CHAPTER 3

METHODOLOGY

DEVELOPMENT OF NNS AND HYBRID SARIMA-NN MODELS

3.1 INTRODUCTION

This chapter discusses the extraction of TEC data from a single GSV4004B GPS Ionospheric Scintillation and TEC Monitor (GISTM) receiver station at Parit Raja, Malaysia. A brief description on the mapping technique that is used to convert the slant TEC into vertical TEC is explained. Continuously, the chapter provides a detailed description on the Neural Network (NN) and Hybrid Seasonal Autoregressive Integrated Moving Average-Neural Network (Hybrid SARIMA-NN) techniques that are used to estimate and forecast the ionospheric TEC, respectively. The first method in this chapter, describes the development of TEC estimation model based on NN. It includes the background material towards the development of a TEC estimation model (i.e. the inputs and output parameters, the architecture and the algorithms used to develop the NN model) over Parit Raja station.

The second method, explains the development of a hybrid model to forecast the ionospheric TEC. In this section, a linear SARIMA technique is hybridised with a non-linear NN technique to forecast the TEC values in advance. The pre-processing approach in the hybrid SARIMA- NN model is demonstrated and outline.

3.2 TEC DATA AT WIRELESS AND RADIO SCIENCE CENTRE (WARAS), PARIT RAJA STATION

In this study, the GSV4004B GPS Ionospheric Scintillation and TEC Monitor (GISTM) receiver is used to collect the ionospheric data. The receiver is able to track up to 11 GPS signals at the L1 (1575.42 MHz) and L2 (1227.60 MHz) frequencies. The primary purposes of this receiver are to measure the phase and amplitude scintillation at 50 Hz rate from the LI frequency and to compute the GPS TEC measurement from the combined of L1 and L2 pseudo-range and carrier phase measurements (GSV4004B, 2004). The equipment computes and provides four pairs of TEC and rate of TEC (ROT) for every 15s. Therefore, this receiver is well suited to be used by the science community to study the TEC and TEC rate. Figure 3.1 shows the process of extracting the ionospheric TEC data from the GISTM receiver. An offline utility program, Parseismr.exe is used to extract and convert the binary format data in the GSV4004B receiver to a comma-delimited format. The data in the comma-delimited format is imported in Microsoft Excel for further editing.

In GSV4004B, the TEC value (in TECU, where 1 TECU = 1 x 10^{16} electrons m^{-2}) is determined as in the Equation (3.1) (GSV4004B, 2004; Zain et al., 2005):

$$TEC = \left[9.483 * \left(P_{L2} - P_{L1} - \Delta_{C/A-P,PRN}\right) + TEC_{Cal}\right]TECU \quad (3.1)$$

where P_{L2} is the L2 pseudo-range in meter, P_{L1} is the L1 pseudo-range in meter, $\Delta_{C/A-P,PRN}$ is the input bias between satellite C/A- and P-code chip transitions in meter, and TEC_{Cal} is the TEC due to internal receiver L1/L2 delay and offset (TECU).

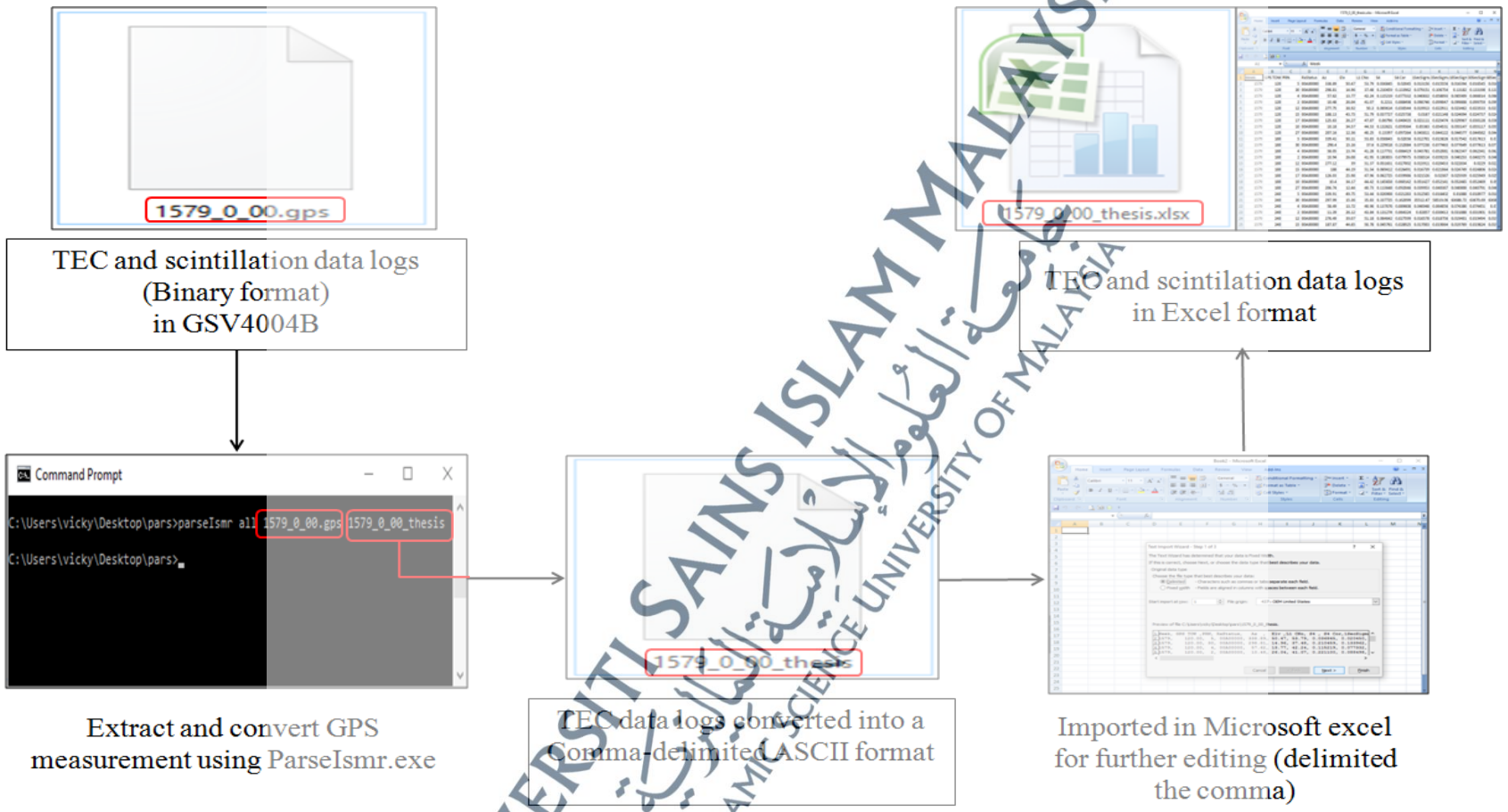


Figure 3.1: Process of Extracting Ionospheric TEC data from the GISTM receiver at Parit Raja station, Malaysia

Accurate and precise TEC value is very important in the ionospheric study, since the parameter is able to describe the Earth's ionosphere behaviours and anomalies. The slant TEC measurements obtained from the GISTM are a sum of real slant TEC, GPS satellite bias and receiver bias (Karia & Pathak, 2011). This derived slant TEC data is corrupted by the instrumental biases of the GPS satellites and the receiver. Therefore, it is essential to remove the satellite-receiver instrumental biases in TEC estimation, because incorrect estimation and biasing measurements (uncalibrated values) may influence the end-use results and lead to false interpretations. The combined satellite and receiver biases can even lead to a negative TEC, where negative TEC values are not physically reasonable. The satellite-receiver instrumental biases are caused by the relative travel times between the L1 and L2 signals from the GPS antenna to the master CPU. Since these biases are frequency dependent, they could not be removed by subtracting different frequency observations (Kao et al., 2013). Thus, these biases must be predicted accurately and removed from the GPS measurements to achieve a reliable TEC data (Sardon & Zarraoa, 1994).

The TEC value in the above Equation (3.1) is corrected for satellite inter-frequency biases but not for the satellite C/A-to-P biases (GSV4004B, 2004; Gwal & Jain, 2011). As stated before, the TEC values are measured on L1 and L2 frequencies, where the C/A code measurements are on the L1 frequency, while the P code measurements are on the L2. Thus, each individual satellite introduces a time delay during the transitions between the C/A- and P- code, which eventually will cause a bias in the recorded TEC values. In order to obtain accurate TEC values, C/A- to P- satellite biases (ns) are added during the hardware initialization to remove the biases or alternatively be corrected during the post-processing by adding these biases (converted in

TECU) to the logged TEC values (GSV4004B, 2004). In this study, these biases are obtained from the file published by the University of Bern at ftp://ftp.unibe.ch/aiub/CODE.

Besides the satellite biases, the other important factor that is known to affect the accuracy of the TEC values is the inter-frequency bias inherited in the GPS receiver. There exists an instrumental delay bias in each signal of the two GPS frequencies. The bias value can be consistent over a few days during the undisturbed conditions and smooth ionospheric behaviours, for instance within the mid-latitude regions. However, in certain regions such as equatorial and auroral zones, the value may vary due to complex and large unpredictable ionosphere variations over these regions (Sardon & Zarraoa, 1994; Manucci et al., 1999). There are various techniques to predict the GPS receiver bias, however, most of the techniques are based on a highly dense GPS receiver network (Mannucci et al., 1998; Otsuka et al., 2002; Ma & Maruyama, 2003; Nayir et al., 2007; Sunehra, 2013). In this study, the calibration technique proposed by Carrano & Groves (2006) for the equatorial region is adopted to predict the inter-frequency bias for a single station receiver based on a minimization process of the TEC variability. In this approach, the bias values are predicted on a daily basis using the TEC data attained before the sunrise, which is between 0300 - 0600 LT when the temporal gradients are generally small. This specific period of time is taken into consideration, since the TEC variability is found to be invariant late at night. Other than that, the vertical TEC (vTEC) from each satellite is assumed to be the same (less scattered) when the biases are correctly removed.

Using the relative sTEC, satellite bias and single layer mapping function (SLM), the vTEC can be defined as a function of receiver bias. The SLM function is used to

convert the slant TEC to vertical TEC and the mapping method is described thoroughly in the following subsection with aid of the diagram. In this study, the height of the SLM is 350km and to minimise the multipath errors, an elevation angle above 20 degrees is used.

A series of trial receiver bias is used independently to calculate the vTEC and the variance of TECs to their mean at each observation time, i. The total variance, $Var(bR)$ of the computed vTEC values are summed over bins (i) of the local time (0300 - 0600 LT) as in Equation (3.2)

$$Var(bR) = \sum Variance[vTEC(bR)]_i \tag{3.2}$$

where vTEC is the vertical TEC, bR is the receiver bias assumed independently, and $[\,]_i$ is the data computed during i^{th} local time bin.

The primary purpose of binning is to reduce the contribution of small temporal gradients in TEC, which may impact the total variance and affect the accuracy of the bias determination.

The best predicted receiver's bias for each day is the value of bR that minimizes the total variance of vTEC in Equation (3.2) between 0300 - 0600 LT. Finally, the predicted receiver bias is computed everyday and smoothed over 14 days to obtain a stable and reliable calibrated TEC, TEC_{Cal} for that day. In certain scenarios, overestimation of the bR value tends to lead a negative vTEC result. Since negative TEC values are not physically reasonable, the bias value is reassumed to enforce that the min(TEC) value is equal to zero. Thus, a well calibrated TEC shows a non-negative TEC

value and the TEC curve from each satellite is very similar and collapse well at night. The predicted TEC_{Cal} values for 2005 and 2006 are tabulated in APPENDIX A.

3.2.1 Single Layer Model and mapping function

The TEC computed from the post-processing (as in the above method) is called as slant TEC (sTEC). It measures the number of free electrons contained in a column with a cross-sectional area of $1m^2$ along the ray path from satellite to the receiver on the Earth's surface through the ionosphere (Todorova et al., 2008). As the slant TEC is dependent on the elevation of the ray path, an equivalent vertical TEC (vTEC) value is required to remove the dependency from the elevation angle. A single layer thin-shell model (SLM) is adopted (Schaer, 1999) to approximate the absolute vertical TEC. The geometry of the SLM is depicted in Figure 3.2.

SLM assumes that all free electrons are concentrated in a thin layer at about 300 to 500 km height (hm) above the Earth's surface. The intersection point between the transmitted GPS signal and the ionosphere shell is known as ionospheric pierce point (IPP). Furthermore, the zenith angle at that ionospheric pierce point is z' and the zenith angle at the receiver location is z. Generally in this thesis, an elevation dependant mapping function (Schaer, 1999), is used to obtain the absolute vertical TEC by considering the zenith angle of the satellites and the expression can be expressed as in Equation (3.3):

$$F(z) = \frac{sTEC}{vTEC} = \frac{1}{\cos z'} = \frac{1}{\sqrt{1 - sin^2 z'}} \tag{3.3}$$

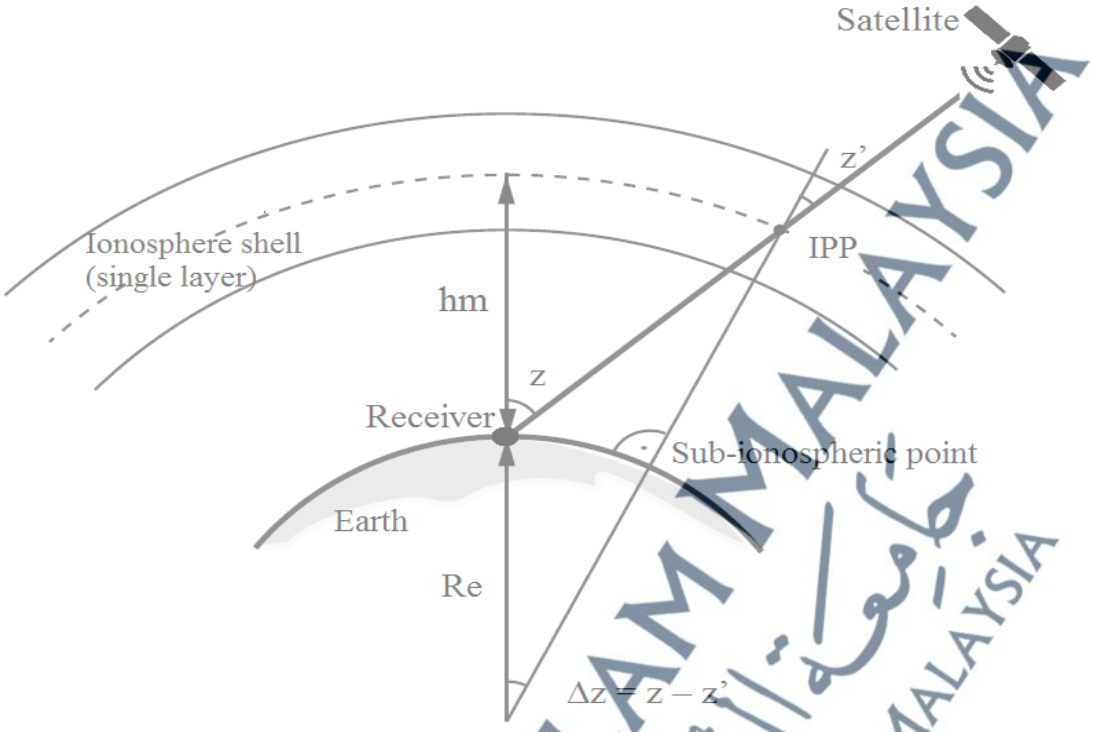


Figure 3.2: Ionospheric Single Layer thin-shell Model (SLM) (Modified from Schaer, 1999)

From Figure 3.2 the relation between z and z' can be derived using the law of sines:

$$\sin z' = \frac{Re}{Re + hm} \sin(z) \tag{3.4}$$

where z, z' represent the zenith angles at the receiver location and at the ionospheric pierce point, respectively, Re is the mean radius of Earth (≈6370 km), and hm denotes the height of the single layer model in km. In this thesis, the slant TEC estimated from GISTM observations is converted to vertical TEC by assuming the ionosphere to be a single layer (ionosphere shell) of fixed height of 350 km above the Earth's surface. Equation (3.4) is substituted into Equation (3.3) to estimate the absolute vTEC value. These estimated and calibrated TEC values are utilised in ionospheric TEC modelling.

3.3 DATA PROCESSING FOR IONOSPHERIC TEC MODELLING

Figure 3.3 shows the flow diagram of the overall methods used for ionospheric TEC modelling in this thesis. The diagram is divided into two sections; estimation and forecasting. The two main models are thoroughly explained with proper illustrations throughout this chapter.

Missing and erroneous data are always a crucial problem that can lead to false conclusions and affect the post processing analysis. The estimation process in the first section offers an opportunity to estimate and fill up the gaps of the lost TEC data over Parit Raja based on a neural network (NN) technique. This method has become a well known tool to model the complex non-linear processes due to its non- linearity property (Haykin, 1999). The incorporation of prior knowledge into NN can help the model to learn, store the information and finally can be used for generalization. Since the prior knowledge is very important in NN modelling, in this chapter the parameters known to influence the ionospheric TEC variability during the medium solar cycle are identified. These parameters are used as the input spaces in the TEC based NN model to estimate accurate TEC data. Furthermore, the architecture of NN for TEC modelling is determined in order to reduce the training time of the NN model. A short term hourly vertical TEC dataset are used to train the NN model and the performance of the model to estimate the ionospheric TEC over a single base station is investigated.

In the following part, forecasting the ionospheric TEC ahead for a short term period is designed. In order to develop a reliable forecasting model, a valid TEC data is required as an input parameter. Hence, the observed TEC and the estimated TEC (NN

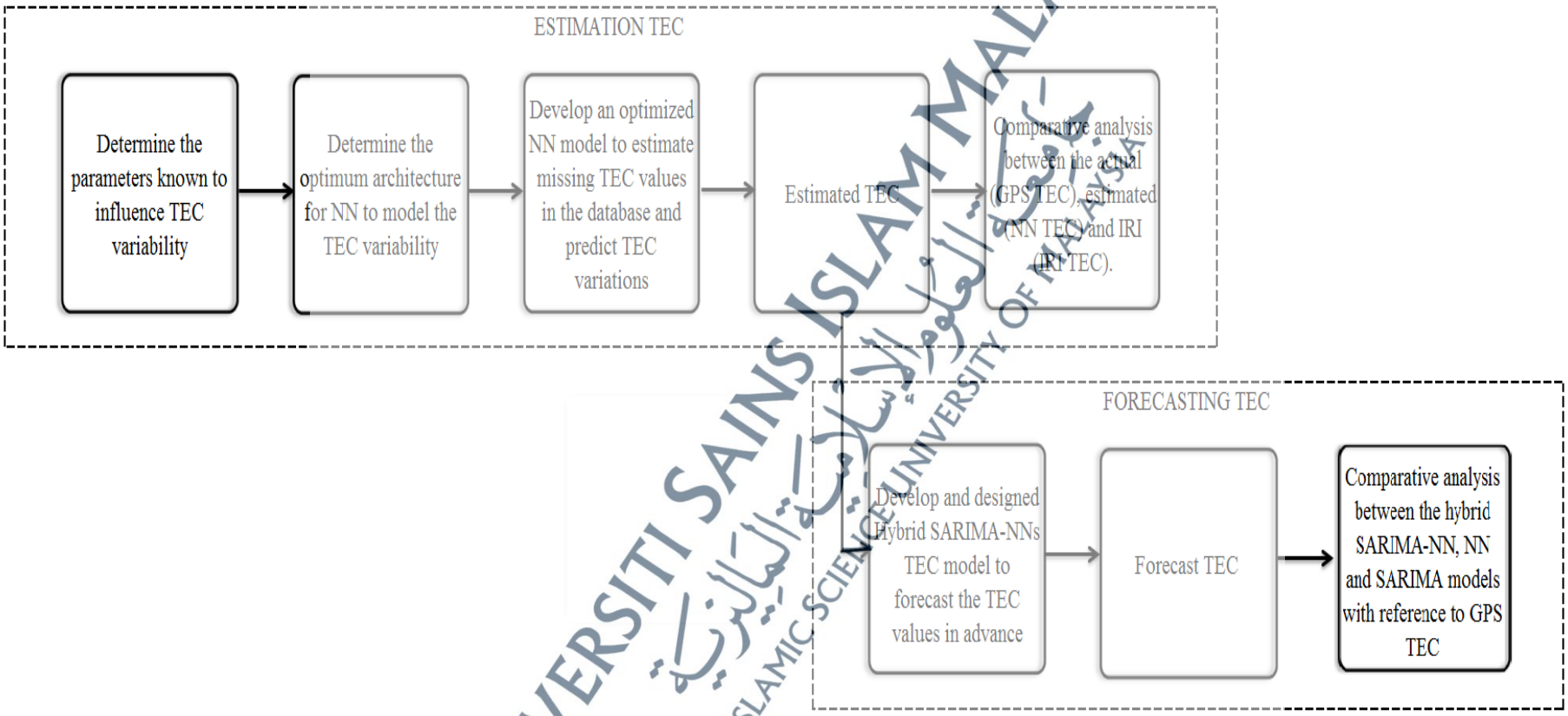


Figure 3.3: Flow diagram for TEC modelling using NN and hybrid SARIMA-NN models

TEC) values are used together (considered no missing or errors in the dataset) to develop a single-station TEC short-term forecast model based on a hybrid SARIMA-NN technique. The hourly TEC data during a medium solar activity period from February 2005 - September 2006 are used for training while the TEC values for the last three months of the year 2006 are used for validation and testing purposes. To evaluate the performance of hybrid model, the model is compared and validated against the forecast values of both the individual models; SARIMA referred as FCAST-SARIMA and NN denoted as FCAST-NN separately. The comparison results are shown and discussed in the results chapter.

3.4 DEVELOPMENT OF NEURAL NETWORK-BASED TEC ESTIMATION MODEL

NN is very adaptive in nature, where the network's generalization capability remains precise and robust even in a stochastic environment whose characteristics may change over time. Besides, being a data driven model, the NN impose few prior assumptions on the underlying process from which data are generated. On account of this property, NN is less prone to model misspecification compared to other non-linear techniques. In addition, the NN has flexible non-linear function mapping capability, which can estimate any non-linear measurable function with arbitrarily desired accuracy (Khashei & Bijari, 2011). Owing to all these properties, the NN technique has become an attractive and well known tool in many non-linear applications, for instance in ionospheric research (Hernandez-Pajares et al., 1997; Conway et al., 1998; Poole & McKinnell, 2000; Sutcliffe, 2000; Leandro & Santos, 2004; Oyeyemia et al., 2006; Habarulema et al., 2007; Mckinnell, 2008; Jean et al., 2009; Acharya et al., 2010).

Knowing the advantages of NN in complex processes, this technique has been adopted and designed to estimate the non-linear ionospheric TEC. In this thesis, the NN is developed using MATLAB version 7.11.0.584 (R2010b) with 64-bit (win 64) is installed in a system which has Intel Pentium 987 processor with the speed of 1.5 GHz and 4 GB memory under Windows7 operating system.

Figure 3.4 depicts the flow diagram that describes the overall process of developing an ionospheric TEC model based on NN. The development process can be divided into three stages: stage 1: Identify the input(s) that highly contributes or influences the ionospheric TEC data over Malaysia. The input data should comprise the full range of the input space so that the designated model will be more reliable, stage 2: determine the number of hidden neurons in the hidden layer of the NN and stage 3: identify the best training algorithm to train the NN for ionospheric TEC estimation. All the three stages are explained thoroughly in subsections 3.4.1, 3.4.2 and 3.4.3, respectively.

The selection of input parameters, the number and arrangement of the neurons in each layer, the algorithm used to train the network and the interconnection links within and between the layers are the features that distinguish the architectures of a NN. These features contribute to the achievement of a trained network. The predictive ability of the NN technique may vary with respect to those features. However, according to Maruyama (2010) and Ghaffari et al. (2006), the initial weights and biases even contribute on the achievement of NN training. In this study, both the weights and biases of the neurons are initialized by random numbers in the NN training, where the values may vary between -1 and 1. To overcome the influence of random initialization in the NN training during

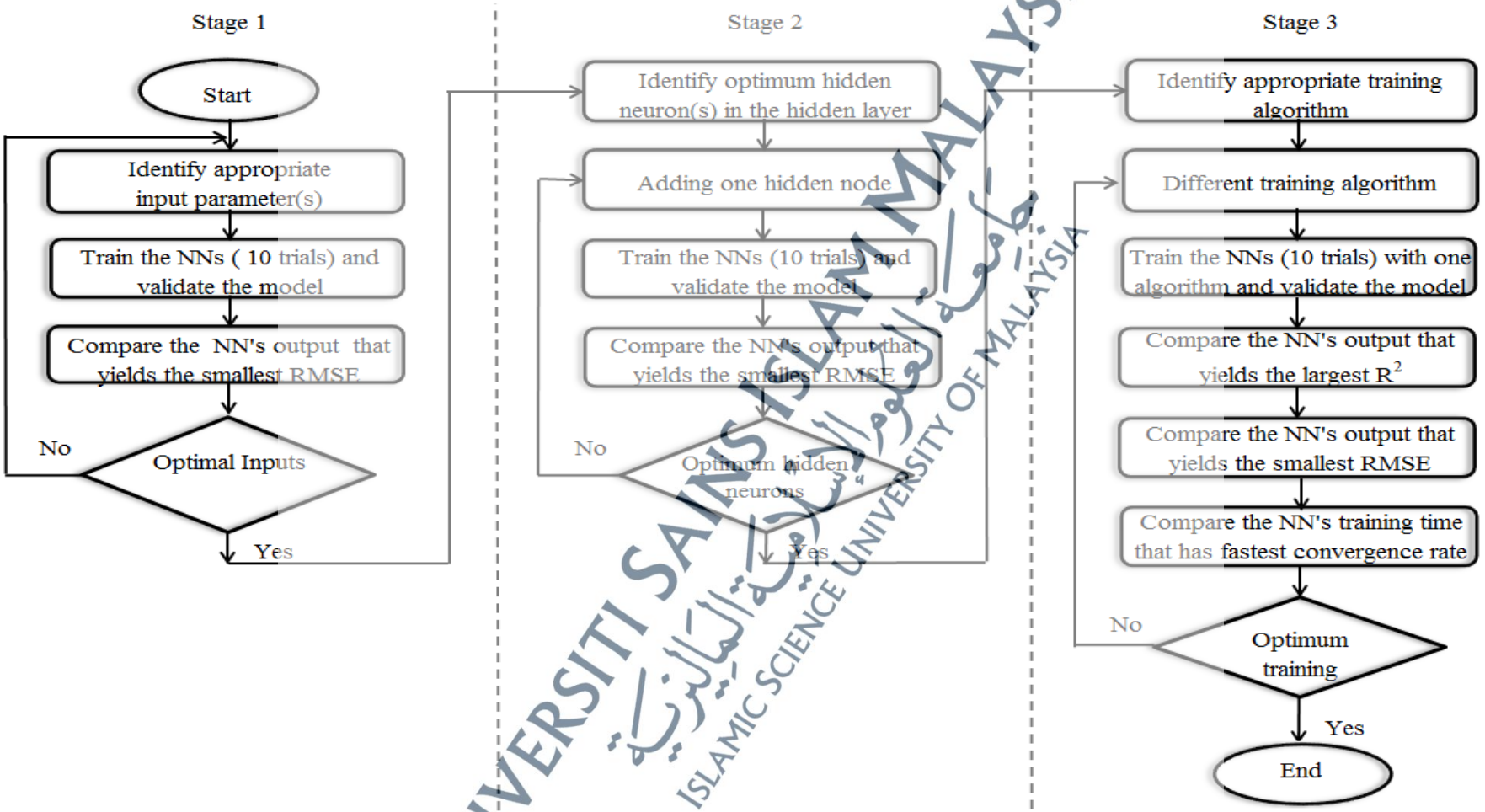


Figure 3.4: Flow diagram for development of ionospheric TEC estimation model based on Neural Network technique

the determination of input parameter(s), hidden neuron(s) and training algorithm, each individual network in the training runs ten trials and the best results are used for comparison and modelling. Finally, the developed NN model based on these three criteria is trained and used to estimate the ionospheric TEC values.

3.4.1 Determination of optimum input space

The production of plasma in the ionosphere is strongly governed by the activity of the sun and the magnetic perturbations. To provide specific information on the population of electron concentration in the ionosphere during the medium solar activity, a quantitative study is carried out to obtain the optimum parameter(s) to represent solar and magnetic activities that are known to influence the TEC variability. The identified parameters are used to develop a local TEC model for a single station at Parit Raja Malaysia (equatorial region). Based on the literature review, it is possible to deduce that not many works have reported on the indices that should be considered during ionospheric TEC modelling over Malaysia. In order to investigate the effective indices, a number of solar and magnetic proxies that control the ionospheric TEC variability are examined. Following are the solar and magnetic indices examined in this thesis:

i. **Solar indices**

Solar radiation is the primary factor that controls the TEC variations. The solar activity is directly proportional with the amount of ionizing radiations from the Sun. Several solar indices were introduced for different objectives to foresee the ionospheric conditions. However, in the empirical modelling of ionosphere system, a few indices have long been used to represent the solar activity such as sunspot number (SSN) and

solar radio flux at 10.7cm ($F_{10.7}$) wavelength. Both of the indices are measured on ground. Besides these indices, some solar indices are obtained from satellites measurements and rockets (e.g. solar X-rays, solar extreme ultraviolet (EUV)). These proxies are commonly absorbed before reaching the lower atmosphere, where the variability of the indices will begin to fade out with lowering of height. Thus, the ground-based instruments are not well-suited to monitor those proxies. Brief descriptions on the solar indices that have been considered in this study are described below.

a. Sunspot Number (SSN)

   Sunspot Number (SSN) is the longest data series among other solar indices. SSN was introduced by Johann Rudolph Wolf in 1848 to express the solar activity in numerical terms (Perrone & De Franceschi, 1998). Generally, SSN is not constructed through any measurement of radiation whereas it is quantified from sum of the numbers of sunspots and sunspots group on the solar disk. It is a daily index of sunspot activity which varies with solar cycle. The SSN index is acquired from the archives of National Geophysical data Centre (NGDC).

b. Solar Radio Flux ($F_{10.7}$)

   Along with the SSN, another most commonly used solar proxy is the solar radio flux, $F_{10.7}$. Tapping & DeTracey (1990) stated that the thermal ionization that originates from the photosphere and chromosphere interface with the magnetic resonance above the plages resulting $F_{10.7}$ index. It is closely related to the amount of ionization and hence the electron concentration in the F2 region. $F_{10.7}$ is the integrated emission from the

entire solar disk at a frequency of 2800 MHz (10.7 cm radio wavelength). The value was recorded routinely by radio telescope near Ottawa, Canada since the first observation in 1947, while in 1991 the observatory was relocated to Penticton, Canada. The physical unit of $F_{10.7}$ in solar flux unit (sfu) is (= $10^{-22}Wm^{-2}Hz^{-1}$) and the value is observed at 20UT (local time at Penticton). It varies with the 11-year solar cycle and is used almost interchangeably with the sunspot index.

c. Solar extreme ultraviolet (EUV) flux ($S_{10.7}$)

The Solar and Heliospheric Observatory (SOHO) was launched in late 1995. Solar EUV Monitor (SEM) and EUV Imaging Telescope (EIT) are two instruments abroad SOHO truly capable of measuring EUV irradiance. University of Southern California (USC) operates the SEM, where they continually monitors the solar extreme ultraviolet (EUV) fluxes in two wavelengths ranges, which are 26 - 34 nm and 0.1 - 50 nm. According to Maruyama (2007), the solar $EUV_{26-34nm}$ has high correlation and coincides with the wavelength of EUV that ionize the Earth's upper atmosphere. The value is normalized using a common time frame mean value of 1.9955×10^{10} photons cm^{-2} s^{-1} and generate a new index $S_{10.7}$ by converting the normalized value to sfu, as $F_{10.7}$ through linear regression (Tobiska et al., 2008).

ii. Magnetic indices

Even though solar proxies are highly correlated with TEC variations, there are some other factors that contribute to rapid and random fluctuations in the ionospheric

TEC values, for instance the space weather disturbances. The occurrence of geomagnetic storms and associated ionospheric storms are mainly due to the interference between the high energy charged particles and the earth's magnetosphere. The interference causes large perturbations in the earth's magnetic field which excite the ions and increase the electron densities in the ionosphere. Thus, magnetic activity indices were designed to describe variation in the geomagnetic field caused by these irregular current systems. In this study, two types of magnetic inputs that have been considered in TEC modelling are described below.

a. Equivalent three hourly Planetary amplitude (ap)

The planetary indices Kp and Ap are two common indices used to indicate the severity of geomagnetic activity and the disturbance to the ionosphere. The non-linear relationship of the K-scale to magnetometer fluctuations and the difficulty in averaging the K-scale which is expressed in "quasi logarithmic, an equivalent "a"-index has been derived from the K-index. The ap index is a 3-hourly "equivalent amplitude" index ranges from 0 - 400 was introduced to obtain a linear index from Kp in nanoTesla (nT). The 8 ap index values are averaged to designate the daily planetary index Ap, represents the degree of global geomagnetic variability of each day. The ap index is retrieved from the archives of World Data Centre (WDC) for Geomagnetism, Kyoto (http://wdc.kugi.kyoto-u.ac.jp/).

b. Disturbance Storm Time (Dst)

The enhancement of ring current intensity and energy which leads to a significant depression in the Dst index is another method to define the

severity of geomagnetic storm. Disturbance storm time (Dst) was introduced and the average behaviour was established by Sugiura (1964). The Dst index is defined as the average value of the horizontal magnetic field component (H) at four observatories near the magnetic equator. The index is calculated at hourly basis and expressed in nT. The Dst values drop to lower negative as the intensity of the magnetic storm increases. It aims to measure the strength of the equatorial ring current. This hourly Dst values were obtained from the National Geophysical Data Centre (NGDC) at their anonymous FTP site: (ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/DST)

Besides these proxies, two external indices; namely day number and hour are considered as a constant parameter (Cp) in this study. The day number (DN) represents as seasonal variations, while the hour (HR) denotes as diurnal variations. Since Parit Raja station lies in the equatorial anomaly, the region is more subjected to large temporal variability. The diurnal and seasonal variations of ionosphere are considered as dominant factors in specifying the TEC variations at this region. Therefore the HR and DN are included as the input spaces in the TEC modelling based NN. The DN and HR are split into trigonometrical components to allow the data continuity and consistency (Poole & McKinnell, 2000) as in Equation (3.5) and Equation (3.6):

$$DNs = \sin\left(\frac{2\pi \times DN}{365.25}\right), DNc = \cos\left(\frac{2\pi \times DN}{365.25}\right) \quad (3.5)$$

$$HRs = \sin\left(\frac{2\pi \times HR}{24}\right), HRc = \cos\left(\frac{2\pi \times HR}{24}\right) \quad (3.6)$$

where the DNs, DNc, HRs and HRc are the sine and cosine components of the day number (DN) and hour (HR), respectively. In the above Equation (3.5), the factor 0.25 is taken into account if there are data from leap years.

For this training, fifteen months (February to April 2006) TEC data are used to train the network while two months data (July 2005 and May 2006) are used to validate the network to obtain the best model among the considered models. The validation data is excluded during the training session to ensure the capability of the NN technique to generalize from the training set and not to memorize the dataset. Root mean square error (RMSE) is used to monitor the performance of the network and to determine the optimum parameters that capable to estimate the TEC with least bias. This method has been used in many areas as a means to determine the optimum parameters in NN estimation (Jean et al., 2009; Oyeyemia et al., 2006; Mckinnell, 2008). The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(NN\,TEC - GPS\,TEC)^2} \tag{3.7}$$

where $NN\,TEC$ is TEC estimated by the neural network, $GPS\,TEC$ is the vertical TEC retrieved from the GPS at Parit Raja station and N is the number of the data sets.

3.4.1.1 Optimum solar proxies

All three solar inputs that describe the TEC variability are investigated using NN model. The solar proxies represented as the input parameters in the NN model to

determine the ionospheric TEC. As mentioned above, the solar proxies are originated at different time and locations around the globe. Thus, the behaviour of these indices varies with short-term and long term variations. According to Tobiska et al. (2008), the long-term and short-term variations of solar indices are desirable and should be taken into consideration during the modelling. The application of three solar proxies with various mean periods is introduced as the input spaces for the NN model in this study. The daily values of sunspot number (SSN), solar radio flux ($F_{10.7}$) and solar extreme ultraviolet flux ($S_{10.7}$) are used in the NN model. Besides the daily values, the indices are smoothed into two different schemes, centered and backward, over two types of time frames; 27- and 81- days. Here, the 27- and 81- days represent one and three solar rotation(s) period, respectively. The periods are shown in Figure 3.5. The backward and centered means are used to smooth these periods because Maruyama (2010) found a delayed response in the ionospheric variations to solar irradiance changes. Thus, it is important and necessary to consider the time lag of the ionospheric TEC variations which correspond to the solar proxies in the TEC modelling.

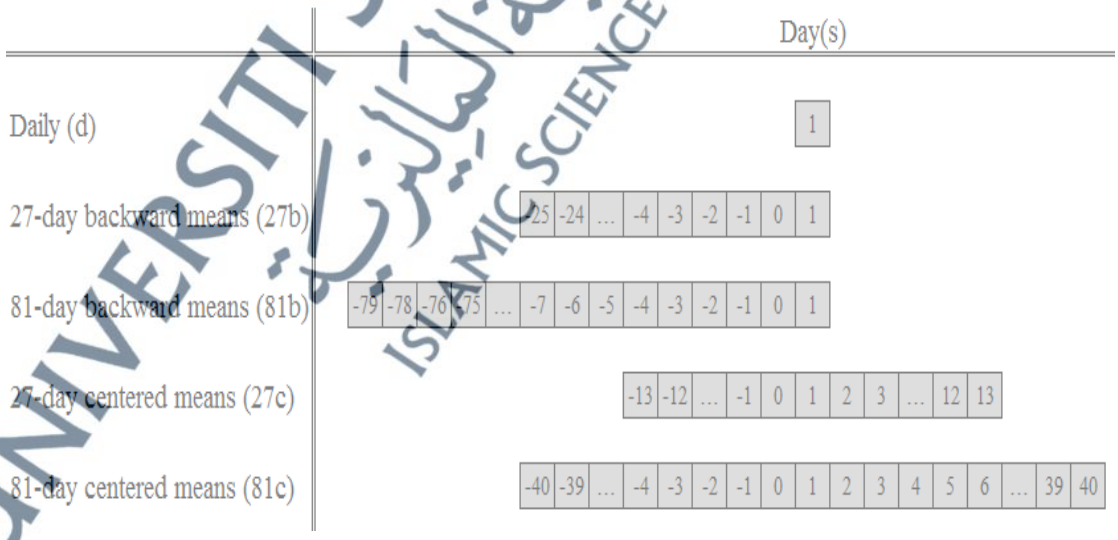


Figure 3.5: Proxies are averaged in different smoothing schemes

There are 15 solar inputs (5 variables/periods for 3 solar proxies) in this study. Figure 3.6 (a), (b) and (c) illustrate the daily, 27 and 81 backward means for SSN, respectively while Figure 3.7 (a), (b) and (c) show the daily, 27 and 81 centered means for SSN, respectively. The smoothing periods for other proxies; $F_{10.7}$ and $S_{10.7}$ are shown in the APPENDIX B. In this study, 13 combinations of period patterns for each solar proxy along with the constant parameter (Cp) are included as the input parameter in the NN model to identify the appropriate proxy that influence TEC.

Figure 3.8 shows the comparison between daily, 27- and 81-days of centered and backward means patterns for SSN, $F_{10.7}$ and $S_{10.7}$. The horizontal axis represents the period and time lag patterns. The plus sign "+" in each pattern describes the concurrent use of those values in NN training. The letters d, b and c along with the periods denoted daily, backward and centered mean, respectively. The predictability of each pattern is assessed by comparing the deviations between the observed (GPS TEC) and estimated (NN TEC) values. The pattern that provides the least RMSE value is considered to be the best pattern for TEC estimation.

In general, the proxies with more than one period produced smaller RMSE than proxies with single period. Neither the centered nor the backward means of a single period have significant effect on the TEC estimation. Furthermore, the NN estimation results deteriorate when only 27- or 81-days with centered or backward means used as the input space in the NN training. For SSN, $F_{10.7}$ and $S_{10.7}$, the combination of daily and 27-day centered mean yielded the least RMSE values. Among the proxies, SSN gives an improvement in TEC estimation when daily and 27c (centered mean) are used concurrently compared to the other patterns in the NN models.

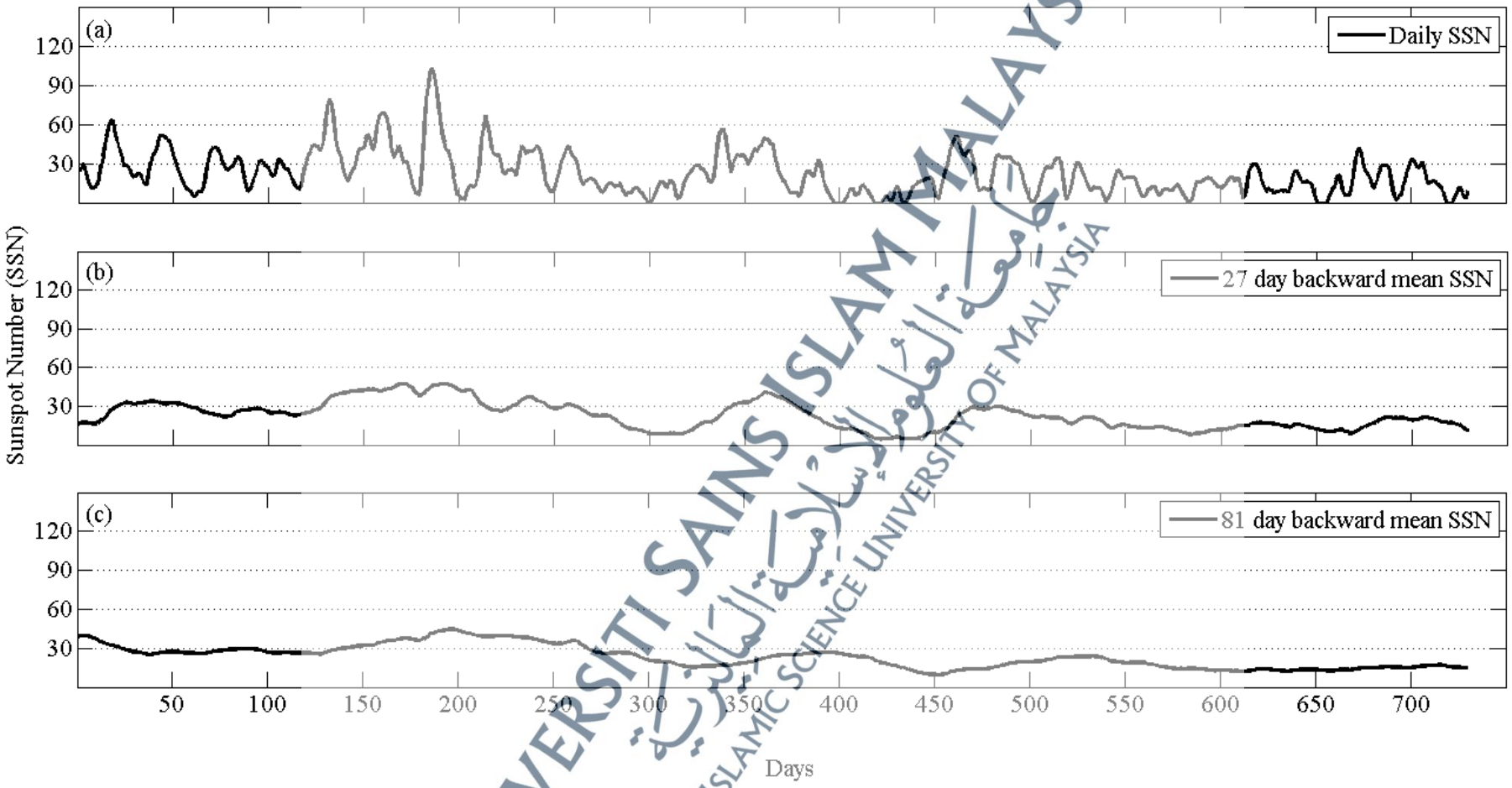


Figure 3.6: (a) Daily, (b) 27 days and (c) 81 days smoothing backward means for SSN during 2005 and 2006

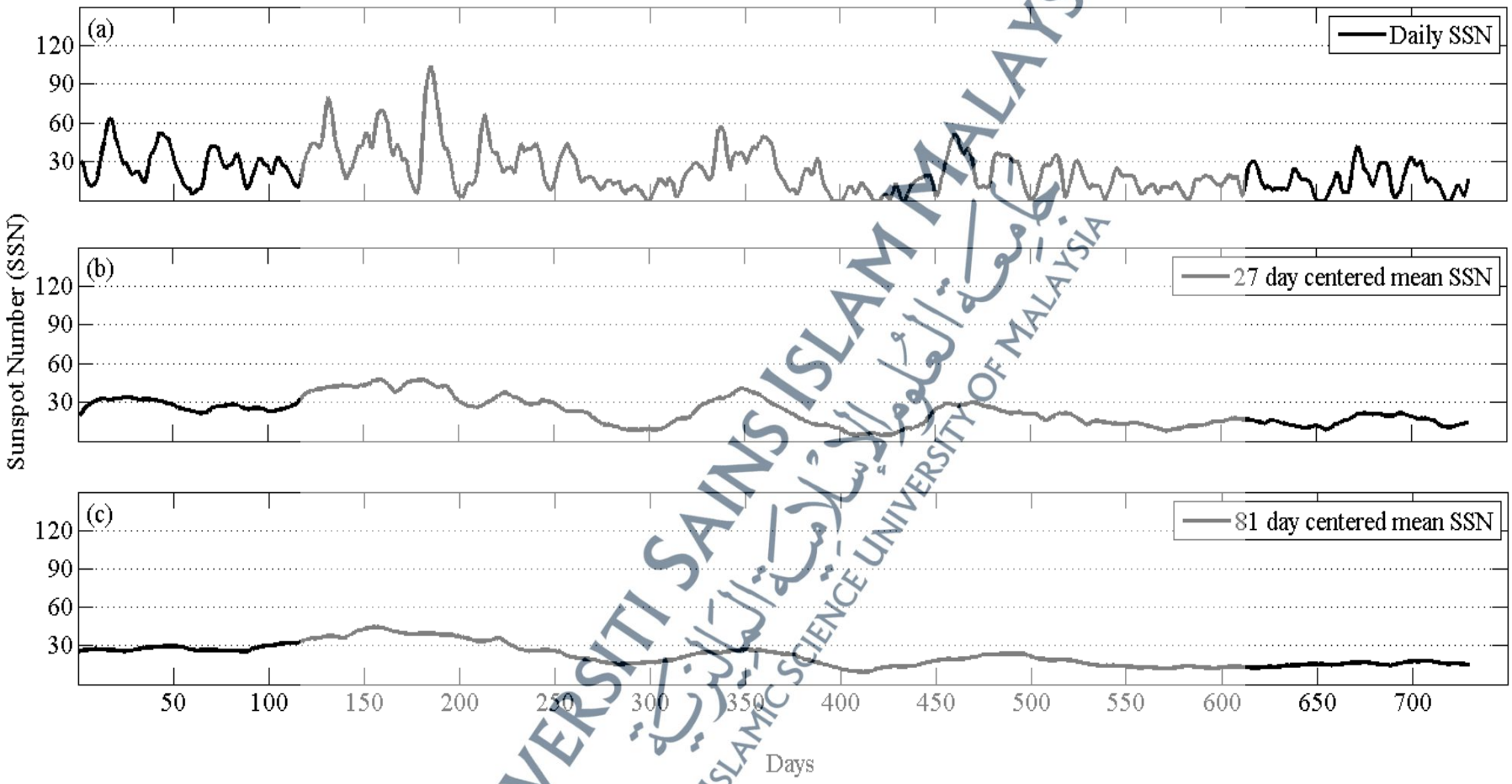


Figure 3.7: (a) Daily, (b) 27 days and (c) 81 days smoothing centered means for SSN during 2005 and 2006

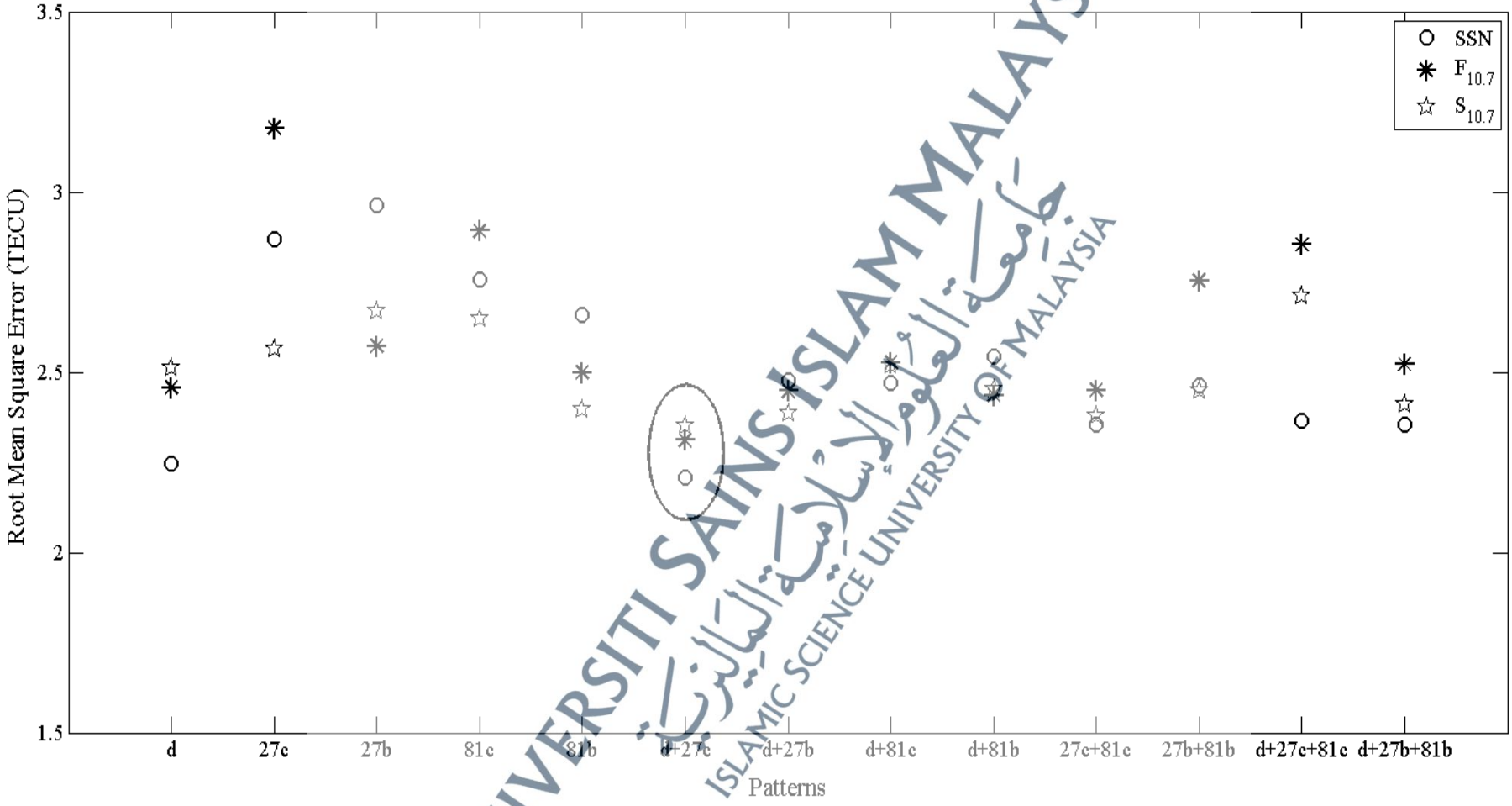


Figure 3.8: Combinations of different periods for SSN, $F_{10.7}$ and $S_{10.7}$

Both the short and long-term variations of proxies are included in the NN models and found that the 27-day which is equal to one solar rotation yielded more accurate results than 81-day. Tobiska et al. (2008); Watthanasangmechai et al. (2012) and Maruyama (2010) have concluded that the long-term means performed better and improved the TEC modelling. Conversely, in this study, the 27-day period is fairly good to describe the TEC variations. This may due to the duration of the dataset. Mostly the aforementioned researchers have used longer period of data set (more than 4 years) in the training process whereas in this research less than two years data set are used in the training process. Thus, the 27-day solar rotation may represent as a better index and more effective for short duration of dataset.

Knowing that the combination of daily and 27-day periods concurrently yielded the least RMSE and shows improvement in TEC modelling, both the periods are used to investigate further. A few studies have investigated the effective solar inputs specifying the ionospheric variations and they concluded that more than one solar proxy describe the ionospheric parameter better than a single proxy (Tobiska et al., 2008; Mckinnell, 2008; Bowman et al., 2008). Thus, a variety of combinations of solar proxies (SSN, $F_{10.7}$ and $S_{10.7}$) with a concurrent use of short term means (daily and 27-day) is investigated. The effects of different solar proxies in TEC modelling using short term periods are compared and summarized in Table 3.1. Long term variations are not included in the modelling, since, the previous result shows that the 81-day centered or backward means of SSN only had a small effect on the TEC modelling. Similarly for EUV and solar radio flux, the inclusive of 81-day means had no improvement on the modelling while the indices deteriorated the estimation accuracy.

Table 3.1: Combinations of solar proxies [i, ii]

No. of run	Proxy			RMSE
	SSN	$F_{10.7}$	$S_{10.7}$	
1	b	b	b	2.56
2	c	c	c	2.74
3		b	b	2.22
4		c	c	2.76
5	b	b		2.27
6	c	c		2.43
7	**b**		**b**	**2.16**
8	c		c	2.19

[i] Each proxy consists of combination of daily and 27 day period
[ii] b denote backward mean; c denote centered mean

Generally the result shows that the combined use of different proxies improved the model efficiency when compared with the cases in which SSN, $F_{10.7}$ and $S_{10.7}$ are used separately. Overall, the backward mean for multiple proxies is better than the centered mean. The combined use of $S_{10.7}$ and SSN yielded the least RMSE than the other types of combinations which are examined. For both SSN and $S_{10.7}$, the combination with $F_{10.7}$ deteriorates the performance of the model. In addition, the combinations of all three proxies are examined and result is worse than the other cases. A correlation study between TEC and solar indices ($F_{10.7}$ and SSN) was conducted by Bilitza (2001) and found that it is desirable to use $F_{10.7}$ (SSN) index in ionospheric modelling for region which exhibits strong solar variations below (above) the F-peak. In a recent study in

Malaysia, Rhazali & Zain (2011) have investigated on true height of F-layer over Parit Raja, Malaysia and stated that the true height density peak (h_{max}) is around 550±50 km during daytime and 300 km from midnight till pre-sunrise hours. Since the h_{max} is above the F-peak, SSN would be a better index to represent solar activity in this region. In the other hand, Maruyama (2010) describes that the wavelength of EUV that ionizes the ionosphere, overlap with the wavelength of $S_{10.7}$. Therefore this index offers the possibility of a more accurate representation of TEC variability.

The results from Table 3.1 clearly show that $S_{10.7}$ and SSN are the best combination of solar proxies to improve the estimation accuracy of ionospheric TEC modelling, which is consistent with the clarification given by Bilitza (2001) and Maruyama (2010). Therefore, in this thesis $S_{10.7}$ and SSN are adopted as a representation of solar proxies for modelling the ionospheric TEC at this region. Other than solar inputs, the other factor that is known to influence the TEC variability is magnetic activity. The following section discussed to find the optimum parameter(s) to represent magnetic activity at this region.

3.4.1.2 Optimum magnetic proxies

The NN models are trained in order to find the effective magnetic indices which attributed to the ionospheric TEC variability in this region. The optimum solar proxies, Cp, and the magnetic proxies described above are used as the input parameters in the NN model to determine the TEC values. A combination of input parameters, which yielded the smallest RMSE is considered to be the optimum parameters for TEC estimation. Table 3.2 shows the RMSE values obtained from each pattern, where an improvement is

Table 3.2: Combinations of magnetic and solar proxies

Pattern	Proxy				RMSE
	Solar Proxies	Cp	Dst	ap	
1	✓	✓	✓		2.22
2	✓	✓		✓	**1.96**
3	✓	✓	✓	✓	2.26

observed in the NN training results when the magnetic index ap is included in the input space of the NN model. The ap index can be considered as the optimum magnetic input parameter since the proxy is fairly good to describe the TECvariations over Parit Raja station. The use of two magnetic indices concurrently deteriorated the estimation accuracy of the NN training. However, the overall result has shown that the inclusive of magnetic indices in the NN training, improved the overall accuracy of the estimation model.

In conclusion, the daily as well as the 27 backward means of $S_{10.7}$ and SSN (solar proxies), ap (magnetic index), HR (diurnal variations) and DN (seasonal variations) are the best combination to improve the TEC estimation accuracy based on NN. Hence all the aforementioned proxies are adopted for TEC modelling over Parit Raja, Malaysia. Even though the geographical position does contribute in TEC variability but in thesis this parameter is not included as in input space, since the work only dealt with a single station. Once the model is trained with sufficient and well-defined information, the NN model is capable of generalizing on new or unseen data.

3.4.2 Determination of optimum hidden neurons

NN is an information processing paradigm with high degree of interconnection processing element known as neurons. Number of hidden layer neuron(s) is an important criterion to develop an optimal NN architecture because inappropriate neurons in the hidden layer may cause the NN model becomes ill-conditioned. If the number of neurons are less as compared to the complexity of the application then "underfitting" may occur. This occurs when there are too few neurons in the hidden layers to adequately detect the signals in a complicated application. Conversely, if unnecessary more neurons are present in the network then "overfitting" may occur. Therefore, it is essential to compute the number of neurons that should be kept in the hidden layer to achieve an optimal architecture.

Several general rules are implemented by a number of researchers to determine the best number of hidden units depending on the application, yet according to Shuxiang & Ling (2008) most of those rules are not applicable in most circumstances. This is because the process of deciding the number of hidden units depends on many criteria e.g. number of inputs and outputs, number of trials, the activation function in the hidden layer, training algorithm, complexity of the application, etc (Kumar, 2004; Shuxiang & Ling, 2008). Alsmadi et al. (2009), has also indicated that there is no explicit specification and experimental work to explain the layout of a network.

Therefore in this thesis, the optimal hidden nodes are attained by varying the number of nodes in the network. It is done by adding one hidden node at a time, then train and validates the network and finally calculates the RMSE between the GPS TEC

and NN TEC. The NN that gives the smallest RMSE is adopted as the optimal number of hidden neurons. In this work, the results show that the number of hidden nodes that provided least RMSE is summarized as follow:

$$Number\ of\ hidden\ neurons = n + 2 \tag{3.8}$$

where *n* is the number of inputs space in the NN.

Further increase or reduce number of neurons in the hidden layer produced high generalization errors due to over-fitting or under-fitting, as well effects the learning quality and time.

3.4.3 Determination of optimum training algorithm

Training algorithm used in the NN is also another factor which can affect and contribute on the improvement of the NN training. Hagan et al. (1996) and Habarulema & McKinnell (2012) have emphasized that a proper selection of training algorithm is able to minimize the generalization error in the estimation model and enhance the predictability of the model. In other words, a network with proper training algorithm does contribute in developing a reliable model. There are several different training algorithms in feed forward NN, which are applicable in various applications like behavioural sciences, financial, medical, engineering, pharmaceutical, physiochemical and other technical fields have shown that different applications require different training algorithms (Plumb et al., 2005; Koker et al., 2007; Ghaffari et al., 2006; Torrecilla et al., 2007). Generally it is hard to determine which of the training algorithms have the highest efficiency with fastest rate of convergence for a specific application.

In this thesis, the NN training is assessed among eight different training algorithms using Neural Network toolbox of MATLAB. The algorithms belong to four major methods: gradient descent, conjugate gradients, quasi-Newton, and Levenberg - Marquardt. The training algorithms shown in Table 3.3 are analysed in this thesis. The eight training algorithms presented here were chosen based on the popularity of these algorithms in many applications (Koker et al., 2007; Plumb et al., 2005; Sakamoto et al., 2005; Ghaffari et al., 2006; Torrecilla et al., 2007; Habarulema, 2007; Leandro & Santos, 2007; Mckinnell, 2008; Alsmadi et al., 2009; Kenpankho et al., 2011; Watthanasangmechai et al., 2012). A comprehensive description on the training algorithms and the features are fully described in other sources (Fausett, 1994; Haykin, 1999).

Table 3.3: Training algorithms used in NN training

Training algorithms	Descriptions	Methods
GDA	Batch gradient descent with variable learning rate	Gradient descent
RP	Resilient back-propagation algorithm	Gradient descent
CGF	Fletcher – Reeves Update conjugate gradient	Conjugate gradient
CGB	Powell – Beale restarts conjugate gradient	Conjugate gradient
CGP	Polak-Ribiere Update conjugate gradient	Conjugate gradient
SCG	Scaled conjugate gradient	Conjugate gradient
BFG	Boyden, Fletcher, Goldfarb and Shanno update	Quasi -Newton
LM	Levenberg-Marquardt algorithm	Levenberg-Marquardt

This study concentrated on the performance of the training algorithms to determine the optimum training algorithm. In order to allow a direct performance comparison between the algorithms, identical dataset and network architecture are used to test the aforementioned algorithms. The best algorithm for TEC estimation based NN is obtained via the comparison. The network architecture which was implemented during the determination of optimum input spaces and number of hidden neuron(s) is used. The assessment of the NN modelling requires evaluation of data excluded from the training set. In this case, the hourly TEC data in July 2005 is used to validate the performance of the model. In order to determine the best training algorithm in estimation of TEC, a few performance criteria are taken into consideration. The NN model is compared in terms of capability, reliability and computational time of the training.

The first criterion is the determination coefficient (R^2) of the linear regression line between the NN TEC and GPS TEC. R^2 is used to study the capability of the network. This measurement denotes the strength of the linear association between the two variables, the estimated and observed values and shows the predictability of TEC using the designed model. The R^2 takes on value between the interval [0, 1] and the greater the R^2 is, the better the designed model able to describe the observational data. A scatter plot is generally used to graphically represent the correlation between the estimated and observed values as well as aids the interpretation of the gradient (m), intercept (c) and R^2. For each training algorithm a scatter plot with the linear regression analysis is computed and illustrated between the NN TEC and GPS TEC in Figure 3.9 and Figure 3.10.

The R^2 gives a simple measurement of a model performance as well as measure the ability of a trained model to estimate. The NN models are tested on the dataset which

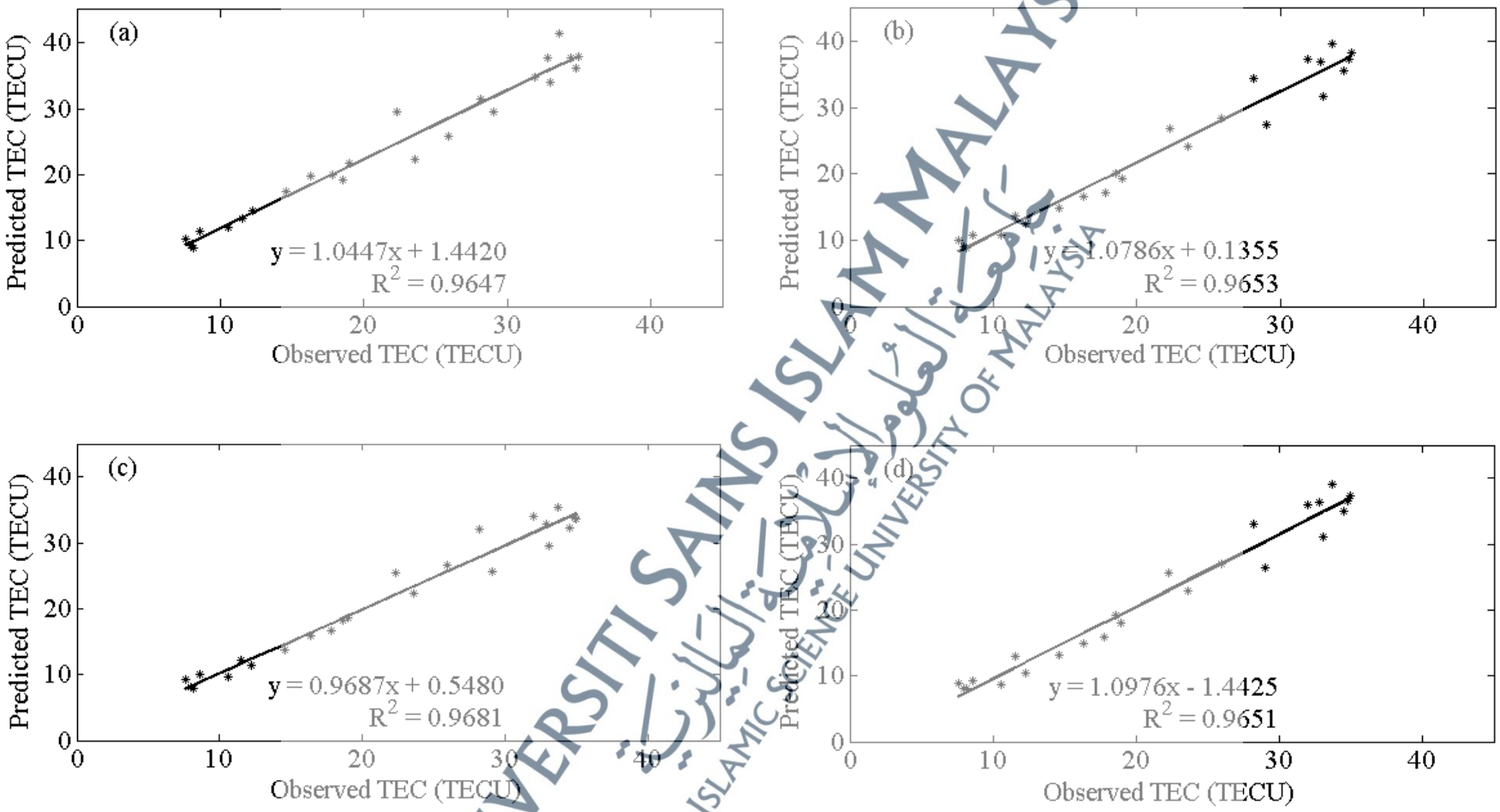


Figure 3.9: Linear regression between the observed and estimated TEC values using (a) GDA, (b) RP, (c) LM and (d) BFG algorithms

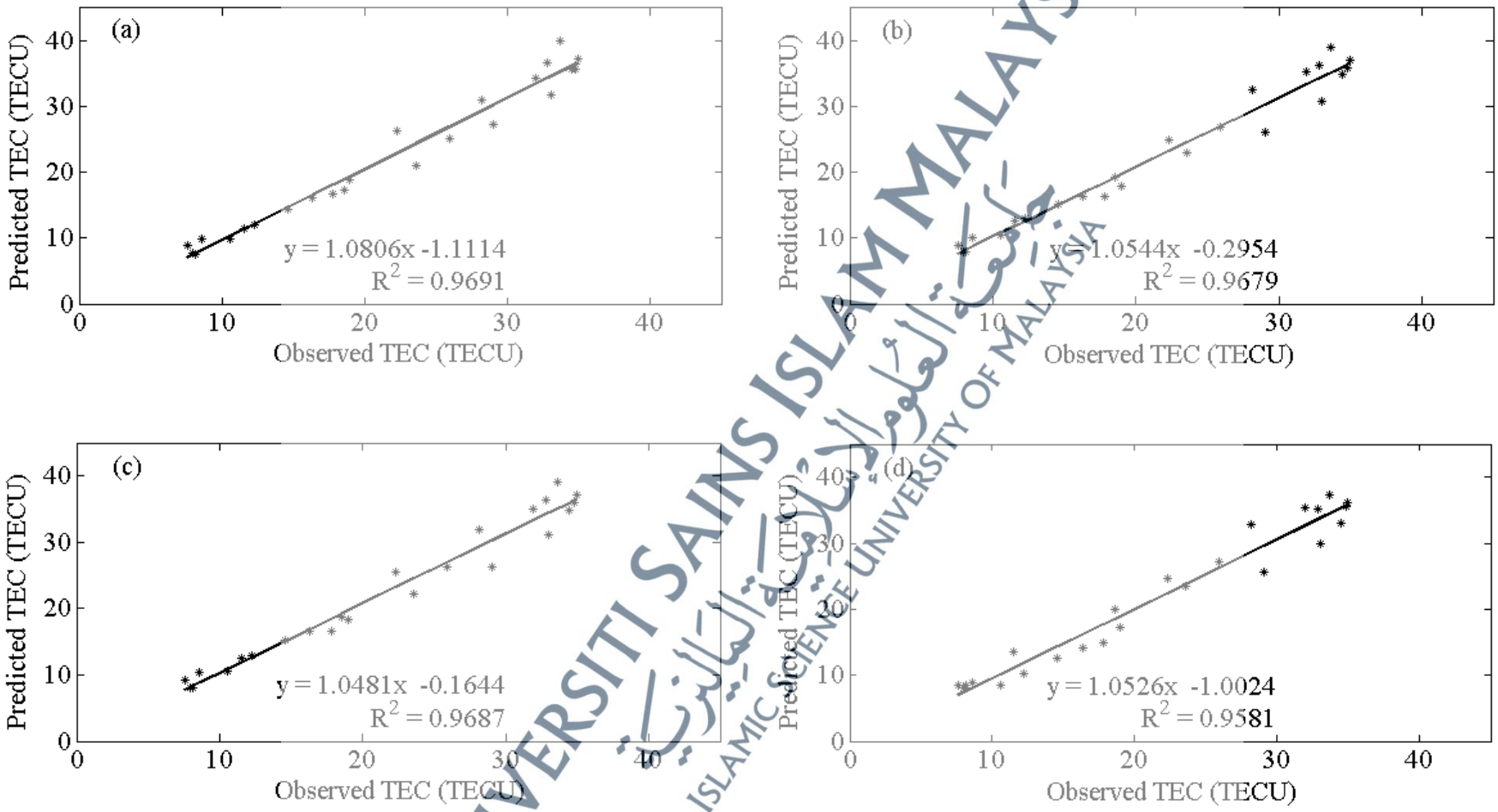


Figure 3.10: Linear regression between the observed and estimated TEC values using (a) CGB, (b) CGF, (c) CGP and (d) SCG algorithms

was excluded from the training, and results shown that all the mentioned training algorithms achieved R^2 value significantly greater than 0.9 in NN training. This is mainly denoted by the characteristic of NN which able to interpolate well within the input space (Habarulema, 2007). Amongst the training algorithms, the NN model trained by CGB training algorithm (R^2 = 0.9691) shows the highest estimation accuracy, while SCG training algorithm shows the lowest estimation accuracy (R^2 = 0.9581). The result indicates that the NN model trained by SCG is not as good as the CGB. However, the overall deviation of the R^2 values among the eight training algorithms is small and the measurement only differs in the range of 0.004 to 0.0044 from the highest determination coefficient. Ghaffari et al. (2006) have mentioned that the R^2 evaluates the robustness of a model and according to them a model that provides R^2 more than 0.9 is considered a well trained model and can be regarded as a good overall fit. In this case all the algorithms well suited in training the NN model to estimate the ionospheric TEC.

The following criterion is the RMSE, which used to distinguish the model's performance between the training algorithms. The training algorithm which gives the minimum RMSE is considered as the most appropriate training algorithm in the NN training. Figure 3.11 and Figure 3.12 show the diurnal variations of TEC and the errors (between GPS TEC and NN TEC) in TECU with respect to hours in Universal Time (UT) for all the training algorithms.

Overall, the NN training results showed a good agreement with the observational data. The RMSE is computed for each training algorithm and their abilities to estimate are in the order of: LM > CGP > CGB > CGF > SCG > BFG > RP > GDA. The LM algorithm yielded the highest accuracy with least error in estimating the TEC values. The

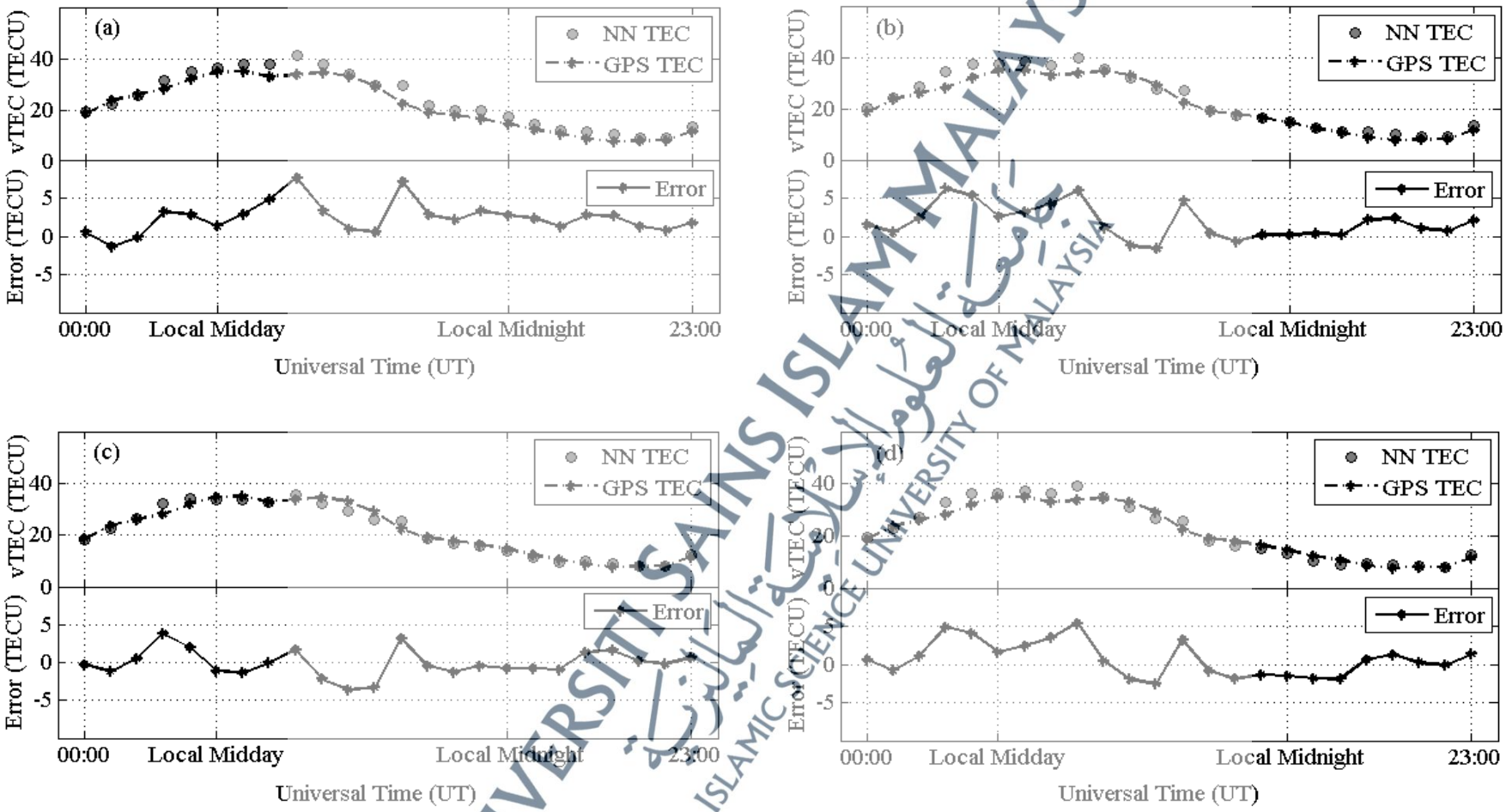


Figure 3.11: Computed error values between GPS TEC and NN TEC based on (a) GDA, (b) RP, (c) LM and (d) BFG

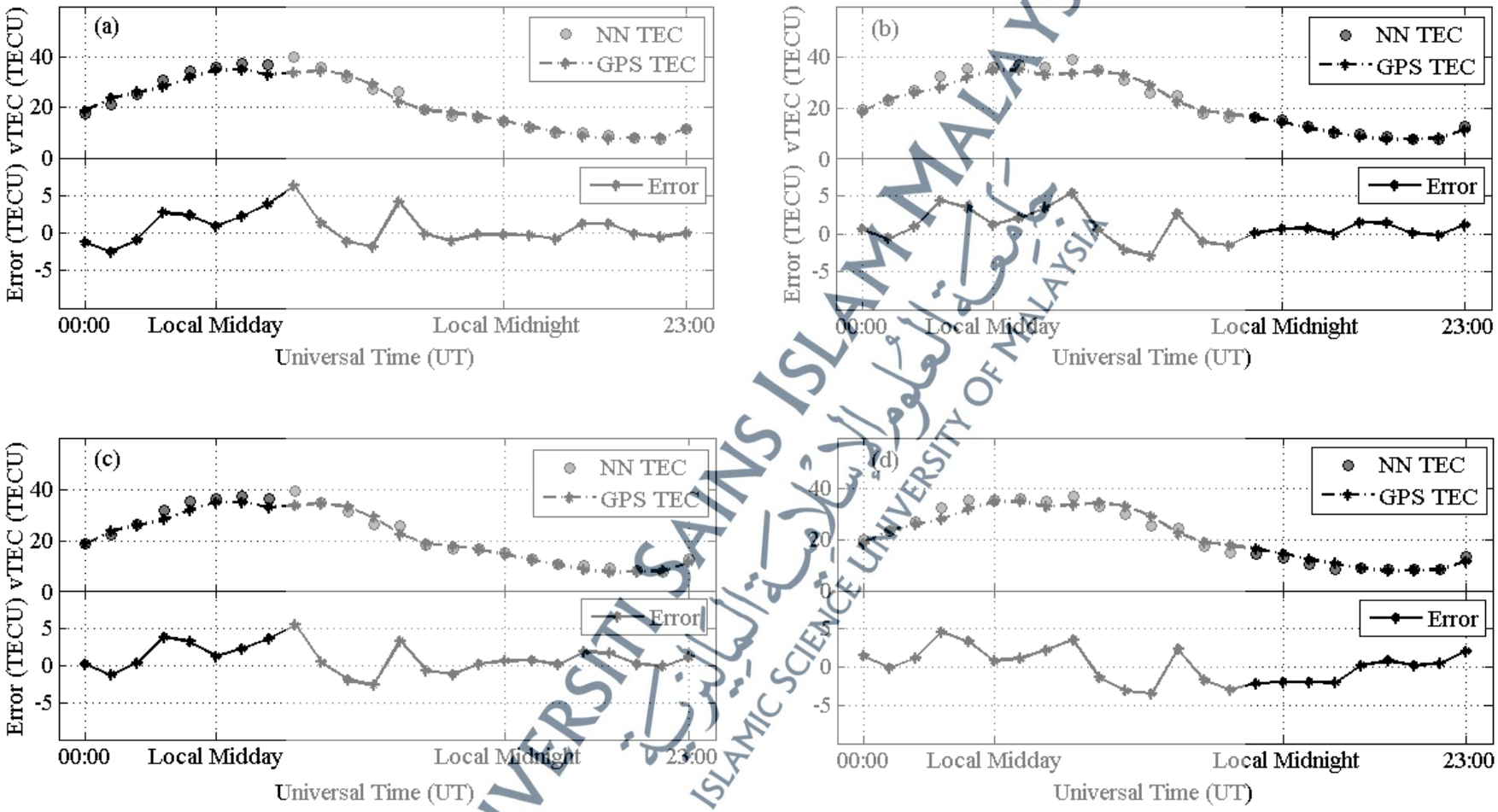


Figure 3.12: Computed error values between GPS TEC and NN TEC based on: (a) CGB, (b) CGF, (c) CGP and (d) SCG

RMSE value for LM is ~1.7534 TECU and for this algorithm the error range is between -4 and 4 TECU, whereas the remaining algorithms have wider range of errors which can be evidently seen in the Figure 3.11 and Figure 3.12. The GDA algorithm had contributed the highest RMSE compared to the rest of the training algorithms. The performance of this algorithm is very sensitive and may disrupt, mainly due to the determination of learning rate in the initial setting of the network (Demuth & Beale, 2002). The difference in the RMSE values between the LM and the gradient descent algorithms, GDA and RP are 1.3802 TECU and 1.0669 TECU, respectively. The conjugate gradient descent algorithms give better estimation accuracy compare to GDA. The models trained with CGB and CGF give improvements in estimation accuracy by 31.65% while CGP and SCG improvements of 33.12% and 29.13% with respect to GDA, respectively.

Finally, the training period of each algorithm is observed and used as a determinant in the choice of the algorithms in the NN. Computation time is an important factor in NN based application since it tells the approximate time required by the model to achieve the convergence during the training process. The designed model that has long training time will lead to over-fit or over-generalize the data. The over-fitting process causes the model begins to adopt random noise in the data, and generalizes poorly to new or "unseen" data. This may consequently affect the performance of the designed model and might as well deteriorate the end results of the model. Thus, the computational time requires by each training algorithm during the training period is considered as a performance criterion in evaluating the network. The algorithm with the fastest convergence rate is used for the further development of the model.

The timeline in Figure 3.13 shows the computational time (seconds) required by

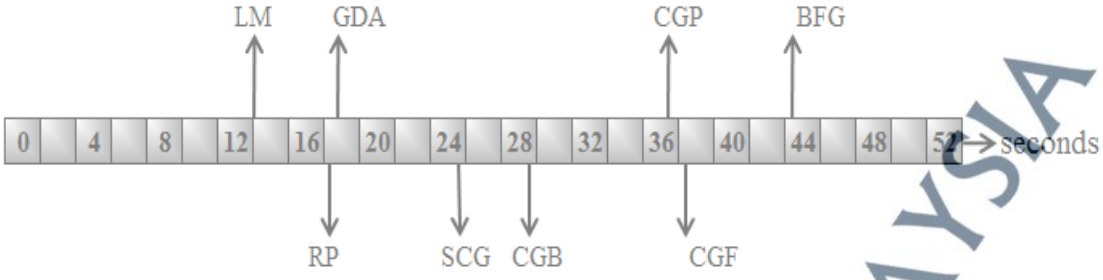


Figure 3.13: Time line of the training algorithms in seconds

each training algorithm. The result exhibits that the LM training algorithm has the fastest convergence among the others. Despite the fact, LM training algorithm requires more memory and high computational complexity, yet the algorithm still yielded the least computation time compared to other training algorithms. Sakamoto et al. (2005) stated that the least computational time is one of the properties that generally motivate the researchers to employ this algorithm in many NN based applications. The convergence speeds of conjugate gradients are generally slow and show significant difference compared to LM. The computation time of each conjugate gradient training algorithm is approximately 2 to 3 times higher than the LM. The convergence speed of LM is 45.49%, 34.69%, 35.69% and 53.7% faster than the computational time required by CGB, CGF, CGP and SCG, respectively. The BFG algorithm has the slowest convergence rate, where the LM algorithm only needed 30% of the BFG's computational time.

According to Hagan et al. (1996) and Koker et al. (2007), the LM training algorithm requires higher storage for large dataset because the algorithm needs to store the approximate Hessian matrix. However, in this study the LM still gives the less number of iterations with the fastest convergence rate compared to other algorithms. This may be due to the small duration of training dataset used in this study. This shows that

different applications and dataset require different algorithm and a specific training algorithm can't fit in all kind of scenarios.

The results of the performance criteria for each training algorithm are summarized in Table 3.4. The training algorithm with highest R^2, lowest deviations root mean square and fastest convergence rate is considered to be the most adequate training algorithm. In terms of correlation coefficient, all the algorithms achieved more than 95% estimation accuracy and can be considered as good algorithms. Since it is essential to attain an optimal training algorithm, all of them are compared over RMSE and computation time.

Table 3.4: Comparison of performance index between the training algorithms

Training algorithm	Performance Index		
	Correlation Coefficient, R^2	RMSE (TECU)	Computation Time (second)
GDA	0.9647	3.1336	17.784
RP	0.9652	2.8203	17.597
CGB	0.9691	2.1417	28.689
CGF	0.9679	2.1418	37.614
CGP	0.9687	2.0956	36.567
SCG	0.9581	2.2209	24.290
LM	**0.9681**	**1.7534**	**13.050**
BFG	0.9651	2.3527	43.432

The LM algorithm is found with least error and has the fastest convergence rate. It is difficult to generalize a specific training algorithm for all types of applications due to the complexity of each research area, but in this study the NN trained with Levenberg - Marquardt provides the optimum TEC model. Although this work does not investigate all types of training algorithms, it serves as a guide to NN users who may want to apply different training algorithms to their datasets, especially for ionospheric applications.

3.4.4 Optimum architecture of neural network

The architecture of NN that gave high correlation coefficient, least RMSE and fastest convergence rate during the determination of optimum input spaces, optimum hidden neurons in the hidden layer and optimum training algorithm is employed in order to estimate the GPS TEC. The NN configuration uses 9 numbers of nodes (SSN (d +27b), $S_{10.7}$ (d+27b), ap, HRs, HRc, DNs, DNc) in input layer, 11 nodes in a single hidden layer and 1 node (ionospheric TEC) in output layer designated as 9:11:1 used in this thesis.

A feed forward multi-layer network associated with Levenberg - Marquardt (LM) back-propagation algorithm is implemented to provide optimum results. Hyperbolic tangent sigmoid function is used as an activation function in both layers; hidden and output layers. The sigmoid function is always preferable in non-linear technique, since the function is continuous and can be easily differentiable at any point. In other words, the function is capable to move forward from a layer to another layer which is a main criterion in a multi layer network. The optimal NN architecture is illustrated in Figure 3.14.

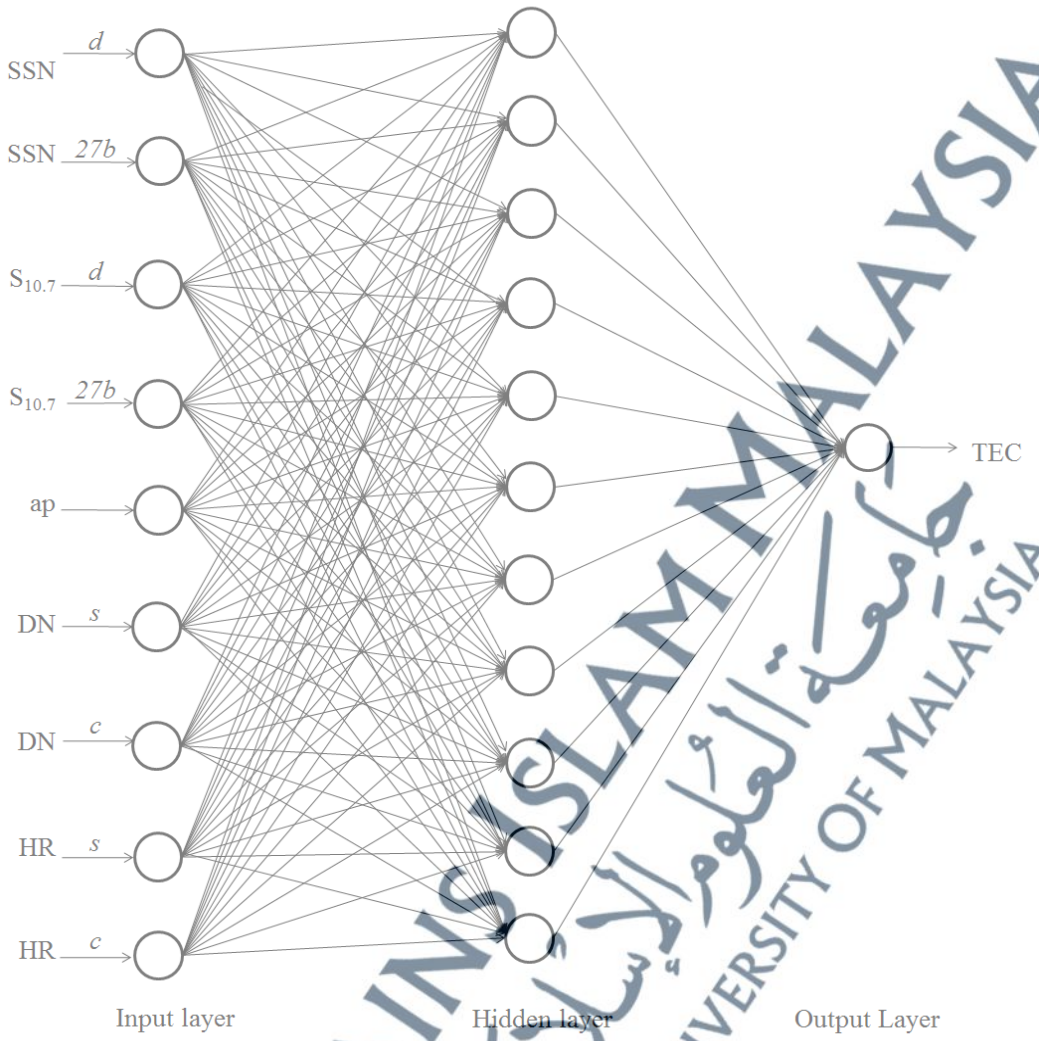


Figure 3.14: Optimal feed forward neural network architecture for TEC estimation model over Parit Raja

In the NN model, all the inputs and the corresponding GPS TEC are scaled to take values between the range of -1 and 1 so that the maximum and minimum value of any input and output is within the range values. This indirectly reduces the computation time of the network during learning and testing phases. The input-output mapping features of the NN develop a relationship between the TEC and the input parameters; solar, geomagnetic, diurnal and seasonal during the learning process with the aid of inter-neuron connections between the layers and the learning algorithm function. The output of

the NN model is known as estimated TEC (NN TEC) and it is expressed as a function of the SSN, $S_{10.7}$, ap, DN and HR, as in Equation (3.9):

$$NN\ TEC\ =\ f(SSN(d+27b), S_{10.7}(d+27b), ap, DNs, DNc\ HRs, HRc) \tag{3.9}$$

where d is the daily value, 27b is the 27 days backward mean, c and s are the cosine sine components, respectively.

Hourly GPS TEC data from 2005 to 2006 are used to train, validate, and test the performance of the NN based TEC model to estimate the ionospheric TEC over Parit Raja. The validation and testing data are excluded from the training process to prevent memorization. Memorization could deteriorate the generalization capability of the model and causes over fitting. Therefore, to further improve the generalization of NN, a performance function is used and monitored during the training process. Mean square error (mse) is represented as a stopping criterion whereby the training is halted when the mse of the dataset starts increasing. Once the NN model is trained, the predictive performance of the model is investigated. The trained NN model is tested on new or "unseen" dataset to verify the model's generalization capability.

In this thesis, the interpolation and extrapolation capabilities of the developed NN model to estimate the missing GPS TEC are investigated. Two NN models, NN1 and NN2 are developed to estimate the hourly ionospheric TEC. The NN1 model is used for interpolation technique. In this network, the March 2006 dataset is employed within the training set period (February 2005 - April 2006), however the data is set aside for testing purposes. The NN2 model is used for extrapolation technique, where in this network, the

March 2006 dataset is outside the training sample (February 2005 - February 2006). In this investigation, the original TEC data are removed and assumed to be missing, so that the real measurements would be still available and can be used for comparative purposes. The findings of this work are presented in Chapter 4, subsection 4.2.1 and 4.2.2.

In order to access the NN's extrapolation capability more thoroughly, the NN2 model is further used to estimate the seasonal TEC and diurnal disturbed TEC. The estimated TEC (NN2 TEC) is compared with the observed TEC extracted from the GPS measurements (GPS TEC) and also with the estimation TEC from the IRI version 2007(IRI TEC) to validate the developed model. The findings of this investigation are presented in Chapter 4, section 4.3 and 4.4.

3.4.5 Verification of NN model

An estimation model requires model verification to quantify the confidence of a developed model's correctness for its intended use. Therefore to verify the creditability of the developed TEC model based NN, part of the training TEC data are used for testing ("seen" data) as suggested by Habarulema et al. (2007) and Habarulema & McKinnell (2012). Correctness, reliability and robustness of the model are ensured via this method. The verification of the model is performed using TEC data in June 2005. The level of agreement between the two variables, the estimated and the observed values is investigated using the determination coefficient, R^2.

Figure 3.15 shows the scatter plot and the R^2 value between the observed and estimated TEC values. It is found that, the correlation coefficient is more than 0.9, which

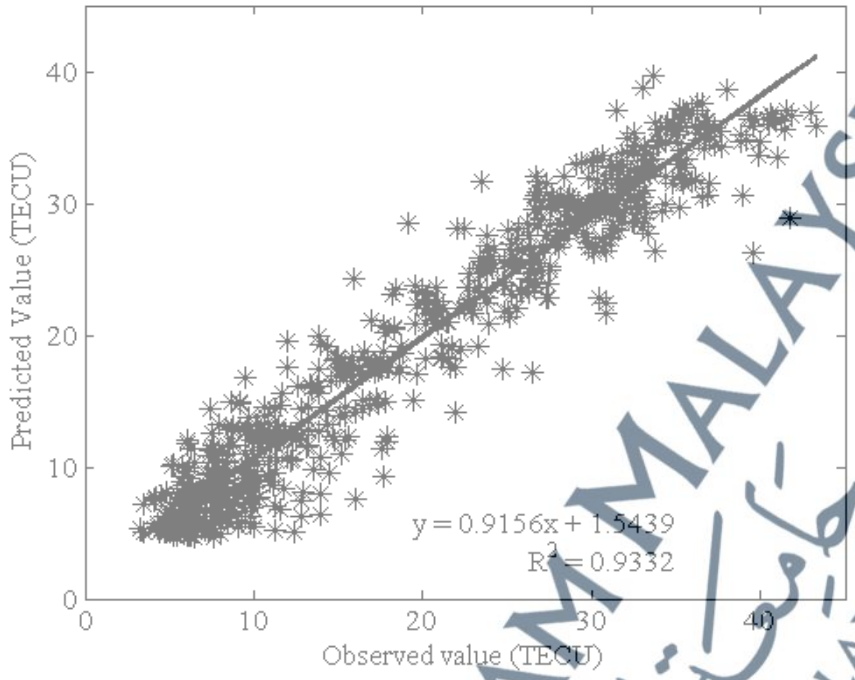


Figure 3.15: Scatter plot and correlation coefficient, R^2 between the observed TEC and estimated TEC

shows that the NN model is able to estimate about 93% of the "seen" GPS TEC values accurately. Overall the performance of the model concludes that the developed NN model can be used for estimation with confidence.

3.5 DEVELOPMENT OF HYBRID SARIMA-NEURAL NETWORK-BASED TEC FORECASTING MODEL

Forecasting the ionospheric parameters ahead is still an ongoing concern even though various time series methods have been used in these applications (Cander et al., 1998; Cander, 1998a; Stankov et al., 2001; Tulunay et al., 2004; Tulunay et al., 2006; Acharya et al., 2009; Garcia-Rigo et al., 2011; Niu et al., 2014). Commonly, the

ionospheric parameters are highly complex and stochastic nature. Hence, it is difficult to completely understand the characteristic of those parameters. In other words, it is difficult to determine an appropriate model to forecast their unique nature ahead. In this type of scenarios, neither non-linear nor linear models can be solely applicable in forecasting the time series patterns.

SARIMA (Seasonal Autoregressive Integrated Moving Average) is a type of linear model modified from ARIMA (Autoregressive Integrated Moving Average). SARIMA models assume that the present data are a linear function of past data points and past errors. They are well known for linear modelling due to their nature of pre-processing the raw data before fitting a linear equation to the data to obtain more accurate predictions. However, using the SARIMA to model the non-linear applications have yielded mixed results because SARIMA do not assume non-linearity. NN (neural network) is an efficient technique for non-linear time series forecasting. They are so adaptive in nature because they are capable of fitting a non-linear function to a given data. The NNs are commonly advantageous compared with SARIMA in many applications because of their non-linearity approach, flexible computing framework and do not require pre-processed data. In order to overcome the limitation of SARIMA models and produce more general and more accurate forecasting model, a non-linear NN technique is hybridized with the SARIMA model to capture both the linear and non-linear components of a complex time series.

The benefits of a hybrid model is supported and agreed by Zhang (2003); Aladag et al. (2009); Khashei & Bijari (2010) and Yolcu et al. (2013). According to them, integrating linear and non-linear models contribute more accurate prediction model than

an individual model in forecasting a complex time series patterns. This may be due to the fact that each individual model carries its own signatures and features which is able to characterize the time series pattern in its own way. The performance of the forecasting model may be enhanced by integrating these different features into a single framework.

Acknowledging the advantage of a combined model, in this work, a feasibility study of a new hybrid method that combines two different techniques, linear SARIMA model and non-linear NN model is carried out to investigate the possibility of the hybrid model to forecast the GPS TEC values ahead. The hybrid model is designed as a function of past observations of the TEC series data to forecast the time series ahead. Figure 3.16 shows the flow chart which describes the overall method of TEC forecasting based on a hybrid technique. The hybrid technique is divided into three phases; phase 1: modelling and forecasting the linear components of the TEC time series using SARIMA model, phase 2: modelling and forecasting the residuals from the SARIMA model (non-linear components) using NN and phase 3: finally, the forecast TEC values are yielded by integrating both the models values and the results are compared with the individual models, SARIMA (FCAST-SARIMA) and NN (FCAST-NN) separately.

3.5.1 Modelling of SARIMA

SARIMA model is an extension of the ARIMA model in order to include the seasonal components which exist in most of the time series. Stationary time series data is a necessary condition in building a SARIMA model used for forecasting. A time series is defined stationary if the statistical properties of the data do not vary with time where the

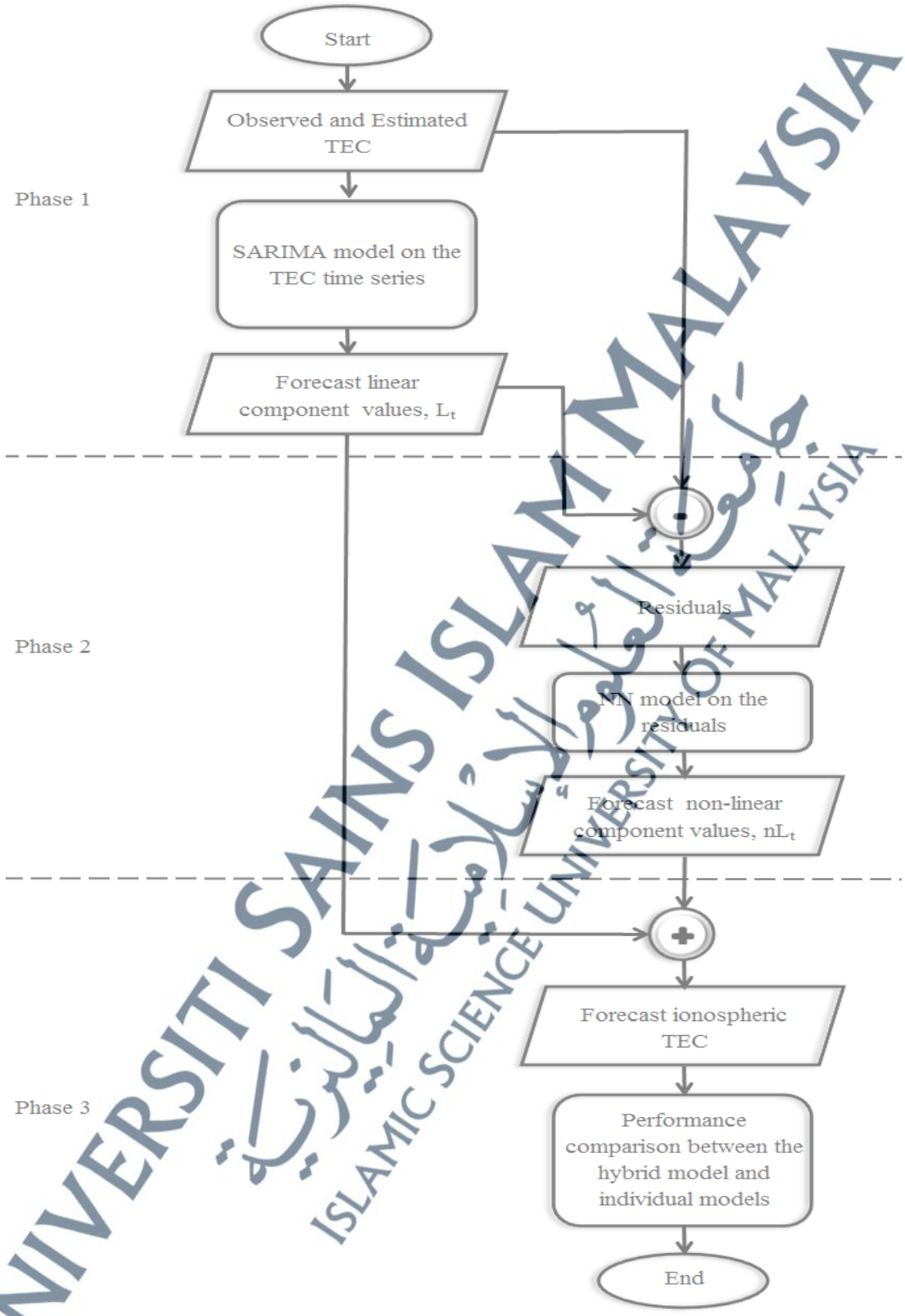


Figure 3.16: Flow chart for development of forecasting TEC model based on the hybrid technique

mean and variance of the data are constant. The time series with trends or seasonality are considered to be non-stationary as explained in Chapter 2 subsections 2.7.4 and 2.7.5. Both the components have significant impact on the time series data, and it is important to identify those components and pre-process it before the data is fitted into the SARIMA model for identification purpose. The pre-process is performed to achieve stationary. Therefore, before fitting the TEC time series into the linear model, it is important to have prior knowledge on the characteristic of ionospheric TEC.

Figure 3.17 is the enlarged portions of some data points of the temporal variation of vTEC in 2005 and 2006, i.e. the diurnal variations of GPS TEC for 15 days. The results show that, almost similar diurnal curve is observed every day. In the diurnal pattern, rapid and stochastic fluctuations in the vTEC exhibit the behaviour of the vTEC in the equatorial anomaly region. The vTEC shows a steady rise in the early morning till noon, maximizing during the post noon and gradually decreases after the sunset. The diurnal behaviour shows large variations in vTEC during the day time while the amplitude of vTEC is found to be invariant at night time. The TEC time series is considered non-stationary because it has both, the stochastic trend and periodic characteristics. The periodic variation occurs over a short period. In this work, the periodicity of the time series is 24 since the TEC variations is recurring daily (an hourly data).

A statistical hypothesis test is conducted on the TEC time series to verify the stationarity of the time series. There are many methods to perform the stationarity test, for instance the Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) test, unit root test, Augmented Dickey-Fuller (ADF) test, and Leybourne-McCabe (LMC) test. In this

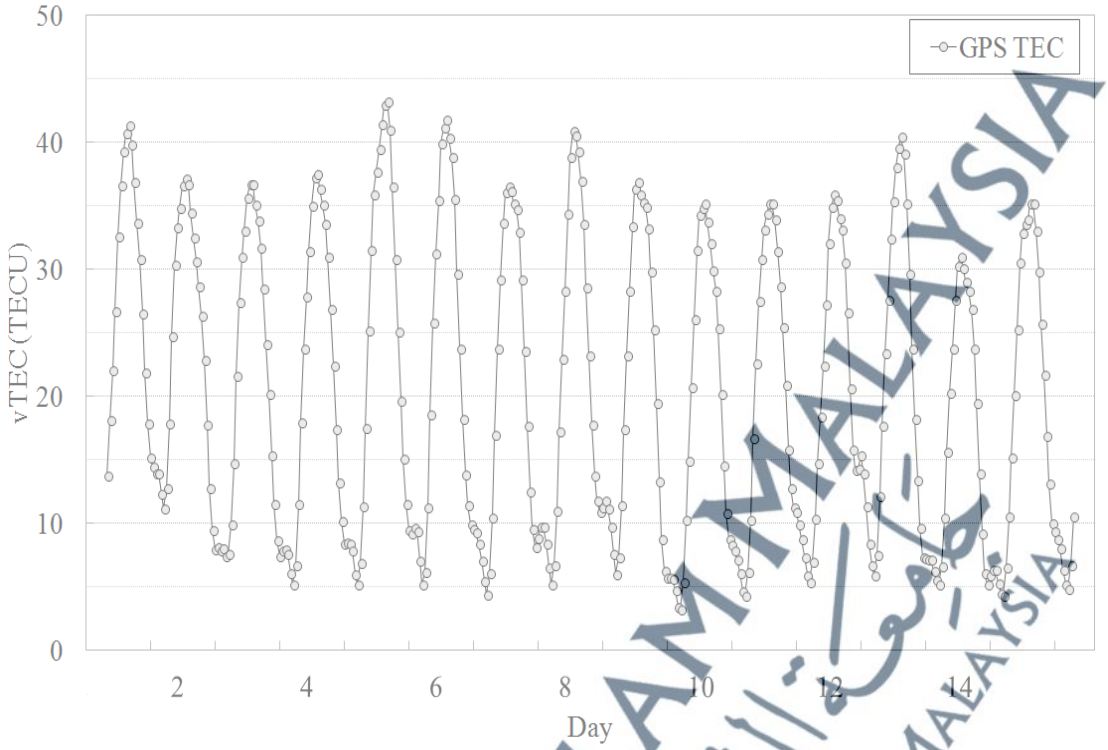


Figure 3.17: Hourly variation of TEC in June 2005 at Parit Raja

analysis, the KPSS test is adopted to examine the stationarity of TEC series using MATLAB. That is:

$$[h] = kpsstest\ (y_t) \tag{3.10}$$

where y_t is the TEC time series, h is a logical value represents the test rejection value. The value of h equal to 0 (null hypothesis) indicates the test takes the null hypothesis that y_t is considered trend stationary. The value of h equal to 1 (alternative hypothesis) indicates the test rejects the null hypothesis that y_t is considered as a non-stationary trend time series. The result from the KPSS test shows that the TEC time series is a non-stationary time series data, where the h value is equal to 1.

In order to further cross check the statistical hypothesis test, an autocorrelation function (ACF) is used to examine the TEC time series. In brief, the ACF gives the visibility to identify the stationarity of a time series. The behaviours and properties of ACF are explained and summarized thoroughly in Chapter 2. The autocorrelation plot of the TEC time series is displayed in Figure 3.18. Commonly, in a non-stationary time series, the ACF is always dominance by the trend and seasonal components compare to the other components in the time series. The sample of autocorrelation function exhibits identical periodicity, where high spikes are observed at the exact seasonal lags; L, 2L, 3L, and 4L. Since it is an hourly data, the L value is equal to 24. This is one of the main properties of a seasonal time series. Besides, there are also other spikes greater than two standard errors at near the seasonal lags, which are L –2, L – 1, L + 1, L + 2, 2L – 2, 2L – 1, 2L + 1, 2L + 2, and so on. Those lags represent the non-seasonal component. According to Theresa (2013), the time series values are considered stationary if the ACF shows cuts off or dies away fairly quickly through different lags at the non-seasonal and seasonal levels, otherwise the time series values are considered non-stationary. Since the ACF pattern shows 24 hours seasonality and does not cut off or slowly decay to zero, the TEC time series is neither stationary in mean nor in variance, thus the data must be transformed to achieve stationary.

An appropriate method is needed to decompose the trend and seasonal components. Mostly differencing and power transformations methodsare applied to the time series data to adjust those components (Zhang, 2003; Khashei & Bijari, 2010). The detrending technique removes the deterministic and stochastic trend of the time series. Besides, this technique helps to stabilize the expectation of the time series. In this study, the power transformation and the combinational differencing methods are implemented

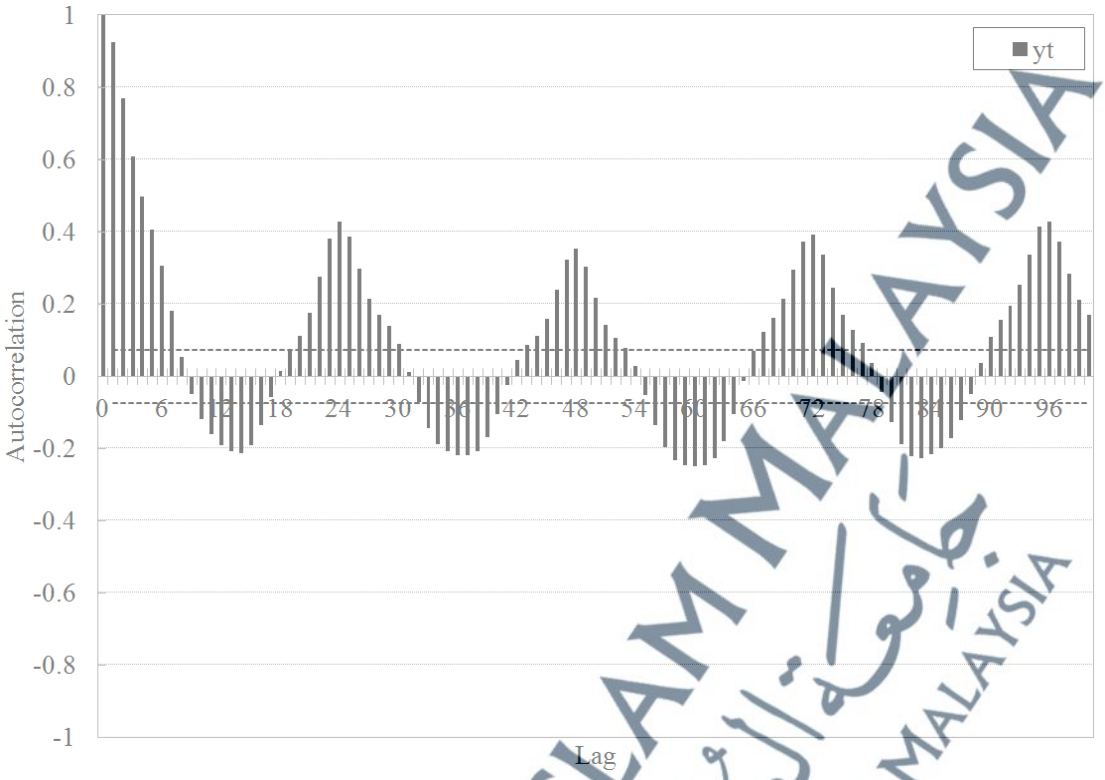


Figure 3.18: Autocorrelogram of the TEC time series in June 2005

in tandem to handle the trend and seasonal time series. Firstly the logarithm (to the base of 10) is used on the TEC time series to stabilize the variance of the data series. Subsequently, the log-data are transformed into stationary time series z_t, by differencing. A mixed of first non-seasonal differencing (regular differencing) and second seasonal differencing are applied on the log-transformed data. Equation (3.11b) expresses the differencing transformation method:

$$z_t = (1-B)(1-B^{24})^2\, y_t \tag{3.11}$$

$$= (1-B)(1-2B^{24}+B^{48})y_t \tag{3.11a}$$

$$z_t = (1-B-2B^{24}+2B^{25}+B^{48}-B^{49})y_t \tag{3.11b}$$

where z_t time series can be also referred as stationary TEC time series or deseasonalised and detrended time series, y_t denotes the non-stationary TEC time series, B denotes as the backward shift operator, $(1-B)$ is the non-seasonal differencing, and $(1-B)^{24}$ is the seasonal differencing. Since $By_t = y_{t-1}$, therefore

$$z_t = y_t - y_{t-1} - 2y_{t-L} + 2y_{t-L-1} + y_{t-2L} - y_{t-2L-1} \tag{3.12}$$

where L represents the seasonal lag which in this case, equals to 24.

The time series achieved stationary with respect to mean and variance once the differencing transformation is implemented. The result of the transformation is plotted in Figure 3.19 along with the original TEC time series, y_t (GPS TEC). Higher degree differencing is required if the time series are still non-stationary. The stationary time series, z_t is further used in model identification. SARIMA model is adopted in this study to model and forecast the linear component in the TEC time series. GPS TEC data over a period of 20 months (February 2005 to September 2006) are used for model calibration to obtain the optimal SARIMA model for the TEC series, while 3 months TEC data (October 2006 to December 2006) are used for model verification, forecasting and comparison purposes. The SARIMA model is implemented by MATLAB version 8.0.0.783 (2012b) with 32-bit (win 32) is installed in a system which has Intel Pentium 987 processor with the speed of 1.5 GHz and 4 GB memory under Windows7 operating system.

Model identification is the first iterative step in building up a seasonal ARIMA model. The shorthand notation for SARIMA model is SARIMA(p,d,q)$(P,D,Q)_s$. The

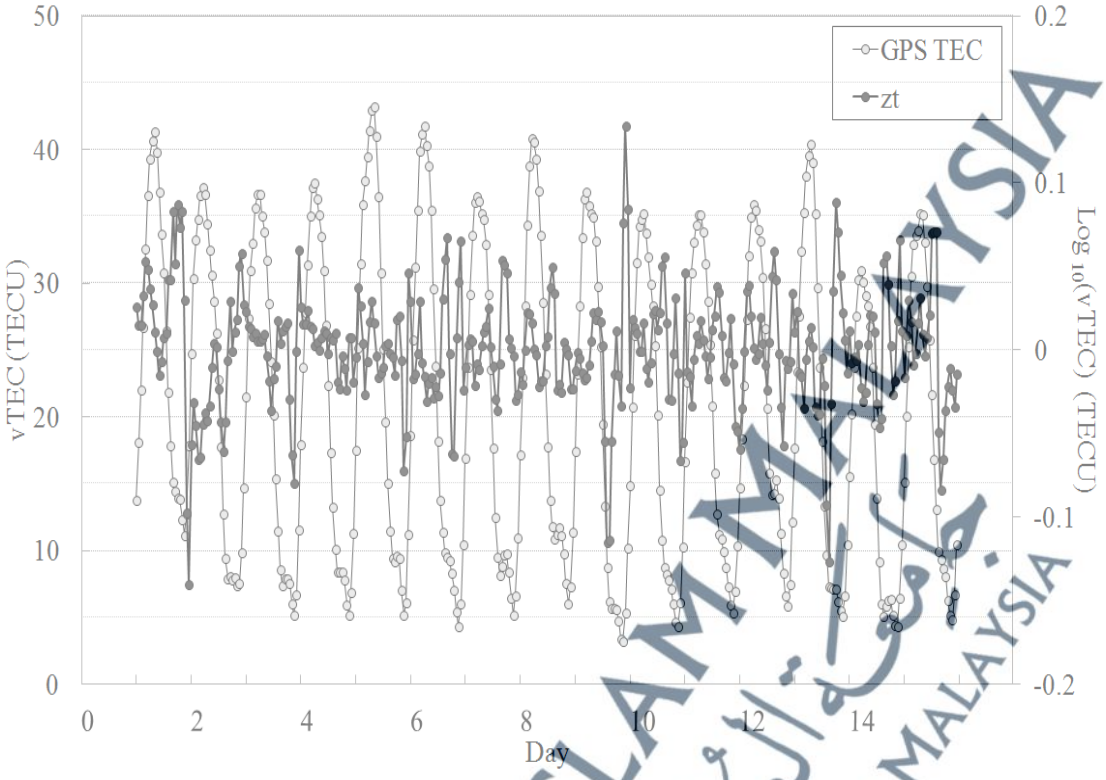
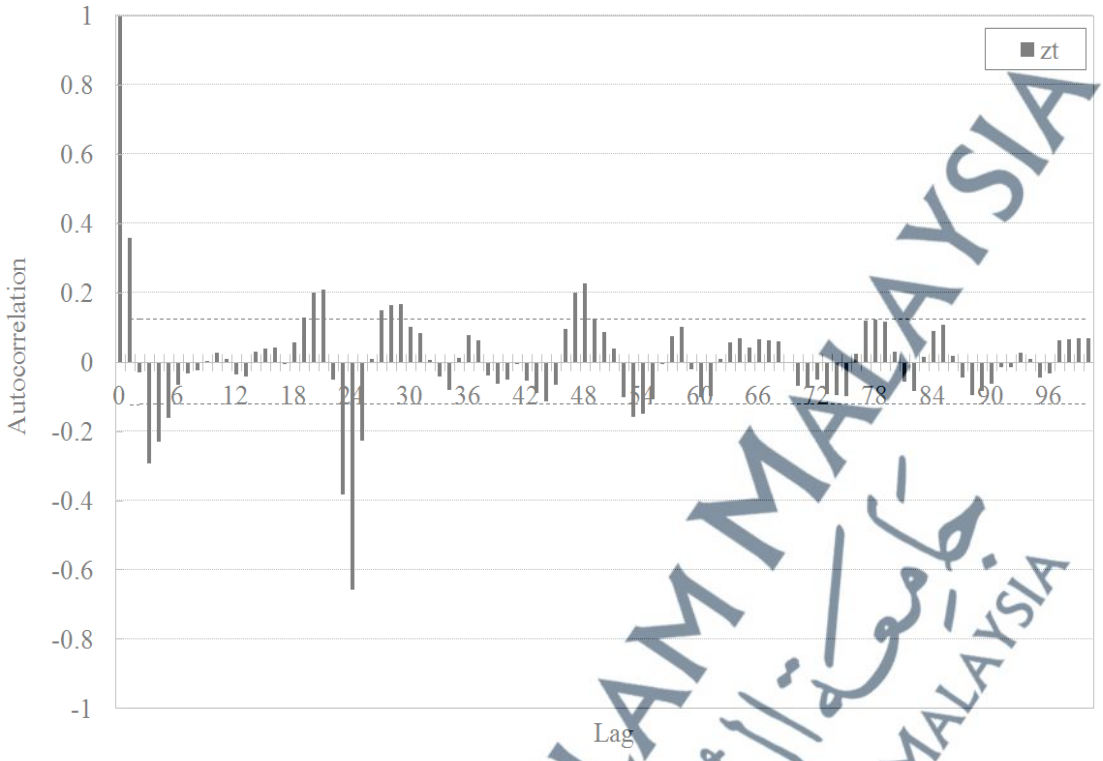


Figure 3.19: Stationary TEC time series, z_t in June 2005

lowercase notation (p,d,q) and uppercase notation (P,D,Q) denote the optimal non-seasonal and seasonal component, respectively, while s represents the length of the seasonal period . Basically in this step, the number of AR terms and number of MA terms are identified for both components; non-seasonal and seasonal by examining the stationary time series, z_t using the autocorrelation function (ACF) and partial autocorrelation function (PACF). The AR refers to the autoregressive function regressed on the values in the previous period while the MA denotes the moving average regressed on the random process. Besides, verifying the stationarity of the time series, the behaviour of ACF and PACF plots along with their theoretical properties are used to identify the plausible models and determine the appropriate p, q, P and Q to fit the stationary TEC time series, z_t.

Figure 3.20 shows the sample of ACF and PACF for z_t along with the approximate 95% confidence intervals which are denoted by the bold dotted line ("- - - -") on the both sides of the horizontal axis. The length of ACF or PACF sample is equals to 100, where it covers four full hours of the seasonal component or seasonal lag. The correlograms illustrate a mixed process where the autocorrelation shows a damped sine wave decay with large spikes at lag 1, suggesting a possible of MA(1) term while the partial autocorrelation depicts an exponential tails off toward zero with significant spikes up to 3 lags suggesting a possible of AR(3) at the non-seasonal components. The plots indicate that the stationary time series depends on past observations (referred as p terms) and past shocks (referred as q terms). In addition, both the correlograms show large spike at lag 24 which indicates the seasonal relationship for hourly data. In ACF, a single significant spike is seen at lag 24 and no any other large spikes at any other seasonal lags whereas in PACF, the pattern shows large spikes at lags 24 and 48. Since, the spike at lag 24 in ACF falls more abruptly than PACF, this may suggestive a seasonal MA(1).

A multiplicative SARIMA model is used since the time series is factored into non-seasonal and seasonal components. The "I" in the SARIMA model known as integration, determine the differencing order in the time series which is represented by d (differencing non-seasonal component) and D (differencing seasonal component). The integration process indicates that the time series data has been transformed into a stationary time series. Since the differencing process is considered during the stationary process, therefore further integration is not required at this stage. The orders of differencing for non-seasonal and seasonal components are equal to 1 and 2, respectively. In conclusion, by matching the visualization of ACF and PACF patterns with the



(a) Autocorrelogram

(b) Partial Autocorrelogram

Figure 3.20: (a) The autocorrelation function, ACF and (b) partial autocorrelation, PACF plots of the stationary TEC time series, z_t for June 2005

theoretical properties which are explained in Chapter 2, number of possible models are identified to select the optimal non-seasonal components (p,d,q) and the seasonal components (P, D,Q) of the SARIMA model. Akaike Information Criterion (AIC) described in Chapter 2 subsection 2.7.6 is used as a model selection criterion. The model that gives minimum value of AIC is considered as the best fitted model. AIC is calculated using Equation (2.30). Table 3.5 shows some possible models that fit the TEC time series with an initial suggestive model that is SARIMA (3,1,1)$(0,2,1)_{24}$ via the analysis above. However, among the tested SARIMA models, SARIMA (3,1,1)$(0,2,2)_{24}$ is found to be to the most appropriate which yielded the least AIC value. The identified (p,d,q)(P,D,Q) are substituted in the mathematical expression of SARIMA model which is expressed in Chapter 2 Equation (2.27).

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)a_t$$

$$\Phi_0(B^{24})\phi_3(B)(1-B)^1(1-B^{24})^2 y_t = \theta_1(B)\Theta_2(B^{24})a_t \quad (3.13)$$

$$(1-\phi_3 B^3)(1-B)(1-B^{24})^2 y_t = (1+\theta_1 B)(1+\Theta_2(B^{24})^2)a_t \quad (3.13a)$$

The stationary TEC time series, z_t as in Equation (3.11) is substituted into Equation (3.13a) to obtain the optimal SARIMA mathematical expression as below:

$$(1-\phi_3 B^3)z_t = (1+\theta_1 B)(1+\Theta_2(B^{24})^2)a_t \quad (3.14)$$

where z_t is the stationary TEC time series, a_t is random error at time period t, B is the backward shift operator, Θ_2 is the second coefficient of seasonal moving average, ϕ_3 is the third coefficient of non-seasonal autoregressive, and θ_1 is the first coefficient of non-seasonal moving average. The random errors or the noise components of the model, a_t are assumed to be independently and identically distributed.

Table 3.5: SARIMA models with the AIC, Ljung-Box statistic and Chi-square[i] values

No	Model	AIC	H_o	H_a	Q^*
1	SARIMA $(3,1,1)(0,2,1)_{24}$	-1199.2	1		14.474
2	SARIMA $(2,1,1)(0,2,1)_{24}$	-1165.2		1	48.954
3	SARIMA $(1,1,1)(0,2,1)_{24}$	-1169.3		1	45.166
4	SARIMA $(3,1,2)(0,2,1)_{24}$	-1174.7		1	60.579
5	SARIMA $(3,1,3)(0,2,1)_{24}$	-1176.4		1	37.419
6	SARIMA $(2,1,3)(0,2,1)_{24}$	-1186.3		1	40.240
7	SARIMA $(1,1,3)(0,2,1)_{24}$	-1201.8		1	35.718
8	**SARIMA $(3,1,1)(0,2,2)_{24}$**	**-1235.2**	**1**		**18.880**
9	SARIMA $(3,1,2)(0,2,2)_{24}$	-1203.0		1	75.641
10	SARIMA $(3,1,3)(0,2,2)_{24}$	-1209.1		1	75.641
11	SARIMA $(2,1,3)(0,2,2)_{24}$	-1217.4		1	51.520
12	SARIMA $(1,1,3)(0,2,2)_{24}$	-1230.9	1		30.497
13	SARIMA $(3,1,1)(1,2,0)_{24}$	-1085.2	1		14.0643
14	SARIMA $(3,1,1)(2,2,0)_{24}$	-992.37	1		16.430
15	SARIMA $(3,1,1)(1,2,2)_{24}$	-1132.9	1		18.964
16	SARIMA $(3,1,1)(2,2,2)_{24}$	-998.42	1		17.807

[i] Chi-square (critical test), $\chi^2 = 31.410$

H_o: The model does not exhibit lack of fit

H_a: The model exhibits lack of fit

In the second stage, the coefficients of the best fitted parameters are estimated once the orders of the SARIMA model are determined. The parameters are estimated such that the overall accuracy of the measured errors is minimized. Maximum likehood estimation (MLE) technique - based MATLAB is used to accomplish this computational. The method is recognised as a standard approach for parameter estimation especially for large samples of data (Myung, 2003). The coefficients of the best fitted SARIMA model for TEC time series are estimated and substituted in the Equation (3.14). These values may vary with respect to the TEC dataset.

$$(1-\phi_3 B^3)z_t = (1+\theta_1 B)(1+\Theta_2(B^{24})^2)a_t$$
$$(1+0.360B^3)z_t = (1+0.385B)(1-0.543(B^{24})^2)a_t \quad (3.15)$$

In the development of SARIMA model, the last step is the diagnostic checking or the residual analysis. The residuals of the fitted model are examined to verify the adequacy of the model. This test is basically to determine if the selected model assumptions about the residuals, a_t are satisfied. The residuals from the model should resemble pure random errors (white noise). Zhang (2003) and other previous works (Zhang et al., 1998; Aladag et al., 2009) mentioned that the residual from the ARIMA model only contains non-linear properties. Therefore, the diagnostic check of the residuals is very important to determine the sufficiency of the SARIMA model. Adequacy of the developed linear model (SARIMA) is considered not sufficient enough if there is still left of linear correlation structure in the residuals. A few statistical tests are suggested in the following studies (Box & Jenkins, 1976; Zhang, 2003; Zhang & Qi, 2005; Durdu, 2010; Khashei & Bijari, 2010) to examine the goodness of fit of the model.

In this analysis, with the aid of ACF correlogram, the correlation or the independence of the residuals is tested. Figure 3.21 illustrates the autocorrelogram of the residual where there is no significant correlation between the present and past values. All the values of the residuals are within the confidence intervals, except a few correlations appeared slightly larger. The residuals of the model exhibit white noise properties and proved that the overall SARIMA model is considered adequate.

Besides the ACF test, the residuals are examined using the Ljung-Box statistic (Q^*) to check the adequacy of the proposed model. The test is mainly conducted to test whether a series of data over time are random and independent. The test statistic is:

$$Q^* = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r(k)^2 \tag{3.16}$$

where n is the number of observations, K is the number of autocorrelations lags, $r(k)$ is the autocorrelation of the residual at lags k.

According to Durdu (2010), the model is considered adequate and the residuals are considered white noise if the value of Q^* statistic is less than the chi-square (critical test value), χ^2 distribution with respective degree of freedom. In this scenario, the null hypothesis, (H_0) is accepted, where the residuals of the fitted model exhibits no correlation for a fixed number of lags. Conversely, if Q^* is more than χ^2, then the alternative hypothesis (H_a) is accepted. This means the tested model is inadequate and the residuals represent a non-white noise. In this study, the Q^* statistic is calculated for all the possible models and presented in Table 3.5 to observe the adequacy of the proposed

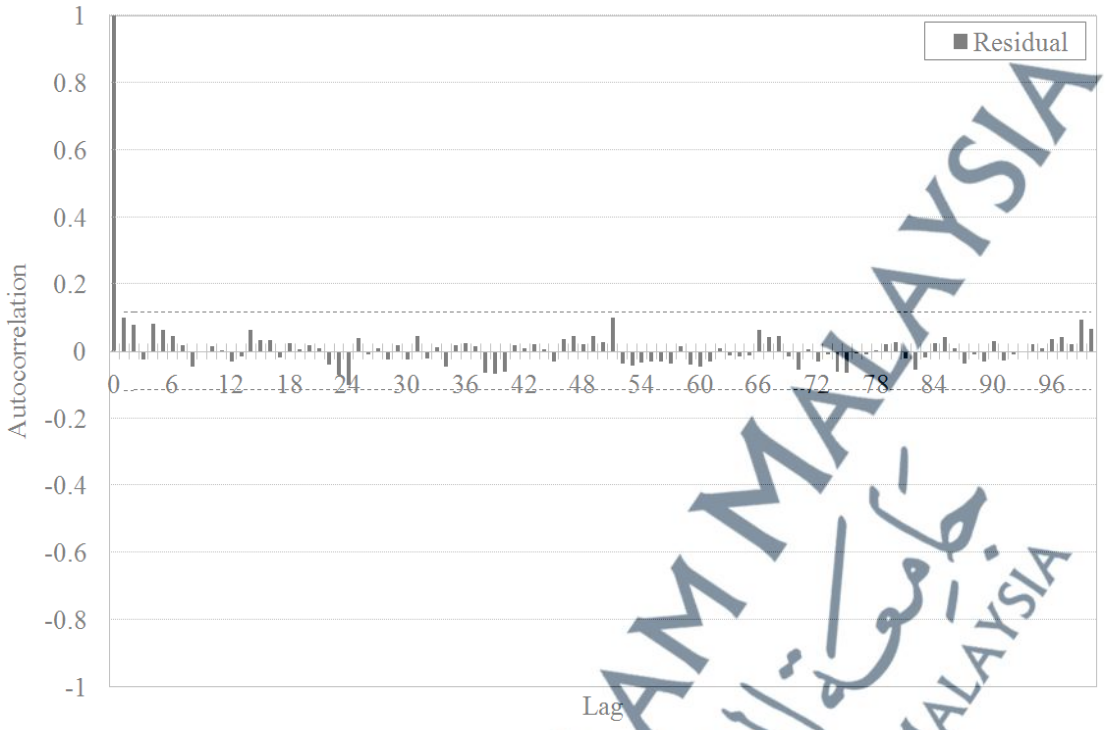


Figure 3.21: Autocorrelogram of the residual

model. The analysis shows that the value of Q^* statistic (18.880) for fitted SARIMA $(3,1,1)(0,2,2)_{24}$ model is less than the χ^2 value (31.410). The result explains the null hypothesis cannot be rejected, thus the fitted model is considered adequate.

Eventually, the SARIMA $(3,1,1)(0,2,2)_{24}$ model is chosen as the most adequate and final model to forecast the linear component ($\hat{L}_t$) of the ionospheric TEC time series. Based on the Zhang's (2003) methodology, to build up the hybrid model, the time series data can be decomposed into a linear and non-linear component as below:

$$y_t = L_t + nL_t \tag{3.17}$$

where y_t denotes the actual time series data, L_t represents the linear component and nL_t denotes the non-linear component. Both the components have to be estimated from the time series actual data, y_t. In this phase, the SARIMA $(3,1,1)(0,2,2)_{24}$ model deals with the linear component of the time series dataset and forecast the linear component ahead. The difference between the actual data, y_t and the forecast linear component from the SARIMA model, $\hat{L}_t$ is known as the residuals. That is:

$$e_t = y_t - \hat{L}_t \tag{3.18}$$

where e_t denotes as the residual at time t from the linear SARIMA model.

In the second phase the residuals are used in neural network to forecast the non-linear components ($n\hat{L}_t$). The NN model that deals with the non-linear component of the dataset is discussed thoroughly in the following section.

3.5.2 Modelling of neural network

The diagnostic check of the residuals in the previous section can be only used to examine the goodness of fitting of the linear model but the analysis is unable to capture the non-linear patterns in the time series under study. In other words, even though the model has passed the diagnostic checking, but the model is still lack off in the non-linear relationship which have not been modelled. This shows the major limitation of the SARIMA model where no non-linear patterns are captured by the model.

Thus, in this phase, the residuals obtained from the SARIMA model are modelled using the NN technique to analyse the non-linear relationship. The NN model is

developed to forecast the non-linear components, since theoretically and practically in many works (Fausett, 1994; Bishop, 1995; Hagan et al., 1996; Hernandez-Pajares et al., 1997; Conway et al., 1998; Zhang et al., 1998; Tulunay et al., 2006; Habarulema et al., 2007; Maruyama, 2007), it is proven that the NN model is able to learn and generalize any complex environment with high accuracy. The concept of NN model was discussed thoroughly in section 3.4 and in Chapter 2. Therefore only the final structure of the model is discussed in this section. A three - layers feed forward network associated with Levenberg - Marquardt (LM) back-propagation algorithm is used to forecast the residuals ahead. In this work, the lagged residuals represent the input nodes while the forecast values are obtained from the output node. The NN model is presented as a function of the past observations of the residuals, where the model can be expressed as:

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-N}) + \varepsilon_t \tag{3.19}$$

where f is the non-linear function determined by the neural network structure, e_t denotes as the residual at time t, ε_t is the random error, and N is the number of lagged residuals in the NN.

The residuals over the period of 17 months (February 2005 to June 2006) are used for training, 3 months (July 2006 to September 2006) and 3 months (October to December 2006) data are used for validation and testing purposes, respectively. All the corresponding residuals are normalised to the range of 0 to 1 so that the maximum and minimum value of any residuals does not exceed the range values. The determination of an optimal number of input neurons and hidden layer neurons are always a complex task in NN modelling which contributes the most to obtain a precise output. The residuals

obtained from the SARIMA model exhibit white noise properties where the residuals are independent of each other. Hence, it is quite complex and difficult to identify the appropriate lagged observations that could contribute the best model. A comparative analysis of the residuals denoted by the hourly lags from 1 to 24 is executed to obtain the optimal number of input nodes. The seasonal lags more than 24, which are 48, 72, 96, 120, 144, 168, etc. values are not taken into consideration since the residuals are assumed to be deseasonalised data and free from the linear correlation structure.

RMSE is used to monitor the performance of the network and to determine the optimum input parameters that capable to forecast the residuals ahead with least error. Each of the architecture is run ten trials individually with random initialization of weights and biases to attain the best performance criteria and the results are used for comparison. MSE is used as a stopping criterion whereby the training is halted when the mse of the validation dataset starts increasing. In the hidden and output layers, the log-sigmoid is used as an activation function. The best fitted model is selected based on the validation sample result.

Table 3.6 summarised the RMSE values obtained from the validation sample of each pattern. In this thesis, various lag numbers are considered, with minimum two and maximum up to 12 different numbers of lags. This may be due to the fact that there is no specific lagged observation that could contribute the most to the random residual time series. Several observations can be made from the summary of the table. Overall, for all patterns, the RMSE values only differ slightly from each pattern. First, as the lagged patterns reached the maximum or minimum lag numbers, the NN model unable to forecast the residuals well. This may due to inappropriate and insufficient number of

Table 3.6: Combinations of residual lag numbers

Pattern	Lag values	RMSE (TECU) (10^{-2})
1	$y_{t-2}, y_{t-4}, y_{t-6}, y_{t-8}, y_{t-10}, y_{t-12}, y_{t-14}, y_{t-16}, y_{t-18}, y_{t-20}, y_{t-22}, y_{t-24}$	1.693
2	$y_{t-1}, y_{t-3}, y_{t-5}, y_{t-7}, y_{t-9}, y_{t-11}, y_{t-13}, y_{t-15}, y_{t-17}, y_{t-19}, y_{t-21}, y_{t-23}$	1.677
3	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8}, y_{t-9}, y_{t-10}, y_{t-11}, y_{t-12}$	1.611
4	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8}, y_{t-9}, y_{t-10}, y_{t-11}$	1.574
5	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8}, y_{t-9}, y_{t-10}$	1.572
6	$\boldsymbol{y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8}, y_{t-9}}$	**1.561**
7	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8}$	1.569
8	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}$	1.576
9	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}$	1.577
10	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}$	1.584
11	$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}$	1.588
12	$y_{t-1}, y_{t-2}, y_{t-3}$	1.590
13	y_{t-1}, y_{t-2}	1.620
14	$y_{t-3}, y_{t-6}, y_{t-9}, y_{t-12}, y_{t-15}, y_{t-18}, y_{t-21}, y_{t-24}$	1.601
15	$y_{t-4}, y_{t-8}, y_{t-12}, y_{t-16}, y_{t-20}, y_{t-24}$	1.602
16	$y_{t-6}, y_{t-12}, y_{t-18}, y_{t-24}$	1.598
17	y_{t-12}, y_{t-24}	1.638

inputs may affect the performance of the NN model. In both case, the RMSE values are above 1.6 x 10^{-2}. Second, the number of lags which varies from 1 to 24 or 2 to 24 with certain numbers of increment e.g., 2, 3, 4, 6, and 12 even deteriorate the forecasting results. These types of lagged inputs generate RMSE values between the range ~1.6 x 10^{-}

^{2} and ~1.7 x 10^{-2}. It is obvious from the Table 3.6 that the consecutive lagged patterns performed better compared to the cases in which lagged patterns with specific increment is used solely. This indicates that applying consecutive lagged patterns is more approachable in modelling and forecasting the random time series. Among the consecutive lagged numbers, the pattern with 9 consecutive lags yielded the least RMSE value in forecasting the random residuals data. The least RMSE result is bolded in the Table 3.6. In conclusion, consecutive of 9 lagged inputs are considered as the best combination of residuals to improve the forecasting accuracy. The NN residual is

$$NN\ residual(e_{t+71}) = f(e_{t-1}, e_{t-2}, e_{t-3}, \dots, e_{t-8}, e_{t-9}) + \varepsilon_t \quad (3.20)$$

where (e_t) is the residual at the present hour t and the output (e_{t+71}) represents the residual values for 72 hours ahead. The optimal hidden nodes are identified as in the subsection 3.4.2 by varying the number of hidden nodes. The best network is attained by adding and reducing one neuron at a time. The architecture that yielded the least error is adopted as the optimal hidden neurons. In this analysis, the optimal NN configuration used to forecast the residuals is 9:7:1, where 9 input nodes to represent the lagged observations, 1 output node to depict the forecast residuals values and 7 nodes in the hidden layer. It is clearly noticeable that the number of hidden nodes vary for both analysis; TEC estimation and TEC forecasting, in this thesis. In conclusion, determining the optimal number of hidden neurons is still an open issue, seem the numbers may vary with the applications. The 3 months testing data are excluded from the training process to prevent memorization.

Finally, the forecast linear component $(\hat{L}_t)$ from the SARIMA and the forecast

non-linear component ($n\hat{L}_t$) from the NN models are combined in the following hybrid section to forecast the ionospheric TEC ahead.

3.5.3 Optimum Hybrid SARIMA-NN model

The hybrid modelling consists of three phases. In the first phase, the SARIMA model is used to analyse and forecast the linear component of the TEC time series. In the second phase, residuals which represent the non-linear component of the TEC time series data are modelled based on the NN to analyse the non-linear relationship. In the last phase, both the forecast components are integrated to obtain the final forecast value. The final output of hybrid SARIMA-NN model can be represented as in Equation (3.21):

$$\text{hybrid SARIMA} - \text{NN TEC} = \hat{L}_t + n\hat{L}_t \tag{3.21}$$

where hybrid SARIMA-NN TEC represent the forecast TEC values, $\hat{L}_t$ represents the forecast linear component and $n\hat{L}_t$ denotes the forecast non-linear component.

The observed TEC along with the estimated TEC (NN TEC) values, "without the gaps" are used in the forecasting model to avoid degradation in the model performance. In this feasibility study, 15 days prior TEC data is utilised to forecast the ionospheric TEC ahead with respect to the result in Chapter 4 subsection 4.2.2. The result shows that the NN2 model is able to estimate or extrapolate the missing TEC data up to 18 days (60% of the overall 30 days data). The performance of NN is model deteriorates for further missing days, where the value of the relative correction (Crel) starts reducing while the RMSE value starts increasing. Since15 days TEC data are still within the

estimation period of the NN2 model, this period is taken into consideration for TEC forecasting model. The development of hybrid SARIMA - NN model is tested using 15 days prior hourly TEC data to forecast the TEC data 3 days ahead (72 points hourly).

To validate the hybrid SARIMA-NN model, the forecast TEC from the hybrid model is compared against the actual values as well as, with the forecast values of the individual models SARIMA and NN, used separately. Both the individual models are described thoroughly in the following section.

3.5.4 Single model techniques

To verify and validate the performance of the hybrid SARIMA-NN model, both the techniques integrated in the hybrid models, SARIMA and NN are developed individually and separately to compare with the developed hybrid SARIMA-NN model. The concept of the SARIMA model was explained thoroughly in the previous subsection 3.5.1 and the developed SARIMA $(3,1,1)(0,2,2)_{24}$ is adopted to represent as the individual SARIMA model to forecast the ionospheric TEC variability. The model is known as forecast SARIMA (FCAST-SARIMA).

In this section the individual NN model used to forecast the TEC is explained briefly. A single hidden layered feed forward network with Lenvenberg- Marquardt back propagation algorithm is used. This NN model has one input layer, one hidden layer and one output layer. The NN model is known as forecast NN (FCAST-NN). The FCAST-NN is represented as a function of 9 consecutive hourly lagged TEC values adopted from the method in subsection 3.5.2 and also combined with the inputs adopted from Tulunay

et al. (2004). Table 3.7 shows the inputs used in NN forecasting TEC technique, where the t values in Table 3.7 represent hour(s) and the value of TEC at the present hour is designated by TEC_t. The FCAST-NN configuration is determined as 12:10:1, where 12 input nodes in the input layer, 10 hidden nodes in the hidden layer and 1 output in the output layer. The output node produces TEC values 3 days or 72 hours in advance.

Table 3.7: Input parameters of the FCAST-NN model

No	Input Parameters	Notation
1	9 consecutive hourly lagged TEC values	$\mathrm{TEC}_{t-1}, \mathrm{TEC}_{t-2}, \mathrm{TEC}_{t-3}, \mathrm{TEC}_{t-4}, \mathrm{TEC}_{t-5}, \mathrm{TEC}_{t-6}, \mathrm{TEC}_{t-7}, \mathrm{TEC}_{t-8}, \mathrm{TEC}_{t-9}$
2	First difference	$\Delta 1_t = \mathrm{TEC}_t - \mathrm{TEC}_{t-1}$
3	Second difference	$\Delta 2_t = \Delta 1_t - \Delta 1_{t-1}$
4	Relative difference	$R\Delta_t = \Delta 1_t / \mathrm{TEC}_t$

Both the individual models, FCAST-SARIMA and FCAST-NN are developed based on the same hourly TEC dataset as in hybrid SARIMA-NN model. Finally, to analyse the performance of the hybrid SARIMA-NN and the individual models, each the models are tested using the identical hourly TEC data over Parit Raja station. The comparison covers three different periods; quiet, moderate, and disturbed conditions, which is from October 2006 to December 2006. The comparison results between the hybrid SARIMA-NN, FCAST-SARIMA and FCAST-NN are presented in chapter 4.

3.6 ASSESSMENT OF RESULTS

Accuracy of a model can be quantified in terms of errors. The discrepancies between the observed and estimated values are indicated as the errors of the model. The predictability of the model is inversely proportional to the error from the model. To assess the predictability of the model, the results are grouped in order to compare between the observed TEC and estimated TEC. The predictability of the forecast model developed in the latter section is also evaluated via these methods. The comparison is provided in terms of absolute and relative errors. The absolute error (Eabs) is computed according to:

$$Eabs = Est_{TEC} - Obs_{TEC} \tag{3.22}$$

where the observed TEC (Obs_{TEC}) is extracted from the GPS measurement and the estimated TEC (Est_{TEC}) measurement is from NN2 and IRI-07 models (Leandro & Santos, 2007. In forecasting model the Est_{TEC} is replaced by F_{TEC} which is obtained from hybrid SARIMA-NN, SARIMA, and NN models. The Eabs is expressed in units of TEC (TECU).

The relative error (Erel) is expressed in terms of percentage where it is defined as the ratio between absolute errors and observed TEC. This error is calculated as follows:

$$Erel = \left(\frac{Eabs}{Obs_{TEC}}\right) \times 100\% \tag{3.23}$$

Finally the effectiveness of the estimation or the forecasting model is quantified in terms of relative correction (Crel) which is defined as follows:

$$Crel = (100 - Erel)\% \tag{3.24}$$

The Crel indicates the estimation accuracy for the NN and IRI model, as well as the forecasting accuracy of the hybrid and individual models. The higher the relative correction, the closer the estimated values to the observed values. A Crel of 100 percent indicates an optimum estimation while Crel approximately zero percentage signifies the model is no more efficient enough.

Other than the absolute and relative errors, the accuracy of the estimation and forecasting models are also quantified in terms of normalized RMSE which used to avoid the effects of large TEC background during comparison. According to Watthanasangmechai et al. (2012), the normalized RMSE is computed as follows:

$$normalized\ RMSE = \frac{RMSE}{Average\ Obs_{TEC}} \tag{3.25}$$

where RMSE and Average Obs_{TEC} is the root mean square error and the average GPS TEC, respectively. In Chapter 4, the suitability and the effectiveness of the constructed NN as well as the forecasting hybrid SARIMA-NN models are investigated with respect to the following assessment. The NN model is tested diurnally and seasonally whereas the hybrid SARIMA-NN model is tested during quiet, moderate and disturbed conditions.

3.7 SUMMARY

The chapter explains the procedure of extracting the GPS TEC data from the GSV4004B receiver. The minimization technique used in estimating the hardware bias in a single station receiver and the mapping function used in TEC conversion from slant TEC to vertical TEC are described in this chapter. Then, the chapter gives a comprehensive discussion on the construction of Neural Network (NN) to estimate the TEC variations and hybrid Seasonal Auto Regressive Integrated Moving Average - Neural Network (hybrid SARIMA-NN) model to forecast the ionospheric TEC variations over Parit Raja, Malaysia. Neural Network (NN) is a well known data driven model used to develop an ionospheric TEC modelling. The optimum parameters that influenced the TEC variability are identified. They are the daily and 27 backward means of SSN and $S_{10.7}$, the hourly planetary amplitude (ap) and together with the diurnal (HR) and seasonal (DN) components. In addition, the optimum NN configuration is determined, where the NN model uses 9 numbers of nodes in input layer, 11 nodes in single hidden layer and 1 node in output layer designated as 9:11:1. In this thesis, a single hidden layered feed forward network with a back-propagation algorithm is used. In order to access the performance accuracy of the learning methods in ionospheric TEC estimation, a number of training algorithms are investigated and found that Levenberg –Marquardt (LM) algorithm is the best optimization method for TEC modelling. This method differs significantly in terms of RMSE and time required to achieve convergence during training. With all the criteria above the optimum NN architecture is constructed to estimate accurate TEC values with least bias.

Continuously, this work is further extended using the recovery GPS TEC, where a

time series forecasting model is developed using a hybrid technique that combines seasonal autoregressive integrated moving average (SARIMA) and neural network (NN). The SARIMA$(3,1,1)(0,2,2)_{24}$ and NN with 9 consecutive lagged residual values with the optimum configuration 9:7:1 are used to forecast the linear and non-linear components, respectively. Eventually both the components are integrated to obtain the forecast TEC values. The hybrid SARIMA-NN model utilized 15 days prior TEC data to forecast the ionospheric TEC 3-day ahead.