Journal of Theoretical and Applied Information Technology
15th February 2017. Vol.95. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SARIMA-EGARCH MODEL TO REDUCE HETEROSCEDASTICITY EFFECTS IN NETWORK TRAFFIC FORECASTING

[1]INDRA HIDAYATULLOH, [2]ISNA ALFI BUSTONI

[1]Department of Informatics, ST3 Telkom Purwokerto, Indonesia

[2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science,

Universitas Gadjah Mada, Indonesia

E-mail: [1]indra@st3telkom.ac.id, [2]isna@ugm.ac.id

ABSTRACT

Difference needs in bandwidth allocations have not been accommodated by static bandwidth allocations that leads to ineffective bandwidth use. There are several previous researches about bandwidth allocations which have been conducted, such as the use of Seasonal Autoregressive Integrated Moving Average (SARIMA) method. However, SARIMA method is not able to overcome various kinds of error problems or heteroscedasticity. Therefore, this research proposes the application of SARIMA-EGARCH (Exponential Generalized Autoregressive Conditional Heteroscedastic) method to generate the more accurate model that is able to overcome heteroscedasticity on the needs of bandwidth forecasting. In addition, this research compares the result of SARIMA to SARIMA-EGARCH examinations. It shows that SARIMA $(1,0,1)(3,1,1)_7$ has 11,38% Mean Absolute Percentage Error (MAPE) and SARIMA-EGARCH $(1,0,1)(3,1,1)_7(1,1)$ has only 9,20%. The comparison shows that applying EGARCH increase the accuracy to 19,15%.

Keywords: *Forecasting, Bandwidth, Heteroscedasticity, SARIMA, EGARCH*

1. INTRODUCTION

Internet has been part of individual and company basic needs to support any kind of business process. It should not only be reliable, but also able to provide adequate access to the users. Forecasting bandwidth, a common method to predict bandwidth allocation therefore bandwidth allocation became more efficient and near to the actual needs.

Many researchers focused on predicting bandwidth using classic statistical technique, one of them is Seasonal Autoregressive Moving Average (SARIMA).SARIMA can be used to forecast time series data contained seasonal effect.

Dong Peng et al [1] analyzed Hadoop backbone data characteristic. In this paper, SARIMA model was used to find the best daily data trace for data characteristic analysis. The result showed daily data trace can be used to determine monthly data trace because the characteristic nearly the same. Unfortunately, this paper research did not calculate the error of SARIMA model when it comes to monthly data.

Another study that applies SARIMA method, analyzes sequences of times with season pattern, generates the daily bandwidth needs which are more flexible and closed to the actual needs [2]. The study shows that there were outliers on monthly data especially during holidays. By applying the model of SARIMA $(0,1,1)(0,1,1)_7C$ with the addition of outlier detection generate 14% of Mean Absolute Percentage Error (MAPE). Even though, using outlier detection on data training can effect prediction because of data loss during replacing or removing outlier.

Dandan Miao[3], in his paper found that Multiplicative SARIMA can be used to model the traffic of monthly data trace, but for daily and weekly the accuracy decreasing because SARIMA treats daily and weekly data as same as monthly data. The study also shown that holidays effect the accuracy of prediction. More advanced studies [4][5][6] also using SARIMA as the main method to generate time series model. But, it still not yet treat the holidays effect on monthly data.

The focus of the modeling and forecasting bandwidth use in those previous researches is to find the right model that represents seasonal data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

traffics. The seasonal pattern appears due to the low or high use of the bandwidth in particular times. For example, the numbers of users on Sunday is differed from the number of users on Wednesday leads to distinguished bandwidth allocations. The conditions lead the heteroscedasticity, or a situation where there is different error in each forecasting result occurs and causes deviation [7] which makes the forecasting model not represent the actual data.

Engle (1982) proposed Autoregressive Conditional Heteroscedastic (ARCH) to overcome the situation. The method was then developed into Generalized Autoregressive Conditional Heteroscedastic (GARCH) in 1986 as it as proposed by Bollersley. In 1991, GARCH was developed into Exponential GARCH (EGARCH) [8].

GARCH method has been applied in the previous studies, such as the application of GARCH model to predict the accurate stock price [9] in which the model (2,2) generated <5% MAPE. In [10] GARCH and EGARCH methods are compared in its application to the property field in the world monetary crises. EGARCH models show a better performance with lower MAPE value and it coped with the asymmetric influences. Empiric study to Buy-Back Rates measured structures with EGARCH modeling [11] shows that EGARCH model generates better results than GARCH model. Even though, GARCH and EGARCH, was not able to consider seasonal effect on time series data.

Therefore, based on the previous studies, this research aims the application of SARIMA-EGARCH method to obtain the right model in Wi-Fi network traffic forecasting in the Department of Computer Science and Electronics, Universitas Gadjah Mada (UGM). The combination of SARIMA-EGARCH method supposed to reduce heteroscedacity effect while seasonal effect applied on network traffic data.

2. DATA

The data is inbound bandwidth data in UGM-Hotspot network at Department of Computer Science and Electronics. The Figure 1 displays time series plot of inbound data that has 150 data in which one datum represents a day.

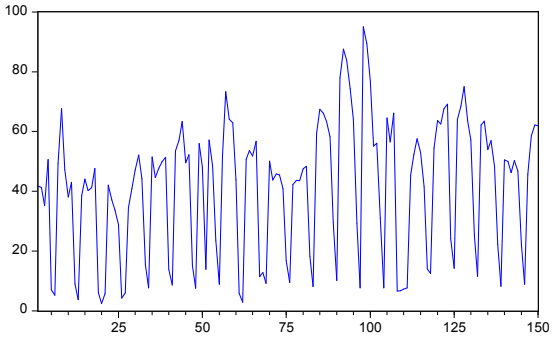

Figure 1: Time Series Plot of Inbound Data

Time series plot of inbound data indicates up-and-down peak trends following trend in a certain period. It indicates that there is seasonal trend so that SARIMA method will be done to the trend of the data.

3. RESEARCH METHODOLOGY

3.1 Seasonal ARIMA (SARIMA)

Seasonal ARIMA (Autoregressive Integrated Moving Average) method is the development of ARIMA method by adding seasonal effects.

ARIMA model describe systematic form of time series using 3 parameters as follows:

1) p : autoregressive (AR-term) order
2) d : differencing order of stationer time series
3) q : moving average (MA-term) order

with equation as follows:

$$\phi p\ (B)(1-B)d\ Zt\ =\ \delta + \Theta q(B)\in t \qquad (1)$$

where B is backshift operator. This equation will be used on Seasonal ARIMA(SARIMA). SARIMA consist of two part: non-seasonal (regular part) and seasonal part. Therefore, SARIMA can be written as follows:

$$ARIMA(p,d,q)(P,D,Q) - model \qquad (2)$$

ARIMA method itself is a Box-Jenkins method that has the following steps [12].

1) Identification

Identification step is conducted to determine the order of all models that maybe used by looking at correlogram of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2) Parameter Estimation

This step is conducted to determine the parameter in SARIMA (p, d, q)(P, D, Q) which is most significantly used.

3) Diagnostic Checking

This step is conducted to examine the properness of used model so that the best model that mostly represents the data is obtained.

4) Forecasting

The last step is completed by using selected model from the previous step.

3.2 EGARCH

ARCH is a method proposed by Engle to reduce heteroscedasticity. There are two models, mean model and variance model. Mean model equation for return value is [13]:

$$r_t = \mu + \varepsilon_t \quad (3)$$

where, r_t is the return value at t (day) and ε_t is independent observation from N (0, σ_t^2).

A model that follow heteroscedasticity follow serial correlation of variance equation bellow:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 \quad (4)$$

with estimation error,

$$v_t = \varepsilon_t^2 - \sigma_t^2 \quad (5)$$

Moreover, ARCH using equation (4) and (5) as equation bellow:

$$\varepsilon_t^2 = \sigma_t^2 + v_t = \omega + \alpha \varepsilon_{t-1}^2 + v_t \quad (6)$$

Generalized ARCH (GARCH) is an ARCH with ARMA (1,1). GARCH(p,q) has following equation.

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \quad (7)$$

EGARCH as further development of GACRH has equation as follows:

$$\log(\sigma_t^2) = \omega + \beta \log(\sigma_{t-1}^2) + \alpha \left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \quad (8)$$

EGACRH steps are decsribed bellow:

1) Heteroscedacity testing

In this step, residual squared and Q-statistic used to determine whether model follow heteroscedasticity or not.

2) Parameter Estimation

Parameter estimation in EGARCH model will be decided using Maximum Likelihood Estimation (MLE).

3) Diagnostic Checking

In this step, Akaike Info Criterion (AIC) will be used. AIC equation is as follows:

$$AIC = -\ln L + p \quad (9)$$

where L is the likelihood for an estimated model with p parameters. Model wth the smallest AIC value consider as the best model.

3.3 SARIMA-EGARCH

This research combined SARIMA method with EGARCH method into SARIMA-EGARCH to produce the most suitable model with network traffic data. The steps of SARIMA-EGARCH were shown in Figure 2.

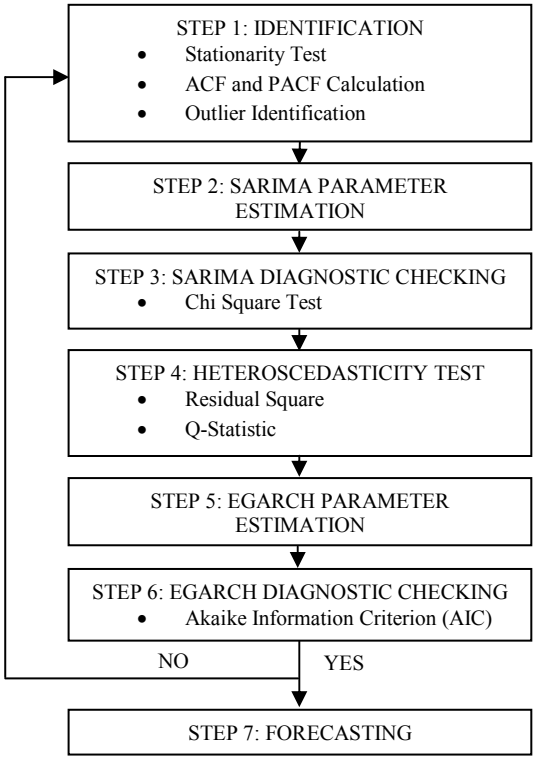

Figure 2: Steps of SARIMA-EGARCH

Journal of Theoretical and Applied Information Technology
15th February 2017. Vol.95. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. SARIMA MODEL FINDING

4.1 Identification

Before doing model identification, first, we identify the normality of time series plot using Anderson Darling test as it is shown in Figure 3.

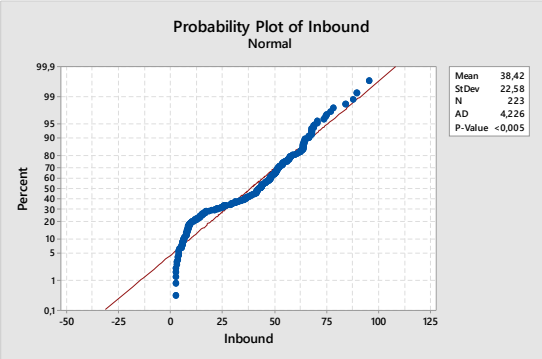

Figure 3: Probability Plot

Anderson Darling statistic is a calculation method on how far a plot point is from fitted line in probability plot [14]. The observation toward Figure 3 shows that data tend to be around the straight line, but an arch appears and be identified as abnormality of data distribution. The tendency of abnormal distribution is also shown in Anderson Darling statistic with big relative value, i.e. 4.226. Value of p-value = 0.005 meant p-value < 5% so that H0 is rejected or it can be said that the data distribution is abnormal.

To overcome abnormal distribution of the data, we perform differencing process. The differencing process that has been conducted is order 1 regular differencing and seasonal differencing. Figure 1 shows that the plot has certain peak trend so it brought out season at 7 lag. Therefore, we perform order 1 seasonal differencing with 7 lag. The following Figure 4 shows the comparisons between regular and order 1 seasonal differencing.

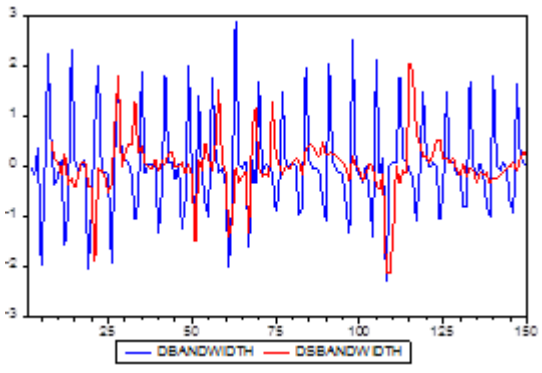

Figure 4: Regular and Seasonal Differencing

Model identification has been done by looking at ACF and PACF plot of the data. ACF and PACF plot with order 1 seasonal differencing using 7 lag are shown in Figure 5 and Figure 6.

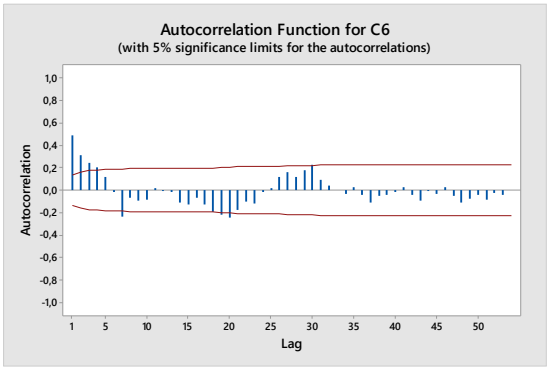

Figure 5: ACF with Order 1 Seasonal Differencing

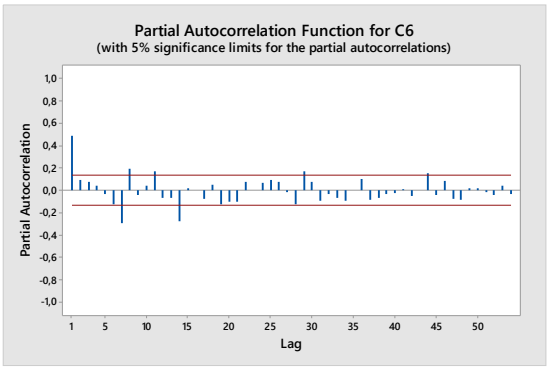

Figure 6: PACF with Order 1 Seasonal Differencing

4.2 Parameter Estimation

Based on ACF and PACF with order 1 seasonal differencing, the proposed models are displayed as follows.

Table 1: Proposed Models

No.	Proposed Models	No.	Proposed Models
1.	$(1,0,0)(1,1,0)_7$	9.	$(1,1,0)(1,1,0)_7$
2.	$(1,0,0)(2,1,0)_7$	10.	$(1,1,0)(1,1,1)_7$
3.	$(1,0,0)(3,1,0)_7$	11.	$(1,1,0)(2,1,0)_7$
4.	$(1,0,1)(1,1,1)_7$	12.	$(1,1,0)(3,1,0)_7$
5.	$(1,0,1)(1,1,1)_7C$	13.	$(1,1,0)(3,1,2)_7$
6.	$(1,0,1)(2,1,2)_7$	14.	$(1,1,1)(1,1,0)_7$
7.	$(1,0,1)(3,1,0)_7$	15.	$(1,1,1)(2,1,0)_7$
8.	$(1,0,1)(3,1,1)_7$	16.	$(1,1,1)(3,1,0)_7$

4.3 Diagnostic Checking

This procedure is used to examine the properness of selected models in Table 1. Diagnostic checking method is conducted through examining the signification of the models either by using constant

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

or not. The examination has been done using Chi-Square method. Model $(1,0,0)(1,1,0)_7$ obtains the following results.

Table 2: Example of Diagnostic Checking Results

Type		Coef	SE Coef	T	P
AR	1	0,5326	0,0720	7,40	0,000
SAR	7	-0,3115	0,0810	-3,84	0,000

The results show that the P value < α (0,05), it means that the model is significant and worth using. After going through the checking process, we obtain results that show all proposed models on the parameter estimation step has already been significant. The next checking process is examining whether there is any heteroscedasticity symptoms in residual model.

5. EGARCH

5.1 Heteroscedasticity Test

Heteroscedasticity test has been performed with Residual Square and Q-statistic method on each model. The first model is $(1,0,0)(1,1,0)_7$.

Table 3: Correlogram of Q-Statistic

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.013	0.013	0.0224	
		2	-0.030	-0.030	0.1437	
		3	-0.071	-0.071	0.8547	0.355
		4	0.068	0.069	1.5003	0.472
		5	0.167	0.163	5.4573	0.141
		6	0.028	0.025	5.5723	0.233
		7	-0.133	-0.121	8.1319	0.149
		8	0.007	0.028	8.1391	0.228
		9	-0.060	-0.085	8.6637	0.278
		10	-0.046	-0.097	8.9816	0.344
		11	0.164	0.187	12.999	0.163
		12	-0.088	-0.069	14.157	0.166
		13	-0.045	-0.047	14.466	0.208
		14	-0.306	-0.291	28.733	0.004
		15	-0.059	-0.065	29.274	0.006
		16	0.046	-0.027	29.609	0.009
		17	-0.034	-0.070	29.791	0.013
		18	-0.123	-0.030	32.201	0.009
		19	-0.053	0.013	32.649	0.012
		20	-0.040	-0.021	32.909	0.017

The correlogram of residual results using Q-Statistic methods shown in Table 3 shows that not all values Prob > α. It means there are autocorrelation symptoms in the residual. Furthermore, we did heteroscedasticity test using Residual Square.

Table 4: Correlogram of Residual Square

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.139	0.139	2.6477	
		2	-0.033	-0.053	2.7958	
		3	-0.005	0.007	2.7994	0.094
		4	-0.036	-0.039	2.9805	0.225
		5	-0.084	-0.075	3.9770	0.264
		6	0.095	0.118	5.2693	0.261
		7	0.235	0.205	13.222	0.021
		8	-0.031	-0.091	13.360	0.038
		9	-0.040	-0.016	13.595	0.059
		10	0.003	0.007	13.596	0.093
		11	-0.061	-0.039	14.158	0.117
		12	-0.026	0.015	14.262	0.161
		13	-0.042	-0.102	14.526	0.205
		14	0.106	0.091	16.245	0.180
		15	-0.006	-0.005	16.251	0.236
		16	0.024	0.032	16.337	0.293
		17	-0.048	-0.066	16.702	0.337
		18	-0.052	-0.021	17.128	0.377
		19	-0.079	-0.053	18.123	0.381
		20	-0.096	-0.074	19.600	0.356

From Table 4 above, we find that the values of Prob < α so that it can be concluded that there are heteroscedasticity symptoms. The following Table 5 shows the results of overfitting on all selected models.

Table 5: Overfitting Results

Model	Non-Auto Correlation	Non-Heteroskedasticity	Normality	AIC
$(1,0,0)(1,1,0)_7$	×	×	√	1.403246
$(1,0,0)(2,1,0)_7$	√	√	√	1.222809
$(1,0,0)(3,1,0)_7$	√	√	√	1.197958
$(1,0,1)(1,1,1)_7$	×	√	√	1.090783
$(1,0,1)(1,1,1)_7$C	×	×	√	1.073115
$(1,0,1)(2,1,2)_7$	√	√	√	0.847227
$(1,0,1)(3,1,0)_7$	√	√	√	1.214246
$(1,0,1)(3,1,1)_7$	√	√	√	0.891324
$(1,1,0)(1,1,0)_7$	×	×	√	1.638349
$(1,1,0)(1,1,1)_7$	×	×	√	1.289860
$(1,1,0)(2,1,0)_7$	×	×	√	1.436434
$(1,1,0)(3,1,0)_7$	×	×	√	1.419483
$(1,1,0)(3,1,2)_7$	×	×	√	1.144925
$(1,1,1)(1,1,0)_7$	×	×	√	1.354323
$(1,1,1)(2,1,0)_7$	√	√	√	1.243339
$(1,1,1)(3,1,0)_7$	√	×	√	1.214077

Table 5 above obtains 3 models with the smallest AIC value, i.e. (1,0,1)(2,1,2) with AIC 0,847227, (1,0,1)(3,1,1) with AIC 0, 891324, and (1,0,1)(1,1,1)C with AIC 1, 073115. Model (1,0,1)(2,1,2) and (1,0,1)(3,1,1) have passed the autocorrelation and heteroscedasticity test, but the model (1,0,1)(1,1,1)C has autocorrelation and heteroscedasticity problem that required further treatment.

Journal of Theoretical and Applied Information Technology
15th February 2017. Vol.95. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.2 Parameter Estimation

In addition, EGARCH parameter is added to overcome the problem of heteroscedasticity in the model. Based on the previous selected models, we choose the best EGARCH parameter with the smallest AIC value. EGARCH order itself has been generated using Maximul Likelihood method. Parameter Estimation produces 2 types of EGARCH, i.e. (1,0) and (1,1), with coefficient of variance equation as shown in table 6 and 7.

Table 6: EGARCH (1,0)

Models	ω	α	γ
$(1,0,1)(1,1,1)_7$C	-1.568795	-1.154560	-1.42342
$(1,0,1)(2,1,2)_7$	-1.409145	-1.345203	-1.465214
$(1,0,1)(3,1,1)_7$	-1.482054	-0.582703	-0.780126

Table 7: EGARCH (1,1)

Models	ω	β	α	γ
$(1,0,1)(1,1,1)_7$ C	-1.025397	0.337330	-0.958339	-1.383337
$(1,0,1)(2,1,2)_7$	-1.112856	0.202106	-1.241718	-1.525605
(1,0,1) (3,1,1)7	-1.104590	0.289426	-1.248159	-1.543103

5.3 Diagnostic Checking

Further, we conduct test on the selected models using equation (8), and coefficient in table 6 also 7. Akaike Information Criterion (AIC) equation (9) will be used to determine best model. The test results are displayed in Table 8.

Table 8: AIC Calculation

Model	AIC with EGARCH			
	Without EGARCH	(1,0)	(1,1)	Max Difference
$(1,0,1)(1,1,1)_7$C	1.0731	0.8051	0.7569	29,54%
$(1,0,1)(2,1,2)_7$	0,8472	0.6853	0.6197	26,85 %
$(1,0,1)(3,1,1)_7$	0.8913	0.9665	0.5195	41,71%

Based on Table 6 above, AIC value of model $(1,0,1)(3,1,1)_7$ with EGARCH (1,1) generates AIC value 41.71% smaller than the model without EGARCH. The smallest AIC value is also on this model so that it can be said that the best model is the model SARIMA $(1,0,1)(3,1,1)_7$ EGARCH(1,1). The final model is as follows:

$$\log(\sigma_t^2) = -1.104590 + 0.289426 \log(\sigma_{t-1}^2) + 1.248159 \left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + (-1.543103)\frac{\varepsilon_{t-1}}{\sigma_{t-1}}$$

6. FORECASTING

The last step is evaluation of selected model SARIMA$(1,0,1)(3,1,1)_7$ EGARCH(1,1) by calculating the model using MAPE standard. Forecasting data that used was the inbound bandwidth data on May 19, 2016 until June 17, 2016. The forecasting results can be seen in Figure 7 as follows.

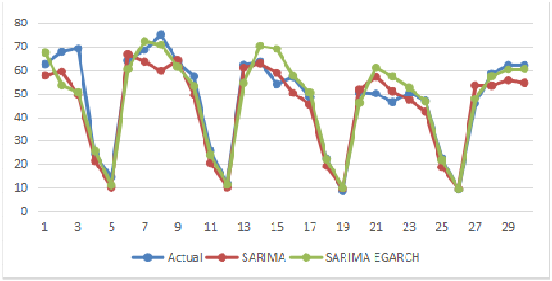

Figure 7: Forecasting Results

The MAPE calculation generates value 11.38% for SARIMA model $(1,0,1)(3,1,1)_7$ without EGARCH and 9.20% for model with EGARCH. It can be seen that the addition of EGARCH parameter to the SARIMA model generates smaller MAPE value by a margin of 19.15% if we compared it to SARIMA model without EGARCH. This result is also smaller than the error calculation results in previous study [2].

Despite having MAPE smaller than the model without EGARCH, SARIMA-EGARCH models still have no fix error variance. From Figure 5, the first results show that SARIMA–EGARCH forecasting has peak lower than the actual value, but the peaks of the third and fourth seasons are higher than the actual value.

7. CONCLUSSION

During this research, Residual Square test found out that there were some indications of heteroscedasticity among the residuals. Therefore, the method must be applied to overcome the heteroscedasticity problems. The use of the combined SARIMA-EGARCH method is able to reduce the value of AIC and produce smaller MAPE than SARIMA model without EGARCH. SARIMA-EGACRH model is able to increase

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

forecasting accuracy by reducing the heteroscedasticity.

This research can be challenging for further exploration. In addition, autocorrelation symptoms are also found in the model examined with Q-Statistic method. Bandwidth data has a unique characteristic, which can be very low or very high. Therefore, outlier, autocorrelation, and heteroscedasticity are likely happened. This symptomps can be explored in the future research to get a better result.

REFRENCES:

[1] Dong Peng, Yuanyuan Qiao, Jie Yang. "Analyzing Traffic Characteristic Between Backbone Network Based on Hadoop". *Proceedings of CCIS 2014*.

[2] Permanasari, Adhistya Erna. Hidayah, Indriana. Bustoni, Isna Alfi. "Forecasting Model for Hotspot Bandwidth Management at Department of Electrical Engineering and Information Technology UGM". *International Journal of Applied Mathematics and Statistics.* Vol. 53; Issue No.4 , Year 2015. ISSN 0973-1377

[3] Dandan, Miao. Xiaowei, Qin. Weidong, Wang. " The Periodic Data Traffic Modeling Based on Multiplicative Seasonal ARIMA Model.". *Sixth International Conference on Wireless Communications and Signal Processing (WCSP).* 2014.

[4] Hanbanchong, Aphichit. Piromsopa, Krerk. "SARIMA Based Network Bandwidth Anomaly Detection." *Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE).* 2012.

[5] Permanasari, Adhistya Erna. Hidayah, Indriana. Bustoni, Isna Alfi. "SARIMA (Seasonal ARIMA) Implementation on Time Series to Forecast The Number of Malaria Incidence." *Proceeding of International Conference on Information Technology and Electrical Engineering (ICITEE)*. 201.

[6] Néstor, González Cabrera. G. Gutiérrez-Alcaraz. Esteban, Gil. "Load Forecasting Assessment Using SARIMA Model and Fuzzy Inductive Reasoning." *Proceding of IEEE IEEM*. 2013.

[7] Williams, Richard. "Heteroskedasticity". [Online]. Available: http://http://www3.nd.edu/~rwilliam/ . [Accessed: 30- Jan- 2015].

[8] Nursalam. "Pemodelan Exponensial GARCH". *Jurnal Sains dan Teknologi UIN Allaudin Makasar.* Vol 5, No.2, July 2011. ISSN: 1979-3154

[9] Annila, Nur. Kritianti, Farida Titik. "Model Garch (Generalized Autoregressive Conditional Heteroscedasticity) untuk Prediksi dan Akurasi Harga Saham Masa Depan". *Jurnal E-Proceeding of Management.* Vol 2 No. 1, April 2015. ISSN:2355-9357

[10] Widayati, Nur. "Penerapan Model GARCH dan Model EGARCH pada Saham Sektor Properti Ketika Krisis Ekonomi Dunia". Departemen Statistika Fakultas MIPA, IPB. Year 2009.

[11] Ning.L. "Emipirical Research on Term Structure of Buy-Back Rates Based on EGARCH Model". *Proceeding of International Symposium on Electronic Commerce and Security.* Volume:1,

[12] Box, E.P. George, Jenkins. M. Gwilym. "Time Series Analisys Forecating and Control". New Prentice Hall. Year 2004

[13] Nelson, Daniel B. "Conditional Heteroskedasticity in Asset Returns: A New Approach", *Econometrica*, Vol. 59,347-370. Year 1991.

[14] T. W. Anderson and D. A. Darling, "A test of goodness of fit", J. Amer. Stat. Assn., 49 765769. 1954