BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2022 Volume 16 Issue 1 Page 355-362
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss1pp355-362

# COMPARISON OF SARIMA, SVR, AND GA-SVR METHODS FOR FORECASTING THE NUMBER OF RAINY DAYS IN BENGKULU CITY

**Novi Puspita[1], Farit Mochamad Afendi[2*], Bagus Sartono[3]**

[1,2,3] *Statistics and Data Science Study Program, FMIPA, IPB University*
*Jln. Meranti Wing 22 Level 4. Kampus IPB Darmaga, 16680, Bogor, Indonesia*

*Corresponding author e-mail: ²* fmafendi@gmail.com*

***Abstract.*** *The number of rainy days is a calculation of the rainy days that occur in one month. In recent years, there has been a decrease in rainy days in some parts of Indonesia. One of the areas at risk of quite a high decreasing number of rainy days is the Bengkulu City area. The decrease in the number of rainy days is one of the impacts caused by climate change. The community will feel the impact of climate change-related to the season, especially those working in the agricultural sector. In compiling the planting calendar, it is necessary to consider the seasons to estimate water availability. This study aimed to forecast the data on the number of rainy days in Bengkulu City in the period January 2000 to December 2020 using the Seasonal Autoregressive Integrated Moving Average (SARIMA), Support Vector Regression (SVR), and Genetic Algorithm Support Vector Regression (GA-SVR) methods. The criteria for selecting the best model used was Mean Absolute Deviation (MAD). The MAD value in the SARIMA method was 4,16, 5,07 in the SVR model, and 3,67 in the GA-SVR model. Based on these results, it can be concluded that the GA-SVR model is the best model for forecasting the number of rainy days in Bengkulu City*.

***Keywords****: number of rainy days, SARIMA, SVR, GA-SVR*

***Article info:***

*Submitted: 12th August 2021* *Accepted: 10th March 2022*

***How to cite this article:***
N. Puspita, F. M. Afendi and B. Sartono, "COMPARISON OF SARIMA, SVR, AND GA-SVR METHODS FOR FORECASTING THE NUMBER OF RAINY DAYS IN BENGKULU CITY", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 1, pp. 355-362, Mar. 2022.

## 1. INTRODUCTION

Climate change is one of the current problems. Climate change has a significant impact on various environmental sectors whose impacts are directly felt by the community. According to [1], ongoing climate changes include an increase in the earth's surface temperature, an increase in sea level, reduced snow cover on land, increased air pollutants, and the progress of the seasons. Changes related to the rainy and dry seasons, such as the frequent occurrence of extreme rains, changes in the volume of rainwater, and the duration of the rainy and dry seasons, are unpredictable [2].

The unpredictable change of seasons creates problems for people who work in the agricultural sector. In planning the planting calendar, it is necessary to consider the seasons related to water availability in the dry season. One of the variables that can describe the occurrence of the dry season or the rainy season is the number of rainy days. The number of rainy days is the accumulation of rainy days that occur in one month based on daily measurements. Based on records [1], changes in the amount of rain volume in Bengkulu City have a downward trend. The publication of the Central Statistics Agency (BPS) in the book "Indonesian Environmental Statistics" from 2017 to 2019 states that the number of rainy days in Bengkulu City continues to decrease. According to [3], in 2017, the number of rainy days in Bengkulu amounted to 201 days per year; in 2018, the number of rainy days decreased to 196 days per year [4], and in 2019 the number of rainy days decreased by 42 days from the previous year with the number of rainy days only 154 days per year [5].

The number of rainy days is time-series data. Time series data is observations based on time series. Between adjacent observations correlates, namely observations of a variable correlated with the variable itself in previous observations [6], and time series analysis is a data analysis procedure that considers the effect of time and is carried out to predict a possible future situation. An important step in analyzing time series data is to find the pattern of time series data so that the appropriate method can be determined to model the data. According to [7], the time-series data pattern is divided into a horizontal, cyclical, trend, and seasonal patterns. Data with a seasonal pattern is characterized by repeated changes in the seasonal period, usually one year for monthly data [8].

Data on the number of monthly rainy days in this study were analyzed using the Seasonal Autoregressive Integrated Moving Average (SARIMA) method and the Support Vector Regression (SVR) method. According to [9], the SVR method is a model based on machine learning that can recognize time series data patterns and provide good forecasting results.

Several relevant studies regarding forecasting using the SARIMA and SVR models, namely research [10] in 2017, compared the SARIMA and SVR models in forecasting the number of foreign tourist visits to Bali. It was found that the SARIMA model is a better method for forecasting the number of foreign tourists visiting Bali. Research [11] in 2020 on forecasting short-term electricity demand also uses the SARIMA and SVR models.

Based on the above studies, this study aimed to forecast the number of monthly rainy days in Bengkulu City in the period January 2000 to December 2020 using the Seasonal Autoregressive Integrated Moving Average (SARIMA), Support Vector Regression (SVR), and SVR method using the Genetic Algorithm optimization method. (GA). GA is used to obtain the SVR model with optimal parameters $C$, $\varepsilon$, and $\gamma$. GA is an optimization technique where the way it works is based on the evolutionary process [12], hereinafter referred to as the model (GA-SVR). The three models were compared to determine which method showed the best performance in forecasting by looking at the Mean Absolute Deviation (MAD) value. The MAD value measures the accuracy of the forecasting results by calculating the average value of the estimator error (absolute value of each error)[13].

.

## 2. RESEARCH MODEL

The time-series data used in this study was secondary data on the number of monthly rainy days in Bengkulu City from January 2000 to December 2020. It was obtained based on records from the Bengkulu Meteorology, Climatology, and Geophysics Agency (BMKG) which can be downloaded on the official BMKG https://dataonline.bmkg.go.id/data_iklim. The analysis in this study was carried out with the assistance of software R with the following stages:

1. Divide the data into training data and test data.
2. Conduct data exploration to see an overview of the data.

3. Check the stationarity of the data by looking at the data plot and checking the stationarity of the data with a formal test, namely stationary in the mean with the Augmented Dickey-Fuller (ADF) test with the following formula.:

$$t_{\mathrm{hit}} = \frac{\hat{\gamma} - \gamma}{SE(\hat{\gamma})}$$

   and check the stationarity of the data in variance with the Bartlett test with the following formula:

$$Q_{\mathrm{hit}} = \frac{f\,\mathrm{ins}^2 - \sum_{i=1}^{k}(f_i\,\mathrm{ins}_i^2)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{f_i}\right) - \frac{1}{f}\right]}$$

4. Develop the SARIMA Model with the following stages:
   a. Determine temporary model based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.
   b. Determine a tentative model based on ACF and PACF plots.
   c. Estimate and check the significance of the tentative model parameters.
   d. Check diagnostic tests using the Ljung-Box with the following formula:

$$Q = n(n+2)\sum_{k=1}^{K}\left(\frac{re_k^2}{n-k}\right)$$

   and check residual autocorrelation using the Kolmogorov Smirnov test with the following formula:

$$D = {}^{\sup}_{x}|F(x) - F_0(x)|$$

   where:

   $F(x)$ : cumulative probability function calculated from sample data
   $F_0(x)$ : cumulative probability function of normal distribution

   while the decision criteria taken is to reject $H_0$ if the D is greater than the Kolmogorov-Smirnov statistic or p-value is less than $\alpha$ (0.05).
   e. Do Overfitting.
   f. Choose the best model using the mean absolute deviation with the following formula:

$$MAD = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{n}$$

   where $y_t$ is the actual data, $\hat{y}_t$ namely data from forecasting, and $n$ is the number of data.
5. Analyze the SVR model with the following steps:
   a. Determine the kernel function used. The Kernel function used in this study was Radial Basis Function (RBF)
   b. Determine the range of parameter values $C$, $\varepsilon$, and $\gamma$ for Hyperplane optimization on training data.
   c. Perform SVR modeling based on ranges parameter value
   d. Forecast the SVR model.
   e. Calculate the MAD value.
6. Conduct SVR analysis using Genetic Algorithm (GA) optimization with the following steps:
   a. Arrange chromosomes by generating 100 chromosomes. The generated chromosomes consist of three genes that show the parameters of the SVR model using the RBF kernel function.
   b. Determine the fitness value. The fitness used was the negative MAD value.
   c. Carry out the selection process of 100 chromosomes from some parents from the population using the Roulette Wheel selection.
   d. Do the crossover process if the random is less than the crossover probability $(P_c) = 0.8$.
   e. Carry out the mutation process if the generated random number value is less than the mutation probability value $(P_m) = 0.01$.
   f. Carry elitist process.
   g. Replace the old population with a new generation by selecting some chromosomes that have the best fitness through selection, crossover and elitism.
   h. Check the solutions that have been obtained. The solution has reached the criteria if the fitness has converged. If this condition is not met, the optimization process will be repeated from step 6c.
   i. Enter the parameters obtained based on GA optimization into the SVR algorithm.
   j. Measure the goodness of the model based on the MAD value.

7. Comparing the SARIMA, SVR, and GA-SVR models based on the MAD value to get the model with the best forecast.

## 3. RESULTS AND DISCUSSION

The data used as training data in this study were data from January 2000 to December 2019 and the data used as test data was data from January to December 2020. The first step in the analysis stage of this research was data exploration that can be seen in Figure 1

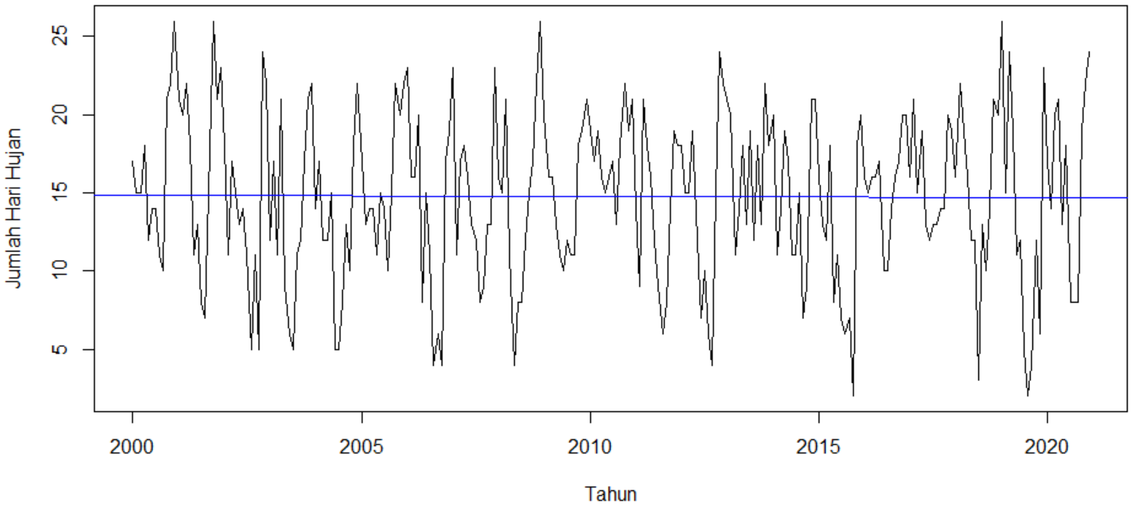


**Figure 1. Number of monthly rainy days in Bengkulu city in 2000 - 2020**

It is known that there are 252 data points in Figure 1 that represent the number of monthly rainy days in the last twenty-one years. Axis X shows the time (year) and the Y shows the number of rainy days. It is known that the number of rainy days in Bengkulu city every year ranges from 138 to 212 days. The highest number of rainy days occurred in 2010 and the least amount in 2015. The average rainy day each month lasts for approximately 15 days. The average value is marked with a blue horizontal line used as a reference to see the distribution pattern of the data. The pattern of stationary data can be identified from the shape of the data that tends to approach the average value and fluctuate around that value. Based on Figure 1, it can be seen that the data is stationary in the mean and variance. However, it needs to be checked again using a formal test. The method used to check the stationarity of the data in the mean and variance is the ADF test and the Bartlett test.

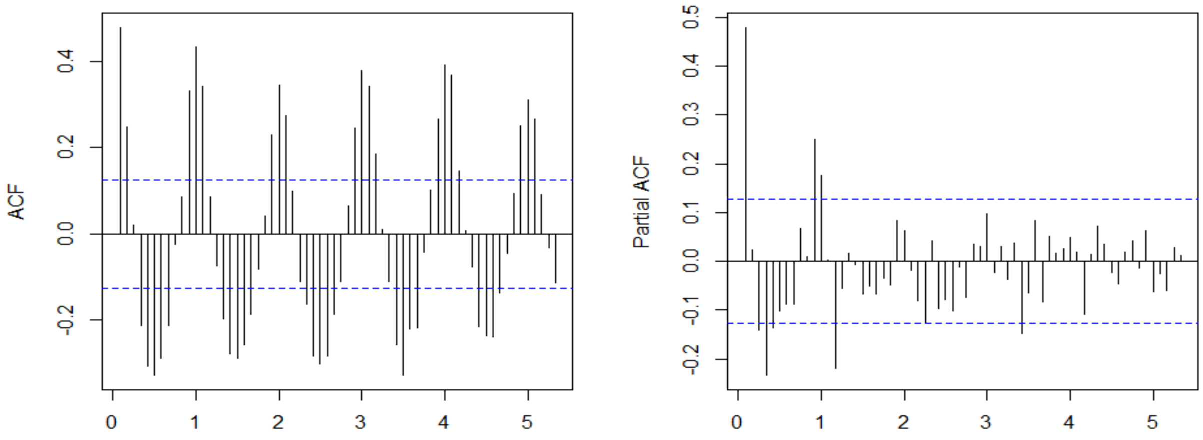
**Table 1. The Result of Stationary test in Mean and Variance**

| Test | Description p-value | ADF |
|---|---|---|
| **ADF** | 0.01 | Stationary in Mean |
| **Bartlett** | 0.15 | Stationary in variance |

Based on Table 1, the ADF test obtained a p-value of 0.01. This value is smaller than the alpha of 0.05. This shows that the data on the number of rainy days in Figure 1 is stationary in mean. Based on the results of the Bartlett test, a p-value is 0.15. It is greater than the alpha $(\alpha = 0.05)$ which means that the data is stationary in variance.

### 3.1 SARIMA

After checking the stationary of the data, the next step is to display the ACF and PACF plots that will be used to determine the order of $p, q, P,$ and $Q$ in the model.



(a) (b)

**Figure 2. ACF (a) and PACF (b) number of monthly rainy days in Bengkulu City**

Order determination in the SARIMA model was carried out by looking at the overall lag in the ACF and PACF plots. Based on Figures 2(a) and 2(b), there are several tentative models that are likely to be used, namely the SARIMA $(1,0,0)(1,0,0)^{12}$, SARIMA $(1,0,1)(1,0,0)^{12}$, SARIMA $(1,0,2)(1,0,0)^{12}$, SARIMA $(1,0,1)(1,0,2)^{12}$ dan SARIMA $(1,0,3)(1,0,0)^{12}$. Furthermore, the significance and white noise for the freedom and normality of the remainder were tested from the several tentative models.

**Table 2. P-value of white noise tentative model test**

| Model | Ljung-Box | Kolmogorov Smirnov | AIC |
|---|---|---|---|
| **SARIMA** $(1,0.0)(1,0,0)^{12}$ | 0,455 | 0,109 | 1394,37 |
| **SARIMA** $(1.0.1)(1.0.1)^{12}$ | 0,759 | 0,520 | 1393,76 |
| **SARIMA** $(1.0.2)(1.0.0)^{12}$ | 0,853 | 0,697 | 1390,34 |
| **SARIMA** $(1.0.1)(1.0.2)^{12}$ | 0,647 | 0,073 | 1352,32 |
| **SARIMA** $(1.0.3)(1.0.0)^{12}$ | 0.986 | 0,613 | 1391,49 |

The five tentative models in table 2 above meet the assumptions of freedom and residual normality according to the p-value of the Ljung-Box and Kolmogorov-Smirnov tests which are greater than the real level ($\alpha = 0.05$). Based on these five models, the SARIMA$(1,0,1)(1,0,2)^{12}$ is the best model with the smallest AIC value.

The next step is to overfit the selected best model. At this stage, the SARIMA$(1,0,1)(1,0,2)^{12}$ , which is feasible based on testing is a tentative model, so it is necessary to combine the SARIMA$(1,0,1)(1,0,2)^{12}$ models to produce several possible SARIMA models which are thought to produce a better model. Based on the overfitting process, several models were obtained as follows.

**Table 3. Comparison of AIC values of SARIMA overfitting model $(1,0,1)(1,0,2)^{12}$**

| model | AIC |
|---|---|
| **SARIMA$(1,0,1)(1,0,2)^{12}$** | 1352,32 |
| **SARIMA$(1,0 ,2)(1,0,2)^{12}$** | 1350,29 |
| **SARIMA$(1,0,1)(2,0,2)^{12}$** | 1352,85 |
| **SARIMA$(1,0,1)(1,0,3)^{12}$** | 1353,60 |

Based on Table 3, the SARIMA model $(1,0,2)(1,0,2)^{12}$ is the best model with the smallest AIC value compared to the other three alleged models. The next step was to check the residual normality test based on Figure 3 below.

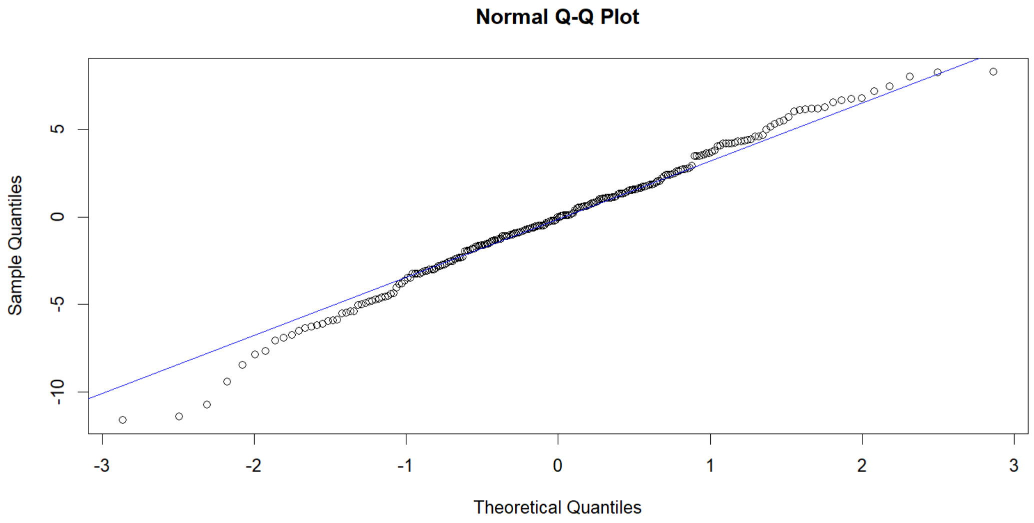


**Figure 3. QQ plot of the SARIMA residual (1,0,2) (1,0,2)$^{12}$**

Based on Figure 3, it can be seen that the data points spread very closely on the diagonal line. It can be said that the assumption of normality has been met, but formal tests can be carried out in advance using the Kolmogorov-Smirnov test. The value is 0.3127 > alpha (= 0.05). It means that the rest of the data is normally distributed. The next step is to check the residual freedom using the Ljung-Box test. The value is 0.8903. This number is greater than the value of (0.05), which means that the SARIMA (1,0,2)(1,0,2)$^{12}$ has independent residuals.

### 3.2 SVR

Support vector regression (SVR) is a development method of support vector machine (SVM). SVR is used for regression and time series modeling. According to [14], an analysis of the SVR model was carried out using the radial basis function (RBF) kernel function. The RBF kernel function has three parameters, cost, epsilon, and gamma. These parameters are used to determine the value of these three parameters, optimization using the grid search. The range of parameters used in this study, respectively for $C, \gamma$, and $\varepsilon$ is (0.01 – 10); (0.6-1.5), and (0.1-1). Based on these initial values, the SVR process gets the optimal parameter values for $C, \gamma$, and $\varepsilon$ that is 4.01; 0.6; and 0.1. Forecasting on the data on the number of rainy days in Bengkulu City for the period January 2020 to December 2020 reach the MAD value for the SVR model of 5.073.

### 3.3 GA-SVR

Modeling the number of rainy days in Bengkulu City used the GA-SVR method with the genetic algorithm as the optimization method. This optimization method was expected to produce better accuracy values using the initialization value, namely the SVR parameter value obtained previously using the grid search optimization method.

The first step in this process was to determine the value needed to optimize of the genetic algorithm. Some values that need to be determined were 100 chromosomes in the population with a maximum iteration set of 100 iterations. Then, the probability of crossing over in this study was set at 0.8, with a probability of mutation between chromosomes of 0.01. The next step was to carry out the optimization process using a genetic algorithm with a range used for $C, \gamma$, dan $\varepsilon$ (1-2); (0,5-0,8) and (0,01- 0,1), respectively. In GA, a measure is needed to evaluate a chromosome. This study used fitness. The fitness used is a negative MAD value. In this process, a chromosome is said to be good and can survive for the next process if it has a minimum MAD value.

The next step was to make a selection using the Roulette Wheel. The fitness obtained based on the previous process was used as a guide for selecting prospective parents of chromosomes. The selected chromosome was a chromosome with a random number value located between the frequency value of the previous chromosome and the chromosome obtained. The chromosomes selected as prospective parents were then given a uniform random(0,1). If the value of the number was less than the probability of crossing over (Pc = 0.8), the chromosome was selected as the parent, and a crossover process occurred so that it is possible to obtain a new chromosome, which was then recalculated its fitness. The next step was to mutate the selected chromosome by replacing one of the parameter values with a random number. If the value of the random number was less than the specified probability of 0.1, the chromosome was the selected chromosome.

### 3.4. Comparison of the Forecasting Model

Figure 4 is a comparison graph of the actual value of the rainy days number in Bengkulu City in 2020 and the forecasting results using the SARIMA, SVR and GA-SVR methods.

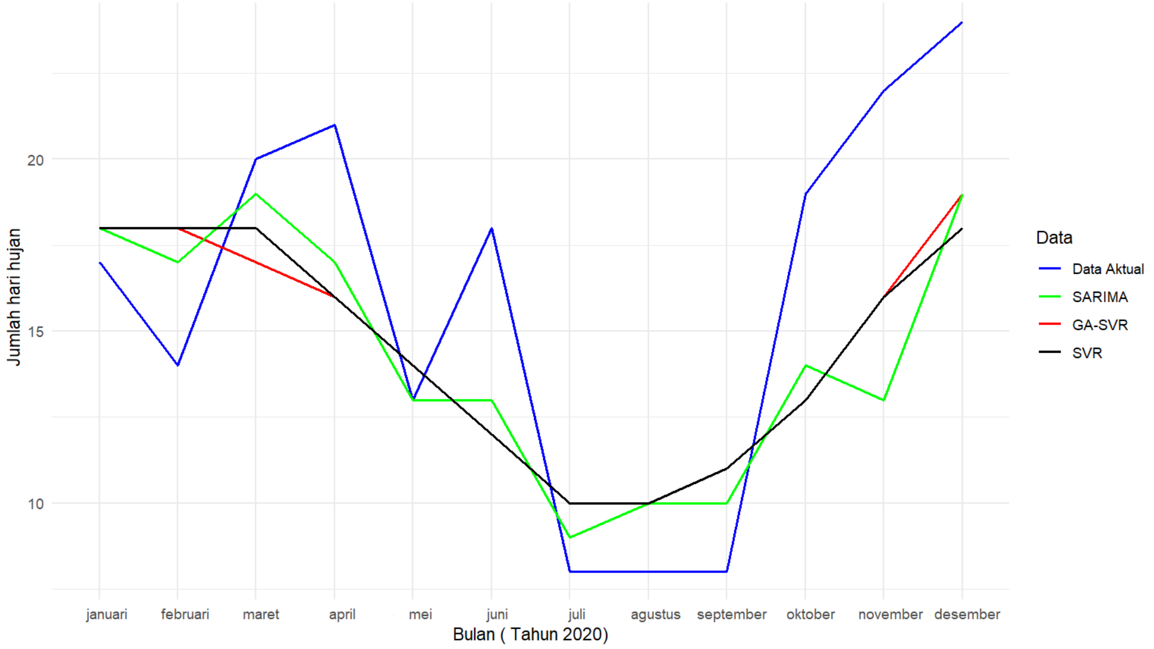


**Figure 4. The Forecasting Value Comparison of the SARIMA, SVR and GA-SVR Models**

Based on Figure 4, the results of forecasting the number of rainy days in Bengkulu City in 2020 will increase at the beginning of the year. The number of rainy days has decreased gradually, reaching the least in June and August. It is estimated that there will be the beginning of the dry season in June and August, wherein one month it only rains as much as 9 days. The largest decrease in the number of rainy days in each month occurred from April to May and June to July, and the increase in the number of rainy days is mostly shown from September to October and November to December.

## 4. CONCLUSION

Based on the research results, it was found that the forecast shown by the SARIMA model had a MAD value of 4.16. The SVR model has a MAD value of 5.07, and the forecast shown by the GA-SVR model has a MAD value of 3.67. These results indicate that the GA-SVR model is the best method for forecasting the number of rainy days in Bengkulu City in 2020.

## REFERENCES

[1] N. Binternagel, “Adaptasi dan Mitigasi Perubahan Iklim Global,” ["Adaptation and Mitigation of Global Climate Change,"] no. October, 2009.

[2] TK Manik, B. Rosadi, and E. Nurhayati, “Mengkaji dampak perubahan iklim terhadap distribusi curah hujan lokal di propinsi lampung.” [“Examining the impact of climate change on local rainfall distribution in Lampung province.”]

[3] Central Bureau of Statistics, “Statistik Lingkungan Hidup Indonesia Enviroment Statistic of Indonesia 2017,” [“Environmental Statistics of Indonesia Environmental Statistics of Indonesia 2017,”] *Badan Pus. stats.*, vol. 91, no. 1, pp. 186–189, 2017, [Online]. Available: http://www.un-ilibrary.org/economic-and-social-development/the-sustainable-development-goals-report-2017_4d038e1e-en.

[4] Central Bureau of Statistics, “Statistik Lingkungan Hidup Indonesia (SLHI) 2018,” [“Indonesian Environmental Statistics

(SLHI) 2018,"] *Badan Pus. stats. Indonesia.*, pp. 1–43, 2018, doi: 3305001.

[5] [BPS] Central Bureau of Statistics, "Statistik Lingkungan Indonesia 2019," ["Indonesian Environmental Statistics 2019,"] *Badan Pus. stats.*, pp. 1-224, 2019,
[Online]. Available: https://www.bps.go.id/publication/2018/12/07/d8cbb5465bd1d3138c21fc80/statistik-lingungan-live-indonesia 2018.html.

[6] S. Wei and CP Soon, "Genetic algorithm-based text clustering technique," *Lect. Notes Comput. science. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4221 LNCS, pp. 779–782, 2006, doi:10.1007/11881070_103.

[7] HR Makridakis S, Wheelwright SC, "1 / the Forecasting Perspective," *Forecast. methods app.*, pp. 1–632, 1997.

[8] Rob J Hyndman and A. George, "Forecasting: Principles and Practice," *Princ. Optimistic. Dec.*, no. September, pp. 421–455, 2018, [Online]. Available: https://otexts.com/fpp2/index.html.

[9] P. Meesad and RI Rasel, "Predicting stock market price using support vector regression," *2013 Int. conf. Informatics, Electron. Vision, ICIEV 2013*, no. May, 2013, doi:10.1109/ICIEV.2013.6572570.

[10] NPN Hendayanti and M. Nurhidayati, "Perbandingan Metode Seasonal Autoregressive Integrated Moving Average (SARIMA) dengan Support Vector Regression (SVR) dalam Memprediksi Jumlah Kunjungan Wisatawan Mancanegara ke Bali," ["Comparison of the Seasonal Autoregressive Integrated Moving Average (SARIMA) Method with Support Vector Regression (SVR) in Predicting the Number of International Tourist Visits to Bali,"] *J. Varian*, vol. 3, no. 2, pp. 149–162, 2020, doi:10.30812/varian.v3i2.668.

[11] Q. Lou, Q. Lyu, Z. Na, D. Ma, and X. Ma, "Short-term electric power demand forecasting using a hybrid model of SARIMA and SVR," *IOP Conf. Ser. Earth Environment. science.*, vol. 619, no. 1, 2020, doi:10.1088/1755-1315/619/1/012035.

[12] M. Shehab, H. Alshawabkah, L. Abualigah, and N. AL-Madi, "Enhanced a hybrid moth-flame optimization algorithm using new selection schemes," *Eng. Comput.*, vol. 37, no. 4, pp. 2931–2956, 2021, doi: 10.1007/s00366-020-00971-7.

[13] EL Amalia, DW Wibowo, HS Pakpahan, and O. Anandiya, "Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error."

[14] S. Sande and ML Privalsky, "Identification of TRACs (T3 receptor-associating cofactors), a family of cofactors that associate with, and modulate the activity of, nuclear hormone receptors," *Mol. Endocrinol.*, vol. 10, no. 7, pp. 813–825, 1996, doi:10.1210/me.10.7.813.

[15] X. Tang, L. Wang, J. Cheng, J. Chen, and VS Sheng, "Forecasting model based on information-granulated GA-SVR and ARIMA for producer price index," *Comput. mater. Contin.*, vol. 58, no. 2, pp. 463–491, 2019, doi:10.32604/cmc.2019.03816.