International Journal of Accounting and Finance Studies
Vol. 2, No. 1, 2019
www.scholink.org/ojs/index.php/ijafs
ISSN 2576-2001 (Print) ISSN 2576-201X (Online)

Original Paper

ARIMA Model Selection for Composite Stock Price Index in Indonesia Stock Exchange

Zul Amry[1*] & Budi Halomoan Siregar[1]

[1] Department of Mathematics, University of Medan, Indonesia

[*] Zul Amry, Department of Mathematics, University of Medan, Indonesia

Received: May 17, 2019 Accepted: May 27, 2019 Online Published: May 30, 2019

doi:10.22158/ijafs.v2n1p31 URL: http://dx.doi.org/10.22158/ijafs.v2n1p31

Abstract

Composite Stock Price Index (CSPI) can be used as a reflection of the national economic condition of a country because it is an indicator to know the development the capital market in a country. Therefore, the movement in the future needs to be forecast. This study aims to build a model for the time series forecasting of Indonesia Composite Index (ICI) using the ARIMA model. The data used is the monthly data of ICI in Indonesia Stock Exchange (IDX) from January 2000 until December 2017 as many as 216 data. The method used in this research is the Box-Jenkins method. The autocorrelation (ACF) and partial autocorrelation function (PACF) are used for stationary test and model identification. The maximum estimated likelihood is used to estimate the parameter model. In addition, to select a model then used Akaike's Information Criterion (AIC). Ljung-Box Q statistics are used for diagnostic tests. In addition, to show the accuracy of the model, we use Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) and the most appropriate model is ARIMA (0, 1, 1).

Keywords

CSPI, Box-Jenkins method, ARIMA model, IDX

1. Introduction

In general, investment managers need to accurately predict the CPSI in order to minimize the risk of the decision. Furthermore, most of the central banks in the world generally use CPSI data as one of the considerations to determine monetary policy. In other words, monetary policy was decided by considering the upcoming CPSI value. One tool to predict the CPSI value is to use a time series model. Capital market is the part of the financial system concerned with raising capital by dealing in shares, bonds, and other long-term investments. Where, it can support the development of the national economy, because it can support the financing of national development. The ups and downs of the capital market stretching can be reflected in the movement of the Composite Stock Price Index (CSPI), in other words, that the CSPI is one indicator of the condition of a market. JCI is a series of historical information about joint stock price movements up to a certain time. Where, it is able to reflect a value that can be used as a performance indicator of a joint stock in the stock exchange. Therefore, when investment managers take a financial decision, they need to predict JCI accurately to reduce the risk of loss from the decision. The Autoregressive Integrated Moving Average (ARIMA) model developed by Box and Jenkins (1976) and has been widely used in various fields as a statistical model, especially related to forecasting problems. Time series model forecasting is a type of forecasting that uses past observational data, investigates its behavior and is extrapolated into the future. This paper focuses on constructing a time series forecasting model with the ARIMA model to be applied to CSPI data forecasting.

2. Method

The material used in this study consisted of CPSI data and a number of theories in statistics and mathematics. CPSI data is a monthly data from January 2000 to December 2017 as many as 216 data and Autoregressive Integrated Moving Average model written ARIMA (p, d, q), with the general formula:

$$\phi_p(B)(1-B)^d \dot{Z}_t = \theta_q(B)a_t \tag{1}$$

Where d ≠ 0 is the level of differentiation, then the method used is the Box-Jenkins method.

The ARIMA model was first popularized by Box and Jenkins (1976), known as the Box-Jenkins method or the Box-Jenkins model. This research is based on Box-Jenkins modeling. Where, the steps are: stationary test based on behavior from ACF graph, identical model using ACF and PACF, parameter estimation using maximum likelihood method, model selection using AIC, diagnostic test using Ljung-Box Q statistics, and test model accuracy using RMSE, MAE and MAPE.

Definition 2.1

For the time series $\{Y_t; t \in Z\}$, the ACF is $\{\hat{\rho}_k,\ k = 0, 1, 2, \ldots\}$, where

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}\left(Y_t - \overline{Y}\right)\left(Y_{t+k} - \overline{Y}\right)}{\sum_{t=1}^{n}\left(Y_t - \overline{Y}\right)^2} \tag{2}$$

Definition 2.2

If the time series $\{Y_t; t \in Z\}$, then the PACF is $\{\hat{\phi}_{kk},\ k = 0, 1, 2, \ldots\}$, where

$$\hat{\phi}_{kk}=\frac{\begin{vmatrix} 1 & \hat{\rho}_1 & \hat{\rho}_2 & \Lambda & \hat{\rho}_{k-2} & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \Lambda & \hat{\rho}_{k-3} & \hat{\rho}_2 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & & \mathrm{M} & \mathrm{M} \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \hat{\rho}_{k-3} & \Lambda & \hat{\rho}_1 & \hat{\rho}_k \end{vmatrix}}{\begin{vmatrix} 1 & \hat{\rho}_1 & \hat{\rho}_2 & \Lambda & \hat{\rho}_{k-2} & \hat{\rho}_{k-1} \\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \Lambda & \hat{\rho}_{k-3} & \hat{\rho}_{k-2} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & & \mathrm{M} & \mathrm{M} \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \hat{\rho}_{k-3} & \Lambda & \hat{\rho}_1 & 1 \end{vmatrix}} \qquad (3)$$

Moreover, the identification of models is determined based on the characteristics of ACF and PACF, shown as in following Table 1:

Table 1. Characteristics of ACF and PACF

Model	ACF ρ_k	PACF ϕ_{kk}
AR(*p*)	Damped exponential and/or sine functions	ϕ_{kk} =0 for k>p
MA(*q*)	ρ_k = 0 for k >q	Dominated by damped exponential and/or sine function
ARMA(*p,q*)	Damped exponential and/or sine functions after lag (q-p)	Dominated by damped exponential and/or sine function after lag (p-q)

Definition 2.3

The joint density function of n random variables $X_1, X_2, \ldots, X_n$ evaluated at $x_1, x_2, \ldots, x_n$ say $f(x_1, x_2, \ldots, x_n; \theta)$, is referred to as a likelihood function. For fixed $x_1, x_2, \ldots, x_n$ the likelihood function is a function of θ and often denoted by $L(\theta)$. If $X_1, X_2, \ldots, X_n$ represents a random sample from $f(x;\theta)$, then the likelihood function is:

$$L(\theta)=\prod_{i=1}^{n} f(x_i,\theta) \qquad (4)$$

The maximum likelihood method is the method used to determine the minimum value. The AIC is defined are as follows (Wei, 2006):

$$AIC = n\ ln\ \hat{\sigma}_a^2 + 2M \qquad (5)$$

Where n is the number of observations, $\hat{\sigma}_a^2$ is the maximum likelihood estimate of σ_a^2 and M is the number of parameters. The best model is given by the model with the lowest AIC value. An overall check of model adequacy is provided by the Ljung-Box Q statistics. The test statistic Q (Wei, 1994) is:

$$Q = n(n+2)\sum_{k=1}^{K}\frac{\hat{\rho}_k^2}{n-k} \sim \chi^2(K-p-q)) \qquad (6)$$

If $Q > \chi^2{}_{1-\alpha}\left(K-p-q\right)$ the adequacy of the model is rejected at the level α. Where n is the sample size, $\hat{\rho}_k^2$ is the autocorrelation of residuals at lag k and K is the number of lags being tested.

There are measures to determine the accuracy of a forecasting model in this research is RMSE, MAE and MAPE defined respectively as follows:

$$RMSE = \sqrt{\frac{ESS}{n}} \tag{7}$$

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

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|}{n} x\ 100\% \tag{9}$$

Where Y_t=The actual value at time t; $\hat{Y}_t$=The forecast value at time t; n=The number of observations and *ESS*=the error sum of square.

3. Results

Based on the original data plot x in Figure 1, the graph shows an increasing trend, this indicates that the time series data is not stationary. Furthermore, this is confirmed by the ACF plot in Figure 2 which shows a slow decline. Therefore, it was concluded that the time series x data set was not stationary. To overcome this non-stationary condition, it is transformed to y; y is the first differences of x:

$$y_t = x_t - x_{t-1} \tag{10}$$

The data plot in Figure 3 indicates that the time series data is stationary, confirmed by the ACF plot in Figure 4 with a muffled sine wave shape. Furthermore, it was concluded that this set of data y was stationary.

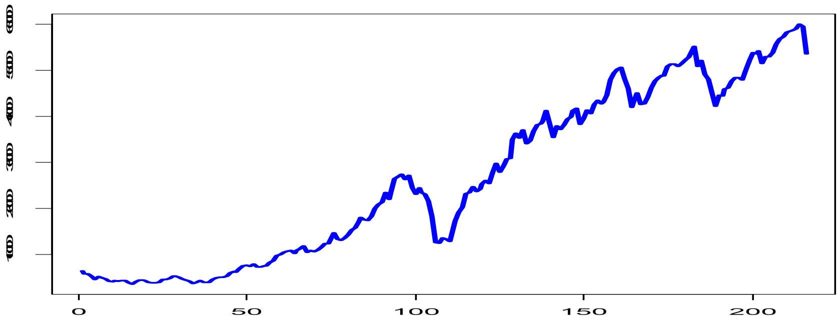

Figure 1. Plot of x

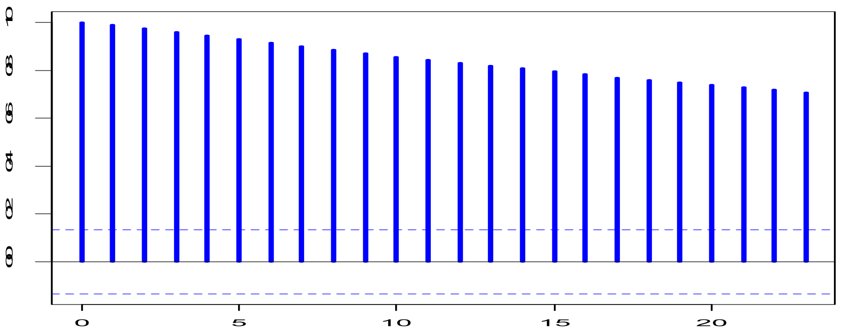

Figure 2. Plot ACF of x

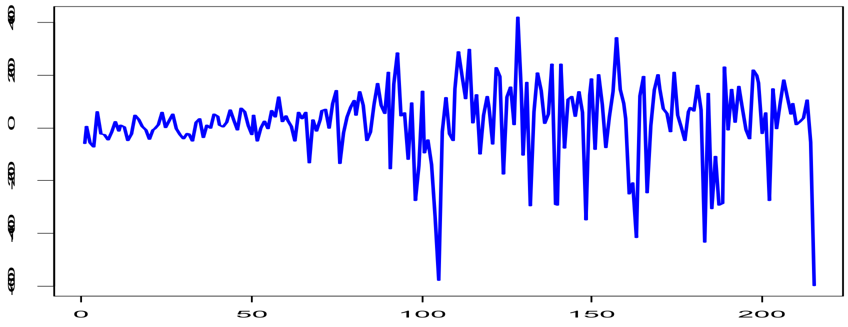

Figure3. Plot of y

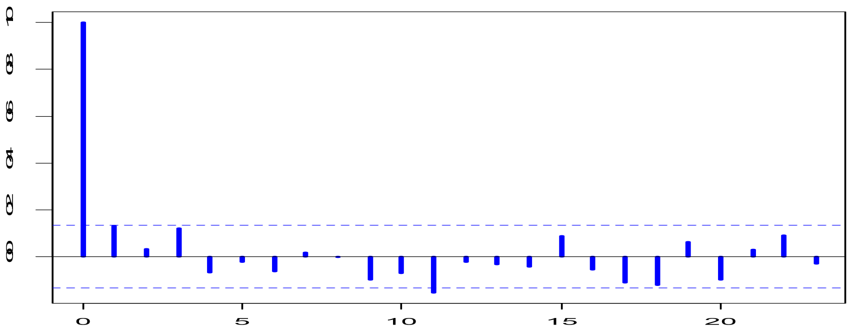

Figure 4. Plot ACF of y

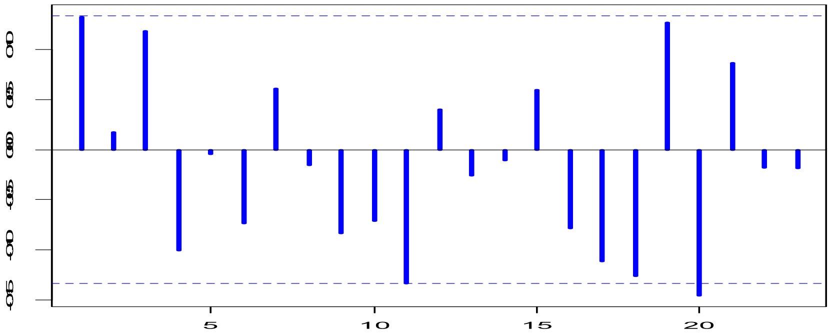

Figure 5. Plot PACF of y

3.1 Model Identification

According to the ACF plot in Figure 4 and the PACF plot in Figure 5, they are interrupted after lag 1, then the possible models for data on time series Y are ARMA (1, 0), ARMA (0, 1) or ARMA (1, 1).

3.2 Parameter Estimation and Model Selection

Furthermore, to estimate parameters, the maximum likelihood method is used, as shown in the following Table 2.

Table 2. The Values of Parameter and AIC

MODEL	$\hat{\phi}_1$	$\hat{\theta}$	AIC
ARMA(1,0)	0.1444	-	2745.32
ARMA(0,1)	-	0.1449	2745. 19
ARMA(1,1)	0.4770	−0.3424	2746. 92

The results in Table 2 show that the smallest AIC value is found in ARMA (0, 1) which reaches 2745.19. Therefore, it can be concluded that the most appropriate model for y data is the ARMA (0, 1) model.

3.3 Diagnostic Test

Furthermore, to find out the appropriate model, the diagnostic check is performed with the basic assumption that the residual is a white noise process; is a random variable that is not mutually correlated with zero mean and constant variance in Figure 6, by examining residual autocorrelation through hypotheses: Testing with Ljung-Box Q Statistics, with the calculation of Chi-squared=0.0633, df=1, p-value=0.8014 shows that Chi-squared<p-value, therefore, it can be concluded that the ARMA (0, 1) model is considered sufficient to be used as a forecasting model for the Indonesia composite stock price index.

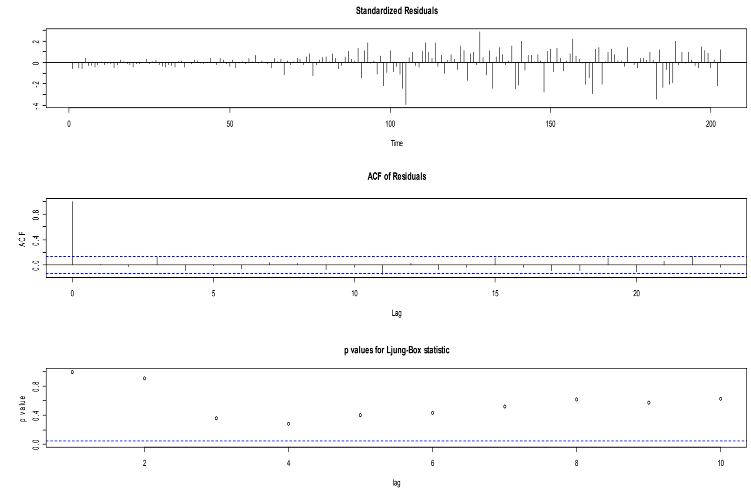

Figure 6. Plot of Residual Diagnostic Test

3.4 Model Accuracy

The measurements used to see the level of accuracy of the model are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), as in Table 3 below:

Table 3. The Values of RMSE, MAE and MAPE

MODEL	Accuracy		
	RMSE	MAE	MAPE
ARMA(1,0)	445.3735	383.0598	6.5621 %
ARMA(0,1)	444.3502	382.3258	6.5499 %
ARMA(1,1)	455.1833	392.7058	6.7291 %

Based on Table 4 it can be concluded that the best model for stationary data y is the ARMA model (0.1). As for the reason is because it has the smallest AIC value and the accuracy of RMSE, MAE and MAPE is also smallest compared to other models. If returned to the original data of x in equation (1), then the model is ARIMA (0, 1, 1) model as follows:

$$
\begin{aligned}
&\phi_0(B)(1-B)^1 Z^{\&}_t = \theta_1(B)a_t \\
&\Leftrightarrow\ 1(1-B)^1 x_t = (1+\theta_1 B)a_t \\
&\Leftrightarrow x_t - Bx_t = a_t + \theta_1 Ba_t \\
&\Leftrightarrow x_t - x_{t-1} = a_t + \theta_1 a_{t-1} \\
&\Leftrightarrow x_t = x_{t-1} + \theta_1 a_{t-1} + a_t
\end{aligned}
\tag{11}
$$

Furthermore, by substituting the parameter value to equation (11), the following model is obtained:

$$x_t = x_{t-1} + 0.1449\ a_{t-1} + a_t \tag{12}$$

4. Conclusions

Indonesia's Monthly Stock Price Index data taken from January 2000 to December 2017 is time series data which is not stationary, while the data on the only level difference is stationary. Therefore, the data analyzed is the one-level difference data. The model obtained was returned to the original data through the research stages on autocorrelation function, partial autocorrelation function, parameter estimation and diagnostic check with Ljung-Box Q statistics. It was concluded that the appropriate model was ARIMA (0, 1, 1) model.

References

Ashik, M. A., & Kannan, S. K. (2017). Forecasting Nifty Bank Sectors Stock Price Using ARIMA Model. *JCRT*, *5*(4), 1360-1365.

Assis, K., Amran, A., & Remali, Y. (2010). Forecasting Cocoa Bean Prices Using Univariate Time Series Models. *Journal of Arts & Commerce*, *I*(I), 71-80.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics* (2nd ed.). Duxbury Press, Belmont, California. https://doi.org/10.2307/2532587

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Enders, W. (1995). *Applied Econometric Time Series*. John Wiley & Son. Inc., New York.

Ihaka, R. (2005). *Time Series Analysis*. Statistics Department University of Auckland.

Jin, R., Wang, S., & Zhu, J. (2015). The Application of ARIMA Model in 2014 Shanghai Composite Stock Price Index. *Science Journal of Applied Mathematics and Statistics*, *3*(4), 199-203. https://doi.org/10.11648/j.sjams.20150304.16

Madsen, H. (2008). *Time Series Analysis*. Chapmann Hall, Informatics and Mathematical Modelling, Technical University of Denmark. https://doi.org/10.1201/9781420059687

Rossiter, D. G. (2013). *Time Analysis in R*. Department of Earth Systems Analysis, University of Twente, Facultyof Geo-Information Science & Earth, Observation (ITC), Enschede (NL).

Salam, M. A., Salam, S., & Feridun, M. (2007). *Modeling and Forecasting Pakistan's Inflationby Using Time Series ARIMA Models*. Paper.

Wei, W. W. S. (2006). *Time Series Analysis Univariate and Multivariate Methods*. Addison Wesley Publishing Company, Inc. Canada.