TUTA/IOE/PCU

*Journal of the Institute of Engineering, 2017, **13(1):** 160-168*

Printed in Nepal

Network Bandwidth Utilization Prediction Based on Observed SNMP Data

Nandalal Rana[1], Krishna P Bhandari[2a], Surendra Shrestha[3]
[1]Himalayan Institute of Science and Technology, Purvanchal University, Nepal
[2]Central Office, Nepal Telecom, Nepal
[3]Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal
[a] Corresponding author: *krishna.bhandari@ntc.net.np*

Received: *April 25, 2017* ***Revised:*** *Dec 21, 2017* ***Accepted:*** *Dec 27, 2017*

Abstract: Bandwidth requirement prediction is an important part of network design and service planning. The natural way of predicting bandwidth requirement for existing network is to analyze the past trends and apply appropriate mathematical model to predict for the future. For this research, the historical usage data of FWDR network nodes of Nepal Telecom is subject to univariate linear time series ARIMA model after logit transformation to predict future bandwidth requirement. The predicted data is compared to the real data obtained from the same network and the predicted data has been found to be within 10% MAPE. This model reduces the MAPE by 11.71% and 15.42% respectively as compared to the non-logit transformed ARIMA model at 99% CI. The results imply that the logit transformed ARIMA model has better performance compared to non-logit-transformed ARIMA model. For more accurate and longer term predictions, larger dataset can be taken along with season adjustments and consideration of long term variations.

Keywords: Bandwidth Prediction, ARIMA, Logit Transformation, MAPE

1. Introduction

The rapid growth in the use of the Internet has led to huge and increasing demand for bandwidth, both international and domestic. Ensuring sufficient bandwidth in their network from core to the customer premises has been the constant challenge for network operators and service providers. In this context, timely and accurate prediction of bandwidth is very helpful to plan the network resources, expansion and upgrades on time so that bottlenecks and QoS degradations are avoided. This, in turn, is crucial for operators to attract, retain and regain customers. Hence, proper network traffic prediction is very useful for operators in initial planning as well as later on for upgrades and optimizations. To efficiently handle increasing subscriber base, service diversity and ensuing increase in bandwidth needs, service providers need to continuously monitor and predict traffic. For this research the reference data is taken from the utilization statistics of links in the Internet network of Nepal Telecom in the Far Western Region of Nepal. The data thus taken has been used to predict traffic using the method proposed here.

2. Literature Review

In a packet based network such as Internet, traffic flows through different links and nodes to reach from source to destination. The capability to predict future traffic demands can help improve the performance of network with timely planning for upgrades and updates. The self-similarity of network traffic has been studied in Local Area Network (LAN) [2]. To fit the bursty nature of the traffic ARIMA/GARCH model, which is a combination of a linear ARIMA (Auto-Regressive Integrated Moving Average) and non-linear Generalized Autoregressive Conditional Heteroskedasticity (GARCH) time series models, was proposed in [8]. According to [4], methodology for network resource management and network traffic congestion control using properly trained Artificial Neural Network (ANN) can better handle non-linear data relationship between the input and output patterns. The prediction model of network traffic based on Milti-Kernel Support Vector Regression (MKSVR) method showed that the proposed model can accurately describe the change trend of network traffic and hence improve the prediction accuracy of network traffic [6].

Several models have been proposed to predict network traffic. The short-term (a few minutes) forecast model using Auto-Regressive Moving Average (ARMA) with 1 sec time-scale data is proposed in [5]. The long-term (1 year) forecast model of Internet backbone traffic using ARIMA with 1 week time-scale data is proposed in [3]. The mid-term (1 day) forecast model using Autoregressive Conditional Heteroskedasticity (ARCH) model with 15 minute timescale data is proposed in [1]. This research focuses on three months prediction of the bandwidth utilization using 30 second time-scale data.

3. Methodology

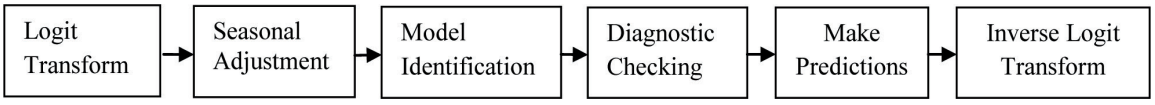

Fig. 1: Research Schematic Model

For this research, real traffic data of 5 months for the Far Western Region network of Nepal Telecom was taken. The data is based on the SNMP capture for NMS visualization of the designated network nodes. This data is used as the training data for creation of the time series dataset that is used for the application of logit transformation and time series functions. The network links associated with the collected reference data is shown in Table 1.

Table 1: Data Source Links

S.N.	Source	Destination
1	Dhangadhi	Butwal
2	Dhangadhi	Nepalgunj
3	Dhangadhi	Hetauda

Daily aggregated traffic data has been taken of the Dhangadhi node for 6 months. This data is based on the NMS records which are aggregated every 30 seconds. The data collected was the inbound and outbound traffic at Dhangadhi node from the links given in Table 1. The inbound and outbound reference data thus collected are shown diagrammatically in Fig. 2 and Fig. 3 respectively. Six

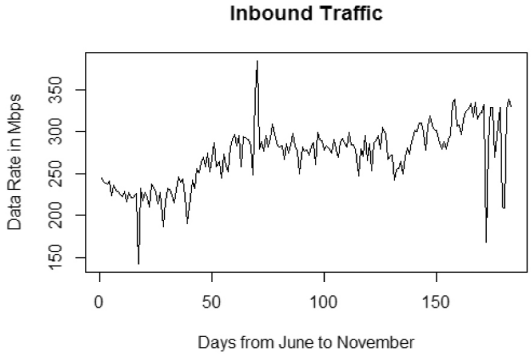

Fig. 2: Inbound Traffic Time Series

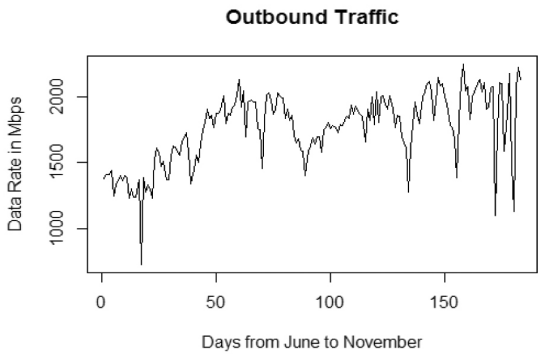

Fig. 3: Outbound Traffic Time Series

The framework for prediction based on collected data is shown in Fig. 4.

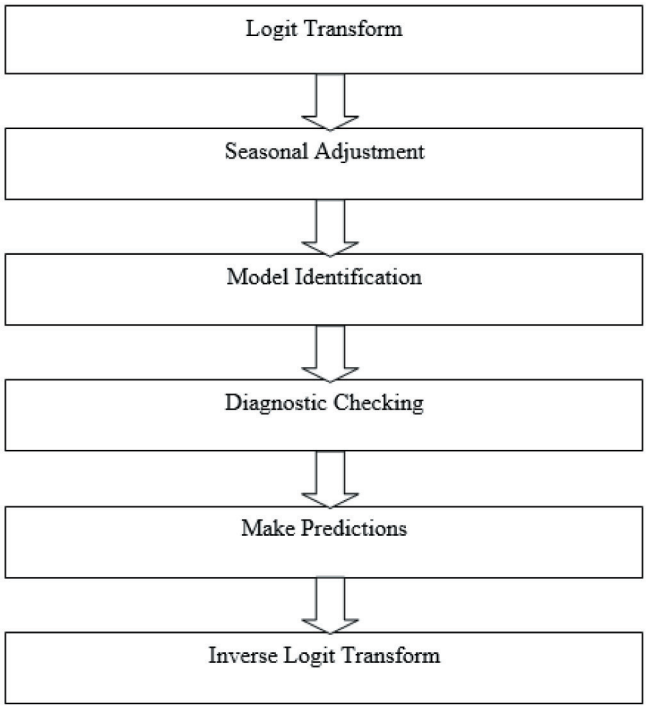

Fig. 4: Framework for Prediction

A univariate linear time series logit transformed ARIMA model is used for bandwidth utilization prediction of network traffic using the observed SNMP data.

3.1 Logit Transformation

Logit transformation is applied to the SNMP data *'x'* to set the lower and upper bounds based on these limits. Time series data *'x'* containing *'n'* observations is transformed to time series data *'y'* with lower bound 'a' and upper bound 'b' (10^{10} bit/second) using the expression shown in Equation (1).

$$y = logit(x) = \log[\frac{(x-a)}{(b-x)}] \quad (1)$$

The logit transformed SNMP data obtained after applying the transformation logic of Equation 1 are tested for seasonal adjustment using Seasonal Decomposition of Time Series by Loess (STL) in R programming.

3.2 Model Identification

The entire data set is split into training data (June to October) and testing data (November) for model validation. The purpose of model validation step is to determine accuracy of the model. For identification of the model, the selection the orders of p, d and q is performed using the "*auto. arima*" function in R programming which satisfied the Akaike's Information Criterion (AIC). The model with the minimum AIC is chosen as the best candidate model. The reason for the choice of AIC is because of the fact that AIC not only evaluates the fit between values predicted by the model and actual measurements, but also penalizes models with larger number of parameters.

3.3 Diagnostic Checking

The ARIMA diagnostic tests are run on the most promising of the model orders. Before fitting the accurate ARIMA model, the diagnostic checks of stationarity and normality are carried out.

Stationarity Test: A stationary time series is one whose properties do not depend on the time at which the series is observed. Time series with trends, or with seasonality, are not stationary. The stationarity of the residuals of the series is to be checked using Kwiatkowski-Phillips-Schmidt-Shin (KPSS) and Augmented Dickey–Fuller (ADF) tests. Small p-values (<0.05) of ADF test confirms the stationarity. Large p-values (>0.05) of KPSS test also confirms the stationarity of the time series.

Normality Test: The non-existence of serial correlation of the residuals of the series indicates the condition of normality. Box-Ljung and Box–Pierce tests are used to confirm the normality. Large p-values (>0.05) of Box-Ljung and Box–Pierce tests confirmed the normality.

3.4 Traffic Predictions

The network traffic is predicted after the order selection using "*auto.arima*" function. Then these predicted values are converted to the original scale using reverse logit transformation using Equation (2).

$$x_h = (b-a)\left[\frac{e^{y_h}}{(1+e^{y_h})}\right] + a \quad (2)$$

A one-step ahead prediction method is used recursively to obtain the subsequent values. The data of six months from June to November of a year of FWDR have been utilized to predict the network traffic.

3.5 Prediction Accuracy Check

A prediction error is the difference between the actual or real and the predicted value of a time series. For prediction accuracy check MAPE is used which is given in Equation (3).

$$MAPE = 100n^{-1}(\sum(|y_t - f_t|)/|y_t|) \quad (3)$$

where y_t is actual traffic and f_t is predicted traffic.

The normality test is carried out to test whether the data is normally distributed or not. When the data is normally distributed, then both mean or the median can be used as measures of prediction errors. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of prediction errors because it is the measure that includes all the values in the data set for its calculation and any change in any of the data will affect the value of the mean. This is not the case with the median or mode.

4. Results and Discussion

The entire data set is split into training data (June to October) and testing data (November) for model validation. The purpose of model validation step is to determine accuracy of the model . For identification of the model which satisfies the minimum Akaike's Information Criterion (AIC), the "*auto.arima*" of R programming is used. The function has the form where p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity and q is the number of lagged forecast errors in prediction equation.

The simulations have been carried out to identify the best ARIMA model and compare the errors of the other time series prediction models using R Studio package of R programming. For model validation, the entire data set is split into training dataset (June to October) and testing data set (November). The training for inbound and outbound are shown in Fig. 2 and Fig. 3 respectively. Similarly, the testing data (actual data of November) for inbound and outbound are shown in Fig. 5 and Fig. 6 respectively.

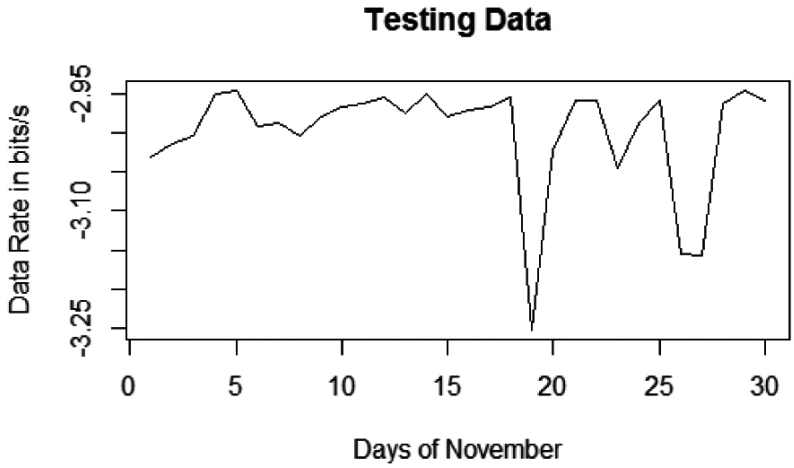

Fig. 5: Inbound Testing Data

An ARIMA model in training data set is implemented and found is a best model as per minimum Akaike information criterion using "*auto.arima*" function of R programming in forecast package. Seasonal components are also tested using "*decompose*" function in *R* and no seasonal components were found.

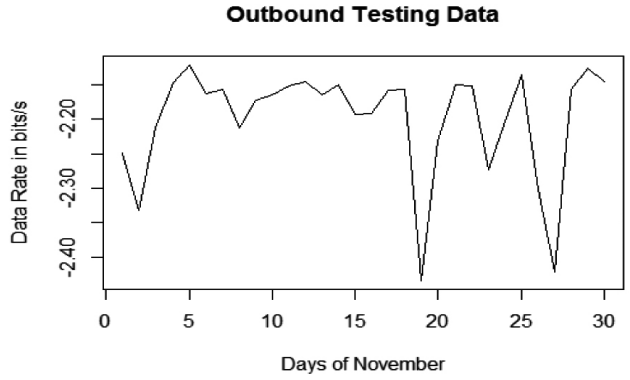

Fig. 6: Outbound Testing Data

KPSS and ADF tests are implemented to check the stationary condition of inbound and outbound traffic time series of the best fitted model which are found that the both inbound and outbound traffic of the best fitted model is stationary. The best fitted model is selected using "*auto.arima*" function and it is found as for both inbound and outbound traffic.

Box-Ljung and Box-Pierce tests are performed for testing normality of the bested fitted model of the inbound traffic time series which are found that the both inbound and outbound traffic time series of the best fitted model are normally distributed.

Fig. 7 and Fig. 9 show the actual traffic from June to October and predicted traffic for November for inbound and outbound respectively. Similarly Fig. 8 and Fig. 10 show the predicted traffic for November with its actual values of inbound and outbound traffic respectively. The dotted line indicates the actual traffic and dark and light shaded indicate the 80 and 95 percent CI prediction.

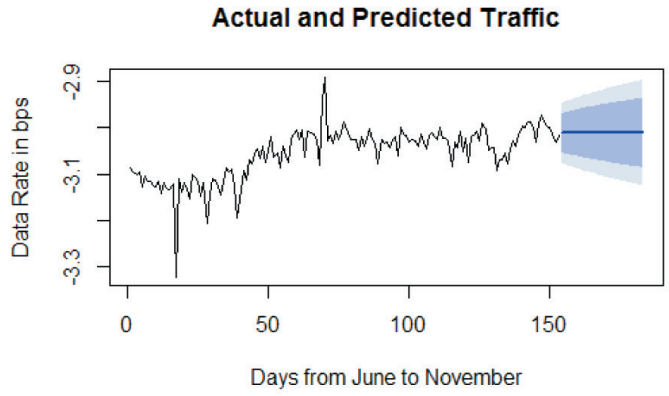

Fig. 7: Actual Data for 5 months and Predicted Range for Sixth Month for Inbound

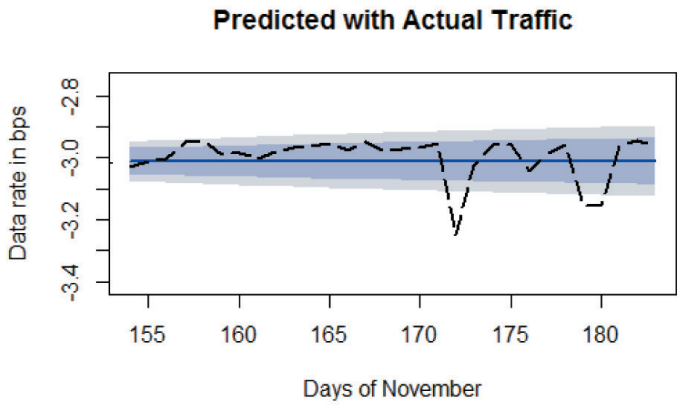

Fig. 8: Actual Data and Predicted Range for Sixth Month for Inbound

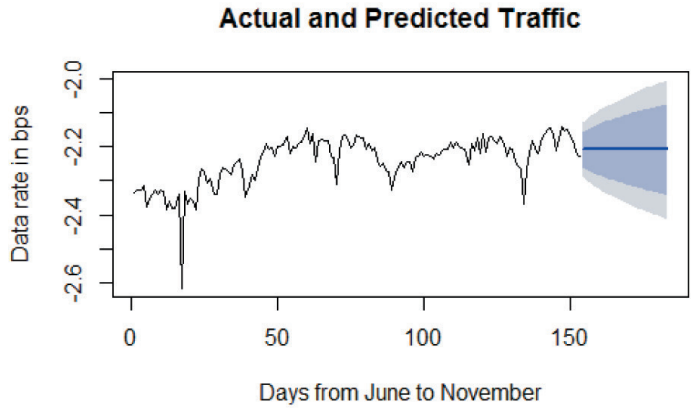

Fig. 9: Actual Data for 5 Months and Predicted Range for Sixth Month for Outbound

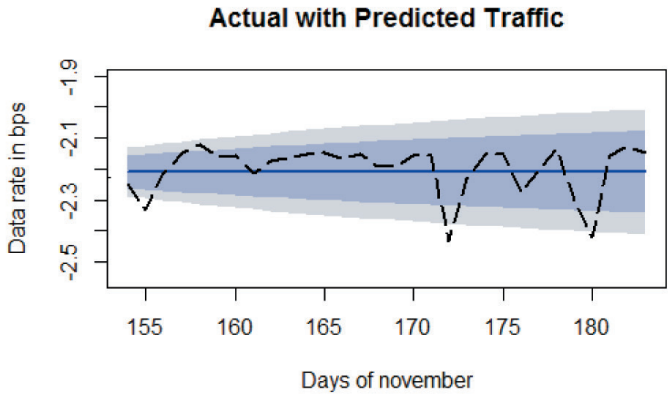

Fig. 10: Actual Data and Predicted Range for Sixth Month for Outbound

MAPE is used for measurement of in-sample errors. On the basis of MAPE, one step ahead

prediction method is chosen because of its minimum MAPE at 99% CI. The prediction of the traffic for the next time instant is performed recursively for the next day taking six months traffic data. This process is recursively performed for the next 90 days. The predicted inbound and outbound traffic time series plots for 90 days at 99% CI are shown in Fig. 11 and Fig. 12 respectively.

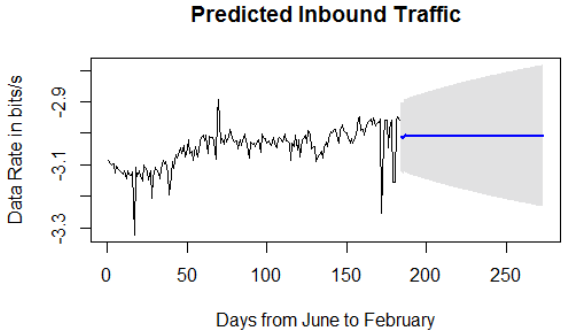

Fig. 11: Predicted Inbound Traffic for 90 Days

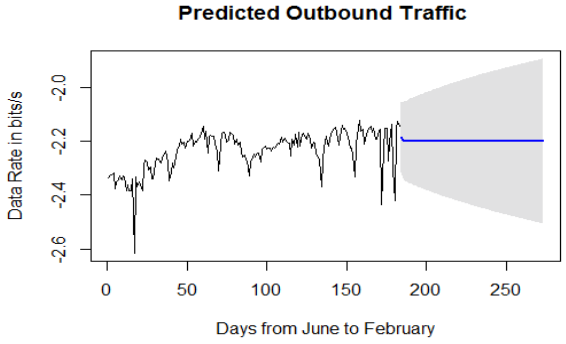

Fig. 12: Predicted Outbound Traffic for 90 Days

A comparative study of the existing ARIMA model with logit transformed and non-logit transformed time series data is carried out to evaluate the prediction accuracy. It is found that the forecast error of logit transformed ARIMA model is significantly smaller than the non-logit transformed ARIMA model which clearly states the prediction efficiency of ARIMA model with logit transformed time series data. The logit transformed ARIMA model fitted the traffic data very well and this is confirmed by the goodness of fit test which are given by stationarity and normality. Such well-fitted statistical evidence is a solid empirical foundation for the prediction. The logit transformed ARIMA model reduces the MAPE by 11.71% and 15.42% respectively as compared to the non-logit transformed ARIMA model at 99% CI of inbound and outbound traffic.

5. Conclusion

A network bandwidth utilization prediction scheme is presented and analyzed using univariate linear time series model with ARIMA model after logit transformation. The logit transformed

ARIMA model reduces the MAPE by 11.71% and 15.42% for inbound and outbound traffic respectively as compared to the non-logit transformed ARIMA model at 99 % CI. The reference data was available only for 6 months. It can be argued that if longer period of data is available, more accurate prediction for even longer duration can be achieved. For further research, larger dataset can be taken and season adjustments and long term variations can be taken into account to carry out predictions with longer periods with better confidence levels.

References

[1] Krithikaivasan B, Deka K and Medhi D (2004), Adaptive Bandwidth Provisioning based on Discrete Temporal Network Measurements, *IEEE INFOCOM 04, Hong Kong, China,* 1786–1796.

[2] Leland WE, Taqqu MS, Willinger W and Wilson DV (1993), On the Self-Similar Nature of Ethernet Traffic, *Proceedings of ACM SIGCOMM,* **93 :** 183–193.

[3] Papagiannaki K, Taft N, Zhant Z and Diot C (2003), Long-Term Forecasting of Internet Backnone Traffic: Observations and Initial Models, *IEEE INFOCOM* **03 :** 1178–1788.

[4] Piedra N, Chicaiza J and Lopez J (2014), Study of the Application of Neural Networks in Internet Traffic Engineering, *International Book Series, Information Science and Computing Advanced Research in Artificial Intelligence,* 33-47.

[5] Sang A and Li SQ (2000), A Predictability Analysis of Network Traffic, *Infocom 2000,* **1 :** 342 – 351.

[6] Xiang C, Qu P and Qu X (2015), Network Traffic Prediction Based on MKSVR, *Journal of Information & Computational Science,* 3185–3197.

[7] Yoo W and Sim A (2014), Network Bandwidth Utilization Forecast Model on High Bandwidth Network, *LBNL Paper LBNL-6677E*.

[8] Zhou B, He D and Shun Z (2005), Traffic Modeling and Perdiction using ARIMA/GARCH Model, *Symposium on Modeling and Simulation Tool for Emerging Telecommunications Networks: Needs, Trends, Challenges and Solutions, Meunchen, Germany*.