VAR and GSTAR-Based Feature Selection in Support Vector Regression for Multivariate Spatio-Temporal Forecasting

Dedy Dwi Prastyo[1(✉)], Feby Sandi Nabila[1], Suhartono[1], Muhammad Hisyam Lee[2], Novri Suhermi[1], and Soo-Fen Fam[3]

[1] Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
dedy-dp@statistika.its.ac.id
[2] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Malaysia
[3] Department of Technopreneurship, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Abstract. Multivariate time series modeling is quite challenging particularly in term of diagnostic checking for assumptions required by the underlying model. For that reason, nonparametric approach is rapidly developed to overcome that problem. But, feature selection to choose relevant input becomes new issue in nonparametric approach. Moreover, if the multiple time series data are observed from different sites, then the location possibly play the role and make the modeling become more complicated. This work employs Support Vector Regression (SVR) to model the multivariate time series data observed from three different locations. The feature selection is done based on Vector Autoregressive (VAR) model that ignore the spatial dependencies as well as based on Generalized Spatio-Temporal Autoregressive (GSTAR) model that involves spatial information into the model. The proposed approach is applied for modeling and forecasting rainfall in three locations in Surabaya, Indonesia. The empirical results inform that the best method for forecasting rainfall in Surabaya is the VAR-based SVR approach.

Keywords: SVR · VAR · GSTAR · Feature selection · Rainfall

1 Introduction

Global warming has caused climate change which affected the rainfall. As a tropical country, Indonesia has various rainfall pattern and different amount of rainfall in each region. The rainfall becomes hard to predict because of this disturbance. The climate change that is triggered by the global warming causes the rainfall pattern becomes more uncertain. This phenomenon affects the agricultural productivity, for example, in East Java province, Indonesia [1], United State [2], and Africa [3]. The capital city of East Java, Surabaya, also suffers climate change as the effect of global warming.

The rainfall has a huge variance in spatial and time scale. Therefore, it is necessary to apply a univariate or multivariate modeling to predict rainfall. One of the multivariate

B. W. Yap et al. (Eds.): SCDS 2018, CCIS 937, pp. 46–57, 2019.
https://doi.org/10.1007/978-981-13-3441-2_4

models commonly used is Vector Autoregressive Moving Average (VARMA), which is an expansion of ARMA model [4]. If the spatial effect from different locations is considered, then Generalized Space Time Autoregressive (GSTAR) model play into role.

In this research, we apply Vector Autoregressive (VAR) and GSTAR to model the rainfall. The VAR model does not involve location (spatial) information, while GSTAR model accommodates the heterogeneous locations by adding the weight to each location. The comparison and application of VAR and GSTAR models has already done Suhartono *et al*. [5] to determine the input in Feed-forward Neural Network (FFNN) as nonparametric approach. There are two types of time series prediction approach: parametric approach and nonparametric approach. Another nonparametric approach which is widely used is Support Vector Regression (SVR) as the modification of Support Vector Machine (SVM) [6–9] which handles the regression task. The main concept of SVR is to maximize the margin around the hyper plane and to obtain data points that become the support vectors.

This work does not handle outliers if they exist. This paper is organized as follows. Section 2 explains the theoretical part. Section 3 describes the methodology. Section 4 informs empirical results and discussion. At last, Sect. 5 shows the conclusion.

2 Literature Review

2.1 Vector Autoregressive (VAR) Model

The VAR model order one, abbreviated as VAR(1), is formulated in Eq. (1) [4]:

$$\dot{\boldsymbol{Y}}_t = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}\dot{\boldsymbol{Y}}_{t-1} + \boldsymbol{\alpha}_t, \tag{1}$$

where $\dot{\boldsymbol{Y}}_t = \boldsymbol{Y}_t - \boldsymbol{\mu}$, with $\boldsymbol{\mu} = E(\boldsymbol{Y}_t)$. The $\boldsymbol{\alpha}_t$ is $m \times 1$ vector of residual at time t, $\dot{\boldsymbol{Y}}_t$ is $m \times 1$ vector of variables at t, and $\dot{\boldsymbol{Y}}_{t-1}$ is $m \times 1$ vector of variables at $(t-1)$. The parameter estimation is conducted using conditional least square (CLS). Given m series with T data points each, then VAR(p) model could be expressed by (2).

$$\boldsymbol{y}_t = \boldsymbol{\delta} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{a}_t. \tag{2}$$

Equation (2) can also be expressed in the form of linear model as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{A} \tag{3}$$

and

$$\boldsymbol{y} = \left(\boldsymbol{X}^T \otimes \boldsymbol{I}_m\right)\boldsymbol{\beta} + \boldsymbol{a}, \tag{4}$$

with $\otimes$ is Kronecker product, $\boldsymbol{Y} = (y_1, \ldots, y_T)_{(m \times T)}$, $\boldsymbol{B} = \left(\boldsymbol{\delta}, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p\right)_{(m \times (mp+1))}$, and $\boldsymbol{X} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_t, \ldots, \boldsymbol{X}_{T-1})_{((mp+1) \times T)}$. The vector of data at time t is

$$\boldsymbol{X}_t = \begin{pmatrix} 1 \\ y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix}_{((mp+1)\times 1)}$$

and $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)_{(m\times T)}$, $\boldsymbol{y} = (\mathrm{vec}(\boldsymbol{Y}))_{(mT\times 1)}$, $\boldsymbol{\beta} = (\mathrm{vec}(\boldsymbol{B}))_{((m^2p+m)\times 1)}$, and $\boldsymbol{a} = (\mathrm{vec}(\boldsymbol{A}))_{(mT\times 1)}$. The *vec* denotes a column stacking operator such that:

$$\widehat{\boldsymbol{\beta}} = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \otimes \boldsymbol{I}_m\right)\boldsymbol{y}. \tag{5}$$

The consistency property and asymptotic normality property of the CLS estimate $\widehat{\boldsymbol{\beta}}$ is shown in the following equation.

$$\sqrt{T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} N\left(0, \boldsymbol{\Gamma}_p^{-1} \otimes \sum\right), \tag{6}$$

where $\boldsymbol{X}'\boldsymbol{X}/T$ converges in probability towards $\boldsymbol{\Gamma}_p$ and $\xrightarrow{\mathrm{d}}$ denotes the convergence in distribution. The estimate for $\boldsymbol{\Sigma}$ is given as follows:

$$\widehat{\Sigma} = (T - (mp+1))^{-1}\sum_{t=1}^{T} \widehat{\boldsymbol{a}}_t\widehat{\boldsymbol{a}}_t', \tag{7}$$

where $\widehat{\boldsymbol{a}}_t$ is the residual vector.

2.2 Generalized Space Time Autoregressive (GSTAR) Model

Given a multivariate time series $\{\boldsymbol{Y}(t) : t = 0, \pm 1, \pm 2, \ldots\}$ with T observations for each series, the GSTAR model for order one with 3 locations is given as [5, 10, 11]:

$$\boldsymbol{Y}(t) = \boldsymbol{\Phi}_{10}\boldsymbol{Y}(t-1) + \boldsymbol{\Phi}_{11}\boldsymbol{W}^{(l)}\boldsymbol{Y}(t-1) + \boldsymbol{a}(t), \tag{8}$$

with $\boldsymbol{Y}(t)$ is $(T \times 1)$ random vector at t, $\boldsymbol{\Phi}_{10}$ is a matrix of coefficient, $\boldsymbol{\Phi}_{11}$ is spatial coefficient matrix, and $\boldsymbol{W}^{(l)}$ is an $(m \times m)$ weight matrix at spatial lag l. The weight must satisfy $\boldsymbol{w}_{ii}^{(l)} = 0$ and $\sum_{i\neq j}^{(l)} \boldsymbol{w}_{ij}^{(l)} = 1$. The $\boldsymbol{a}(t)$ is vector of error which satisfies i.i.d and multivariate normally distributed assumption with $\boldsymbol{0}$ vector mean and variance-covariance matrix $\sigma^2\boldsymbol{I}_m$.

Uniform Weighting

Uniform weighting assumes that the locations are homogenous such that:

$$W_{ij} = \frac{1}{n_i}, \tag{9}$$

where n_i is the number of near location and W_{ij} is the weight location i and j.

Inverse Distance Weighting (IDW)
The IDW method is calculated based on the real distance between locations. Then, we calculate the inverse of the real distance and normalize it.

Normalized Cross-Correlation Weighting
Normalized cross-correlation weighting uses the cross-correlation between locations at the corresponding lags. In general, the cross-correlation between location i and location j at time lag k, i.e. the $corr\left[Y_i(t), Y_j(t-k)\right]$, is defined as follows:

$$\rho_{ij}(k) = \frac{\gamma_{ij}(k)}{\sigma_i \sigma_j}, k = 0, \pm 1, \pm 2, \ldots, \tag{10}$$

where $\gamma_{ij}(k)$ is cross-covariance in location i and location j. The sample cross-correlation can be computed using the following equation.

$$r_{ij}(k) = \frac{\sum_{t=k+1}^{T}\left(Y_i(t) - \overline{Y}_i\right)\left(Y_j(t-k) - \overline{Y}_j\right)}{\sqrt{\sum_{t=1}^{T}\left(Y_i(t) - \overline{Y}_i\right)^2 \sum_{t=1}^{T}\left(Y_j(t) - \overline{Y}_j\right)^2}}. \tag{11}$$

The weighting is calculated by normalizing the cross-correlation between locations. This process generally results in location weight for GSTAR (1_1) model, which is as follows:

$$w_{ij} = \frac{r_{ij}(1)}{\sum_{j \neq i}\left|r_{ij}(1)\right|} \text{for } i \neq j. \tag{12}$$

2.3 Support Vector Regression (SVR)

The SVR is developed from SVM as a learning algorithm which uses hypothesis that there are linear functions in a high dimensional feature space [6–9]. SVM for regression uses $\varepsilon -$ insensitive loss function which is known as SVR. The regression function of SVR is perfect if and only if the deviation bound equals zero such that:

$$f(x) = w^T \varphi(x) + b, \tag{13}$$

where w is weight and b is bias. The notation $\varphi(x)$ denotes a point in feature space $\mathcal{F}$ which is a mapping result of x in an input space. The coefficients w and b is aimed to minimize following risk.

$$R(f(x)) = \frac{C}{n}\sum_{i=1}^{n} L_\varepsilon(y_i, f(x_i)) + \frac{1}{2}\|w\|^2 \tag{14}$$

Where

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} 0 & ; |y_i - f(x_i)| \le \varepsilon, \\ |y_i - f(x_i)| - \varepsilon & ; \text{otherwise.} \end{cases} \tag{15}$$

The L_ε is a ε–insensitive loss function, y_i is the vector of observation, C and ε are the hyper parameters. The function f is assumed to approximate all the points (x_i, y_i) with precision ε if all the points are inside the interval. While infeasible condition happens when there are several points outside the interval $f \pm \varepsilon$. The infeasible points can be added a slack variable ξ, ξ^* in order to tackle the infeasible constrain. Hence, the optimization in (14) can be transformed into the following.

$$\min \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\xi_i + \xi_i^*\right), \tag{16}$$

with constrains $(w^T\varphi(x_i) + b) - y_i \le \varepsilon + \xi_i^*; y_i - (w^T\varphi(x_i) - b) \le \varepsilon + \xi_i$ and $\xi, \xi^* \ge 0$, and $i = 1, 2, \ldots, n$. The optimization in that constrain can be solved using primal Lagrange:

$$\begin{aligned} &L\left(w, b, \xi, \xi^*, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*\right) = \\ &\tfrac{1}{2}\|w\|^2 + C\left(\textstyle\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)\right) - \textstyle\sum_{i=1}^{n}\beta_i\left[w^T\varphi(x_i) + b - y_i + \varepsilon + \xi_i^*\right] - \\ &\textstyle\sum_{i=1}^{n}\beta_i^*\left[y_i - w^T\varphi(x_i) - b + \varepsilon + \xi_i^*\right] - \textstyle\sum_{i=1}^{n}\left(\alpha_i\xi_i + \alpha_i^*\xi_i^*\right) \end{aligned} \tag{17}$$

The Eq. (17) is minimized in primal variables w, b, ξ, ξ^* and maximized in the form of non-negative Lagrangian multiplier $\alpha_i, \alpha_i^*, \beta_i, \beta_i^*$. Then, we obtain a dual Lagrangian with kernel function $K\left(x_i, x_j\right) = \varphi(x_i)^T\varphi\left(x_j\right)$. One of the most widely used kernel function is *Gaussian radial basis function* (RBF) formulated in (18) [6]:

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j^2\right\|}{2\sigma^2}\right) \tag{18}$$

$$\begin{aligned} \partial(\beta_i, \beta_i^*) = &\sum\nolimits_{i=1}^{n} y_i(\beta_i - \beta_i^*) - \varepsilon\sum\nolimits_{i=1}^{n}(\beta_i + \beta_i^*) \\ &-\frac{1}{2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}(\beta_i - \beta_i^*)(\beta_j - \beta_j^*)K\left(x_i, x_j\right). \end{aligned} \tag{19}$$

Then, we obtain the regression function as follows.

$$f\left(x, \beta_i, \beta_i^*\right) = \sum_{i=1}^{l}(\beta_i - \beta_i^*)K\left(x_i, x_j\right) + b. \tag{20}$$

The SVM and SVR are extended for various fields. It is also developed for modeling and analyzing survival data, for example, done by Khotimah *et al*. [12, 13].

2.4 Model Selection

The model selection is conducted using out-of-sample criteria by comparing the multivariate *Root Mean Square Error* (RMSE). The RMSE of a model is obtained using the Eq. (11) for training dataset and (12) for testing dataset, respectively.

$$RMSE_{in} = \sqrt{MSE_{in}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}, \tag{21}$$

where n is the effective number of observation in training dataset.

$$RMSE_{out} = \sqrt{\frac{1}{n_{out}}\sum_{i=1}^{n_{out}}\left(Y_{n+l} - \widehat{Y}_n(l)\right)^2}, \tag{22}$$

with l is the forecast horizon.

3 Data and Method

The dataset that is used in this research is obtained from *Badan Meteorologi, Klimatologi, dan Geofisika* (BMKG) at Central of Jakarta. The rainfall is recorded from 3 stations: Perak I, Perak II, and Juanda for 34 years and 10 months. The dataset is divided into in-sample (training) and out-of-sample (testing) data. The in-sample spans from January 1981 to December 2013. The data from January 2014 to November 2015 is out-of-sample for evaluating the forecast performance.

The analysis is started by describing the rainfall pattern from 3 locations and model them using VAR and GSTAR. Once the right model with significant variables is obtained, it is continued with VAR-SVR and GSTAR-SVR modeling using the variables obtained from the VAR and GSTAR, respectively [5]. This kind of feature selection using statistical model was also studied by Suhartono *et al*. [14].

4 Empirical Results

4.1 Descriptive Statistics

The descriptive statistics of accumulated rainfall from three locations is visualized in Fig. 1. The means of rainfall at station Perak 1, Perak 2, and Juanda are respectively 45.6 mm, 43.1 mm, and 58.9 mm. These means are used as threshold to find period when the rainfall is lower or greater at each location. Figure 1 shows that from April to May there is a shift from rain season to dry season. There is a shift from dry season to rain season in November. The yellow boxplots show the rainfall average at that period is lower than overall mean, while the blue ones show the opposite. The yellow and blue boxplots are mostly in dry season and rain season, respectively.

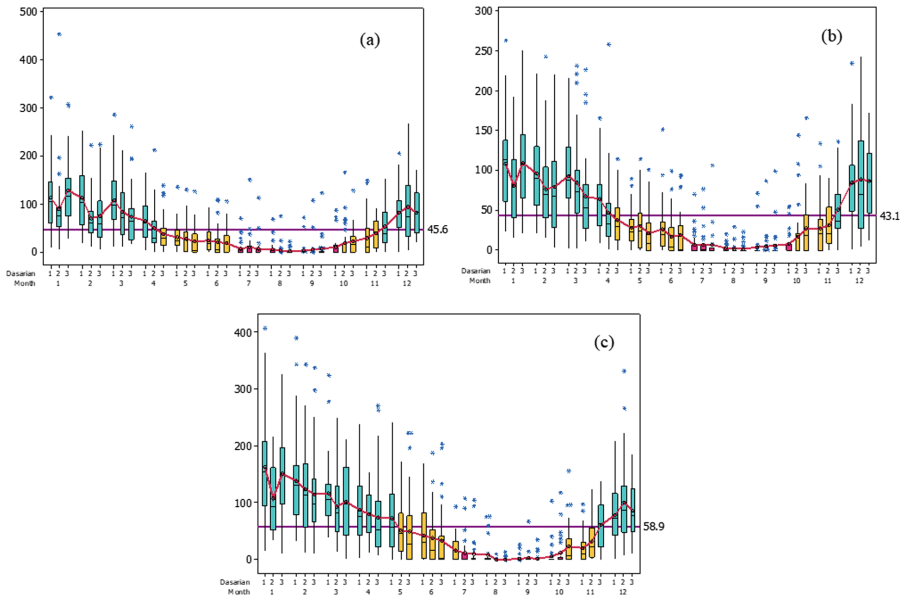

Fig. 1. The boxplots for rainfall per month and dasarian at station Perak 1 (a); Perak 2 (b); and Juanda (c) (Color figure online)

4.2 VAR Modeling

Order identification in VAR is conducted based on partial cross-correlation matrix from stationary data, after being differenced at lag 36. Lag 1 and 2 are significant such that we use non-seasonal order 2 in our model. We also use seasonal orders 1, 2, 3, 4, and 5 since we can see that lags 36, 72, 108, and 144 are still significant. Hence, we have 5 candidate models: VARIMA $(2,0,0)(1,1,0)^{36}$, VARIMA $(2,0,0)(2,1,0)^{36}$, VARIMA $(2,0,0)(3,1,0)^{36}$, VARIMA $(2,0,0)(4,1,0)^{36}$, and VARIMA $(2,0,0)(5,1,0)^{36}$.

VAR's residual must satisfy white noise and multivariate normality assumptions. The test results show none of the model satisfies the assumptions for $\alpha = 5\%$. The Root Mean Square Error (RMSE) of the out-of-sample prediction for five models is summarized in Table 1.

Table 1. Cross tabulation before and after merging stadium category.

Model	Location			Overall RMSE
	Perak 1	Perak 2	Juanda	
VARIMA $(2,0,0)(1,1,0)^{36}$	62.32484	47.51616	56.44261	55.76121
VARIMA $(2,0,0)(2,1,0)^{36}$	45.70456	41.19213	48.4053*	45.19871
VARIMA $(2,0,0)(3,1,0)^{36}$	40.13943	37.04678*	51.56591	44.76059
VARIMA $(2,0,0)(4,1,0)^{36}$	39.44743*	36.86216	49.98893	42.48063*
VARIMA $(2,0,0)(5,1,0)^{36}$	41.60190	37.73775	49.20983	43.11405

*Minimum RMSE

Table 1 shows that VARIMA $(2,0,0)(4,1,0)^{36}$ has the smallest overall RMSE. Hence, we choose it as the best model. The equation of VAR model for location Perak 1 is given as follows.

$$\begin{aligned}\widehat{y1_t} = \; & y1_{t-36} + 0.08576(y1_{t-1} - y1_{t-36}) - 0,1242(y1_{t-2} - y1_{t-38}) + \\ & 0.15702(y2_{t-2} - y2_{t-38}) + 0.0564(y3_{t-2} - y3_{t-38}) - 0.73164(y1_{t-36} - \\ & y1_{t-72}) - 0.60705(y1_{t-72} - y1_{t-108}) - 0.50632(y1_{t-108} - y1_{t-144}) + \\ & 0.14915(y2_{t-108} - y2_{t-144}) - 0.21933(y1_{t-144} - y1_{t-180})\end{aligned}$$

VAR model for location Perak 1 show that the rainfall in that location is also influenced by the rainfall in other location. The equation of VAR models for location Perak 2 and Juanda are given as follows, respectively.

$$\begin{aligned}\widehat{y2_t} = \; & y2_{t-36} + 0,07537(y1_{t-1} - y1_{t-36}) - 0.10323(y1_{t-2} - y1_{t-38}) + \\ & 0.09673(y2_{t-2} - y2_{t-38}) + 0.0639(y3_{t-2} - y3_{t-38}) + 0.16898(y1_{t-36} - \\ & y1_{t-72}) - 0.8977(y2_{t-36} - y2_{t-72}) + 0.07393(y1_{t-72} - y1_{t-108}) - \\ & 0.70884(y2_{t-72} - y2_{t-108}) - 0.38961(y2_{t-108} - y2_{t-144}) - \\ & 0.20648(y1_{t-144} - y1_{t-180})\end{aligned}$$

$$\begin{aligned}\widehat{y3_t} = \; & y3_{t-36}0.0767(y1_{t-1} - y1_{t-36}) + 0.06785(y3_{t-1} - y3_{t-36}) - 0.09053(y1_{t-2} - y1_{t-38}) \\ & + 0.11712(y3_{t-2} - y3_{t-38}) - 0.74691(y3_{t-36} - y3_{t-72}) - \\ & 0.07787(y2_{t-72} - y2_{t-108}) - 0.58127(y3_{t-72} - y3_{t-108}) - \\ & 0.32507(y3_{t-108} - y3_{t-144}) - 0.16491(y3_{t-144} - y3_{t-180})\end{aligned}$$

The rainfall at station Perak 2 and Juanda are also influenced by the rainfall in other locations.

4.3 GSTAR Modeling

We choose GSTAR ([1,2,3,4,5,6,36,72]1)-I(1)$(1)^{36}$ as our model. Residual assumption checking in GSTAR shows that this model does not satisfy assumptions for $\alpha = 5\%$. The prediction for out-of-sample is done with two scenarios: using all the variables and using only the significant variables. The results are shown in the Table 2.

Table 2. The RMSEs of out-of-sample from STAR ([1,2,36,72,108,144,180]1)-I$(1)^{36}$

Location	Model with all variables			Model with only significant variables		
	Uniform	Inverse distance	Cross correlation	Uniform	Inverse distance	Cross correlation
Perak 1	271.2868	62.2626	61.7325	68.7310	61.7713	61.0599*
Perak 2	55.1563*	56.1579	56.1562	66.2657	56.1511	55.6316
Juanda	79.1952	75.5786*	76.776	109.9268	76.0599	76.1471
Total	166.2688	66.3322	66.2888*	84.0611	65.2016	64.8629*

*Minimum RMSE

The GSTAR model equations for locations Perak 1, Perak 2, and Juanda are given in the following, respectively.

$$\begin{aligned}\widehat{y1_t} = \; & y1_{t-36} + 0.0543608(y2_{t-1} - y2_{t-37}) + 0.042241(y3_{t-1} - y3_{t-37}) + \\ & 0.0444963(y2_{t-2} - y2_{t-38}) + 0.034576(y3_{t-2} - y3_{t-38}) + \\ & (-0.81419(y1_{t-36} - y1_{t-72})) + (-0.69679(y1_{t-72} - y1_{t-108})) + \\ & (-0.57013(y1_{t-108} - y1_{t-144})) + 0.039493(y2_{t-108} - y2_{t-144}) + \\ & 0.030688(y3_{t-108} - y3_{t-144}) + (-0.34571(y1_{t-144} - y1_{t-180})) + \\ & (-0.21989(y1_{t-180} - y1_{t-216})) + 0.038646(y2_{t-180} - y2_{t-216}) + \\ & 0.0300298(y3_{t-180} - y3_{t-216})\end{aligned}$$

and

$$\begin{aligned}\widehat{y2_t} = \; & y2_{t-36} + 0,035769(y1_{t-1} - y1_{t-37}) + 0.040401(y3_{t-1} - y3_{t-37}) + \\ & 0.029355(y2_{t-2} - y2_{t-38}) + 0.033157(y3_{t-2} - y3_{t-38}) + (-0.84879(y2_{t-36} - \\ & y2_{t-72})) + 0.022607(y1_{t-36} - y1_{t-72}) + 0.0255357(y3_{t-36} - y3_{t-72}) + \\ & (-0.72905(y2_{t-72} - y2_{t-108})) + (-0.55416(y2_{t-108} - y2_{t-144})) + \\ & (-0.36629(y2_{t-144} - y2_{t-180})) + (-0.20078(y2_{t-180} - y2_{t-216})).\end{aligned}$$

and

$$\begin{aligned}\widehat{y3_t} = \; & y3_{t-36} + 0.103666(y3_{t-1} - y3_{t-37}) + 0.089536(y3_{t-2} - y3_{t-38}) + \\ & (-0.78084(y3_{t-36} - y3_{t-72})) + (-0.6609(y3_{t-72} - y3_{t-108})) + \\ & (-0.44531(y3_{t-108} - y3_{t-144})) + (-0.31721(y3_{t-144} - y3_{t-180})) + \\ & (-0.18268(y3_{t-180} - y3_{t-216})).\end{aligned}$$

4.4 Forecasting Using VAR-SVR and GSTAR-SVR Model

The VAR-SVR and GSTAR-SVR modeling use grid search method to determine the hyper parameters, i.e. epsilon, sigma, and cost. Finding these hyper parameters values is in purpose of to obtain the minimum RMSE of out-of-sample data. VAR-SVR model uses the variables of VARIMA $(2,0,0)(4,1,0)^{36}$, which is the best VAR model, as the inputs. Then, GSTAR-SVR model uses the significant variables of GSTAR $([1,2,3,4,5,6,36,72]1)\text{-I}(1)(1)^{36}$ with normalized cross-correlation weight. The prediction of the out-of-sample data (testing data) are given in Table 3. It shows that the RMSE of VAR-SVR model at Perak 2 is the smallest. It means that VAR-SVR model performs better at Perak 2 than other locations.

Table 3. The VAR-SVR model with smallest RMSE

Location	Epsilon	Cost	Sigma	RMSE Out sample
Perak 1	8.67×10^{-4}	2270	1.3×10^{-7}	38.57858
Perak 2	8.65×10^{-4}	2100.1	1.09×10^{-7}	34.03217
Juanda	8.69×10^{-4}	3001	1.08×10^{-7}	47.75733

The results in Table 4 also show that the RMSE of GSTAR-SVR model at Perak 2 is the smallest. Compared to GSTAR-SVR, the best model with the smallest overall RMSE is VAR-SVR. The VAR-SVR and GSTAR-SVR models are used to forecast the rainfall from November 2015 to November 2016. The forecast results in testing dataset as well as one year ahead forecasting are given in Figs. 2 and 3.

Table 4. The GSTAR-SVR model with smallest RMSE

Location	Epsilon	Cost	Sigma	RMSE out sample
Perak 1	9×10^{-5}	355	5×10^{-7}	41.68467
Perak 2	8×10^{-8}	450	3×10^{-7}	32.90443
Juanda	10^{-9}	280	7×10^{-7}	50.33458

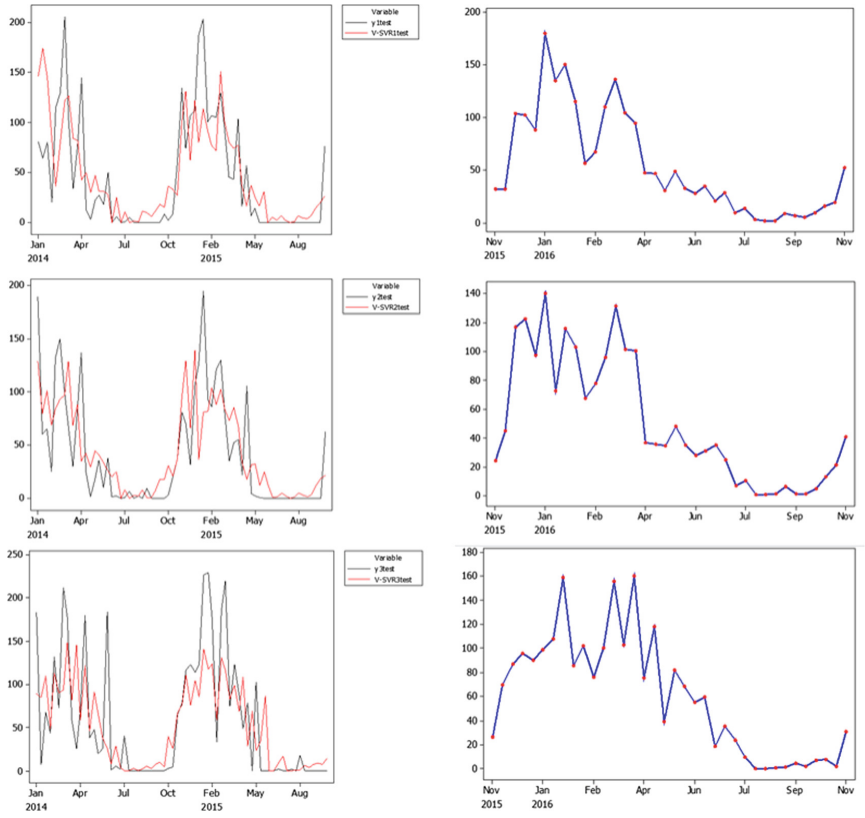

Fig. 2. The rainfall observation (black line) and its forecast (red line) at testing dataset using VAR-SVR model at station Perak 1 (top left); Perak 2 (middle left); Juanda (bottom left); and one-year forecasting (right) at each location. (Color figure online)

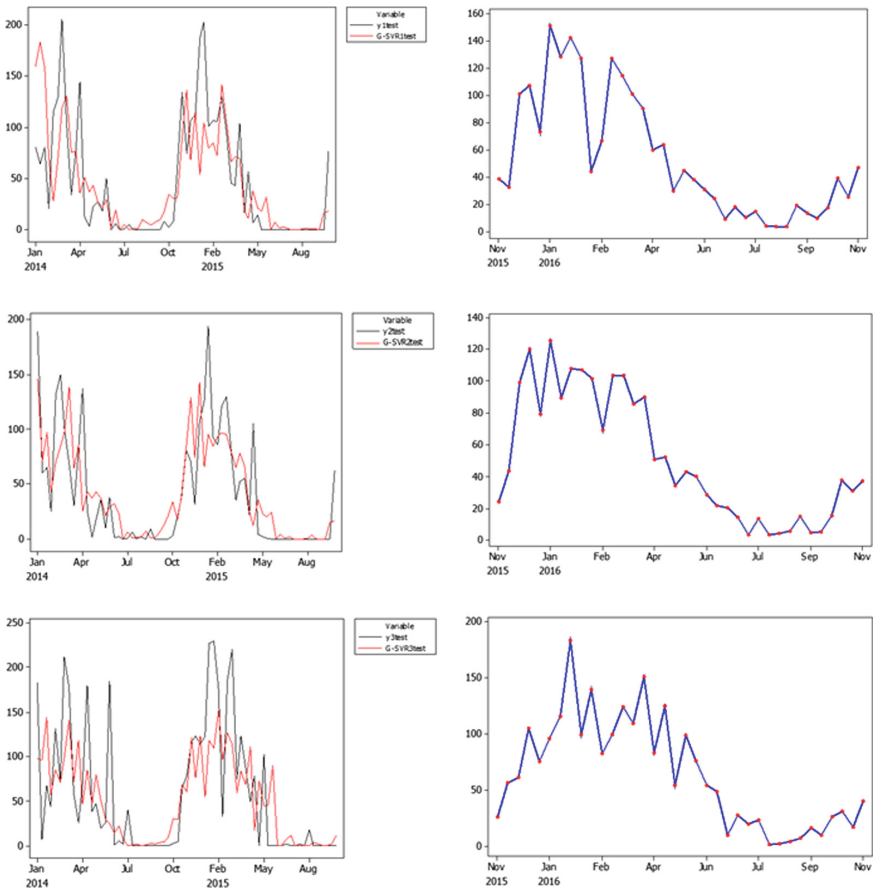

Fig. 3. The rainfall observation (black line) and its forecast (red line) at testing dataset using GSTAR-SVR model at station Perak 1 (top left); Perak 2 (middle left); Juanda (bottom left); and one-year forecasting (right) at each location. (Color figure online)

5 Conclusion

First, the best VARIMA model used to forecast rainfall in Surabaya is VARIMA $(2,0,0)(4,1,0)^{36}$. Second, the forecast of GSTAR $([1,2,3,4,5,6,36,72]1)$-$I(1)(1)^{36}$ using the only significant input (restricted form) and normalized cross-correlation weight resulted in the smallest RMSE than the other GSTAR forms. Third, the hybrid VAR-based SVR model with VARIMA $(2,0,0)(4,1,0)^{36}$ as feature selection produced smallest RMSE than other models. Thus, the spatial information does not improve the feature selection of SVR approach used in this analysis.

Acknowledgement. This research was supported by DRPM under the scheme of "Penelitian Dasar Unggulan Perguruan Tinggi (PDUPT)" with contract number 930/PKS/ITS/2018. The

authors thank to the General Director of DIKTI for funding and to the referees for the useful suggestions.

References

1. Kuswanto, H., Salamah, M., Retnaningsih, S.M., Prastyo, D.D.: On the impact of climate change to agricultural productivity in East Java. J. Phys: Conf. Ser. **979**(012092), 1–8 (2018)
2. Adams, R.M., Fleming, R.A., Chang, C.C., McCarl, B.A., Rosenzweig, C.: A reassessment of the economic effects of global climate change on U.S. agriculture. Clim. Change **30**(2), 147–167 (1995)
3. Schlenker, W., Lobell, D.B.: Robust negative impacts of climate change on African agriculture. Environ. Res. Lett. **5**(014010), 1–8 (2010)
4. Tsay, R.S.: Multivariate Time Series Analysis. Wiley, Chicago (2014)
5. Suhartono, Prastyo, D.D., Kuswanto, H., Lee, M.H.: Comparison between VAR, GSTAR, FFNN-VAR, and FFNN-GSTAR models for forecasting oil production. Matematika **34**(1), 103–111 (2018)
6. Haerdle, W.K., Prastyo, D.D., Hafner, C.M.: Support vector machines with evolutionary model selection for default prediction. In: Racine, J., Su, L., Ullah, A. (eds.) The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, pp. 346–373. Oxford University Press, New York (2014)
7. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM, Pittsburgh (1992)
8. Smola, A.J., Scholköpf, B.: A tutorial on support vector regression, statistics and computing. Stat. Comput. **14**(3), 192–222 (2004)
9. Suykens, J.A., Vandewalle, J.: Least squares support vector machines classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
10. Borovkova, S., Lopuhaä, H.P., Ruchjana, B.N.: Consistency and asymptotic normality of least squares estimators in Generalized STAR models. Stat. Neerl. **62**(4), 482–508 (2008)
11. Bonar, H., Ruchjana, B.N., Darmawan, G.: Development of generalized space time autoregressive integrated with ARCH error (GSTARI - ARCH) model based on consumer price index phenomenon at several cities in North Sumatra province. In: Proceedings of the 2nd International Conference on Applied Statistics (ICAS II). AIP Conference Proceedings 1827 (020009), Bandung (2017)
12. Khotimah, C., Purnami, S.W., Prastyo, D.D., Chosuvivatwong, V., Spriplung, H.: Additive survival least square support vector machines: a simulation study and its application to cervical cancer prediction. In: Proceedings of the 13th IMT-GT International Conference on Mathematics, Statistics and their Applications (ICMSA). AIP Conference Proceedings 1905 (050024), Kedah (2017)
13. Khotimah, C., Purnami, S.W., Prastyo, D.D.: Additive survival least square support vector machines and feature selection on health data in Indonesia. In: Proceedings of the International Conference on Information and Communications Technology (ICOIACT). IEEE Xplore (2018)
14. Suhartono, Saputri, P.D., Amalia, F.F., Prastyo, D.D., Ulama, B.S.S.: Model selection in feedforward neural networks for forecasting inflow and outflow in Indonesia. In: Mohamed, A., Berry, M., Yap, B. (eds.) SCDS 2017. CCIS, vol. 788, pp. 95–105. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7242-0_8