Comparison of SARIMA and SVM model for rainfall forecasting in Bogor city, Indonesia

A S Abdullah1*, B N Ruchjana2, I G N M Jaya3,4, and Soemartini3

1Department of Computer Science, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang, km 21 Jatinangor, Sumedang 45363, Indonesia
2Department of Mathematics, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang, km 21 Jatinangor, Sumedang 45363, Indonesia
3Department of Statistics, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang, KM21 Jatinangor, Sumedang 45363, Indonesia
4Faculty of Spatial Sciences, Groningen University, the Netherlands

*atje.setiawan@unpad.ac.id

Abstract. Heavy rainfall and flash floods are still significant for several cities in Indonesia, especially Bogor City. Rainfall prediction is essential to develop early warning systems for early detection of massive and flash floods. The accurate forecasting method is needed to produce accurate predictions of rainfall. Several methods have been developed to get a precise rainfall forecast, such as the classical approach, such as seasonal autoregressive integrated moving average (SARIMA), and machine learning approach, for example, gradient boosting and support vector machine (SVM). In this study, we introduce for forecasting rainfall at Bogor city, Indonesia. The SVM provides an accurate result with a small MAPE.

1. Introduction
Rain is the greatest of nature’s gifts for our daily lives [1] and the most crucial climate element affecting human lives with complex systems [2, 3]. Rain provides water for all living things. But in many places in the world, the rain has become a menace. Heavy rainfall and flash floods are still significant issues for several cities in Indonesia, especially Bogor City. Rainfall forecasts are frequently used to develop an early warning system of the flash flood [4]. The accurate forecasting method is needed to produce accurate predictions of rainfall. Several methods have been introduced for univariate time series data. In classical statistics, the Box-Jenkin autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) for seasonal time series data have been used frequently [5]. The time series models are also developed based on machine learning approaches such as extreme Gradient Boosting [6, 7] and support vector machine (SVM) [1]. The classical method, such as ARIMA or SARIMA, needs a tight assumption of stationarity to provide a good prediction. In contrast, machine learning methods do not require this assumption for forecasting purposes.

The objective of this study is to examine whether SVM without stationary assumption can outperform the classical ARIMA. This study provides insight of SVM for rainfall forecasting by performing training and testing data method. We did not use the simulation method for models comparison. We use the real data and compare the model quality based on several criteria, including
mean absolute error (MAE), root means squared error (RMSE), and mean absolute percentage error (MAPE).

The paper is structured as follows: Section 2, we material and method, move on with result and discussion in Section 3. Finally, Section 4 presents the conclusion.

2. Material and Method

2.1 Material

The monthly rainfall data over the period of January 2011 – June 2020 are obtained from http://dataonline.bmkg.go.id/home. The data are observed from four observation rainfall stations in Bogor. The data are presented in Figures 1 and 2.

![Rainfall data in Bogor, January, 2011 – June, 2020 by years](image1.png)

(a) Rainfall data in Bogor, January, 2011 – June, 2020 by years

![Rainfall data in Bogor, January, 2011 – June, 2020](image2.png)

(b) Rainfall data in Bogor, January, 2011 – June, 2020

Figure 1 Rainfall data, January 2011 – June 2020

Figure 1 (a) shows the similar pattern for four stations and figure 1 (b) shows the seasonal pattern of the rainfall data.

2.2. ARIMA

ARIMA is the most popular classical time series model for stationary and non-stationary data [8]. The basic ARIMA is defined as follows [7, 9]:

\[ y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} - \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}, \quad t = 1, ..., T, \quad i = 1, ..., p \text{ and } j = 0, 1, ..., q \]  

(1)

where \( y_t \) is the rainfall data for time \( t \), \( \alpha_0 \) denotes an intercept, \( \alpha_i \) autoregressive coefficient for lag \( i \), and \( \varepsilon_{t-j} \) error term with \( j \) time lag and the coefficients moving average \( \beta_j \). The error term usually assume to be withe noise, i.e., \( \varepsilon_t \sim N(0, \sigma^2) \). Then, the \( i \) and \( j \) terms are autoregressive and moving average parts of the model. Integrated with number differencing \( d \) can be applied to satisfy the stationarity assumption. Eq.1 can be developed to account seasonal pattern and the model becomes \( SARIMA(p,d,q)(P,D,Q)[s] \), where \( P, D, Q \) are similar to ARIMA\((p,d,q)\) but here as a part of seasonal components with \( s \) denotes the seasonal order. The model can be estimated using maximum likelihood method.

2.3. SVM

SVM through nonlinear eps-regression model provides flexibility in forecasting purpose without stationary assumption needed. SVM allows to define how big error is acceptable in prediction and find the best model to fit the data [10]. Let consider regression ordinary least square (OLS)
\[ y_t = \alpha + w_t x_t + \varepsilon_t \quad t = 1, \ldots, T \]  

(2) where \( y_t \) is the target, \( \alpha \) is an intercept, \( w_t \) is the coefficient, and \( x_t \) is the predictor (feature) and \( \varepsilon_t \) and error term. The objective of linear models is to minimize the sum of squared errors as follows:

\[
\min \sum_{t=1}^{T} (y_t - \alpha - w_t x_t)^2.
\]

(3) SVM provides the flexibility to define how big error is acceptable, the objective function becomes:

\[
\min \frac{1}{2} \|w\|^2
\]

(4) with constraint:

\[
|y_t - \alpha - w_t x_t| < \varepsilon.
\]

(5) These deviations have the potential to solve and we would still like to minimize them as much as possible by adding these deviations to the objective function

\[
\min \frac{1}{2} \|w\|^2 + C \sum_{t=1}^{T} |\xi_t|
\]

(6) with constraint:

\[
|y_t - \alpha - w_t x_t| < \varepsilon + |\xi_t|.
\]

(7) For nonlinear SVM, the radial kernel is widely used defined as:

\[
\exp\left(-\frac{\gamma}{2} \|x_t - x_{\xi}\|^2\right)
\]

(8) where \( C \) is the cost parameter used to control overfitting and \( \gamma \) is precision parameter to control the variance of the data.

2.4. Forecast evaluation methods

In order to compare the performance of SARIMA and SVM, we use some prediction performance criterion and goodness of fit index \([11, 12]\). First criteria is mean absolute error (MAE) as presented below:

\[
MAE = \frac{1}{H} \sum_{h=1}^{H} |y_{t+h} - \hat{y}_{t+h}|
\]

(9) where \( y_{t+h} \) denotes data testing at \( t + h \) period and \( \hat{y}_{t+h} \) is the forecast value for \( t + h \) period with \( H \) is length of forecast. The second and third criteria are root means squared (RMSE) and mean absolute percentage deviance (MAPE) given below:

\[
RMSE = \sqrt{\frac{1}{H-1} \sum_{h=1}^{H} (y_{t+h} - \hat{y}_{t+h})^2}
\]

(10)
\[ MAPE = \frac{1}{H} \left( \sum_{h=1}^{H} \left| \frac{y_{t+h} - \hat{y}_{t+h}}{y_{t+h}} \right| \right) \times 100\%. \] (11)

The smallest values of MAE, RMSE, and MAPE indicate the best model.

For goodness of fit index we use r-Pearson correlation:

\[ r = \frac{H \sum_{h=1}^{H} y_{t+h} \hat{y}_{t+h} - \sum_{h=1}^{H} y_{t+h} \sum_{h=1}^{H} \hat{y}_{t+h}}{\sqrt{\left[ n \sum_{h=1}^{H} y_{t+h}^2 - (\sum_{h=1}^{H} y_{t+h})^2 \right] n \sum_{h=1}^{H} \hat{y}_{t+h}^2 - (\sum_{h=1}^{H} \hat{y}_{t+h})^2}}. \] (12)

The higher r is better fit model [12].

All computation process will be done by R-software using forecast [13] and e1071 [14] packages.

3. Result and discussion

3.1. Data exploration

Diagnostic is the first step to define the temporal pattern and to test the stationarity. We present the autocorrelation function (ACF) and partial autocorrelation function (PACF) plot for identifying the temporal pattern and the Dicky Fuller test for testing stationarity assumption. For the purpose of model identification and validation, we use early 75% data as data training and the rest 25% for data testing.

![ACF and PACF plots](image)

**Figure 2.** (a) ACF and (b) PACF for training data (January 2011 – December 2017)

Figure 2 present the autocorrelation function with a clear seasonal pattern at time lag 12. It shows that the rainfall data in the city of Bogor have seasonal pattern for every year (s=12 months). The Dicky Fuller stationarity test rejects the null hypothesis (H0: data series is not stationary) with a p-value of 0.01.

3.2. ARIMA

By evaluating the ACF and PACF plot, there are two candidates model were evaluated with the forecast results are presented below:
3.3 SVM

We use nonlinear epsilon-regression techniques using the radial kernel in SVM to forecast rainfall data in Bogor for 18 months. There are two different models we evaluate. First model without seasonal constraint by defining time index $t=1,2,\ldots,T$ and model with seasonal constraint by defining cycling time index $t=1,\ldots,12,\ldots,1\ldots12$. In nonlinear epsilon-regression techniques using the radial kernel, there are two parameters that have to be optimized. The first parameter is Cost parameter to control the overfitting and the second parameter is the gamma parameter to control the variance of the parameter estimate. The epsilon is fixed=0.1. The optimization results is presented in Figure 3.

| Model               | Cost | Gamma | Best performance |
|---------------------|------|-------|------------------|
| Without seasonal constraint | 1    | 966   | 36,846.28        |
| With seasonal constraint  | 4    | 2     | 36,521.54        |

(a) Without seasonal constraint  
(b) With seasonal constraint
Figure 5. (a) Model without seasonal constraint and (b) model with seasonal constraint.

Figure 5 shows the model without seasonal constraint is better for the training data set but has the worst performance for testing data. It provides constant forecast values over times. The model with seasonal constraints has good performance for training and testing data.

Table 2. Models comparison

| Model        | Parameter           | MAE     | RMSE   | MAPE   | r-Pearson |
|--------------|---------------------|---------|--------|--------|-----------|
| SARIMA       | (1,0,0,1,0,0)       | 128.366 | 156.767| 93.480 | 0.204     |
|              | (1,0,1,1,0,1)       | 128.173 | 155.401| 92.835 | 0.257     |
| SVM          | Without seasonal constraint (Cost=1; Gamma=996) | 131.074 | 158.749| 94.954 | 0.122     |
|              | With seasonal constraint (Cost=4; Gamma=2)     | 97.16   | 121.62 | 63.28  | 0.655     |

Table 2 shows the model comparison between SARIMA and SVM. Based on prediction performance criterion: MAE, RMSE, and MAPE, and goodness of fit index using r-Pearson, we found SVM with seasonal constraint and parameter cost=4 and Gamma=2 is the best model. It has the lowest MAE, RMSE, and MAPE, and highest r-Pearson. Figure 2 and Table 6 present the forecast output based on SARIMA (1,0,1,1,0,1) and SVM (With seasonal constraint).

Figure 6. Forecasting result comparisons between SARIMA and SVM
Table 3. Forecasting results based on SVM and SARIMA

| Year | Month  | SVM    | SARIMA  | Deviance |
|------|--------|--------|---------|----------|
| 2020 | July   | 161.918| 274.111 | -112.194 |
| 2020 | August | 146.098| 302.662 | -156.564 |
| 2020 | September | 260.286| 282.458 | -22.173  |
| 2020 | October | 404.852| 354.825 | 50.027   |
| 2020 | November | 423.086| 292.323 | 130.763  |
| 2020 | December | 348.107| 369.280 | -21.173  |
| 2021 | January | 352.381| 383.770 | -31.389  |
| 2021 | February | 424.098| 368.298 | 55.800   |
| 2021 | March   | 426.443| 444.888 | -18.445  |
| 2021 | April   | 395.571| 347.509 | 48.062   |
| 2021 | May     | 339.090| 341.684 | -2.593   |
| 2021 | June    | 250.626| 333.697 | -83.071  |
| 2021 | July    | 161.918| 342.401 | -180.483 |
| 2021 | August  | 146.098| 337.828 | -191.730 |
| 2021 | September | 260.286| 341.824 | -81.538  |
| 2021 | October | 404.852| 328.735 | 76.117   |
| 2021 | November | 423.086| 340.228 | 82.858   |
| 2021 | December | 348.107| 326.189 | 21.917   |

Figure 6 and Table 3 shows the forecasting results based on SVM and SARIMA and its deviance. By introducing seasonal constraints in the SVM model, the forecasting result has a similar pattern with the historical data. Although ARIMA also includes seasonal information, forecast results tend to be constant overtimes. It indicates, SVM with seasonal constraint outperforms SARIMA because it can fit the data better and taking the historical pattern into the forecast outputs.

4. Conclusion
The application of SVM in forecasting rainfall data is explored. The accuracy of the forecasts, in terms of MAE, MSE, MAPE, and goodness of fit based on r-Pearson have been compared with classical SARIMA models. From the model's comparison and evaluate the rainfall phenomena in Indonesia, particularly in Bogor city, SVM model significantly outperform SARIMA in term of model comparison criterion: MAE, MSE, MAPE, and r-Pearson. We have also shown that the prediction performance of the SVM increase after including the seasonal constraints as the time prediction (covariate). The SVM model was commonly used for model classification. It has relatively new for forecasting purposes. This research has clearly shown the potential of SVM in predicting time series data such as climate extrapolation. Understanding the pattern of the time-series data becomes very useful in order to define the input in the SVM model. The SVM method is a promising alternative method for forecasting. The SVR works best for any type of data whether linear or non-linear [15].

Acknowledgments
We thanks to Rector Universitas Padjadjaran. Financial support was received from ALG Unpad contract: 1427/UN6.3.1/LT/2020.
References

[1] Samui P, Mandla V R, Krishna A, Teja T 2011 Prediction of rainfall using support vector machine and relevance vector machine *Earth Science India*, 4 188

[2] Hastenrath S, Greischar L, Heerden J v 1995 Prediction of the summer rainfall over South Africa *American Meteorological Society* 8, 1511

[3] Jaya I G N M, Ruchjana B N, Abdullah A S, Andriyana Y 2018 Monitoring the underlying probability of daily rainfall occurrence in West Java over space and time *International Journal of Geoinformatics* 17 7

[4] Sene K 2013 Rainfall Forecasting in *Flash Floods*, (Dordrecht:Springer, pp 101-132)

[5] Hyndman R J, Koehler A B, Ord J K, Snyder R D 2008 *Forecasting with Exponential Smoothing: The State Space Approach* (Berlin: Springer)

[6] Chen T, C Guestrin 2018 *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785

[7] Agata R, Jaya I G N M 2019 A comparison of extreme gradient boosting, sarima, exponential smoothing, and neural network models for forecasting rainfall data, *Journal of Physics Conference Series* 1397 1

[8] Afrifa-Yamoah E, Saeed B I I, Karim A 2016 Sarima modelling and forecasting of monthly rainfall in the Brong Ahafo region of Ghana *World Environment* 6 1

[9] Kohzadi N, Boyd M S, Kermanshahi B, Kaastra I 1996 A comparison of artificial neural network and time series models for forecasting commodity prices *Neurocomputing* 10 169

[10] Awad M, Khanna R 2015 *Efficient Learning Machines Theories, Concepts, and Applications, for Engineers and Systems Designers* (Berkeley: Apress Open)

[11] Box G E P, Jenkins G M, Reinsel G C, Ljung G M 2016 *Time Series Analysis Forecasting and Control Fifth Edition* (Hoboken: John Wiley & Sons, Inc)

[12] Pal R 2017 *Validation methodologies in Predictive Modeling of Drug Sensitivity* (Amsterdam: Elsevier Inc pp 83-107)

[13] Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E 2020 *Forecasting Functions for Time Series and Linear Models* (CRAN-R)

[14] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C C, Lin C C 2019 *Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)* (TU Wien: CRAN-R)

[15] Ojemakinde B T 2006 *Support Vector Regression for Non-Stationary Time Series* (Knoxville: Master's Thesis, University of Tennessee - Knoxville)