CHAPTER 2

LITERATURE RIVIEW

2.1 INTRODUCTION

The chapter introduces to the prior knowledge of the variation and distribution of ionosphere physics. After some introductory to ionosphere, this particular chapter is extended on the ionospheric TEC, which is one of the most important parameters to characterize the ionosphere properties. Furthermore, the historical works on TEC estimation and forecasting models are briefly described. Theoretical descriptions and the algorithms of the Neural Network (NN) and SARIMA methods are discussed in detail. This chapter provides further explanation on the GPS and the refraction theories of electromagnetic waves propagation through the ionosphere. At the end of this chapter, it gives a brief description on the algorithm or method that is used to derive the TEC from the GPS measurements based on the two L-band frequencies, prior to the development of TEC modelling over Parit Raja, Malaysia.

2.2 EARTH'S IONOSPHERE

The name "ionosphere" was superseded from the older term "conducting layer" or the "Kennelly-Heaviside Layer". Arthur E.Kennely and Oliver Heaviside

predicted the existence of the free electric charges in the upper atmosphere that enable radio waves reflection. In 1924, Edward V.Appleton proved the existence of the reflecting layers from the “frequency change” through experiments. The ionosphere is an ionized region of the Earth's atmosphere, where it is formed by the photoionization of atoms and molecules. The ionization is mainly owed by extreme ultraviolet (EUV) and X-ray radiation from the sun that able to ionize one or more atmospheric constituents (Rishbeth & Garriott, 1969). Besides the photoionization, the corpuscular ionization produced by precipitation of energetic charged particles of magnetospheric, solar or cosmic origin and by burst of solar EUV and X-ray radiation ,which enter the earth's atmosphere at high latitudes is also a source of ionization in the atmosphere.

Ionospheric layered structure is organized by the density number of the plasma with altitude variations. Figure 2.1 exhibits the distinct layers denoted by D, E, F1 and F2 layers. It even shows the dominant ion density profiles with respect to the heights. The existence of each layer depends on the solar spectrum energy at various altitudes relying on the absorption of atmosphere, different physical processes and altitudes variations in the atmospheric neutral composition. The production of ionization through absorption of solar photon radiation and the destruction of ionization through recombination process of electron and atomic ions are the main physical processes that control the ionosphere. Mostly these processes are dominant in the lower ionosphere, the D and E regions which are known as “photochemical” processes.

The D-region is the lowest part of the ionosphere, which has an altitude range of ~50 to ~90 km above the Earth's surface. The recombination process of ions and electron is relatively quick at D, which causes the layer has low electron density compared to

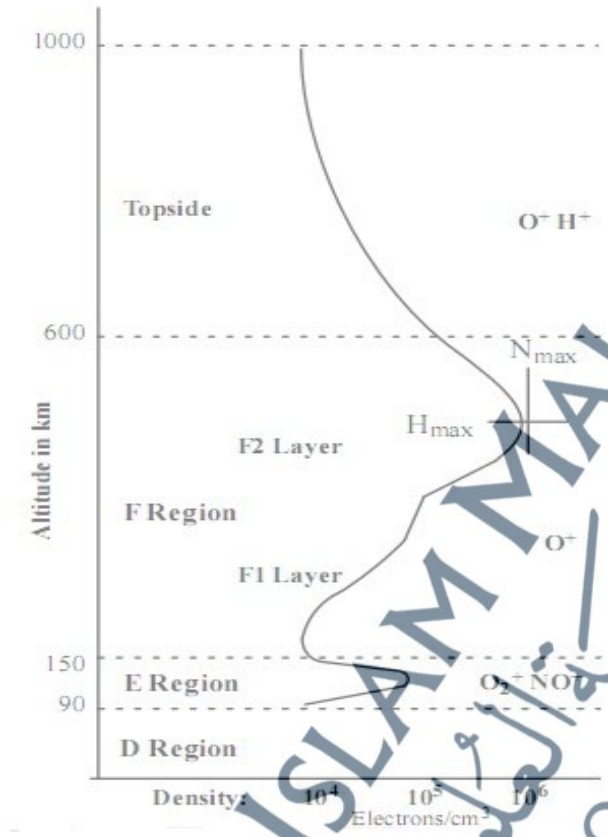

Figure 2.1: Dominant ion population and ionospheric plasma density with various layers (http://www.swpc.noaa.gov/info/Iono.pdf, August 2013).

other layers in the ionosphere. Therefore after the sunset, the ionization level reduces significantly and causes the layers essentially disappear in the night.

The layer above the D-region with an altitude range of ~90 to ~150 km is referred as the E-region. This region is mainly dominated by oxygen ions which are formed by soft X-rays and UV solar radiation of the sun dissociating the molecular oxygen (Baumjohann & Treumann, 1997).

The F region with an altitude range of ~150 to 500 km is often subdivided into two sub-layers during the day denoted by F1 - layer and F2 - layer, whereas in the night

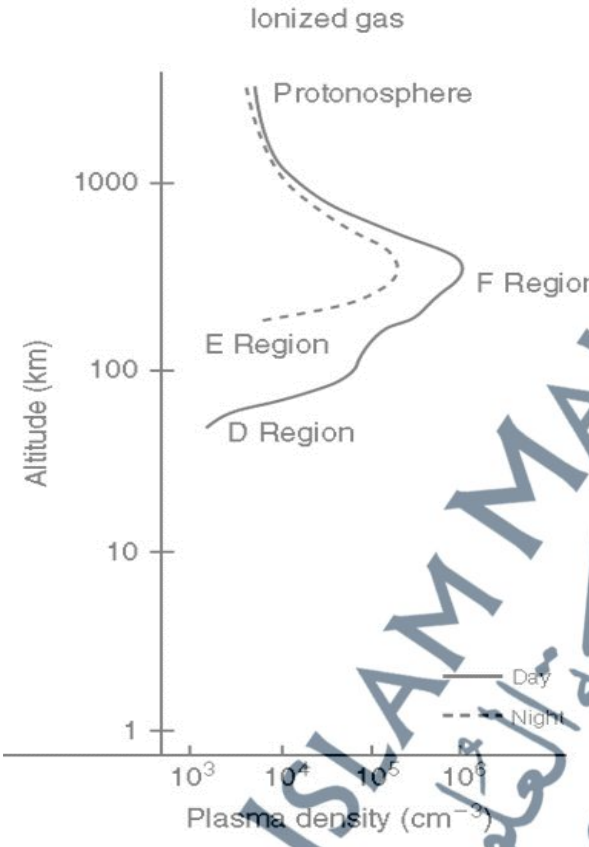

Figure 2.2: Profiles of plasma density during day and night (Kelly, 2009)

the two sub-layers merge back into a single F layer. The maximum electron density occurs at this region. The photochemical processes are still dominant in the F1 layer, the lowest layer of the F region while the F2 layer is the uppermost layer of the F region, lies at a level of transition between the photochemical and "transport" processes such as diffusion. Plasma diffusion is the dominant process in controlling the electron density distribution in the F2 layer. The recombination process is much slower, thus the ionization still persists till overnight (Rishbeth & Setty, 1961; Rishbeth & Garriott, 1969). Figure 2.2 shows typical profiles of plasma density during day and night. The disappearance of the lower layers and mergence of the F1 and F2 layers are clearly seen in this figure.

2.3 IONOSPHERIC VARIATIONS

The factors that describe the ionospheric variability have been extensively studied around the global (Rishbeth & Setty, 1961; Alex, 1987; Davies, 1990; Balan et al., 1996; Loewe & Prolss, 1997; Abdu et al., 2008; Bagiya et al., 2009; Aggarwal, 2011; Liu et al., 2011; Aggarwal et al., 2012; Richa et al., 2013). The results concluded that the major factors that influence the ionospheric variability are always related with the sun and its activities under different space weather conditions. The factors are as follow:

i. Solar activity

Solar cycle variation, also known as "sunspot cycle", is the primary factor that influences the ionospheric variability. It describes the periodic changes in the sun's activities with an average duration of ~11 years. The period of maximum and minimum sunspot counts is referred as high and low solar activity variations respectively. Moreover, the increase in photoionization processes during high solar activity will yield more variations in the ion, electron densities, temperatures, neutral winds, and electric fields in the ionosphere due to high intensity of solar electromagnetic radiation. Conversely, changes in the intensity of sun radiation during low solar activity reduce the ionospheric variability.

ii. Geomagnetic activity

Apart from solar activity, other events e.g. geomagnetic storms, ionospheric and solar storms affect the ionospheric variations. Geomagnetic storm is defined as a large and persistent perturbation of the Earth's magnetosphere. 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. These disturbance events have a profound influence on the Earth’s upper atmosphere, where the ionosphere exhibits complex and unpredictable behaviours resulted from the intensity of the storms.

iii. Diurnal variation

Diurnal variation of the ionosphere mainly describes the variation of electron density and ions in the ionosphere throughout a day (day to night). The variations mainly depend on the rotation of the Earth about it's axis with respect to the sun. The photoionization and decay of ionization processes control the variations of electron density in the ionosphere during day and night, respectively.

iv. Seasonal variation

The ionosphere also varies with seasons where the ionosphere's seasonal variation is the result of the Earth revolving around the sun. The solar radiation and solar zenith angle are the factors that contribute to the seasonal anomaly (Wu et al., 2004), which cause the enhancement or depletion of electron densities in the ionosphere. For example at equator, the sub-solar point or the sun is perceived to be directly overhead the equator during equinoxes. Hence, the incoming solar radiation is able to control the electron population and can produce more electrons during both solar maximum and minimum activities. However, during the solstices, the sub-solar point moves toward higher latitudes in the hemispheres. As a result, the photoelectron at the equator decreases and the effect reduces electron density in the solstices. Apart from the sub-solar point, the change of the neutral atmosphere's composition and direction of the neutral winds

are other possible mechanisms for causing seasonal variations (Rishbeth & Setty, 1961; Liu et al., 2013; Bhuyan & Borah, 2007).

v. Geographical variation

The latitude variations of the ionosphere show a significant distinctive characteristic. This is due to the physical processes associated with different geomagnetic field lines configurations governed by the ionospheric plasma especially in the equatorial region. The electron densities are expected to be higher over the equatorial region compared to the mid- and high latitude regions.

2.4 IONOSPHERIC TOTAL ELECTRON CONTENT (TEC)

The space plasma variability and irregularities effects on modern applications that use radio waves, for instance satellite communication, positioning, navigation and radar systems. The high frequency electromagnetic signals emitted by the GPS satellites, reflected from or propagated through the upper atmosphere (highly variable propagation medium) are highly affected. These signals interact with large number of free electrons and positively charged "ions" in the ionosphere, resulting in modifications of the signals like amplitude and phase fluctuations, rapid changes in propagation speed as well as the direction of the signals (Radicella & Tulunay, 2004; Jakowski et al., 2004; Watthanasangmechai, 2011). The ionospheric effects on the radio wave technologies are depicted in Figure 2.3.

Prominent changes in the electron density and varying population of electron can induce severe ionospheric disturbances and may give a significant impact on any

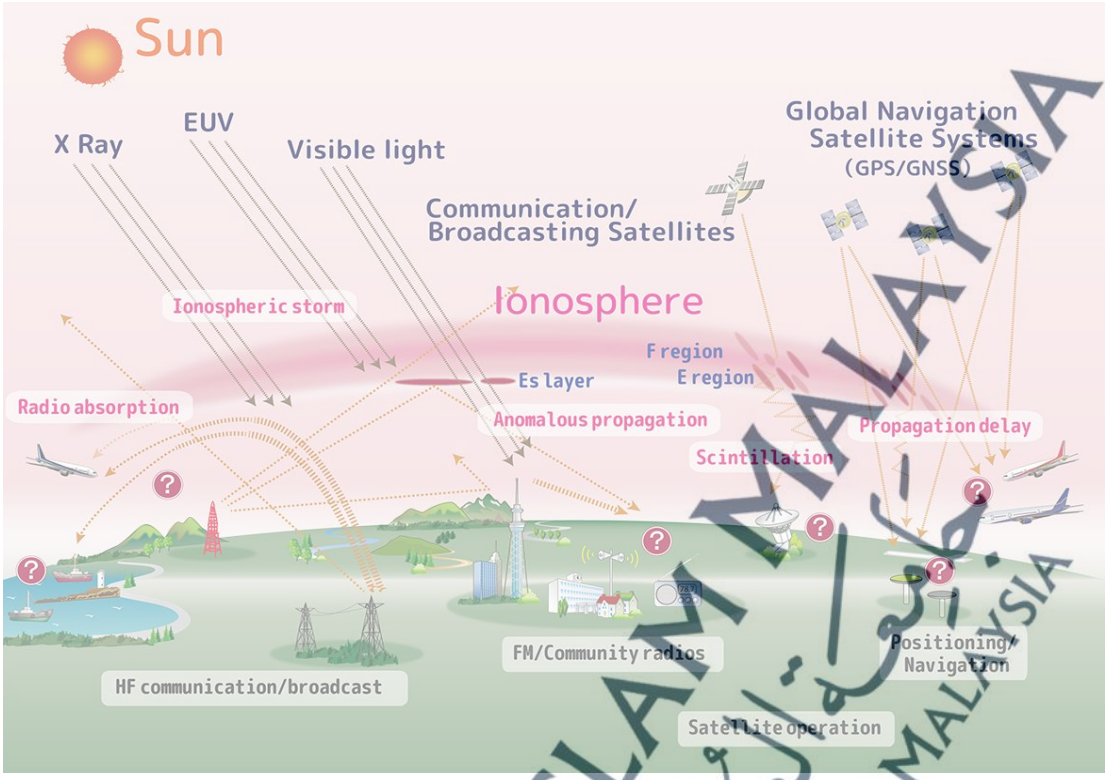

Figure 2.3: Ionospheric effects on the radio wave technologies (Watthanasangmechai, 2011)

operating system which involves radio waves that transverse the ionosphere. Due to the inhomogeneity, the Earth's ionosphere is referred as a dispersive medium, which causes the refractive index in the medium changes along the signal's path, resulting the propagation speed of the radio waves throughout the path to be inconstant. Thus, the GPS signals that travel through this medium, refracted or bent from the shortest path (a prefect straight line between satellite and receiver at ground station), cause delay in the signal propagation as depicted in Figure 2.4. The ionospheric time delay encountered by the radio wave is found to be the largest effect on the GPS signal. In addition, increasing in the propagation delay may influence the performance of high precision satellite positioning, navigation and surveillance systems (Watthanasangmechai, 2011) as well as degrade the position accuracy and affects the signal integrity (Jakowski et al., 2004).

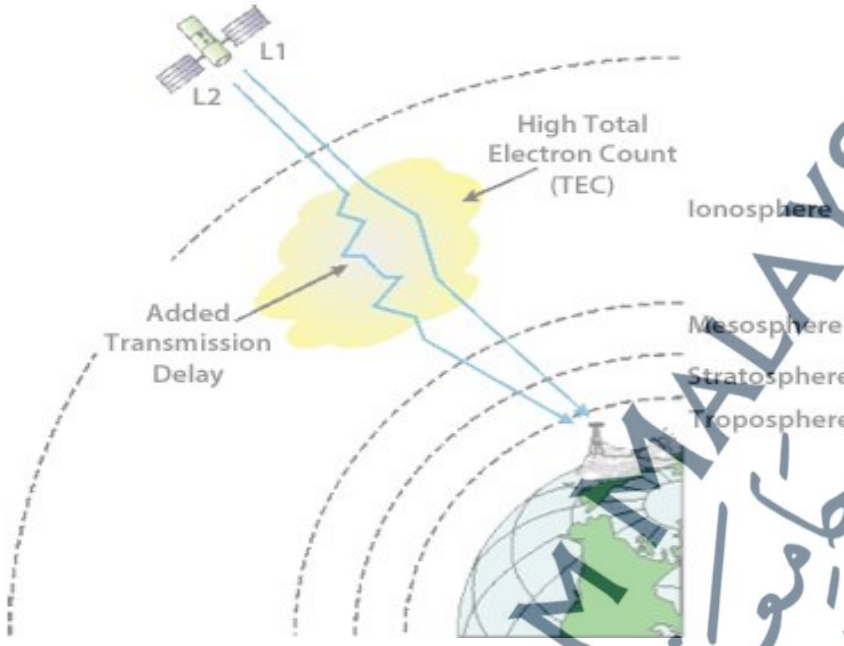

Figure 2.4: Propagation of radio waves is disrupted by the population of electron in the Earth's ionosphere (http://reflexions.ulg.ac.be/cms/c_358355/en/the-erroneous-gps-signal, February 2014)

Total electron content (TEC) is an important ionospheric parameter that affects the radio waves that propagate through this inhomogeneity region. The ionospheric TEC is defined as the total number of electrons in a column of one square meter ($1m^2$) in area along the propagation path of the radio waves between the satellite and receiver. It is a key parameter of ionosphere used to describe the ionosphere electron density and ionization rate. Furthermore, the time delay of the electromagnetic radio waves is in proportion to the varying number of free electron and inversely in proportion to the square of its frequency. Thus, in order to correct the ionosphere's time delay error, in-depth knowledge in TEC is required since TEC modelling is important for the estimation of the ionospheric time delay in the applications involving GPS (Sarma & Madhu, 2005).

Historically, the ionospheric TEC studies were widely studied with the aid of

measurements obtained from different observing techniques. Monitoring the polarization twist imposed on VHF radio waves by the ionosphere using a lunar technique was the earliest method used to study and determine TEC (Klobuchar, 1975). TEC is obtained to describe the first temporal and spatial behaviour of the topside of the ionosphere (Klobuchar, 1989). Furthermore, in the satellite era, the advent of radio beacon on board artificial earth satellites, has led to a new method to measure TEC from the VHF radio waves transmitted from the satellites. Few techniques are used to derive the ionospheric TEC i.e. Faraday polarization rotation, group delay and differential carrier phase (Klobuchar, 1989; Davies, 1990; Bhuyan et al., 2004). Other than the aforementioned techniques, in the recent decades, the number of operating Global Positioning System (GPS) receivers is growing extensively. Owing to its large population and continuous operation, GPS has served as a valid and powerful instrument to study and determine the ionospheric TEC. Knowing to its validity, many studies have been carried out in different locations around globe to investigate TEC (Komjathy, 1997; Bhuyan & Borah, 2007; Abdullah et al., 2009; Bagiya et al., 2009; Mukherjee et al., 2010; Purohit et al., 2011; Adewale et al., 2012). In this thesis, the measurements of TEC obtained from the dual frequency GPS receiver are investigated and the derivation of TEC from the GPS data is explained thoroughly in subsection 2.8.3. Beside satellites, vertical incidence ionosonde technique is also used to measure the vertical ionospheric electron content. This technique is able to provide an ionogram derived electron content, which gives information to compute the vertical electron density profile up to the F2 layer peak (Alex, 1987; Reinisch & Huang, 2001; Mosert et al., 2007; Mushini et al., 2009).

TEC is measured in Total Electron Content Unit (TECU), with 1 TECU equivalent to 10^{16} electrons meter per square area (10^{16} electrons/m^2)(Kersley et al.,

2004). The nominal range of TEC value is from 10^{16}(minima) to 10^{19}(maxima) and the value of TEC depends on many variables, including long and short term changes in solar ionizing flux (solar activity), magnetic activity, season, time of day, and viewing direction (elevation and azimuth of the satellite). The electron population mainly builds up when direct sunlight interacts with the upper atmosphere. Besides temporal variation, ionospheric TEC values show distinctive characteristics at different geographical locations (polar, mid-latitude and equatorial regions). Among the regions, the equatorial ionosphere is considered to induce the high range error in positioning and navigation systems, due to large electron density irregularities and complex variation in this region(Wu et al., 2004; Bagiya et al., 2009; Liu et al., 2011; Xu et al., 2012; Liu et al., 2013).

The equatorial ionosphere differs significantly in the characteristics and dynamic structures compared to other latitudes. Equatorial Ionization Anomaly (abbreviated as EIA) or "Appleton anomaly" is an interesting phenomenon in this region. The anomaly is characterized by two ionization crests on the either side of the magnetic equator at about ± 15˚ magnetic latitude and reduced electron concentration (trough) at the magnetic equator. For instance, EIA is produced by the "fountain effect" which has been explained in the electrodynamics drift and diffusion theories (Rishbeth & Garriott, 1969; Sethia et al., 1979). According to the drift theory, the east-west electric field perpendicular with the north-south geomagnetic field at the magnetic equator produces a plasma fountain which lifts the plasma from the magnetic equator to higher altitudes. Once it has been lifted to a certain height, the drifting plasma loses its momentum and diffuses along the magnetic field lines, due to the combined influence of pressure gradient forces and gravity. It also forms the crests at the higher latitudes from the magnetic equator (Bhuyan

& Borah, 2007; Kumar & Singh, 2009; Aggarwal et al., 2012). Thus, this region is subjected to large spatio-temporal variations that severely affect the accuracy of the radio signals (Obrou et al., 2009). Since Malaysia lies in the EIA region, it is important to study and investigate TEC as it is more prone to the anomalies.

Knowing the importance of TEC, many space researchers have shown interest in the ionospheric study. These groups have started to comprehend the behaviour of the ionospheric TEC at various locations around the globe. A number of prediction or estimation TEC models were developed regionally and globally to provide valuable information on the ionization level of the ionosphere. The results from the predicted model are compared between the observed measurement and the existing global models. Apart from the estimation models, a number of approaches were introduced to develop the ionospheric TEC forecasting models to forecast the ionospheric state in advance, which is a crucial requirement in space-borne and ground-based technological systems as well as on human life and health. These ionospheric TEC models are briefly described in the following section.

2.5 ESTIMATION AND FORECASTING IONOSPHERIC TEC MODELS

The ionospheric space community and engineering groups are endeavouring to increase the ionospheric TEC models (single station, regional and global) for estimation and forecasting at different locations and time intervals with greater accuracy. In this thesis, two categories of ionospheric TEC models will be discussed. In subsection 2.5.1, a number of ionospheric TEC estimation models used to estimate the TEC will be explained while in the following subsection 2.5.2, efforts that have been made to develop

a number of forecast model to forecast the ionospheric TEC values in advance (short- and long- term) will be described.

2.5.1 Models for estimating the ionospheric TEC

Among the estimation TEC models, one of the most well known empirical ionospheric global models is International Reference Ionosphere (IRI). It is a combined international project established by the International Union of Radio Science (URSI) and the Committee of Space Research (COSPAR), primarily based on worldwide measurements from ground and space (i.e. ionosonde stations, International Satellites for Ionospheric Studies (ISIS) and Alouette topside sounder satellites, incoherent scatters radars, rocket measurements, satellite data from in situ, etc) in the late 60s (Bilitza et al., 1993; Bhuyan & Borah, 2007; Bilitza & Reinisch, 2008). Both working groups are responsible in developing and improving the IRI model version with new data and modelling techniques. As a result, the empirical IRI model has been through several major milestone editions and has reached a high level of reliability especially in the mid-latitude ionosphere, since the early development of this model was primarily based on the mid latitudes and with only certain data from the low and equatorial latitudes (Obrou et al., 2009). However, in recent years, the IRI model has concentrated to become a reliable ionospheric representation in low and high latitudes (Zhang et al., 2010). The empirical IRI model is widely used for standard specification of ionospheric parameters which describes the median and average electron density, electron, neutral and ion temperatures, ion composition (O^+, H^+, He^+, NO^+, N^+), and ion drift as functions of year, month, day, local or universal time (LT/UT), height, geographic or geomagnetic coordinates and solar zenith angle. Besides those parameters, the IRI model describes the ionospheric TEC and

the values are obtained by integration the electron density from 50 km to 2000 km. An online web interface for computing and plotting the ionospheric parameter values is accessible from the IRI homepage to estimate or calculate the TEC values and other ionospheric parameters using the appropriate independent variable inputs to the model at http://iri.gsfc.nasa.gov/. Fewer databases from a particular region may cause the estimation accuracy of the IRI results can largely deviate from the observation data, since the model depends fundamentally on the accessibility of registered data for a specific region and time period. This difficulty has been encountered by introducing a number of regional approaches to estimate the ionospheric total electron content and the existing regional TEC estimation models are summarized in Table 2.1.

One of the approaches is an empirical technique using neural network (NN). This technique is proven to be relatively efficient for modelling complex and non-linear processes such as the ionosphere processes (Poole & McKinnell, 2000; Sutcliffe, 2000; Sarma & Madhu, 2005; Oyeyemia et al., 2006; Mckinnell, 2008; Jean et al., 2009). With sufficient of historical data, NN is found to be a promising tool in the ionospheric TEC modelling. Leandro & Santos (2004, 2007) used the NN technique to predict the vertical TEC values at any locations covered by the GPS network in Brazil during low and high solar activities based on two input parameters (i.e. latitude and longitude). The estimation accuracy of their model was approximately 85%. Besides, the NN technique has been used extensively in estimating the ionospheric TEC over South Africa. A feasibility study on the TEC estimations over South Africa based on NNs was carried out on three GPS stations separately by Habarulema et al. (2007a). The TEC values were estimated (NN TEC) as a function of day number (DN), hour (HR), a 4-month running mean of the daily sunspot number (R4) and the running mean of the previous eight 3-hourly magnetic A-

Table 2.1: Summary of regional TEC estimation models

Author-year	Country	Method	Data
Liu et al. (1991)	China	a) Empirical model based on Fourier harmonic analysis	i. Data from 1981-1985 ii. Single-station
Leandro & Santos (2004, 2007)	Brazil	a) NN	i. During LSA & HSA[i] ii. Multi-station
Habarulema et al. (2007, 2009a, 2009b, 2009c, 2010)	South Africa	a) Feed forward NN	i. Data from 2000-2004 ii. Multi-station
		b) Recurrent -Elman NN	i. Data from 2000-2007 ii. 2 stations
		c) Feed forward NN	i. Data from 2000-2005 ii. Multi-station
Acharya et al. (2010)	India	a) Adaptive recurrent network with in – situ learning algorihm	i. Data August 2006 ii. Three different locations: magnetic equator, equatorial anomaly crest, outside the anomaly range
Konstantinidis et al. (2011).	Cyprus	a) Genetic Programming (GP) with Multi-objective Evolutionary Algorithm based on Decomposition characteristics (GP-MOEA/D	i. Data from 1998-2009
Watthanasangmechai et al. (2012)	Thailand	a) Feed forward NN	i. During LSA ii. Data from 2005-2009 iii. Single-station

[i] HSA denotes high solar activity; LSA denotes low solar activity

index values (A8). Based on the following inputs, Habarulema et al. (2009b) employed recurrent Elman neural network (ENN) technique to determine the TEC values. The estimated TEC was used as a reference to quantify the solar wind effects on GPS TEC by including solar wind velocity (Vsw), proton number density (Np) and north-south direction of the interplanetary magnetic field (IMF Bz) as separate inputs. The results showed that the estimation accuracies were close to the reference values, leading them to strongly believe that the solar, magnetic, diurnal and seasonal variability are the strongest mechanisms for TEC variations. Continuously, Habarulema et al. (2009c) developed a multi station or regional NN model to determine the TEC values at any locations over South Africa by including data from different stations at different latitudes within South Africa. They incorporated the geographical position of the receivers together with the parameters adopted in Habarulema et al. (2007) as the input parameters. This developed regional NN model (National NN model) was used to estimate the TEC dynamics during magnetic storms over a five-year period (2000 to 2004) (Habarulema et al., 2010a). In most of Habarulema's modelling, the results from the NN model were compared with those TEC values derived from GPS, IRI and ionosonde for validation. Acharya et al. (2010) have done a modelling of ionospheric TEC using an adaptive recurrent neural network with in-situ learning algorithm. The output is the estimated TEC values for three different stations located at the magnetic equator, equatorial anomaly crest and also outside the anomaly range. Other than that, a feed forward NN technique has been applied by Watthanasangmechai et al. (2012) to estimate the TEC values during low solar activity period (2005-2009) for a single station at Chumpon (10.72˚N, 99.37˚E), equatorial latitude Thailand by using the day number, hour and sunspot number as the input data. The hourly, diurnally and seasonally TEC variations were estimated and the results were compared with the TEC values estimated by IRI-2007 model for validation.

Besides NN techniques, a few other approaches have been introduced on the estimation of the ionospheric TEC. Liu et al. (1991) proposed a new empirical model based on Fourier harmonic analysis for estimating TEC values from 1981 to 1985 over Xinxiang (35.3˚N, 113.9˚E). Diurnal TEC values were estimated based on the annual coefficients generated from this analysis. A feasibility study on the TEC estimations over Cyprus using a Genetic Programming (GP) approach with a Multi-objective Evolutionary Algorithm based on Decomposition characteristics (GP-MOEA/D) was designed using TEC data that covers one full solar cycle (1998 - 2009). The model successfully estimated the vertical TEC with a good approximation (Konstantinidis et al., 2011).

Other than estimation, forecasting the ionospheric parameters a step ahead is also a vital issue in space research due to fast and unpredictable changes in the ionospheric condition. To re-establish readers' knowledge on the techniques used in forecasting the TEC, the following sections may worth for a reading.

2.5.2 Models for forecasting the ionospheric TEC

The important tasks in processing and analysing the ionospheric parameters are to monitor, forecast and detect the ionospheric anomalies earlier. Ionospheric perturbations are always associated to the complex solar-terrestrial and acoustic-gravity waves interaction in the course of space weather and natural hazard events, respectively. Thus, recently, the forecasting models are being used as the background models to produce the preliminary real-time data of the ionospheric parameters. In addition, forecasting the ionospheric parameters such as critical frequency f_0F2, electron density and total electron content ahead have long been an attractive solution for many diverse applications, both in

civil and military applications such as communications, radar, positioning, navigations systems, and seismology.

A short-term forecasting of ionospheric TEC that forecast the ionospheric state in time scale of hours or days is very demanding due to fast changes in the ionospheric condition (diurnal variation) and has great impact on many scientific and technological applications (Garcia-Rigo et al., 2011). There has been a great effort in the ionospheric research community, where several forecasting techniques have been proposed using the ground- and space- based ionospheric TEC data. The existing forecasting models are summarized in Table 2.2.

Neural network-based techniques are implemented to forecast the ionospheric TEC ahead (Cander et al., 1998; Cander, 1998a; Tulunay et al., 2004, 2006). Cander et al. (1998) presented a hybrid time-delay multi-layer percepton neural network to forecast the TEC values for 1-hour ahead, f(t+1) based on the TEC data retrieved from Faraday rotation at Florence (43.8˚N, 11.2˚E), using the signal of the OTS-2 satellite. The model was designed based on 12 input parameters with two hidden layers to perform a short-term forecasting ability. The input parameters used to construct the NN model are shown in Table 2.3. To represent the production of electron population in the ionosphere, the model uses daily sunspot, R_{12} and Dst as the input parameters to describe the TEC variability. Since the model has solar index, the model's output shows a good agreement during the daytime compared to night because solar activity does not influence the electron contents during night hours. Tulunay et al. (2004) from Middle East Technical University (METU) in Ankara developed a NN model to forecast the TEC values for the intervals ranging from 10 minutes to 24-hours ahead during high solar activity period

Table 2.2: Summary of regional TEC forecasting models

Author-year	Country	Method	Data	Output
Cander et al. (1998,1998a)	Italy	a) A hybrid time-delay multi-layer percepton neural network	i. Data 1990 ii. Single-station: Florence	i. 1 hour ahead TEC values
Krankowskia et al. (2005)	Europe	a) Autoregressive Moving Average (ARMA)	i. During LSA, MSA, HSA ii. Data November 1997, September 1999, March-April 2001 iii. Dual station: Borowiec, Matera	i. 1- to 3- hours ahead TEC values
Tulunay et al. (2004, 2006)	Turkey	a) NN	i. During HSA ii. Data from 2000-2002 iii. Dual-station: Chilbolton, Hailsham	i. 10 minutes to 24 hours ahead TEC values
		b) NN	i. Data November 2003 ii. Multi-station	

Acharya et al. (2009)	India	a) Kalman filter	i. During LSA – September 2005 ii. Multi-station	i. 1-, 3- and 5-minutes ahead TEC values
Yan & Yamin (2011)	China	a) Superposition analysis of periodical wave variance	i. Data January and February 2008	i. 2 days ahead TEC values
Ratnam et al. (2014)	India	a) Holt-winters : Additive and Multiplicative	i. Data 2013 ii. Single-station	i. 2 days ahead TEC values
Mane et al. (2014)	India	a) Autoregressive Moving Average (ARMA)	i. Data 2013	i. 24 hours ahead TEC values
Niu et al. (2014)	China	a) Combinational of seasonal and Autoregressive Moving Average (ARMA) model	i. Data 2013	i. 2 days ahead TEC values

[i] HSA denotes high solar activity; LSA denotes low solar activity; MSA denotes medium solar activity

Table 2.3: The parameters used in (Cander et al. 1998) forecasting model

No	Input parameters	Notations
1	TEC at time t	*f(t)*
2	Mean TEC at t	*Mf(t)*
3	Mean TEC at t-1	*Mf(t-1)*
4	Mean TEC at t-23	*Mf(t-23)*
5	Mean TEC at t-47	*Mf(t-47)*
6	Mean TEC at t+1	*Mf(t+1)*
7	Delta f(t)	*f(t)-Mf(t)*
8	Delta f(t-1)	*f(t-1)-Mf(t-1)*
9	Delta f(t-23)	*f(t-23)-Mf(t-23)*
10	Delta f(t-47)	*f(t-47)-Mf(t-47)*
11	R_i	Daily sunspot number
12	Dst	Hourly ring current index

(2000 – 2001) at Chilbolton (51.8˚N, 1.26˚W) station. The developed model was further validated by forecasting the TEC values at Hailsham (50.9˚N, 0.3˚E) in 2002 whose data was not included in the training phase. Tulunay et al. (2006) presented ionospheric TEC forecast mapping based on NN technique during November 2003 using the parallel temporal parameters and past TEC values for 104 grid locations over Europe. The forecast TEC values from 10 minutes to 1-hour ahead at these 104 grids are used to generate the TEC maps over Europe using Bezier surfaces.

Besides NN technique, Acharya et al. (2009) have put forward the Kalman filter approach for short-term ionospheric TEC forecasting using the TEC data from different GPS Aided GEO Augmented Navigation (GAGAN) stations in 2005. The output of this model forecast the TEC for different intervals (1-, 3-, 5-minutes) at different stations. The model tends to forecast fairly for stations located out of the equatorial anomaly region as well as for shorter estimation intervals. Garcia-Rigo et al. (2011) proposed ionospheric vertical TEC forecasting approach to forecast the TEC 2-days ahead at global scale using GPS data from the International Global Navigation Satellite Systems (GNSS) Service (IGS) Ionospheric Working Group (IGS Iono-WG). This model was based on the Discrete Cosine trasform (DCT), where the linear regression method was applied to predict the past and future values of each DCT coefficient.

In addition, a few statistical based methods were introduced in modelling the ionospheric total electron content. Traditional methods such as Autoregressive (AR) and Autoregressive Moving Average (ARMA) techniques were developed to forecast the ionospheric TEC time series ahead using single station GPS measurements (Krankowskia et al., 2005; Karthik et al., 2012; Mane et al., 2014). Yan & Yamin (2011) from China developed an improved superposition analysis of periodical wave variance forecast method using TEC data obtained from IGS in 2008. The model is used for short-term forecast of the ionospheric TEC. Subsequently, a weighted method was implemented in this model to improve the variance analysis. The model successfully forecast the ionospheric TEC with higher precision accuracy compared to the existing forecast model in China. A study by Ratnam et al. (2014) proposed Holt-Winters method to forecast the ionospheric TEC ahead for a single station in a low latitude region. In this work, Additive and Multiplicative Holt-Winter models were adopted to forecast the TEC values 2-days

ahead based on the history of the TEC values. The output of these models showed that Holt-Winters additive model was able to forecast the TEC values more accurately than the multiplicative model over low latitude region. Niu et al. (2014) proposed a new combination technique based on seasonal and ARMA models to forecast the ionospheric TEC ahead and the output of this model was compared with the traditional model. They found that the combination method produces better results compared to a single traditional model.

A great number of ionospheric estimation and forecasting TEC models (for a single station or regional) in this direction, shows that the essentialness of developing an ionospheric TEC model in this field. Besides, different approaches have been introduced and developed in different countries as the Earth's ionosphere behaviours are very geographical dependant, especially the equatorial ionospheric characteristics. Therefore, it is necessary to develop estimation and forecasting TEC model based on the local measurements at the equatorial region due to the numerous complexities. The importance of developing local estimation and forecasting ionospheric TEC models in this region has been highlighted in Chapter 1. In this work, a NN technique is used for estimation of ionospheric TEC, while a combination method based on SARIMA and NN models is developed for forecasting the GPS TEC values ahead for different condition days. The following sections describe the theories, concepts and algorithms of the NN and SARIMA methods in detail.

2.6 NEURAL NETWORK (NN) MODEL IN NON-LINEAR APPROXIMATION

Neural Network (NN) is advent of modern neuroscience. The networks are often described based on the brain concept. Neural network has massive interconnection of processing elements referred as 'neurons' and an artificial neuron is similar to the nervous system in the human brain. The network requires knowledge from its environment to accomplish a complex task that corresponds to the desired target. Through the learning process or known as “experience” in human language, the network has ability to set up its own rules according to the specific task. Indeed as “experience”, the neurons in the network are trained to learn about the environment. In addition, during the learning process, the synaptic weight refers to the inter-neurons' connection strength in the network is adjusted by the learning algorithm function to produce an optimum outcome. Eventually, after storing the sufficient experiential knowledge, the network is capable to generalize. NN can produce an optimum and satisfactory result for the unseen inputs, which are excluded during the learning and validation processes. In other words, a new sample set of input is fed into the trained network to generalize new desired outcome (Haykin, 1999; Fausett, 1994; Kumar, 2004).

Three fundamental aspects of an artificial NN are the connection links, the adder and the activation function. Connections between the processing units are known as synaptic weights, w_{kj}. In the connection, an effective element input, x_j is transmitted to neuron k and it is multiplied by weight w_{kj}. The weighted values from the parallel inputs are fed into a summing junction to sum up the input signals. Finally, the activation function performs a mathematical operation on the output signal to control the amplitude

range of the output signal to some finite values based on the activation function assigned to the network. Besides weights, there is also another factor that has effects on the input signals of the activation function known as bias, b. It is an external input of the network with a positive or negative value (Krose & Van der Smagt, 1996; Haykin, 1999). Figure 2.5 depicts a general model of an artificial NN with bias.

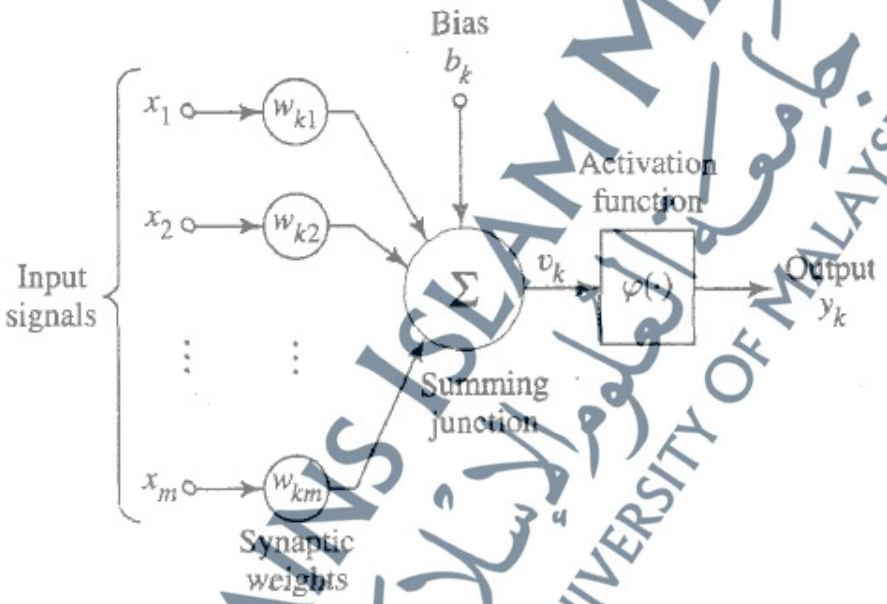

Figure 2.5: A general neuronal model with bias (Haykin, 1999)

An appropriate activation function is able to scale the output signals to a desired range of values. There are a few activation functions that are commonly used in designing the network. Linear function, piecewise linear function, threshold function are known as the linear activation functions while the non-linear activation functions are sigmoid (S-shaped) function and tangent hyperbolic function. Mostly the non-linear activation functions are used to comprehend complex task in a network. There are a few factors that cause the non-linear activation functions become well known than the others.

In a multilayer network, the non-linear activation functions permit the input signals to propagate through many layers, whereas the linear activation functions present similarly as in a single-layer network. Therefore, a non-linear function does not limit the capabilities of a network. Furthermore, a network which is trained with backpropagation algorithm is always beneficial with a non-linear activation function. This is mainly because the function has two significant criteria, continuity and monotonicity (Kumar, 2004). In addition, this function can be easily differentiated at any point, resulting in the reduction of computational process time during training (Fausett, 1994; Habarulema et al., 2007). In general, the sigmoid function is divided into two categories, binary sigmoid and bipolar sigmoid. The binary sigmoid is defined as in Equation (2.1) where the x represents the input vector and the σ corresponds to the slope parameter.

$$f(x) = \frac{1}{1 + e^{-\sigma x}} \tag{2.1}$$

The desired range of the output values for a binary sigmoid are in between 0 and 1. However in certain scenarios, the network requires a larger range of output values and in this case the bipolar sigmoid function is expanded from the binary sigmoid. It has a wider range than the binary sigmoid function. This function most commonly is used in the network which requires a desired output range in the interval between -1 to 1 and the function terms is as in Equation (2.2):

$$f(x) = \frac{2}{1 + e^{-\sigma x}} - 1 \tag{2.2}$$

The bipolar sigmoid is closely related to tangent hyperbolic function which has the similar output range. The tangent hyperbolic function is derived as in Equation (2.3):

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$= \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2.3)$$

2.6.1 Architecture of Neural Network

In general, NN has various kinds of architectures and the usage of particular architecture commonly depends on the types of applications that the network is required. The arrangement of the neurons in each layer and the interconnection link within and between the layers are the two key factors that distinguish different types of architectures in NN. Fundamentally, there are two types of network architectures:

i. Single-layer network

A single-layer perceptron is known as a single-layer network. It is the simplest form of network which consists of only an input layer with input nodes that corresponds to the number of inputs fed into NN and an output layer with output nodes transmit the compute signals of the network to real world. The signals propagate in forward direction and the computation is only executed in the output layer, therefore the network is referred as a single-layer.

ii. Multilayer network

A multilayer network is referred as multilayer perceptrons. It is an expanded form of a standard single-layer perceptron. A layered network consists of an input layer and an output layer. In addition a multilayer network usually has one or more

hidden layers located between the input layer and the output layer (Habarulema, 2007; Alsmadi et al., 2009). Both the input and output nodes are connected to external environment or to the real world. The internal connections correspond to number of hidden neuron to the hidden layer is not directly accessible. Therefore, the hidden neurons are the intermediation between the source nodes and the output nodes. The outcome from the output layer is a response to each of the input pattern specified in the network (Haykin, 1999).

The number of hidden layers and hidden neurons per layer are still a critical issue in a multilayer network due to lack of theoretical approach to determine the appropriate number of the processing elements and hidden layers (Kumar, 2004; Alsmadi et al., 2009). Alsmadi et al. (2009), has also indicated that there is no explicit specification and experimental work to explain the layout of a network. There are certain general rules are implemented by the researcher to design a network depending on the application. Kumar (2004) has suggested a cross validation approach to estimate the number of hidden neurons to obtain a reliable network. In certain NN applications, increasing the hidden layers might ease the training (Fausett, 1994), but the accuracy of the results may still not vary significantly (Haykin, 1999). The fully interconnected multilayer network with a single hidden layer is shown in Figure 2.6. The multilayer network is able to solve complicated tasks more effectively but it requires longer computational time compared to a single-layer network. The following section focuses on the algorithm in a feed forward multilayer network. There are several algorithms available for multilayer NN applications, yet in this thesis only the feed forward network trained by back propagation is explored and briefly discussed.

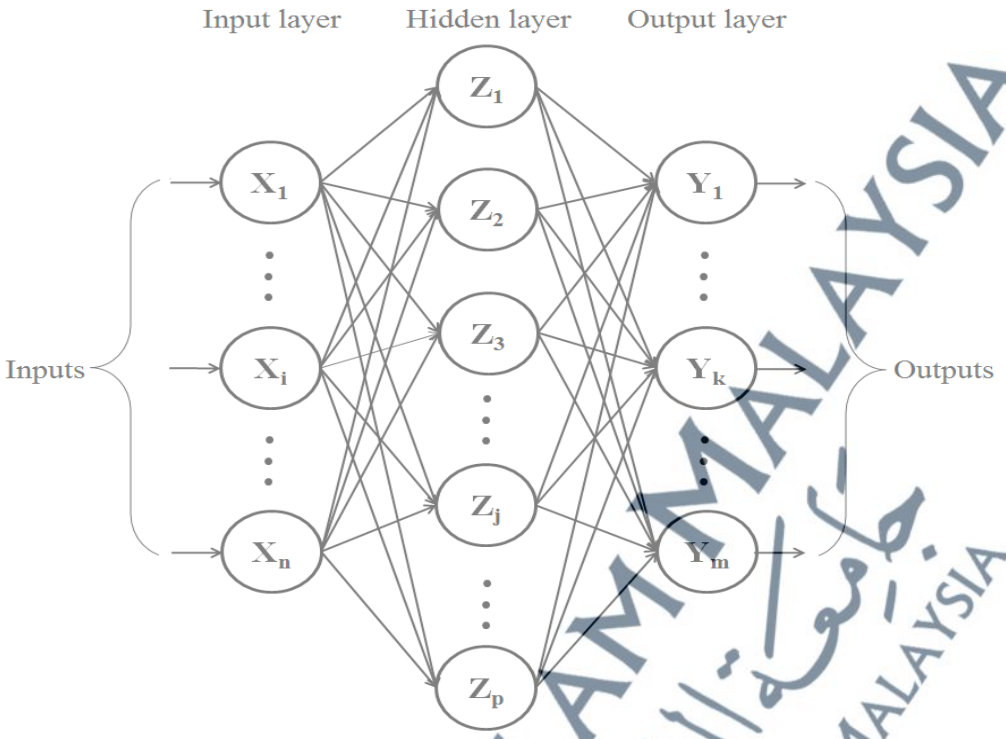

Figure 2.6: A multilayer network with a single hidden layer

2.6.2 Feed forward with back propagation algorithm

A feed forward network is a form of NN which exists in both single-layer and multilayer networks. An example of a multilayer feed forward network with a single hidden layered is illustrated in Figure 2.6. Overall in a feed forward with back propagation algorithm learning phase, the signals flow from the input layer to the output layer in a forward direction without permits any closed paths (closed-loop) from a layer back to itself, while the error signals are sent in the backward direction during back propagation (Fausett, 1994). The theoretical and mathematical formulations of feed forward multilayer network trained by back propagation (of errors) algorithm with a randomized weights are referred and adopted from Fausett (1994). In this section an example of the basic method of back propagation based on an optimization technique

known as gradient descent learning method, which is used to estimate the errors and alter the weights in all the layers during each training pattern is presented with the objective of minimizing the total squared error of the output computed by the network (Fausett, 1994; Kumar, 2004).

A feed forward network with back propagation algorithm has three standard stages but prior to the main stages, random initialisation of weight is executed. Following are the stages of this algorithm:

i. Stage 0: Random weight (w) initialisation

   The weights in each layer are randomly initialised and the values of the weight falls within the interval range of $-1 \leq w \leq 1$.

ii. Stage 1: Feeding the input signals in forward direction (feed forward)

   a. In this network, the input neuron(s) (X_i, $i = 1,...,n$) received the input signals and forwarded the signals to the hidden neuron(s) (Z_j, $j = 1,...,p$) in the hidden layer. Each hidden neuron extracted the signals and summed its weighted input signals;

$$z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij} \tag{2.4}$$

   where z_in_j is the net input to Z_j, v_{0j} denotes as the bias on the hidden neuron j, x_i is the input signal of the X_i, and v_{ij} represents as the weight connection between the input neuron i and hidden neuron j.

b. Once summed up the weights, as in Equation (2.5) each hidden neuron then computed its net input (z_in_j) using its specific activation function that has been assigned to this neuron;

$$z_j = f(z\_in_j) \tag{2.5}$$

where z_j is the output activation signal of Z_j and f is the assigned activation function.

c. The output signals from the hidden layer are forwarded to the output neuron(s) (Y_k, k = 1,...,m) as the input signals to the output layer. Each output neuron extracted the signals and summed its weighted input signals as done by the hidden neuron previously;

$$y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk} \tag{2.6}$$

where y_in_k is the net input to Y_k, w_{0k} denotes as the bias on the output neuron k, z_j is the output signal of the Z_j, and w_{jk} represents as the weight connection between the hidden neuron j and output neuron k.

d. Finally, each output neuron calculated its net input (y_in_k) using its activation function and produced the output signal at the output neuron as in Equation (2.7):

$$y_k = f(y\_in_k) \tag{2.7}$$

where y_k is the output activation signal of Y_k and f is the assigned activation function. The feed forward process ends here and as the training continues, the error signals are estimated and propagated in back direction.

iii. Stage 2: Computation and back propagation of the errors

Since a supervised training method is employed in the network, both known inputs and desired outputs are used in the training phase. Therefore each output neuron has its own target value, t_k . In this stage,

a. Firstly each output neuron (Y_k, k = 1,...,m) compared its output signal, y_k to the corresponding targeted value t_k and computes the associated error. From the error, δ_k is calculated as in Equation (2.8):

$$\delta_k = (y_k - t_k) f'(y\_in_k) \tag{2.8}$$

where δ_k is the error information term of the output neuron k, Y_k and f' is the derivative of the activation function. The weight correction term, Δw_{jk} (between the output and hidden) and bias correction term Δw_{0k} are computed using the δ_k as in Equation (2.9) and Equation (2.10), respectively.

$$\Delta w_{jk} = \alpha \delta_x z_j \tag{2.9}$$

$$\Delta w_{0k} = \alpha \delta_x \tag{2.10}$$

where α is the learning rate. Both the weight and bias correction terms are calculated to update the random weights and biases between the hidden and output layers later as in Equation (2.15) and (2.16) in stage 3. Then the

computed δ_k is sent back to the neurons in the previous layer connected to Y_k, where in this case the previous layer is the hidden layer with hidden neuron j, Z_j.

b. Secondly in the hidden layer, each hidden neuron (Z_j, j = 1,...,p) sum its δ_k input from the output layer

$$\delta\_in_j = \sum_{k=i}^{m} \delta_k w_{jk} \tag{2.11}$$

where δ_in_j is the sum of all delta input units at hidden neurons j. Based on the δ_in_j, the error information term for each hidden neuron, δ_j is calculated as follow:

$$\delta_j = \delta\_in_j f'(z\_in_j) \tag{2.12}$$

The δ_j is only used to update the weights and biases between the hidden and input layer and not to propagate the error back to the input layer. Using the same procedure, the weights and biases between input and hidden layer are updated using the correction terms. The Δv_{ij} and Δv_{0j} are computed as in Equation (2.13) and (2.14), respectively.

$$\Delta v_{ij} = \alpha \delta_j x_i \tag{2.13}$$

$$\Delta v_{0j} = \alpha \delta_j \tag{2.14}$$

where Δv_{ij} is the weight correction term between input and output layers and Δv_{0j} is the bias correction term between input and output layers.

iv. Stage 3: Adjustment of the synaptic weights

a. In the final stage using the correction terms, the weights and bias (j = 0,...,p) between the hidden and output layers are updated as follows:

$$w_{jk}(new) = w_{jk}(previous) + \Delta w_{jk} \quad (2.15)$$

b. Using the same procedure as above, the weights and bias (i = 0,...,n) between the input and hidden layers are updated as follows:

$$v_{ij}(new) = v_{ij}(previous) + \Delta v_{ij} \quad (2.16)$$

The adjustments to all weights in the layers are done simultaneously during the training phase. The weights are updated after each training pattern is executed. In other words, the algorithm updated the adjusted weights iteratively until the network reaches its stopping condition(s).

Overall the stages show the complexity of the algorithm during NN applications. Due to the complexity, a number of epochs are required to train the back propagation network, where an epoch is one cycle of the entire training set. The combination of sigmoid functions and the network trained by back propagation algorithm simply benefits the network, because the simple relationship between the value of the function at a point and the value of the derivative at that point reduces the computational burden during training (Fausett, 1994).

2.7 SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA) MODEL IN LINEAR APPROXIMATION

Autoregressive Integrated Moving Average (ARIMA) is a traditional time series forecasting technique, which is mainly used in developing linear models. ARIMA has been established based on the Box- Jenkins methodology (Chatfield, 2000) where sometimes it is also referred as Box- Jenkins models. ARIMA is a combination of three different functions comprising autoregressive (AR), moving average (MA) and an integrated (I). The theoretical and mathematical formulations of the ARIMA model are referred and adopted from Box and Jenkins (1976).

2.7.1 Autoregressive (AR) processes

The autoregressive function regressed on the values in the previous periods. The autoregressive model with p lags, (abbreviated AR(p)) with a constant can be written in the form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t \tag{2.17}$$

where a_t represents a purely random process with zero mean and constant variance of σ^2. Using the backward shift operator B, where $By_t = y_{t-1}$ the expression in Equation (2.17) can be expressed in the form:

$$\begin{aligned} \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) y_t &= a_t \\ (1 - \sum_{i=1}^{p} \phi_i B^i) y_t &= a_t \end{aligned} \tag{2.18}$$

where $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ *or* $1 - \sum_{i=1}^{p} \phi_i B^i$ is polynomial in B in order p. Finally the AR process in order of p is written in the form:

$$\phi_p(B) y_t = a_t \tag{2.19}$$

2.7.2 Moving average (MA) processes

The moving average function regressed on the random process or errors. MA(q) is a moving average model with q lags and expressed in the form:

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{2.20}$$

where a_t represents a purely random process with zero mean and constant variance of σ^2. The Equation (2.20) can alternatively be expressed as in the Equation (2.21) by substituting the backward shift operator, B.

$$\begin{aligned} y_t &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t \\ &= (1 - \sum_{j=1}^{q} \theta_j B^j) a_t \end{aligned} \tag{2.21}$$

where $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ *or* $1 - \sum_{j=1}^{q} \theta_j B^j$ is polynomial in B in order q. Finally the MA process in order of q is expressed in the form:

$$y_t = \theta_q(B) a_t \tag{2.22}$$

2.7.3 Autoregressive moving average (ARMA) processes

The combination of autoregressive moving average model with p autoregressive terms and q moving average is known as an ARMA (p, q) model Chatfield (2000). The function of the ARMA model can be expressed in the form

$$\phi_p(B)y_t = \theta_q(B)a_t \tag{2.23}$$

where $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of finite order p and q, respectively. Equation (2.23) is a mixed combination of Equation (2.19) and Equation (2.22).

2.7.4 Autoregressive integrated moving average (ARIMA) processes

A time series is said to be stationary if the statistical properties of the data do not vary with time where the mean (no trend overtime) and variance of the data are constant. However in many practices, the time series data exhibits seasonal and trend variations, which are the major components of a non-stationary time series. A trend component is defined as a long term movement in the time series data. This means that the observed data has an underlying direction (upwards or downwards). This type of trend is classified as a deterministic trend, where a simple linear or polynomial model can be used to detrend the model. Apart from that, there is also time series shows stochastic trend which is unable to be easily modelled.

Autoregressive integrated moving average (ARIMA) model is a combination of ARMA model with an integrated "I" function, which allow the time series to be

stationary by differencing. The main requirement for developing an ARIMA model is that the process requires the original time series to be stationary. The first differencing of the original time series can be expressed in the form.

$$(1-B)^1 y_t = y_t - By_t$$
$$= y_t - y_{t-1} \quad (2.24)$$

and d[th] non-seasonal differencing can be written as $(1-B)^d y_t$, where y_t is the original time series and B is the backward shift operator. An ARIMA model is abbreviated as ARIMA (p,d,q) when the non-stationary data series is differenced at d times before the data is fitted into the ARMA (p,q) model. The mathematical expression of the ARIMA(p,d,q) model is as in Equation (2.25):

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)a_t \quad (2.25)$$

where $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of finite order p and q, y_t is the original time series or non-stationary time series and a_t represents a purely random errors with zero mean and constant variance of σ^2. The major limitation of this model is that, it is unable to comprehend the non-linear patterns and pre-assumed linear form of the model. Thus, the future value of a variable is assumed to be a linear function in an ARIMA model.

2.7.5 Seasonal autoregressive integrated moving average (SARIMA) processes

Seasonality is another significant component that is often accompanied with the trend time series. Seasonal time series is a periodic and recurrent pattern that always has

a fixed and known period e.g. hourly, weekly, monthly, quarterly and yearly. In seasonal time series data, deseasonalisation is very important. According to Zhang and Qi (2005) and the references therein, seasonal fluctuation has high influence on the time series data. Without a proper seasonal adjustment, seasonal fluctuation can cause difficulties in capturing the other time series components e.g. cyclic and irregular components.

SARIMA model is an extension of the ARIMA model in order to include the seasonality component, which is particularly present in most of the time series. The first seasonal differencing can be expressed in the form:

$$(1 - B^s)^1 y_t = y_t - B^s y_t$$
$$= y_t - y_{t-s} \tag{2.26}$$

and the D[th] seasonal differencing can be written as $(1 - B^s)^D y_t$, where *s* is the length of the seasonal period. The mathematical expression of ARIMA model can be explained as SARIMA (p,d,q) $(P,D,Q)_s$ where (p,d,q) represent the non- seasonal part of the model while $(P,D,Q)_s$ is the seasonal part of the model, is mentioned in Equation (2.27):

$$\Phi_P(B^s)\phi_p(B)(1 - B)^d(1 + B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2.27}$$

where

$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$

$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$

$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$

B is the backward shift operator, *p* is the order of non-seasonal autoregressive, *q* is the order of non-seasonal moving average, *d* is the order of differencing non-seasonal, *P* is the order of seasonal autoregressive, *Q* is the order of seasonal moving average, *D* is the order of differencing seasonal, *s* is the length of seasonal period, Φ is the coefficient of seasonal autoregressive, Θ is the coefficient of seasonal moving average, ϕ is the coefficient of non-seasonal autoregressive, and θ is the 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.

2.7.6 SARIMA model development based on Box-Jenkins technique

The Box and Jenkins (1976) methodology involves three iterative steps for SARIMA model selection. Figure 2.7 exhibits the 3 steps on the Box-Jenkins approach to develop a SARIMA model. Each step describes its possible contribution to develop an accurate linear forecasting model. Following are the three steps to develop the SARIMA model:

i. Step 1: Model identification

   Identification of the model consists of two stages: (1) In the first stage, data transformation is required to stationarize a time series. Appropriate transformation such as logs, differencing or detrending is performed to achieve stationary time series. Stationary is a necessary conditions in modelling a SARIMA model in order to attain its statistical characteristic that (mean and variance) does not change over time or have a pronounced trend. The "I" or known as "Integrated" which identifies the degree of differencing SARIMA

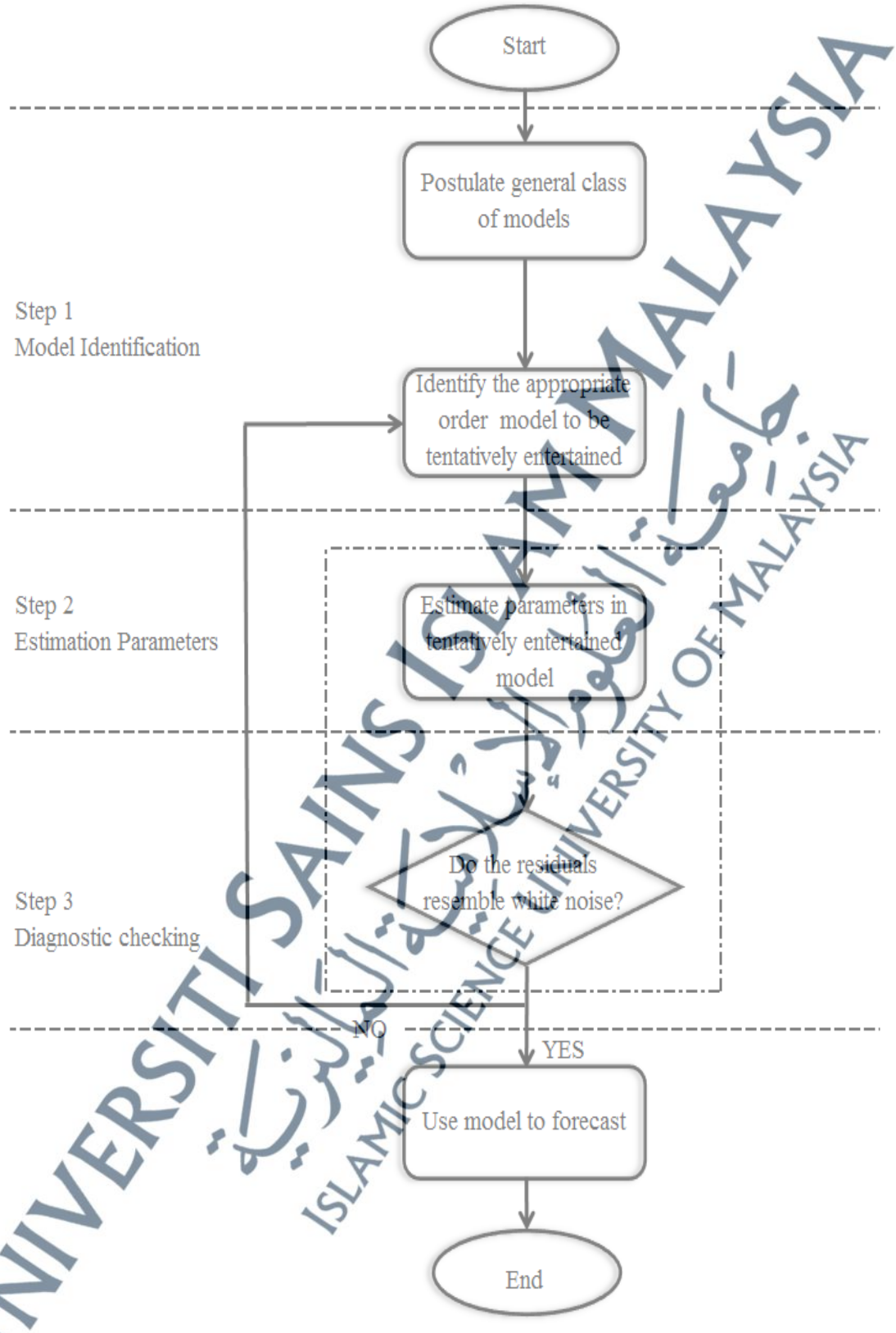

Figure 2.7: Box-Jenkins modelling approach (Makridakis,1983)

model indicates that the modelling data series has been transformed into a stationary time series. (2) In the second stage, the order of the SARIMA model (the order of appropriate autoregressive, moving average, seasonal autoregressive and seasonal moving average) terms is identified by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the sample data (Box & Jenkins, 1976).

ACF is a basic statistical tool used to define the patterns in the data based on the number of time-lags in a time series. The ACF values are in the range between -1 and 1 computed from the time series at different lags to measure the correlation between the same variable with its present and past observations. Equation (2.28) is a general formula for ACF at lag k (Chatfield, 2000):

$$ACF(k) = r(k) = \frac{\gamma_k}{\gamma_0} = \frac{\sum_{t=1}^{N-k}(y_t - \bar{y}_t)(y_{t+k} - \bar{y}_t)}{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2} \quad (2.28)$$

where γ_k is the set of auto-covariance coefficients for k = 0, 1, 2, . . . , γ_0 represents the variance of the original time series y_t, $\bar{y}_t$ denotes the mean of the time series and y_{t+k} represents the lag values of time series data at *k*. PACF is another potentially useful tool in model identification. PACF essentially measures the excess correlation at lag τ which has not been accounted by autocorrelations at lower order lags. PACF at lag τ is given by the following Equation (2.29) with $\tau > 1$:

$$\pi_\tau = \frac{r_\tau - \sum_{j=1}^{\tau-1} \pi_{\tau-1,j} \cdot r_{\tau-j}}{1 - \sum_{j=1}^{\tau-1} \pi_{\tau-1,j} \cdot r_{\tau-j}} \quad (2.29)$$

where r and π is the autocorrelation and partial autocorrelation functions, respectively. The ACF coefficient r and PACF coefficient π at zero lag are equal to 1, $r_0 = \pi_0 = 1$, while both the coefficients are same at first lag, $r_1 = \pi_1$.

Table 2.4 shows the summary of the basic three types of the non-seasonal theoretical ARIMA models; moving average model, autoregressive model, and mixed autoregressive moving average model. In the SARIMA model, the seasonal parameters; seasonal moving average model, seasonal autoregressive model, and seasonal mixed autoregressive moving average model follow the similar theoretical property as per the ARIMA model to identify the order of the seasonal lags, P and Q. Furthermore, the theory shows that the ACF function clearly indicates a moving average process of order q, (MA(q)) through the number of non-zero terms in ACF. Similarly, PACF can be used to identify the AR processes, clearly indicates the order p of the autoregressive process AR(p) via the number of zero terms for lags greater than p. In addition, the visualizations of the ACF and PACF along with their theoretical autocorrelation

Table 2.4: Non-seasonal theoretical ARIMA models

Model	ACF	PACF
Moving Average of order q	Cuts off after lag q	Dies away
Autoregressive of order p	Dies away	Cuts off after lag p
Mixed Autoregressive Moving Average (p,q)	Dies away	Dies away

autocorrelation properties are used to identify the possible model and determine the appropriate p, q, P and Q to fit the stationary TEC time series. Usually, by combining the autocorrelation patterns with the theoretical ones, it is possible to determine one or more than one potential model for a given time series.

Among the plausible identified models, the most appropriate SARIMA model that provides adequate description of the stationary time series is determined by using the Akaike Information Criterion (AIC). This index is used to estimate the quality of the model and to determine the best fitted model among various competing SARIMA models ((Zhang, 2003; Khashei & Bijari, 2010). The model that gives minimum value of AIC is considered as the best fitted model. The AIC can be calculated using equation (2.30) and the formula is referred and adopted from Durdu (2010):

$$AIC = n\left(ln\left(\frac{2\pi * SSE}{n}\right) + 1\right) 2m \tag{2.30}$$

where *m* is equal to (*p*+*q*+*P*+*Q*) is the number of orders estimated in the model selection, *n* represents the number of sample in the estimation data set, and *SSE* represents the sum of squared errors.

In model identification, other than ACF and PACF proposed by Box and Jenkins (1976), there are some other new techniques that have been proposed widely in model identification which can be found in Zhang et al. (1998) and Khashei & Bijari (2010). Those techniques are based on validity criteria and intelligent paradigms to improve the accuracy of order selection of ARIMA models.

ii. Step 2: Parameter estimation

Once the time series is transformed and the orders of the SARIMA model are tentatively determined, the coefficients of the best fitted parameters are estimated in the second step. In Box-Jenkins approach, this stage is considered the most straightforward. The parameters are estimated as such that the overall accuracy of model is maximized or in other words, the measure errors of the model are minimized. This accuracy can be accomplished using a non-linear optimization procedure (Zhang et al., 1998; Aladag et al., 2009; Durdu, 2010; Khashei & Bijari, 2010), e.g. based on steepest descent method, maximum likehood method and least squares method.

iii. Step 3: Diagnostic checking

Finally, the last step of building up a SARIMA model is the diagnostic check to determine the model adequacy. In this step the residuals e_t generated from the fitted model are examined and verified to find out if the model is an adequate one for the time series. Several tests can be employed to examine the goodness of fit of the fitted model. Residual autocorrelation function (RACF) is one of the tests that is commonly used to verify the left over residuals from the tentatively entertained model (Durdu, 2010). This test is basically conducted to determine if the selected model assumptions about the residuals e_t are satisfied, where the residuals from the fitted model should resemble pure random errors (white noise). In this case, the white noise means that the autocorrelation of the residuals are uncorrelated and normally distributed around a zero mean. Other than the correlogram of the residuals, the diagnostic statistics tools among those such as Ljung- Box Q test (Q*), Box-Pierce test (Q_{BP}) are used to test whether the

residuals over time are random and independent. If the test statistics Q* or Q_{BP} recedes the critical value χ^2 (chi-squared distribution with degree of freedom), then the model is considered adequate.

Lastly, if the fitted model exhibits lack of fit through the plots or statistics test, all the three steps for the SARIMA model's building process; identification, estimation and diagnostic checking are repeated until a satisfactory model is designed. The final fitted model is used for forecasting purposes.

2.8 IONOSPHERIC TEC BASED ON GPS MEASUREMENTS

The Global Positioning System (GPS) is the first satellite-based navigation system in operation in the mid-20th century. This satellite network system consists of a constellation of at least 24 reliable satellites at the height of about 20,200 km above the Earth's surface. All the GPS satellites are distributed equally in six orbital planes with an inclination of about 55° relative to the equator to ensure worldwide coverage (El-Rabbany, 2002). Typically, the GPS receiver requires at least four different satellites in order to provide accurate information on position or location. Initially, the system was developed for military purposes in the United States Department of Defense (USDoD), but was later made available for civilian use as well.

The GPS consists of three major segments; the space segment, the control segment and the user segment. As aforementioned, the space segment is formed by a constellation of 24 operational satellites for transmitting radio signals to users, receiving and retransmitting navigation message to the control segment. Each satellite has a

nominal period of 12 sidereal hours (Navstar, 1996). The control segment comprises monitor and control stations. The monitor station tracks all the GPS satellites in view and accumulates the ranging data of each satellite meanwhile, the master control station processes the gathered information to control the satellite constellation and updates the satellites' navigational messages (Navstar, 1996; Komjathy, 1997). The user segment is utilised by the military as well as civilian which consists of GPS antennas and receivers. In order to allow the users receive the GPS signals, the GPS receiver is connected to the GPS antenna to provide accurate positioning, velocity and timing information to the end users (El-Rabbany, 2002).

The signals from GPS satellites are composed of two carrier L-band frequencies, L1 (1575.42 MHz) and L2 (1227.60 MHz) modulated by two pseudorandom noise code, C/A-code (coarse acquisition code) and P-code (precise code) and a navigation message. Both the carrier frequencies are generated based on the fundamental frequency, f_o (10.23 MHz), where L1 = $154f_o$ and L2 = $123f_o$. Even though all the GPS satellites transmit the same carrier frequencies but still each satellite differs significantly from one another due to the modulation code. The P-code is modulated onto L1 and L2 frequencies while the C/A code is modulated onto L1 frequency only. The GPS network provides two types of positioning and timing services; (i) Precise Positioning Service (PPS) accessible for military and authorised users. It uses one of the transmitted GPS codes, called as P(Y) code (encrypted P-code into Y-code), and whereas (ii) Standard Positioning Service (SPS) is accessible for civilian users and the service is considered less precise than PPS (El-Rabbany, 2002; Navstar GPS, 1996; Habarulema, 2010). The C/A code is used for this service, which is free to all users (authorised and unauthorised users). In order to provide accurate positioning services as mentioned above, the GPS receiver is able to

access the codes simultaneously from at least four visible satellites at any time from the user position. The system is structured in such a way, so that it can use three satellites to compute the receiver's location, and the fourth satellite is used to remove the receiver clock offset from GPS time (Komjathy, 1997).

Discontinue of Selectivity Availability (SA) feature, the ionosphere has become the largest source of error in the GPS positioning and navigation systems (Klobuchar, 1991). For high precision GPS positioning, the ionospheric effect must be eliminated from the GPS observation. The next section gives a brief knowledge on the effects of ionosphere's refractive index on the propagation of electromagnetic waves.

2.8.1 Radio waves refraction in ionosphere

The electromagnetic waves that transverse the Earth's atmosphere are mostly affected by the F2 region in the ionosphere. Moreover, the ionosphere, which acts as a dispersive medium due to the number of free electrons, causes the speed of the propagation signals change as a function of its frequency and delay in the modulating signal. To understand and describe the behaviour of radio signals in the ionosphere, the ionosphere's refractive index, which is mainly related to the speed of an electromagnetic waves propagating in this medium should be understood.

Basically, the waves that propagate through a refractive medium can be associated to its refractive index of the medium, *n* which can be defined as a function of propagation velocity in the free space, *c* over the speed in the medium, *v*:

$$n = \frac{c}{v} \tag{2.31}$$

where c is the speed of light (3 x 10^8 ms^{-1}). According to Langley (2000), in a dispersive medium, the phase velocity (phase speed) , v_ϕ of propagation is a function of the wave's frequency and the refractive index that influence the velocity is called as phase refractive index, n_ϕ as defined in Equation (2.32) while the group refractive index, n_g is defined as in Equation (2.33):

$$n_\phi = \frac{c}{v_\phi} \tag{2.32}$$

$$n_g = \frac{c}{v_g} \tag{2.33}$$

The n_g can be also defined in terms of n_ϕ (Langley, 2000) as:

$$n_g = n_\phi + f\frac{dn_\phi}{df} \tag{2.34}$$

where $\frac{dn_\phi}{df}$ defined the changes of phase refractive index with respect to the frequency.

The complex refractive index n of an ionized medium in a magnetic field is usually associated with Appleton-Hartree formula (magnetionic dispersion equation) as in Equation (2.35),which was derived in early 1930's by Sir Edward Appleton and Douglas Hartree (Davies, 1990).:

$$n^2 = (\mu - i\chi)^2$$

$$= 1 - \frac{X}{(1 - jZ) - \left(\frac{Y_T^2}{2(1 - X - jZ)}\right) \pm \left(\frac{Y_T^4}{4(1 - X - jZ)^2} + Y_L^2\right)^{\frac{1}{2}}} \quad (2.35)$$

where μ and χ are the real and imaginary component of the refraction index respectively, while the dimensionless quantities *X*, *Y* and *Z* are defined by the following expressions:

$$X = \frac{\omega_N^2}{\omega^2} \quad (2.36)$$

$$Y = \frac{\omega_B}{\omega} \quad (2.37)$$

$$Z = \frac{v}{\omega} \quad (2.38)$$

$$Y_L = \frac{\omega_L}{\omega} = \frac{\omega_B cos\theta}{\omega} \quad (2.39)$$

$$Y_T = \frac{\omega_L}{\omega} = \frac{\omega_B sin\theta}{\omega} \quad (2.40)$$

where

ω_N : Angular plasma frequency with $\omega_N^2 = \frac{Ne^2}{\varepsilon_0 m}$, N is the electron density, e is the electron charge, ε_0 referred as electric permittivity of free space and m denoted as electron mass,

ω_B : Electron gyrofrequency with $\omega_B = \frac{Be}{m}$, B is the magnetic field strength,

ω : Angular frequency of the propagating wave,

ω_L : Longitudinal component of the electron gyrofrequency,

ω_T : Transverse component of the electron gyrofrequency,

v : Angular collision frequency between electrons and heavier particles,

θ : Angle between the geomagnetic field vector and the propagation wave direction.

A few assumptions on the properties of the medium are considered in the Appleton - Hartree equation (Davies, 1990):

i. The medium is electrically neutral.
ii. Number of positive ions and electrons are equal in the ionosphere with no resultant space charge.
iii. Constant external magnetic field.
iv. Refractive index varies inversely proportional to the ion mass, thus the effect of the heavy ions on the radio waves is negligible during high frequencies

For high frequencies, the refractive index increases with low concentration of electron density. During this period, the collisions are negligible where $\nu \backsimeq 0$, thus $Z \backsimeq 0$ and the Appleton Hartree equation gives the following expression:

$$n^2 \cong \mu^2 = 1 - \frac{X}{1 - \left(\frac{Y_T^2}{2(1-X)}\right) \pm \left(\frac{Y_T^4}{4(1-X)^2} + Y_L^2\right)^{\frac{1}{2}}} \tag{2.41}$$

and when the wave propagates in the direction nearly perpendicular to the magnetic field (i.e. $\theta = 90^{\circ}$), the $Y_L \simeq 0$ and the equation is simplified to

$$\mu^2 \cong 1 - \frac{X}{1 - \left(\frac{Y_T^2}{2(1-X)}\right) \pm \left(\frac{Y_T^4}{4(1-X)^2}\right)^{\frac{1}{2}}} \tag{2.42}$$

The plane polarized wave will be splitted into two waves which can be derived from Equation (2.42). The wave with positive sign "+ve" is known as an ordinary wave:

$$\mu_+^2 \cong 1 - X \tag{2.43}$$

while the one with a negative sign "-ve" is referred as an extraordinary wave:

$$\mu_-^2 \cong 1 - \frac{X(1-X)}{1-X-Y_T^2} \tag{2.44}$$

Among the two waves, the refractive index of extraordinary wave depends on the magnetic field. However according to Langley (2000) when the frequency of the radio waves are significantly higher than the plasma frequency (i.e. $f \gg f_N$), the effect of the magnetic field is ignorable. In this case, the expressions of both ordinary and extra ordinary waves will be same:

$$\mu_\pm^2 \cong (1 - X) \tag{2.45}$$

Since, $\mu = \sqrt{1-X}$ and the expansion of the real part gives (using a Binomial expansion):

$$\mu \cong (1-X)^{\frac{1}{2}}$$

$$\cong 1 + \frac{1}{2}(-X) + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)}{2!}(-X)^2 + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{3!}(-X)^3 + \cdots$$

$$\cong 1 - \frac{1}{2}X - \frac{1}{8}X^2 - \frac{1}{16}X^3 - \cdots$$

Neglecting higher powers of X

$$\mu \approx \left(1 - \tfrac{1}{2}X\right) \tag{2.46}$$

and substitute the $X = \frac{\omega_N^2}{\omega^2} = \frac{Ne^2}{\varepsilon_0 m(2\pi f)^2}$ the earlier expression in Equation (2.46) gives

$$\mu = 1 - \frac{Ne^2}{8\pi^2 \varepsilon_0 m f^2} \tag{2.47}$$

where N is the electron density (electrons m^{-3}), e is the electron charge (-1.602×10^{-19} C), ε_0 referred as electric permittivity of free space (8.854×10^{-12} Fm^{-1}), m denoted as electron mass (9.107×10^{-31} kg), and f is signal frequency (Hz). Substituting the respective values in Equation (2.47) and the refractive index or also known as phase refractive index (n_ϕ) gives

$$n_\phi = \mu_\phi = 1 - \frac{40.3N}{f^2} \tag{2.48}$$

By substituting Equation (2.48) into Equation (2.34) and the group refractive index can be re-written as:

$$n_g = \mu_g = 1 + \frac{40.3N}{f^2} \tag{2.49}$$

Substitution of Equation (2.48) and (2.49) into Equation (2.32) and (2.33), respectively, gives the phase velocity as:

$$v_\phi = \frac{c}{1 - \frac{40.3N}{f^2}} \tag{2.50}$$

and group velocity as:

$$v_g = \frac{c}{1 + \frac{40.3N}{f^2}} \tag{2.51}$$

The dispersive effect of the ionosphere shows differences between the phase and group refractive indices as well as in the velocities.

2.8.2 Derivation of Total Electron Content from ionospheric refraction

The ionosphere's refractive index is an important and basic quantity in deriving the ionosphere TEC. The velocity of radio signals within the ionosphere is mainly contributed by varying refractive index as the radio waves traverse through different layers or regions in the ionosphere. Therefore, the signal along the ray path between satellite and receiver are affected by the cumulative effect of the ionosphere and the time required by the signal, *t* to reach a receiver on the ground can be determined by integrating the overall path of the signal as in Equation (2.52):

$$t = \int_{Tx}^{Rx} \frac{n}{c} dx \tag{2.52}$$

where Rx is the receiver on the Earth's surface, Tx is the satellite, *n* is the refractive index, which can be represented either as the phase or the group refractive index, *c* is the speed of light, and *dx* is path of the radio signal travelled.

The geometric range or distance, ρ of the signal is equal to speed of light multiplied by time required by the radio signal (Langley, 2000; Hofmann-Wellenhof et al., 2001) and can be expressed in the form:

$$\rho = tc = \int_{Tx}^{Rx} n\, dx \tag{2.53}$$

The measured geometric phase range can be derived by substituting the expression of phase refractive index, $n_\phi = 1 - \frac{40.3N}{f^2}$ in the Equation (2.53):

$$\begin{aligned}
\rho_\phi &= \int_{Tx}^{Rx} n_\phi\, dx \\
&= \int_{Tx}^{Rx} \left(1 - \frac{40.3N}{f^2}\right) dx \\
&= \int_{Tx}^{Rx} dx - \int_{Tx}^{Rx} \frac{40.3N}{f^2}\, dx \\
&= \int_{Tx}^{Rx} dx - \frac{40.3}{f^2} \int_{Tx}^{Rx} N\, dx
\end{aligned} \tag{2.54}$$

and the measured geometric group range can be obtained by inserting the expression of phase refractive index, $n_g = 1 + \frac{40.3N}{f^2}$ in the Equation (2.53):

$$\begin{aligned}
\rho_g &= \int_{Tx}^{Rx} n_g\, dx \\
&= \int_{Tx}^{Rx} dx + \frac{40.3}{f^2} \int_{Tx}^{Rx} N\, dx
\end{aligned} \tag{2.55}$$

where $\int_{Tx}^{Rx} dx$ is the true satellite-receiver geometric range (ρ) which is obtained when $n = 1$ and $\int_{Tx}^{Rx} N\, dx$ is the integrated electron density along the signal path and known as total electron content (TEC).

The ionospheric distance correction, $\Delta\, d_{ion}$ is determined by differencing the measured (phase or group) and geometric ranges and the expression can be expressed in the form (Alizadeh et al., 2013):

$$\Delta d_{ion} = \rho_{\phi/g} - \rho = \mp \frac{40.3}{f^2} TEC \qquad (2.56)$$

where Δd_{ion} is the ionospheric distance correction or ionospheric path delay, "$-ve$" is the sign for carrier phase measurements, "$+ve$" is the sign for pseudo-range measurements. The above expression concludes that the existence of ionosphere reduces the carrier phase measurements (phase is advanced) and increases the pseudo-range measurements (signal is delayed). Both the measurements generate the same results (dimension of length), but with opposite signs.

2.8.3 Derivation of Total Electron Content from GPS measurements

In practice, GPS does not provide direct measurements of ionospheric parameters, but measurements from which ionospheric parameters can be estimated. Single or dual frequency GPS receiver is used to obtain and record the GPS measurements. A dual-frequency is used to estimate TEC, which can be derived through the differences between both the carrier phase (Φ) and pseudo-range (P) measurements at L-band frequencies, L1 and L2. The advantage of dual-frequency P-code in the GPS is found to be a very promising method in computing and correcting the ionospheric range errors (ionospheric path delay), by combining the pseudo-ranges observed on L1 and L2 (Habarulema, 2010). From this ionospheric delay, valuable information on the properties, temporal and spatial variations of the ionosphere can be inferred (e.g. the ionization level of the

ionosphere). Therefore, GPS is a predominant technique for TEC measurement.

Basically, the GPS derived TEC data can be computed from the dual frequency based on either pseudo-range (P) or carrier phase (Φ) measurements and can be expressed mathematically as follows (Gao & Liu, 2002; Ya'acob et al., 2009):

Pseudo-range measurement at L1 and L2 frequencies:

$$P_1 = \rho + c\left(\varepsilon_R(t) - \varepsilon_S(T)\right) + \Delta_{Orbit} + \Delta_{Trop} + \Delta I_{P_1} + bS_{P_1} - bR_{P_1} + \Delta_{multipath,P_1} + \varepsilon(P_1) \quad (2.57)$$

$$P_2 = \rho + c\left(\varepsilon_R(t) - \varepsilon_S(T)\right) + \Delta_{Orbit} + \Delta_{Trop} + \Delta I_{P_2} + bS_{P_2} - bR_{P_2} + \Delta_{multipath,P_2} + \varepsilon(P_2) \quad (2.58)$$

Carrier phase measurement at L1 and L2 frequencies:

$$\Phi_1 = \rho + c\left(\varepsilon_R(t) - \varepsilon_S(T)\right) + \Delta_{Orbit} + \Delta_{Trop} + \lambda_1 N_1 - \Delta I_{\Phi_1} + bS_{\Phi_1} - bR_{\Phi_1} + \Delta_{multipath,P_1} + \varepsilon(\Phi_1) \quad (2.59)$$

$$\Phi_2 = \rho + c\left(\varepsilon_R(t) - \varepsilon_S(T)\right) + \Delta_{Orbit} + \Delta_{Trop} + \lambda_2 N_2 - \Delta I_{\Phi_2} + bS_{\Phi_2} - bR_{\Phi_2} + \Delta_{multipath,P_2} + \varepsilon(\Phi_2) \quad (2.60)$$

where

ρ : is the line of sight range between the receiver and satellite (m),

c : is the speed of light in vacuum (ms^{-1})

$\varepsilon_R(t), \varepsilon_S(T)$: are the receiver and satellite clock errors with respect to the GPS time (s), respectively,

Δ_{Orbit} : is the satellite orbit error (m),

Δ_{Trop} : is the tropospheric induced error (m),

$\Delta I_{P_1}, \Delta I_{P_2}$: are the pseudo-range ionospheric induced errors (m) at L1 and L2 frequencies, respectively,

$\Delta I_{\Phi_1}, \Delta I_{\Phi_2}$: are the carrier ionospheric induced errors (m) at L1 and L2 frequencies, respectively,

λ_1, λ_2 : are the wavelength of the signal (m) at L1 and L2 frequencies, respectively,

N_1, N_2 : are the carrier phase integer ambiguities between satellite and receiver (cycle) at L1 and L2 frequencies, respectively,

bS_{P_1}, bS_{P_2} : are the pseudo-range satellite hardware delays (m) at L1 and L2 frequencies, respectively,

bS_{Φ_1}, bS_{Φ_2} : are the carrier phase satellite hardware delays (m) at L1 and L2 frequencies, respectively,

bR_{P_1}, bR_{P_2} : are the pseudo-range GPS receiver hardware delays (m) at L1 and L2 frequencies, respectively,

bR_{Φ_1}, bR_{Φ_2} : are the carrier phase GPS receiver hardware delays (m) at L1 and L2 frequencies, respectively,

$\Delta_{multipath,P_1}$, $\Delta_{multipath,P_2}$: are the pseudo-range multipath effects (m),

$\Delta_{multipath,\Phi_1}$, $\Delta_{multipath,\Phi_2}$: are the carrier phase multipath effects (m),

$\varepsilon(P_1), \varepsilon(P_2)$: are the pseudo-range measurement noises (m),

$\varepsilon(\Phi_1), \varepsilon(\Phi_2)$: are the carrier phase measurement noises (m).

By differencing the code observations at L1 (1575.42 MHz) and L2 (1227.60 MHz) frequencies and by ignoring the effect of multipath and other measurement errors (e.g. thermal noise), the pseudo-range measurements can be described as follows:

$$P_2 - P_1 = \Delta I_{P_2} - \Delta I_{P_1} + bS_{P_2} - bS_{P_1} - (bR_{P_2} - bR_{P_1}) \tag{2.61}$$

where

$I = \Delta I_{P_2} - \Delta I_{P_1}$: pseudo-range ionospheric path delay,

$bS_P = bS_{P_2} - bS_{P_1}$: differential satellite hardware delay (satellite inter-frequency bias) between L1 and L2 frequencies,

$bR_P = bR_{P_2} - bR_{P_1}$: differential receiver hardware delay (receiver inter-frequency bias) between L1 and L2 frequencies,

while the carrier phase measurements becomes:

$$\Phi_1 - \Phi_2 = \lambda_1 N_1 - \lambda_2 N_2 + \Delta I_{\Phi_2} - \Delta I_{\Phi_1} + bS_{\Phi_1} - bS_{\Phi_2} - (bR_{\Phi_1} - bR_{\Phi_2}) \tag{2.62}$$

where

$I = \Delta I_{\Phi_2} - \Delta I_{\Phi_1}$: phase ionospheric path delay,

$\lambda N = \lambda_1 N_1 - \lambda_2 N_2$: differential integer ambiguities,

$bS_\Phi = bS_{\Phi_1} - bS_{\Phi_2}$: differential satellite hardware delay (satellite inter-frequency bias) between L1 and L2 frequencies,

$bR_\Phi = bR_{\Phi_1} - bR_{\Phi_2}$: differential receiver hardware delay (receiver inter-frequency bias) between L1 and L2 frequencies,

Both the differencing pseudo-range and carrier phase measurements eliminate the geometric range, the receiver and satellite clock errors, the satellite orbital error, as well as the delay induced by the troposphere and eventually provide the "geometry-free" linear combination. The difference in the code measurements between the two frequencies, L1 and L2, can be derived by substituting the expression of Equation (2.56) into Equation (2.61) and Equation (2.62) and express as below:

$$\begin{aligned} P_2 - P_1 &= I + bS_P - bR_P \\ &= \frac{40.3}{f^2} TEC + bS_P - bR_P \end{aligned}$$

$$P_2 - P_1 = 40.3 \left[\frac{1}{L2^2} - \frac{1}{L1^2}\right] TEC + bS_P - bR_P \tag{2.63}$$

and

$$\begin{aligned} \Phi_1 - \Phi_2 &= \lambda N + I + bS_P - bR_P \\ &= -\frac{40.3}{f^2} TEC + \lambda N + bS_P - bR_P \end{aligned}$$

$$\Phi_1 - \Phi_2 = -40.3 \left[\frac{1}{L2^2} - \frac{1}{L1^2}\right] TEC + \lambda N + bS_P - bR_P \tag{2.64}$$

Commonly, TEC from the pseudo-range measurements has relatively high noise compared to TEC from the carrier phase measurements, though the carrier phase measurements are introduced to unknown ambiguity, where the signal from L1 and L2 may not be easily differentiated each other. In the pseudo-range measurements, TEC is ambiguity-free. Thus, combining both the code measurements linearly provide more accurate TEC measurements although the output TEC is not the absolute TEC (Habarulema et al., 2007). To estimate the absolute TEC, the differential satellite and

receiver biases should be taken into consideration concurrently with the ionospheric delay parameters. In the following chapter, the errors are estimated for a single receiver station at Parit Raja to obtain the absolute TEC and the values are used in ionospheric TEC modelling.

2.9 SUMMARY

This chapter discussed briefly on the existence of Earth's ionosphere and the major factors that influence its variability, e.g. solar and magnetic activities, time and latitudinal variations of the ionosphere. Furthermore, the chapter presented the basic definition of ionospheric TEC and its effects on electromagnetic waves. A brief literature on the techniques (globally and regionally) that have been used for ionospheric TEC modelling is presented. Besides, a comprehensive discussion on the theories and concepts of neural network and seasonal autoregressive integrated moving average techniques along with their mathematical expressions have been presented. Finally the chapter provided a brief description on the GPS system and described how the GPS signals are influenced by the ionosphere's refractive index. The chapter ended with the derivation of ionospheric TEC from the GPS measurements based on the pseudo range (P) and carrier phase (Φ) observables via L1 and L2 frequencies.