages of MAPE and MAE on assessing STLF are due to their scale-independency and interpretability but the disadvantage is the probability of creating infinite or undefined values for zero (Kim & Kim, 2016). RMSE has the advantage of penalizing undesirable large error due to it gives a relatively high weight to large errors. Standard deviation is used to indicate the spreading of data compare to the mean in order to determine the system consistency.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This project presents the performance analysis of three numerical short term load forecast techniques such as multiple linear regression technique, curve fitting technique and bagged tree regression technique. The use of techniques will contribute in different accuracy forecasting result in the same circumstance. To ensure fair comparison of STLF techniques' performance, same load data will be extracted and sampled for data processing. Then, the methodology is arranged in the following sequence – data pre-processing, modelling of STLF techniques, scaling and training of load data, simulation and error analysis between forecasted load and actual load to examine the accuracy performance of STLF techniques.

The flow diagram of general methodology is expressed in figure 3.1:

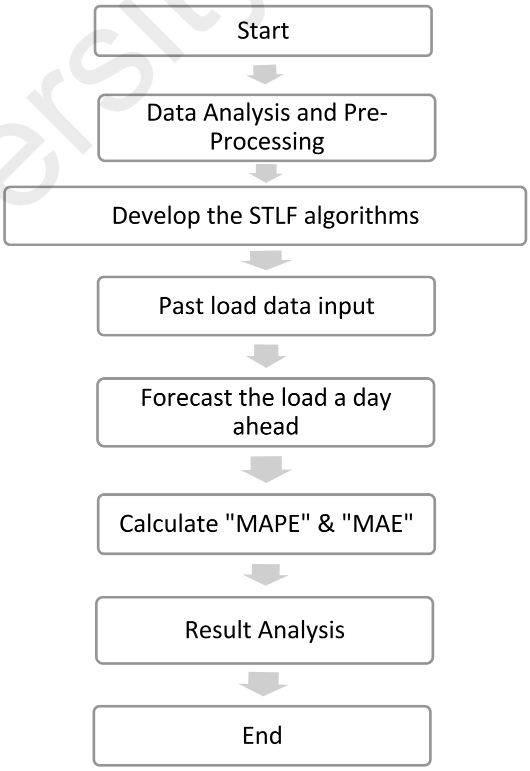
Start
Data Analysis and Pre-Processing
Develop the STLF algorithms
Past load data input
Forecast the load a day ahead
Calculate "MAPE" & "MAE"
Result Analysis
End

Figure 3.1: Flow diagram of proposed STLF methodology

3.2 Data Pre-processing

In this project, load data is obtained from the database source from Global Energy Forecasting Competition 2012 website (IEEE Power & Energy Society, 2012) which comprise hourly load for a year. The dataset containing hourly load from 1 January 2004 to 31 December 2004 are used for training and sampling purpose. Load data of 1 January 2005 is used for leave-one-out cross validation purpose. For each day, the datasets are separated into twenty-four subset which represent a specific hour of the day. This will allow us to perform comparison analysis between the load of present hour and the load of previous day or week same hour.

At the beginning of project, we convert and store the calendar variables of training load dataset into strings comprises year, month, day and hour respectively. Then, we further categorize the day parameter into corresponding day of week for grouping purpose. Meanwhile, the training loads as well are imported from the dataset and convert into strings according to the time sequence in hour basis. Later, the loads are sort and stored as previous day same hour load, previous week same hour load and previous twenty-four-hour average load. All missing load data of each category is preset as dummies variable (NaN) before the training process.

Besides, the validation dataset is also converted and stored in string form. The last 24 load data before end of string are sorted and stored into the validated previous day same hour load, validated previous week same hour load and validated previous twenty-four-hour average load to evaluate predictive model. In this case, both training set and testing set are ready to proceed with the next step – modelling of forecast technique.

3.3 Modelling for Short Term Load Forecasting (STLF) Techniques

3.3.1 Multiple Linear Regression Technique

The implementation of multiple linear regression technique for this forecasting study is based on hourly load which modelled as two main components. One of the main components is the time of observation which represent the day of week. Another main component is the load for different time period, such as previous week load and previous day load. The relationship between the time of observation and the historical load profile are expressed in multiple linear equation.

The multiple linear regression model is expressed as

$$y(t) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Where:

$\beta_1, \beta_2, \beta_3, \beta_4$ = partial regression coefficients

$x_1 = L\ (t\text{-}24)$: represent previous day same hour load

$x_2 = L\ (t\text{-}720)$: represent previous month same hour load

$x_3 = \frac{\sum_{t-24}^{t} L\ (t)}{24}$: represent previous day average hour load

$x_4 = d$: represent day of week

(1 to 7 corresponding to Monday to Sunday)

Previous day same hour load illustrates the seasonality impacted load which due to high day and low night temperature as well as human activity at daily resolution. The relationship between load and temperature for monthly resolution can be described and modelled by previous month same hour load due to the variance of weather sensitivity component is reflecting to the load. Day of week is modelled to assign the load into respective Mondays to Sundays to analyze the cross effect between the load profile and hour in the particular day (Pahasa & Theera-Umpon, 2007).

The simulation of multiple linear regression is based on the past one-year load profile in hour basic to predict a daily load for one day ahead. The aforementioned four partial regression coefficients are calculated by using the hourly data in respective day of week for every time interval. These parameters are used for cross-validation with the leave-one-out testing set to forecast a daily load curve which consists the generalization error estimate.

3.3.2 Curve Fitting Technique

The implementation of curve fitting is modelled based on nonlinear curve fit comprises fourier function and root mean square error (RMSE) function. At first, a fourier series equation is created and specified through “fittype” function for the training load data as a data analysis input to create a curve fit. For validation data, we transform the non-linear daily input into a polynomial sequence with coefficient through regression method.

The fourier equation is described as $f(x) = a_0 + a_1 \cos(\omega x) + b_1 \sin(\omega x)$. a_0, a_1 and b_1 are a dummy coefficient that keep varying until a good fit is obtained. The goodness of fit is based on the RMSE comparing the previous daily load with the present fit model within same hour interval. RMSE states the boundary limitation for the dummy coefficients to converge till the best-fit fourier equation.

The best fit curve represents a smoothing curve that had trained with the regression coefficient along the entire day range. Then, the predicted load is obtained by considering the trend of the best fit curve and the left-one-out testing sample. For a day, load data is split into 24 hours scatter plot. The vertical distance from the load of the best fit curve interpolates into testing sample to generate the predicted load.

The algorithm equation between power and its training parameter is expressed in

$$y(t) = a_1 + a_2.P_1 + a_3.P_2 + a_4P_3$$

Where:

a_1, a_2, a_3, a_4 : the coefficients derived from the previous load data of similar hour in day, month and the day type.

P_1 : previous day same hour load

P_2 : previous week same hour load

P_3 : day type

3.3.3 Bagged Tree Regression Technique

The implementation of bagged tree regression starts from generate predictor of tree and initiate the prune size for constructing a type of bagged tree technique which is called as random forest. The predictor of tree consists of historical load data and time of observation. With these training observations, an ensemble of bagged regression tree with specified 50 numbers of weak learner is trained. In random forest technique, weak learner is commonly defined as the input of decision tree for training. The numbers of weak learner determine how accurate and strong the regression tree model can be when combining all of them through ensemble model.

In this paper, the model is built by 24 random forest predictors standing for the 24 hours of the day. The load predictor is affected by few input parameters such as previous day same hour load, previous month same hour load, type of the day, previous 24 hours average load. The output is the load of day at same hour. With these random forest predictors, the regression tree model can predict the 24 coming hours of load demand.

To obtain the optimal size of the tree without overfitting the training data and generate accurate information, the minimum leaf size observation is set as 1 to grow deep tree for large training sample size. Once the regression tree had trained, out-of-sample prediction is used to obtain the validation observation for each of several subtrees. The terminal nodes of the regression tree contain the predicted output variable values. Analyzing through this parameter in each regression tree with the left-out testing sample, the most accurate cross validated predictions can be produced. In other word, the regression tree is constructed with large number of terminal nodes to optimize the performance of predictor. The flow diagram of bagged tree regression technique is expressed in following figure 3.2:

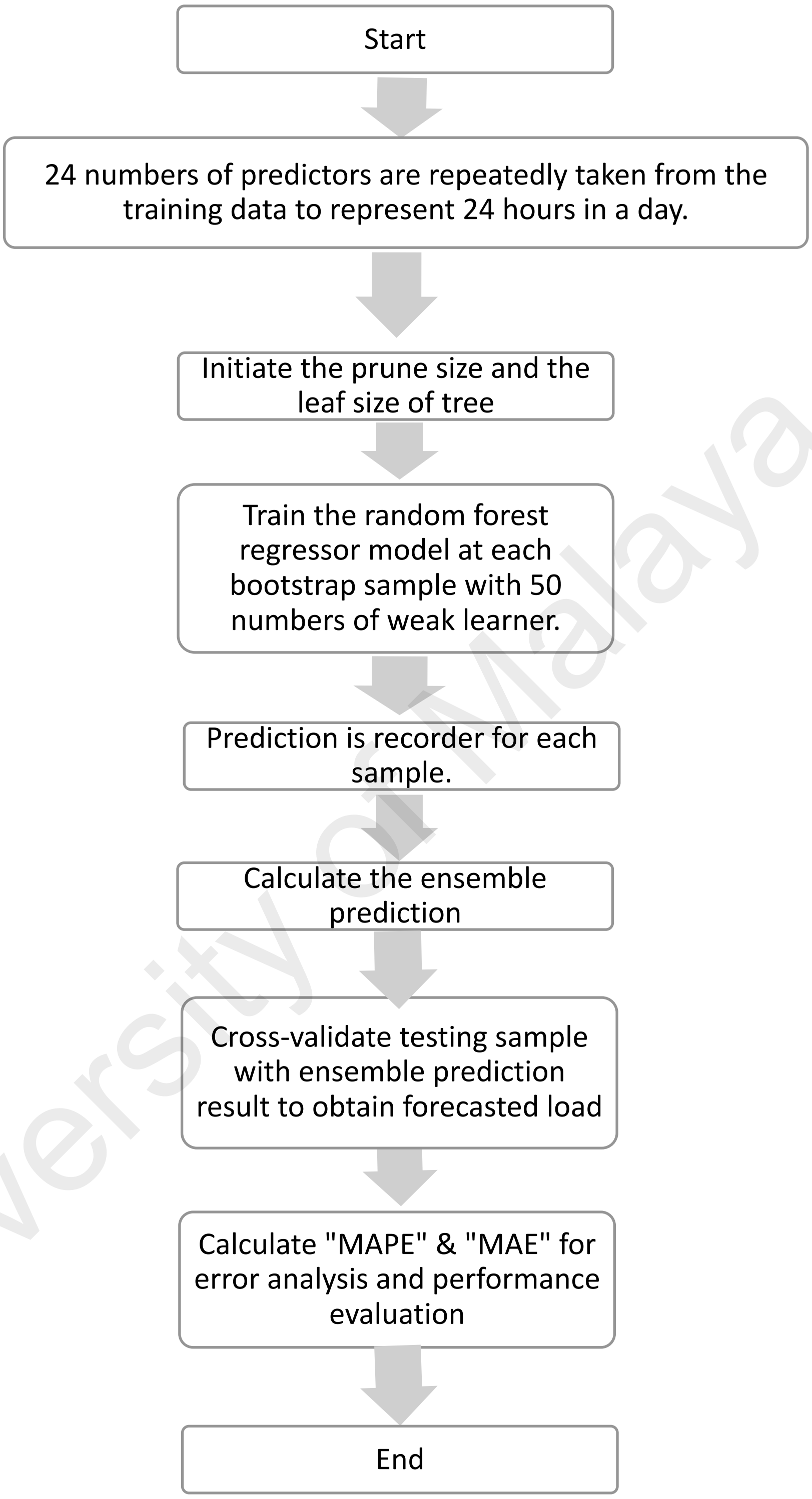
Start
24 numbers of predictors are repeatedly taken from the training data to represent 24 hours in a day.
Initiate the prune size and the leaf size of tree
Train the random forest regressor model at each bootstrap sample with 50 numbers of weak learner.
Prediction is recorder for each sample.
Calculate the ensemble prediction
Cross-validate testing sample with ensemble prediction result to obtain forecasted load
Calculate "MAPE" & "MAE" for error analysis and performance evaluation
End

Figure 3.2: Flow diagram of bagged tree regression technique

3.3.4 Evaluation Metric And Error Analysis

The evaluation metric is to assess the performance of STLF model through analyze the error between predicted and actual load. In this report, there are two main error analysis technique had applied such as mean absolute percentage error (MAPE) and mean absolute error (MAE). The error analysis techniques are defined in the following Equation (1) and (2):

$$MAE\ (y,p) = \frac{1}{n}\sum_{t=1}^{n}|y_t - p_t| \tag{1}$$

$$MAPE\ (y,p) = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - p_t}{y_t}\right| \tag{2}$$

Where

y_t : actual load value

p_t : forecasted load value

n : number of samples

t : hour

For each STLF techniques, the forecasted load curve is displayed on a graphic including all forecasted hour in the validation phase. The forecasted load curve is used to examine the error which reflected in load curve with the actual load curve for a day ahead. The variance of load for each hour is assessed with the average percentage relative error curve.

3.4 Summary of Chapter

This chapter describes the research methodology used to model STLF techniques and analyse the data to achieve the research objective. First, data pre-processing introduces the source of load data, transform the raw data into understandable and readable format for coding and modelling tasks. The modelling of three STLF techniques is one of the main research aims to generate consistent and comprehensive system to make the analysis easier to understand and visualize. After that, the assessment of STLF techniques through simulation is presented. Lastly, the error analysis is carried out to examine the performance of STLF techniques.

CHAPTER 4: RESULT AND DISCUSSION

4.1 Introduction

The chapter first presents the forecasted load curve and actual load curve for both multiple linear regression technique, curve fitting technique and bagged tree regression technique using MATLAB Toolbox. Then, the result of evaluation metric such as MAPE and MAE are calculated and presented. For each respective technique, the relative error between forecast load and actual load in per hour basic is presented. In the last section, the comparison of numerical STLF technique is presented.

4.2 Result of Forecasted Load in STLF

This section presents the result of the forecasted load in short term load forecasting technique including multiple linear regression, curve fitting and bagged tree regression.

4.2.1 Multiple Linear Regression Technique

In this section, the result of multiple linear regression with four input parameters including previous day same hour load, previous month same hour load, previous day average hour load and day of week are presented including the load curve, percentage relative error curve and error metric. The model is derived with the regression parameters which have been calculated using historical hourly load data. These parameters are shown in Table 4.1:

Table 4.1: Regression parameter of multiple linear regression technique

Regression Parameter	Simulation Result
β_1	0.6957
β_2	0.230
β_3	0.0705
β_4	15.209

For every hourly time interval, the multiplication of regression parameter and validation load data generates the forecasted load. The forecasted load against actual load for one day ahead are simulated and illustrated in Figure 4.1.

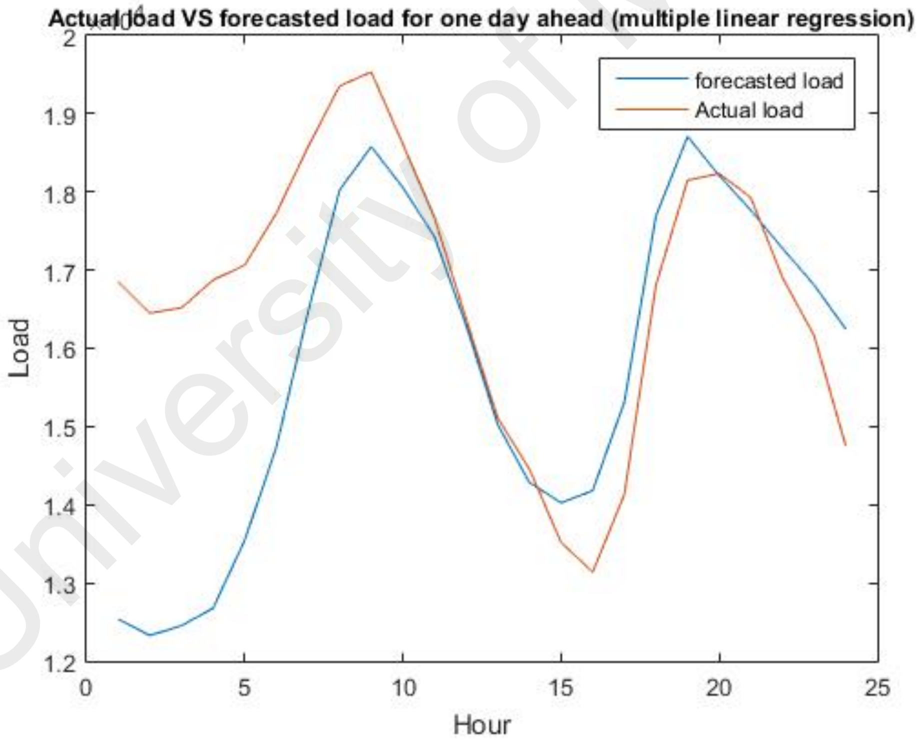

Figure 4.1: Comparison of actual load and forecasted load for one day ahead in multiple linear regression technique

To evaluate the performance of MLR technique for STLF, the relative error between forecasted load and actual load are computed and shown in Table 4.2. Besides, the mean absolute percentage error (MAPE) was calculated as 8.88% while the mean absolute error (MAE) was calculated as 1484.15kW. The percentage relative error curve for MLR is presented in Figure 4.2 to depict the error variation that was generated through MLR.

Table 4.2: Actual, forecast load and absolute percentage relative error for MLR

Time	Actual Load, W	Forecast Load, W	Absolute Percentage Relative Error, %
0100	16853	12536	25.61
0200	16450	12328	25.06
0300	16517	12453	24.60
0400	16873	12676	24.87
0500	17064	13544	20.63
0600	17727	14732	16.90
0700	18574	16459	11.39
0800	19355	18023	6.88
0900	19534	18579	4.89
1000	18611	18055	2.99
1100	17666	17427	1.35
1200	16374	16294	0.49
1300	15106	15022	0.56
1400	14455	14285	1.18
1500	13518	14024	3.75
1600	13138	14180	7.93
1700	14130	15310	8.35
1800	16809	17693	5.26
1900	18150	18708	3.07
2000	18235	18208	0.15
2100	17925	17765	0.89
2200	16904	17278	2.21
2300	16162	16809	4.00
2400	14750	16245	10.13
		MAPE	8.88

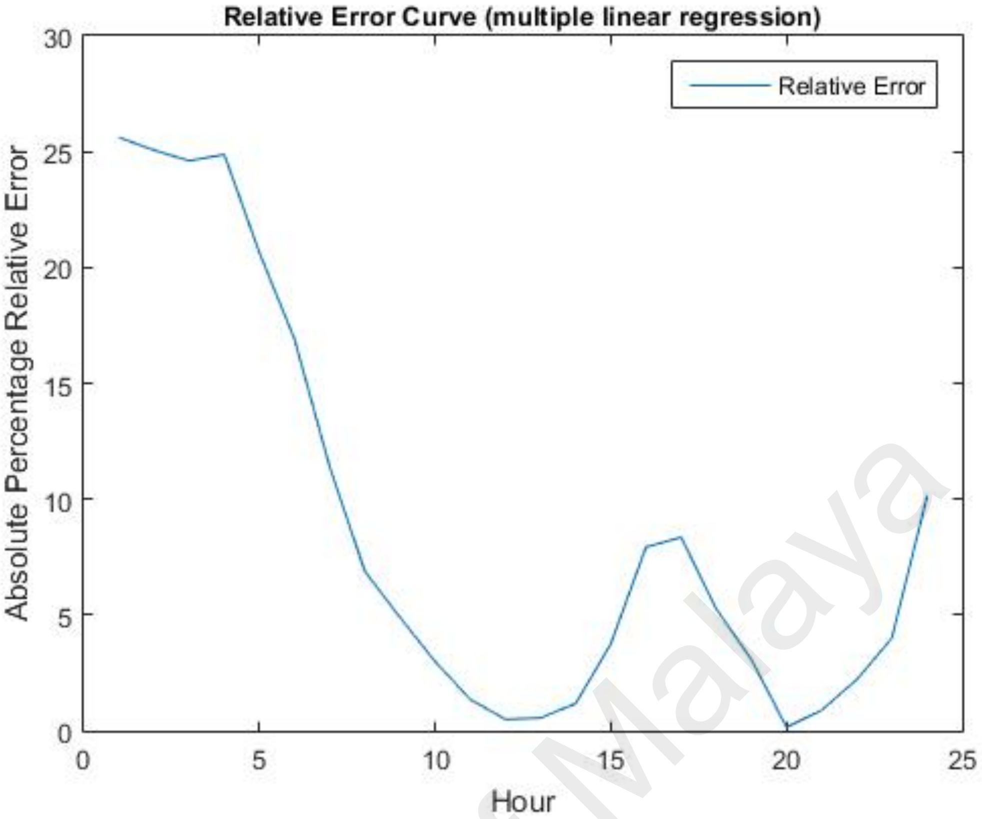

Figure 4.2: Relative Error Curve (Multiple Linear Regression)

In figure 4.2, the highest absolute percentage relative error is 25.61% during early morning while the lowest absolute percentage relative error is obtained at night time which is 0.49%. The relative error went down during peak load period and spike up when the load demand drops during off peak period.

4.2.2 Curve Fitting Technique

In this section, the result of curve fitting which implemented in fourier equation is presented including the load curve, percentage relative error curve and error metric. The model is based on the "goodness to fit" principle which initially built with the linear regression parameters as the primary input. After few iterations, the best-fit forecasted load against actual load curve for one-day ahead was achieved and shown in figure 4.3.

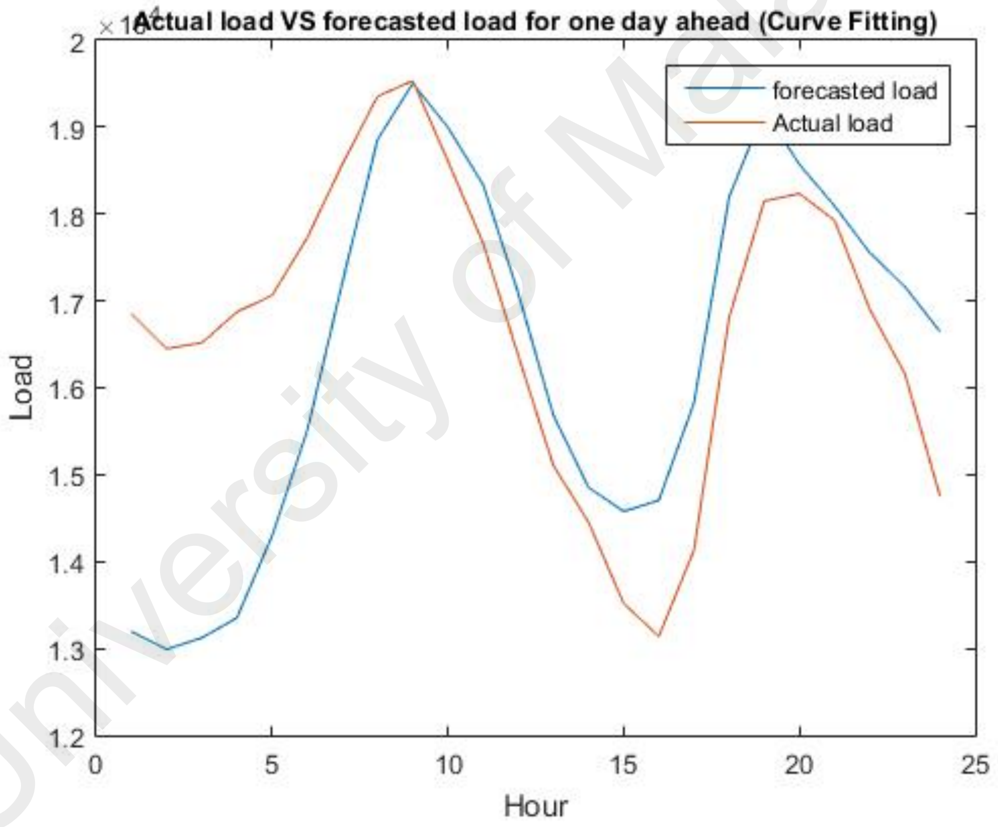

Figure 4.3: Comparison of actual load and forecasted load for one day ahead in curve fitting technique

To evaluate the performance of curve fitting technique for STLF, the relative error between forecasted load and actual load are computed and shown in Table 4.3. Besides, the mean absolute percentage error (MAPE) was calculated as

8.803% while the mean absolute error (MAE) was calculated as 1436.37kW. The percentage relative error curve for curve fitting is presented in Figure 4.4 to depict the error variation that was generated through curve fitting.

Table 4.3: Actual, forecast load and absolute percentage relative error for curve fitting

Time	Actual Load, W	Forecast Load, W	Absolute Percentage Relative Error, %
0100	16853	13194	21.71
0200	16450	12989	21.04
0300	16517	13121	20.56
0400	16873	13352	20.86
0500	17064	14294	16.23
0600	17727	15504	12.54
0700	18574	17212	7.34
0800	19355	18857	2.57
0900	19534	19503	0.16
1000	18611	18997	2.07
1100	17666	18339	3.81
1200	16374	17091	4.38
1300	15106	15686	3.84
1400	14455	14852	2.75
1500	13518	14577	7.83
1600	13138	14704	11.92
1700	14130	15824	11.99
1800	16809	18195	8.24
1900	18150	19192	5.74
2000	18235	18570	1.84
2100	17925	18093	0.93
2200	16904	17555	3.85
2300	16162	17163	6.19
2400	14750	16646	12.86
		MAPE	8.80

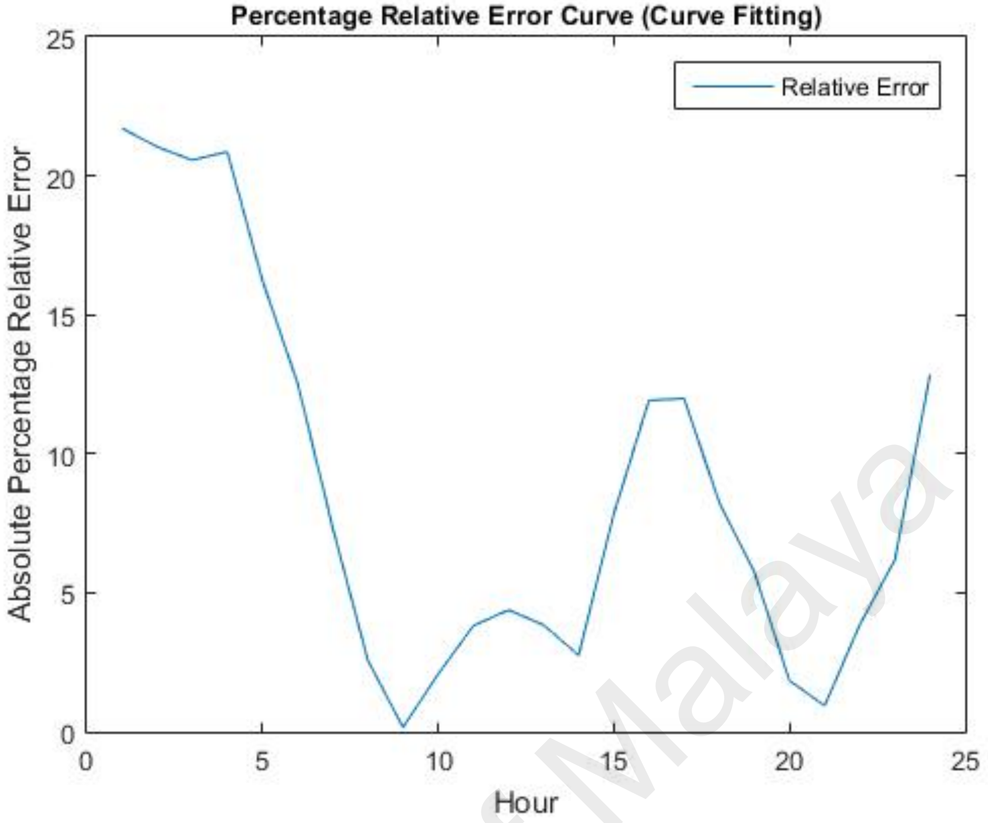

Figure 4.4: Comparison of actual load and forecasted load for one day ahead in curve fitting technique

For curve fitting technique, the highest absolute percentage relative error is 21.71% during early morning while the lowest absolute percentage relative error occurred in the late morning at 0.16%. The relative error curve of curve fitting illustrates similar pattern to regression method which varies according to the load demand. During peak load time, respective lower relative error is obtained especially in the transition period from low load demand to high load demand. In opposite, the relative error goes up and fluctuate between low load usage period.

4.2.3 Bagged Tree Regression Technique

In this section, the result of bagged tree regression which applied with ensemble learning method is presented including the load curve, percentage relative error curve and error metric. Multiple weak learners in ensemble term groups together to form a strong learner, so called as random forest. The final result may either be a weighted mean or mean of all of the terminal nodes that are reached. In this case, the optimal forecasted load of bagged tree regression against actual load curve for one-day ahead was achieved and shown in figure 4.5.

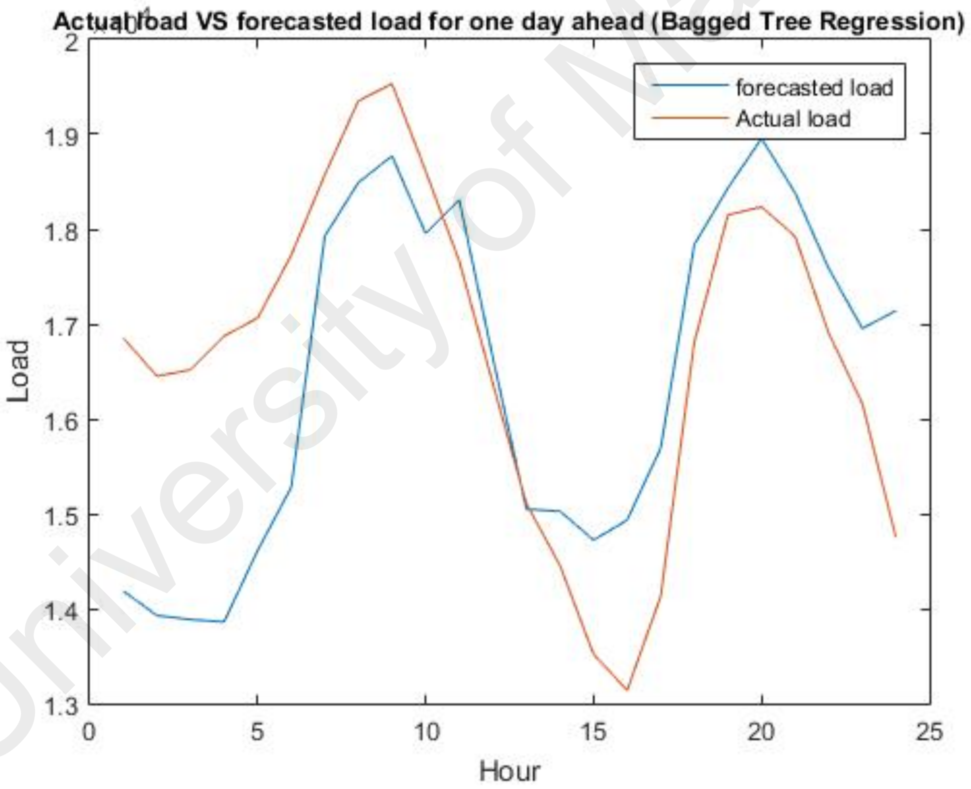

Figure 4.5: Comparison of actual load and forecasted load for one day ahead in bagged tree regression technique

To evaluate the performance of bagged tree regression technique for STLF, the relative error between forecasted load and actual load are computed and shown in Table 4.4. Besides, the mean absolute percentage error (MAPE) was calculated as 7.96% while the mean absolute error (MAE) was calculated as 1296.82kW. The percentage relative error curve for bagged tree regression is presented in Figure 4.6 to depict the error variation that was generated through bagged tree regression method.

Table 4.4: Actual, forecast load and absolute percentage relative error for bagged tree regression

Time	Actual Load, W	Forecast Load, W	Absolute Percentage Relative Error, %
0100	16853	14184.91	15.83
0200	16450	13926.96	15.34
0300	16517	13884.27	15.94
0400	16873	13860.32	17.86
0500	17064	14613.74	14.36
0600	17727	15279.52	13.81
0700	18574	17928.36	3.48
0800	19355	18492.66	4.46
0900	19534	18771.71	3.90
1000	18611	17954.35	3.53
1100	17666	18309.70	3.64
1200	16374	16646.33	1.66
1300	15106	15047.82	0.39
1400	14455	15026.53	3.95
1500	13518	14722.64	8.91
1600	13138	14933.39	13.67
1700	14130	15692.78	11.06
1800	16809	17838.04	6.12
1900	18150	18433.18	1.56
2000	18235	18956.86	3.96
2100	17925	18382.17	2.55
2200	16904	17585.97	4.03
2300	16162	16952.34	4.89
2400	14750	17140.40	16.21
		MAPE	7.96

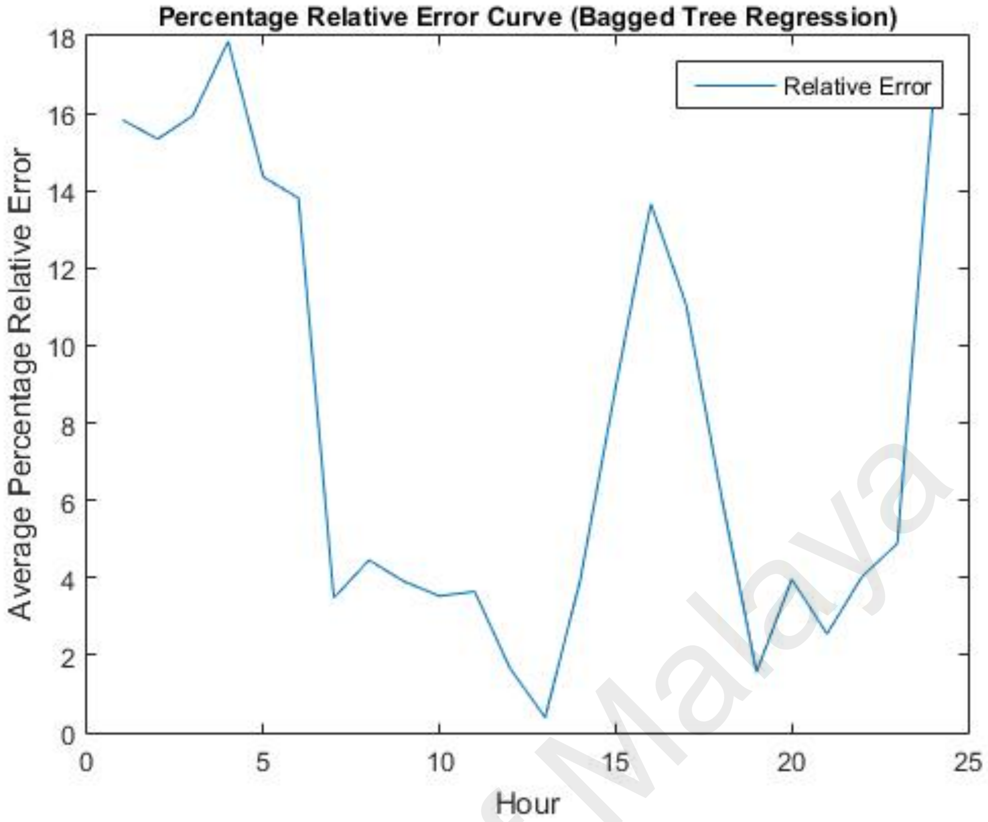

Figure 4.6: Comparison of actual load and forecasted load for one day ahead in bagged tree regression technique

For tree bagged regression technique, the highest absolute percentage relative error is 15.83% during early morning while the lowest absolute percentage relative error occurred in the late morning at 0.39%.

Due to the characteristic of bagged tree regression method as a random forest, every program execution creates a subset of the data randomly to form a forest with varies tree size. In addition of the random predictor variables are selected differently from node to node along with the higher possibility of noise, the final terminal node that is reached might different to the previous execution. Hence, the tree bag regression with random forest algorithm predicts the forecast load through optimization process.

4.3 Performance Evaluation Among STLF Techniques

Comparing the relative error and mean absolute percentage error for three discussed STLF techniques, the overall error is close but slightly different from time to time. As it shown in the relative error curve for all three techniques, the index can efficiently evaluate the accuracy of forecast load under the same input load data. The comparative STLF of the load has been tabulated below Table 4.5 to show the variation of load with respect to the diagnosis statistics and the comparison of relative curve among three STLF techniques is illustrated in Figure 4.7:

Table 4.5 Diagnosis Statistics

STLF Techniques	MAPE (%)	StdAPE (%)	MAE (W)	STDAE (W)
Multiple Linear Regression	8.88	8.96	1484.15	1510.36
Curve Fitting	8.80	6.99	1436.37	1155.79
Tree Bagged Regression	7.96	5.82	1296.82	930.29

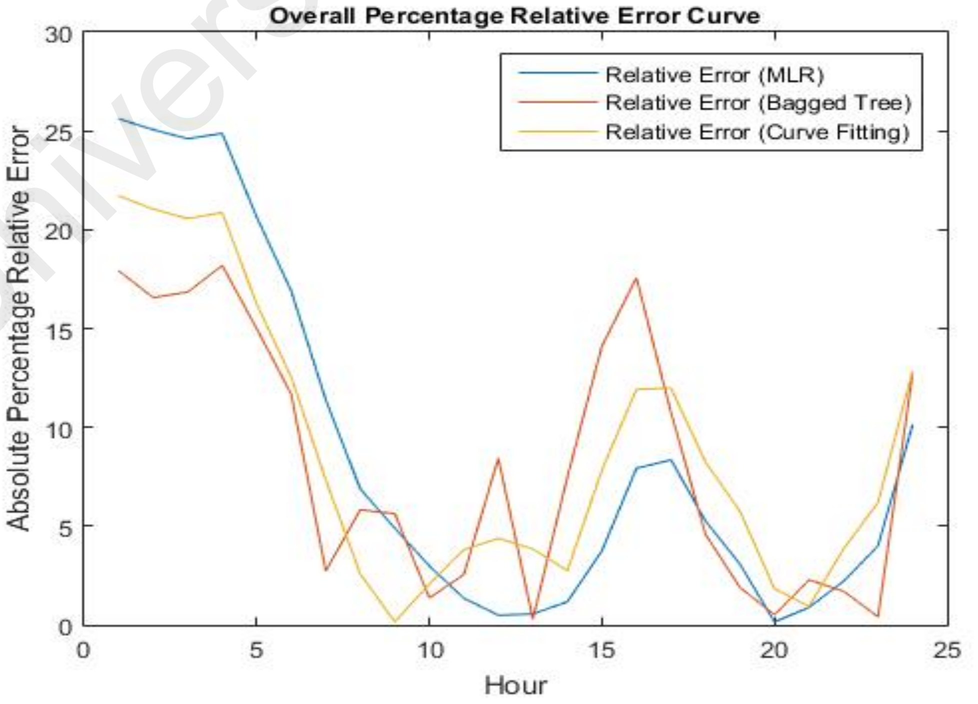

Figure 4.7: Comparison of relative error among three STLF techniques

To have better understanding on the STLF techniques' performance, we further introduce the standard deviation diagnosis statistic to study how the load data is distributed about the mean value throughout the observation period. From Table 4.5, we know that tree bagged regression generate the forecast load curve with lowest StdAPE and StdAE which indicate that this technique capable to provide the most accurate prediction among three STLF techniques.

Meanwhile, curve fitting and multiple linear regression have a very close MAPE but the multiple linear regression produces a load prediction with greater standard deviation. The reason is mainly due to the wide range of predictor parameter creates white noise and also MLR technique unable to forecast beyond the training data. Thus, the forecasted result may contain unnecessary noise error and the equation only limited to the particular regression coefficient without cover all affecting factors.

For curve fitting technique, the best-fit curve is obtained through fitting multiple input parameters into the defined fourier equation on top of the basic linear regression. Therefore, the standard deviation of average percentage error seems to be slightly lower compare to the multiple linear regression technique. In the other words, the number of prediction variable determines how efficient and consistent the performance of STLF techniques are to generate accurate forecast load curve.

Overall, three STLF techniques have interpreted similar load profile for cross validation. The high relative error which has occurred on the morning and early evening indicate the training data consists numbers of undesired intercorrelation. Moreover, the prediction during the peak load period has the lowest relative error, therefore, the forecasted peak load is in high precision. Comparing the diagnosis statistics of numerical STLF techniques that had computed, tree bagged regression has the greater advantage to generate precise load prediction in short term basis.

4.4 Summary of Chapter

In this chapter, the findings of research study are discussed with the simulation result. First of all, the validation result for one day ahead load is presented in the comparison curve of forecasted load and actual load. Next, the relative error is tabulated and presented in graphical form to analyse the mean absolute percentage error (MAPE) and the mean absolute error (MAE) for all three techniques such as multiple linear regression, curve fitting and bagged tree regression. Besides, the performance evaluation with the diagnosis statistic is conducted with StdAPE and StdAE index.

CHAPTER 5: CONCLUSION

5.1 Conclusion

This project serves the purpose of analyse the short term load forecasting (STLF) techniques and evaluate the performance of respective technique for one day ahead load prediction. The project took one year historical load data as the training data for load forecasting simulation.

First, the load data from Global Energy Forecasting Competition 2012 is extracted and sampled for left-one out cross validation. These data are then stored in excel file in hour basis as time interval to analyse the daily load profile. For three numerical STLF techniques such as multiple linear regression, curve fitting and bagged tree regression, the respective algorithms are modelled in MATLAB Toolbox with several predictors selected. With this, the objective of STLF technique demonstration had been achieved.

Furthermore, the simulation result that had been obtained through cross validation comprises of a day ahead forecasted load curve and the relative error between the prediction result and actual load over the time of observation. Evaluating the result allows us to study the characteristic of the STLF techniques and how precise the simulation is done. The key index of MAPE and MAE which determine the error had been generated through simulation. As such, the analysis of each STLF technique had been conducted as Objective 2.

Last but not least, the comparison findings of numerical STLF techniques' performance have been analyzed with the statistics results of StdAPE and StdAE. The strength and weakness of each STLF technique had been discussed for a specified period of time ahead. It had concluded that with the same historical data as the training parameter, tree bagged regression had deliberately reflected the best-fit forecasted load result with the lowest error made. Objective 3 which aims to identify the best common practice numerical short-term load forecasting technique has been achieved.

In short, all objectives which had been stated in Chapter 1 have been achieved to address the problem statements of this research project.

5.2 Contribution of Research

a) **Discovery that the bagged tree regression technique led to optimal advancement for one-day-ahead load prediction.**

This project had modelled the bagged tree regression which is known as one of the most popular tree-based ensemble machine learning method. The combination of several weak learner into a decision tree with large leaf size permutate the training data for decision making. The simulation result of proposed predictive ensemble bagged tree regression method indicates it has the good performance to reduce the variance of error. The comparison results revealed that this suggested method could significantly increase the forecast accuracy and reliability in forecasting daily non-linear load profile.

b) **Discovery that the curve fitting technique is highly dependent on the regression coefficient obtained from basic linear regression method.**

This project had studied the capability of curve fitting technique for short term load forecasting with modelling in MATLAB Toolbox. Even though the interpretation of prediction method differs to the multiple linear regression (MLR) technique, however, the initial predictors are derived in the same way with linear regression method therefore the similar regression coefficients are obtained at the beginning. The variation between curve fitting and MLR is only the prediction variables are not as straightforward to interpret as linear regression, but provide the best-fit to the specified curve type in our modelling.

5.3 Recommendations

This research project had concluded that the capability of multiple linear regression, curve fitting and bagged tree regression to produce one day ahead forecasting with one year historical load data. Moreover, the consideration factors for data training in this research project are limited to the previous load samples and time of observation. For further research purpose, the modelling with temperature and weather condition factors can be conducted to explore more reliable and comprehensive result.

Furthermore, the project considered only the conventional load profile throughout the static period of time. Future study for renewable energy environment and advanced electric load can be conducted to analyze the capability of STLF techniques to produce accurate prediction.

REFERENCES

A Farahat, M., & Talaat, M. (2012). *A New Approach for Short-Term Load Forecasting Using Curve Fitting Prediction Optimized by Genetic Algorithms* (Vol. 2).

Abbas, S. R., & Arif, M. (2006, 23-24 Dec. 2006). *Electric Load Forecasting Using Support Vector Machines Optimized by Genetic Algorithm.* Paper presented at the 2006 IEEE International Multitopic Conference.

Aman, S., Simmhan, Y., & Prasanna, V. K. (2015). Holistic Measures for Evaluating Prediction Models in Smart Grids. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 475-488. doi:10.1109/TKDE.2014.2327022

Amral, N., Ozveren, C. S., & King, D. (2007, 4-6 Sept. 2007). *Short term load forecasting using Multiple Linear Regression.* Paper presented at the 2007 42nd International Universities Power Engineering Conference.

Cocianu, C. (2013). *Kernel-Based Methods for Learning Non-Linear SVM* (Vol. 47).

Debnath, K. B., & Mourshed, M. (2018). Forecasting methods in energy planning models. *Renewable and Sustainable Energy Reviews, 88*, 297-325. doi:https://doi.org/10.1016/j.rser.2018.02.002

del Carmen Ruiz-Abellón, M., Gabaldón, A., & Guillamón, A. (2018). *Load Forecasting for a Campus University Using Ensemble Methods Based on Regression Trees* (Vol. 11).

Dudek, G. (2015). Short-Term Load Forecasting Using Random Forests. In (Vol. 323, pp. 821-828).

Fengxia, Z., & Shouming, Z. (2011, 8-10 Aug. 2011). *Time series forecasting using an ensemble model incorporating ARIMA and ANN based on combined objectives.* Paper presented at the 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC).

Ferrera, E., Hu, X., Tomasi, R., & Pastrone, C. (2014). *Evaluation-of-Short-Term-Load-Forecasting-Techniques-Applied-for-Smart-Micro-Grids* (Vol. 8).

García-Ascanio, C., & Maté, C. (2010). Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy, 38*(2), 715-725. doi:https://doi.org/10.1016/j.enpol.2009.10.007

Hahn, H., Meyer-Nieberg, S., & Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research, 199*(3), 902-907. doi:https://doi.org/10.1016/j.ejor.2009.01.062

Hambali, M., Akinyemi, Oladunjoye, M., & N, Y. (2017). *Electric Power Load Forecast Using Decision Tree Algorithms* (Vol. 7).

Han, H.-G., Chen, Q., & Qiao, J. (2010). *Research on an online self-organizing radial basis function neural network* (Vol. 19).

Hong, T., Gui, M., Baran, M. E., & Willis, H. L. (2010, 25-29 July 2010). *Modeling and forecasting hourly electric load by multiple linear regression with interactions.* Paper presented at the IEEE PES General Meeting.

IEEE Power & Energy Society, K. (2012). Global Energy Forecasting Competition 2012 - Load Forecasting. Retrieved from https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting

Jain, M. B., Nigam, M. K., & Tiwari, P. C. (2012, 30 Oct.-2 Nov. 2012). *Curve fitting and regression line method based seasonal short term load forecasting.* Paper presented at the 2012 World Congress on Information and Communication Technologies.

Jingfei, Y., & Stenzel, J. (2005, 29 Nov.-2 Dec. 2005). *Historical load curve correction for short-term load forecasting.* Paper presented at the 2005 International Power Engineering Conference.

José Montaño Moreno, J., Palmer, A., & Muñoz Gracia, P. (2011). *Artificial neural networks applied to forecasting time series* (Vol. 23).

Kafazi, I. E., Bannari, R., Abouabdellah, A., Aboutafail, M. O., & Guerrero, J. M. (2017, 4-7 Dec. 2017). *Energy Production: A Comparison of Forecasting Methods using the Polynomial Curve Fitting and Linear Regression.* Paper presented at the 2017 International Renewable and Sustainable Energy Conference (IRSEC).

Khair, U., Fahmi, H., Hakim, S. A., & Rahim, R. (2017). Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. *Journal of Physics: Conference Series, 930*, 012002. doi:10.1088/1742-6596/930/1/012002

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669-679. doi:https://doi.org/10.1016/j.ijforecast.2015.12.003

Kumar, S., Mishra, S., & Gupta, S. (2016, 12-13 Feb. 2016). *Short Term Load Forecasting Using ANN and Multiple Linear Regression.* Paper presented at the 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT).

Kuster, C., Rezgui, Y., & Mourshed, M. (2017). *Electrical load forecasting models: A critical systematic review* (Vol. 35).

Lahouar, A., & Slama, J. B. H. (2015, 24-26 March 2015). *Random forests model for one day ahead load forecasting.* Paper presented at the IREC2015 The Sixth International Renewable Energy Congress.

Lajevardy, P., Parand, F.-A., Rashidi, H., & Rahimi, H. (2015). *A HYBRID METHOD FOR LOAD FORECASTING IN SMART GRID BASED ON NEURAL NETWORKS AND CUCKOO SEARCH OPTIMIZATION APPROACH* (Vol. 5).

Lee, K. Y., Cha, Y. T., & Park, J. H. (1992). *Short-term Load Forecasting Using an Artificial Neural Network* (Vol. 7).

Lee, W., Jung, J., & Lee, M. (2017, 16-20 July 2017). *Development of 24-hour optimal scheduling algorithm for energy storage system using load forecasting and renewable energy forecasting.* Paper presented at the 2017 IEEE Power & Energy Society General Meeting.

Mat Daut, M. A., Hassan, M. Y., Abdullah, H., Rahman, H. A., Abdullah, M. P., & Hussin, F. (2017). Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. *Renewable and Sustainable Energy Reviews, 70*, 1108-1118. doi:https://doi.org/10.1016/j.rser.2016.12.015

Molugaram, K., & Rao, G. S. (2017). Chapter 5 - Curve Fitting. In K. Molugaram & G. S. Rao (Eds.), *Statistical Techniques for Transportation Engineering* (pp. 281-292): Butterworth-Heinemann.

Nie, H., Liu, G., Liu, X., & Wang, Y. (2012). Hybrid of ARIMA and SVMs for Short-Term Load Forecasting. *Energy Procedia, 16*, 1455-1460. doi:https://doi.org/10.1016/j.egypro.2012.01.229

Pahasa, J., & Theera-Umpon, N. (2007, 3-6 Dec. 2007). *Short-term load forecasting using wavelet transform and support vector machines.* Paper presented at the 2007 International Power Engineering Conference (IPEC 2007).

Papadopoulos, V., Delerue, T., Ryckeghem, J. V., & Desmet, J. (2017, 28-31 Aug. 2017). *Assessing the impact of load forecasting accuracy on battery dispatching strategies with respect to Peak Shaving and Time-of-Use (TOU) applications for industrial consumers.* Paper presented at the 2017 52nd International Universities Power Engineering Conference (UPEC).

Patel, H., Pandya, M., & Aware, M. (2015, 26-28 Nov. 2015). *Short term load forecasting of Indian system using linear regression and artificial neural network.* Paper presented at the 2015 5th Nirma University International Conference on Engineering (NUiCONE).

Reddy Cheepati, K., & Nageswara Prasad, T. (2016). *Performance Comparison of Short Term Load Forecasting Techniques* (Vol. 9).

Singh, A. K., Ibraheem, K. S., & Muazzam, M. (2013). *An overview of electricity demand forecasting techniques* (Vol. 3).

Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews, 16*(2), 1223-1240. doi:https://doi.org/10.1016/j.rser.2011.08.014

Tealab, A., Hefny, H., & Badr, A. (2017). Forecasting of nonlinear time series using ANN. *Future Computing and Informatics Journal, 2*(1), 39-47. doi:https://doi.org/10.1016/j.fcij.2017.05.001

Vazquez, R., Amaris, H., Alonso, M., López, G., Moreno, J., Olmeda, D., & Coca, J. (2017). *Assessment of an Adaptive Load Forecasting Methodology in a Smart Grid Demonstration Project* (Vol. 10).

Wang, F., He, T., & Nie, H. (2017). *Power load prediction based on multiple linear regression model* (Vol. 55).

Wen, Z., Li, Y., Tan, Y., Cao, Y., & Tian, S. (2015, 15-18 Nov. 2015). *A combined forecasting method for renewable generations and loads in power systems.* Paper presented at the 2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC).

Widodo, & Fitriatien, S. (2016). *ARTIFICIAL NEURAL NETWORK FOR ELECTRIC LOAD FORECASTING.*

Willis, H. L., Powell, R. W., & Wall, D. L. (1984). Load Transfer Coupling Regression Curve Fitting for Distribution Load Forecasting. *IEEE Power Engineering Review, PER-4*(5), 42-42. doi:10.1109/MPER.1984.5526044

Zhang, Z., Li, C., Cao, Y., Tang, L., Li, J., & Wu, B. (2012, 21-24 May 2012). *Credibility assessment of short-term load forecast in power system.* Paper presented at the IEEE PES Innovative Smart Grid Technologies.