Advances in Social Science, Education and Humanities Research, volume 550
Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMED 2020)

Modeling Cabbage Production in Malang East Java with GSTAR Approach

Muhammad Syahfitra[1], Ni Wayan S. Wardhani[2], Atiek Iriany[3]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Since, Brawijaya University, Malang, Indonesia*

Email: Syahfitra144@gmail.com, wswardhani@ub.ac.id, Atiek_iriany@ub.ac.id

ABSTRACT
Based on the Directorate Report General's Horticulture, the contribution of vegetable horticulture agriculture tends to increase, where the GDP of vegetable horticulture has increased by 9.86%. In 2016, cabbage is a vegetable horticultural commodity that has the highest production amount in Indonesia, and the poor district is one of the major producers of commodities cabbage in eastern Java. Generalized Space-Time Autoregressive (GSTAR) is a multivariate time series model that considers site aspects with heterogeneous location characteristics. The purpose of this study was to model cabbage production in Malang Regency using the GSTAR model. Selection criteria for the best model to use the value of *the root mean square error* (RMSE) and the value of $R^{2.}$ The results showed that the GSTAR model (1,2) is the best model for modeling cabbage production and has good forecasting accuracy to predict cabbage production in Malang Regency.

Keywords: *Generalized Space-Time Autoregressive* (GSTAR), Cabbage, *root mean square error* (RMSE), R^2.

1. INTRODUCTION

Time series analysis is an analysis that considers the effect of time used to predict future observations. In general, there are two time-series data modeling, namely univariate and multivariate [1]. The univariate time series model that is often used is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA model only involves one time series variable. While the multivariate time series model is a model that involves more than one time series variable, for example, the Vector Autoregressive Integrated Moving Average (VARIMA) model.

The ARIMA model does not contain a location element in its forecasting model, so another model that contains time is needed, and the location of the space-time model is a model that combines elements of time and location linkages in time and location data series. The space-time model is a model that combines elements of time and location linkages in time and location data series. The space-time model developed by [2]. Pfeifer and Deutsch adopted the steps developed by [3] for ARIMA modeling, which includes identification, estimation, and diagnostic checks into STARIMA (Space-Time Autoregressive Integrated Moving Average) modeling.

The model *space-time* developed by [4] has a weakness in the flexibility of parameters that explain the relationship between locations and different times in a time series data and location. This weakness is corrected by [5] through a model is known as the GSTAR (Generalized Space-Time Autoregressive) model.

This research was conducted with the aim of applying the GSTAR model to cabbage production data in the Malang Regency. In this study, it is hoped that an appropriate model can be obtained, so that it can be used to obtain accurate forecast values, and can explain the relationship between cabbage production in one location and cabbage production in other locations.

2. REVIEW OF LITERATURE

2.1. Stationarity Test Stationarity

Testing in this study used the unit root test (Dickey-Fuller). The hypothesis used in testing whether there is a unit root problem is:

$H_0: \phi_1 = 1$ (non-stationary data)

$H_1: \phi_1 < 1$ (stationary data)

The test statistic used is the t-test statistic. However, under H_0, the t-test statistic does not have at distribution, but instead distributes τ:

$$\tau = \frac{\hat{\phi}_1^*}{SE\,(\hat{\phi}_1^*)} \sim \tau_n \quad (1)$$

where,

$\hat{\phi}_1^*$: the estimated value of the parameter *autoregressive* (AR)

$$SE\,(\hat{\phi}_1^*) = \frac{\sigma_{\hat{\phi}_1^*}}{\sqrt{n}} \quad is\ the\ standard\ error\ \hat{\phi}_1^* \quad (2)$$

2.2. Space-Time

Model The space-time model (*space-time*) is a model that can combine elements of time and location dependence on data *multivariate time series*. The concept of multivariate time series data is that there is more than one time series variable, in this case, there are several research locations that are used as research variables. This model is a modeling of a number of observations $Z_i(t)$ contained at each N location in space $(i = 1,2,\dots,N)$ against t time period.

2.3. Inverse matric distance

weighting serves to describe the spatial relationship between regions. The location weight used in this study is the inverse location weight of the distance. The weight of the inverse distance location is obtained from calculations based on the actual distance between locations. This weight gives a smaller coefficient of weight for long distances, and vice versa. The calculation of this weight is:

$$W_{ij} = \frac{1/d_{ij}}{\sum_{i \neq j} 1/d_{ij}} \quad (3)$$

where d_{ij} is the distance between location *i* and location *j*.

2.4. GSTAR

GSTAR or Generalized pace-Time Autoregressive model is a space-time model that aims to increase the flexibility of STAR parameters. The GSTAR model is a more specific form of the VAR (Vector Autoregressive) model. The most basic difference is in spatial dependent and matrix weights. [6] stated that the GSTAR model is more realistic because there are more models with different model parameters for different locations. This model was introduced [5].

Determination of the order of the GSTAR model is the same as the VAR model, namely using the MPACF (Matrix Partial Autocorrelation Function) [1]. The partial autocorrelation matrix on lag *k is* denoted by $P\,(k)\ is$ defined as follows

$$P(k) = [D_v(k)]^{-1} V_{vu}\,(k) [D_u(k)]^{-1} \quad (4)$$

where,

$D_v(k)$: The diagonal matrix of size m x m with the to-i diagonal element is the root of the to-i diagonal element of the var ($v_{k-1,}$t).

$D_u(k)$: diagonal matrix of size m x m with the i-th diagonal element is the root of the i-th diagonal element of the var matrix $\left(u_{k-1,t+1}\right)$

$V_{vu}(k)$: matrix the covariance matrix of $v_{s-1,t}$ and $u_{s-1,t+1}$ which measures m x m.

$$v_{k-1,t} = Z_t - \beta_{k-1,1} Z_{t+1} - \dots - \beta_{k-1,k-1} Z_{t+k-1}$$
$$u_{k-1,t+k} = Z_{t+k} - \alpha_{k-1.1} Z_{t+k+1} - \dots - \alpha_{k-1,k-1} Z_{t+1}$$

The autoregressive order is determined from a significant matrix element to the lag to-p [1]. The GSTAR model with autoregressive order p and spatial order (λ), or denoted by GSTAR $(p,_\lambda)$ can be written in the following equation:

$$\boldsymbol{Z}_{(t)} = \sum_{k=1}^{p} \left[\Phi_{k0} Z_{(t-k)} + \sum_{h=1}^{\lambda} \Phi_{kh} W^{(g)} Z_{(t-k)} \right] + \varepsilon_{(t)} \quad (5)$$

with

λ : spatial order on the parameter *autoregressive*

$Z_{(t)}$: random vector of size $(m \times 1)$ at time *t*.

Φ_{k0} : diag $(\Phi_{k0}^1,\dots,\Phi_{k0}^m)$ that is, the diagonal matrix of autoregressive parameters at time lag *k* and spatial lag 0

Φ_{kg} : diag $(\Phi_{k0}^1, \dots,\Phi_{k0}^m)$ that is, the diagonal matrix of autoregressive parameters at lag time -*k* and spatial lag *g*.

$W^{(g)}$: size weighted matrix $(m \times m)$ on spatial lag *g*, where $w_{ii}^{(g)} = 0$ and $\sum_{i \neq j} w_{ij}^{(g)} = 1$.

$\varepsilon_{(t)}$: the size error vector (m × 1) is white noise and normally multivariate distribution.

m : the number of locations used.

In the GSTAR model, the model parameter is a matrix with diagonal elements which states the autoregressive parameter and the changing space-time parameter for each location. GSTAR has limitations, namely that it can only be used for stationary and non-seasonal space-time data. This condition tends not to be fulfilled on data that is not stationary and contains seasonal patterns.

2.5. *Estimation of Parameters*

[7] used the Ordinary Least Square (OLS) method to estimate the parameters of the GSTAR model. The OLS method is a method used to estimate the parameters of a model by minimizing the number of squares of errors, namely minimizing $\sum_{i=1}^{T} e_t^2$. The estimation method of least squares to $\hat{\theta}$ be

$$\hat{\theta} = (X'X)^{-1}X'Z \quad (6)$$

For each location $i = 1,2,\ldots,m$ have partial linear model such as equation (5). This means that the estimation of the least-squares θ_i for each location can be calculated separately.

2.6. Goodness Model Criteria

Criteria for selecting the best model to use the value of the root mean square error (RMSE) and the value of R^2.RMSE has a function to obtain the amount of difference that appears between the actual value and the predicted value. RMSE value is obtained from the following formula

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Z_t - \hat{Z}_t)^2} \quad (7)$$

R^2 stating how much the diversity of the dependent variable can be explained by the independent variable.value R^2 obtained from the following formula

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Z_t - \hat{Z}_t)^2}{\sum_{t=1}^{n}(Z_t - \underline{Z})^2} \quad (8)$$

3. RESEARCH METHODS

This study uses secondary data obtained from the Department of Agriculture and Horticulture Malang. This study used variables, namely cabbage production (quintal). This study models cabbage production in Malang Regency with the GSTAR-OLS model using the inverse matric distance. In addition, the simulation also applies location weighting, namely the normalization of statistical inference results in partial cross-correlation at the appropriate time lag to determine the optimal location weight in the GSTAR model.

4. RESULTS AND DISCUSSION

4.1. Description of Cabbage Production Data

Data used in this study were cabbage production in three locations in Malang Regency during the period 2013 - 2017. In Table 1, the descriptive statistics of the data are presented below.

Table 1. Descriptive Statistics Production Data Cabbage in Three Locations Malang

No	Location	N	Mean	Min	Max	Standard Deviation
1	Poncokusumo	60	20533.33	2500	42000	9510.417
2	Wajak	60	4341	630	22050	3092.141
3	Tumpang	60	3359.7	150	25 200	3836.008

Plot production data cabbage in Malang Regency period Month January 2013 until December 2017 is presented in Figure 1 below:

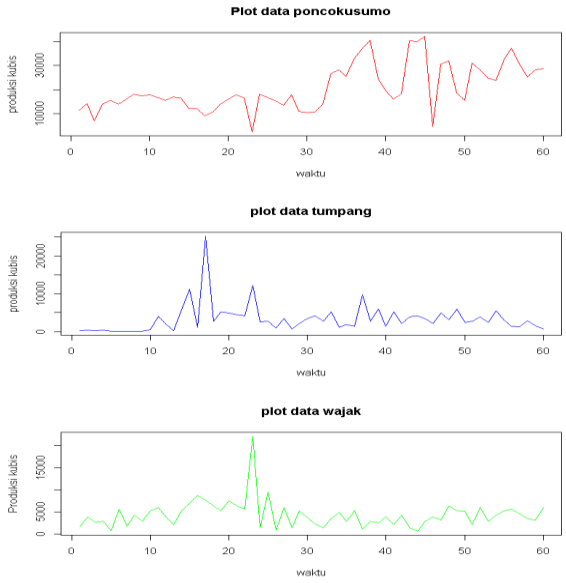

Figure 1. Data plot of Cabbage Production

Figure 1 shows that the data is not stationary, so the data is processed *differencing*. Data already done process *differencing* is denoted as $Z_i(t) = ln\ ln\ Y_i(t)$ and it is said to be stationary if the data structure from time to time has fixed or constant data fluctuations and does not change.

4.2. Stationarity Test

After the data was transformed, it was tested using Augmented Dickey-Fuller (ADF). The test results are presented in Table 2 below:

Table 2. Stationarity Test Results of Cabbage Production Data in Three Locations of Malang Regency

No	Location	P-value	Information
1	Poncokusumo	0.095	Accept
2	Wajak	0.020	Reject
3	Tumpang	0.078	Accept

Because the *p-value* in Poncokusumo and Tumpang Districts in Table 2 is greater than $\alpha = 0.05$ then is accepted, in other words, the cabbage

production data in Poncokusumo and Tumpang Districts are not stationary yet, so it needs to be stationary.

4.3. *Location Weight GSTAR (1,2)*

The results of the weighted matrix calculation with the inverse distance weight are as follows:

$$W_{ij} = \begin{bmatrix} 0.000000 & 0.457335 & 0.542665 & 0.520193 \\ 0.000000 & 0.479807 & 0.562642 & 0.437358 & 0.000000 \end{bmatrix}$$

4.4. *Estimation Results of GSTAR Model*

Parameters (1,2) Three Locations

Determine the parameters of the GSTAR model (1,2) using Equation (5), where the parameter values are calculated using SAS software. The results of parameter estimation for cabbage production data in the three research locations obtained the parameter estimation values for the period January 2013 to December 2017 are as follows:

Table 3. Estimation Results of Cabbage Production Data Parameters in Three Locations of Malang Regency

No	Location	Variabel	DF	Parameter Estimate	Standard Error	t-Value	Pr > |t|
1	PONCOKUSUMO	Z1t_1	1	0.7733	0.1436	5.39	<.0001
		Z1t_2	1	0.0386	0.1503	0.26	0.7981
		V1t_1	1	1.6479	0.9818	1.68	0.0990
		V1t_2	1	-0.1076	0.9414	-0.11	0.9094
2	WAJAK	Z2t_1	1	0.1881	0.1060	1.77	0.0818
		Z2t_2	1	0.6082	0.0998	6.10	<.0001
		V2t_1	1	0.0471	0.0319	1.47	0.1462
		V2t_2	1	-0.0029	0.0339	-0.09	0.9299
3	TUMPANG	Z3t_1	1	0.2745	0.1239	2.21	0.0310
		Z3t_2	1	0.4646	0.1211	3.84	0.0003
		V3t_1	1	0.0429	0.0468	0.92	0.3630
		V3t_2	1	0.0091	0.0479	0.19	0.8505

Based on Table 3 can be seen that for the District Poncokusumo variables that significantly affect the production of cabbage is the result of cabbage production in the previous period with a p-value of 0.0001. for Wajak Regency that had a significant effect on cabbage production at time *to- t* was the cabbage production yield in the previous two periods (*t-2*). For Tumpang District, what affects cabbage production at time *to- t* is the yield of cabbage in the previous period *(t-2)* and the two previous periods *(t-2)*.

Based on Table 3, the GSTAR equation of cabbage production data from three locations is:

$$\begin{bmatrix} \hat{Z}_1(t) \\ \hat{Z}_2(t) \\ \hat{Z}_3(t) \end{bmatrix} = \begin{bmatrix} 0.7733 & 0.0000 & 0.0000 \\ 0.0000 & 0.1881 & 0.0000 \\ 0.0000 & 0.0000 & 0.2745 \end{bmatrix} \begin{bmatrix} Z_1(t-1) \\ Z_2(t-1) \\ Z_3(t-1) \end{bmatrix} + \begin{bmatrix} 0.0386 & 0.0000 & 0.0000 \\ 0.0000 & 0.6082 & 0.0000 \\ 0.0000 & 0.0000 & 0.4646 \end{bmatrix} \begin{bmatrix} Z_1(t-2) \\ Z_2(t-2) \\ Z_3(t-2) \end{bmatrix} + \begin{bmatrix} 1.6479 & 0.0000 & 0.0000 \\ 0.0000 & 0.0471 & 0.0000 \\ 0.0000 & 0.0000 & 0.0429 \end{bmatrix} \begin{bmatrix} 0.000000 & 0.457335 & 0.542665 \\ 0.520193 & 0.000000 & 0.479807 \\ 0.562642 & 0.437358 & 0.000000 \end{bmatrix} \begin{bmatrix} V_1(t-1) \\ V_2(t-1) \\ V_3(t-1) \end{bmatrix} + \begin{bmatrix} -0.1076 & 0.0000 & 0.0000 \\ 0.0000 & -0.0029 & 0.0000 \\ 0.0000 & 0.0000 & 0.0091 \end{bmatrix} \begin{bmatrix} 0.000000 & 0.457335 & 0.542665 \\ 0.520193 & 0.000000 & 0.479807 \\ 0.562642 & 0.437358 & 0.000000 \end{bmatrix} \begin{bmatrix} V_1(t-2) \\ V_2(t-2) \\ V_3(t-2) \end{bmatrix}$$

$$\begin{bmatrix} \hat{Z}_1(t) \\ \hat{Z}_2(t) \\ \hat{Z}_3(t) \end{bmatrix} = \begin{bmatrix} 0.7733 Z_1(t-1) \\ 0.1881 Z_2(t-1) \\ 0.2745 Z_3(t-1) \end{bmatrix} + \begin{bmatrix} 0.7536\, V_2(t-1) + 0.8943\, V_3(t-1) \\ 0.0245\, V_1(t-1) + 0.0226\, V_3(t-1) \\ 0.0241\, V_1(t-1) + 0.0188\, V_2(t-1) \end{bmatrix} + \begin{bmatrix} -0.0492\, V_2(t-2) - 0.0584\, V_3(t-2) \\ -0.0015\, V_1(t-2) - 0.0014\, V_3(t-2) \\ 0.0051\, V_1(t-2) + 0.0040\, V_2(t-2) \end{bmatrix}$$

Based on Equation, the GSTAR (1,2) model can be written for each location as follows:

- Poncokusumo

$$\hat{Z}_1(t) = 0.7733 Z_1(t-1) + 0.7536\, V_2(t-1) + 0.8943\, V_3(t-1) - 0.0492\, V_2(t-2) - 0.0584\, V_3(t-2)$$

- Wajak

$$\hat{Z}_2(t) = 0.1881 Z_2(t-1) + 0.0245\, V_1(t-1) + 0.0188\, V_2(t-1) - 0.0015\, V_1(t-2) - 0.0014\, V_3(t-2)$$

- Tumpang

$$\hat{Z}_3(t) = 0.2745Z_3(t-1) + 0.0241\,V_1(t-1) + 0.0188\,V_2(t-1) + 0.0051\,V_1(t-2) + 0.0040\,V_2(t-2)$$

From this equation, it can be seen that the cabbage production data at time t correlates with the cabbage production data at the previous time and is influenced by the cabbage production at other places. In other words, Poncokusumo, Wajak and Tumpang cabbage production data influence each other.

4.5. Goodness model

Goodness GSTA Rused in this study using RMSE and R^2 are appropriate [6] and [7] whose results are presented in Table 4.

Table 4. Criteria Goodness Model

No	Location	RMSE	R^2
1	Poncokusumo	2.5261	0.9035
2	Wajak	0.3675	0.6682
3	Tumpang	1.0182	0.3408

The accuracy of the measurement error estimation method is indicated by the presence of a small RMSE. Models that have a smaller RMSE are said to be more accurate than models that have a larger RMSE. Based on the value of RMSE and R^2above it can be concluded that the level of accuracy of the model for Wajak It can be seen from the smallest RMSE value of 0.3675 and R^2 most large, namely 90.35%.

5. CONCLUSION

The GSTAR model (1,2) is the best model that can be used to predict cabbage production data in three locations, namely Poncokusumo, Wajak, and Tumpang, from the evaluation results of the GSTAR model (1,2) it shows that the R^2 value for each district is very good, especially for Poncokusumo District of 90.35%. while for the Tumpang District the resulting model is not good, this can be caused by the influence of other factors outside of the research variables so that it affects the accuracy of the resulting model.

REFERENCES

[1] Wei, WWS (2006). *Time Series Analysis: Univariate and Multivariate* (2nd ed.). USA: Pearson Education Inc.

[2] Pfeifer, PE and Deutsch, SJ, A Three Stage Iterative Procedure for Space-Time Modeling, Technometrics, Vol. 22, No. 1, p. 35-47, 1980a.

[3] Pfeifer, PE and Deutsch, SJ, Identification and Interpretation of First Order Space-Time ARMA Models. Technometrics, Vol. 22, No. 1, p. 397-408, 1980b

[4] Box, GEP and Jenkins, GM, Time Series Analysis: Forecasting and Control, Second Edition, pp. 135-141, San Francisco: Holden-Day, San Francisco, 1976

[5] Borovkova, SA , Lopuhaa, HP, and Nurani, B., Generalized STAR model with experimental weights, In M Stasinopoulos & G Tou-loumi (Eds.), Proceedings of the 17th International Workshop on Statistics Modeling, Chania, p. 139-147, 2002

[6] Wutsqa, DU, & Suhartono. (2010). Seasonal Multivariate Time Series Forecasting on Tourism Data with the VAR-GSTAR Model. Journal of Basic Science Vol. 11, 101-109.

[7] Ruchjana, B. (2002). The Generalized Space Time Autoregressive Order One Model and Its Application to Oil Production Data. [Dissertation, unpublished]. Bandung: Department of Mathematics, Bandung Institute of Technology.