Forecasting Network Activities Using ARIMA Method

Haviluddin and Rayner Alfred

***Abstract*—This paper presents an approach for a network traffic characterization by using an ARIMA (*Autoregressive Integrated Moving Average*) technique. The dataset used in this study is obtained from the internet network traffic activities of the Mulawarman University for a period of a week. The results are obtained using the Box-Jenkins Methodology. The Box-Jenkins methodology consists of five ARIMA models which include ARIMA (2, 1, 1) (1, 1, 1)^{12}, ARIMA (1, 1, 1) (1, 1, 1)^{12}, ARIMA (2, 1, 0) (1, 1, 1)^{12}, ARIMA (0, 1, 0) (1, 1, 1)^{12}, and ARIMA (0, 1, 0) (1, 2, 1)^{12}. In this paper, ARIMA (0, 1, 0) (1, 2, 1)^{12} was selected as the best model that can be used to model the internet network traffic.**

***Index Terms*—Network traffic, ARIMA, time series, forecasting.**

I. Introduction

Today, internet plays an important role in any organizations. Particularly, internet is used in most universities to support the teaching and learning activities. These activities are part of *civitas-academica*, and thus the internet infrastructure must be well organized in order to optimize the usage of internet. In this sense, Information Technology department plays an important role to ensure a high quality internet service is accessible to all university staff.

As part of the organization's performance, a good planning should be outlined in order to provide an acceptable internet service. Furthermore, internet performance services are normally measured based on the capacity of the internet network and the traffic requirements based on the university activities. As a result, forecasting the demand of the internet service used to support teaching and learning activities is very crucial in order to ensure the effectiveness and efficiency of the teaching and learning processes. In this paper, the characterization of the network activities is performed in order to visualize the internet network traffic requirements [1]-[3]. The knowledge about the network traffic requirements can be used to plan for a better infrastructure development in order to provide high-quality services and design optimization. The modeling or forecasting methodology is one of the best phenomenological used to solve problems related to construction planning and evaluation of internet services provider in the future.

The purpose of this paper is to model and forecast the internet network traffic activities by using a time series analysis. The organization of this paper is arranged as follows. Section II discusses the theoretical basis relevant to the work and the techniques used to perform time series analysis. Section III presents the experimental design and results obtained from the analyzed series, and Section IV concludes this paper with some recommendations on future works.

Manuscript received December 5, 2013; revised March 19, 2014.

Haviluddin is with the School of Natural Science, Department of Computer Science, Universitas Mulawarman, 75119 Samarinda, Indonesia (e-mail: haviluddin@gmail.com).

Rayner Alfred is with the COESA, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia (e-mail: ralfred@ums.edu.my).

II. Network Traffic Data and ARIMA (Autoregressive Integrated Moving Average)

A forecasting method that is used to model data that continues to evolve is called a time series forecasting method. This method performs the forecasting process by analyzing the relationship between the variables to be estimated and the time variable. The time series model assumes that some patterns or combination of patterns will be repeated all the time. Thus, by identifying and extrapolating these patterns, it can be predicted. The time series forecasting method focuses on the type or pattern of data. There are four kinds of time series data patterns, namely;

1) trend (*T*); patterns is a trend toward data in the long term can be either an increase or a decrease,
2) seasonal variation (*S*); fluctuations of the data that is periodically occurred within one year, such as monthly, weekly, or daily,
3) cycles (*C*); fluctuations of the data for a period of more than one year, and
4) random component (*R*); time series data that are influenced by seasonal variation, trends, cycles and random factors [1], [3], [4], Fig. 1a- Fig. 1d.

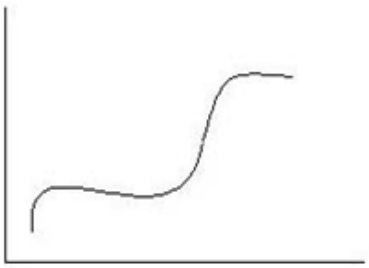

Fig. 1a. Illustrates the types of patterns that are normally identified in a time series data patterns *Trend*.

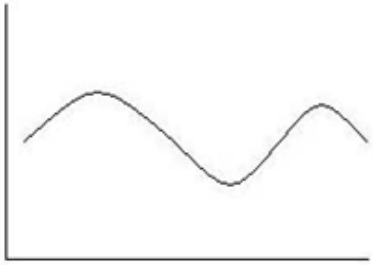

Fig. 1b. Illustrates the types of patterns that are normally identified in a time series data patterns *Seasonal Variation*.

DOI: 10.7763/JACN.2014.V2.106

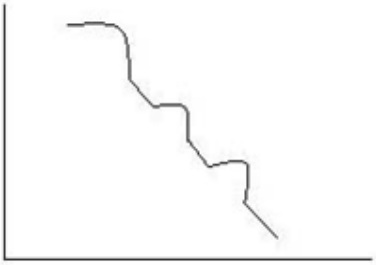

Fig. 1c. Illustrates the types of patterns that are normally identified in a time series data patterns *Cycles*.

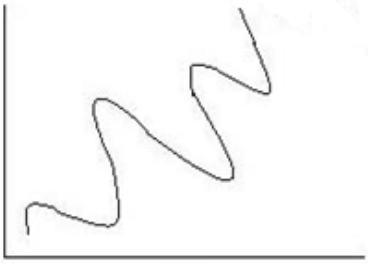

Fig. 1d. Illustrates the types of patterns that are normally identified in a time series data patterns *Random Component*.

A. Network Traffic Data

According to [3], network traffic can be defined as a large network packet datasets, which are generated during the data communication over time. It can be expressed that network traffic pattern is seasonal variation where the pattern has developed significant tend at specific seasonal period. Therefore, network traffic datasets can be analyzed as a time series [2], [3], [5].

The manageable switch or core switch is a central of network traffic. Thus, the core switch network can be customized, such as configuring switches to become a VLAN (*virtual local area network*) that is able to send the packets very quickly. Its main function is to determine *access core*, *routing*, and *filtering*.

In this research, each network traffic data was captured by the CACTI software. The network traffic has two graphic areas, indicated by green and blue colors. The green color indicates the *inbound* traffic and the blue color indicates the *outbound* traffic. The *Inbound* traffic shows data that comes from outside (e.g., a computer) into the network. On the other hand, the *outbound* traffic shows data that goes out from the network. Fig. 2 and Fig. 3 indicate the network traffic plot. The internet network traffic is a seasonal time series and in this paper, the ARIMA of order (p, d, q) $(P, D, Q)^s$ will be used to model the network traffic activities.

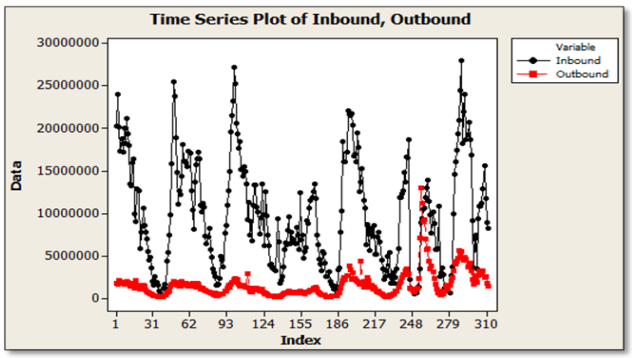

Fig. 2. Network traffic data, capture by CACTI software.

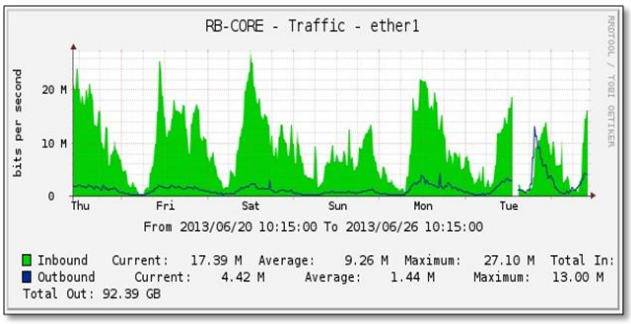

Fig. 3. Network traffic plot.

B. ARIMA

A time series is an ordered sequence of observations and many ways are used to forecast the time series data. In principle, time series model is used to predict the values of data (y_t +1, y+2, ..., y_{t-n}) based on the data (x_t +1, x_t+2, ..., x_{t-n}). In learning a time series data, one may use a method that is based on mathematical or machine learning concepts. A time series method based on a mathematical model is one of the oldest models used in forecasting a time series data. However, this method has some difficulties when applied to a linear or non-linear model, because in fact, a time series data is rarely a linear or non-linear and often contain both. However, this model still has a very good level of accuracy for short-term forecasting with the terms time series data is non-stationary, at the linear moment [6]. On the other hand, the machine learning models can be used to perform a time series forecasting for data which is non-linear, because this model ignores stationary data. The basic principle of this method is based on real life, where a lot of problems with the data containing linearity and non-linearity simultaneously [4]-[6].

One of the famous methods used in forecasting a time series data is ARIMA. The ARIMA method is used to analyze a time series data in which it is designed by integrating the AR (*autoregressive*) and MA (*moving average*) methods. The ARIMA (p, d, q) is a general method that is formulated with respect to the data series that are stationary only, where, p is the number of processes in AR, d is the number of differencing a time series of data to be stationary, and finally, q is the number of processes in MA [6].

In this section, the ARIMA model will be briefly described. In addition, time series condition using ARIMA is series observation must be stationary. The series observation stationer expressed when the processes are not changing along with time changing, meaning that the series observation average and the time throughout are always constant.

In general, a time series that is not stationary has two factors that need to be considered; *means* and *varians*. If the time series is not stationary, the next processes are transformation which is performed in *varians* and differencing which is performed in *means*. In *varians*, the rule for the transformation of the time series model consists of (1) For series only, Z_t are positive, (2) Transformation before *differencing* process, and (3) λ value is chosen based on the *Sum of Square Error* (SSE) obtained from the results of *series* transformation. Usually, the smallest SSE value has a constant *varians*. Meanwhile, in *means*, the differencing process is performed between a specific period of data with another period of data.

TABLE I: 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:[4]

Autocorrelation is a correlation that exists between the observations of a time series separated by k time units. Thus, the autocorrelation (ACF) data is plotted based on the lag k time units itself. Meanwhile, PACF are correlations that exist between sets of ordered data pairs of a time series. However, the values of p, d and q can be determined based on ACF and PACF, as shown Table I.

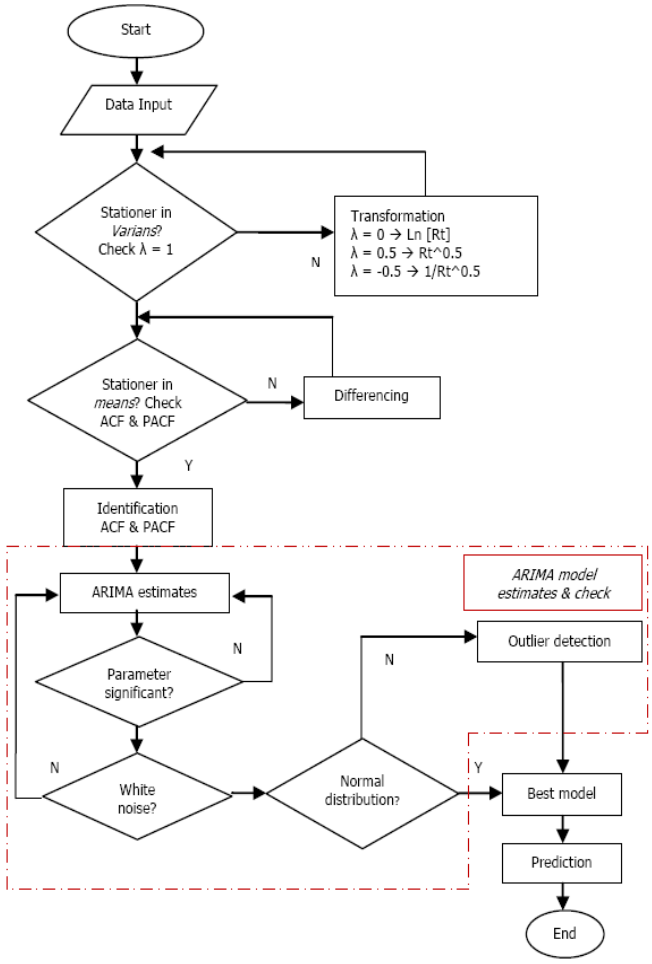

Fig. 4. ARIMA (box-jenkins) of stages.

According to the Box-Jenkins [6] methodology, there are four forecasting stages, that includes; (1) identification model, (2) parameter estimation, (3) model checking (hypothesis and diagnostic test), and (4) forecasting, as shown in Fig. 4.

Stage 1: Model identification

In any kind of time series analysis, the first step is to plot the data. The data series will be carefully examined in order to determine whether the series contains a trend, seasonality, cycles or random phenomena. After that, the sample ACF and PACF of the original series are computed and examined in order to further confirm that the time series data is stationary. If the sample ACF decays very slowly, it indicates that *differencing* processes is needed.

The risk of producing incorrect model identification will be incorrect model estimation produced and model re-identification will be needed.

Stage 2: Parameter estimation

Once the ARIMA parameters estimations are obtained, the model identification can be validated. The purpose of model validation is to ensure that the right model is used. This can be done by using the t -statistics and p -value.

Stage 3: Model checking (hypothesis and diagnostic test)

The proposed model needs to be checked for its competency before it can be used for forecasting. Many model tests in statistics can be used. One of the famous models test is based-on the Ljung-Box Q statistic. This is tested for *white noise* checked with $p\text{-}value > \alpha\ 0.05$ and Kolmogorov Smirnov tested for normal distribution with $p\text{-}value > \alpha\ 0.05$ condition. Nevertheless, $p\text{-}value < \alpha\ 0.05$ or null in the ARIMA estimation should be used, but a few of normal distribution data does not spread in line.

Stage 4: Forecasting

The last step in the Box-Jenkins method is forecasting. The results of the ARIMA forecasting processes can be analyzed in three different options for upper limit, lower limit, and forecast values. The upper and lower limits provide 95% confidence interval. In other words, that forecast values in the confidence limit will be accepted.

III. RESULTS AND DISCUSSION

The results of the network traffic inbound event analysis that has taken for three days (June 20-23, 2013) with samples of 170 data, Fig. 5. The first ARIMA steps is transformation process with *Box-Cox* with twice process, thus stationer data in *varians* λ = 1 is obtained, Fig. 6.

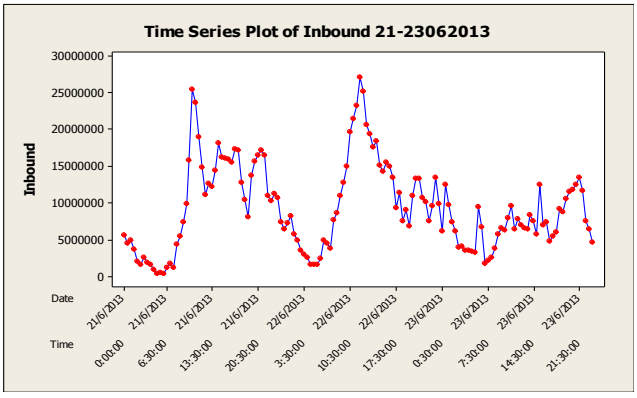

Fig. 5. Network traffic *Inbound* plots in three day June 20 – 23, 2013.

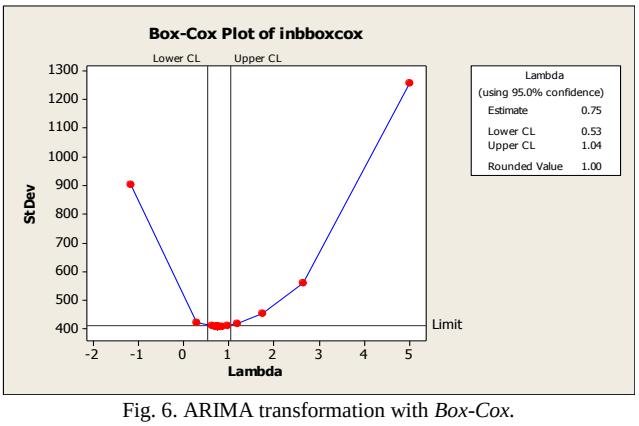

Fig. 6. ARIMA transformation with *Box-Cox*.

The second step is to perform the autocorrelation checking by observing the ACF and PACF plots with $\lambda > 0.05$. Since the data is still not stationary in *means*, both non-seasonal and seasonal differencing processes are performed. For a non-seasonal, a *differencing* process with lag 1 is performed based on the result obtained from the transformation *Box-Cox*. Meanwhile, a seasonal *differencing* process with lag 12 is performed as shown in Fig. 7 and Fig. 8.

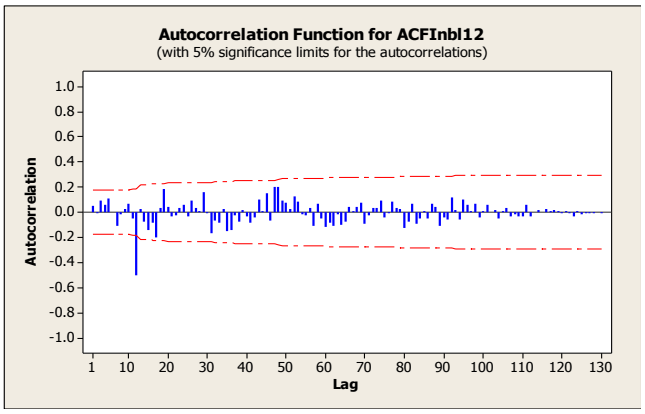

Fig. 7. ACF plot data $\lambda = 1$

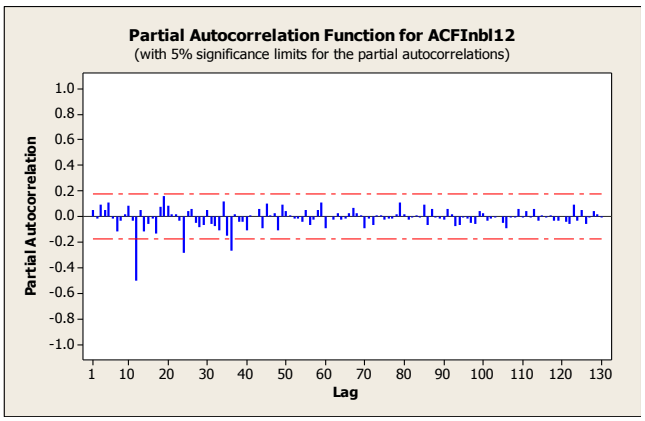

Fig. 8. PACF plot data $\lambda = 1$

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
SAR  12    -0.6373   0.0782  -8.15  0.000
SMA  12     0.8725   0.0798  10.93  0.000
Constant    -2.155    8.164  -0.26  0.792


Differencing: 1 regular, 2 seasonal of order 12
Number of observations:  Original series 144,
after differencing 119
Residuals:     SS =  37300144 (backforecasts
excluded)
               MS =  321553  DF = 116


Modified Box-Pierce (Ljung-Box) Chi-Square
statistic

Lag             12     24     36     48
Chi-Square     9.2   47.6   51.5   66.9
DF               9     21     33     45
P-Value      0.418  0.001  0.021  0.019
```

Fig. 9. Final estimates of ARIMA $(0,1,0)(0,2,1)^{12}$ parameters.

The third step is called the ARIMA estimation and model checking processes. The results of ARIMA $(1, 1, 1)^{12}$ are estimated, in which it may consist of five models; ARIMA $(2, 1, 1)\ (1, 1, 1)^{12}$, ARIMA $(1, 1, 1)\ (1, 1, 1)^{12}$, ARIMA $(2, 1, 0)\ (1, 1, 1)^{12}$, ARIMA $(0, 1, 0)\ (1, 1, 1)^{12}$, and ARIMA $(0, 1, 0)\ (1, 2, 1)^{12}$. Afterward, the most significant model is identified and tested by using significant parameters in *white noise* test. The test results determine a good model for analysis and forecasting. The best result in this test is obtained from the ARIMA $(0, 1, 0)\ (1, 2, 1)^{12}$, as shown in Fig. 9. Afterward, the model normality is checked by using *p-value* > *α 0.05* as shown in Fig. 10.

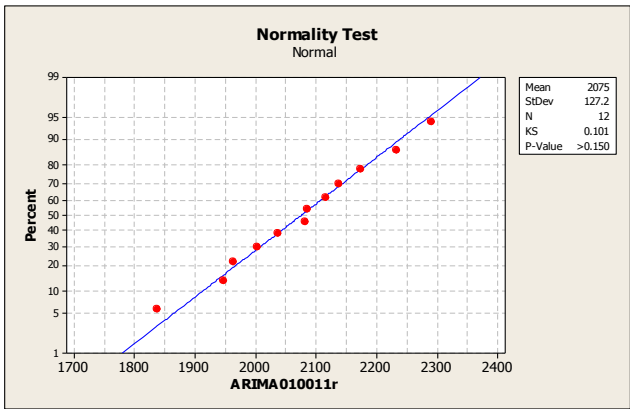

Fig. 10. Normality test parameters.

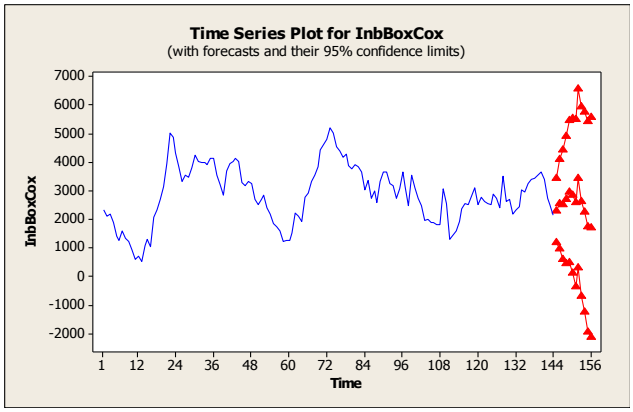

Fig. 11. Forecasting plots.

The fourth step is forecasting. From the best result, internet network traffic forecasting plot in Fig. 11.

IV. Conclusion and Future Research

In this paper, the ARIMA method is applied to forecast the internet network traffic usage. The best result is obtained from the ARIMA $(0, 1, 0)\ (0, 2, 1)^{12}$ with normality test of normal and parameters are estimated with *p-value* < *α 0.05*. However, the results obtained shows that there exists a weak relationship residual value because of *p-value* > 0.019. However, the ARIMA estimation is quite reliable for forecasting because the iteration expression convergence relative change in each estimate is less than 0.0010.

For future works, in order to improve the time series forecasting model, combining methods in machine learning and statistical concepts will be the best option since a time series data consists of both linear and non-linear data [1], [5]-[9] with the assumptions:

$$Z_t = L_t + N_t \quad (1)$$

where:

Z_t= time series
L_t= linier component model
N_t= non-linier component model

REFERENCES

[1] T. C. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 164-181, 2011.

[2] E. J. Palomo, J. North, D. Elizondo, R. M. Luque, and T. Watson, "2012 special issue: application of growing hierarchical SOM for visualisation of network forensics traffic data," *Neural Networks*, vol. 32, pp. 275–284, 2012.

[3] A. C. F. Santos, J. D. S. D. Silva, L. D. S. Silva, and M. P. D. C. Sene, "Network traffic characterization based on Time Series Analysis and computational intelligence," *Journal of Computational Interdisciplinary Sciences*, vol. 2, pp. 197-205, 2011.

[4] W. W. S. Wei, *Time Series Analysis Univariate and Multivariate Methods Second Edition*, 2006.

[5] C. Li and T. W. Chiang, "Complex neurofuzzy ARIMA forecasting—a new approach using complex fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 21, 2013.

[6] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis Forecasting and Control*, 4th ed., 2008.

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

[8] W. Qiu, X. Liu, and H. Li, "A generalized method for forecasting based on fuzzy time series," *Expert Systems with Applications*, vol. 38, pp. 10446–10453, 2011.

[9] G. S. D. S. Gomes and T. B. Ludermir, "Optimization of the weights and asymmetric activation function family of neural network for time series forecasting," *Expert Systems with Applications*, vol. 40, pp. 6438–6446, 2013.

Haviluddin was born in Loa Tebu, East Kalimantan, Indonesia, on May 28th, 1973. He graduated from STMIK WCD Samarinda in 2005 in the field of management information, and he completed a master at Gadjah Mada University, Yogyakarta in 2009 in the field of computer science. He is also a lecturer in the Faculty of Natural Science, University Mulawarman, East Kalimantan, Indonesia. Currently, he is pursuing his PhD in the fieald of computer science at the School of Engineering and Information Technology, Universiti Malaysia Sabah, Malaysia.

Rayner Alfred was born in Kota Kinabalu, Sabah. He completed a PhD in 2008 looking at intelligent techniques to model and optimize the complex, dynamic and distributed processes of knowledge discovery for structured and unstructured data. He holds a PhD degree in computer science from York University (United Kingdom), a master degree in Computer Science from Western Michigan University, Kalamazoo (USA) and a computer science degree from Polytechnic University of Brooklyn, New York (USA).

Dr. Rayner leads and defines projects around knowledge discovery and information retrieval at Universiti Malaysia Sabah. One focus of Dr. Rayner's work is to build smarter mechanism that enables knowledge discovery in relational databases. His work addresses the challenges related to big data problem: How can we create and apply smarter collaborative knowledge discovery technologies that cope with the big data problem.

Dr. Rayner has authored and co-authored more than 75 journals/book chapters and conference papers, editorials, and served on the program and organizing committees of numerous national and international conferences and workshops. He is a member of the Institute of Electrical and Electronic Engineers (IEEE) and Association for Computing Machinery (ACM) societies.