Internet Access Link Speed between ISPs' Myths and Subscribers' Expectations

Mohammad Z. Masoud[1], Yousef Jaradat[1], Ismael Jannoud[1], Omar Al-heyasat[2]
[1]Computer and Communication Engineering Department
Engineering and Technology Faculty
Al-Zaytoonah University of Jordan
130 Amman 11733 Jordan
[2]Computer Engineering Department
Al-Balaqa' Applied University
Al-Salt- Jordan
{m.zakaria, y.jaradat, Ismael.jannoud}@zuj.edu.jo

Abstract: - Due to the eruption of Internet contents and the need to interact with these contents quickly, Internet speed and subscribers' download and upload bandwidths become important limiting issues. Increasing the purchased bandwidth from Internet Service Provider (ISP) can mainly solve these issues. However, the purchased bandwidth is not the real bandwidth subscribers obtained and paid for. Many factors impact the real obtained bandwidth. One of these important factors is the time of the day a subscriber access the Internet.

In this work, we attempt to measure Internet connection speed over time. The recorded values proved the impact and effect of time on the speed and quality of the purchased bandwidth. To this end, we proposed a prediction models based on time series analysis and nonlinear autoregressive with exogenous input (NARX) neural network to predict the actual Internet connection bandwidth over time. Two NARX models have been implemented; the first one is for the download bandwidth and the other is for the upload bandwidth. These models obtained 86% and 88% in the validation test. These models can be utilized to predict the actual bandwidth an ISP can offer to a customer.

Key-Words: - Time Series, Nonlinear autoregressive with external input (NARX), Bandwidth, Internet Speed, Neural Network

1 Introduction

Due to Internet content inflation, Internet access speed is one of the important issues around the Internet. Multimedia, VoIP, VoD and online streaming are carried across the Internet. These contents are bandwidth hungry.

Network bandwidth is defined as the number of bits that network carries per second [1]. Among non-technical users, network bandwidth is known as network speed. This is misleading information. When users purchase an Internet connection from an Internet service provider (ISP), they ask about the Internet connection or access speed (IAS). ISP's customers' service staff usually answers with vague words, like, the speed is *up to* 2 Mbps. However, this answer is only the download bandwidth of the access link or the last mile link. Moreover, what is the meaning of "*up to"*?

IAS consists of three main parts, latency, upload and download bandwidth. Measuring these three parts exactly will give us the speed of our connection. Download is the downstream direction of the data (from ISP to subscriber). Upload is in the opposite direction. Finally, latency is the time required by a packet to cross the distance between a source and a destination. Latency can be measured easily with *ping* protocol. Latency varies with the physical location of the destination. Local sites have lower latency. Latency is very important to some network applications, such as, VoIP calls and network gaming [2]. Latency can be affected by many factors like congestion.

Measuring the download and upload bandwidths requires special collaboration between a client and a server. The server generates a random file that the client downloads it. Subsequently, the client generates a file and uploads it to the server. The

result is the bandwidth of subscriber's access connection. This bandwidth also varies according to different factors, such as, distance from ISP, congestion, time of day, throttling and server side issues. This is why ISP staff uses "up to" phrase when customers purchase Internet access links. However, are we getting the speed we pay for it?

Figure 1 shows the official US government national broadband map of East Coast [3]. The map has two main colors; dark pink and green. Dark pink means that the advertised connection speed is higher than the actual real speed. Green shows that the actual speed is equal to the purchased one. We can observe that even in US the actual bandwidth is lower than the purchased one.

In this work, we attempt to measure our local Internet access bandwidth to prove that it varies with time. To this end, a web crawler has been implemented to measure the download and upload bandwidths over time. Subsequently, we propose a model based on time series analysis and artificial neural network to predict the bandwidth of an access link over the time. This prediction can be utilized to show the best time to download offline content. Two nonlinear autoregressive with exogenous input (NARX) neural network models have been implemented to predict the upload and download bandwidths of the access link.

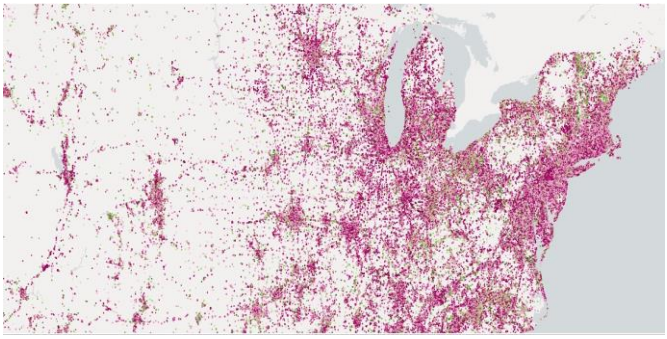

Figure 1: Broadband Map of US East Coast

The rest of this paper is organized as follows; section 2 overviews some of the related works that have been conducted in this area. Section 3 introduces time series analysis and NARX model. Section 4 demonstrated the conducted experiment. Section 5 shows the results. Finally, section 6 concludes this paper.

2 Related Works

Although Internet connections have reached everywhere, users and subscribers still suffer from low-quality connections in third world countries [4-8]. Moreover, a massive technical gap is found in broadband connection among world countries [9]. Nevertheless, Internet subscribers attempt to bear these connections. Researchers and developers attempted to search for methods to tackle this issue. Many methods have been emerged, such as, offline downloading [10]. Two approaches have been proposed in this paradigm; cloud-based and smart access points. In cloud-based approach, companies install massive storage server pools in each Internet service provider (ISP). These servers cache the contents of all ISPs subscribers to download the content from their ISPs in a fast model. This approach has been followed in China, such as Baidu CloudDisk [11] and Tencent Xuanfeng [12]. In [6], a system similar to offline cloud-based download has been implemented and tested. The authors reported a massive performance of the quality of Internet connections. However, this model requires that content providers install servers in each ISP. This is a very hard and complex process when different services from different companies occur around the word.

The second offline downloading approach is smart access points. These access points attempt to download user content before even requested. Subsequently, users can download this content at home LAN speed. Two main smart access points have been developed in China; HiWiFi [13], MiWiFi [14]. However, such a system requires customizing according to users requirements around the world. For example, in Chine a video on demand peer-to-peer system named Fengxing [15] is used. This system can be customized in smart AP. However, in other countries around the word, such as, Middle East, this system is not used. Moreover, sometime customizing not only depends on geographical areas only.

In this work, we attempt to tackle the issue of Internet speed by predicting the best time to download content. This work differs from previous works in two main points. First, we did not proposing a complex or structure enhancement to ISP nor content providers. Second, our system predicts the best time to download content and does not attempt to enhance download bandwidth.

3 NARX Neural Networks

Time series analysis is defined as a collection of methods and procedures that find coherency among events occurred over a period of time. Subsequently, it can predict the occurrence of new events. Time series analysis can be divided into two main paradigms; statistical and intelligent methods.

Statistical methods [16], such as, fractional difference model, Structure model and Bayesian method are easy to understand and implement. However, they are not tractable in complex time series and complex evolvements [17]. On the other hand, intelligent methods [18, 19], such as, NARX, multilayer perceptron's with back propagation and neural networks are better for time series analysis with missing and incomplete data. Moreover, they can model non-linear problems [20]. Finally, they have been utilized in time predication for different issues [21]

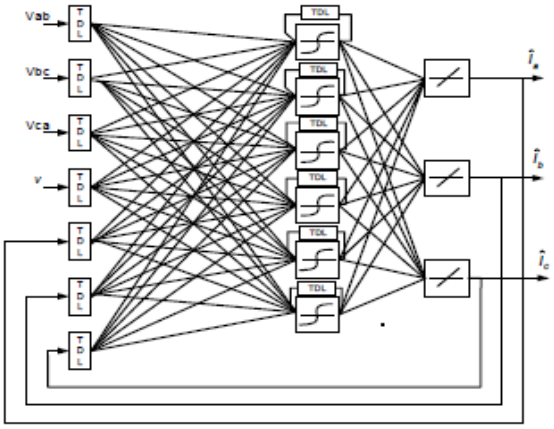

Figure 2: NARX Neural Network

In this work, NARX has been utilized to generate our prediction networks model. NARX or nonlinear autoregressive exogenous model is a nonlinear autoregressive model which has exogenous inputs. Exogenous means that there is one or more feedback inputs to the model. NARX can be implemented utilizing back propagation neural network (BPNN). The use of BPNN model requires the old output values to be fed back to the input. Figure 2 shows a neural network NARX model. One thing to be mention is that back propagation is utilized to optimize the values of network weights. However, the back propagation utilized in NARX is an extended version of the classical algorithm.

4 Experiment

The conducted experiment consists of two parts; data harvesting and prediction model construction. The following sections demonstrate these parts.

4.1 Data harvesting

To collect multi-bandwidth measurements, OOKLA speed test has been utilized [22]. OOKLA is one of the most popular Internet speed test over the globe. OOKLA website is constructed utilizing *flash* technology. To obtain a new data, BEGIN TEST button should be clicked. Subsequently, file downloading and uploading processes should finish before obtaining the results. This process must be automated to collect the readings over time. However, crawling flash webpages is complex. It is harder than static and dynamic webpages.

In static webpages, HTML code can be read to obtain the required information. Moreover, dynamic webpages can be mimicked to execute *JavaScripts* to obtain links to more data. However, *flash* is an application. To harvest these records, *Autohotkey* scripting language [23] has been utilized to mimic the behavior of OOKLA webpage users.

Atypical OOKLA webpage user would first, open a new browser page. Then the link of OOKLA should be called. After that, BEGIN TEST button should be clicked. A delay should be inserted. Finally, copy the results and paste them in a text file. The script mimicked these steps and executes them once every 20 minutes. We started our script on the 24th of June, 2015 and stopped after two weeks. Our Internet connection in this experiment was WiMAX connection with a 2 Mbps of speed. Approximately 990 bandwidth records have been collected. Some of the readings have failed because of issues in our delay time or issues in the webpage.

Finally, a *Python* script has been written to process the generated text file to obtain the download speed, uploaded speed and time records.

4.2 Prediction models construction

The NARX neural network models generated in this work for predicting the upload and the download bandwidth have the same structure and procedure. So, we will demonstrate only one of them. So, let's choose and explore the process of generating the NARX model of the download bandwidth.

To generate the NARX model, the following steps have been conducted. First, the recorded values have been averaged for each hour. In other words, every three values have been averaged to

obtain a single value for each hour. The output of this process is a vector of 330 values. The difference between neighbor hour values has been calculated for two times. Two differences have been used to average the data around zero can be shown in Figure 3. The output of this process has been normalized by dividing the data over the maximum recorded value. Figure 3 shows the normalized output values.

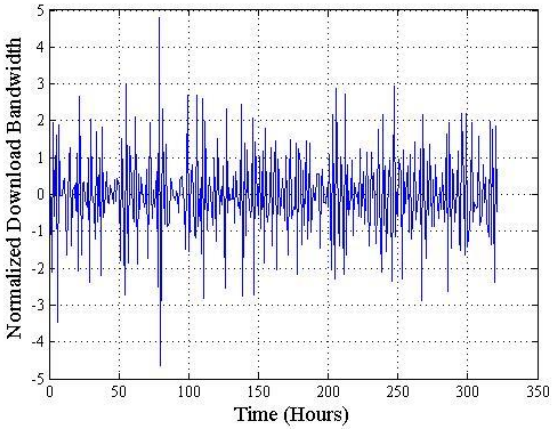

Figure 3: Normalized Download Bandwidth

Secondly, the normalized values should be divided to input and output data to train the model. To divide the normalized data, the correlation distance between values is required. The correlation distance is defined as the number of previous values over time required to predict a new value. In other words, how many time steps the memory of the model requires generating or predicting a new step?

To answer these questions autocorrelation and partial autocorrelation of the collected normalized data should be calculated. Figure 4 and Figure 5 show autocorrelation and partial autocorrelation of the data. According to the analysis in [24] two sparks in partial autocorrelation associated with the pattern in autocorrelation figure demonstrate that 2 old values are enough to produce or predict a new value. This means that the inputs of the model should be two old values and the output is the new value.

According to autocorrelation analysis, the data has been arranged in a matrix with three columns. The first two columns are the input data, the last column is the output data. Moreover, the output of the first raw should be the second input of the second raw. The first input of the second raw is the second input of the first raw. In other words, if the recoded data are $<X_1, X_2, X_3, \ldots\ldots X_n>$, the first raw of the training matrix will be $<X_1,X_2,X_3>$ the second raw $<X_2,X_3,X_4>$ and raw L $<X_L,X_{L+1},X_{L+2}>$. This training matrix has been utilized to train the NARX neural network model, which has the configuration parameters in Table 1. Alg. 1 shows the algorithm utilized to arrange training data.

Algorithm 1: Arranging Training Data

```
n: Total number of collected data
ara[n]: Input data that have been harvested
x[n, 3]: Input data to NARX network, n rows and 3 columns
y[n]: Output data to train the model
Train(x,y): Train NARX model with x input and y output
i=1
FOR i<n  DO
        j=0
        FOR j<3 DO
                x[i,(j+1)]=ara[i+j]
        y[i]=ara[i+j+1]
Train(x,y)
```

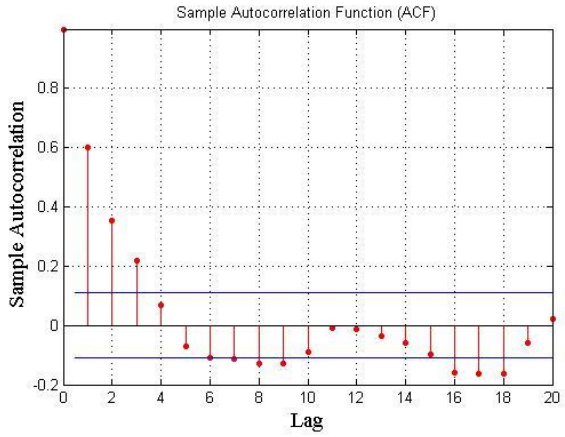

Figure 4: Autocorrelation Function (ACF)

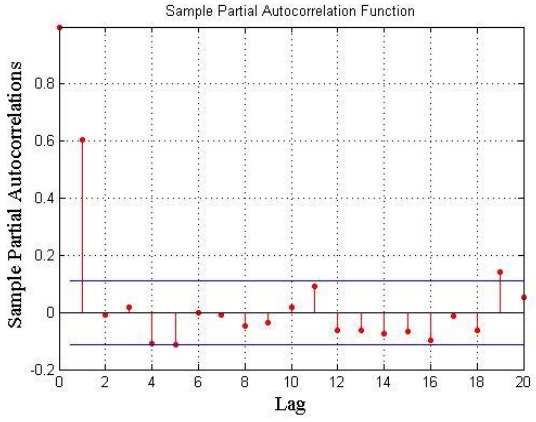

Figure 5: Partial Autocorrelation Function

5 Experiment Results

Experiment results are divided into two sections. The first section shows the quality of sustainability of the download bandwidth. The second section shows the results of the proposed method.

5.1 Internet Connection quality

The recorded download bandwidth values are shown in Figure 6. The figure shows the fluctuation and variation in the download bandwidth. This variation can be translated as the quality of the purchased connection. It can also be noticed from the figure that the time of the day is one important parameter that impact the quality of the connection (maximum download bandwidth is recorded in the morning). The *x* axis is a logged scale axis. The values of *x* axis start at 1 PM every day.

Figure 7 shows the CDF of the recorded download bandwidth values. We can observe from the figure that our Internet connection gave us less than 50% of the purchased connection 50% of the time. Moreover, we can also observe that less than 2% of the time we got the real bandwidth of our connection. In addition, we can observe that the recorded bandwidth can be less than 20% of the real bandwidth.

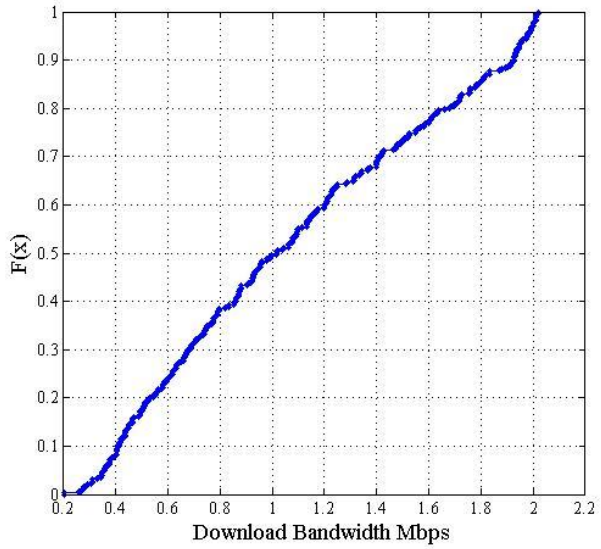

Figure 7: CDF of Download Bandwidth

Table 1: Model Parameters

Parameter	Value
Number of hidden layers	1
Number of neuron in input layer	2
Number of neuron in hidden layer	10

5.2 The Proposed Models

After constructing the models, the Input matrix and the output vectors have been divided into two parts. The first part, we call it the first matrix, has been utilized to train the constructed models. The training matrix has been divided into three parts; 40% for training, 30% for testing and finally, 30% for validation The second part of the data, the second matrix, has been utilized to validate our model. We divided our records to 300 records for training matrix and 20 records for validation.

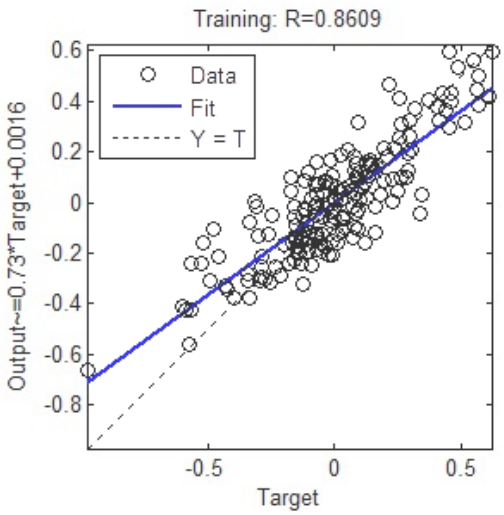

Figure 8: Regression Output

After training the first model- the download bandwidth predictor- we obtained 86% in the validation test. Figure 8 shows the regression value of this model. Subsequently, we utilized the model to obtain the output data from the last 20 inputs.

For the second model - upload bandwidth predictor- we utilized the same previous steps. We obtained a validation of 88%.

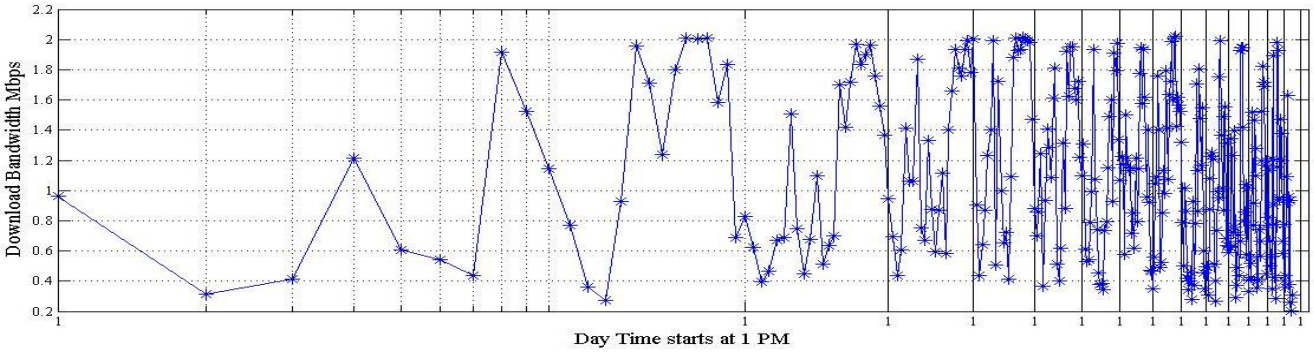

Figure 6: Recorded download bandwidth values

Figure 9 shows the comparison of the real values and the predicted ones.

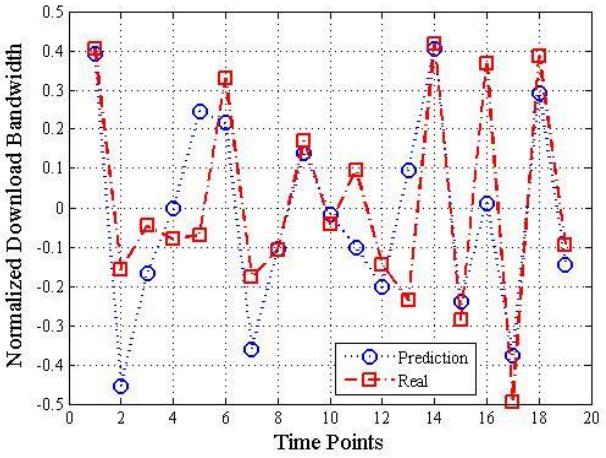

Figure 9: Comparison of predicted and real download bandwidth values

6 Conclusion

Access line speed is the obsession of subscribers and ISPs. In this work, speed test experiment has been conducted to record Internet bandwidth over time. A web-crawler has been constructed to harvest data that is used to train two proposed NARX neural network models. These models are used to predict the bandwidth of access link over time. The predicted values may be utilized by users to make decisions when to download and when to stop. We validate our models and we obtained 86% and 88% values for the download and upload bandwidths respectively.

7 Acknowledgment

This research was supported in part by Al-Zaytoonah University of Jordan fund (2/11/2014). We would like to thank the University for the Equipment and tools they provided for this work.

REFERANCES

[1] W. Stallings, Wireless Communications & Networks. 2004, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[2] Songun Na and Seungwha Yoo. 2002. Allowable Propagation Delay for VoIP Calls of Acceptable Quality. In Proceedings of the First International Workshop on Advanced Internet Services and Applications (AISA '02). Springer London, UK, UK, 47-56.

[3] official US government National Broadband Map east coast http://www.broadbandmap.gov/speedtest

[4] Marshini Chetty, Srikanth Sundaresan, Sachit Muckaden, Nick Feamster, and Enrico Calandro. 2013. Measuring broadband performance in South Africa. In Proceedings of the 4th Annual Symposium on Computing for Development (ACM DEV-4 '13). ACM, New York, NY, USA, , Article 1 , 10 pages.

[5] Fahad, Amal, et al. "An evaluation of web acceleration techniques for the developing world." Presented as part of the 6th USENIX/ACM Workshop on Networked Systems for Developing Regions. 2012.

[6] Johnson, David L., Elizabeth M. Belding, and Consider Mudenda. "Kwaabana: File sharing for

rural networks." Proceedings of the 4th Annual Symposium on Computing for Development. ACM, 2013.

[7] Sarthak Grover, Mi Seon Park, Srikanth Sundaresan, Sam Burnett, Hyojoon Kim, Bharath Ravi, and Nick Feamster. 2013. Peeking behind the NAT: an empirical study of home networks. In Proceedings of the 2013 conference on Internet measurement conference (IMC '13). ACM, New York, NY, USA, 377-390.

[8] Yasir Zaki, Jay Chen, Thomas Pötsch, Talal Ahmad, and Lakshminarayanan Subramanian. 2014. Dissecting Web Latency in Ghana. In Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14). ACM, New York, NY, USA, 241-248.

[9] FCC raises broadband definition to 25Mbps, Chairman mocks ISPs, http://www.extremetech.com/mobile/198583-fcc-raises-broadband-definition-to-25mbps-chairman-mocks-isps

[10] Li, Zhenhua, et al. "Offline downloading in China: A comparative study." Proceedings of the 2015 ACM Conference on Internet Measurement Conference. ACM, 2015.

[11] Baidu CloudDisk offline downloading system, http://pan.baidu.com

[12] Xuanfeng offline downloading system, http://xf.qq.com

[13] HiWiF smart access point, http://www.hiwifi.com

[14] MiWiFi smart AP, http://www.miwifi.com

[15] Fengxing, http://www.fun.tv

[16] Chakhchoukh Y, Panciatici P, Mili L. Electric load forecasting based on statistical robust methods. IEEE T Power Syst 2011; 26: 982-991.

[17] Zhang X, Xing L. The multi-rule & real-time training neural network model for time series forecasting problem. In: International Conference on Machine Learning and Cybernetics; 13-16 August 2006; Dalian, China. New York, NY, USA: IEEE. pp. 3115-3118.

[18] Wang D, Li Y. A novel nonlinear RBF neural network ensemble model for financial time series forecasting. In: Third International Workshop on Advanced Computational Intelligence; 25-27 August 2010; Suzhou, Jiangsu, China. New York, NY, USA: IEEE. pp. 86-90.

[19] Xinxia, L, Anbing Z, Cuimei S, Haifeng W. Filtering and multi-scale RBF prediction model of rainfall based on EMD method. In: 1st International Conference on Information Science and Engineering; 26-28 December 2009; Nanjing, China. New York, NY, USA: IEEE. pp. 3785-3788.

[20] Kalogirou A. Artificial intelligence in renewable energy applications in buildings. In: International Conference on the Integration of the Renewable Energy Systems into the Building Structures; 2005; Patra, Greece. pp. 112-26.

[21] Mohammad Masoud, Yousef Jaradat, Ismael Jannoud "A measurement study of Internet Exchange Points (IXPs): history and future prediction", Turkish Journal of Electrical Engineering and computer science, 2015

[22] Metrics, Ookla Net. "Speedtest (2014)." Akamai Technologies (2013).

[23] Mallet, Chris. "AutoHotKey." (2009).

[24] Fuller, Wayne A. Introduction to statistical time series. Vol. 428. John Wiley & Sons, 2009.