Discharge prediction of Amprong river using the ARIMA (autoregressive integrated moving average) model

Wiwin Sri Rahayu¹, Pitojo Tri Juwono², Widandi Soetopo²

¹Graduate Student of Water Resources Engineering, Brawijaya University, Malang, 65145, Indonesia
²Water Resources Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, 65145, Indonesia

E-mail: wiwinsrirahayu13@gmail.com

Abstract. An accurate determination of water availability of Amprong River has an important role in the planning system to support the agricultural production process in the Kedungkandang Irrigation Area, because if the availability of water is not precisely determined, there will be an error in regulating irrigation water. To overcome these problems, a good analysis system is needed. One of the time-series models is the ARIMA (Autoregressive Integrated Moving Average) model. The model was built by discharge data from 9 periods from 2008/2009 to 2016/2017, and its purpose was to predict the discharge of the 2017/2018 period. There were only five models feasible for use. The best model is the ARIMA model (2,0,1) (1,2,1) with values of MSE = 22.90; KR = 6.00; MSD = 8.05; MAD = 2.04; MAPE = 18.53; and MPE = -8.98.

Keywords: ARIMA, Discharge Prediction, Model, Time-Series.

1. Introduction

River discharge prediction is required in the application of hydrology, including management and planning of water resources. Information on the average discharge in a period represents the potential of water resources that can be utilized from a watershed, so the utilization plan must be arranged appropriately.

Accurate determination of water availability in the 10-day period of the Amprong River has a very important role in the planning of cropping patterns to support the process of agricultural production in the Kedungkandang irrigation area. If the availability of water is not precisely determined, there will be an error in the regulation of irrigation.

To overcome these problems, an analysis system is needed that is able to predict well. River discharge has a repetitive behaviour in the same period, and thus the creation of a time-series model allows the river discharge pattern to be represented in a mathematical formula. One of these time-series models is the ARIMA (Autoregressive Integrated Moving Average) model developed by Box and Jenkins (1976).

The pattern of river discharge data is often unclear, but with the ARIMA model, the pattern can be identified so that it can be used to forecast future patterns. The ARIMA model has a very good accuracy for short-term forecasting, but for long-term forecasting, its accuracy of forecasting is not good.
ARIMA is a black box model. This model is not for finding out the factors that influence a system. The system is merely considered as a process generator. The main purpose of this method is to predict what is coming, not knowing why it happens.

The ARIMA model is a model that completely ignores independent variables to make forecasts. ARIMA uses past and present values of the dependent variable to produce accurate short-term forecasting. ARIMA is suitable if observations from the time-series are statistically related to each other (dependent). Forecasting with the Box-Jenkins ARIMA method in general gives better results than other forecasting methods, because this method does not ignore rules in time-series data [1].

Nigam et al. (2009) stated that the ARIMA model is the right approach for hydrological data models, which often show auto-correlation with time and need an exact explanation of fundamental dynamics [2]. This is not possible with simple statistical forecasting methods such as regression analysis. Their predicted rainfall data showed a very good correlation with existing data. This showed that the chosen ARIMA model has a good level of trust [3]. Valipour et al. (2012) stated that the ARIMA model has better performance than the ARMA model because it results in a stationary time-series in both the modification and forecasting phases [4]. The ARIMA model could be used for forecasting monthly inflow discharge that is suitable for the next 12 months. The accuracy of both the ARMA and ARIMA models increased compared to previous studies, due to the increase in the number of autoregressive and moving average parameters in the model. The ARIMA method produced a better model than ARMA [5]. The results of analysis of the Sengguruh Reservoir operating pattern showed that the actual and the forecasted discharge did not have significant differences, which showed that the forecasting method with the ARIMA model is good enough for use [6].

The purpose of this study is to determine a prediction of the discharge of Amprong River in the next one-year period by using the ARIMA model, and then to compare it to the observed discharge.

2. Materials and Methods
The location of this study is the Kedungkandang Dam on the Amprong River in Kedungkandang Sub-District, City of Malang. The Amprong watershed is located in the City of Malang and Malang Regency, with an area of approximately 24,984 ha. The water from the Amprong River is utilized by the Kedungkandang to the irrigation area, which is 5,169 ha in size, through the Kedungkandang Dam. The study location is illustrated in Figure 1.

2.1. Materials

2.1.1. Data
The needed data is composed of discharge of the Kedungkandang Dam from 10 periods (2008/2009 to 2017/2018), which was obtained from the PSDA Technical Execution Unit in Malang, by the Department of Water Resource Public Works of the Province of East Java.

2.1.2. ARIMA Model
The ARIMA model is divided into 3 elements: the autoregressive (AR), moving average (MA), and integrated (I) models. These three elements are modified to form new models, for example the autoregressive and moving average (ARMA) model. The general form is ARIMA (p, d, q) where p represents the autoregressive order, d represents the integrated order, and q represents the moving average order.

Autoregressive means that the value of x is influenced by the x value of the previous period up to the p-period. Thus, what matters here is the variable itself. Moving average means that the value of the variable x is influenced by the error of the variable x. Integrated means that differences are stated from the data. This means that in the making of ARIMA models, the requirement that must be met is data stationarity. If the data are stationary at the same level, the order is 0, but if they are stationary at the first difference, the order is 1, and so on.
2.2. Analysis Method
2.2.1. Data Preparation
Data from 9 periods (2008/2009 to 2016/2017) were used in the forecasting process, while data for 2017/2018 were used for the calibration process or model reliability testing. To adjust the data to the Kedungkandang irrigation area cropping pattern, which is conducted within a period of 10 days, the average daily discharge data were converted to 10-day average discharge data.

2.2.2. Data Identification
10-day average discharge data patterns were determined by plotting graphs of data by using the Minitab 16 software worksheet.

After the data pattern was known, the data was then tested for stationarity. This is because the data must meet the conditions for the ARIMA model, in that the data must be stationary to the variance and average. The Box-Cox plot was carried out to test the stationarity of the data against variance. The data is stationary to the variance if the value of $\lambda = 1$. If the data was not stationary to the variance, then the data needed to be transformed to achieve stationarity against variance.

Stationary testing of the average was conducted if the data is stationary to the variance. The ACF (Auto Correlation Function) plot was performed to determine the stationarity of the data against the average. Data that are stationary to the average are marked with a lag that is not patterned (random) and does not contain seasonal elements. If the data was not stationary to the average, then the data needed to be made stationary by differentiating until stationary data was obtained.

2.2.3. Determining Temporary Models
At this stage, p, d, and q were determined. The process of determining p and q was assisted by an autocorrelation correlogram (ACF) and a partial autocorrelation correlogram (PACF). Meanwhile, d was determined from the stationarity level. ACF measures the correlation between observations with a lag of k while PACF measures the correlation between observations with a lag of k and controls the correlation between two observations with a lag less than k.
2.2.4. Determining the Final Model

The final model was determined by fitting the model evaluation criteria:

- Residual of forecasting must be random. To ensure that the model met this requirement, a Ljung-Box Statistics indicator was used. From this indicator, furthermore, it could be seen that the P-value for this statistical test is greater than 0.05, which indicated that the residual is random.
- The model must be in the simplest form (a parsimonious model)
- Conditions of invariability or stationarity must be met. This was indicated by the value of MA or AR coefficients, which must be less than 1.
- The model must have a small MS and SS.
- The ACF and PACF graphs of residuals showed a cut-off pattern, which meant that the residuals are random.

2.2.5. Forecasting

At this stage, the selected model was inputted in the Minitab 16 software, and then the forecasting process was carried out by the software.

2.2.6. Calibration/Model Reliability Testing

Discharge prediction results were compared with comparative data that had been prepared beforehand to determine the reliability of the model in forecasting. Calibration was performed by obtaining the values of MSD (Mean Squared Deviation), MAD (Mean Absolute Deviation), MAPE (Mean Absolute Percentage Error), MPE (Mean Percentage Error), and RE (Relative Error).

3. Results and Discussion

3.1. 10-Day Average Discharge

The 10-day average discharge of the Kedungkandang Dam that had been converted from the daily average discharge was plotted in a graph as presented in Figure 2, where the graph indicates a seasonal pattern. The data was not stationary on average and the data variance was too large. The data was also not stationary on variance. The stationarity of variance was proven from the Box-Cox plot, while the stationarity of the average was proven by the ACF plot.

![Figure 2](image)

Figure 2. Plot of the 10-Day Average Discharge Data of Kedungkandang Dam

3.2. Stationary Test for Variance

Figure 3 showed that the 10-day average discharge data of Kedungkandang Dam, after being tested by the Box-Cox plot, had a value of $\lambda \neq 1$, which means that the data was not stationary to the variance. The Box-Cox transformation was then performed until the value of $\lambda = 1$ was obtained. The results of the first transformation, as shown in Figure 4, showed that after the first transformation, the data became stationary to the variance, indicated by the value of $\lambda = 1$. 
3.3. Stationary Test for the Average
Stationary testing of the average was conducted when the data became stationary to the variance. The ACF plot in Figure 5 shows that the data was not stationary to the average, characterized by lags that were still patterned (not random) and contained seasonality. Therefore the data needed to become stationary through the first differentiating. Another ACF plot was created from the data that had been differentiated once. Figure 6 shows that the data was still not stationary to the average, and thus a second differentiating needed. Figure 7 shows that the ACF plot of the data that had been differentiated twice became stationary to the average, which is shown by irregular patterns.
3.4. **ARIMA Model (p,d,q)(P,D,Q)S**

From the ACF plot after the second differentiating in Figure 7, there was a cut-off after lag 1, and the result was that the obtained tentative model for non-seasonal MA is $q = 1$ and tentative model for seasonal MA is $Q = 1$. Because the differentiating process was performed twice, the order of $D = 2$. Seasonal cut-off lag occurred at 36th lag, the order of $S = 36$. A PACF plot was created to determine the orders of $p$ and $P$ of the AR model. The PACF plot in Figure 8 shows that the observation points of the average 10-day discharge of the Kedungkandang Dam died down. The tentative model has non-seasonal AR orders of $p = 1, 2, 3, 4, \text{ and } 5$, while the seasonal AR order is $P = 1$. 

![Figure 6. ACF Plot of First Differentiating](image)

![Figure 7. ACF Plot of Second Differentiating](image)

![Figure 8. PACF Plot of Second Differentiating](image)
From these stages, the values of each parameter of the ARIMA model was obtained, which then resulted in several tentative models, as presented in Table 1.

The feasibility test of the models in Table 1 was carried out using the Ljung-Box statistical test. The model is not feasible if the p-value is less than 0.05 on one or all of the lags. The results of the feasibility test of the model showed that there were only five tentative models that could be used.

The best model was selected at the model calibration stage. Model calibration was performed by comparing the discharge data from the ARIMA forecasting method to the existing discharge data.

3.5. Best Model Selection

The relative error of a feasible ARIMA model was used to choose the best model. The best ARIMA model was chosen from the model that had the smallest relative error. Relative error was calculated by comparing the model discharge with historical discharge.

From the historical discharge and the ARIMA model forecasting discharge, the Relative Error (RE) of each model was then calculated. This is the percentage ratio of the difference between the volume of the model and the historical volume, compared to the historical volume. Relative Error (RE) was used to select the best model from the five models that met the Ljung-Box test statistical requirements.

Table 1. Tentative ARIMA Model (p,d,q)(P,D,Q)^S

| No | Tentative ARIMA Model | p-value at lag | Conclusion |
|----|-----------------------|----------------|------------|
|    |                       | 12  | 24  | 36 | 48 |           |
| 1  | (1,0,1)(0,2,1)^Sb     | 0.166 | 0.335 | 0.000 | 0.000 | Model is not feasible |
| 2  | (1,0,1)(1,2,1)^Sb     | 0.965 | 0.898 | 0.692 | 0.857 | Model is feasible |
| 3  | (2,0,1)(0,2,1)^Sb     | 0.151 | 0.349 | 0.000 | 0.000 | Model is not feasible |
| 4  | (2,0,1)(1,2,1)^Sb     | 0.972 | 0.899 | 0.639 | 0.824 | Model is feasible |
| 5  | (3,0,1)(0,2,1)^Sb     | 0.098 | 0.301 | 0.000 | 0.000 | Model is not feasible |
| 6  | (3,0,1)(1,2,1)^Sb     | 0.947 | 0.826 | 0.545 | 0.777 | Model is feasible |
| 7  | (4,0,1)(0,2,1)^Sb     | 0.228 | 0.600 | 0.000 | 0.000 | Model is not feasible |
| 8  | (4,0,1)(1,2,1)^Sb     | 0.903 | 0.821 | 0.452 | 0.712 | Model is feasible |
| 9  | (5,0,1)(0,2,1)^Sb     | 0.674 | 0.853 | 0.000 | 0.000 | Model is not feasible |
| 10 | (5,0,1)(1,2,1)^Sb     | 0.811 | 0.748 | 0.416 | 0.684 | Model is feasible |

Table 2. Summary of ARIMA Modelling Test Results

| ARIMA Model | MSE | RE | MSD | MAD | MAPE | MPE |
|-------------|-----|----|-----|-----|------|-----|
| (1,0,1)(1,2,1)^Sb | 23.05 | 6.62 | 8.54 | 2.11 | 18.89 | -9.31 |
| (2,0,1)(1,2,1)^Sb | 22.90 | 6.00 | 8.05 | 2.04 | 18.53 | -8.98 |
| (3,0,1)(1,2,1)^Sb | 23.12 | 5.86 | 8.58 | 2.08 | 18.86 | -8.99 |
| (4,0,1)(1,2,1)^Sb | 24.01 | 7.27 | 11.65 | 2.39 | 21.32 | -10.38 |
| (5,0,1)(1,2,1)^Sb | 24.22 | 7.03 | 11.79 | 2.39 | 21.38 | -10.19 |

Furthermore, testing of the accuracy of the model was performed by calculating the values of MSD (Mean Squared Deviation), MAD (Mean Absolute Deviation), MAPE (Mean Absolute Percentage
Error), MPE (Mean Percentage Error), and RE (Relative Error). Table 2 presents a summary of the ARIMA model test results.

The (2,0,1)(1,2,1)\textsuperscript{36} ARIMA model was chosen as it had the lowest MSE, MSD, MAD, MAPE, and MPE values. The model with the lowest RE value is the (3,0,1)(1,2,1)\textsuperscript{36} ARIMA model, which is 5.86%, but this is not much different from the RE value for the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA model, which is 6.00%. The chosen model was therefore the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA model. The parameters of the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA model obtained from Minitab software output are shown in Figure 9.

Table 2. Summary of the ARIMA Model Test Results

| Period     | ARIMA Discharge Model (m³/s) | Period     | ARIMA Discharge Model (m³/s) |
|------------|-----------------------------|------------|-----------------------------|
| Nov 2017   |                             | May 2018   |                             |
| 1          | 9.00                        | 1          | 13.99                       |
| 2          | 8.80                        | 2          | 10.55                       |
| 3          | 8.86                        | 3          | 9.98                        |
| Dec 2017   |                             | Jun 2018   |                             |
| 1          | 12.68                       | 1          | 11.37                       |
| 2          | 7.23                        | 2          | 12.43                       |
| 3          | 11.48                       | 3          | 14.76                       |
| Jan 2018   |                             | Jul 2018   |                             |
| 1          | 8.84                        | 1          | 10.27                       |
| 2          | 17.99                       | 2          | 6.20                        |
| 3          | 19.02                       | 3          | 5.58                        |
| Feb 2018   |                             | Aug 2018   |                             |
| 1          | 25.16                       | 1          | 5.06                        |
| 2          | 18.54                       | 2          | 8.74                        |
| 3          | 29.31                       | 3          | 4.85                        |
| Mar 2018   |                             | Sep 2018   |                             |
| 1          | 16.36                       | 1          | 6.20                        |
| 2          | 12.95                       | 2          | 5.46                        |
| 3          | 12.88                       | 3          | 6.91                        |
| Apr 2018   |                             | Oct 2018   |                             |
| 1          | 12.98                       | 1          | 9.35                        |
| 2          | 17.53                       | 2          | 8.80                        |
| 3          | 15.43                       | 3          | 6.74                        |

Figure 9. Output Parameters of the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA Model

The following is thus the form of the obtained mathematical equation for the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA model:

\[ Z_t = 0.5136 Z_{t-1} + 0.1237 Z_{t-2} + 1.3486 Z_{t-36} - 0.6926 Z_{t-37} - 0.0375 Z_{t-74} - 0.6514 Z_{t-108} + 0.3346 Z_{t-109} + 0.0516 \alpha_{t-37} \]

3.6. Forecasting of Discharge for the 2017/2018 Period

Discharge forecasting was performed using the selected best ARIMA model. The model was utilized to predict the discharge for the next period.

Discharge forecasting used the Minitab 16 software. Data input in the Minitab 16 software consisted of historical discharge data of 9 periods, from 2008/2009 to 2016/2017. Forecasting results with the best ARIMA model are presented in Table 3 and Figure 10.

Table 3. Forecasting Results of the (2,0,1)(1,2,1)\textsuperscript{36} ARIMA Model
Figure 10. Plot of Forecasting Results of the (2,0,1)(1,2,1)36 ARIMA Model

4. Conclusion
There were 5 ARIMA models that are feasible to predict the discharge of Amprong River. Of the five feasible models, the best model selected model is the (2,0,1)(1,2,1)36 ARIMA model, which is used to predict the discharge of the Amprong River for the 2017/2018 period using the Minitab 16 software.

From the results of the accuracy testing of the model, it could be seen that the discharge prediction using the (2,0,1)(1,2,1)36 ARIMA model was not very different from the observed discharge in the period of 2017/2018, and thus the (2,0,1)(1,2,1)36 ARIMA model can be applied in predicting the discharge of the Amprong River for the period of 2017/2018.

References
[1] Mulyana, 2004. Analisis Data Deret Waktu [Time-Series Data Analysis], (Buku Ajar), Universitas Padjadjaran, Jawa Barat.
[2] Nigam, Rashmi et al. 2009. Time Series Modeling and Forecast of River Flow. Journal of Current World Environment, Vol. 4 (1), 2009, 79-87.
[3] Zakaria, Saleh et al. 2012. ARIMA Models for Weekly Rainfall in the Semi-arid Sinjar District at Iraq. Journal of Earth Sciences and Geotechnical Engineering, Vol. 2, No. 3, 2012, 25-55.
[4] Valipour, Mohammad et al. 2012. Parameters Estimate of Autoregressive Moving Average and Autoregressive Integrated Moving Average Models and Compare Their Ability for Inflow Forecasting. Journal of Mathematics and Statistics, Vol. 8 (3), 2012, 330-338.
[5] Valipour, Mohammad et al. 2013. Comparison of The ARMA, ARIMA and the Autoregressive Artificial Neural Network Models in Forecasting The Monthly Inflow of Dez Dam Reservoir. Journal of Hydrology, Vol. 476, 2013, 433-441.
[6] Hanggara, Ikrar et al. 2015. Analisa Peramalan Debit Sungai Menggunakan Metode ARIMA (Auto Regressive Integrated moving Average) di Sungai Brantas Hulu [Analysis of River Discharge Forecasting Using the ARIMA (Autoregressive Integrated Moving Average) Method at Brantas Hulu River]. Jurnal Teknik Pengairan, Volume 6, Nomor 2, Desember 2015, Halaman 197-205.