JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 21, No. 3, May 2025, pp. 847-854
DOI: 10.20956/j.v21i3.43613

Stock Price Forecasting Using Autoregressive With Exogenous Variable Support Vector Regression (ARX – SVR)

Erlyne Nadhilah Widyaningrum*, Rizka Amalia Putri[2], Morina A. Fathan[3], Nur Rezky Safitriani[4]

[*] Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, 75124, Indonesia
[2]Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Riau University, Pekanbaru, 28293, Indonesia
[3,4]Department of Statistics, Faculty of Mathematics and Natural Sciences Tadulako University, Palu, 94119, Indonesia
***Email**: [*]erlynenadhilah@fmipa.unmul.ac.id, [2]rizkaamaliaputri@lecturer.unri.ac.id, [3]morinafathan@untad.ac.id, [4]rezky.untad@untad.ac.id*
**Corresponding author*

Received: 26 March 2025, revised: 23 April 2025, accepted: 25 April 2025

Abstract

Stock prices move fluctuate continuously and dynamically at all times, so stock price predictions are needed to maximize profits for investors and avoid losses due to the characteristic of stock prices. Autoregressive (AR) model is a forecasting method and has weaknesses against nonlinear patterns. In addition to using linear modeling, forecasting stock prices can use the Support Vector Regression model which offers a global optimal solution that works with data maps to high-dimensional spaces and has good performance with time series problems. The addition of exogenous variables X to the model can also improve forecasting accuracy. Forecasting will be done using significant lags as input to Support Vector Regression. The modeling results show that the ARX-SVR model with X as an outlier exogenous variables provides the best out-of-sample data forecasting results for the case study of stock closing price forecasting. This model provides forecasting results with Symmetric Mean Absolute Percentage Error (sMAPE) 5.382430%.

Keywords: Exogenous Variable, Forecast, Stock, Support Vector Regression

1. INTRODUCTION AND PRELIMINARIES

1.1 Introduction

Stocks are one of the most popular capital market instruments. Issuing stocks is one of the strategies of a company or business entity to raise funds for the company. Stock price financial data is time series financial data that has very high volatility. The characteristics of this financial data make stock price data forecasting require precise and accurate forecasting methods to make forecasts [9]. Stock price forecasting is an analysis technique to determine stock prices in the future by using past stock price history. Stock price forecasting is very useful for investors engaged in stock trading. This technique is to avoid losses due to the nature of stock prices that move fluctuative and tend to be dynamic at all times, so stock price predictions are needed to maximize profits for investors [13].

A time series model that has been carried out in previous studies is the Vector Autoregressive Moving Average (VARMA) which is an extension of the ARMA model. This model explains the relationship between observations on the variable itself at previous times, and also its relationship with observations on other variables at previous times [16]. The VARIMA modeling procedure was introduced by Tiao & Box starting from determining a tentative model, estimating parameters, and checking diagnostics [15]. The Vector Autoregressive (VAR) and Vector Autoregressive with Exogenous Variable (VARX) models are time series models used to predict the future based on linear functions of past observations. According to Suhartono et al. (2018), the addition of exogenous variables X to the model can also increase forecasting accuracy [14].

In addition, other time series data models that can model past data include using Machine Learning to improve model performance based on the prediction result plot, the SVR method shows better performance in predicting CPI data compared to ARIMAX [5]. In addition, SVM has been successfully applied in stock prediction with an accuracy rate of around 60%–70% for the Support Vector Regression model. After calculating the results accurately using the conventional model, it was found that Support Vector Regression has a higher accuracy rate than linear regression [2]. Support Vector Regression (SVR) can provide an alternative in predicting stock prices. Support Vector Regression (SVR) emerged as a new technique in the process of machine learning from data to solve regression problems. Support Vector Regression (SVR) is a machine learning method that is very suitable for solving problems with data that has a high dimensional data [11]. The concept of Support Vector Regression works as a linear classifier, then developed for nonlinear problems by inserting a kernel in a high-dimensional workspace [3]. The Support Vector Regression method finds a globally optimal solution and works by mapping training data to a high-dimensional space [6]. The SVR method is used to analyze various parameters of stock market performance, such as daily and monthly returns, monthly cumulative returns, their volatility properties, and the risks associated with them [4].

Based on the description above, Autoregressive will be used with the addition of exogenous variables for time series modeling which is used to predict the future based on linear functions from past observations. Then it will be continued by using Machine Learning, namely Support Vector Regression to estimate stock prices. This study aims to contribute to the field as a consideration for investors in investing in stock.

1.2 Dataset Description

The data used in this study is the closing price of stocks from construction and building companies. The variable used in the research is arranged in Table 1.1 as follows:

Table 1.1 Research Variables

No.	Data	Notation
1.	Closing Price of ADHI	Y_t

Input and output variables are describe as follows:

1) Output Series
   Y_t : stock prices in construction and building sector
2) Input Series
   X : significant lag of Autoregressive with Exogenous Variable (ARX)

1.3 Autoregressive

An Autoregressive model represents a forecast as a function of previous values of a specific time series [10]. The Autoregressive (AR) model with order p is denoted as AR(p). The general form of the time series model for Autoregressive is expressed as:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{1.1}$$

1.4 Support Vector Regression

The basic idea of SVR is to determine the data set which is divided into in-sample and out-sample data. Then from the in sample data a regression function is determined with certain deviation limits so that it can produce predictions that are close to the actual value. If there $n$is a data set in sample, $(\boldsymbol{x}_i, y_i)$then is $\boldsymbol{x}_i = \{x_1, \dots, x_n\}^T \in R^n$a $i = 1,2,\dots,n$vector in the input space and$y = \{y_1, \dots, y_n\} \epsilon\, R$ is the output value based on $\boldsymbol{x}_i$the corresponding, then the linear SVR method regression function is obtained in equation (1.2)

$$y = f(\boldsymbol{x}_i) = \boldsymbol{w}.\boldsymbol{x}_i + b \tag{1.2}$$

Where $\boldsymbol{w}$ is the weight vector and b is the bias. To get a good generalization, the coefficients $\boldsymbol{w}$ and b estimated by minimizing the risk function as follows [7]

$$R(f(\boldsymbol{x}_i)) = \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\sum_{i=1}^{n} E_\varepsilon\,(y_i - f(\boldsymbol{x}_i)) \tag{1.3}$$

With $\left|\left|\boldsymbol{w}\right|\right|$is the regularization of the function which is minimized to make the function as *flat as* possible. The constant $C(Cost) > 0$ is the trade off between the thinness of the function f and the upper limit of the deviation ε that is more than can be tolerated [1]. E_ε is $\varepsilon - insensitive\ loss\ function$ defined as follows.

$$E_\varepsilon\,(y_i - f(\boldsymbol{x}_i)) = \begin{cases} |y_i - f(\boldsymbol{x}_i)| - \varepsilon, & for\ |y_i - f(\boldsymbol{x}_i)| \geq \varepsilon \\ 0, & others \end{cases} \tag{1.4}$$

By using the Lagrange multiplier and optimality conditions, the SVR function can be written in equation 1.5

$$f(\boldsymbol{x}_i) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{1.5}$$

Where $(\boldsymbol{x}_i, \boldsymbol{x})$is defined via a kernel function and $w = (\alpha_i - \alpha_i^*)$. In the case of real data , not all data is linear, so finding a linear separator is difficult. SVR can be used for nonlinear cases with an alternative approach through data mapping $\boldsymbol{x}$ from input space to feature space with higher dimensions of a function φ so that $\varphi\!: \boldsymbol{x} \rightarrow \varphi(\boldsymbol{x})$**.** The equation of the nonlinear SVR regression function can be seen through equation 1.6

$$f(\boldsymbol{x}_i) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\varphi(\boldsymbol{x}_i).\varphi(\boldsymbol{x}) + b \tag{1.6}$$

The transformation φ is generally unknown and difficult to understand so this problem is solved by using the kernel function in equation

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \varphi(\boldsymbol{x}_i).\varphi(\boldsymbol{x}) \tag{1.7}$$

Thus, the equation of the nonlinear SVR regression function using the kernel function is obtained as equation 1.8

$$f(\boldsymbol{x}_i) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{1.8}$$

1.5 Model Training and Evaluation

To find out the forecast results are made, a method of measuring the performance of prediction or forecasting results is needed. One of the best model selections through the training data approach is by using RMSE (Root Mean Square Error). RMSE is used as a tool to select the best model based on the size of the squared error [8]. The formula for calculating RMSE can be seen in equation 1.9

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2} \tag{1.9}$$

In addition to using RMSE, selecting the best model through testing data approach can use sMAPE(Symmetric Mean Absolute Percentage Error). Performance measurement using SMAPE will produce a value in the form of a percentage. The smaller the SMAPE value, the better the forecast will be [12]. The formula for calculating sMAPE can be seen in equation 1.10

$$sMAPE = \frac{1}{N} \sum_{l=1}^{N} \left| \frac{Y_l - \hat{Y}_l}{(Y_l - \hat{Y}_l)/2} \right| \times 100\% \tag{1.10}$$

2. MAIN RESULTS

The government is currently implementing equal development and planning the construction of the National Capital in Kalimantan. This government policy has the potential to attract investors to invest in construction and building sector companies. Based on data from the Indonesia Stock Exchange (IDX), currently the State-Owned Enterprises engaged in this sector and also included in the LQ45 stock index category are PT. Adhi Karya Tbk. ADHI is the stock code for PT Adhi Karya (Persero) Tbk. So, ADHI stock price data will be used to predict stock prices. The characteristics of stocks price data in construction and building sector are obtained through descriptive statistical analysis by exploring information in data without making inference. Figure 2.1 shows the stock price data (ADHI) from August 2017 to August 2022. There is up and down trend pattern that explains the non-statioary mean data and the PACF plot for stationary data is shown in Figure 2.2

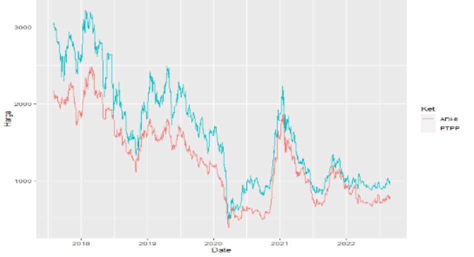


Figure 2.1 Time Series Plot of Stocks Price

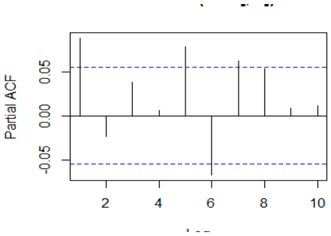


Figure 2.2 Plot PACF ADHI

The results of the PACF plot identification in Figure 2 show that significant lags are found at lags 1,5,6,7 for ADHI. So the possible AR models are ARI(1), ARI(5), ARI(6), ARI(7) and ARI[1,5,6,7] for ADHI. The AIC values for each possible model are shown in Table 2.1

Table 2.1 AIC Value

ADHI	AIC
ARI(1,1)	12631.56
ARI(5,1)	12629.19
ARI(6,1)	12625.44
ARI(7,1)	12622.57
ARI([1,5,6,7],1)	12620.11

After obtaining a significant AR model estimate in Table 2.1, The selection of the best model is determined by considering the smallest AIC value, based on the Table 2.1 ARI model ([1,5,6,7],1) was selected as the best model. The next step was diagnostic checking which carried out on the residuals included the normal distribution assumption and the white noise assumption as follows in Table 2.2

Table 2.2 Diagnostic Checking on Residuals

Model	White Noise Test		Normality Test
	Lag	p-value	p-value
ADHI	6	0.2261	0.000
	12	0.0471	
	18	0.1182	
	24	0.1127	

Table 2.3 Parameter Estimation in ADHI model

y	Parameter	Estimation	Std.Error	t	p-value	Variable
$y_{1,t}$	ϕ_1	0.09531	0.02791	3.41	0.0007	$y_{1,(t-1)}$
	ϕ_5	0.08342	0.02790	2.99	0.0028	$y_{1,(t-5)}$
	ϕ_6	-0.07124	0.02810	-2.54	0.0114	$y_{1,(t-6)}$
	ϕ_7	0.05897	0.02796	2.11	0.0351	$y_{1,(t-7)}$

Based on the Table 2.3 the model formed for ADHI has a p-value <0.05, so the initial hypothesis decision is rejected. Likewise, the normal distribution test obtained is a p-value <0.05 so that the assumption of white noise and normal distribution has not been met. Autoregressive (AR) modeling is continued with outlier observation detection to overcome the problem of unmet assumptions. In this study, two types of outliers will be observed, namely level shift and additive as many as 5 observations. The outlier examination obtained the following results in Table 2.4

Table 2.4. Outlier Detection

Observations	Type	Variable
127	Shift	${I_t}^{(127)}$
195	Shift	${I_t}^{(195)}$
676	Shift	${I_t}^{(676)}$
695	Shift	${I_t}^{(695)}$
882	Additive	${I_t}^{(882)}$

Next, after obtaining outlier observations, each Autoregressive (AR) model is added with a dummy variable representing the outlier observation (I_t), then continued with the Autoregressive (AR) model estimation with outliers. The parameter estimation results can be displayed in Table 2.5.

Table 2.5. Parameter Estimation of Autoregressive (AR) Model with Outliers in ADHI Model

y	Parameter	Estimation	Std.Error	t	p-value	Variable
$y_{1,t}$	ϕ_1	0.09748	0.02798	3.48	0.0005	$y_{1,(t-1)}$
	ϕ_5	0.07406	0.02801	2.64	0.0083	$y_{1,(t-5)}$
	ϕ_6	-0.06383	0.02820	-2.26	0.0238	$y_{1,(t-6)}$
	ϕ_7	0.04789	0.02818	-1.99	0.0470	$y_{1,(t-7)}$
	ω_{LS}	1.94917	32.66349	6.08	<.0001	${I_{127}}^{(127)}$
	ω_{LS}	-4.09974	32.63191	-7.21	<.0001	${I_{195}}^{(195)}$
	ω_{LS}	-11.83293	32.65926	-2.86	0.0044	${I_{676}}^{(676)}$
	ω_{LS}	15.39641	32.67049	3.68	0.0002	${I_{695}}^{(695)}$
	ω_{AO}	-108.19908	22.27514	-3.73	0.0002	${I_{882}}^{(882)}$

The Autoregressive (AR) model equation with outliers for the closing price of ADHI stocks can be described as follows

$$\begin{aligned} y_t = \; & 0.09748\left(y_{(1,t-1)} - y_{(1,t-2)}\right) + \; 0.07406\left(y_{(1,t-5)} - y_{(1,t-6)}\right) \\ & - 0.06383\left(y_{(1,t-6)} - y_{(1,t-7)}\right) \\ & + 0.04789\left(y_{(1,t-7)} - y_{(1,t-8)}\right) + 1.94917\, {I_t}^{(127)} \\ & - 4.09974\, {I_t}^{(195)} - 11.83293\, {I_t}^{(676)} + 15.39641\, {I_t}^{(695)} \\ & - 108.19908\, {I_t}^{(882)} + a_{1,t} \end{aligned} \qquad (2.1)$$

Diagnostic checking is carried out on the residuals which includes the assumption of normal distribution and the assumption of white noise as follows.

Table 2.6 Diagnostic Checking on Residual Autoregressive (AR) Models with Outliers

Model	White Noise Test		Normality Test
	Lag	Nilai *p*	Nilai *p*
ADHI	6	0.4543	0.000
	12	0.0509	
	18	0.1248	
	24	0.0822	

The assumption of normality that has not been met is suspected because the Autoregressive (AR) model contains non-linear components. Therefore, the approach with a non-linear model is also applied in this study, namely by modeling the lag of the Autoregressive (AR) model with the Support Vector Regression (SVR) model.

Table 2.7 Input Lag Support Vector Regression (SVR)

Model		Input Lag
ADHI	ARI([1,5,6,7],1)	$f(y_{1,(t-1)}, y_{1,(t-5)}, y_{1,(t-6)}, y_{1,(t-7)})$
	ARI([1,5,6,7],1) + $\omega {I_t}^{(T)}$	$f({I_t}^{(127)}, {I_t}^{(195)}, {I_t}^{(676)}, {I_t}^{(695)}, {I_t}^{(882)}, y_{1,(t-1)}, y_{1,(t-5)}, y_{1,(t-6)}, y_{1,(t-7)})$

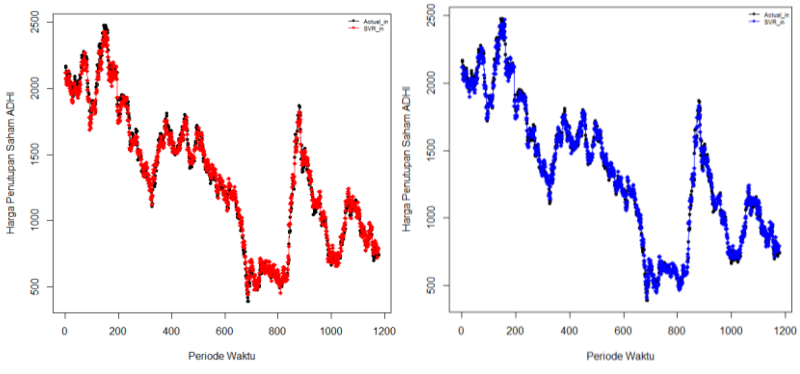
The parameter range used for the Support Vector Regression (SVR) model is $C = \{8^{-1} - 2^5\}, \varepsilon = \{0.01 - 0.1\}$ and $\gamma = \{2^1 - 2^4\}$. Based on the results of the combination of these values, the optimal Support Vector Regression (SVR) parameters obtained for C, ε and γ are 31.125, 0.1 and 2^1. After the model is obtained, modeling will be carried out for the next five months on the testing data. The comparison of sMAPE values on the testing data is as in Table 2.8.

Table 2.8. Comparison of sMAPE Grid Search Optimization

	Input = AR			Input = AR with X outlier		
	RMSEin	RMSEout	sMAPEout	RMSEin	RMSEout	sMAPEout
ADHI	24.06737	40.42749	5.492957%	24.24414	39.70523	5.38243%

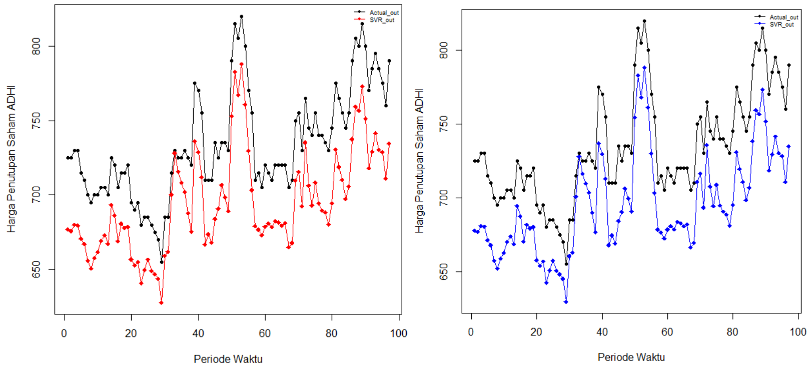
The SVR model formed has the following equations

$$\hat{y}_i{}^{(1)} = (\alpha_1 - \alpha_1^*)\exp(-2 \times \|\boldsymbol{x}_1 - \boldsymbol{x}\|^2) + \cdots \\ + (\alpha_{1178} - \alpha_{1178}^*)\exp(-2 \times \|\boldsymbol{x}_{1178} - \boldsymbol{x}\|^2) - 0.4882775 \tag{2.2}$$



(a) (b)

Figure 2.1 SVR Training Data Plot with (a) input = AR and (b) input = AR with X Outlier



(a) (b)

Figure 2.2 SVR Testing Data Plot with (a) input = AR and (b) input = AR with X outlier

From the Table 2.8, it is found that the Autoregressive (AR) input model with X outlier has good accuracy in predicting the closing price of stocks in ADHI companies. This is indicated by the sMAPE value which is smaller than the other models. Therefore, the SVR model with significant lag input from the Autoregressive (AR) model with X outlier performs better in predicting the closing price of ADHI stocks.

3. CONCLUSION

This paper proposed SVR to provide a more accurate model for stock prices forecast. The proposed method is a solution combining the AR model with exogenous variable. Comparative results of different techniques such as SVR with input AR (AR-SVR) and SVR with Autoregressive input model with exogenous variable (ARX-SVR) confirm the out performance of the proposed approach. The Autoregressive input model with exogenous variables (ARX) has met the white noise assumption so that the best model for predicting stock closing prices is the Support Vector Regression model with the ARX input model (ARX-SVR). The model provides the best forecasting results with sMAPE ADHI = 5.38243%. Further development can be done on other optimization methods so that the results can be compared.

REFERENCES

[1] Bernhard., Scholkopf. & Smola, A.J., 2002. *Learning With Kernel*. MIT Press, Massachusetts.

[2] Chhajer, P., Shah, M. & Kshirsagar, A., 2022. The Applications of Artificial Neural Networks, Support Vector Machines, and Long–Short Term Memory for Stock Market Prediction. *Decision Analytics Journal*, Vol. 2, 100015.

[3] Cortes., Corinna. & Vapnik, V., 1995. Support-Vector Networks. *Machine Learning*, Vol. 20, 273–297.

[4] Dash., Kumar., Nguyen,T.N., Cengiz, K. & Sharma, A., 2023. Fine-Tuned Support Vector Regression Model for Stock Predictions. *Neural Computing and Applications*, Vol. 35, No. 32, 23295–23309.

[5] Ghofur, A.F., Dewi, Y.S. & Anggaraeni, D., 2022. Comparison Of Support Vector Regression And Autoregressive Integrated Moving Average With Exogenous Variable On Indonesia Consumer Price Index. *Science and Research Journal*, Vol. 5 , No. 3, 144-148.

[6] Gunn., Steve R., 1997. *Support Vector Machines for Classification and Regression*. University of Southampton.

[7] Haykin, Simon. 2009. *Neural Network and Learning Machines*, Third Edition. Pearson Education Inc., New Jersey.

[8] Hillmer, Steven C. & William W. S., 1991. Time Series Analysis: Univariate and Multivariate Methods. *Journal of the American Statistical Association*, Vol.86, No.413, 245.

[9] Kewat., Pooja., Sharma, R., Singh, U. & Itare, R., 2017. Support Vector Machines through Financial Time Series Forecasting. *International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, Vol. 2, 471–477.

[10] Makridakis, Spyros, Wheelwright, S.C., McGee, V.E., Andriyanto,U.S. & Basith,A., 1999. *Metode Dan Aplikasi Peramalan Jilid 1*, Edisi Kedua. Bina Rupa Aksara., Jakarta.

[11] Meesad., Phayung. & Rasel, R.I., 2013. Predicting Stock Market Price Using Support Vector Regression. *International Conference on Informatics, Electronics and Vision (ICIEV)*, 1–6.

[12] Pai, Feng, P. & Lin, C.S., 2005. A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting. *Omega*, Vol.33, No. 6, 497–505.

[13] Rahmadayanti, Cipta, Rabbani, H. & Rohmawati, A.A., 2018. Model GARCH Dengan Pendekatan Conditional Maximum Likelihood Untuk Prediksi Harga Saham. *Indonesia Journal on Computing (Indo-JC)*, Vol.3, No. 2, 21–28.

[14] Suhartono., Gazali, M.M. & Prastyo,D.D., 2018. VARX and GSTARX Models for Forecasting Currency Inflow and Outflow with Multiple Calendar Variations Effect. *Matematika*, Vol. 34, No. 3, 57–72.

[15] Tiao., George, C. & Box, G.E.P., 1981. Modeling Multiple Time Series with Applications. *Journal of the American Statistical Association*, Vol. 76, No. 376, 802–816.

[16] Wutsqa. & Urwatul, D., 2010. Seasonal Multivariat Time Series Forecasting on Tourism Data by Using Var-Gstar Model. *Jurnal Ilmu Dasar*, Vol.11, No.1, 101–109.