Rice Harvest Area Modeling With GSTARIMA on Six Provinces in Indonesia

Januar Harun Paulus Messakh[1], Muhammad Nur Aidi[2], Farit Mochamad Afendi[3]

Abstract— In Indonesia rice has become most consumed food ingredient of people almost in all area with consumption per capita per year attain about 85 kilograms. One of the problem that should be encountered due to the increase of population is the fullfillment of domestic food needs. Increasingly population certainly needs to be followed up with strategies and policies about fullfillment foods needs. Due to policies and appropriate decision making availability of accurate data is required. Some necessary indicator required related to food data is rice harvest area which related and affect little or big rice crop production number. Besides accurate data appropriate methods is required to predict about future, one of them is forecasting. GSTARIMA is space time forecasting model which is combined time series analysis with association between locations that represented by spatial weight matrix. There are six provinces in Indonesia which has biggest rice harvest area than others those are West Java, Central Java, East Java, South Sulawesi, South Sumatera and North Sumatera. GSTARIMA model to forecast rice harvest area on six provinces in Indonesia had built is GSTARIMA (2,0,0)*(0,1,1)12 with spatial weight matrices are distance inverse weight and cross corellation weight. The result of compared both matrices to forecast 12 periods (months) ahead could be viewed from MAPE Total values, model with cross correlation weight gives better accuracy than model with distance inverse weight.

Index Terms— Space-Time Model, GSTARIMA, Forecast, Spatial weight Matrix, Rice Harvest Area.

1 INTRODUCTION

According to the FAO report in 2015, Indonesia was third major producer of rice in the world behind China and India with the production of about 70.8 million tonnes per year. In Indonesia, rice had becomes most consumed food ingredient of peoples in almost in all area with consumption per capita per year attain about 85 kilograms.

Statistics of Indonesia reported population projection in 2013 showed that total population of over twenty-five years from 2010 rised from 238.5 million to 305.6 million in 2035. One of the problem that should be encountered due to the increase of population is the fullfillment of domestic food needs. Increasingly population certainly needs to be followed up with strategies and policies about fullfillment foods needs. Due to policies and appropriate decision making availability of accurate data is required. Some necessary indicator required related to food data is rice harvest area which related and affect little or big rice crop production number.

To draw up a strategy on food policy and effort to prevent the problems about food needs that may arise, accurate data and some appropriate methods is required to address such matters. One of them is forecasting that is able to predict event will be occured in the future. Most commonly forecast method is ARIMA known as Box-Jenkins Method. In 1979, Pfeifer and Deutsch[2] proposed space-time data analysis which is developmet of time series analysis with considered associations between locations called STARIMA. Ruchjana[1] developed a model GSTAR which is an extension of STARIMA. Difference between both is in parameters in GSTAR is assumed to be different or heterogeneous at each location. GSTARIMA is a common form of GSTAR which containing Moving Average. The association between locations in model represented by spatial weighted matrix.

Planting or harvesting frequency pattern in a region depend on several factors some of them are climate and availability of water so that among locations there may be has similiar or difference patterns each others. Hence, forecasting with attention to association of locations as known as space time analysis which is GSTARIMA can be applied. There are six region in Indonesia which had biggest rice harvest area among others those are West Java, Central Java, East Java, South Sulawesi, South Sumatera and North Sumatera with vast percentage of harvest in 2015 attained 64.65 percent of total area of the national harvest.

Association among locations in GSTARIMA represented by spatial weight matrix, in this paper spatial weight matrices should be used are distance inverse weight which describe the influence of the distance toward forecasting result and cross correlation weight which does not emphasize the distance among locations but on cross correlations of responsed variable from each locations. This paper focused on comparing both spatial weight matrices in harvest area modeling with GSTARIMA to obtain best forecast.

2 BACKGROUND

2.1 Spatial Weight Matrix

Spatial weight matrix is tool to represent relationships of location ussualy denoted by W, inverse disctance calculated from latitude and longitude center point coordinate distances of observed locations. The formula for distances is similiar to Euclid distance, element of distance inverse weight matrix is obtained as follows:

1) Author name is currently pursuing masters degree program in applied statistics in Bogor Agriculture University, Indonesia. E-mail: gaussjordan89@gmail.com
2) Co-Author name is currently lecturer in statistics departement in Bogor Agriculture University, Indonesia. E-mail: nuraidi@yahoo.com
3) Co-Author name is currently lecturer in statistics departement in Bogor Agriculture University, Indonesia. E-mail: fmafendi@gmail.com

$$w_{ij} = \begin{cases} \dfrac{c_{ij}}{\sum_j c_{ij}} & \text{for i} \neq \text{j} \\ 0 & \text{for i = j} \end{cases} \quad (1)$$

Cross correlation weight was proposed by suhartono and Atok in Suhartono and Subanar[3], this weight use general cross correlation between two location i and j at time lag k, estimated by:

$$r_{ij}(k) = \frac{\sum_{t=1}^{n-k}\left(Z_{i,t} - \bar{Z}_i\right)\left(Z_{j,t+k} - \bar{Z}_j\right)}{\left[\left(\sum_{t=1}^{n}\left(Z_{i,t} - \bar{Z}_i\right)^2\right)\left(\sum_{t=1}^{n}\left(Z_{j,t} - \bar{Z}_j\right)^2\right)\right]^{1/2}} \quad (2)$$

Element of cross correlation weight matrix is obtained as:

$$w_{ij} = \frac{r_{ij}(k)}{\sum_{i\neq j}^{n}\left|r_{ij}(k)\right|} \quad (3)$$

And satisfy condition $\sum_{i\neq j}\left|w_{ij}\right| = 1$, this weight give all possibilities at appropriate time lag also has flexible on the sign.

2.2 Generalized Space-Time Autoregressive Integrated Moving Average

Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) is a model that combined time series analysis with spatial dependent developed from STARIMA model which parameters of each location are assumed heterogenous[4]. Thus in this model, the autoregressive and moving average parameters denoted by matrix such as Φ and Θ. If W denoted spatial weight matrix then GSTARIMA model can be written as:

$$\nabla Z_t = \sum_{s=1}^{p}\left[\Phi_{s0}\nabla Z_{t-s} + \sum_{k=1}^{\lambda_s}\Phi_{sk}W^{(k)}\nabla Z_{t-s}\right] - \sum_{s=1}^{q}\left[\Theta_{s0}e_{t-s} + \sum_{k=1}^{v_s}\Theta_{sk}W^{(k)}e_{t-s}\right] + e_t \quad (4)$$

Where: $\Phi_{sk} = diag\left(\phi_{sk}^{(1)},...,\phi_{sk}^{(N)}\right)$ and $\Theta_{sk} = diag\left(\theta_{sk}^{(1)},...,\theta_{sk}^{(N)}\right)$. S denoted time series order and k denoted spatial order which in this paper restrained at lag 1.

3 Stage of Analysis

1. Explore data
2. Split data into two parts which are training and testing. Training data is used for building prediction models and testing data is used for measure accuracy of models
3. Checking stationarity of data with Augmented Dickey Fuller (ADF) test, if data is not satisfy stationary assumption then differencing should be carried out
4. Calculating spatial weight matrices which are distance inverse and cross correlation
5. Identifying autoregressive and a-moving average order of GSTARIMA through smallest value of Akaike's Information Criterion Corrected (AICC) and highest order of ARIMA models from each location basic on training data
6. Building GSTARIMA models based on distance inverse weight and cross correlation weight using training data and estimate the parameters of models
7. Validation of models to diagnose assumption of residuals whether have behaviour like white noise
8. Evaluate model by comparing Mean Absolute Percentage of Error (MAPE) of both models use testing data

4 Model Building to Forecast Rice Harvest Area

Before time series modeling first step have to be done is checking stationarity of data. Test of stationarity can be carried out by perform Augmented Dickey-Fuller (ADF) test to training data with null hypothesis (H_0) data is not stationary. Rice harvest area has seasonal pattern in twelve months because very depend on rainfall, waters supply and climate. Table 1 shows result for ADF test, data from all location is satisfy stationarity assumption in non-seasonal component, it can be seen from p-value which is less than 0.05 imply reject null hypothesis, in other hand data is not stationary in seasonal component (s = 12) which is p-value more than 0.05, hence differencing has to be carried out at lag 12.

TABLE 1. ADF TEST FOR EACH LOCATION

Province	Non-Seasonal		Seasonal (s = 12)	
	Rho	P-Value	*Rho*	P-Value
West Java	-18.8597	0.0018	-4.9846	0.2296*
Central Java	-20.9243	0.0009	-5.6423	0.2003*
East Java	-31.8096	< 0.0001	-4.7455	0.2411*
South Sulawesi	-30.7318	< 0.0001	1.3822	0.6484*
South Sumatera	-44.3035	< 0.0001	-1.2984	0.4499*
North Sumatera	-9.4145	0.0313	-1.1084	0.4635*

Association between locations represented by spatial weight matrix. Distance inverse weight matrix describes geographically distance influence toward forecasting model. Distance inverse matrix calculated for six province for West Java (Z1), Central Java (Z2), East Java (Z3), South Sulawesi (Z4), South Sumatera (Z5), North Sumatera (Z6) is:

	Z1	Z2	Z3	Z4	Z5	Z6
Z1	0	0.4080	0.2107	0.0834	0.2122	0.0857
Z2	0.3500	0	0.3728	0.0876	0.1266	0.0630
Z3	0.2229	0.4598	0	0.1319	0.1180	0.0674
Z4	0.1949	0.2386	0.2913	0	0.1589	0.1164
Z5	0.3106	0.2159	0.1632	0.0996	0	0.2107
Z6	0.2059	0.1762	0.1529	0.1196	0.3455	0

Fig. 1. Distance Inverse Weight Matrix

Cross correlation weight matrix describes correlationally association among variable between locations influenced toward forecasting. Calculated result of cross correlation weight matrix for six province above is:

	Z1	Z2	Z3	Z4	Z5	Z6
Z1	0	0.2477	0.2607	0.2053	0.2284	0.0579
Z2	0.2668	0	0.2929	0.1204	0.2055	0.1144
Z3	0.2741	0.2860	0	0.1397	0.2176	0.0827
Z4	0.3232	0.1759	0.2091	0	0.2889	–0.0029
Z5	0.2359	0.1970	0.2136	0.1895	0	0.1640
Z6	0.1436	0.2632	0.1950	–0.0046	0.3936	0

Fig. 2. Cross Correlation Weight Matrix

ARIMA model from each location determined based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) cut off at certain lag and model with smallest value of Akaike's Information Criterion (AIC) is selected from tentative models. Table 2 shows ARIMA model for each location and obtained all location have similiar order at seasonal component but different for two locations from others at non-seasonal component such as West Java and South Sulawesi

TABLE 2. SELECTED MODELS OF ARIMA FOR EACH LOCATION

Province	Order	AIC
West Java	(2,0,2)*(0,1,1)12	2282.879
Central Java	(2,0,0)*(0,1,1)12	2283.466
East Java	(2,0,0)*(0,1,1)12	2356.695
South Sulawesi	(0,0,1)*(0,1,1)12	2170.466
South Sumatera	(2,0,0)*(0,1,1)12	2111.815
North Sumatera	(2,0,0)*(0,1,1)12	2043.426

Identifying order of GSTARIMA model can be determined through Akaike's Information Criterion Corrected (AICC) by smallest value. Selected order of AICC used for non-seasonal pattern and can not be used for seasonal pattern. This procedure is performed for practicality reason than Matrix Autocorrelation Function (MACF) and Matrix Partial Autocorrelation Function (MPACF) which need experiences and subjective judgment of researcher. Table 3 shows that selected order of GSTARIMA for non-seasonal component is follows autoregressive process at order 2 or simply written AR (2) which has smallest value of AICC.

TABLE 3. AICC FOR TENTATIVE GSTARIMA MODEL

Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	121.2515	120.9154	121.0755	121.3752	121.7476	122.54
AR 1	120.2888	120.4237	120.9727	121.5597	122.1696	123.3198
AR 2	**120.1161**	120.5071	121.502	122.2456	123.2046	124.7704
AR 3	120.6582	121.1109	122.2907	123.0794	124.4825	126.2925
AR 4	121.3003	121.9343	123.4403	124.5805	126.6177	128.9511
AR 5	122.4503	122.9362	124.3858	126.2488	129.1583	133.203

Seasonal order of GSTARIMA determined by highest order of ARIMA model from each location. From table 2 order for seasonal pattern of GSTARIMA can be obtained follows moving average process at order 1 or simply written MA (1) with s = 12 because all location have similiar order at seasonal component.

Thus selected order of GSTARIMA models obtained from process above is (2,0,0)*(0,1,1)12. The model can be written as follows:

$$\nabla^{12}Z_t = \Phi_{10}\nabla^{12}Z_{t-1} + \Phi_{11}W^{(1)}\nabla^{12}Z_{t-1} + \Phi_{20}\nabla^{12}Z_{t-2} + \Phi_{21}W^{(1)}\nabla^{12}Z_{t-2} - \Theta_{12,0}e_{t-12} - \Theta_{12,1}W^{(1)}e_{t-12} + e_t \quad (5)$$

5 Comparing Model Based on Spatial Matrices Weight

GSTARIMA model based on distance inverse and cross correlation used for forescast rice harvest area to 12 periods (months) ahead. The result of forecasting paired with testing data used to calculate Mean Absolute Percentage Error (MAPE) . Table 4 shows MAPE values for each location almost all location have under 25 % except one location which is South Sumatera. The result of MAPE from model with distance inverse weight and cross corelation at each location tends to balanced. In other hand from MAPE Total values model with cross correlation has better accuracy in forecasting because it has smaller value than model with cross correlation. Thus it can be interpreted that rice harvest area forecasting with GSTARIMA more influenced by correlationally association among variable between locations and not depend on geographically distances.

TABLE 4. MAPE GSTARIMA

Province	*GSTARIMA*	
	Distance Inverse	Cross Correlation
West Java	23.857	22.888
Central Java	21.868	21.588
East Java	17.356	18.863
South Sulawesi	23.663	24.741
South Sumatera	42.511	41.870
North Sumatera	21.427	22.092
MAPE Total	**13.421**	**13.243**

6 Conclusion

Rice harvest area has seasonal pattern which has to accomodated in forecast modeling. GSTARIMA models to forecasting rice harvest area built from autoregressive process in non-seasonal component which is AR (2) and moving average process in seasonal component which is MA (1). Generally GSTARIMA model with with cross correlation weight has better accuracy than model with distance inverse weight. It can be interpreted that forecasting rice harvest area with GSTARIMA is more depend on correlationally associations among variables between locations

References

[1] B.N. Ruchjana, "Pemodelan Kurva Produksi Minyak Bumi Menggunakan Model Generalisasi S-TAR", Forum Statistika dan Komputasi. Bogor, Indonesia, 2002.

[2] P.E. Pfeifer, S.J. Deutsch, "A STARIMA Model-Building Procedure With Application to Description and Regional Forecasting. Transactions of the Institute of Bristish Geographers", Vol. 5, No. 3 (1980), pp. 330-349, 1979.

[3] Suhartono, Subanar, "The Optimal Determination of Space Weight in GSTAR Model by using Cross-Correlation Inference". Journal of Quantitative Method: Journal Devoted to The Mathematical and

Statistical Application in Various Fields, Vo, 2, No. 2, pp. 45-53, 2006.

[4] X. Min, J. Hu, Z, Zhang. "Urban Traffic Network Modeling and Short-term Traffic Flow Forecasting Based on GSTARIMA Model", Annual Conference on Intelligent Transportation Systems Madeira Island, Portugal, September 19-22, 13th International IEEE, 2010.