Comparison of SSA and SARIMA in Forecasting the Rainfall in Sumatera Barat

F Fitri1, R Rahmat2

1Statistics Department, Universitas Negeri Padang, Jalan Prof. Dr. Hamka Padang, 25171, Indonesia
2Climate Variability Division, Indonesia Agency for Meteorology, Climatology and Geophysics (BMKG), Jalan Angkasa 1 Jakarta, 10610, Indonesia

corresponding author: fadhilahfitri@fmipa.unp.ac.id

Abstract. Rainfall data in West Sumatra in 2001–2018 is assumed to have a seasonal pattern. Several methods can be used to model this data, including SSA and SARIMA. Each method has a different approach but has its own advantages. Therefore, a comparison will be made between the two based on the MAPE. According to the forecasting result, obtained MAPE for SSA is 17% and 22.75% for SARIMA.

1. Introduction

Indonesia has two seasons, dry and rainy season. The difference between these two seasons is the intensity of rainfall [1]. In the dry season, it is still possible to rain, while in the rainy season, the amount of water will be abundant so that the rainwater must be managed. Hence, it is necessary to do forecasting to find out how much rainfall intensity in the next several periods so that rainwater management facilities can be better prepared.

Several methods can be used to predict rainfall. Two of them are the Singular Spectrum Analysis (SSA) method and the Seasonal Autoregressive Integrated Moving Average (SARIMA). The SSA method is a flexible method because it uses a non-parametric approach. In performing the analysis, SSA does not make any statistical assumptions regarding signal or noise [2]. Meanwhile, SARIMA is a method used when it is known that seasonality is present in time series data. This method was popularized by George Box and Gwilym Jenskin around the 1970s. The SARIMA model is an expansion model of the Autoregressive Integrated Moving Average (ARIMA) model.

One comparative study between SSA and other methods has been performed previously (see [3]) comparing SSA, SARIMA, ARAR, and Seasonal Holt-Winter using monthly accidental deaths in the USA. The comparison shows that SSA is more accurate for these data. In this study, SSA and SARIMA will be compared to determine which method is more suitable in predicting rainfall data in West Sumatra from 2001 to 2018 obtained from BMKG (Minangkabau Meteorological Station). Comparisons are made by looking at the accuracy of the forecast results. Research on rainfall forecasts in West Sumatra using SSA has been conducted before (see [4]). So in this article, the forecasting process using SSA will not be discussed in detail.

2. Singular Spectrum Analysis (SSA)

Singular Spectrum Analysis is a popular method used to predict time series data in recent years because of its ability to decipher time series data patterns in a form that is considered good enough in...
producing forecast data [5]. Deciphering time series data is a basic concept of SSA. The SSA method involves two complementary stages, decomposition and reconstruction; each of the stages have two separate steps. Decomposition consists of two-step: embedding and singular value decomposition, whereas reconstruction divided into grouping and diagonal averaging step. Time series data decomposition is carried out to facilitate the interpretation of the characteristics of the time series data then make a reconstruction of the data to be used for forecasting data points in the next several periods.

SSA will decompose the time series data into a data summary so that each component contained in the data summary can be identified as either trend, periodic, or noise. Then performed a reconstruction of the actual time series data. After the reconstruction phase is carried out, it is followed by forecasting using the reconstructed data to obtain new data values. The rainfall forecast, as in [4], can be seen as follows

| Table 1. Forecast Results using SSA |
|-------------------------------------|
| Period (in 2019) | Rainfall Forecast |
|-----------------|------------------|
| July            | 335.3            |
| August          | 394.9            |
| September       | 320.3            |
| October         | 294.1            |
| November        | 306.6            |

3. Seasonal Autoregressives Integrated Moving Average (SARIMA)

In general, Seasonal ARIMA or ARIMA (p, d, q)(P, D, Q) has a mathematical model as can be seen below [6]:

\[
\phi(B)\Phi(B^s)Y_t = \Theta(B)\Theta(B^s)Z_t
\]

Where:

\[
\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p
\]

\[
\Phi(z) = 1 - \Phi_1 z - \Phi_2 z^2 - \ldots - \Phi_P z^P
\]

\[
\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \ldots - \theta_q z^q
\]

\[
\Theta(z) = 1 - \Theta_1 z - \Theta_2 z^2 - \ldots - \Theta_Q B^Q
\]

\[
Y_t = (1 - B)^d (1 - B^s)^D X_t
\]

The forecasting process using SARIMA method begins by creating a time series plot for the data.

![Plot of Rainfall Data Recorded at Stasiun Minangkabau Station](image)

**Figure 1.** The plot of Rainfall Data Recorded at Minangkabau Station
Based on Figure 1, it can be seen that the monthly rainfall at Minangkabau Station, West Sumatra has a seasonal pattern, where in several periods there is a significant increase in rainfall. Then proceed to the next step, as follows:

3.1. Perform stationarity testing on data

A data set is called stationary if the mean and variance of the data are constant. Data stability is related to the estimation method used. If the data is not stationary, the model to be estimated will be less good. Stationarity test was performed on mean and variance.

3.1.1. Stationarity in variance

The method to be used is the Box-Cox transformation. The transformation is defined [6] as:

\[ T(X_i) = \begin{cases} \frac{X_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln X_i, & \lambda = 0 \end{cases} \]

\( \lambda \) is called the transformation parameter. The type of transformation used depends on the estimated value \( \lambda \). The commonly used values of \( \lambda \) and their associated transformations are shown in Table 2 [7].

**Table 2.** Box-Cox Transformation Type

| \( \lambda \) | Transformation |
|---------------|---------------|
| -1.0          | \( \frac{1}{X_i} \) |
| -0.5          | \( \sqrt{X_i} \) |
| 0.0           | \( \ln X_i \) |
| 0.5           | \( \sqrt{X_i} \) |
| 1.0           | \( Z_i \) |

In this study obtained \( \lambda \) of 0.54, so that the \( \sqrt{X_i} \) transformation is performed on the data.

3.1.2. Stationarity in mean

The test to be performed is the Augmented Dickey Fuller (ADF) Test. The hypothesis for this test:

- \( H_0: \delta = 0 \), the data is nonstationary
- \( H_1: \delta < 0 \), the data is stationer

\[ \tau = \frac{\tilde{\delta} - \bar{\delta}}{SE(\tilde{\delta})} \]

Test criteria: Reject \( H_0 \) if \( |\tau| \geq |df| \), accept in other ways.

If the data is not stationary in mean, then do the differencing process. This method is done by subtracting data in a period with data in the previous period. The differencing process is performed until stationary data is obtained. In general, if there is differencing for the \( d \) period to achieve stationarity, it can be stated by

\[ X_i^d = (1 - B)^d X_i \]

The Dickey-Fuller statistic value equals to -7.0154 with p-value 0.01 therefore \( H_0 \) is rejected. The rainfall data is stationary.

3.2. Model Identification

Model identification is done by looking at the ACF and PACF plots from stationary data. The ACF and PACF for this data can be seen in Figure 2.
The ACF dan PACF plots of Rainfall in West Sumatra

Beside that, model identification can be done instantly by using “auto.arima” function in R Software. It will give several option for the model, one best temporary model will be selected. The choosen one is ARIMA (0,0,0)(1,0,1)_{12}.

3.3. Parameter Significance

T-test will be used in testing the significance of the parameters [8].

\[ t = \frac{\hat{\delta}}{SE(\hat{\delta})} \]

The result are as follow

**Table 3. The Result of the Parameter Significance Test**

| Parameter | p-value  |
|-----------|----------|
| \( \Phi_1 \) | < 2.2^{-16} |
| \( \Theta_1 \) | < 2.2^{-16} |
| intercept | < 2.2^{-16} |

From Table 3, it can be seen that all the parameters for ARIMA (0,0,0)(1,0,1)_{12} are significant.

3.4. Diagnostic Check

In this step, an examination of the white noise assumption will be carried out, the fulfillment of the distribution normality in the residuals, and the overfitting model.

3.4.1. White Noise

Residual is said to be white noise when it fulfills two characteristics, identical, where the variance is constant, and independent, meaning that the residuals are not correlated with ,mean equal to zero. The statistical test used to test the white noise assumption is the Ljung-Box statistic with the formula [7]:

\[ Q = n(n + 2) \sum_{k=1}^{K} \frac{\hat{\rho}_k^2}{n-k} \]

The result are as follow

**Table 4. The Result of the Parameter Significance Test**

| Number of lag | p-value |
|---------------|---------|
| 5             | 0.60    |
| 10            | 0.89    |
| 12            | 0.92    |
| 15            | 0.93    |

Table 4 above shows the p-value of the white noise assumption test for several lags, namely lag 5,10,12, and 15. Based on testing, all p-values are greater than alpha 0.05, so it can be concluded that
there is no autocorrelation in the residuals. After obtaining white noise residuals, then we check whether the residuals are normally distributed or not.

### 3.4.2. Normality

The test used to test whether the residuals are normally distributed is Kolomogorov Smirnov Test [9]

\[
D = \max \left| F_0(X) - S_N(X) \right|
\]

with:

- \(F_0(X)\): the theoretical cumulative distribution under \(H_0\) whose value is the proportion of cases expected to have score equal to or less than \(X\)
- \(S_N(X)\): the observed cumulative frequency distribution of \(N\) observations

Test criteria: Reject \(H_0\) if \(D \geq D_{\text{table}}\), accept in other ways

It turns out that the statistical value \(D = 0.045146\) with \(p\)-value = 0.7708, then \(H_0\) is accepted. It means that the residuals are normally distributed.

### 3.4.3. Overfitting

Overfitting is one part of the diagnostic check, for example using more parameters than needed [10]. This procedure is done by increasing or decreasing the value of the parameter. Here we will use the possible models obtained from the auto.arima function so that there are 10 models to be compared as can be seen in Table 5

#### Table 5. The Obtained Models

| ARIMA Models                        |  
|-------------------------------------|
|-------------------------------------|
| (2,0,2)(1,0,1)\(^{12}\)            |
| (1,0,0)(1,0,0)\(^{12}\)            |
| (0,0,1)(0,0,1)\(^{12}\)            |
| (0,0,0)(1,0,0)\(^{12}\)            |
| (0,0,0)(0,0,1)\(^{12}\)            |
| (0,0,0)(1,0,1)\(^{12}\)            |
| (1,0,0)(0,0,1)\(^{12}\)            |
| (1,0,0)(1,0,1)\(^{12}\)            |
| (0,0,1)(1,0,0)\(^{12}\)            |
| (0,0,1)(1,0,1)\(^{12}\)            |

### 3.5. Forecasting Accuracy

When analyzing the errors of the forecasting method used, it is best if the MAPE is calculated because it is easier to interpret the value, which is a percentage of the whole. Calculation of the Mean Absolute Percentage Error (MAPE) value can be done using the following formula [10]

\[
MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{X_i - F_i}{X_i} \right|
\]

With

- \(X_i\): actual data
- \(F_i\): forecasting result

Based on the calculation of the 10 models in Table 5, the following results were obtained
Table 6. MAPE of the Models

| ARIMA Models          | MAPE (%) |
|-----------------------|----------|
| (2,0,2)(1,0,1)$^{12}$ | 23.03    |
| (1,0,0)(1,0,0)$^{12}$ | 29.90    |
| (0,0,1)(0,0,1)$^{12}$ | 30.00    |
| (0,0,0)(1,0,0)$^{12}$ | 29.91    |
| (0,0,0)(0,0,1)$^{12}$ | 29.71    |
| (0,0,0)(1,0,1)$^{12}$ | 22.75    |
| (1,0,0)(0,0,1)$^{12}$ | 29.97    |
| (1,0,0)(1,0,1)$^{12}$ | 23.16    |
| (0,0,1)(1,0,0)$^{12}$ | 29.91    |
| (0,0,1)(1,0,1)$^{12}$ | 23.00    |

In Table 6, it can be seen that the model that has the highest level of accuracy is the ARIMA (0,0,0)(1,0,1)$^{12}$ with MAPE of 22.75%.

3.6. Forecasting for the Period Ahead

The next step is forecasting the rainfall in West Sumatra for July to November 2019 (as in [4]). The result as in Table 7 follows

Table 7. Forecast Results using SARIMA

| Period (in 2019) | Rainfall Forecast |
|-----------------|------------------|
| July            | 277.1            |
| August          | 306.6            |
| September       | 330.4            |
| October         | 384.3            |
| November        | 445.4            |

4. Comparison of SSA and SARIMA

SARIMA is a time series data analysis method used for data that has seasonal patterns, as well as SSA. SSA is an advanced time series method and quite powerful, especially for dealing with time series that contain seasonal patterns [5]. The two methods have very different approaches, therefore these models are formally not comparable. But, for real-world time series whose models are unknown, SARIMA and SSA can be numerically compared [11].

The comparison of forecast results using the SSA and SARIMA methods can be seen in Figure 3

![Figure 3. The Forecasting Result Using SSA and SARIMA](image-url)
The comparison between the two will be seen through the forecasting accuracy value (MAPE). The method that has the smallest MAPE is the most suitable method for this data. The obtained MAPE of the two methods can be seen in Table 8

| Method       | MAPE   |
|--------------|--------|
| SSA          | 17%    |
| SARIMA       | 22.75% |

The result in Table 8 show that SSA has a smaller value of MAPE than SARIMA.

5. Conclusion
SSA and SARIMA have different approach in forecasting the time series data. Formally they are not comparable but numerically can be compared to find the most suitable model for the data. In this study we compare SSA and SARIMA by the obtained MAPE. MAPE for SSA is 17% and MAPE for SARIMA is 22.75%. Therefore, we can conclude that SSA is more suitable for forecasting the rainfall in West Sumatra.

References
[1] S Wirjohamidjoyo and Y Swarinoto 2010 *Iklim Kawasan Indonesia (Dari Aspek Dinamik-Sinoptik)* (Jakarta: Badan Meteorologi Klimatologi dan Geofisika)
[2] N Golyandina, V Nkerutkin, and A Zhigljavsky 2001 *Analysis of Time Series Structure: SSA and Related Techniques* (London: Chapman & Hall/CRC)
[3] H Hassani 2007 *Singular Spectrum Analysis: Methodology and Comparison* *Journal of Data Science* 5 239-257
[4] F Fitri *et al* 2020 Forecasting of Rainfall in Sumatera Barat: Singular Spectrum Analysis (SSA) Application *Journal of Physics: Conference Series* 1554 012047
[5] H Khaeri, E Yulian and G Darmawan 2018 Penerapan Metode Singular Spectrum Analysis (SSA) pada Peramalan Jumlah Penumpang Kereta Api di Indonesia Tahun 2017 *Jurnal Euclid* 5 1 8-20
[6] D C Montgomery, C L Jennings, M Kulahci 2008 *Introduction to Time Series Analysis and Forecasting* (New Jersey: John Wiley & Sons, Inc).
[7] W W S Wei 2006 *Time Series Analysis Univariate and Multivariate Methods*, Second Edition (Boston: Pearson Education)
[8] Aswi and Sukarna 2006 *Analisis Data Deret Waktu: Teori dan Aplikasi* (Makassar: Andira Publisher)
[9] S Siegel 1956 *Nonparametric Statistics for The Behavioral Sciences* (New York: McGraw-Hill Book Company, Inc.)
[10] S Makridakis, S C Wheelwright, V E McGee 1983 *Forecasting: Methods and Applications, 2nd Edition* (New York: John Wiley & Sons, Inc)
[11] N Golyandina and A Korobeynikov 2014 Basic Singular Spectrum Analysis and forecasting with R *Computational Statistics and Data Analysis* 7 1 934-954