Vector Autoregression (Var) Model for Rainfall Forecast and Isohyet Mapping in Semarang – Central Java – Indonesia

Adi Nugroho  
Faculty of Information Technology  
Satya Wacana Christian University  
Salatiga, Indonesia

Subanar  
Faculty of Mathematics and Natural Science  
Gadjah Mada University  
Jogyakarta, Indonesia

Sri Hartati  
Faculty of Mathematics and Natural Science  
Gadjah Mada University  
Jogyakarta, Indonesia

Khabib Mustofa  
Faculty of Mathematics and Natural Science  
Gadjah Mada University  
Jogyakarta, Indonesia

Abstract—Agricultural and plantation activities in Indonesia, especially in Semarang, Central Java, Indonesia rely on water supply from the rainfall. The rainfall in the future is basically influenced by rainfall patterns, humidity and temperature in the past. In this case, Vector Autoregression (VAR) multivariate model is applied to forecast the rainfall in the future, in which all along Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) generally uses ARIMA model (Autoregressive Integrated Moving Average) to carry out the same thing. The study applied the data, comprising the data of rainfall, humidity and temperature taken on a monthly basis during 2001-2013 periods from 5 measurement stations. Plotting of rainfall forecast result with VAR method is portrayed in the form of isohyet contour map to see the correlation between rainfall and coordinates of the area of the rainfall. The forecast result shows that VAR method is quite accurate to use for rainfall forecast in the study area as well as better than ARIMA method to forecast the same thing as having smaller Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

Keywords—Rainfall Forecast; VAR; Multivariate Time Series; Isohyet

I. INTRODUCTION

Indonesia has abundant natural resources in tropical areas with quite high rainfall as the islands are surrounded by the vast oceans, fairly high daily temperature and humidity [20]. Currently, there are approximately 40.6 million hectares of agricultural and plantation areas in Indonesia [22] which mostly rely the water supply on the rainfall. Regarding this matter, the western and northeast parts of Indonesia have geological condition and fertile soil which enables the agriculture/plantation can virtually be done as long as the water supply from the rainfall is sufficient [21].

Semarang, Central Java, Indonesia (the study area) is geographically located in Java island in the western part of Indonesia. The study area lies on the geographic position of 6° 5’ – 7° 10’ S and 110° 34’ – 110°35’ E with a total area of 37,366,838 hectares or about 373.7 km²[19]. Generally, the rainfall in the study area follows the pattern of 2 seasons, namely dry season (April – September) and wet season (October – March).

Fig. 1. Study Area Map

[20]. The conducted study tried the seasonal rainfall forecast (dry and wet season) 1 year forward (2014) based on monthly rainfall data taken along the span of the previous 13
years (2001-2013). The rainfall forecast is conducted by using multivariate time series method, or more particularly: Vector Autoregression (VAR). Furthermore, for the sake of decision making accuracy, the rainfall forecast is portrayed in the form of isohyet contour map to see the portrait of rainfall amounts in each sub-district in Semarang.

II. PREVIOUS STUDY

VAR method that is introduced by Christopher A. Sims, a Nobel Prize winner in econometrics, is previously pretty much used to develop econometric models [5] [13], such as to see the correlation of Gross Domestic Product (GDP) to the inflation level in a country, the tourist arrival level to a country which is influenced by many factors, the correlation between Composite Stock Price Index (Indeks Harga Saham Gabungan/IHSg), currency exchange rates and prices [11], the market response to the marketing mix [16], the correlation between public expenditure and economic growth [1], etc. For natural phenomena, especially for rainfall forecasting, Dewi Retno, et al [14] have conducted a study of correlation between rainfall in a region and rainfall in other nearby regions. Meanwhile, in the study that we conducted, the rainfall is connected to the rainfall in the previous periods as well as its correlation with humidity and temperature data.

In general, rainfall (precipitation) is a part of hydrologic cycle (water cycle) consisting of: (1) evaporation and/or evapotranspiration, (2) precipitation, and (3) surface water flow [12]. At each stage, the air humidity (water vapor percentage in an air volume) and the temperature would be very influential. When the temperature is relatively high due to the sunlight, there would be an evaporation/evapotranspiration of surface water/vegetation and water vapor in an air volume will be formed in which this water vapor in a certain altitude, in turn, would form a core of condensation to form clouds. Due to the influence of low temperature and relatively immense droplet, the clouds then would drop back to the ground in the form of rain, snow, dew, fog, etc., which in turn would form surface water flow to restart the hydrologic cycle.

In Indonesia, the studies related to the rainfall according to Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) were mostly conducted using ARIMA method in which the study area has 22,13% MAPE value [6]. The study that we conducted aims to find out if VAR method that we applied can reduce this MAPE value, where the result would enhance the prediction accuracy. Moreover, in our study, we intend to connect the values of rainfall forecast to geographic coordinates of the area where the rainfall would drop by portraying it in the form of isohyet contour map.

III. RESEARCH METHODOLOGY

Time series is basically a measurement data taken in chronological order in certain time [9]. In the conducted study, based on the characteristic of each time series with some different kinds of data (rainfall, humidity and temperature), Vector Autoregression (VAR) method is applied. VAR is basically a combination of Autoregressive (AR) method and frequently known as Box-Jenkins method as developed by George Box and Gwilym Jenkins in 1976 [9].

For instance, the following is the time series of AR.

\[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p+1} + \beta_p y_{t-p} + \varepsilon_t \]  

(1)

In which \( y_t \) is the current value, while \( p \) is lag in autoregressive process; \( y_{t-1} \) to \( y_{t-p} \) is the measurement values from t-1 to t-p; \( \beta_0 \) is intercept value and \( \beta_1 \) to \( \beta_p \) is regression coefficient from t-1 to t-p; and \( \varepsilon_t \) is error value or frequently known as white noise which is assumed to be a normal distribution, independent of \( y_{t-1} \) and constant variance of \( \sigma^2 \) or equal to 0 [3] [7]. In terms of the use of VAR model in this conducted study, it required applying stationary condition criteria, which is defined as a condition where its mean and variance are constant and the covariance is not time-dependent [11]. The stationarity in this study was tested using unit root test with Augmented Dickey-Fuller (ADF) method. The ADF test has the following mathematical equation [8].

\[ \Delta y_t = \beta_0 + \theta y_{t-1} + \sum_{i=2}^{p} \Phi_i \Delta y_{t-i+1} + \varepsilon_t \]  

(2)

where \( \Delta y_t \) is time series value at the t-time minus time series value in 1 previous measurement period (the t-1 time), \( \theta \) is constant-valued (\( \beta_1 + \ldots + \beta_p \) - 1) [8] which is used to determine whether or not the unit roots exist with hypothesis \( H_0: \theta = 0 \) (the data contain unit roots) and \( H_1: \theta < 0 \) (the data do not contain unit roots). Meanwhile, \( \Phi_i \) is trend coefficient on the time series data of which the value is equal to \( \Phi_i = -\sum_{j=1}^{p} \beta_j \) [8]. In this case, the criteria that can be drawn are non-stationary data will have unit roots, while stationary data will not have unit roots. If the data are not stationary, it required to conduct differencing process once or several times on the related data until the data turn out to be stationary [10] [17].

Prior to VAR model was completely formed eventually, the accuracy level should be evaluated by calculating its lag value, which is generally indicated by its p-value. In the conducted study, in order to assess the feasibility level of the rainfall forecast model, it required applying Aikake’s Information Criterion (AIC) calculation for some k independent variables where the AIC value is generally defined using the following mathematical equation [15].

\[ AIC = log\sigma_k^2 + \frac{n+k}{n} \cdots \]  

(3)

Where \( \sigma_k^2 = \frac{SSE}{n} \) with \( SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 \)

In which \( y_i \) is observed value at the i-time; k is the number of parameters in the model; \( \bar{y} \) is mean; and n is the number of observation times. In this case, it can be stated that in case the AIC calculation value is smaller, the taken lag value is the better lag value [5] [15] as well as can be used as forecasting basis.

After accomplishing to determine the (p) lag value with AIC, the VAR mathematical equation system which theoretically does not distinguish the number of dependent and independent variables [17], the combination of rainfall, humidity and temperature variables by considering the accuracy of autoregressive equation (equation (1)) can be noted in the form of a matrix equation as follows [5] [13].
The length of time series of observation can also be noted as follows (for rainfall).

\[
\begin{align*}
\{y_{1,t}\} & = \begin{bmatrix} c_1 \\ a_{1,1} \end{bmatrix} + \begin{bmatrix} a_{1,1} \\ a_{1,2} \end{bmatrix} y_{1,t-1} + \cdots + \\
\{y_{2,t}\} & = \begin{bmatrix} c_2 \\ a_{1,1} \end{bmatrix} + \begin{bmatrix} a_{1,1} \\ a_{1,2} \end{bmatrix} y_{2,t-1} + \cdots + \\
\{y_{3,t}\} & = \begin{bmatrix} c_3 \\ a_{1,1} \end{bmatrix} + \begin{bmatrix} a_{1,1} \\ a_{1,2} \end{bmatrix} y_{3,t-1} + \cdots + \\
\{y_{1,t-p}\} & = \begin{bmatrix} c_1 \\ a_{p,1} \end{bmatrix} + \begin{bmatrix} a_{p,1} \\ a_{p,2} \end{bmatrix} y_{1,t-p} + \cdots + \\
\{y_{2,t-p}\} & = \begin{bmatrix} c_2 \\ a_{p,1} \end{bmatrix} + \begin{bmatrix} a_{p,1} \\ a_{p,2} \end{bmatrix} y_{2,t-p} + \cdots + \\
\{y_{3,t-p}\} & = \begin{bmatrix} c_3 \\ a_{p,1} \end{bmatrix} + \begin{bmatrix} a_{p,1} \\ a_{p,2} \end{bmatrix} y_{3,t-p} + \cdots + \\
\{x_{e,1}\} & = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_{p,1} \end{bmatrix} \\
\{x_{e,2}\} & = \begin{bmatrix} \varepsilon_2 \\ \varepsilon_{p,2} \end{bmatrix} \\
\{x_{e,3}\} & = \begin{bmatrix} \varepsilon_3 \\ \varepsilon_{p,3} \end{bmatrix}
\end{align*}
\]

The matrix equation above (equation (4)), using regression notation can also be noted as follows (for rainfall).

\[
y_{1,t} = c_1 + a_{1,1} y_{1,t-1} + a_{1,2} y_{2,t-1} + a_{1,3} y_{3,t-1} + \cdots + a_{p,1} y_{1,t-p} + a_{p,2} y_{2,t-p} + a_{p,3} y_{3,t-p} + \varepsilon_{1,t}
\]

In which \(y_{1,t}\) is the rainfall at the \(t\) time; \(y_{2,t}\) is the humidity at the \(t\) time; and \(y_{3,t}\) is the temperature at the \(t\) time. Meanwhile, \(c\) is the constant indicating the intercept; \(\varepsilon\) is the errors level; and \(p\) is the lag length. In this case, the parameter values \(a_{1,1}\) to \(a_{p,3}\) can be estimated by using Ordinary Least Square (OLS) method, by minimizing the value of squared error (minimizing \(\varepsilon^2\) value) [13] [17]. The determination of the parameter values can be started from the \(p\) matrix, and then recursively defined to the other parameters.

Once the best model was obtained and able to be used for forecasting, the forecast accuracy level of the model can also be evaluated mathematically using the following mathematical equation [5] [15] [17].

- Computing Mean Absolute Error (MAE)

MAE is a computation of mean absolute error to see how close the values between the forecast and the real value. MAE is generally defined as the following equation.

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |F_t - Y_t|
\]

Where \(F_t\) is the forecast value; \(Y_t\) is the actual data; and \(n\) is the length of time series of observation.

<table>
<thead>
<tr>
<th>Table I</th>
<th>Rainfall, Humidity and Temperature Stationarity Test in Various Measurement Stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall Station</td>
<td>ADF Of Rainfall</td>
</tr>
<tr>
<td>Ungaran Semarang</td>
<td>-8.8005</td>
</tr>
<tr>
<td>Brigin-Salatiga</td>
<td>-3.0617</td>
</tr>
<tr>
<td>Adisumarmo Boyolali</td>
<td>-2.8473</td>
</tr>
<tr>
<td>Borobudur Magelang</td>
<td>-3.6135</td>
</tr>
<tr>
<td>Semarang Kota</td>
<td>-3.0887</td>
</tr>
</tbody>
</table>

Note: S (stationary hypothesis) and not requiring differencing process.

<table>
<thead>
<tr>
<th>Table II</th>
<th>Table of AIC for Various VAR Computations</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR (3)</td>
<td>VAR (4)</td>
</tr>
<tr>
<td>Ungaran - Semarang</td>
<td>35</td>
</tr>
<tr>
<td>Brigin - Salatiga</td>
<td>25</td>
</tr>
<tr>
<td>Adisumarmo - Boyolali</td>
<td>26</td>
</tr>
<tr>
<td>Borobudur - Magelang</td>
<td>29</td>
</tr>
<tr>
<td>Semarang Kota</td>
<td>28</td>
</tr>
</tbody>
</table>

Computing Mean Absolute Percentage Error (MAPE)

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - F_t|}{Y_t} \times 100
\]

Where \(F_t\) is the forecast value and \(Y_t\) is the actual data.

In terms of the MAE and MAPE computation, the good model will have the smallest possible value of MAE as well as MAPE (less than or equal to 10%) [5] [15].

After the rainfall forecast value for the following year was accomplished, in order to portray the rainfalls in the study area related to the geographical position which is different from the measurement station, the isolayet map of Semarang then needs to be composed. The isolayet map is actually a regular contour map drawn to connect the dots in Semarang map with the same rainfall (forecast) values [18].

IV. Results and Discussion

The rainfall, humidity and temperature data used in this study were taken from Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) of Central Java, Indonesia. The obtained data are the data of monthly rainfall throughout 2001-2013 periods from 5 measurement stations, comprising Ungaran-Semarang, Semarang Kota, Brigin-Salatiga, Adisumarmo-Boyolali, and Borobudur-Magelang. At each measurement station, the rainfall forecast was carried out using VAR model, then (after the data from those 5 measurement stations were processed) the isolayet map for dry and wet season was composed by computing coordinates/geographic location of each station.

In order to apply VAR method to the existing data, we necessarily figure out the stationarity of data in each station initially using ADF calculation as the VAR[15]. ADF calculation for each time series data and interpretation of stationarity in each station are shown in Table 1.
By following the stages that have been described previously, based on the rounded AIC calculation (Table 2), the final VAR model which can be obtained is VAR (6). VAR (6) with the smallest AIC value (shaded area) can be represented using mathematical equation model obtained through Ordinary Least Square (OLS) approach in the following equation (4).

- For measurement station of Ungaran-Semarang, \( Y_{1,t} = -1.40,785,646 + 0.182,222,205 Y_{1,t-1} - 0.014,095,96 Y_{1,t-2} - 0.096,583,67 Y_{1,t-3} - 0.100,907,471 Y_{1,t-4} + 0.118,868,817 Y_{1,t-5} - 0.118,000,286 Y_{1,t-6} + 0.198,268,972 Y_{2,t-1} + 0.287,944 Y_{2,t-2} - 0.044,638,981 Y_{2,t-3} + 0.013,124,160 Y_{2,t-4} - 0.017,253,552 Y_{2,t-5} + 0.139,557,736 Y_{2,t-6} + 0.010,990,09 Y_{3,t-1} + 0.103,491 Y_{3,t-2} + 0.487,549 Y_{3,t-3} - 0.104,888,805 Y_{3,t-4} + 0.032,615,013 Y_{3,t-5} + 0.131,957,736 Y_{3,t-6}.\)

- For measurement station of Bringin – Salatiga, \( Y_{1,t} = -349,089 + 0.297,818,92 Y_{1,t-1} - 0.563,277,09 Y_{1,t-2} + 0.038,723,242 Y_{1,t-3} - 0.119,070,471 Y_{1,t-4} + 0.107,368,817 Y_{1,t-5} - 0.107,002,00 Y_{1,t-6} - 0.083,972,126 Y_{1,t-1} + 0.293,267,51 Y_{1,t-2} - 0.255,211,53 Y_{1,t-3} - 0.012,024,160 Y_{2,t-1} - 0.017,253,552 Y_{2,t-2} + 0.139,557,736 Y_{2,t-3} + 0.296,468,20 Y_{3,t-1} + 0.017,938,93 Y_{3,t-2} - 0.102,788,34 Y_{3,t-3} - 0.103,788,805 Y_{3,t-4} + 0.031,515,013 Y_{3,t-5} + 0.139,557,736 Y_{3,t-6}.\)

- For measurement station of Adi Sumarmo - Boyolali, \( Y_{1,t} = -468,738,158 + 0.490,067,870 Y_{1,t-1} - 0.095,125,17 Y_{1,t-2} - 0.033,233,273 Y_{1,t-3} - 0.122,907,471 Y_{1,t-4} + 0.125,368,817 Y_{1,t-5} + 0.129,000,286 Y_{1,t-6} + 0.422,738,18 Y_{2,t-1} - 1.184,634,56 Y_{2,t-2} + 0.473,073,87 Y_{2,t-3} - 0.012,244,160 Y_{2,t-4} - 0.017,475,552 Y_{2,t-5} + 0.151,757,376 Y_{2,t-6} + 3.275,684,0 Y_{3,t-1} + 1.815,258,2 Y_{3,t-2} + 2.241,157 Y_{3,t-3} - 0.103,999,805 Y_{3,t-4} + 0.053,715,013 Y_{3,t-5} + 0.158,579,936 Y_{3,t-6}.\)

- For measurement station of Borobudur - Magelang, \( Y_{1,t} = -272,493 + 0.461,066,9 Y_{1,t-1} - 0.430,968,25 Y_{1,t-2} - 0.352,013,72 Y_{1,t-3} - 0.010,097,471 Y_{1,t-4} + 0.107,368,817 Y_{1,t-5} + 0.107,000,00 Y_{1,t-6} + 0.992,350,069 Y_{2,t-1} - 0.526,134,70 Y_{2,t-2} - 1.169,981,37 Y_{2,t-3} - 0.018,084,160 Y_{2,t-4} - 0.151,053,552 Y_{2,t-5} + 0.139,225,736 Y_{2,t-6} - 0.180,966,27 Y_{3,t-1} - 0.221,797,260 Y_{3,t-2} + 0.176,700,05 Y_{3,t-3} - 0.103,566,805 Y_{3,t-4} + 0.031,515,123 Y_{3,t-5} + 0.139,553,536 Y_{3,t-6}.\)

Where

- \( Y_{1,t} \) is rainfall value at the-t time.
- \( Y_{2,t} \) is humidity value at the-t time.
- \( Y_{3,t} \) is temperature value at the-t time.

### TABLE III. MAE AND MAPE CALCULATION

<table>
<thead>
<tr>
<th>Rainfall Station</th>
<th>MAE</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ungaran - Semarang</td>
<td>6.95128</td>
<td>2.139876</td>
</tr>
<tr>
<td>Bringin - Salatiga</td>
<td>13.15702</td>
<td>6.008978</td>
</tr>
<tr>
<td>Adisumarmo - Boyolali</td>
<td>10.23671</td>
<td>6.558561</td>
</tr>
<tr>
<td>Borobudur - Magelang</td>
<td>13.02692</td>
<td>5.713805</td>
</tr>
<tr>
<td>Semarang Kota</td>
<td>12.36196</td>
<td>6.707421</td>
</tr>
</tbody>
</table>

For calculating the accuracy of data processing, the calculation of MAE and MAPE value for each measurement station of which the result can be seen in Table 3 above is required. In general, the data processing result present the quite well values (the relatively small value of MAE and MAPE value that are below the range of 10%), hence VAR (6) model can be stated to be pretty well to forecast the rainfall in the study area (Semarang).
in certain area (sub-district) in dry and wet season as well, hence the plant that will be planted in the related area can be determined afterwards regarding the characteristic (water requirement) of the plant.

V. CONCLUSION

VAR (6) model can be applied well to forecast the rainfall in Semarang in dry and wet season. VAR (6) is used for the reason that through ADF calculation, each time series in the existing measurement stations is all stationary. Meanwhile, VAR (6) is taken as regarding the AIC calculation, this model has the lowest/smallest value than the other VAR models. VAR (6) can be applied well in Semarang as having the relatively small values of MAE and MAPE (valued below 10%) and smaller than ARIMA model used by BMKG (valued about 22,13%). Based on the mathematical model formulated with VAR (6) model based, isohyet map for each dry and wet season was made. It will be beneficial for the decision making stages as showing the correlation between certain areas (in this case – certain sub-districts in Semarang) and the rainfall (forecast) in the related areas.

REFERENCES


AUTHOR PROFIIL

Adi Nugroho earned a bachelor's degree from Geological Engineering of Bandung Institute of Technology (ITB), Indonesia and a master's degree majoring in Information System Management in Gunadarma University, Jakarta, Indonesia. He is currently finishing his doctoral program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia. He has interests in programming, databases, and software engineering. He is currently a lecturer in Faculty of Information Technology of Satya Wacana Christian University, Salatiga, Indonesia.

Subanar earned a bachelor's degree from Gadjah Mada University, Yogyakarta, Indonesia and a doctoral degree from Wisconsin University, United States. His field of research is statistics. He currently serves as a professor in Doctoral Program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia.

Sri Hartati earned a bachelor's degree from Department of Electronics and Instrumentation of Gadjah Mada University, Indonesia and a master's degree majoring in Computer Science (Spatial Processing) in New Brunswick University, Canada. Her doctoral degree was also earned from New Brunswick University, Canada majoring in Computer Science (Artificial Intelligence). She currently serves as a professor in Doctoral Program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia.

Khabib Mustofa earned a bachelor's and a master's degree from Department of Computer Science of Gadjah Mada University, Indonesia and a doctoral degree from Vienna University of Technology, Austria majoring in Computer Science. His fields of research are Web Technology, Semantic Web, Software Engineering, and Information Management. He currently serves as a lecturer in Doctoral Program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia.