Available online at www.ejournals.uofk.edu

UofKEJ Vol. 12 Issue 1 pp. 23- 27(February 2022)

UNIVERSITY OF KHARTOUM ENGINEERING JOURNAL (UofKEJ)

Forecasting of Data Traffic in Sudan Internet Exchange Point Using Autoregressive Integrated Moving Average Model

Asma Awad[1], Khalid Hassan [2]

Department of Electronics
University of Khartoum
Khartoum, Sudan

Department of Electronics
University of Garden City
Khartoum, Sudan

(E-mail: asmaabcd15@gmail.com, khalidhmm@gmail.com*)*

Abstract: Network traffic prediction is an important issue that has received much interest recently from computer network community. The network traffic prediction is one of the typical issues useful for monitoring network, network security, avoid congestion and increase speed of networks. Different techniques are used for network traffic prediction; first are linear time predictors such as Last Value (LV) Predictor, Windowed Moving Average (MA), Double Exponential Smoothing (DES), Auto Regressive (AR), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA). The Second techniques are nonlinear time predictor such as generalized autoregressive conditional heteroscedasticity (GARCH) and the last one is hybrid model techniques, and it is combination between two or more models. From literature review, the ARIMA model is the best model to predict data traffic that has time series specification and linear growth. Data traffic collected from Sudan Internet Exchange Point (SIXP) and ARIMA is used to model data and forecast the next five years values. In addition to forecasting values, upper, and lower values are also founded. Furthermore, ARIMA model findings are compared with Monte Carlo forecast and it is found that the results are nearly typical. Finally, when the forecasted results are compared to actual data traffic on January 2020, the predicted values are shown to be very close to the measured ones.

Keywords*: Internet Exchange point (IXP); ARIMA; forecasting data traffic.*

1. INTRODUCTION

The Internet is one of the most complex technology infrastructures of modern civilization. Defined as a network of networks, the Internet has been generally increasing every year in terms of usage and geographical range since the early days of its development. Latest technology advances along with the global population growth have resulted in more bandwidth and access speed. Thus, the communication equipment and underlying technical issues for Internet Service Providers (ISPs) and backbones would have to be determined accordingly and, ideally, in advance. The total number of Internet users across the world increases every year and so does the aggregate traffic which they generate on a continent and on a global scale. Therefore, modeling and predicting the growth aspects of Internet traffic is technically and economically important as well as challenging and has received wide attention from researchers mainly since the 1990s [1].

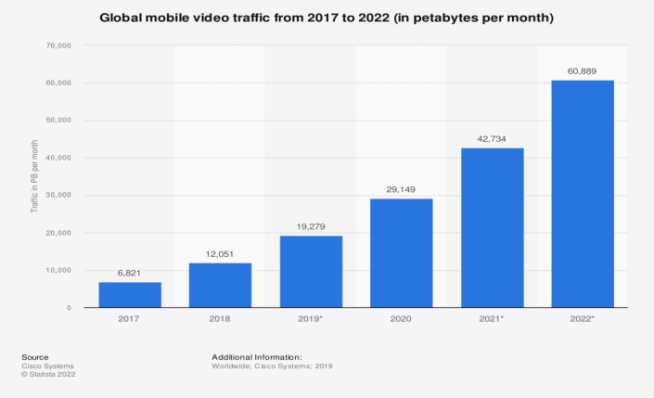


Fig. 1. Global internet traffic growth per year

Internet exchange points (IXP) created in the early 90s [2], are physical locations where different networks are connected to Exchange Internet traffic via common switching infrastructures. They are key parts of the Internet ecosystem and represent a vital way to increase the affordability and quality of connectivity in local communities [3]. The number of IXPs in the world is 1061 IXP from which only 79 are in Africa [4]. IXPs usually; are dispersed across Countries to enable local networks to efficiently exchange information by eliminating the need to exchange local Internet traffic overseas. IXPs create efficient interconnection points that encourage network operators to Connect in the same location in search of beneficial peering arrangements, cheaper and better traffic exchange, and other information and communication services. The presence of an IXP also can attract out-of-country service operators.

A single Connection to an IXP provides out-of-country service operators with lower collective access costs to multiple potential local customers. Because of this, IXPs are uniquely able to encourage the development of a region's communications infrastructure, including national and international fiber cables and local data center development [3].

There are two types of IXP where Layer 2 and Layer 3 IXPs commonly exist. Layer 2 uses a common network medium like Ethernet (10/100/1000 Base technologies) and all traffic is exchanged outside routers that are connected to a shared media, members bring their own routers and circuits from their backbone.

No transit or customer connections and also members of the IXP determine who they peer with? No one have to peer with everyone, which is depicted in Figure 2. In layer 3 all traffic is

exchanged *inside* a router (One extra as hop between peers); Layer 3 IXPs network which is illustrated in Figure 3 has limited the autonomy of the members also ISPs don't have control with whom they can peer with, Layer 3 IXP today is marketing concept used by transit ISPs where real Internet exchange points are only Layer 2 [4].

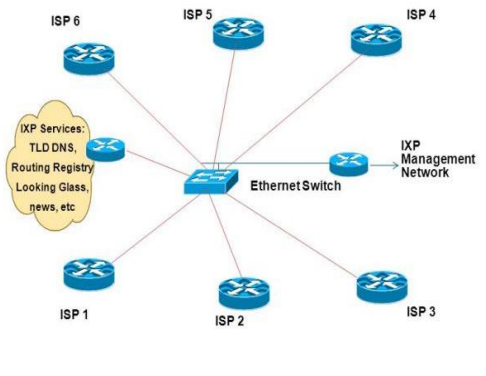


Fig. 2. Layer 2 IXP network topology

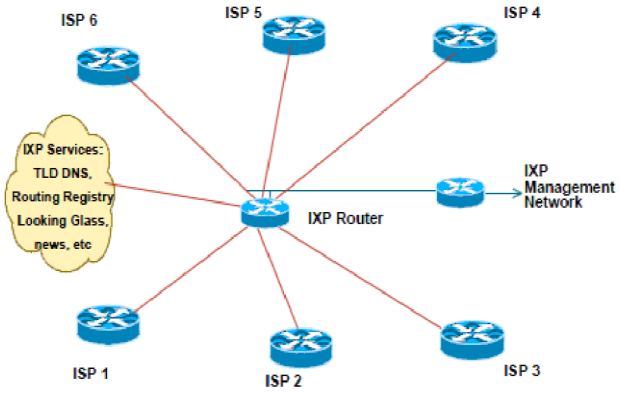


Fig. 3. Layer 3 IXP network topology

Network traffic prediction is an important issue that has received much interest recently from computer network community. The network traffic prediction is one of the typical issues that is useful for monitoring network. Network security, is used to avoid congestion and increase speed of networks. Different techniques are used by researchers for network traffic prediction such as ARMA and ARIMA models [6].

The predictability of network traffic is of significant interest in many domains. Two categories of prediction can be distinguished: long and short period predictions. Traffic prediction for long periods provides a detailed forecasting of the workload and traffic patterns to assess future capacity requirements, and therefore allows for more accurate planning and better decisions. Short period prediction (milli-seconds to minutes) is relevant for dynamic resource allocation. It can be used to improve the Quality of Service (QoS) mechanisms as well as congestion and resource control by adapting the network parameters to traffic characteristics. It can also be used for routing packets [7].

2. Sudan internet exchange point

Sudan Internet Exchange Point (SIXP) was founded in 2011[7]. Like other IXPs, SIXP aggregates, most ISPs are operating in the country. SIXP evolved and now has eight members (Sudatel, Canar, MTN, ZAIN, Maxnet, Sudan internet society (SIS). However, national data center (NDC) and Sudanese Universities Information Network (SUIN) exchange around 603.23 Mbps of two-way traffic in the peak hour [8]. The Sudan internet exchange point network topology is clearly demonstrated in Figure 4.

In addition, Figure 5 shows a (partial) snapshot of a web page, it contains the data from ZAIN which is a large access provider. The information was time-stamped on 13/11/2019 at 12:52 pm [7], with a port capacity 5 out of 1Gbps (value Max Speed in the graph), listing the average, max and current demand for both the inbound and outbound traffic, on daily, weekly(shown), monthly and yearly basis (not shown in Figure 2).

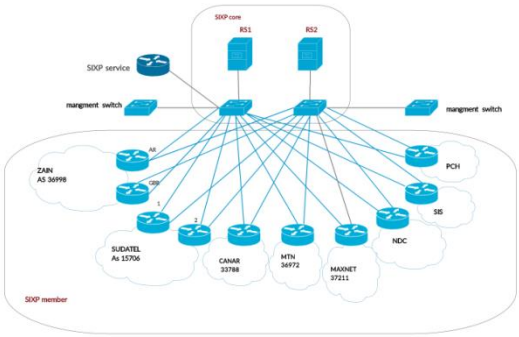


Fig. 4. Sudan internet exchange points network topology

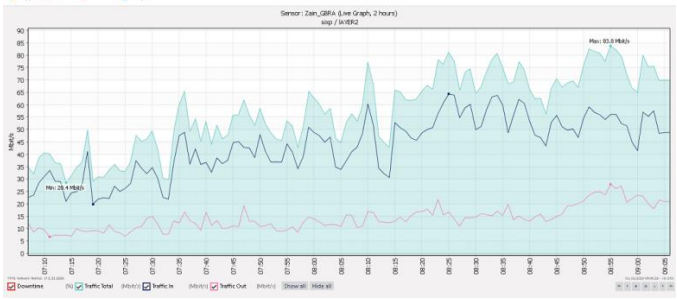


Fig. 5. Sample of SIXP daily traffic

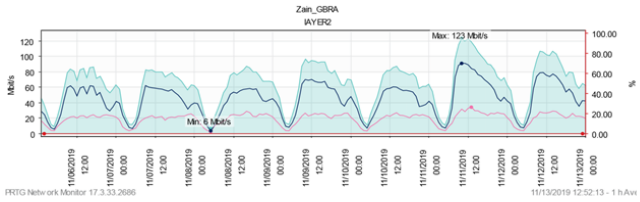


Fig. 6. Autoregressive Integrated Moving Average Model

ARIMA model is an integrated ARMA model. also known as the famous Box-Jenkins Model, which was a time series modeling, founded in the early 70s by Box and Jenkins, and it was a mixed form of AR and MA, which can be extended to analyze the time series including seasonal trends, ARIMA play an important role with time series model to predict network traffic[8].

AR (p):

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + ... + a_p x_{n-p} + \varepsilon_n \quad (1)$$

MA (q):

$$\varepsilon_n = e_n - \theta_1 e_{n-1} - \theta_2 e_{n-2} - .... - \theta_q e_{n-q} \quad (2)$$

ARMA (p, q):

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + ... + a_p x_{n-p} + e_n - \theta_1 e_{n-1} - \theta_2 e_{n-2} - .... - \theta_q e_{n-q} \quad (3)$$

Where AR is autoregressive, MA is moving average P is autoregressive order, q is moving average, X is observation value, a is predication coefficients, e is prediction error, θ is error coefficients, Ɛ is white noise and n represents the time domain.

If taking a non-stationary time series by d time's difference, which becomes stable, then with a smooth ARMA (p, q) as its generation model, it can be said that the original time series is an autoregressive integrated moving average time series, denoted as ARIMA (p, d, q). Where AR is the auto regressive, p is the factor of auto regression, MA is the moving average, q is the moving average life cycle, d is the frequency difference since time series become stationary [2], then the prediction model can be written as in equation (3). In the formula, x_n are sample values in time domain n, a_i and θ_i are coefficients to be estimated, which are the model parameters, where the i takes values in the range between 1 and p. ε_n is the white noise sequence complied with normally distributed, meanwhile p, *d and q* are the orders of model.

ARIMA model which is used for forecasting has 5 steps described as follows [9]:

Step 1: Data validation: Based on time series' histograms, autocorrelation and partial autocorrelation function are plotted in order to test of its variance, trend and seasonal variation by Augmented Dickey-Fuller (ADF) Test Unit root, to identify stationary of the sequence.

Step 2: Stationary processing: when the data are non-stationary, autocorrelation function (ACF) and partial autocorrelation (PACF) are used to process the non-stationary sequence to stationary state.

Step 3: The ACF and PACF are analyzed, then through the AKAIKE information criterion (AIC). AIC is used to determine the model's order parameters which are p, d, q.

Step 4: Parameters estimation and model diagnostics,: using maximum likelihood estimation criterion is used to obtain all model parameters (a_i), then these parameters are tested in the model, then diagnosis whether the residual series error is white noise or not in order to determine the appropriateness of the model, if it does not appropriate then it is required to re-estimate parameters.

Step 5: These appropriate parameters are used to predict the future values. However, this research uses the ARIMA model for the purpose of predication.

3. data traffic predication FRAMEWORK

The prediction model for each input at time n− 1, the model calculates $\hat{x}_n$ which is the prediction value for x_n. The inputs can be previous observations (i.e., lags of x_{n-i} or any exogenous variable measured at time n− 1). Generally, the methodology is categorized to four steps as shown in Figure 7.

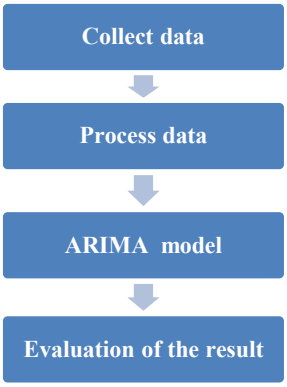


Fig. 7. Predication steps

1- In our predication, data traffic is collected from SIXP using Simple Network Management Protocol (SNMP) protocol with PRTG software. Data traffic from January to December for three years from 2017 to 2019 ago were collected as depicted in Table 1.

Table 1: Data Traffic for SIXP members

ISP name	2017 /Tb	2018 /Tb	2019 / Tb
MTN	13.06	29.01	171.04
ZAIN	12.83	50.23	288.91
CANAR	18.12	19.21	66.98
SUDATEL	14.81	37.05	180.89
MAXNET	0.82	5.56	18.64
NDC	0	0	6.24
PCH	0	0	2.817
Total	59.64	141.06	735.487

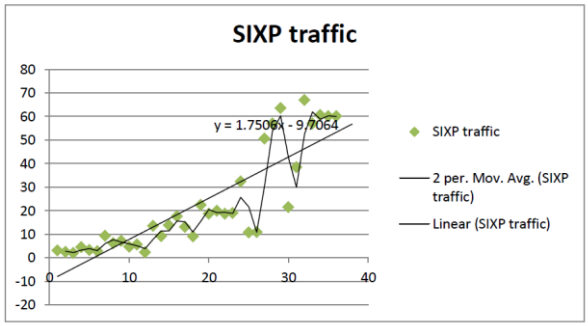


Fig. 8. SIXP data traffic

2-Data processing:

The data was processed using log transformation to provide stationary data over time. The model is identified based on time domain and frequency domain analysis i.e. autocorrelation function (ACF), partial autocorrelation function (PACF) and spectral density function. A graph of autocorrelation function determines whether the series is stationary or not. The time series is considered stationary if the graph of ACF values either cuts off fairly quickly or dies down fairly quickly. The series is considered as non-s tationary if the graph of ACF dies down extremely slowly. In case of the non-stationary series, it can be converted to a stationary series by successive differencing.

3-ARIMA model

In network traffic prediction techniques, many different metrics are used to investigate the quality of time series forecasting. The detection rate, false positive rate, accuracy and time cost metrics are employed for measuring the performance of classifier for different data set. A number of metrics exist obtained to express prediction accuracy. The metric used in this paper the mean squared error which is defined in equation (4)[8]:

$$MSE = \sqrt{\frac{\sum_{n=1}^{p}(x_n - \hat{x}_n)}{p}} \qquad (4)$$

where

x_n is the actual value and

$\hat{x}_n$ is the predicted value

The prediction interval can be conventionally formulated as in equation (5):

$$[\mu - z\sigma, \mu + z\sigma] \qquad (5)$$

Where

μ is the mean, σ is the standard deviation and Z is predication interval.

For example, to calculate the 95% prediction interval for a normal distribution with a mean (μ) of 5 and a standard deviation (σ) of 1, then z is approximately 2. Therefore, the lower limit of the prediction interval is approximately 5 – (2·1) = 3, and the upper limit is approximately 5 + (2·1) = 7, thus giving a prediction interval of approximately 3 to 7 [5]. However, Table 2 illustrates these concepts of calculating z for different range of precision [9, 10].

Table 2: predication interval Z value [5]

Prediction interval	Z
75%	1.15
90%	1.64
95%	1.96
99%	2.58

2- Evaluation the of result

In this step prediction result is evaluated by comparing between ARIMA and Monte Carlo simulation.

Monte Carlo method is a statistical method of understanding complex physical or mathematical systems by using randomly generated numbers as input into those systems to generate a range of solutions. The likelihood of a particular solution can be found by dividing the number of times that solution was generated by the total number of trials. By using larger numbers of trials, the likelihood of the solutions can be determined more and more accurately. The Monte Carlo method is used in a wide range of subjects, including mathematics, physics, biology, engineering, and finance, and in problems in which determining an analytic solution would be too time-consuming [9, 10].

Findings and Discussion

In the first step data are collected from SIXP over a peri od of three years (36 month) as shown in Figure 9 (a), and the Figure 9 (b) show the Log transformation of data traffic to provide stationary data over long time (i.e., old data, recent historical data as well as newly acquired data).

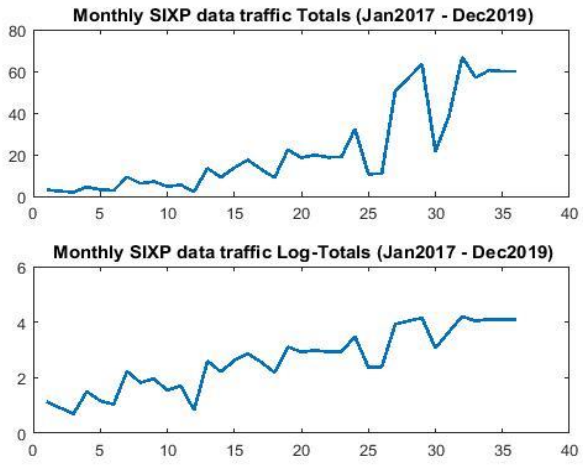


Fig. 9. SIXP data traffic from January 2017 – December 2019

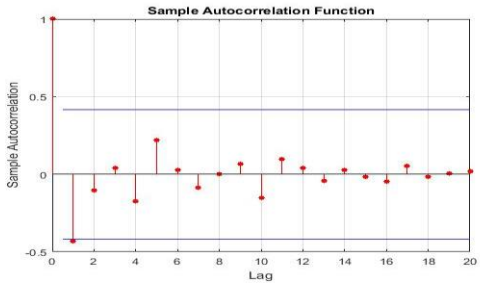
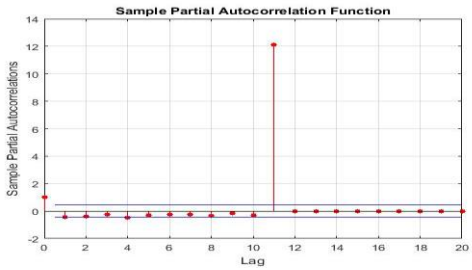


Fig. 10 .sample autocorrelation function



Fi g. 11 .sample partial autocorrelation function

The next step is to fit the ARIMA model with parameter as demonstrated in Table 3, and use the result of this model with MSE to predict the upper, lower limits, and forecast data traffic as figured in Figure 12; however, it is found that the value of the error increases as time increases.

Using ARIMA (0, 1, 1) model seasonally integrated with seasonal MA (12) and Gaussian Conditional Probability Distribution, the flowing results which are shown in Table 3 are obtained:

Table. 3. Model performance in terms of monthly data traffic

Standard			T
Parameter	**Value**	**Error**	**Statistic**
Constant	0	Fixed	Fixed
MA{1}	-0.88638	0.0940479	-9.42477
SMA{12}	-0.0042299	0.274616	-0.015403
Variance	0.2152	0.0289247	7.43999

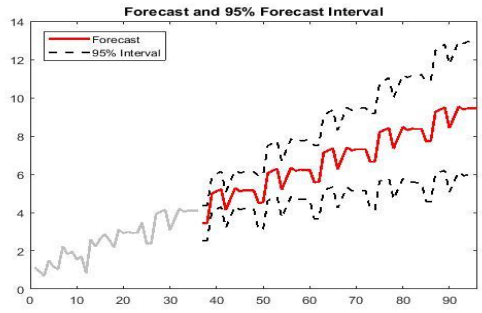


Fig. 12. Forecasted data traffic

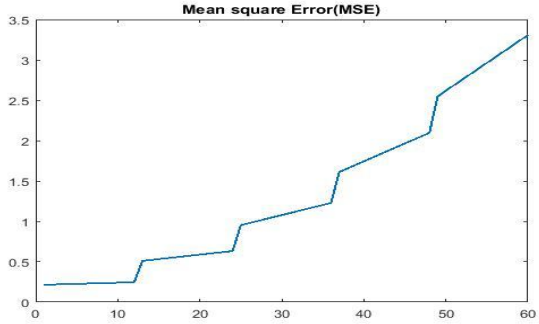


Fig. 13. Minimum mean square error

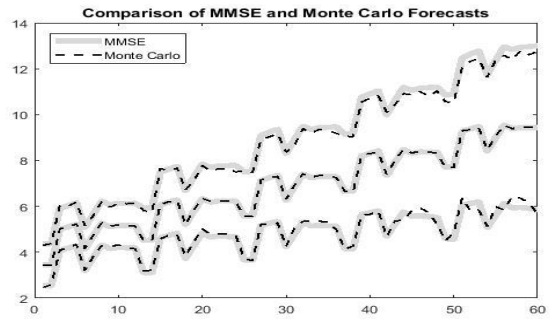


Fig. 14. Comparison of MMSE and Monte Carlo forecasting

Conclusion

Predication data traffic is the current focus of several researchers due to the growing interest in internet data traffic. However, data traffic forecasting method, which needs to be as accurate as possible, is needed to help decision makers to make frame strong strategic plans, in this paper, ARIMA model is used to predict the monthly data traffic with seasonal lag and mimic the seasonality and monthly cyclic nature of data traffic in Sudan internet exchange point. The model was evaluated using the comparison between ARIMA model and Monte Carlo forecasts and it was found that the result is similar. In addition, when the forecasting result is compared with data traffic in January 2020, it is found that the actual data traffic is 82 Tetra-Byte meanwhile the predicted data traffic is 80 Tetra-Byte.

Future Work

It seems that the internet demand growth is a nonlinear growth, meanwhile the ARIMA model is linear time series model and the demand of the internet data traffic is not linear and this fact can be affected in the result of the forecasting. Therefore, it is recommended to use nonlinear prediction model in order to get more precious forecasting data.

REFRENCE

[1] Nikolaos Vlachos, Internet Traffic Volumes Characterization and Forecasting, Submitted for the degree of Doctor of Philosophy, HeriotWatt University, School of Mathematical and Computer Sciences, Department of Computer Science, May 2016.

[2] Gaurab Raj Upadhaya ,Internet Exchange Point ,RIPE NCC Regional Meeting ,Bahrain, 15 November 2006

[3] CHAOBA NIKKIE ANAND ,M.E., INTERNET TRAFFIC MODELING AND FORECASTING USING NON-LINEAR TIME SERIES MODEL GARCH, Bangalore University, India, 2004 .

[4] https://www.pch.net/resources/papers.php, last access is on on 8 Jan 2023.

[4] An Internet Society Public Policy Briefing , Internet Exchange Points ,30 October 2015

[5] Manish R. Joshi1, School of Computer Sciences, North Maharashtra University, Jalgaon (M.S) India ,joshmanish@gmail.com ,Theyazn Hassn Hadi2 , School of Computer Sciences, North Maharashtra University,

Jalgaon (M.S) India,thhadi@nmu.ac.in ,A Review of Network Traffic Analysis and Prediction Techniques , 2015 .

[6] Mohamed FatenZhani and Halima Elbiaze ,Department of Computer Sciences, University of Qu´ebec in Montr´eal, Canada ,{zhani.mohamedfaten, elbiaze.halima}@courrier.uqam.ca Date??Farouk KamounNational School of Computer Sciences, Manouba 401 0, Tunisia kamoun@planet.tn . Analysis and Prediction of Real Network Traffic , JOURNAL OF NETWORKS, VOL. 4, NO. 9, NOVEMBER 2009

[7] National Information center, Sudan internet exchange point, last access is on 8 January 2023.

[8] Ni Lihua, Chen Xiaorong, Huang Qian , College of Computer Science and Information Guizhou University ,Guiyang, China , nilihua8707@126.com, xrchengz@163.com,ARIMA Model for Traffic Flow Prediction Based on Wavelet Analysis, 2010 IEEE

[9] Paulo Cortez, Miguel Rio, Miguel Rocha, and Pedro Sousa, 2006, Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006 Internet Traffic Forecasting using Neural Networks, International Joint

[10] Sanam Narejo and Eros Pasero, An Application of Internet Traffic Prediction with Deep Neural Network, August 2018, DOI:10.1007/978-3-319-56904-8_14.