Network Traffic Time Series Performance Analysis using Statistical Methods

Purnawansyah [a, 1], Haviluddin [b, 2, *], Rayner Alfred [c, 3], Achmad Fanany Onnlita Gaffar [d, 4]

[a] *Faculty of Computer Science, Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM5, Makassar 90231, Indonesia*
[b] *Faculty of Computer Sci. and Information Tech, Mulawarman University, Jl. Kuaro no.1, Samarinda 75123, Indonesia*
[c] *Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu 88400, Malaysia*
[d] *Dept. of Information Tech., Samarinda State Polytechnic, Jl. DR. Ciptomangunkusumo, Samarinda 75131, Indonesia*

[1] *purnawansyah@gmail.com; [2] haviluddin@gmail.com*; [3] ralfred121@gmail.com; [4] onnygaffar212@gmail.com*
corresponding author

ABSTRACT

This paper presents an approach for a network traffic characterization by using statistical techniques. These techniques are obtained using the decomposition, winter's exponential smoothing and autoregressive integrated moving average (ARIMA). In this paper, decomposition and winter's exponential smoothing techniques were used additive and multiplicative model. Then, ARIMA based-on Box-Jenkins methodology. The results of ARIMA (1,0,2) was shown the best model that can be used to the internet network traffic forecasting.

This is an open access article under the CC BY-SA license
(https://creativecommons.org/licenses/by-sa/4.0/).

I. Introduction

The remarkable and high accuracy of forecasting result is indeed required to take a decision [1, 2]. In this paper, three statistical models i.e. Decomposition, Winter's exponential smoothing and autoregressive integrated moving average (ARIMA) were used to make forecasting on the use of daily internet traffic. In which, the data traffic constitutes a time series. Furthermore, time series comprises a series of observation pursuant to time. Employed time series, principally, for making forecasting is a data series of $(y_{t+1}, y_{t+2}, ..., y_{t-n})$ in accordance with $(x_{t+1}, x_{t+2}, ..., x_{t-n})$ in particular time range [2-4].

Then, the primary factor influencing forecasting technique determination relies on identification and approach to determine pattern data which basic notation of forecasting Y_t: time series data value during the period of t, $\hat{Y}_t$: forecasting value of Y_t and $et = Y_t - Y_t$: surplus or error in forecasting. Time series comprises of (1) trend (T); data characteristic tend to be high or low, (2) seasonal variation (S); periodical fluctuated data in a year such as monthly, weekly, and daily data, (3) cycles (C); fluctuated data in more than a year, (4) random component (R); data combination from seasonal variation, trends, cycles and random factor are required to be taken into account within forecasting method [5-7].

This present study aims at juxtaposing forecasting result using time series data in accordance with three statistical methods i.e. Decomposition, Winter's exponential smoothing and ARIMA. This paper consists of four different part. The first part deals with the issue on why the authors were intrigued on conducting such a study. Then, it is followed by the second part which exposes several related theories and technique on time series forecasting. The third part presents the results of the study and the fourth part discusses the results and draws a conclusion.

II. Methods

In making forecasting, numerous statistical methods and literature review are available and have been employed by several researchers. These statistical methods have been widely used in financial and demography aspects. In making forecasting, the employment of statistical methods are considerably influenced by time series pattern generated, hence, initial observation in making forecasting shall oversee and analyze the type of the data since every statistical method possesses different working phase [8].

Internet data traffic is characterized as a seasonal variation which fluctuates periodically. Thus, three statistical methods considered as the most applicable to make forecasting are decomposition, Winter's exponential smoothing and ARIMA [1-3]. Below, the three methods employed in this present study is briefly explained

A. Decomposition

Decomposition method comprises of two models which include the additive and multiplicative model. Additive model constitutes by (1).

$$Y_t = Trend + Seasonal + Error \tag{1}$$

While multiplicative model constitutes by (2).

$$Y_t = Trend * Seasonal * Error \tag{2}$$

In which, Y_t observation towards time t.

The basic principle of time series decomposition method is to disintegrate time series data in several patterns and identified those time series segregated, then discovered it separately. After, the data is discovered, integrate the data to make a forecasting. The disintegration of the data is conducted to improve the accuracy of forecasting and attain better time series data attitude comprehension [4, 7].

B. Winter's exponential smoothing

Exponential smoothing method is the procedure for continuous improvement in forecasting the recent object observations. This method provides an average weighted exponential moving at the entire last observed values. Winter's exponential smoothing recognizes three constants as determinants of outcome data forecasting, it is composed of α as a smoothing constant, γ as the trend component and d as δ seasonal component, where the magnitude of a constant between 0 and 1. To generate accurate forecasting, it is determined some combination of values smoothing constant.

Winter's exponential smoothing method for forecasting time series comprises of two models i.e., multiplicative and additive. Multiplicative, principally, contains duplication between the trend component and the seasonal component and it is used when the data in a particular season proportional to the previous season. The formula used (3).

$$y_t = (b_1 + b_2 t) S_t + \varepsilon_t \tag{3}$$

where b_1 is the permanent component; b_2 is linear; trend component; S_t is the multiplicative seasonal factors; ε_t is error component.

While the additive model containing the sum of the trend component with the seasonal component and is used if the difference data reaches a relatively constant in every season, (4).

$$y_t = (b_1 + b_2 t) S_t + \varepsilon_t \tag{4}$$

where b_1 is the permanent component; b_2 is a linear *trend* component; S_t is *additive seasonal* factors; ε_t is the *error* component [1, 2, 4]

C. Autoregressive integrated moving average (ARIMA)

ARIMA method is used to analyse the time series consisting of *autoregressive* (AR) and *moving average* (MA). The methods of ARIMA *(p, d, q) (P, D, Q)* [s] is used with the provisions of the time series which is stationary, where *p* is the process in AR, *d* is the process of *differencing* to convert the data into stationary type, and, finally, *q* is processed on the MA [1, 5].

In general, the time series is considered not to be stationary in the means and variances. If the time series is not stationary, then the transformation process should be carried out in variance and the

Table 1. ACF and PACF identification

Model	ACF	PACF
AR (*p*)	dies down	cut-off after *lag p*
MA (*q*)	cut-off after *lag q*	dies down
ARMA (*p,q*)	dies down	dies down
AR (*p*) or MA (*q*)	cut-off after *lag q*	cut-off after *lag p*

Source: [1]

differencing process is performed in the means. In variance, rules of transformation, namely (1) only for series Z_t are positive, (2) the process of transformation is done before the process of difference, and (3) the value of λ serving as a standard is seen from the Sum of Square Error (SSE) in the transformation process. Normally, the smallest value of SSE variance indicating transformation process has been successfully carried out. Meanwhile, the means, the process of differencing will show specifically the period between data, Table 1.

According to the Box-Jenkins methodology, there are four stages in doing the forecasting using ARIMA model, i.e.; (1) identification of models and patterns; It visually looks the data pattern to be analyzed and check actual data validity, (2) parameters determination; can be done using statistical t-test and p-value, (3) model check (hypothesis testing and diagnostics); testing model that is widely used is the Ljung-Box Q statistic, to check the white noise with the provisions of the p-value> α of 0.05 and Kolmogorov-Smirnov test to check for normal distribution with the provisions of the p-value> α 0:05, and (4) forecasting; the results of ARIMA process will be analyzed in three parts, namely the upper limit, the lower limit should be worth 95%, and forecast values. The finest ARIMA model for forecasting is the model with the smallest error value [1, 2].

D. Dataset Testing

Daily usage of data internet traffic is the main indicator of telecommunication usage in a particular network. Daily usage data internet traffic is used for the network technicians in controlling and managing the use of the network. In this study, daily usage data of internet traffic used is data of daily usage of internet traffic in the network at Mulawarman University taken from the main server using CACTI software. These data were taken in the span of 21 to 24 June 2013. Prior to the forecasting process is done, the original data is normalized to speed up the counting process without eliminating the actual data value [6].The normalization formula as (5).

$$\bar{X} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

In which, $\bar{X}$ is the original data; X_{max} is the maximum data value; X_{min} is the minimum data value. Table 2 presents the original data of daily usage of internet traffic. While Fig. 1 exposes daily usage of internet traffic plot.

E. Determining the finest forecasting model

The selection of the finest time series method is determined by an indicator measuring the accuracy of the data through a specific method of analysis. In the statistical method of determining the finest

Table 2. Original data traffic on 21-24 June 2013

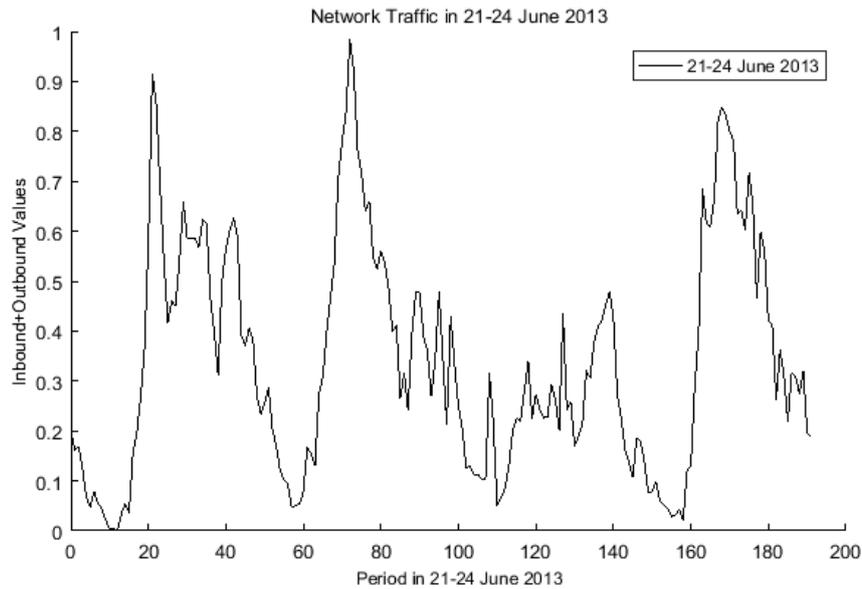| Date		Time	Inbound + Outbound	Date		Time	Inbound + Outbound
6/21	1	00:00	6293000	6/23	97	00:00	10517000
	2	00:30	5185000		98	00:30	6715000
…	…	…	…	…	…	…	…
	48	23:30	11661000		144	23:30	5236000
6/22	49	00:00	8390000	6/24	145	00:00	4528000
…	…	…	…	…	…	…	…
	96	23:30	14530000		192	23:30	5969000

Fig. 1. Daily usage of internet traffic plot

indicator is set to a certain size, among other things, *mean absolute error* (MAE) or *mean absolute deviation* (MAD), mean absolute percentage error (MAPE), *mean square error* (MSE) or *mean square deviation* (MSD), *root mean square error* (RMSE) and *mean percentage error* (MPE). The data test results indicator provided by values such as MAE / MAD, MAPE, MSE / MSD, RMSE, and MPE are the smallest error values. Where the value indicates the error value of testing a method. Therefore, the determination of the finest model is performed by selecting the smallest error value. Hence, the forecasting result having the smallest value is the finest model since it will give the test results closer to the actual value data [7-11].

In this study, the method of measuring the accuracy of forecasting is using MAPE, MAD, MSD. Where each method has a formula. First, MAPE formula is as follows, (6).

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{(Y_t - Y_t')}{Y_t} \tag{6}$$

Second, the formula of MAD, (7).

$$MAD = \sum \frac{(Y_t - Y_t')}{n} \tag{7}$$

Third, the formula of MSD, (8).

$$MSD = \sum_{t=1}^{n} \left| \frac{(Y_t - \hat{Y}_t)}{n} \right|^2 \tag{8}$$

where, Y_t observation value; Y_t' forecasting value; and n the amount of observation.

This present study deals with a comparative study on the result of predetermined statistical model testing; decomposition, Winter's exponential smoothing and ARIMA. The following Fig. 2 illustrates the flow of undertaken study.

III. Results and Discussion

In this study, the observations were contrived to the daily usage of Internet traffic *(*inbound and outbound*)* at a state university. The data was collected for forecasting in June 2013 for 4 days (21-24 June 2013) amounting to 192 data samples. Further, the data were analysed and observed using a predetermined statistical method includes decomposition, Winter's exponential smoothing, and ARIMA. The aforementioned methods were determined and undertaken due to seasonal variation of daily usage of internet traffic. SPSS 19 and Minitab 16 were utilized to assist the data analysis.
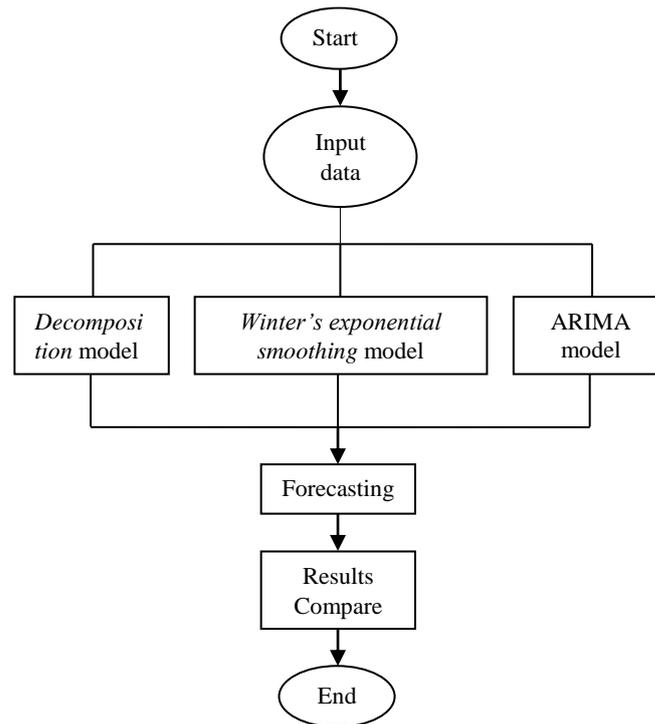
Fig. 2. Research flow

A. Decomposition Analysis

The first stage undertaken is to test the data using decomposition model. In this study, decomposition models used consisted of two models of decomposition includes additive and multiplicative decomposition. Simultaneously, the process of *network traffic* analysis was done by dividing the *dataset* into two parts, namely the data *inbound* and *outbound*. The data were analysed separately. Then, the analysis results are re-consolidated. The fairly-decent error rate of forecasting was obtained by *decomposition additive* which MAPE is worth 4.69E + 01, MAD is worth 1.65E-01, and MSD is worth 4.02E-02.

B. Winter Exponential Smoothing Additive Analysis

The second stage is to test the data using Winter's exponential smoothing additive models. In this study, Winter's exponential smoothing consisted of additive and multiplicative models were used. The process of analysis is done identically to the decomposition model. In this study, respectively trend and smoothing of the data set are worth 0.2 and 0-1 to get a satisfying forecasting accuracy. The fairly-decent error rate of forecasting was obtained by Winter's exponential smoothing additive which MAPE is worth 2.35E + 01, MAD worth 1.89E + 06, and MSD is worth 2.58E + 06.

C. ARIMA Analysis

The last stage is to test data using ARIMA model. The data testing within ARIMA phase was done by stationer processes, thus the data converting into variance (transformation) and means (differencing) to obtain the ARIMA model (1,0,0), (1,1,0), (1,1,1), (1,0,1) and (1,2,1). After checking the model (hypothesis testing and diagnostics) with the test model Ljung-Box Q statistic, to check the white noise with the provisions of the p-value> α of 0.05 and continued with the Kolmogorov-Smirnov to check the normal distribution with the provisions of the p-value> α 0:05. Then the ARIMA model (1,0,2) has qualified which upper limit, the lower limit is worth 95%. A fair forecasting error rate is obtained with ARIMA Where MAPE is worth 2.78E + 01, MAD is worth 2.54E + 06, and MSD is worth 1.89E + 06. The results of the forecasting are illustrated in Fig. 2, 3, and 4 and the comparison of MAPE, MAD, and MSD are exposed in Table 3.

Table 3. Comparison of three predetermined statistical model analysis

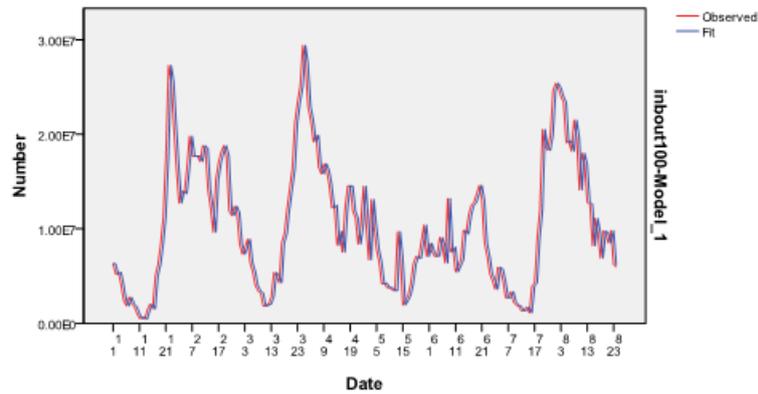| Models	MAPE	MAD	MSD
Decomposition			
• Additive model	4.69E+01	1.65E-01	4.02E-02
• Multiplicative model	5.69E+01	1.66E-01	4.03E-02
Winter's exponential smoothing			
• Additive model	2.35E+01	1.89E+06	2.58E+06
• Multiplicative model	2.39E+01	2.05E+06	3.07E+06
ARIMA (102)	2.78E+01	2.54E+06	1.89E+06

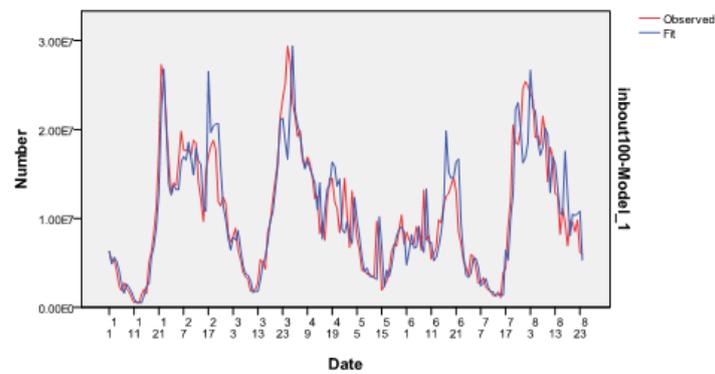Fig. 3. Winter's exponential smoothing additive Plot

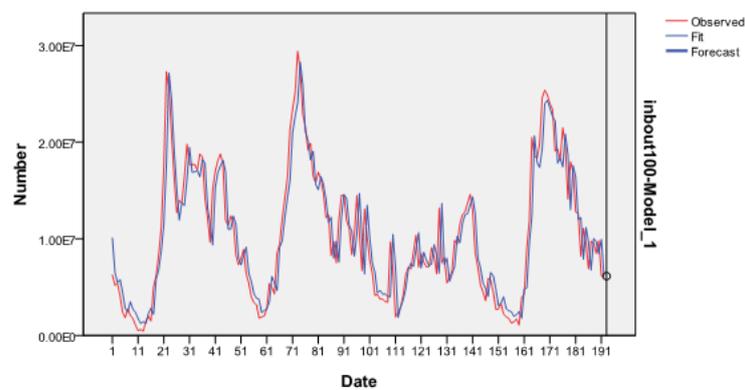Fig. 4. Winter's exponential smoothing additive Plot

Fig. 5. ARIMA (1,0,2) plot

IV. Conclusion

This present study utilizing statistical methods Decomposition, Winter's exponential smoothing and ARIMA to forecast the usage of Internet traffic on Mulawarman University. In order to identify the results of forecasting, the three predetermined models, MAPE, MAD, and MSD were employed. The test results of the three methods confirm that ARIMA model (1,0,2) has a fair forecasting error rate which is calculated with the smallest value of MSD is 1.89E + 06. This indicates that the accuracy of the ARIMA forecasting accuracy approaches the actual data. However, the ARIMA model cannot accommodate the increase or decrease of internet users' frequencies. In addition, if the data sample is large then the forecasting result will be constant. Along with the widespread development of computational intelligence, then the future undertaken study will employ forecasting using one of the machine learning methods that are considered in accordance with the seasonal variation time series.

References

[1] Box, G.E.P., G.M. Jenkins, and G.C. Reinsel, *Time Series Analysis Forecasting and Control Fourth Edition*. 2008, Copyright © 2008 by John Wiley & Sons, Inc. All rights reserved.

[2] Wei, W.W.S., *Time Series Analysis Univariate and Multivariate Methods Second Edition*. 2006, Pearson Education, Inc. All rights reserved.

[3] Santos, A.C.F., et al., *Network traffic characterization based on Time Series Analysis and Computational Intelligence.* Journal of Computational Interdisciplinary Sciences, 2011. **2**(3): p. pp. 197-205.

[4] Brockwell, P.J. and R.A. Davis, *Introduction to Time Series and Forecasting Second Edition*, G. Casella, Editor. 2002, © 2002, 1996 Springer-Verlag New York, Inc.

[5] Li, C. and T.-W. Chiang, *Complex Neurofuzzy ARIMA Forecasting—A New Approach Using Complex Fuzzy Sets.* IEEE Transactions on Fuzzy Systems, 2013. **21**(NO. 3, JUNE 2013).

[6] Bernacki, J. and G. Kołaczek, *Anomaly Detection in Network Traffic Using Selected Methods of Time Series Analysis.* I. J. Computer Network and Information Security, 2015. **9**: p. 10-18.

[7] Sermpinis, G., et al., *Forecasting and trading the EUR/USD exchange rate with stochastic Neural Network combination and time-varying leverage.* Decision Support Systems, 2012. **54, (2012)**: p. 316–329.

[8] Khashei, M. and M. Bijari, *An artificial neural network (p, d,q) model for timeseries forecasting.* Expert Systems with Applications, 2010. **37 (2010)**: p. 479–489.

[9] Khashei, M. and M. Bijari, *A new class of hybrid models for time series forecasting.* Expert Systems with Applications, 2012. **39**(2012): p. 4344–4357.

[10] Gomes, G.S.d.S. and T.B. Ludermir, *Optimization of the weights and asymmetric activation function family of neural network for time series forecasting.* Expert Systems with Applications, 2013. **40**(2013): p. 6438–6446.

[11] Haviluddin and R. Alfred, *Forecasting Network Activities Using ARIMA Method.* Journal of Advances in Computer Networks (JACN), 2014. **2, (3) September 2014**: p. 173-179.