Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

# Hybrid SARIMA-ELM-based Data Traffic Forecasting: The Case of UMTS Network in Addis Ababa, Ethiopia

**By**

**Getinet Tesfaye**

**Advisor**

**Dr. -Ing. Dereje Hailemariam**

A Thesis Submitted to the Addis Ababa University

Addis Ababa Institute of Technology in Partial Fulfillment of the Requirements for the

Degree of

Masters of Science in Telecommunications Engineering

October 25, 2018

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

# Hybrid SARIMA-ELM-based Data Traffic Forecasting: The Case of UMTS Network in Addis Ababa, Ethiopia

**By**

**Getinet Tesfaye**

Approval by Board of Examiners

______________________________ ____________

Chairman, school Graduate committee Signature

Committee

Dr. -Ing. Dereje Hailemariam ____________

Advisor Signature

______________________________ ____________

Internal Examiner Signature

______________________________ ____________

External Examiner Signature

# Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in this or any other university, and all sources of materials used for the thesis have been fully acknowledged.

Getinet Tesfaye ______________

Name Signature

Place: Addis Ababa

Date of Submission: _______________

This thesis has been submitted for examination with my approval as a university advisor.

Dr. -Ing. Dereje Hailemariam ______________

Advisor's Name Signature

# Abstract

In Universal Mobile Telecommunications Service (UMTS) network planning, data traffic demand is one critical input in deciding dimension of network elements. Past data collected from deployed UMTS network can be used to forecast future demand. In the context of ethio telecom, the sole telecom service provider in Ethiopia, the future demand forecast is, however, based on number of subscribers growth forecast obtained from marketing section. This approach assumes uniform data demand per subscriber to obtain the total data demand. Understandably, it does not utilize the data growth information which is already available in the network.

Forecasting the traffic demand based on historical data from network can enhance the marketing inputs and the traffic model accuracy. In this regard, taking data from ethio telecom's UMTS network, a prior research has used Seasonal Autoregressive Integrated Moving Average (SARIMA) model to forecast a one month data traffic demand. However, the research did not consider the non-linearity observed in the data traffic. This thesis handles this non-linearity via a hybrid model that accounts the linearity with SARIMA model and the non-linearity via Extreme Learning Machine (ELM) model; here after called the hybrid SARIMA-ELM model.

A one and half year (i.e., from April 2015 – June 2016) data traffic collected from five Radio Network Controllers (RNCs) of the UMTS network in the city of Addis Ababa is used for the forecast. The forecasting performance metrics are: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Square Error (MASE). The results indicate that the hybrid SARIMA-ELM model with SARIMA order of $(0,1,1)(1,0,1)_7$ is selected with 3.75% increase in forecast than SARIMA only model. The outperform SARIMA-MLP, which has the second lower error, with 24.8% percentage error reduction.

**Keywords**— *time series, traffic, prediction, model, forecast, linear, non-linearity, hybrid, residual SARIMA model, UMTS, ELM model*.

# Acknowledgment

This thesis wouldn't have been finalized without the kind support of people who contributed to it. The first and foremost thanks goes to, Dr. -Ing. Dereje H/Mariam, followed by the almighty God, for his advice, giving directions and unreserved feedbacks. I also like to thank Ms. Bethelhem Seifu for providing me valuable insights to the research. Finally, I would like to thank my wife Mrs. Haymanot Andualem for her all time support and Mr. Dereje Gemechu for providing me the data for the UMTS data traffic used as an input for the thesis.

# Table of Content

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| AA | Addis Ababa |
| ACF | Auto Correlation Function |
| AIC | Akakie's Information Criterion |
| AICc | Akakie's Information Criterion controlled |
| AR | Auto Regressive |
| ARIMA | Auto Regressive Integrated Moving Average |
| ARMA | Auto Regressive Moving Average |
| BIC | Bayesian Information Criterion |
| ELM | Extreme Learning Machine |
| EXP | Exponential |
| FARMA | Fractional Auto Regressive Moving Average |
| KPI | Key Performance Indicator |
| MA | Moving Average |
| MAE | Mean Absolute error |
| MAPE | Mean Absolute Percentage Error |
| MASE | Mean Absolute Squared Error |
| MLP | Multi-Layer Perception |
| NN | Neural Network |
| PACF | Partial Auto Correlation Function |
| RMSE | Root Mean Squared Error |
| RNC | Radio Network Controller |
| SARIMA | Seasonal Auto Regressive Moving Average |
| UMTS | Universal Mobile Telecommunication System |

# 1. Introduction

The increase in penetration rate of smart phones and tablets has driven exponential growth of mobile data traffic [1]. From operators’ perspective, this growing mobile data traffic is an opportunity as it leads to revenue maximization. However, two of the resulting challenges are more radio spectrum usage and huge energy consumption. To addresses these challenge operators use options such network optimization and network expansion. The latter case calls for efficient network planning, where one input is accurate traffic forecast.

Forecasts can be obtained in qualitative or time series analysis approach. The qualitative approach is mainly based on judgements and opinions; whereas, time series analysis approach is through the idea that, the evolution in the past will continue into the future. The UMTS data traffic demand is mainly driven by user’s behavior of data usage, tariff and the nation’s development. Thus, exactly knowing where to reach the data traffic volume level in the next periods could be difficult. Hence, a causal relationship is needed. UMTS data traffic is random, impossible to predict what will happen. On the other hand, if it is observed some historic correlations on the data, possible to use these correlations to make forecasts, thus UMTS data traffic forecasting based on past data is preferable [2]. There are different time series forecasting models that can be used to predict network traffic demand based on available historical data, of these: the Box-Jenkins variants of different methodology and Machine Learning of Neural Network variants (NN) are the most commons [3], [4].

Traffic data manifests both trend and seasonality. The trend indicates pattern of gradual change in a condition, output, or process, or an average or general tendency of a series of data points to move in a certain direction over time, represented by a line or curve on a graph. Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar in the year.

These and other mobile data traffic forecasting  models and techniques work well in estimating the data trend and seasonal characteristics, though largely undertake modelling the data across time series observed at the Radio Network Controllers (RNCs). In doing so, most of the models lack the capability of capturing the linearity and non-linearity characteristics of the data and their overall impact on the forecast together. Researches have been conducted to forecast A.A UMTS data traffic through modelling the data only capturing the linear characteristics of the data only; however, the data traffic contains both linear and non-linear characteristics together.

## 1.1 Statement of the Problem

Ethio telecom has a rich big data, manipulating and extracting past traffic data information from it can support the network planning to be optimal and accurate in regards of demand forecast. However, ethio telecom does not utilize this big data to extract the traffic demand information. There is no traffic prediction performed using modeling techniques in ethiotelecom, rather rely on the marketing forecast information for its wireless network planning [6] [7].

Extracting historic traffic demand information from big data to manipulate and use it to forecast the future demand is becoming a trend with telecom service providers globally. Forecasting traffic demand from the available ethio telecom’s big data through hybrid SARIMA-ELM model that can capture the data characteristics (i.e. linearity and non-linearity) is done with this thesis.

## 1.2 Objective

### 1.2.1 General objective

The main goal of the thesis is to propose and investigate hybrid SARIMA-ELM model as an alternative way to forecast UMTS data traffic of Addis Ababa city.

### 1.2.2 Specific objectives

The specific objectives of this thesis are:-

- To review appropriate forecast models for the UMTS data traffic.
- Gather/collect data, pre-process the data, analyze and visualize.
- Finding a model that fits into the data (linear part) and forecast the linear part with this model, then feed this model fit residuals to Auto regressive (AR), Multi-Layer Perception (MLP) and Extreme Learning Machine (ELM) machine learning neural networks to forecast the non-linear part. The resulting models is called hybrid forecast model.
- Compare the forecast performance of the hybrid model with forecast error metrics (RMSE, MAE, MAPE and MASE).
- Select a model based on minimum forecast error.

## 1.3 Methodology

In order to provide measurement data based forecasting, a one year data traffic in an hourly basis, i.e. from April 2015 – June 2016, has been collected from the five Radio Network Controllers (RNCs) of Addis Ababa UMTS network. Considering the peak values during the day, the data was analyzed and based on that an appropriate model has been selected. The weekly seasonality and trend of the data were able to be captured with Seasonal Auto Regressive Moving Average (SARIMA) models. Candidates of the

SARIMA models of different orders are selected based on the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) lag values. From the candidates, a model with minimum Akakie's Information Criterion (AIC), Akakie's Information Criterion Controlled (AICc) and Bayesian Information Criterion (BIC) values were selected. Model fitting and forecasting the future values with the selected model and extract the residuals. The residuals from this selected model fit are taken as an error to be modeled and forecasted with non-linear model forecasting. The forecast from both were added to hybridize the selected SARIMA model with different NN machine learning model variants. Hybrid model with minimum Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Square Error (MASE) were selected. Present the selected forecasting model as an alternative way of forecasting Addis Ababa city UMTS mobile data traffic. R statistical software tool is used for the modelling and processing of the data.

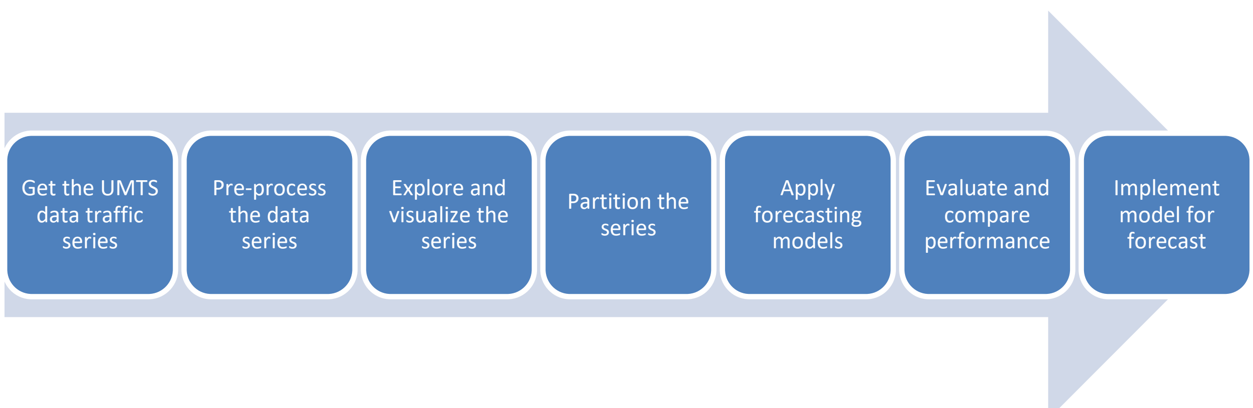


**Figure 1**. Methodology followed

## 1.4 Literature review

Accurate traffic demand forecast is one of the critical inputs to optimal network planning that can be applied based on historical data traffic through different statistical and scientific modeling and forecasting techniques [4]. Ethio telecom usually follows bench marking and traditional approaches for demand forecast. For A.A UMTS Telecom Expansion Project similar approach had been followed [6], [7] and there were no application of known scientific forecasting models and techniques that enables to consider the traffic growth, thus internationally conducted related studies had been reviewed.

Yanhua Yu, et al [5], performed a traffic forecast based on the present traffic data through integrated Auto Regressive Moving Average (ARMA) and Fractional Auto Regressive Moving Average (FARMA) modelling process, on a live UMTS mobile network in a province of china. Thus obtained a prediction result very close to the actual traffic.

Samuel M., et al [11], showed how SARIMA is an alternative way of forecasting Addis Ababa UMTS data traffic based on the past UMTS data traffic analysis. Collected UMTS data traffic from ethio telecom for the duration of October 2015 to June 2016 monitored on a daily bases. The authors analyzed the data and chose SARIMA model. Select the Auto regressive and Moving Average parameters of this model, estimate the orders of the model and forecast the traffic for 30 days using the model. To compare for 5 days SARIMA with two different orders actual value and forecast values compared with their MAPE then SARIMA forecasting model is proposed and applied. From the possible candidates based on AIC, BIC and AICc statistics, SARIMA with order $(2,0,1)x(0,1,1)_7$ is selected to as the, lowest error metrics value, model where a forecast pattern for the next 30 days ahead of the original data was generated. Comparing the forecast values with available data, from June 1, 2016 to June 9, 2016, the maximum error was obtained to be less than 2.62%.The obtained experimental results verifies the effectiveness of the prediction method, which is adequate to be used to forecast UMTS data traffic.

Oladeji E.O, et al. [12], establish the traffic pattern over 24 hours in Nigeria for a period of 5 months, determine the busy hour and also develop a model for traffic forecasting, using an operator switch in the north-central part of Nigeria with 6 Base Station Controllers (BSC). The authors analyzed the obtained data using Artificial Neural

Network (ANN) which maps input to output. Thus able to analyze the data and the output, enabled them to forecast the traffic based on the different busy hour per BSCs.

Roselina S., et al. [31], Constructed a Hybrid linear and non-linear time series data modeling (GRAN-ARIMA) approach & show its out-performance from the individual & conventional hybrid TS data modeling approach for forecasting, the authors used two different data sets, small scale data set of 13 observations of annual China gross grain crop yields from 1990 to 2003 with 10 affecting factors and Large scale data set of Kuala Lumpur Stock Exchange (KLSE) containing 200 observations of daily KLSE close price from 4 January 2005 till 21$^{St}$ October 2005 with 14 affecting factors. Used performance metrics RMSE, MSE, MAD & MAPE for performance evaluation of the models and got a result GRANN_ARIMA is always better than MR, ARIMA, and ANN model hence can handle linear & non-linear behavior of the data set.

Mergani K., et al. [13] used non-linear time series data that has uncertain behavior and combine from ARIMA, EXP & MLP in a novel hybrid model for greater performance & accuracy of forecasting used Additive & weighting combination. They used a data of exchange Rate of euro from commercial bank of Sudan of 6 months and applied two methods (additive and linear regression) to combine forecasts made by EXP, ARIMA, (linear) and MLP (non-linear) models &investigate performance based on MSE, RMSE, MAE, MAPE &SD as performance metrics. With the single models, MLP is best; however with the additive method (MLP+EXP &MLP+ARIMA), MLP+ARIMA is best and from the linear regression method (MLP+EXP+ARIMA) is best and overall (MLP+EXP+ARIMA) is superior in accuracy & performance in forecasting exchange rate of euro in Sudan.

Xin Dong, et al. [14], with their research used the autoregressive integrated moving average ARIMA model and exponential smoothing model to predict the throughput in a single cell and whole region in an LTE network. They used a data set includes records of Internet downloads and uploads in Hong Kong collected from 1352 cell sites across the city over 21 days between February and March 2014. Modeled the throughput as a time series and then predicted using an ARIMA model and exponential smoothing method. They analyze with two practical scenarios, in the first scenario, each cell is divided into regions, and the throughput of an entire region is predicted. In the second scenario, the throughput of a single cell is predicted according to historical data. They analyzed the throughput data in a weekdays and weekends base, for both scenarios they chose the ARIMA model of different orders. With their final result The ARIMA model is better than exponential smoothing for predicting throughput on weekday in a whole region, and exponential smoothing model is much better than the ARIMA model for predicting throughput on weekends in a whole region. Exponential smoothing is more accurate than the ARIMA model for predicting throughput in a single cell. Throughput prediction based on time series models can be used in the design, management, planning, and optimization of networks.

Huimin A., et al. [15], proposes a Block Regression (BR) model for mobile traffic forecasting. With their work they used a data set from a Chinese company containing the hourly data traffic of about 1000 GSM base stations from a major city with period of one month. They clean some faulty data and analyzed, chose block regression (BR) model and made a differentiation, employ the sliding window method to extract historic information and formulate the feature vectors for the BR model then made a normalization which followed by the forecasting. Finally they evaluate the BR method comparing it with Seasonal ARIMA and Block Regression models based on Normalized Root Mean Square Error (NRMSE) and propose a Block Regression model for traffic forecasting in wireless communication networks.

All reviewed papers dealt with forecasting of a data, based on historical data collected. In all the papers except paper [31] used a single linear or non-linear models; however combining models in a hybrid mode increases a forecast performance as shown on [31].

## 1.5 Scope and Limitation

### 1.5.1 Scope of the thesis

Both the GSM & UMTS mobile network technologies are deployed in the whole of Ethiopia; however this thesis only focuses on Addis Ababa and its UMTS mobile network. The UMTS mobile technology provides both data and voice traffic; however, this thesis work focuses only with the data service part hence the UMTS technology services are more to avail the data service. This thesis also considered only the UMTS data traffic demand forecasting of Addis Ababa, as planning input perspective.

### 1.5.2 Limitation of the thesis

There are lots of factors that can be considered and make the forecasting resolution (accuracy) high, like customer behavior modelling and treating the data as multivariate; however, this thesis treated the UMTS data traffic as univariate data, considering the allocated time. Working with the multivariate data considering many factors needs more time because of the complexity.

## 1.6 Contribution of the Thesis

This thesis used a hybrid SARIMA-ELM, with minimum forecast error, model to forecast A.A UMTS data network based on historical data collected. This model increases the performance of forecasting than individual or single models.

Thus, ethio telecom, to use this hybrid SARIMA-ELM with SARIMA $(0,1,1)(1,0,1)_7$ that captures both linear and non-linear characteristics of the UMTS data traffic and its minimum forecast error.

## 1.7 Thesis organization/ Layout

The thesis paper contains 5 chapters. Chapter one deals with the main intention of the thesis. In this chapter the statement of the problem, objective of the study, limitations, scope and contribution of the thesis are stated. Chapter two discuss about the UMTS network and infrastructures. Chapter three deals broadly about the UMTS data traffic types, behavior, characteristics, and stochastic process of this data type as well preparation of the data for modeling and forecast. In this chapter linear, non-linear and hybrid models for forecasting is also explained. Chapter four is the main body of the thesis which the hybrid SARIMA-ELM based forecasting, the system model used, the UMTS traffic data set, model validations and forecasting results discussed. Chapter five explains the findings, concluded by the result of this thesis hybridizing the linear and non-linear model forecasts and implementation of the model to forecast one month ahead values. Finally recommendations and future work are done.

# 2. UMTS network

## 2.1 Introduction

After GSM developed and deployed, the eyes of the development community started to look at the next cellular developments which would provide greater more functionality and greater levels of efficiency. The UMTS 3G history shows how these basic ideas turned into reality and changed the way in which mobile telecommunications was used.

The history shows that despite many setbacks, UMTS was able to become established as the major 3G technology providing new standards in cellular telecommunications performance, functionality, and convenience. UMTS became the dominant 3G technology, setting the foundations for a single worldwide 4G standard in future years [8], [9].

**3G beginnings and IMT-2000**

The International Mobile Telecommunications-2000, IMT-2000 standard is actually a family of standards for third generation (3G) wireless communications. It defines the broad outlines and requirements for standards that can be called 3G standards. It was set in place by the International Telecommunications Union (Radio Communications section), ITU-R [10].

In the 1980s work started on looking at, what was termed in the ITU-R the "Future Public Land Mobile Telecommunications System". However with the deployment on GSM and other 2G technologies the impetus for the development of the next generation system was not present.

It was not until the early 1990s that progress was seen. A working group was set up and also the 1992 World Administrative Radio Conference (WARC'92) allocated 230 MHz of spectrum between 1885 and 2025 and 2110 and 2200 MHz [10].

A number of organizations recognized the need for a global standard for the next generation of mobile telecommunications services. ETSI in Europe moving towards what they termed their Universal Mobile Telecommunications System, UMTS and in Japan the forerunner of the Association of Radio Industries and Businesses, ARIB undertaking a study. To enable a single standard to be adopted the ITU-R requested each regional Standards Development Organisation (SDO) to submit proposals for a Radio Transmission Technology [8], [10].

As a result, between 1996 and 1998 companies and regional SDOs worked towards their proposal submissions.

A total of 17 different proposals were submitted. Of these eleven were for terrestrial systems and the remaining six were for satellite systems. The evaluation of the proposals was completed during 1998 but during early 1999 it was necessary to gain some form of consensus. Once this was complete, by the end of 1999 the specification for the radio Transmission Technology was released by the end of 1999.

Although many proposals were submitted there were several that were considerably more important than others. These included:

***UMTS / WCDMA***: The Universal Mobile Telecommunications System using wideband CDMA was the successor to the highly successful GSM system that was initially deployed around Europe, but was spreading rapidly worldwide.

***CDMA2000***: This scheme was the successor to the cdmaOne system defined under Interim Standard IS-95 which was the first system to be deployed using CDMA technology.

***TDS-CDMA***: This was a scheme developed in China that adopted many elements of the GSM / UMTS technology but was optimized for Time Division Duplex.

The Universal Mobile Telecommunications System represents a complete system. That means, it includes cell phones (and other mobile equipment), the radio infrastructure needed to provide call and data session services, the core network equipment for transporting user calls and data, the billing systems, and the security systems, among others.

Because UMTS has deep GSM roots, it is sometimes called 3GSM. However, as the name W-CDMA implies, it also makes extensive use of CDMA technology.

Upgraded UMTS networks around the globe are able to provide fast download speeds of up to 14 Mbps via the HSDPA (High-Speed Downlink Packet Access) protocol. Faster uplink speeds of up to 5.7 Mbps are currently underway via the HSUPA (High-Speed Uplink Packet Access) protocol. Both HSDPA and HSUPA are part of a larger family of protocols known as High Speed Packet Access (HSPA).

Because UMTS is built on GSM networks, it enjoys the same global roaming capabilities. Practically all UMTS phones are capable of switching to GSM mode. That means, if you're using a UMTS device and happen to wander away from a UMTS network and into a GSM network, you can still avail of cellular services using the same phone (provided of course the necessary roaming agreements are in place).

Like their GSM predecessors, UMTS phones also come with an upgraded SIM (Subscriber Identity Module) known as the USIM (Universal SIM). UTMS phones can work with either SIMs or USIMs.

With the wireless industry now moving from 3G to 4G, UMTS serves as the basis of the 3GPP's new set of radio technologies, known as Long Term Evolution (LTE).

Networks upgrading from GSM to the Universal Mobile Telecommunications System are able to reuse a number of network elements, including: the Home Location Register, Visitor Location Register, Mobile Switching Center, and the Authentication Center, to name some. However, a new Base Station Controller and Base Transceiver Station is required. In most cases, 3G and 2G networks will be made to operate side by side.

## 2.2 UMTS network architecture

The UMTS 3G architecture is required to provide a greater level of performance to that of the original GSM network. However as many networks had migrated through the use of GPRS and EDGE, they already had the ability to carry data. Accordingly many of the elements required for the WCDMA / UMTS network architecture were seen as a migration [9]. This considerably reduced the cost of implementing the UMTS network as many elements were in place or needed upgrading.

With one of the major aims of UMTS being to be able to carry data, the UMTS network architecture was designed to enable a considerable improvement in data performance over that provided for GSM.

**3G UMTS network constituents**

The UMTS network architecture can be divided into three main elements [8]:

*User Equipment (UE):* The User Equipment or UE is the name given to what was previous termed the mobile, or cellphone. The new name was chosen because the considerably greater functionality that the UE could have. It could also be anything between a mobile phone used for talking to a data terminal attached to a computer with no voice capability.

*Radio Network Subsystem (RNS):* The RNS also known as the UMTS Radio Access Network, UTRAN, is the equivalent of the previous Base Station Subsystem or BSS in GSM. It provides and manages the air interface fort the overall network.

Core Network: The core network provides all the central processing and management for the system. It is the equivalent of the GSM Network Switching Subsystem or NSS.

The core network is then the overall entity that interfaces to external networks including the public phone network and other cellular telecommunications networks [8].

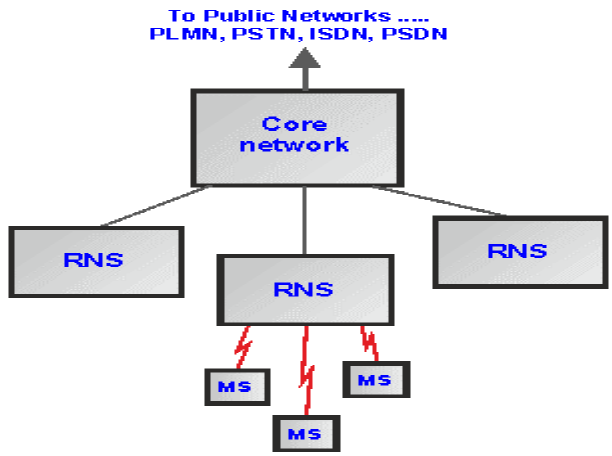


**Figure 2**. UMTS Network Architecture Overview

**3G UMTS Radio Network Subsystem**

This is the section of the 3G UMTS / WCDMA network that interfaces to both the UE and the core network. The overall radio access network, i.e. collectively all the Radio Network Subsystem is known as the UTRAN UMTS Radio Access Network.

The radio network subsystem is also known as the UMTS Radio Access Network or UTRAN [10].

***3G UMTS Core Network***

The 3G UMTS core network architecture is a migration of that used for GSM with further elements overlaid to enable the additional functionality demanded by UMTS.

In view of the different ways in which data may be carried, the UMTS core network may be split into two different areas:

*Circuit switched elements*: These elements are primarily based on the GSM network entities and carry data in a circuit switched manner, i.e. a permanent channel for the duration of the call.

*Packet switched elements*: These network entities are designed to carry packet data. This enables much higher network usage as the capacity can be shared and data is carried as packets which are routed according to their destination.

Some network elements, particularly those that are associated with registration are shared by both domains and operate in the same way that they did with GSM.

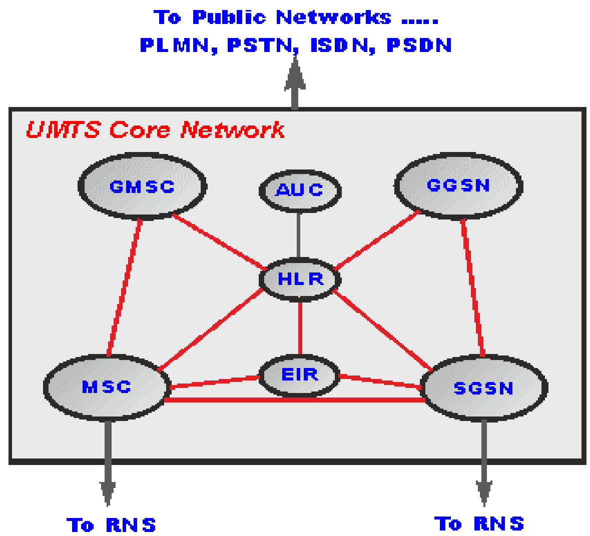


**Figure 3**. UMTS core network

***Circuit switched elements***

The circuit switched elements of the UMTS core network architecture include the following network entities:

*Mobile switching center (MSC)*: This is essentially the same as that within GSM, and it manages the circuit switched calls under way.

Gateway MSC (GMSC): This is effectively the interface to the external networks.

***Packet switched elements***

The packet switched elements of the 3G UMTS core network architecture include the following network entities:

*Serving GPRS Support Node (SGSN)*: As the name implies, this entity was first developed when GPRS was introduced, and its use has been carried over into the UMTS network architecture. The SGSN provides a number of functions within the UMTS network architecture.

Mobility management When a UE attaches to the Packet Switched domain of the UMTS Core Network, the SGSN generates MM information based on the mobile's current location.

*Session management*: The SGSN manages the data sessions providing the required quality of service and also managing what are termed the PDP (Packet data Protocol) contexts, i.e. the pipes over which the data is sent.

Interaction with other areas of the network: The SGSN is able to manage its elements within the network only by communicating with other areas of the network, e.g. MSC and other circuit switched areas.

*Billing*: The SGSN is also responsible billing. It achieves this by monitoring the flow of user data across the GPRS network. CDRs (Call Detail Records) are generated by the SGSN before being transferred to the charging entities (Charging Gateway Function, CGF).

*Gateway GPRS Support Node (GGSN)*: Like the SGSN, this entity was also first introduced into the GPRS network. The Gateway GPRS Support Node (GGSN) is the central element within the UMTS packet switched network. It handles inter-working between the UMTS packet switched network and external packet switched networks, and can be considered as a very sophisticated router. In operation, when the GGSN receives data addressed to a specific user, it checks if the user is active and then forwards the data to the SGSN serving the particular UE.

***Shared elements***

The shared elements of the 3G UMTS core network architecture include the following network entities:

*Home location register (HLR)*: This database contains all the administrative information about each subscriber along with their last known location. In this way, the UMTS network is able to route calls to the relevant RNC / Node B. When a user switches on their UE, it registers with the network and from this it is possible to determine which Node B it communicates with so that incoming calls can be routed appropriately. Even when the UE is not active (but switched on) it re-registers periodically to ensure that the network (HLR) is aware of its latest position with their current or last known location on the network.

*Equipment identity register (EIR)*:   The EIR is the entity that decides whether a given UE equipment may be allowed onto the network. Each UE equipment has a number known as the International Mobile Equipment Identity. This number, as mentioned above, is installed in the equipment and is checked by the network during registration.

*Authentication center (AuC)*:   The AuC is a protected database that contains the secret key also contained in the user's USIM card.

Physical layer within UMTS / WCDMA is totally different to that employed by GSM. It employs a spread spectrum transmission in the form of CDMA rather than the TDMA transmissions used for GSM. Additionally it currently uses different frequencies to those allocated for GSM.

The UTRA, UMTS radio access is the technology that is the radio interface, and the network, or UMTS Radio Access Network is known as the UTRAN. Sometimes the UTRAN may also be known as the Radio Network Subsystem, or RNS.

UMTS radio access network, UTRAN

The UMTS Radio Access Network, UTRAN, or Radio Network Subsystem, RNS comprises two main components:

*Radio Network Controller, RNC*:   This element of the UTRAN / radio network subsystem controls the Node Bs that are connected to it, i.e. the radio resources in its domain. The RNC undertakes the radio resource management and some of the mobility management functions, although not all. It is also the point at which the data encryption / decryption is performed to protect the user data from eavesdropping.

*Node B*:   Node B is the term used within UMTS to denote the base station transceiver. This part of the UTRAN contains the transmitter and receiver to communicate with the

UEs within the cell. It participates with the RNC in the resource management. NodeB is the 3GPP term for base station, and often the terms are used interchangeably [9].

In order to facilitate effective handover between Node Bs under the control of different RNCs, the RNC not only communicates with the Core Network, but also with neighbouring RNCs.

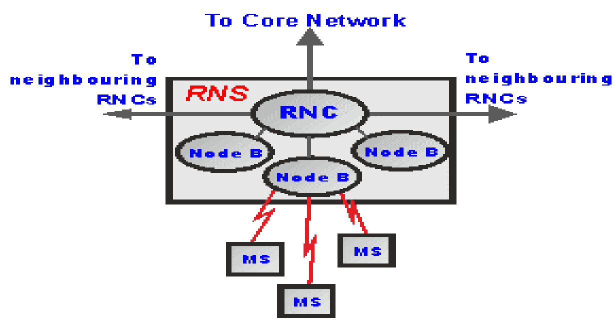


**Figure 4**. UMTS UTRAN architecture

# 3. Mobile Data Traffic Forecasting

A three-step iterative procedure is used to build a linear model to forecast mobile data traffic. First, a tentative model is identified through analysis of historical data. Second, the unknown parameters of the model are estimated. Third, through residual analysis, diagnostic checks are performed to determine the adequacy of the model, or to indicate potential improvements. However in non-linear machine learning models the procedure is to partition the data set in to training and test parts then learn the network with the training set then pass the test data part for the modeling [19],[21].

### *Model identification*

Model identification efforts start with preliminary efforts in understanding the type of process from which the data is coming and how it is collected. The process's perceived characteristics and sampling frequency often provide valuable information in this preliminary stage of model identification [4]. Recommendations that are 50 or preferably more observations should be initially considered. Before engaging in rigorous statistical model-building efforts, analysis and "creative" plotting of the data, such as the simple time series plot and scatter plots of the time series data $y_t$ versus $y_{t-1}$, $y_{t-2}$ and so on is a prerequisite. Simple time series plots should be used as the preliminary assessment tool for stationarity and components identification. If non-stationarity is suspected, the time series plot of the first or $d^{th}$ difference should also be considered. The unit root test by Dickey and Fuller [1979] can also be performed to make sure that the differencing is indeed needed. Once the stationarity can be presumed, the sample Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) of the time series of the original time series (or its $d^{th}$ difference if necessary) should be obtained [19]. Depending on the nature of the autocorrelation, the first 20-25 sample autocorrelations and partial autocorrelations should be sufficient.

The identification of ARMA models would require more care, as both the ACF and PACF will exhibit exponential decay and/or damped sinusoid behavior.

### *Parameter estimation*

There are several methods such as methods of moments, maximum likelihood, and least squares that can be employed to estimate the parameters in the tentatively identified model, most ARIMA models are linear models and require the use of a linear model fitting procedure. However, this is usually automatically performed by sophisticated software packages such as R, JMP etc.…. In some software packages, the user may have the choice of estimation method and can accordingly choose the most appropriate method [21].

### *Diagnostic checking*

After a tentative model has been fit to the data, examining its adequacy and, if necessary, suggest potential improvements may needed. This is done through residual analysis. The residuals for an ARIMA (p, d, q) process can be obtained through extraction of the fitted model errors.

If the specified model is adequate and hence the appropriate orders p and q are identified, it should transform the observations to a white noise process. Thus the residuals should behave like white noise. If the model is appropriate, then the residual sample autocorrelation function should have no structure to identify. That is, the autocorrelation should not differ significantly from zero for all lags greater than one [19], [21].

## 3.1 Mobile data Traffic Types

Mobile traffic data can be obtained from the core network or access networks. Data traffic measured at the core network is an aggregate and includes the GSM hence both GSM and UMTS systems share the core network whereas if measured at the RNCs it only reflects the UMTS data. The UMTS data traffic which have been obtained by measuring a phenomenon at regular intervals of time on the RNCs known as time series data. The term time series refers to series of data $y_t$, $y_{t-1}$, $y_{t-2}$,…, $y_{t-n}$ collected at a successive discrete sequence of time t in which the analysis can be handled with a stochastic process. The objective of time series analysis is to infer the characteristics of the stochastic process from the data sample (i.e. partial realization of the process) and any additional information we have about the process [4].

## 3.2 Characteristics of Time series data

### 3.2.1 Time series data types and components

Time series data can be univariate (single variable) or multivariate (multiple variable) with linear or non-linear characteristics and can be categorized as discrete, continuous, repetitive, periodic, deterministic or non-deterministic. Discrete time series data is a data which exists only at discrete instants of time. Continuous time series is the one which exists at all instants of time during which it occurs [19]. Repetitive time series is the one which contains a pattern that recurs for all time over which the time series exists. Periodic time series is a repetitive time series in which the repetition occurs at uniformly spaced time intervals. The deterministic time series is one which can be expressed explicitly by an analytic expressions with no random or probabilistic aspects. In mathematical terms, it can be described exactly for all time in terms of a Taylor series expansion provided that all its derivatives are known at some arbitrary time. Its past and future values are completely specified by the values of these derivatives at that time. If so, then we can always predict its future behavior and state how it behaved in the past. A nondeterministic time series is one which cannot be described by an analytic expression. It has some random aspect that prevents its behavior from being described explicitly. A time series may be nondeterministic because:

1. All the information necessary to describe it explicitly is not available, although it might be in principle.

Or

2. The nature of the generating process is inherently random.

Since nondeterministic time series have a random aspect, they follow probabilistic rather than deterministic laws. Random data are not defined by explicit mathematical relations, but rather in statistical terms, i.e. by probability distributions and averages of various forms, such as means and variances.

The factors that are responsible for bringing about changes in the data such as trend, seasonality, cycle and irregularity are called components of time series data. These play the main role in the analysis of the data described as below:

- ***Trend***: - It is a smooth, regular and long-term movement of the series either in incremental or detrimental style.
- ***Seasonality***: - repetition of peak data in the series which arises from seasonal effects such as tariff reduction times, holidays, occasions and weekends.
- ***Cycle: -*** a predictable repetition of data behavior in a cyclic manner like sinusoid signal wave.
- ***Irregularity***: - informal or random cycles and seasonality appearing in an irregular manner on the series and not predictable that looks like a noise data series.

## 3.2.2 Time series data analysis

When a model fitted to a given time series and the corresponding parameters are estimated using the collected data values, the procedure of fitting the time series to the model is termed as *Time Series Analysis* [20]. The procedure comprises methods that attempt to understand the nature of the series is often useful for future forecasting.

In time series forecasting, past observations are collected and analyzed to develop a suitable mathematical model which captures the underlying data generating process for the series [20], [21]. The future events are then predicted using the model. This approach is particularly useful when there is not much knowledge about the statistical pattern followed by the successive observations or when there is a lack of a satisfactory explanatory model. Time series forecasting has important applications in various fields including Telecommunications. Valuable strategic decisions and pre-cautionary measures are taken based on the forecast results. Thus making a good forecast, i.e. fitting an adequate model to a time series is very important.

### 3.2.3 Stochastic behavior of time series data

A stochastic (random) process is a statistical phenomenon consisting of a collection of random variables ordered in time. In principle stochastic process is a collection of random variables. The UMTS data traffic is also a non-deterministic time series which can be analyzed by assuming the manifestations of stochastic (random) processes.

### 3.2.4 Time series data stationarity

Stationarity is a stochastic process for time series data whose ensemble statistics are the same for any value of time. A time series is said to be stationary if there is no systematic change in mean (no trend), if there is no systematic change in variance, and if it contains no strictly periodic variations.

The assumption of stationarity is often important for statistical analysis. It can be either strict where all statistical measures on it are stationary despite the change in $t$, or weak (or second-order stationary) where the distribution mean and variance is kept constant regardless of $t$.

Seasonality and trends make time series non-stationary with an effect changing mean or variance over time. Therefore these trend and seasonality needs to be removed to make the time series data stationary (i.e. zero mean and constant variance). Removing the trend and seasonality is done with a single or repetitive differencing of the time series data.

## 3.3 Forecasting models

In statistics forecasting, models can be chosen as linear or non-linear models based on the time series data type that needs to be forecasted. Linear models such as ARMA, ARIMA SARIMA and non-linear models such as Autoregressive Conditional Heteroskedasticity (ARCH), Non-linear Autoregressive (NAR) and Machine learning variants of Artificial Neural Network models exist [3], [19]. In this thesis SARIMA model and Machine learning variants of artificial neural network models (i.e. Auto regressive Neural Network, Multi-Layer Perception and Extreme Learning Machine) are used to model the UMTS data traffic.

### 3.3.1 ARIMA models

ARIMA is a general and combined model that is used to capture the non-stationarity property of a data and provide next stage/s forecast. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values [19]. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The integration or differencing can be applied one or more times to eliminate the non-stationarity and make the time series data stationary. I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as much as possible.

Assuming a polynomial that has a factor *(1 - L)* of multiplicity *d* $(1-\sum_{i=1}^{p'}\alpha_i L^i)$ in the ARMA equation.

Thus it can be written as $(1-\sum_{i=1}^{p'}\alpha_i L^i) = (1-\sum_{i=1}^{p'-d}\phi_i L^i)(1-L^i)^d$

Thus An ARIMA (p,d,q) process expresses this polynomial factorization property with $p=p'-d$, and is given by: $(1-\sum_{i=1}^{p'-d}\phi_i L^i)(1-L^i)^d X_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t$

Generalizing this formula results:- $(1-\sum_{i=1}^{p'-d}\phi_i L^i)(1-L^i)^d X_t = \delta + (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t$

Defines the ARMA (p,d,q) process differentiated with $d$ times and become stationary.

In order to determine the proper model for the given time series, determine the values of p, d and q, it is necessary to carry out *Auto Correlation Function* (ACF) and *Partial Auto Correlation Function* (PACF) analysis. These statistical measures reflect how the observations in a time series are related to each other. For modeling and forecasting purpose it is often useful to plot the ACF and PACF against consecutive time lags to help in determining the order of AR and MA terms.

The AR and MA parameters/orders are estimated from lags of the differenced series of the time series data PACF and ACF plot of peak values a respectively; whereas the difference parameter simply obtained from the number of differencing which makes the data series stationary.

### 3.3.1.1 SARIMA model

For time series data with seasonality components present, it is possible to consider an extension of ARIMA model known as SARIMA This model has non-seasonal and seasonal parts both having the AR, Integrated/ difference and MA parameters.

By definition, SARIMA model with non-seasonal terms of order (p, d, q) and seasonal terms of order (P,D,Q) is abbreviated as SARIMA(p,d,q) × (P,D,Q)s model and may be written as:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-\sum_{j=1}^{q}\phi_j L^{jX\delta})(1-L)^d(1-L^{\delta})^D x_t = (1+\sum_{i=1}^{q}\theta_i L^i)(1+\sum_{j=1}^{Q}\Theta_j L^{jX\delta})a_t$$

$$y_t = (1-L)^d(1-L^{\delta})^D x_t$$

Where:

- $X_t$ is the original non-stationary output at time t.

- $y_t$ is the differenced (stationary) output at time t.
- d is the non-seasonal integration order of the time series.
- p is the order of the non-seasonal AR component.
- P is the order of the seasonal AR component.
- q is the order of the non-seasonal MA component.
- Q is the order of the seasonal MA component.
- s is the seasonal length.
- D is the seasonal integration order of the time series.
- $a_t$ is the innovation, shock or the error term at time t.
  - { $a_t$ }time series observations are independent and identically distributed (i.e. i.i.d) and follow a Gaussian distribution (i.e. $\Phi(0,\sigma^2)$ )

Assuming $y_t$ follows a stationary process with a long-run mean of $\mu$, then taking the expectation from both sides, we can express $\phi_0$ as follows:

$$\phi_0 = (1-\phi_1-\phi_2-...-\phi_p)(1-\Phi_1-\Phi_2-...-\Phi_P)$$

Thus, the SARIMA (p,d,q) (P,D,Q)s process can now be expressed as:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-\sum_{j=1}^{P}\Phi_j L^{j\delta})(y_t-\mu) = (1+\sum_{i=1}^{q}\theta_i L^i)(1+\sum_{j=1}^{Q}\Theta_j L^{j\delta})a_t$$

$$Z_t = y_t - \mu$$

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-\sum_{j=1}^{P}\Phi_j L^{j\delta})Z_t = (1+\sum_{i=1}^{q}\theta_i L^i)(1+\sum_{j=1}^{Q}\Theta_j L^{j\delta})a_t$$

In sum, $z_t$ is the differenced signal after subtracting its long-run average.

The order of the seasonal or non-seasonal AR and MA components are determined by, the ACF and PCF plots peak values behavior (i.e. decay to zero) of the lags. [4].

### 3.3.1.2 Order Selection Criterion

Model selection criterion allow the best model to be fit the data by striking a balance and finding a model that neither under-fits nor over-fits the data. In model selection, the idea is to find the smallest set of variables which provides an adequate description of the data, this is based on the minimum values of AIC, BIC and AICc.

## 3.3.2 Artificial Neural Networks

There are many non-linear machine learning algorithms that can be used in statistical forecasting depending on the data-set and its nature. Artificial Neural Networks (ANN) are a class of models within the general machine learning context, inspired by structure, processing method and learning ability of biological brain. As ANNs are data-driven and self-adaptive in nature, it is not required to provide any a priori assumption about the statistical distribution of the data or specify a particular model form.

The common type of ANNs consists of three layers, of units/nodes. The *input layer* of nodes feeds the input variables into the network and the *hidden layer* is a bridge between the input layer and the *output layer*. The hidden layers learn the relationship between inputs and outputs in a way similar to that of the human brain by adjusting the weights during the training process. The nodes in each layer (excluding the input layer) will compute a weighted sum of the inputs and perform a nonlinear transformation (functional mapping) on the sum using different activation functions. A specific ANN is determined by its topology, learning paradigm and learning algorithm [23].

Outputs-----------------------------→

Output nodes----------------------→

Processing /hidden nodes-------→

Input nodes-------------------------→

Inputs --------------------------------→

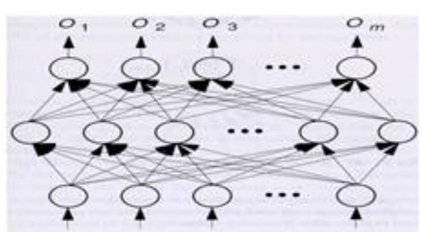

**Figure 5**. Artificial Neural Network topology

For a single layer ANN with single output, the general expression can be [18]

$$y(X,W) = g[\sum_{j=1}^{M} W_j \phi_j (X)]$$

Where *g* is a non-linear activation function such as linear, sigmoid, hyperbolic tangent, Gaussian, etc; $\phi_j$ (x) is fixed and non-adaptive functions referring to the inputs to the nodes and *w* is weight of the functions.

There are different variety of learning algorithms in NN that are based on gradient-descent or iterative approaches like regression, classification etc....which widely used in various applications [25].

### 3.3.2.1 Auto Regressive Neural Network

Auto-Regressive Neural Network ARNN / NNAR is neural network type model to forecast non-linear data series types where lagged values of the time series are used as inputs. The *p* lagged inputs and *k* nodes in the single hidden layer of the network are represented as NNAR (p, k). Similar to linear ARIMA models, it is possible to have non-seasonal NNAR (p, 0) model that is equivalent to an ARIMA (p, 0, 0) and Seasonal NNAR (p, P, k) model with inputs ($y_{t-1}$, $y_{t-2}$, . . . , $y_{t-p}$, $y_{t-m}$, $y_{t-2m}$, $y_{t-Pm}$) that is equivalent to an SARIMA (p, 0, 0)(P,0,0)$_m$ model, without stationarity restrictions in both cases [23], *where P is seasonal AR value of NNAR*.

### 3.3.2.2 Multi-Layer Perceptron

Multi-Layer Perceptron Network (MLP) is a neural network machine learning variant of multi layers of learning algorithm. In MLP the first layer involves a linear combinations of different dimensional inputs [28].

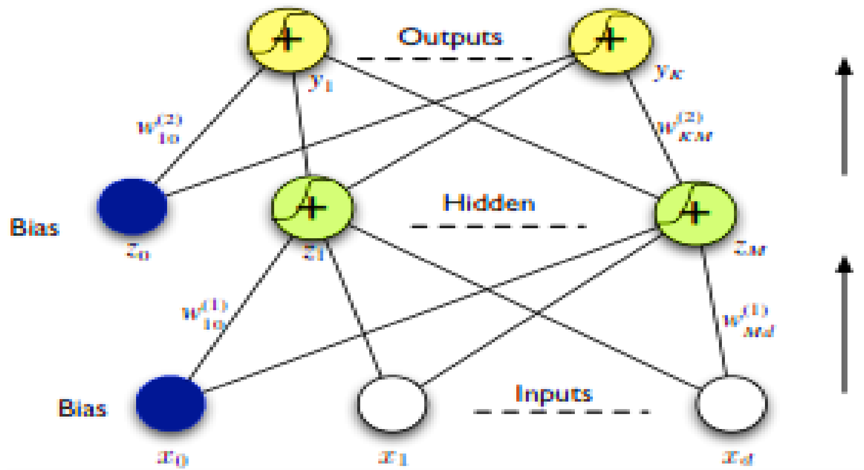


**Figure 6**. Network diagram MLP with two layers of weights.

MLP has linear activation function in all neurons that maps weighted inputs to the output of each neuron. In MLP some neurons use a non-linear activation function developed to model the frequency of action potentials (nerve impulse), or firing of biological neurons. The two common activation functions are hyperbolic & sigmoid (S-shaped mathematical curve) of equation, $y(v_i) = \tanh(v_i)$ and $y(v_i) = (1 + e^{-v_i})^{-1}$.

The MLP consists three or more layers (an input and output layer with one or more hidden layers) of non-linearly activating nodes. Each node connects in a mesh with weight $w_{kj}$ to the next layer.

The learning occurs in the perception by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. The error in the output node $j$ in the n$^{th}$ data point (training example) represented by $e_j(n) = z_j(n) - y_j(n)$, where $d$ is the target value and $y$ is the value produced by the perception [28].

As illustrated in figure 6 above, the outputs $z_j$ correspond to the outputs of the activation function. In the context of neural networks, the quantities $z_j$ are interpreted as the output of hidden units so called, because they do not have values specified by the problem (as is the case for input units) or target values used in training (as is the case for output units).

In the second layer, the outputs of the hidden units are linearly combined to give the activations of the *K* output units:

$$a_k = \sum_{j=0}^{M} W_{kj}^{(2)} Z_{kj} \quad \text{, Where k = 1,2,....K}$$

When $z_0$=1, corresponding to the bias. The transformation is the second layer of the neural network parameterized by weights $W^{(2)}{}_{kj}$. The output units are transformed using an activation function [28]:

$$y_k = g(a_k) = \frac{1}{1+\exp(-a_k)},$$

These equations may be combined to give the overall equation that describes the forward propagation through the network, and describes how an output vector is computed from an input vector, given the weight matrices:

$$y_k = g\left[\sum_{j=0}^{M} W_{kj}^{(2)} h\left[\sum_{j=0}^{d} W_{jk}^{(1)} x_j\right]\right]$$

### 3.3.2.3 Extreme Learning Machine

Extreme learning machine (ELM) is an ANN with a unified learning platform developed to improve the efficiency for single layered Feedforward Networks (SLFNs). The simple ANN discussed in section 3.3.2, which can also be refereed as SLFN, the learning requires manual tuning of control parameters (such learning rate, learning epochs, etc.) whereas in ELM the learning without iteratively tuning hidden neurons.

In general, given any bounded non-constant continuous activation function *g* (integrable for all nodes) for any continuous target function *f* and any randomly generated sequence $\{(a_L, b_L)\}$, $\lim_{x\to\infty} \| f(x) - f_L(x) \| = 0$ holds with probability one, if $\beta_i$ is chosen to minimize the $\| f(x) - f_L(x) \|$, *i= 1,2…L* [28].

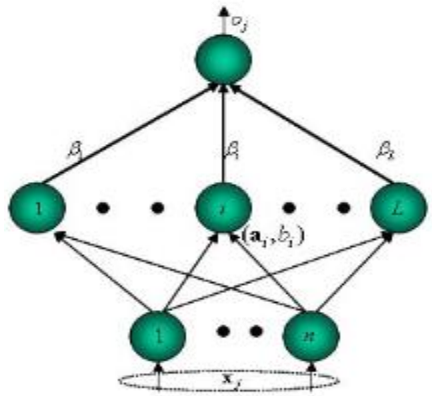


**Figure 7**. Feedforward network architecture to implement ELM algorithm

In ELM, given a training set $N = \{(X_i, t_i) | X_i \in R^n, t_i \in R^m, i = 1, ..., N|\}$, activation function $g$, and the number of hidden nodes $L$, the learning is performed in three steps [29]:

1. Assign randomly input weight vectors or centers $a_i$ and hidden node bias or impact factor $b_i$, $i = 1, \cdots, L$.
2. Calculate the hidden layer output matrix $H$.
3. Calculate the output weight $\beta : \beta = H^{\dagger}T$ where $T = H\beta$

Where $H^{\dagger}$ the moore penrose is generalized inverse of hidden layer output matrix $H$.
ELM has the following main features with respect to real world regression problems:

- It is a simple tuning-free algorithm.
- The learning speed of ELM is extremely fast

## 3.4 Hybrid Model

Hybrid models are a combination of various linear and non-linear models to handle different behaviors of a time series data thus to increase the forecast performance. Hybridizing models can be done through simple or weighted addition, multiplication, and ensemble averaging.

For a time series data $y_t$, a hybrid model comprising a linear and a non-linear component has been employed in experiments [13]:

$$y_t = L_t + N_t$$

Where $L_t$ is the linear AR component and $N_t$ is the non-linear component.

In this case, the linear part is modeled first by fitting an AR function in to the time series. Then, the residuals from the AR model are modeled using neural networks.

Let $r_t$ be the residual of the linear component, then: $r_t = y_t - \hat{L}_t$

where $\hat{L}_t$ is the estimate of the linear AR component.

For non-linear components modeled with neural networks; the model output, $\hat{r}_t$, is expressed as: $\hat{r}_t = f(r_{t-1}, r_{t-2}, \ldots, r_{t-q})$

where $q$ is the number of input delays and $f$ is the non-linear function.

Therefore, the forecast done by integrating the two results will be:

$$\hat{y}_t = \hat{L}_t + \hat{r}_t + \varepsilon_t$$

where $\varepsilon_t$ is the error of the combined model.

Since the non-linearity observed in the data would not be captured by the linear models, we assume that non-linearity will be reflected on the residuals of the linear model. Therefore, by considering the non-linear model to forecast the residuals and integrating the results from the two models, thus exploiting the strength of both components improve the accuracy of the forecast [30].

# 4. Hybrid SARIMA-ELM based forecasting

## 4.1 System model

As described in chapter 3 time series data modeling generally consists three stages: model identification, parameters estimation and diagnostic checking, in this thesis also these principles are followed to model the linear part of the UMTS data traffic.

In forecasting principle, the forecasting model builted is expected to capture all characteristic and components of the predictor data set. The system model designed in the thesis divides the data set in to linear and nonlinear part where the non-linear part is to be extracted as residual (error) from the linear part model fitting of the UMTS data traffic. The system model to perform the modeling and forecast is illustrated in Figure 8.

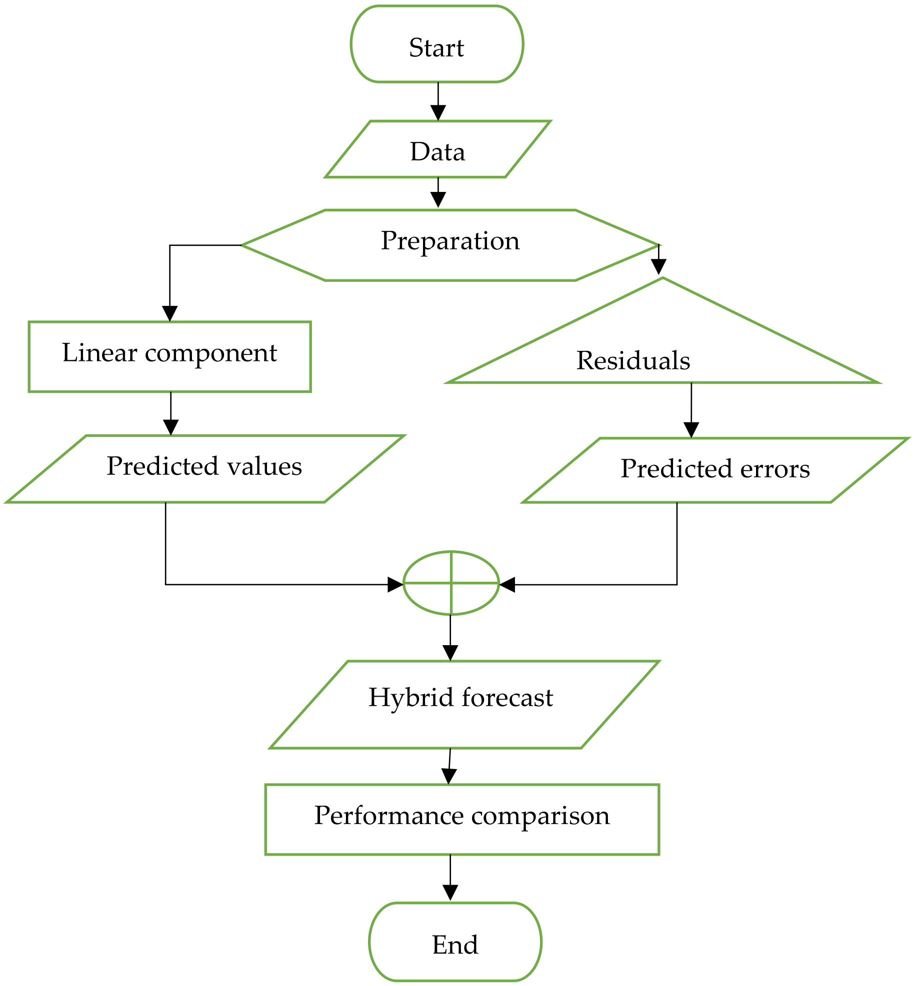


**Figure 8**. The proposed system model flaw chart

## 4.2 UMTS Traffic Data Set

Data traffic collected from five RNCs of A.A UMTS network for the duration April 2015 – June 2016 are analyzed. In order to perform the data analysis, model building and forecasting, an open source and free statistical software called R software with its built-in and modified packages had been used.

***Data exploration and visualization***

The analyzed data set shows, the data has comprised linear and non-linearity in it, for the data collection period of April 2015 and June 2016.

During visual data analysis, we were able to find a data missing for the month May 2015 and April 2016 (circled with blue), as presented in Figure 9. Thus, Kalman filter algorithm of linear Gaussian model estimation is used for the smoothing of the series as well as estimate the missed values thus to keep consistency and completeness of the data for analysis. These missed values estimation will then go as residual when the linear model fits and will be handled by the machine learning modeling part.

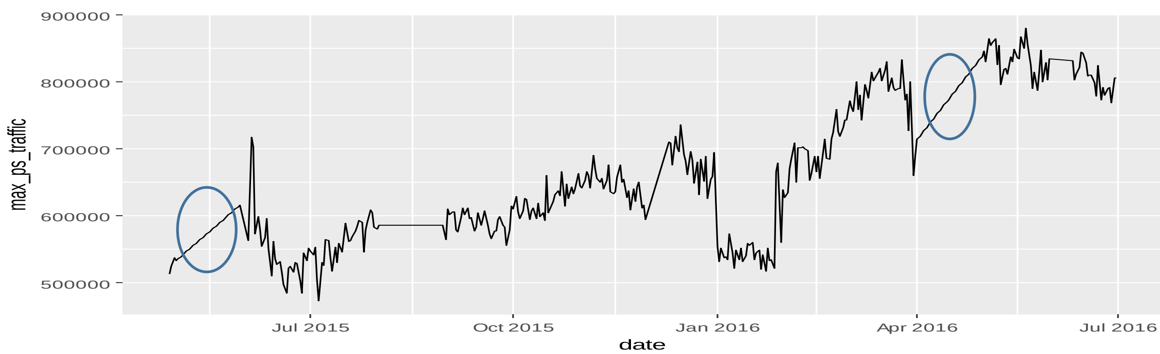


**Figure 9**. Plot of UMTS data traffic collected in daily basis

*Components identification*

Figure 10 illustrated the decomposition of the series which exhibits non-stationarity as a form of increasing trend and weekly seasonality.

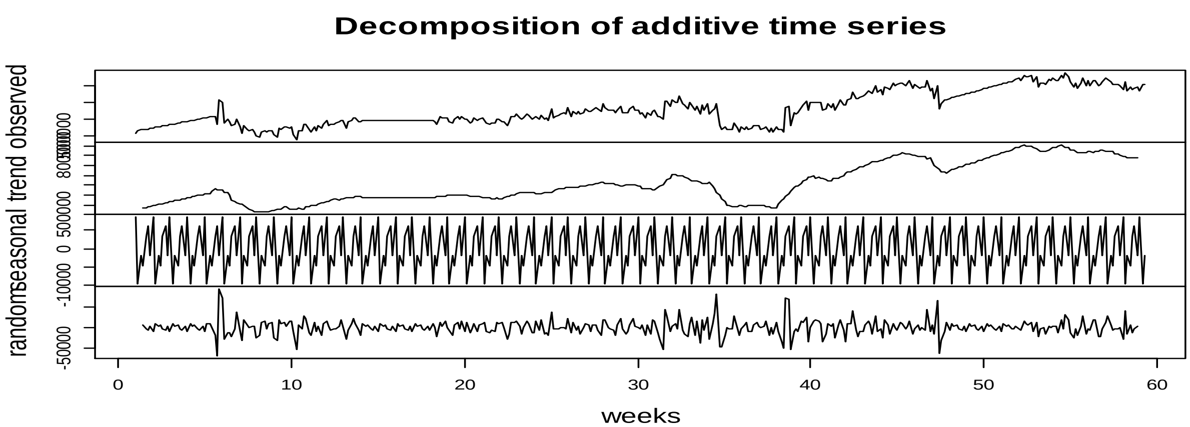


**Figure 10**. Decomposition of the UMTS data traffic in to time series components

*Stationarity*

The linear model, ARIMA, requires the stationarity of the data. Therefore, prior to mode l building seasonality and increasing trend, made the data non-stationary, has to be rem oved with a process of differencing. In order to test stationarity of the differenced series presented in Figure 11, null hypothesis test such as augmented Dickey–Fuller test (ADF ) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) were used confirms stationarity.

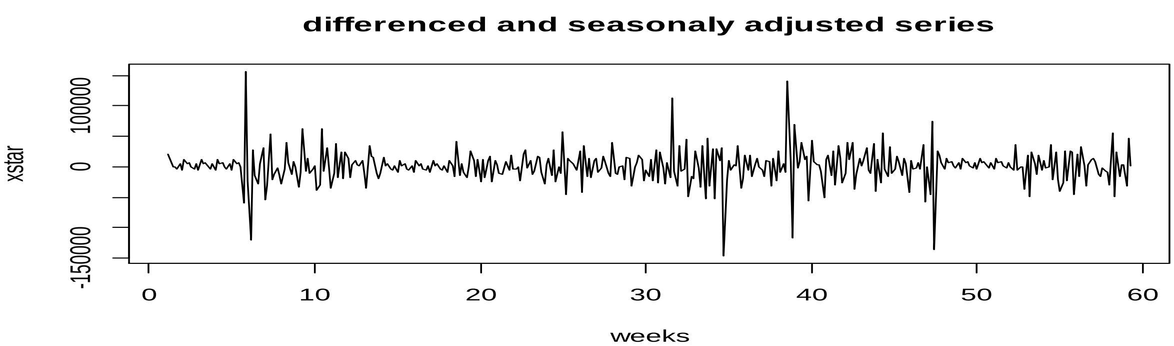


**Figure 11**. Seasonally adjusted and differenced UMTS data traffic

## 4.3 Model identification and parameter estimation

### 4.3.1 Linear modeling of the UMTS peak data traffic

The seasonal components on the UMTS traffic data implicated that the use of Seasonal ARIMA (p,d,q) (P,D,Q)[s] model for the linear modeling. The data series became stationary by removing the trend and seasonality with differencing and seasonal adjusting respectively in one step. Next step is to estimate the model parameters or orders (p, d, q, P, D, Q) values to put candidate SARIMA models of different orders and select the best model to fit the data series.

*Parameters estimation*

Determining or estimating the model parameters can be performed with a simple ACF and PACF analysis. From ACF and PACF plots presented in Figure 12 and table1 of annex , ACF plot shows positive spikes at lag 7, 14, 21… indicating a seasonality of 7 (weekly) and p values decayed to zero from large negative at lag 1, and no other peaks at positive, which indicates the non- seasonal MA and seasonal MA values to be 1. The PACF value has negative peaks decaying from at lags of 1 to 3, which indicates the non-seasonal AR value to be in range from 0 to 3 and a peak at positive of the PACF value at lag 7, indicates the seasonal AR value to be 7.

The data series become stationary with seasonal adjusting of one-time differencing which indicates the non-seasonal differencing value *d* to be 1 and seasonal differencing value *D* to be 0 or 1. Therefore, SARIMA$(0,1,1)(1,0,1)_7$ to $(3,1,1)(1,0,1)_7$ and SARIMA$(0,1,1)(1,1,1)_7$ to $(3,1,1)(1,0,1)_7$ are identified as candidate models for fitting.

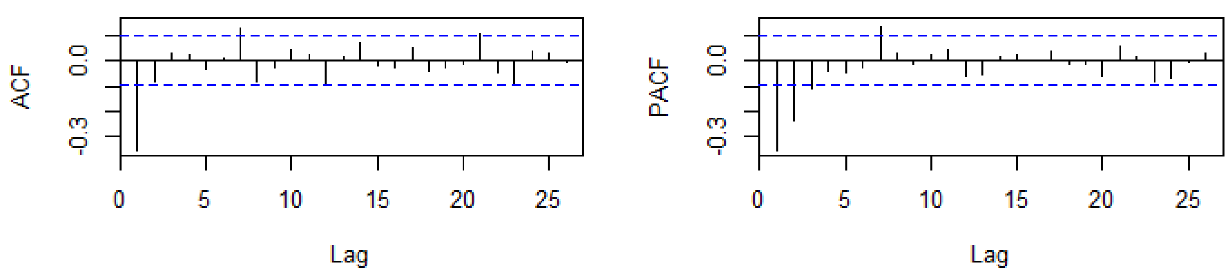


**Figure 12**. The ACF and PACF plot of the UMTS data traffic

### 4.3.1.1 The linear model Validation

***Model selection and fit***

The SARIMA candidate models identified above are compared based on minimum values of AIC, BIC and AICc as presented in Table 1. As indicated on the table ARIMA $(0, 1, 1)(1, 0, 1)_7$ exhibits a minimum values of AIC, BIC and AICc, thus it is selected as the best fit linear model for the UMTS traffic data set under investigation.

From the fitting plot of the selected ARIMA $(0, 1, 1)(1, 0, 1)_7$ model shown in Figure 13, the spikes out of the curve fit indicates the existence of non-linearity on the UMTS peak data traffic. These non-linearity to be extracted as a residual or errors and modeled with

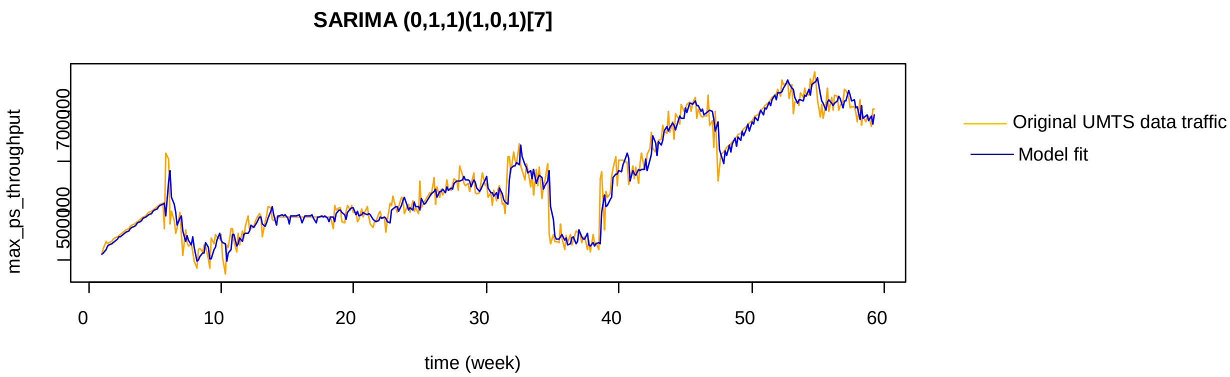


**Figure 13**. SARIMA (0, 1, 1) (1, 0, 1)$_7$ model fit to the UMTS data

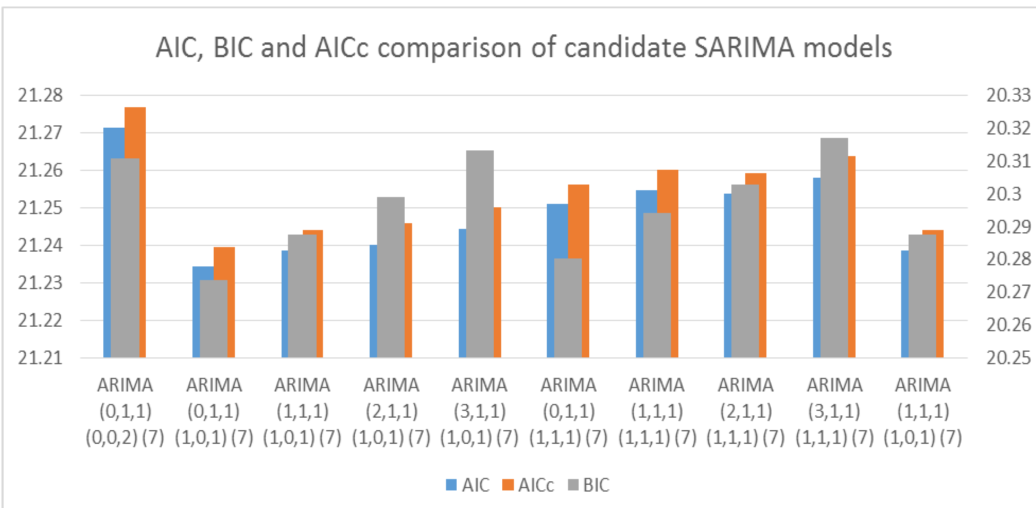


**Figure 14** Candidate SARIMA models comparison for selection

### *Diagnosis of model fit from its residuals*

From the figure 14 below, the ACF plots of the residuals lag values lie within p value of 0.1 and -0.1 also the residuals Q-Q plot shows residuals are normally distributed, which is informative that the selected model fitted well.

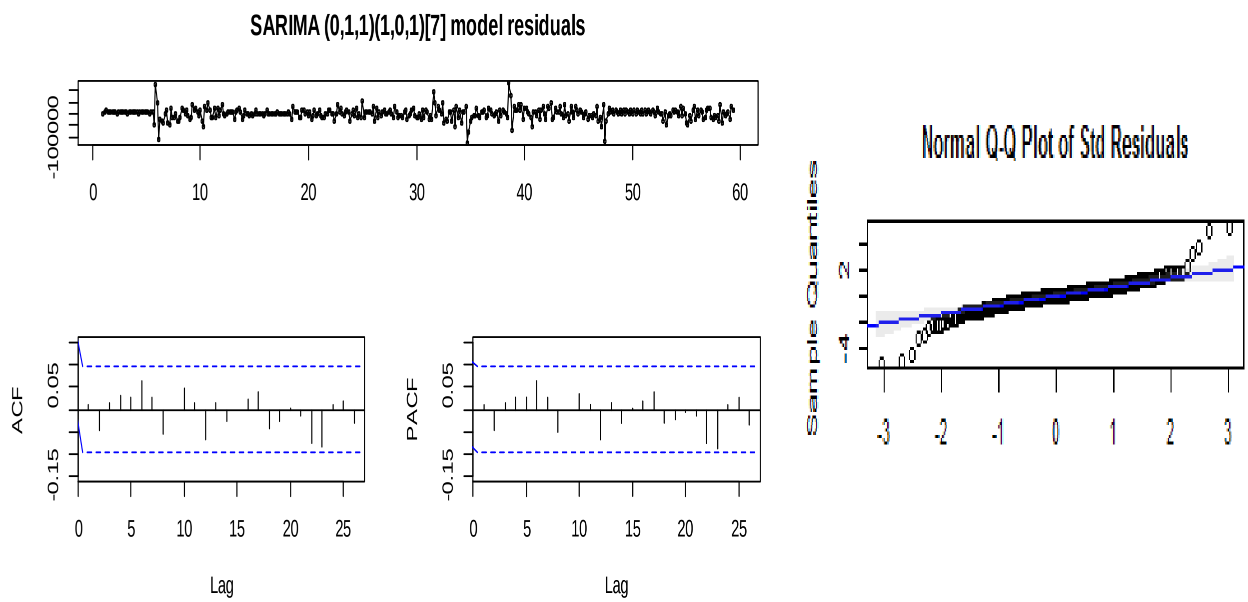


**Figure 15**. ACF and Q-Q plots of SARIMA (0,1,1)(1,0,1)$_7$ model fit residuals

### 4.3.1.2 Linear model forecast results

The UMTS data traffic was forecasted for one month ahead with SARIMA (0,1,1)(1,0,1)$_7$ model as shown in figure 17.The forecast result plot in figure 16, shows the UMTS peak data traffic throughput were 709Mbps at the end of June 2016; however, the forecast on the end of July 2016 reached 806Mbps, which implies the data traffic throughput is increasing on average 3.2Mps per day.

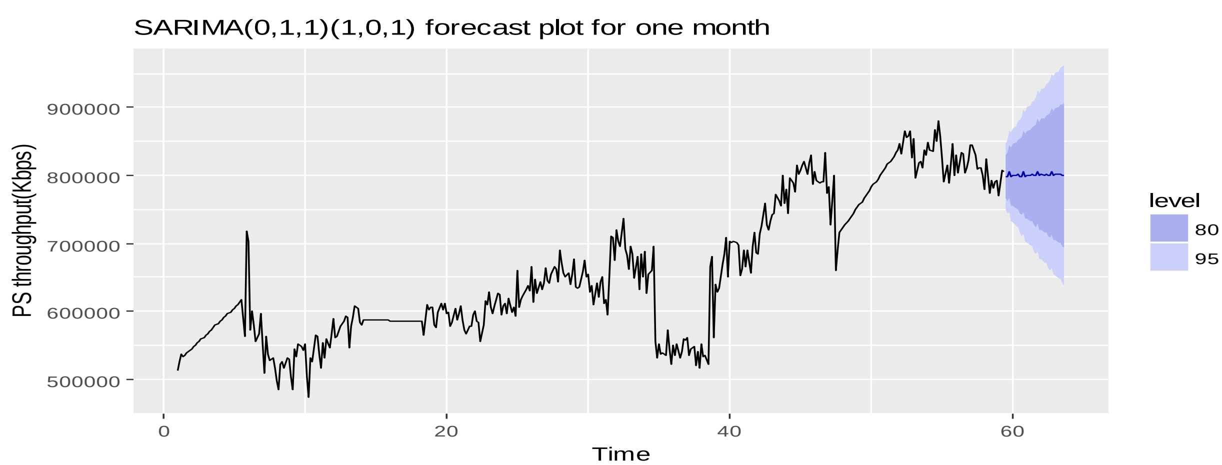


**Figure 16**. SARIMA (0,1,1) (1,0,1)$_7$ forecast plot

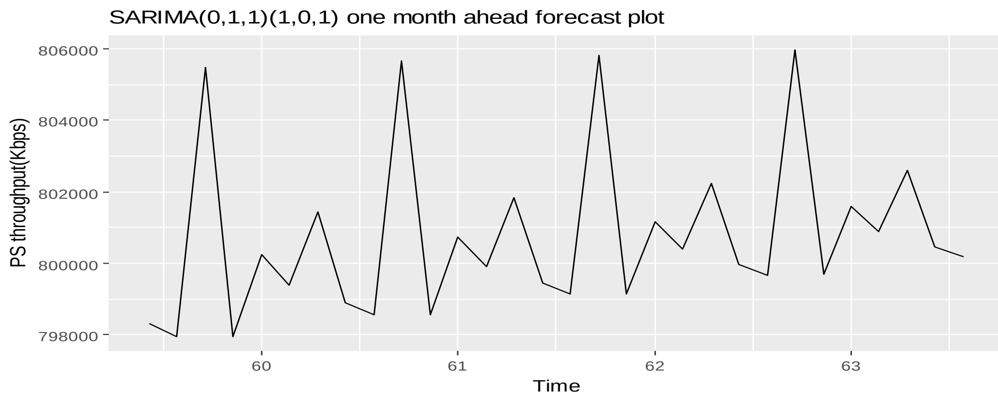


**Figure 17**. SARIMA (0,1,1)(1,0,1) $_7$ one month forecast result plot

## 4.3.2 Non-linear Modeling of the Residuals

The residuals, which can be taken as an errors, from the SARIMA (0,1,1)(1,0,1) $_7$ model fitting can be treated as non-linear components of the UMTS traffic data set and will be modeled with non-linear ANN variant models. The three ANN variants considered for non-linear modeling are ARNN, MLP and ELM.

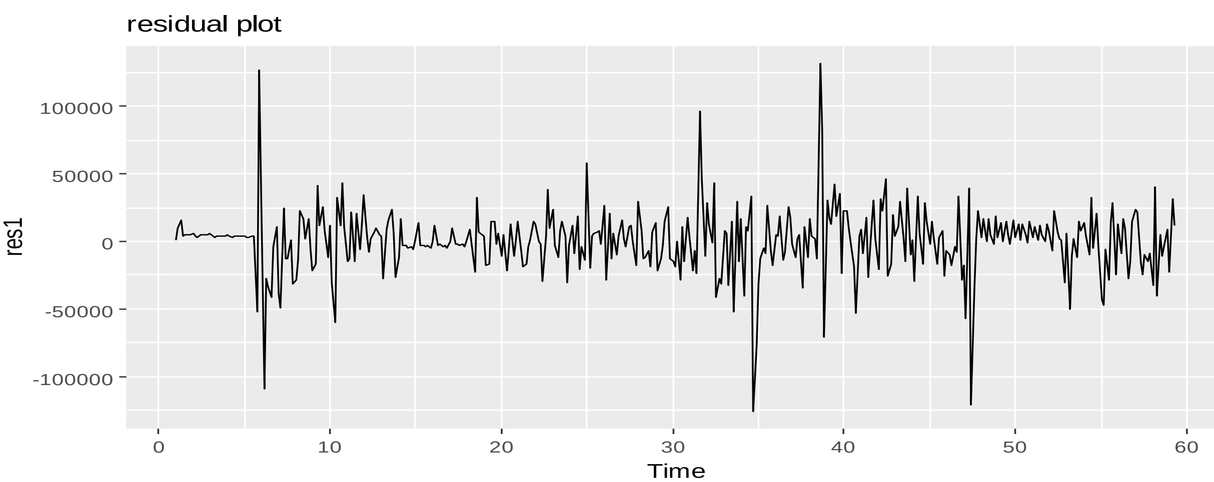


**Figure 18**. SARIMA (0, 1, 1) (1, 0,1)$_7$ residuals plot

## 4.3.2.1 Residuals forecast results

*ARNN model forecast result*

The ARNN based month ahead forecast of the residuals is illustrated with blue line in Figure 18. The code executed for the ARNN, automatically produced an ARNN (22, 1, 12)$_7$ model, with 22 non-seasonal lagged input values, 1 seasonal lagged input value and 12 hidden layer neuron nodes of seasonality 7 (a week), thus ARNN (22, 1, 12)$_7$ is equivalent to an ARIMA (22,0,0)(1,0,0)$_7$.

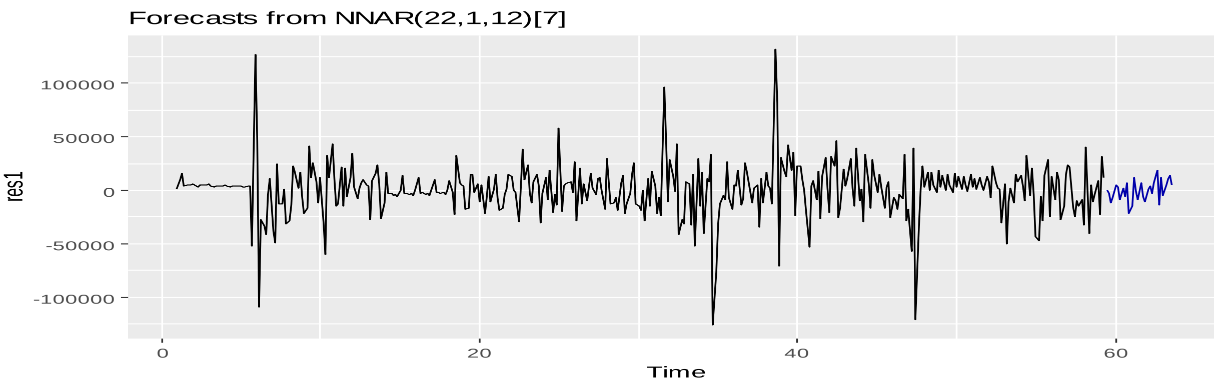


**Figure 19**. Month ahead residuals forecast plot of ARNN

*MLP model forecast result*

The MLP algorithm, neural network diagram with 5 hidden neural nodes as shown with figure 19, with one input of the UMTS data traffic SARIMA model fit residuals passed on 5 layers as perception outputs the forecast results as figure 21.

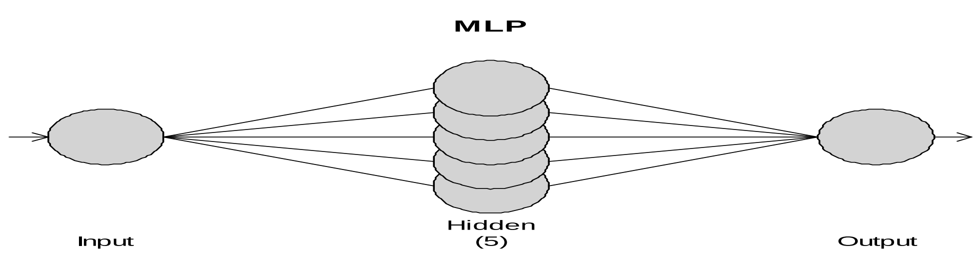


**Figure 20**. The MLP diagram for residuals modeling

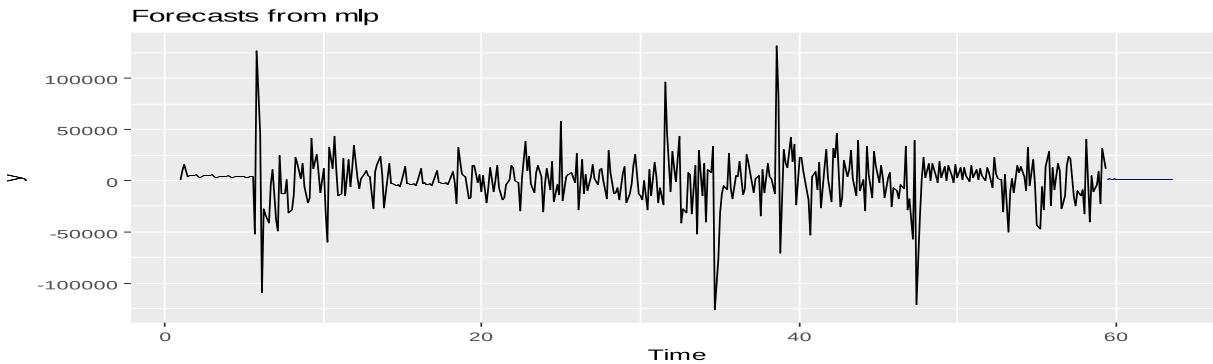


**Figure 21**. Month ahead residuals forecast plot of MLP

***ELM model forecast result***

ELM algorithm based neural network with 100 hidden layer, extreme learning layers, as presented in figure 22 below, with an input of the UMTS data traffic SARIMA model fit residuals for a month ahead forecast results as presented in figure 23.

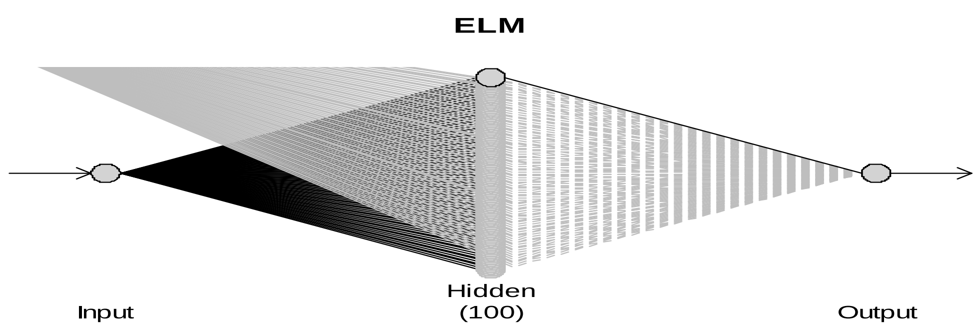


**Figure 22**. The ELM diagram for residuals modeling

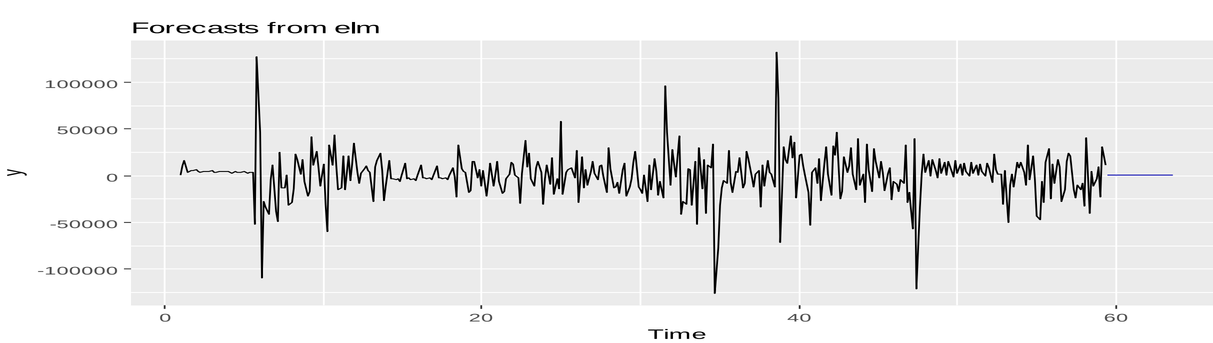


**Figure 23**. Month ahead residuals forecast plot of ELM

## 4.3.3 Hybrid model forecast results

***Hybrid model forecast and results***

Hybridizing the linear and non-linear models is done through a simple addition. The selected ARIMA (0,1,1)(1,01)$_7$ model forecast is added with its residuals forecast, modeled either with ARNN, MLP and ELM as presented in the next sections. The best performing hybrid model with a minimum forecasting error metrics (i.e. RMSE, MAE, MAPE, and MASE) is then selected.

***SARIMA-ARNN forecast result***

As shown in Figure 24, the forecast for one month ahead as of June 2016 using hybrid SARIMA-ARNN model, shows the A.A UMTS network throughput to reach 830Mbps from 790Mbps; however, as it can be seen from the plot, the forecasted UMTS data traffic series has lost the components of seasonality and trend that were observed in the original UMTS data traffic series hence the forecast did not include the predictor data set components inside, SARIMA-ARNN model not selected a hybrid model.

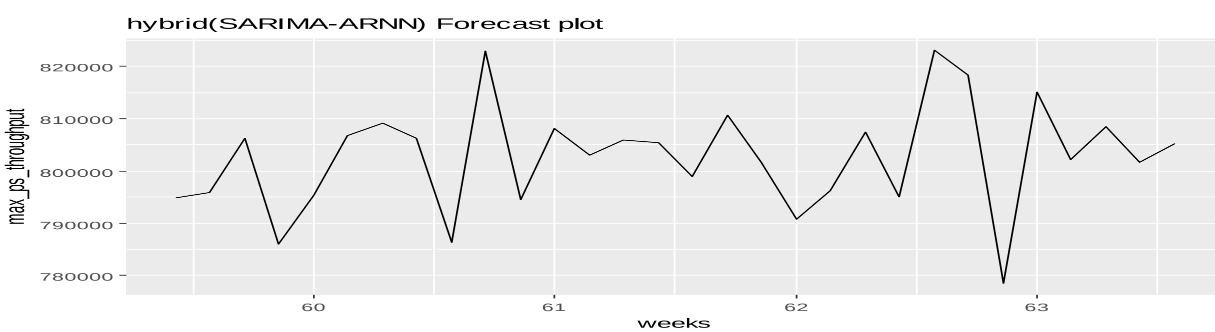


**Figure 24**. SARIMA-ARNN forecast plot for one month

***SARIMA-MLP forecast result***

As shown in Figure 25, the forecast for one month ahead as of June 2016 using hybrid SARIMA-MLP model, shows the A.A UMTS network throughput to reach 807.5Mbps from 790Mbps. As it can be seen from the plot, the forecasted UMTS peak data traffic series also comprised the components (seasonality and trend) that were observed in the original UMTS data traffic series.

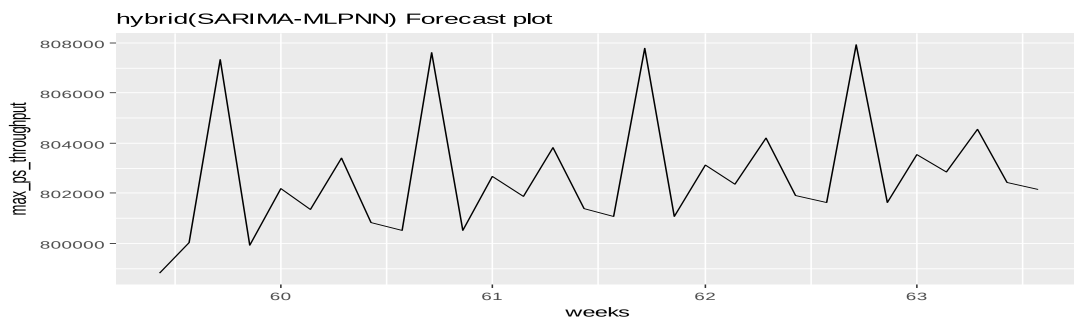


**Figure 25**. SARIMA-MLP forecast plot

***SARIMA-ELM forecast result***

As shown in Figure 25, the forecast for one month ahead as of June 2016 using hybrid SARIMA-ELM model, shows the A.A UMTS network throughput reached 808.5Mbps from 790Mbps that were at the end of June. As it can be seen from the plot, the forecasted UMTS peak data traffic series comprised the components (seasonality and trend) that were observed in the original UMTS data traffic series.

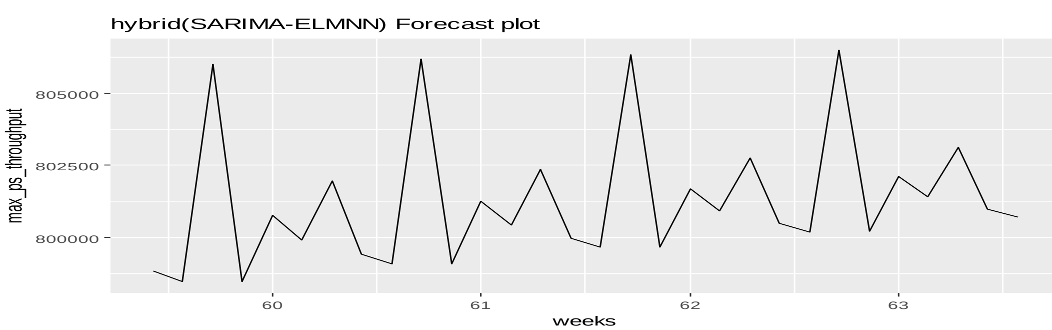


**Figure 26**. SARIMA-ELM forecast plot of UMTS data traffic for one month

***Forecast model selection based on forecast Performance metrics comparison***

Figure 27 and annex table 2, presents the forecast error performance of hybrid models of SARIMA with ANN variants based on RMSE, MAE, MAPE, and MASE error measuring metrics. As it can be observed the SARIMA (0, 1, 1) $(1, 0, 1)_7$ – ELM hybrid model has minimum values of the error metrics. The model is selected as the hybrid forecasting model with an average of 12.4 % percentage error improvement, compared to the second best hybrid model, which is SARIMA (0, 1, 1) (1, 0, 1)$_7$ – MPL.

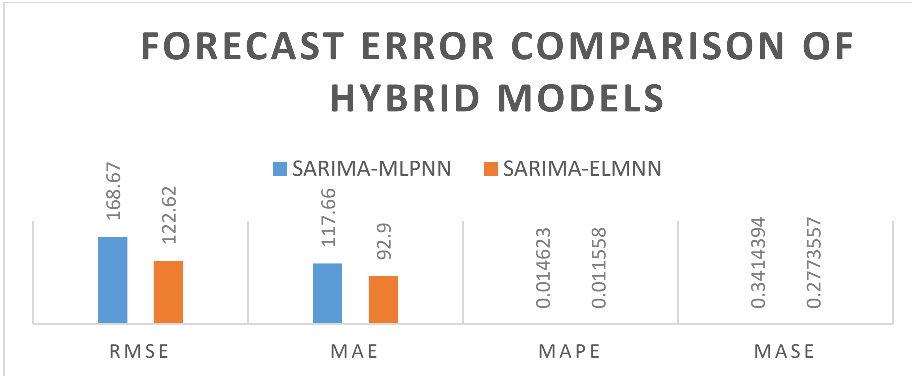


**Figure 27** Candidate SARIMA models comparison for selection

***Hybrid SARIMA-ELM Forecast result***

The 30 days (one month ahead as of June 2016) UMTS data traffic forecast has been conducted with the hybrid SARIMA(0,1,1)(1,0,1)$_7$ - ELM model for Addis Ababa network as presented in Figure 28. The plot also shows the forecast fails within 80 and 95 % confidence intervals of forecast. The hybrid SARIMA-ELM forecast indicated the UMTS data traffic throughput reached 808.5Mbps at the end of July 2016 as it can be observed in figure 26, which increasing on average of 3.32Mbps per day, which is different from the forecast obtained from the forecast done by SARIMA(0,1,1)(1,0,1)$_7$ model alone by 3.75%.

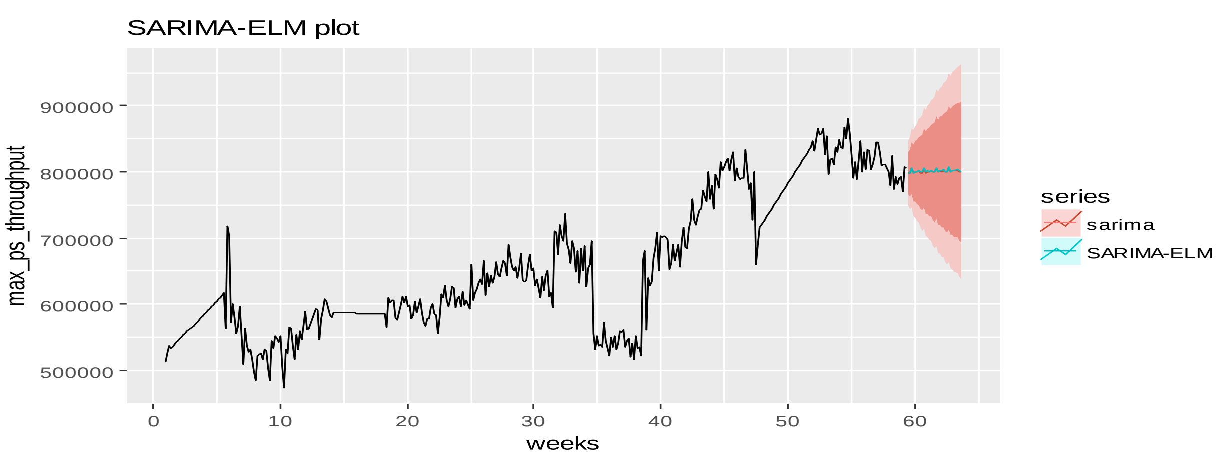


**Figure 28** Hybrid ARIMA (0, 1, 1) (1, 0, 1) $_7$ - ELM plot

# 5. Conclusion, Recommendation and Future work

## 5.1 Conclusion and Recommendation

The main aim of this thesis is to model and forecast UMTS data traffic demand based on historical data collected from the live network, with linear Univariate time series method, namely SARIMA, then extract the errors as residuals of the SARIMA model fit and model it with a non-linear ELMNN model. Forecast the two modeled data and hybridize their forecasts and produce a hybrid model so as to increase the accuracy of the forecast.

For the linear process differencing to remove both the seasonal and trend components has been performed. The SARIMA model candidates identified to fit UMTS data-traffic has been selected with parameter estimations. The best fit SARIMA model has been selected based on the AIC, BIC and AICc criterion from the candidates. The errors from this SARIMA model fit has been extracted as residuals and modeled with ARNN, MLPNN and ELMNN non-linear models. The forecast from the selected SARIMA then hybridized with these NN variants as SARIMA-ARNN, SARIMA-MLPNN & SARIMA-ELMNN, thus based on lowest error metric values, hybrid SARIMA-ELMNN model selected.

With the results of the processing, modeling and forecast performed, the following conclusion is drawn

- The A.A UMTS data traffic comprises non-linearity, which may be due to data service filtering and blockage were happened with in the duration of the data collection period.
- The SARIMA model (ARIMA with seasonality) selected because of seasonality observed in the UMTS data traffic and performed well for every forecasting period, according to the pattern. The plotted UMTS forecasted data-traffic seemed to exhibit a seasonal pattern of every week but ACF indicated a possible seasonal

component on the 7th day of the week. When fitting the SARIMA (0, 1, 1) (1, 0, 1)$_7$ model with this seasonality, it fit the UMTS data-traffic with low error terms, and a lower value of AIC, BIC and AICc.

- Seasonal and trend component exists in data-traffic flow, which can be removed by one-step difference of the original data. The differenced UMTS data-traffic flow data are one-step correlated. In other word, the increase or decrease of data-traffic flow can influence the change of data in the next time step. The prediction of UMTS data-traffic can be made simpler by studying the differenced original data-traffic.
- The ADF and KPSS diagnostic tests of the differenced time series confirms the differenced series as a stationary.
- The residuals of the SARIMA (0,1,1)(1,0,1)$_7$model fit ACF plot shows that there is no peak values at each lag that extends the 0.1 and -0.1 p value boundary for all lags, which indicates the model fits well the original UMTS data traffic series.
- The Q-Q plot shows residuals are normally distributed, also indicates the model fits well with the original UMTS data traffic series.
- With only SARIMA (0,1,1)(1,0,1)$_7$model forecast resulted the UMTS peak data traffic throughput reached 806Mbps as of July 2016 from 709Mbps that were on the end of June2016.
- Model and forecast the residuals from the linear model (SARIMA) fit (as an error with non-linear NN model variants) and hybridize with the SARIMA (0,1,1)(1,0,1) $_7$forecast results: -
    - SARIMA-ARNN forecast couldn't include the components that were in the original UMTS data traffic series, implies ARNN neglect seasonality and trend components in hybridizing with the SARIMA.
    - SARIMA-MLP forecast includes seasonality and trend components that were in the original UMTS data traffic series, and the traffic to reach 807.5Mbps as of July 2016 from 709Mbps that were on the end of June2016.

  - SARIMA-ELM forecast includes seasonality and trend components that were in the original UMTS data traffic series, and the traffic reached 808.5Mbps as of July 2016 from 709Mbps that were on the end of June2016.
- The hybrid SARIMA-ELM forecast is different from the forecast obtained from SARIMA $(0,1,1)(1,0,1)_7$ model alone by 3.75%.
- According to the RMSE, MAE, MAPE and MASE error metrics comparison of the hybrid SARIMA-ARNN, SARIMA-MLP and SARIMA-ELMNN models forecast; SARIMA-ELMNN model forecast scores a minimum error value of these metrics and an average of 24.8 % percentage error improvement than the SARIMA $(0,1,1)x(1,0,1)_7$ –MLP model, which has the next minimum value of the error metrics.

SARIMA-ELM model selected as a hybrid model that handles the linear and non-linear behavior exists in the UMTS data traffic of ethio telecom's Addis Ababa network. This SARIMA-ELM forecasting model thus to be used by ethio telecom to forecast the UMTS data traffic demand in such a way to use as a planning inputs.

## 5.2 Future work

As all international telecom operators, the main business strategy of ethio telecom is to have its system operates more efficiently, securely and economically. Thus to meet this strategy, the behavior of the UMTS mobile data-traffic load must be well understood.

The UMTS mobile data traffic demand forecasting plays a role in assisting network planning, marketing analysis and management decisions on investments beforehand.

This thesis work on a univariate (single variable) UMTS data traffic aiming in hybridization of fitted SARIMA model forecast, with modeling and forecast of its residuals (errors) in three machine learning neural networks to enhance the forecast performance.

Future works further can conduct the forecast the UMTS data traffic and in depth: -

- Including some more variants of neural networks
- Considering the UMTS data traffic as multivariate (multiple variable) data series
- Including different factors as features like customers’ behavior modeling and other features, so as to make the forecast more sophisticated.

# 6. References


[1]. Ericsson, "Sub Saharan Africa Ericsson Mobility Report" , November 2016

[2]. Roland Beutler, Swr and Darko Ratkaj, Ebu "Crystal Ball, Tea Leaves OR Mathematics - FORECASTING DATA TRAFFIC FOR MOBILE SERVICES", Q1 2014, technical Review

[3]. Zhang, G.-P. (2003). "Time series forecasting using a hybrid ARIMA and neural network model ". Neurocomputing, pages, 159–175

[4]. Ratnadip Adhikari, R. K. Agrawal "An Introductory Study on Time Series Modeling and Forecasting", pages, 18 – 29.

[5]. Yanhua Yu, Meina Song, Yu Fu, and Junde Song, "Traffic Prediction in 3G Mobile Networks Based on Multifractal Exploration"

[6]. Marketing input data for TEP, from ethio telecom marketing department internal document.

[7]. The High Level Design for RF LOT1 Addis Ababa Swap and Build, ethio telecom internal document.

[8]. Andreas Mitschele-Thiel, Jens Mückenheim, "UMTS networks", October 19, 2010

[9]. Ivan Stojmenovic, "Handbook of Wireless Networks and Mobile Computing", Copyright © 2002 John Wiley & Sons, Inc.

[10]. Ajay R.Mishra, "cellular technologies for emerging markets 2G, 3G and beyond", Nokia siemens networks.

[11]. Samuel Medhn, Bethelhem Seifu, Amel Salem, Dereje Hailemariam "Mobile Data Traffic Forecasting in UMTS Networks Based on SARIMA Model: The Case of Addis Ababa, Ethiopia", IEEE Africon 2017 Proceedings.

[12]. Oladeji E.O, Onwuka E.N and Aibinu A.M, "Determination of Voice Traffic Busy Hour and Traffic Forecasting in Global System of Mobile Communication (GSM) in Nigeria"

[13]. Mergani Khairalla, Xu Ning, Nashat AL-Jallad "Modelling and optimization of effective hybridization model for TS data forecasting", School of Computer and Science and Technology, Wuhan University of Technology, Wuhan 430070,People's Republic of China and School of Science and Technology, Nile Valley University, Atbara, Sudan

[14]. Xin Dong, Wentao Fan, Jun Gu, "Predicting LTE Throughput Using Throughput Using Traffic Time SeriesTraffic Time Series", Beijing University of Posts and Telecommunications, Beijing, 100876, China

[15]. Huimin Pan, Jingchu Liu, Sheng Zhou, and Zhisheng Niu, "A Block Regression Model for Short-Term Mobile Traffic Forecasting", Tsinghua National Laboratory for Information Science and Technology Department of Electronic Engineering, Tsinghua University Beijing 100084, China

[16]. Capacity and coverage Analysis for Addis Ababa and North and East telecom circle, report by ethio telecom Engineering department, optimization section.

[17]. Liang Guo, Jie Zhang and Carsten Maple, "Coverage and Capacity Calculations for 3G Mobile Network Planning", Department of Computing and Information Systems University of Luton, Luton, LU1 3JU, U.K.

[18]. Miguel, Antonio & Artur, "3G Radio Network Planning for A Mobile Network Operator", Master thesis, December 2014

[19]. Peter J. Brockwell, Richard A. Davis, "Introduction to Time Series and Forecasting", springer texts in statistics.

[20]. Mike West, Jeff Harrison, "Bayesian Forecasting and Dynamic Models", springer texts, second edition

[21]. Anasse Bari, PhD, Mohamed Chaouchi, and Tommy Jung, "Predictive analytics for dummies", a welly brand.

[22]. Amit R. Wasukar, "Artificial Neural Network – An Important Asset for Future Computing", International Journal For Research In Emerging Science And Technology, Vol. 1, Issue-1, Jun, 2014.

[23]. A.P. Cheema and Mohd. Sabir, "The Learning Methods and Applications of Artificial Neural Network", Evolving Ideas Computing, Communication and Networking, Global Vision Publishing House, 2012.

[24]. Tugba Taskaya-Temizel, Matthew C. Casey, "A Comparative Study of Autoregressive Neural Network Hybrids", School of Electronics and Physical Sciences, Department of Computing.

[25]. Fakhri Karray, "Fundamentals of Artificial Neural Networks", University of waterloo.

[26]. Kevin Gurney, "An introduction to Neural Networks", University of Sheffield, 1997

[27]. Jan Larsen, "Introduction to Neural Networks", First edition, November 1999, University of Sheffield.

[28]. Hiroshi Shimodaira, "Multi-layer Neural Networks", 17, 20 March 2015.

[29]. Guang-Bin HUANG, "Introduction to Extreme Learning Machines", Hands-on Workshop on Machine Learning for BioMedical Informatics 2006, National University of Singapore, 21 Nov 2006

[30]. Christiane Lemke, "Combinations of Time Series Forecasts: When and Why Are They Beneficial? ", Bournemouth University, PHD dissertation.

[31]. Roselina Sallehuddin, Siti Mariyam Shamsuddin, and Siti Zaiton Mohd Hashim, "Hybridization model of linear and non-linear TS data forecasting", 2008 IEEE, Second Asia International Conference on Modelling & Simulation.

# ANNEX

| MODEL | REMARK | AIC | AICc | BIC |
|---|---|---|---|---|
| ARIMA (0,1,1) (0,0,2) (7) | from auto ARIMA | 21.27146 | 21.27672 | 20.31072 |
| ARIMA (0,1,1) (1,0,1) (7) | single | 21.23431 | 21.23956 | 20.27356 |
| ARIMA (1,1,1) (1,0,1) (7) | single | 21.23865 | 21.24405 | 20.28772 |
| ARIMA (2,1,1) (1,0,1) (7) | single | 21.24027 | 21.24585 | 20.29916 |
| ARIMA (3,1,1) (1,0,1) (7) | single | 21.24452 | 21.25029 | 20.31322 |
| ARIMA (0,1,1) (1,1,1) (7) | single | 21.251 | 21.25614 | 20.28044 |
| ARIMA (1,1,1) (1,1,1) (7) | single | 21.25484 | 21.26009 | 20.29409 |
| ARIMA (2,1,1) (1,1,1) (7) | single | 21.25389 | 21.25929 | 20.30296 |
| ARIMA (3,1,1) (1,1,1) (7) | single | 21.25808 | 21.26366 | 20.31697 |
| ARIMA (1,1,1) (1,0,1) (7) | from previous work--single | 21.23865 | 21.24405 | 20.28772 |

**Table 1**. AIC, BIC and AICc values of candidate SARIMA models

| MODEL | REMARK | RMSE | MAE | MAPE | MASE |
|---|---|---|---|---|---|
| ARIMA (0,0,1) (1,0,1) (7) | auto ARIMA | 33975.65 | 24447.41 | 3.764547 | 0.8537834 |
| ARIMA (0,1,1) (1,0,1) (7) | single | 24500.01 | 16299.69 | 2.522379 | 0.5692384 |
| ARIMA (1,1,1) (1,0,1) (7) | single | 24492.05 | 16288.55 | 2.52042 | 0.5688495 |
| ARIMA (2,1,1) (1,0,1) (7) | single | 24450.82 | 16261.19 | 2.515302 | 0.5678939 |
| ARIMA (3,1,1) (1,0,1) (7) | single | 24443.83 | 16247.78 | 2.513211 | 0.5674255 |
| ARIMA (0,1,1) (1,1,1) (7) | single | 24545.76 | 16284.78 | 2.51 | 0.5687178 |
| ARIMA (1,1,1) (1,1,1) (7) | single | 24532.77 | 16268.16 | 2.506875 | 0.5681373 |
| ARIMA (2,1,1) (1,1,1) (7) | single | 24461.27 | 16219.11 | 2.497542 | 0.5664244 |
| ARIMA (3,1,1) (1,1,1) (7) | single | 24452.79 | 16213.66 | 2.496904 | 0.566234 |
| ARIMA (1,1,1) (1,0,1) (7) | previous work | 24574.23 | 16393.63 | 2.52686 | 0.5725191 |
| ARIMA (0,1,1) (7,0,1) (7) | single | 24282.30 | 16271.63 | 2.516147 | 0.5682586 |
| SARIMA-ARNN | HYBRID1 | 10233.88 | 8186.54 | 1.017230 | 0.6744662 |
| SARIMA-MLPNN | HYBRID2 | 168.67 | 117.66 | 0.014623 | 0.3414394 |
| SARIMA-ELMNN | HYBRID3 | 122.62 | 92.90 | 0.011558 | 0.2773557 |

**Table 2** Forecast performance error metrics results