Forecasting time series data containing outliers with the ARIMA additive outlier method

Lilis Laome\textsuperscript{1*}, Gusti Ngurah Adhi Wibawa\textsuperscript{1}, Rasas Raya\textsuperscript{1}, Makkulau\textsuperscript{1} and Abdul Rahman Asbahuna\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, Universitas Halu Oleo, Indonesia
\textsuperscript{2}Student of Department of Mathematic, Universitas Halu Oleo, Indonesia

ilhis2laome@gmail.com (corresponding author)

Abstract. ARIMA method is often used in time series data forecasting. However, when an outlier occurs or there is an observation value that is far from a set of data, forecasting with this method will provide a large residual so that the normality assumption is not met. One method developed from the ARIMA Method that overcomes outliers, is called the ARIMA Additive Outlier Method (ARIMA AO). The purpose of this study is to forecast time series data containing outliers with the ARIMA AO Method. This method can reduce the presence of outliers with iterative procedures. The data used is data on the number of foreign tourists visiting the Port of Tanjung Priok. The results obtained are that the model has a normal distribution residual and forecasting accuracy value with MSE and MAPE is smaller than the ARIMA.

1. Introduction

Time series analysis is one of the statistical procedures that is applied to predict the probability structure of future conditions in the framework of decision making. Time series modeling is often associated with the process of forecasting a certain characteristic value in the coming period. Forecasting is guessing or estimating a situation in the future based on the past and present conditions that are needed to determine when an event will occur, so that appropriate action can be taken [1].

In general, time series forecasting is done using the Autoregressive Integrated Moving Average (ARIMA) method, exponential smoothing, decomposition or regression. Although this kind of approach is efficient for time series forecasting, it still shows shortcomings when noise disturbances or extreme fluctuations occur. Extreme fluctuating data can indicate an intervention or outlier. These extreme fluctuations can be caused by various factors, both external and internal factors, such as natural disasters, government regulations, economic stability, riots, and terrorism.

The existence of outliers often makes forecast values far from the original data or has a very large residual value, therefore procedures are needed to detect and reduce the effect of outliers [2]. Chang et al [3] developed a method to detect the presence of outliers in time series data through the iterative detection method. These extreme data can be assessed by ARIMA Box-Jenkins analysis and outlier detection by iterative procedures.

Outliers are observations that are clearly different from other observations [6]. Outliers cause conclusions from the analysis of the resulting data is invalid, so the procedure of detecting and reducing the effects of outliers is very important in data analysis. In time series data, outlier data are
classified into Innovation Outlier (IO), Additive Outlier (AO), Temporary Change (TC), and Level Shift (LS). Additive Outlier (AO) only affects the time $T$, while three other types of outliers, namely Innovation Outlier (IO), Level Shift (LS) and Temporary Change (TC) affect $T-T$, $T+1$, ... and so on. AO is a frequent occurrence and is often found in time series data compared to other types of outliers.

Research on ARIMA AO has been carried out [4,5], but [4] is limited to simulation data, [5] the application of inflation data that contains many outliers. This study limits the incidence of outliers of less than 10 and their application to data on the number of foreign tourists at the Tanjung Priok Port in Jakarta.

2. ARIMA Additive Outlier Model
ARIMA Additive Outlier Method (ARIMA AO) is the development of the ARIMA Method when an outlier occurs at a particular time. In general, the AO Model can be written as:

$$\hat{Z}_t = Z_t + \omega I_t^{(T)}$$

where:

$$I_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$$

$I_t^{(T)}$ is the indicator variable AO which represents the presence or absence of outliers at time $T$.

$$e_t \sim N(0, \sigma^2)$$

$\omega$ is a type magnitude estimator of AO ($\hat{\omega}_{AT}$),

$$\hat{\omega}_{AT} = \frac{e_t - \sum_{j=1}^{T} \pi_j e_{t+j}}{\sum_{j=0}^{T-1} \pi_j^2}$$

3. Data and Analysis
The data used in this study is secondary data regarding the number of foreign tourist arrivals at the Port of Tanjung Priok (January 2008 to December 2017), taken from [9]. The variable used is the number of tourists per month ($Z_t$) for 120 months ($n$). The steps in modeling ARIMA AO with iterative procedures are:

- **Time series model ($Z_t$)** assumed to be free outlier, then calculate residuals($\hat{e}_t$) from the estimation model:

$$\hat{e}_t = \hat{\pi}(B) Z_t - \frac{\hat{\phi}(B) \pi_B}{\hat{\theta}(B)} Z_t$$

and $\sigma^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{e}_t^2$

- Calculate $\hat{\gamma}_1^{(T)}$ dan $\hat{\gamma}_2^{(T)}$ for $t=1,2,...,n$ using estimation models, then specify $\hat{\gamma}_T = \max_{1 \leq t \leq n} \left\{ \hat{\gamma}_1^{(T)} \right\}$ where $T$ is the time when the maximum value occurs. Outlier of AO happens if $\hat{\gamma}_T = |\hat{\gamma}_1^{(T)}| > C$, where $C=3$, then $Z_t$ is modified using $\hat{Z}_t = Z_t - \hat{\omega}_{AT} I_t^{(T)}$ then calculate the new residual with

$$\hat{e}_t = \hat{e}_t - \hat{\omega}_{AT} \hat{\pi}(B) I_t^{(T)}$$

- Recalculate $\hat{\gamma}_1^{(T)}$ dan $\hat{\gamma}_2^{(T)}$ from $Z_t$ and modified residuals then calculate $\hat{\sigma}^2$, after that repeated step 2 until all outlier are detected. The initial estimation of $\pi(B)$ is not changed [8].

Furthermore, data analysis is performed on foreign data. Outliers are a problem that often occurs in time series data. This problem greatly affects the analysis of data related to the mean and standard deviation. Outliers can cause data intervals to have wide ranges, large data variances and residuals generated in the modeling are also large. There are several methods that can be used to detect outliers,
including: scatter diagrams, box plots, and standardized residual values. Following are the results of checking the data on the number of tourists using scatter plot and boxplot methods.

**Figure 1.** Outlier detection on foreign tourist data

Based on Figure 1, there can be seen some data that is far from a set of data, and detected to contain outliers. The next step is ARIMA and ARIMA AO modeling. The ARIMA AO model was carried out after obtaining some initial ARIMA models. In general, ARIMA modelling consists of four steps, namely identification of the initial model, parameter estimation, diagnostic testing, and forecasting. The results obtained in ARIMA modelling are as in Table 1.

**Table 1.** The Summary of The Initial ARIMA Model

| ARIMA Model | Parameter Significance | White Noise Distribution | Residual Normality Test | Mean Square Error (MSE) |
|-------------|------------------------|--------------------------|-------------------------|------------------------|
| (2,1,1)     | No                     | Yes                      | No                      | 559182                 |
| (2,1,0)     | Yes                    | No                       | No                      | 637769                 |
| (0,1,1)     | Yes                    | Yes                      | No                      | 569140                 |

In Table 1, there are three initial models obtained and all models have not fulfilled the normal distribution of residual assumptions. This is because the data still contains outliers. Next is the ARIMA AO modelling. The ARIMA model that allows to do ARIMA AO modelling is ARIMA (0,1,1), because the MSE value is small. The results of the iteration of the ARIMA AO model are presented in Table 2.

**Table 2.** Modelling of ARIMA (0,1,1) with Additive Outlier

| Iteration Stage | Additive Outlier  | Parameter Significance | White Noise Distribution | Residual Normality Test | Mean Square Error (MSE) |
|-----------------|-------------------|------------------------|--------------------------|-------------------------|------------------------|
| Iteration I     | AO75, AO85, AO23  | Yes                    | No                       | Yes                     | 332983                 |
| Iteration II    | AO86              | Yes                    | No                       | Yes                     | 304220                 |
| Iteration III   | AO47, AO98       | Yes                    | Yes                      | Yes                     | 246238                 |
From Table 2 ARIMA AO modeling was carried out 3 times iteration, so that in the 3rd iteration there were no outliers and had fulfilled all the assumptions namely parameter significance, white noise residuals, normal distributed residuals and MSE values were smaller compared to the two previous iterations. Thus, the best ARIMA model (0,1,1) was obtained with the addition of 6 outliers as follows:

\[
\hat{Z}_t = Z_{t-1} - 0.7980e_{t-1} + 3377.278I_t^{(75)} + 2885.711I_t^{(85)} - 2353.5I_t^{(23)} + 1806.23I_t^{(86)} + 1723.76I_t^{(47)} + 1751.12I_t^{(98)}
\]

with MSE and MAPE decreasing respectively to 242544 and 7.37%.

From Figure 2 it can be seen that ARIMA AO generates an estimated value close to real data. Then forecasting is done in one period to the next, obtained forecasting the number of foreign tourists in January 2018 amounted to 5402 visitors.

4. Conclusion

Based on the analysis using the ARIMA method for data on foreign tourists at the Port of Tanjung Priok, it is assumed that the ARIMA model (0,1,1) with the equation:

\[
\hat{Z}_t = Z_{t-1} - 0.9665e_{t-1}
\]

with MSE and MAPE values are respectively 569140 and 10.05%. Whereas with the ARIMA AO method, the ARIMA + 6 Outlier estimation model is obtained by the equation:

\[
\hat{Z}_t = Z_{t-1} - 0.7980e_{t-1} + 3377.278I_t^{(75)} + 2885.711I_t^{(85)} - 2353.5I_t^{(23)} + 1806.23I_t^{(86)} + 1723.76I_t^{(47)} + 1751.12I_t^{(98)}
\]
with MSE of 242544 and MAPE of 7.37%. This can be seen with the ARIMA AO method has a greater forecasting accuracy than the ARIMA method. The forecasting of foreign tourist data for January 2018 was 5402 visitors.

References
[1] Makridakis, 2003. Metode dan Aplikasi Peramalan. Jilid 1. Edisi Revisi. Jakarta: Binarupa Aksara.
[2] Wei. 2006. Time Series Analysis, Univariate and Multivariate Methods. United States of America: Pearson Education Inc.
[3] Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. Technometrics 3, 193-204
[4] Ansari, S. A., 2018. Modeling Data Containing Outlier using ARIMA Additive Outlier (ARIMA AO). IOP Conf. Series: Journal of Physics 954 (2018) 012010.
[5] Suparti and Alfi Faridatus Sa’diyah. 2015. Analisis Data Inflasi Indonesia Menggunakan Model Autoregressive Integrated Moving Average (ARIMA) dengan Penambahan Outlier. Media Statistika, Vol.8 No.1, Juni 2015 : 1-11.
[6] Soemartini. 2007. Outlier (Pencilan). Jatinangor: Penerbit Universitas Padjadjaran Bandung.
[7] Aswi and Sukma. 2006. Analisis Deret Waktu: Teori dan Aplikasi. Makassar: Andira Publisher.
[8] Kusman, S. 2008. Pendeteksian Pencilan Aditif dan Inovatif dalam Data Deret Waktu melalui Metode Iteratif. Jurnal Forum Statistika dan Komputasi, Vol.13 No. 2. ITB, Bogor.
[9] Badan Pusat Statistik. (2018). Wisatawan Mancanegara di Indonesia. www.bps.go.id.