CAUCHY - JOURNAL OF PURE AND APPLIED MATHEMATICS
Volume 4(2) (2016), Pages 57-64

Seemingly Unrelated Regression Approach for GSTARIMA Model to Forecast Rain Fall Data in Malang Southern Region Districts

Siti Choirun Nisak

Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia

Email: nisak.statistika@gmail.com

ABSTRACT

Time series forecasting models can be used to predict phenomena that occur in nature. Generalized Space Time Autoregressive (GSTAR) is one of time series model used to forecast the data consisting the elements of time and space. This model is limited to the stationary and non-seasonal data. Generalized Space Time Autoregressive Integrated Moving Average (GSTARIMA) is GSTAR development model that accommodates the non-stationary and seasonal data. Ordinary Least Squares (OLS) is method used to estimate parameter of GSTARIMA model. Estimation parameter of GSTARIMA model using OLS will not produce efficiently estimator if there is an error correlation between spaces. Ordinary Least Square (OLS) assumes the variance-covariance matrix has a constant error $\varepsilon_{ij} \sim NID(\mathbf{0}, \boldsymbol{\sigma^2})$ but in fact, the observatory spaces are correlated so that variance-covariance matrix of the error is not constant. Therefore, Seemingly Unrelated Regression (SUR) approach is used to accommodate the weakness of the OLS. SUR assumption is $\varepsilon_{ij} \sim NID(\mathbf{0}, \boldsymbol{\Sigma})$ for estimating parameters GSTARIMA model. The method to estimate parameter of SUR is Generalized Least Square (GLS). Applications GSTARIMA-SUR models for rainfall data in the region Malang obtained GSTARIMA models ((1)(1,12,36),(0),(1))-SUR with determination coefficient generated with the average of 57.726%.

Keywords: Space Time, GSTARIMA, OLS, SUR, GLS, Rainfall.

INTRODUCTION

Time series data is data based on the sequence of time within a certain time span and modeled by time series models either univariate or multivariate [1]. If the data consists of the elements of time and space is modeled using multivariate models of space-time. One of the multivariate models are most often used for data modeling space-time is a Generalized Space-Time Autoregressive (GSTAR) model introduced by [2]. GSTAR has limitations that can only be used for data space-time that are stationary and non-seasonal. This condition tends to not be met at the data that is not stationary and containing a seasonal pattern.

GSTARIMA first implemented by Sun, et al. (2010) for forecasting traffic flow data in Beijing. GSTARIMA is the development of GSTAR model for data non-stationary and seasonal. GSTARIMA more flexible and practical for every space observatory has its own real-time parameter and not influenced by other changes in the observation location.

A method that can be used in parameter estimation of GSTARIMA include Ordinary Least Square method (OLS) and approach to the system of equations Seemingly Unrelated Regression (SUR). OLS assumes the variance-covariance matrix has a constant error. Thus, the equations system of Seemingly Unrelated Regression (SUR) is used to overcome the weakness of the OLS.

Submitted: 17 February 2016 Reviewed: 1 March 2016 Accepted: 1 May 2016

SUR equations system is used because it can accommodate correlations between space of the rainfall observatories. Parameter estimation of SUR can use the Generalized Least Square (GLS) method. Discussion and implementation of the model GSTARIMA still a bit to do so in this research will be modeling GSTARIMA approach SUR forecasting rainfall data in the region Malang with an observation location Jabung, Tumpang, Turen, Tumpuk Renteng, Tangkilsari, Wajak, Blambangan, Bululawang, Tajinan and Poncokusumo.

LITERATURE REVIEW

GSTAR (Generalized Space Time Autoregressive) Model

GSTAR with order autoregressive (p) and spatial order (λ_p), or represented by GSTAR (p, λ_p) can be written in the equation:

$$\boldsymbol{z_{(t)}} = \sum_{k=1}^{p} \left[\boldsymbol{\Phi_{k0} W^{(0)} z_{(t-k)}} + \sum_{s=1}^{\lambda p} \boldsymbol{\Phi_{ks} W^{(s)} z_{(t-k)}} \right] + \boldsymbol{e_{(t)}} \quad (1)$$

where :

$\boldsymbol{z_{(t)}}$: vector of location m at time t
λ_p : spatial order from form of autoregressive p
$\boldsymbol{W^{(s)}}$: weighted matrix with size of $m \times m$
$\boldsymbol{\Phi_{ks}}$: autoregresive parameter at time lag k and spatial lag s
$\boldsymbol{e_{(t)}}$: error vector with white noise and normal multivariate distribution

GSTARIMA (Generalized Space Time Integrated Autoregressive and Moving Average) Model

Min, et al. (2010), defines the model GSTARIMA as follows:

$$\Delta_{\boldsymbol{d}} \boldsymbol{z(t)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{p}} \sum_{\boldsymbol{l=0}}^{\boldsymbol{\lambda_k}} \boldsymbol{\Phi_{kl} W^{(l)}} \Delta_d z(t-k) - \sum_{\boldsymbol{k=1}}^{\boldsymbol{q}} \sum_{\boldsymbol{l=0}}^{\boldsymbol{m_k}} \boldsymbol{\Theta_{kl} W^{(l)}} \epsilon\,(t-k) + \boldsymbol{\epsilon_t} \quad (2)$$

with:

$$\boldsymbol{\Phi_{kl}} = diag\left({\phi_{kl}}^1, {\phi_{kl}}^2, \dots {\phi_{kl}}^N\right) = \begin{bmatrix} {\phi_{kl}}^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {\phi_{kl}}^N \end{bmatrix}$$

$$\boldsymbol{\Theta_{kl}} = diag\left({\theta_{kl}}^1, {\theta_{kl}}^2, \dots {\theta_{kl}}^N\right) = \begin{bmatrix} {\theta_{kl}}^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {\theta_{kl}}^N \end{bmatrix}$$

where, p and q is the order of the autoregressive and moving average, λ_k and m_k is the spatial order k to autoregressive and moving average, and $\Delta_{\boldsymbol{d}} \boldsymbol{z(t)}$ is the observation vector $\boldsymbol{z(t)}$ in order differencing to d. $\boldsymbol{\phi_{kl}}$ and $\boldsymbol{\theta_{kl}}$ are autoregressive and moving average parameters on the lag time to spatial lag k and l. $\boldsymbol{W^{(l)}}$ is the weighted matrix of size N x N and $\boldsymbol{\epsilon(t)}$ is the vector of the residual that is random and normal with size n x 1.

Seemingly Unrelated Regression (SUR)

Seemingly Unrelated Regression (SUR) is an equation parameter estimation using General Least Square (GLS). SUR Model with m equations stated,

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_m} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X_m} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_m} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_m} \end{pmatrix} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3}$$

The assumption of the model is a residual ε_{ir} is independent at all times, but between equality / contemporary correlated location. $E[\varepsilon_{ir}\varepsilon_{js}|X] = 0$ when $r \neq s$ and $E[\varepsilon_{ir}\varepsilon_{jr}|X] = \sigma_{ij}$. Variance-covariance matrix is denoted by Ω.

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11}\mathbf{I}_R & \sigma_{12}\mathbf{I}_R & \cdots & \sigma_{1m}\mathbf{I}_R \\ \sigma_{21}\mathbf{I}_R & \sigma_{22}\mathbf{I}_R & \cdots & \sigma_{2m}\mathbf{I}_R \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\mathbf{I}_R & \sigma_{m2}\mathbf{I}_R & \cdots & \sigma_{mm}\mathbf{I}_R \end{bmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I}_R \tag{4}$$

SUR estimation by adding variance covariance matrix residual is stated as follows,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_{\boldsymbol{R}})\boldsymbol{X})^{-1}\boldsymbol{X}'(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_{\boldsymbol{R}})\mathbf{Y} \tag{5}$$

With

$$\begin{aligned} r(\widehat{\boldsymbol{\beta}}) &= E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2\right] \\ &= E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= E[\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\}'] \\ &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'^{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}. \end{aligned}$$

assuming $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma})$. SUR estimators use the information system more efficient than OLS because of the diversity in each equation is smaller, [3].

The precision model with MSE and R^2

Criteria for the good of the model can be determined based on the residual is the Mean Square Error (MSE).

$$Mean\ Square\ Error\ (\text{MSE}) = \frac{1}{\text{N}}\sum_{\text{t}=1}^{\text{N}} \text{e}_\text{t}{}^2 \tag{6}$$

where $\text{e}_\text{t} = \text{Z}_{\text{n}+1} - \hat{Z}_n(l)$ and N is account of data.

The coefficient of determination stating how great the diversity of the dependent variable (Y) can be explained by the independent variable (X). According Makridakis, et al. (1999), R^2 is obtained from,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(Z_{i(t)} - \hat{Z}_{i(t)}\right)^2\right)}{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(Z_{i(t)} - \bar{Z}_{i(t)}\right)^2\right)} \tag{7}$$

where :

$Z_{i(t)}$: dependen variable i

$\bar{Z}_{i(t)}$: mean of $Z_{i(t)}$

$\hat{Z}_{i(t)}$: predicted value of $Z_{i(t)}$

METHODOLOGY

Location of the study is the rainfall observatory stations in the in malang southern region districts. The stations are Tumpang, Wajak, Tajinan, Jabung, Poncokusumo, Turen, Tumpuk Renteng, Tangkilsari, Blambangan, and Bululawang. Data used is 10 days period of rainfall (*dasarian*) since the beginning of the dry season is determined based on the amount of rainfall in a single dasarian (10 days) of less than 50 mm and is followed by several subsequent dasarian. Meanwhile, the beginning of the rainy season is determined based on the amount of rainfall in a single dasarian (10 days) is equal or more than 50 mm and is followed by several subsequent dasarian [4].

Rainfall data used to build GSTARIMA-SUR model is a sample data for forecasting (*insample*). The *insample* data is dasarian rainfall data for the period may 2000 to April 2015. The other data is called *outsample* data that is used to validate the GSTARIMA-SUR models of the data in the period May-June 2015.

Steps taken to form GSTARIMA-SUR model is started by exploration of rainfall data in the ten rainfall observatory stations, identification of univariate model (ACF and PACF) and multivariate model (MPACF and MCCF) to determine the order GSTARIMA, determination parameters of the model, model validation, and the last is forecasting rainfall data in the malang southern region districts and the surrounding region.

RESULTS AND DISCUSSION

Exploration data is used to easy for viewing the information on the data. exploration data can be presented in the form of graphs and descriptive data. Graph data movement precipitation 10 locations this year May 2000 to April 2015 is shown in the form of a time series plot as in Figure 4.1 below.

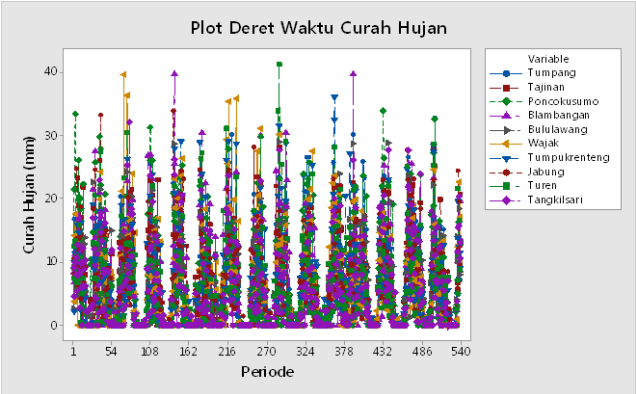

Figure 4.1. Time series Plot of Rainfall data

Based on Figure 4.1, known that movement pattern of rainfall data in 10 locations tend to be similar. High and low altitude change of rainfall data recorded in each period indicates the same pattern every year.

Identification of the model is used to find the order of autoregressive and moving average for GSTARIMA model. The order of autoregressive lag is obtained from the identification of the real MPACF and the order of moving average lag is obtained from the identification of the real MACF, then from some real lag is chosen the best use of AIC. Lag which has the smallest AIC value will be used as the order of autoregressive and moving average for GSTARIMA model. Moreover to identification seasonal pattern use univariate identification by seeing ACF and PACF plot. Results of identification rainfall in malang southern region districts is GSTARIMA((1)(1,12,36),(0),(1))-SUR.

Estimation of parameters is done by inserting a weighted into the equation to describe the spatial relationship between the location of the post of rain. The weight of the locations used in this study is the inverse distance weighting location.

Table 4.1 Inverse Distance Weighted Value 10 Pos Rain

Pos Hujan	turen	tumpuk renteng	wajak	tumpang	jabung	poncokusumo	bululawang	tajinan	tangkilsari	blambangan
turen	0	0.286	0.140	0.057	0.044	0.068	0.103	0.087	0.092	0.122
tumpuk renteng	0.226	0	0.179	0.055	0.041	0.067	0.111	0.098	0.101	0.121
wajak	0.123	0.200	0	0.082	0.055	0.118	0.097	0.130	0.103	0.090
tumpang	0.059	0.073	0.097	0	0.191	0.223	0.075	0.128	0.090	0.064
jabung	0.063	0.074	0.089	0.260	0	0.141	0.083	0.121	0.097	0.072
poncokusumo	0.072	0.090	0.141	0.227	0.105	0	0.078	0.129	0.090	0.069
bululawang	0.068	0.093	0.073	0.048	0.039	0.049	0	0.115	0.280	0.235
tajinan	0.069	0.099	0.117	0.098	0.068	0.097	0.138	0	0.218	0.097
tangkilsari	0.062	0.087	0.080	0.059	0.046	0.058	0.288	0.187	0	0.133
blambangan	0.099	0.125	0.084	0.050	0.041	0.053	0.290	0.100	0.159	0

GSTARIMA-SUR models for rainfall in the Malang southern region districts can be expressed as follows:

$$\boldsymbol{Z(t) = \Phi_{01}^{(1)} Z(t-1) + \Phi_{11}^{(1)} WZ(t-1) + \Phi_{01}^{(2)} Z(t-12) + \Phi_{11}^{(2)} WZ(t-12) + \Phi_{01}^{(3)} Z(t-36) + \Phi_{11}^{(3)} WZ(t-36) + e(t) - \theta_{01}^{(1)} e(t-1) + \theta_{11}^{(1)} We(t-1)}$$

Based on the results of parameter estimation GSTARIMA ((1)(1,12,36),(0),(1))-SUR to Tumpang rain post is as follows:

$$\begin{aligned}\hat{z}_{1(t)} = {} & 0.236z_{1(t-1)} + 0.102z_{2(t-1)} + 0.038z_{3(t-1)} + 0.024z_{4(t-1)} + 0.015z_{5(t-1)} + 0.007z_{6(t-1)} \\ & + 0.05z_{7(t-1)} + 0.026z_{8(t-1)} + 0.037z_{9(t-1)} + 0.041z_{10(t-1)} + 0.026z_{1(t-12)} \\ & + 0.009z_{2(t-12)} - 0.017z_{3(t-12)} + 0.003z_{4(t-12)} + 0.005z_{5(t-12)} - 0.004z_{6(t-12)} \\ & + 0.005z_{7(t-12)} - 0.002z_{8(t-12)} - 0.004z_{9(t-12)} + 0.001z_{10(t-12)} + 0.147z_{1(t-36)} \\ & + 0.121z_{2(t-36)} + 0.023z_{3(t-36)} + 0.007z_{4(t-36)} + 0.006z_{5(t-36)} + 0.02z_{6(t-36)} \\ & + 0.029z_{7(t-36)} + 0.014z_{8(t-36)} + 0.022z_{9(t-36)} + 0.019z_{10(t-36)} + e_{1(t)} \\ & + 0.081e_{1(t-1)} + 0.009e_{2(t-1)} + 0.019e_{3(t-1)} + 0.002e_{4(t-1)} + 0.004e_{5(t-1)} \\ & - 0.005e_{6(t-1)} + 0.011e_{7(t-1)} + 0.012e_{8(t-1)} + 0.003e_{9(t-1)} + 0.015e_{10(t-1)}\end{aligned}$$

Rainfall in the Tumpang Rain Post is influenced by the rain heading more in ten days and twenty days earlier and was influenced by weighting the location. In addition there is seasonality in the rainfall in the period a quarter of the year and annually.

Based on the above model prediction results can be obtained in the data sample for the Tumpang rain post as follows:

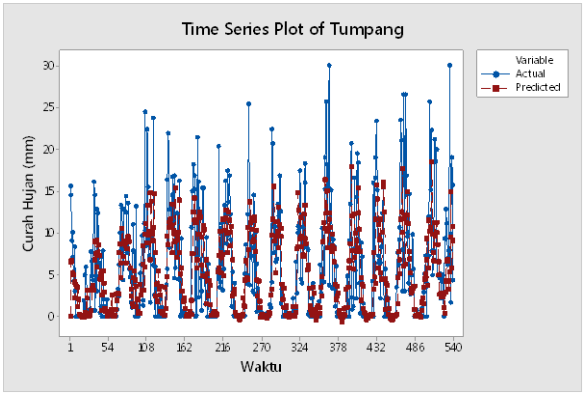

Figure 4.2 Forecasting using GSTARIMA((1)(1,12,36),(0),(1))-SUR in Tumpang Rain Post

Inspection accuracy of the model or the model validation is done by looking at the value of RMSE and R^2 of the model.

Table 4.2 Accuracy Test Model of GSTARIMA ((1),(1,12,36)(0)(1))-SUR Model in ten Rain Post

Location	RMSE	R^2 *Prediction*
Tumpang	5.668	0.5052
Tajinan	5.51	0.5716
Poncokusumo	5.916	0.5146
Blambangan	5.697	0.5935
Bululawang	5.175	0.5805
Wajak	5.899	0.5136
Tumpuk Renteng	6.007	0.6476
Jabung	5.337	0.5671
Turen	6.119	0.6308
Tangkilsari	5.263	0.6481

R^2 *prediction* for 10 observation locations is more than 50%. The greater the value of R^2 prediction obtained the greater, model can explain the rainfall distribution. The biggest prediction of R^2 is Tangkilsari, that is equal to 0.6481. It can meaned that about 64.81% in Tangkilsari rainfall distribution can be explained by the model GSTARIMA ((1), (1,12,36) (0) (1))-SUR.

The lowest R^2 prediction is in Tumpang Area rainfall, but the prediction obtained R^2 can still be said to be good for rainfall prediction for prediction obtained R^2 values> 50%. If viewed from the RMSE and R^2 prediction can be said that the rainfall distribution in ten observation locations have the same result by using modeling GSTARIMA ((1) (1,12,36) (0) (1))-SUR.

Result forecast rainfall using GSTARIMA ((1),(1,12,36)(0)(1))-SUR model is:

Table 4.3 Forecasting Rainfall Period of May-June 2015

Month	May			June		
Locations	1	2	3	1	2	3
Tumpang	7.157	8.254	6.642	5.503	3.241	1.438
Tajinan	3.318	15.397	10.754	1.504	1.842	1.495
Poncokusumo	5.702	3.069	6.390	6.610	3.309	1.179
Blambangan	1.180	1.515	1.100	1.382	1.259	0.397
Bululawang	9.392	2.592	6.521	4.620	3.943	0.490
Wajak	1.509	1.143	1.343	1.401	1.468	0.031
Tumpuk Renteng	5.360	3.440	3.025	4.704	2.654	1.073
Jabung	5.981	1.204	1.291	0.851	1.090	0.962
Turen	8.348	7.754	3.159	5.744	3.885	1.703
Tangkilsari	7.406	1.766	2.010	1.669	1.947	1.160

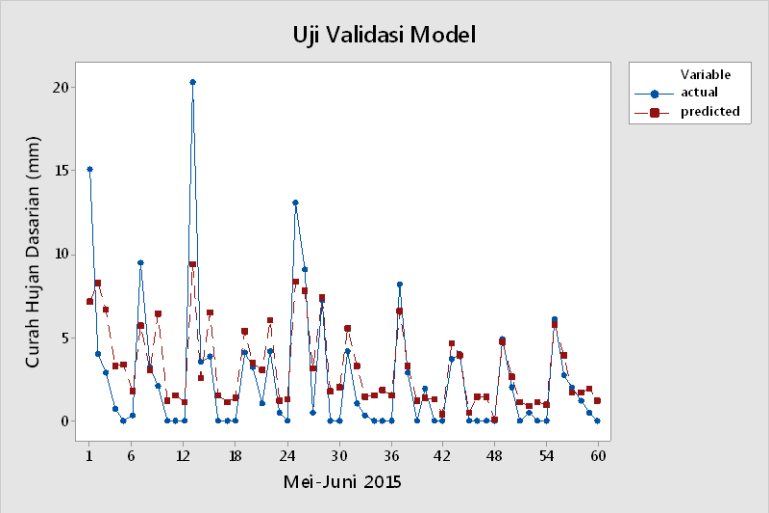

Figure 4.3 Forecast for Period May-June 2015 in Malang Southern Region Districts

Model validation is done by comparing the actual data of rainfall dasarian the period from May to June 2015 on the results of data using models forecasting GSTARIMA ((1),(1,12,36)(0)(1))-SUR. If seen from Figure 4.3 can be said that result of forecasting data can approach the actual data. To better know the data equation then done two paired samples t test. Based on the results of two sample paired t test obtained by value t amounted to 1,958 with significant value 0095. t table with db = 59 values obtained 2,301, for t <t table (1.9585 <2,301) and the p-value is more than 0.05 (0095> 0.05) it was decided to accept H0. The conclusion of this validation test is that the rainfall dasarian data forecasting results do not differ significantly from the actual data, so the model GSTARIMA-SUR formed from in sample data can be used for forecasting rainfall dasarian next period.

CONCLUSION

Our analysis found that the GSTARIMA ((1)(1,12,36)-SUR model can be used to forecast the dasarian seasonal rainfall in the Malang Southern Region Districts. By using seasonal patterns in lag 1,12,36 more accurate forecasting result is obtained. To validate the model, MSE and R^2 prediction can be used. In this research, the largest R^2 prediction is 57.726%.

REFERENCE

[1] J. Rosadi, Pengantar Analisis Runtun Waktu, Diktat Kuliah, Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gajah Mada, Yogyakarta, Universitas Gajah Mada, 2006.

[2] S. Borovkova, H. Lopuhaa and B. Ruchjana, "Generalized STAR model with experimental weights," in *Proceedings of the 17th International Workshop on Statistical Modelling*, 2002.

[3] H. R. Moon and B. Perron, "Seemingly unrelated regressions," *The New Palgrave Dictionary of Economics,* pp. 1-9, 2006.

[4] BMKG, Modul Diklat Badan Meteorologi Klimatologi dan Geofisika Karangploso Malang, BMKG, 2000.

[5] A. Zellner, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American statistical Association,* vol. 57, no. 298, pp. 348-368, 1962.

[6] P. E. Pfeifer and S. Jay Deutsch, "Stationarity and invertibility regions for low order starma models: stationarity and invertibility regions," *Communications in Statistics-Simulation and Computation,* vol. 9, no. 5, pp. 551-562, 1980.

[7] X. Min, J. Hu and Z. Zhang, "Urban traffic network modeling and short-term traffic flow forecasting based on GSTARIMA model," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, 2010.

[8] A. Iriany, Suharningsih, B. N. Ruchjana and Setiawan, "Prediction of Precipitation data at Batu Town using the GSTAR (1,p)-SUR Model," *Journal of Basic and Application Scientific Research,* vol. 3, no. 6, pp. 860-865, 2013.