Paddy Availability Modeling in Indonesia Using Spatial Regression

Yuliana Susanti[1], Sri Sulistijowati H.[2], Hasih Pratiwi[3], and Twenty Liana[4] *

Abstract—**Paddy is one of Indonesian staple food in which its availability is highly needed. A prediction model of paddy availability in future such as by means of spatial regression is deemed necessary. The purpose of this study is to construct a spatial regression model to predict paddy production in Indonesia. The result of the research showed that paddy production could be presented using lag spatial regression with some influencing factors including harvested area, monthly average temperatures and numbers of workers.**

Keywords: spatial regression, Lagrange multiplier, lag spatial regression, error spatial regression.

1 Introduction

Paddy is Indonesian staple food consumed by most of Indonesian citizens. It is a commodity that has political and strategic value since the stability of its availability and commodity price become the indicators to measure Indonesian government success. For the reason, availability and adequacy of paddy must be continuously well concerned. The too high price of paddy most frequently stimulates the social issues.

According to BPS [2] paddy production in 2011 decreased by 23 tons (1.85 percent) compare to the one in 2010. In 2012 (prediction number I), paddy production was projected to rise until 21 tons (1.72 percent) compare to paddy production last 2011. Until recently, the efforts to fulfill national staple food needs through food self-sufficiency of strategic commodity of paddy have not shown an optimal result. This situation is clearly reflected in availability levels of some domestic food commodities that are relied highly upon the import such as soybean at 70 percent while paddy is only about 5 percent. Nevertheless, under the development program of national food security in 2012, the paddy production as the commodity of carbohydrates food increased by 2.74 percent. The policy of stable food price is one of important instrument to maintain national food security. Considering that food price is highly determined by the availability or food production, an attempt to predict paddy production in future is deemed essential. There are some methods that can be applied to predict paddy production and to analyze its influencing factors, one of which is through regression analysis.

Regression analysis is one of the methods applied to observe the relation between variables and to estimate or predict. Zhang [7] stated that the use of the least square method in regression analysis will not be accurate in solving problems which contain outliers or extreme observation. In this case, paddy production which is reached beyond general production can be categorized as outliers; thus, the use of the least square method to estimate the regression parameters is considered inappropriate (Susanti and Pratiwi [5], Yohai [6]). Besides, paddy production is influenced by geographic and demographic factors. The relation of the factors to paddy production can be identified through spatial regression analysis. Something closely related to the others rather than distant relation has a more significant influence (Anselin [1]). Lag spatial regression model is a regression model which input spatial or position impacts. This model follows both an autoregressive process which can be observed from dependency between one area with another related area and error regression model which focuses on error values dependency of an area with errors in another related one. Based on that, regression model of paddy production with spatial regression is needed.

2 RESEARCH METHOD

This research uses secondary data gathered from related government department, which is Indonesian Ministry of Agriculture and Statistics Indonesia in 2012. The data included the amount of paddy production in province i-th (Y_i), harvested area of paddy in province i-th (X_{1i}), average monthly rainfall in province i-th (X_{2i}), average monthly humidity in province i-th (X_{3i}), average monthly temperature in province i-th (X_{4i}), average monthly time of irradiation in province i-th (X_{5i}), a-verage rice price in consumers level in province i-th (X_{6i})

*[1]Mathematics Department, Faculty of Mathematics and Natural Sciences Sebelas Maret University, Jl. Ir. Sutami 36A Surakarta, Indonesia Email: yuliana.susanti@ymail.com [2]Mathematics Department, Faculty of Mathematics and Natural Sciences Sebelas Maret University, Jl. Ir. Sutami 36A Surakarta, Indonesia Email: ssulistijowati@yahoo.com [3]Mathematics Department, Faculty of Mathematics and Natural Sciences Sebelas Maret University, Jl. Ir. Sutami 36A Surakarta, Indonesia Email: hpratiwi@mipa.uns.ac.id [4]Assessment Institute for Agricultural Technology of Kalimantan Tengah, Jl. G. Obos Km. 5 Palangkaraya, Indonesia Email: twentyliana@yahoo.com

and number of workers in sub sector of food plantation in province i-th (X_{7i}).

In this research, the data was provided for each area unit or province; thus, the spatial effects testing become critical. In contrast, the ignorance of spatial effect can make the estimation inefficient and conclusion inappropriate. This study was started by determining the best regression model by means of gradual regression method and regression model assumption testing. Further, it was followed by testing on spatial autocorrelation, position and space effects using Lagrange multiplier test. When finding spatial autocorrelation and space impact, spatial regression modeling might be conducted using spatial weighting matrix. It was continued by conducting spatial regression model assumption testing and finally by predicting the paddy availability in Indonesia.

3 SPATIAL REGRESSION MODEL

Common model of spatial regression (Spatial Autoregressive Moving Average, SARMA) in matrix form (Lesage [4], Anselin [1]) could be written as

$$\begin{aligned} y &= \rho W_y + X\beta v + u \\ u &= \lambda W_u + \varepsilon \\ \varepsilon &\sim N(0,\sigma^2) \end{aligned} \tag{1}$$

with y as dependent variable vector of $n \times 1$, X as independent variable matrix of $n \times (k+1)$, β as regression parameter coefficient vector of $(k+1)\times 1$, ρ as spatial lag dependent variable of coefficient parameter, λ as spatial lag error of coefficient parameter, u, ε as error vector of $n \times 1$, W as weighting matrix of $n \times n$, n as the numbers of location, k as the numbers of independent variable ($k = 1, 2, \dots, l$) and I as identity matrix of $n \times n$. Spatial regression model could be improved from model (1), that is if $\rho = 0$ and $\lambda = 0$ so model (1) is regression linear model $y = X\beta + \varepsilon$, if value $\rho \neq 0$ and $\lambda = 0$ so model (1) is lag spatial regression, $y = \rho Wy + X\beta + \varepsilon$, and if value $\rho = 0$ dan $\lambda \neq 0$ so model (1) is error spatial regression model, $y = X\beta + \lambda Wu + \varepsilon$. Next Moran index scatter diagram could be applied to test spatial dependency. This diagram is a diagram used to see a relation between observation value in a location and average observation value based on neighborhood locations and related locations (Lee and Wong, [3]).

To decide spatial regression model used, Lagrange lag multiplier and error testing should be done. Lagrange lag multiplier testing was used to observe spatial dependency in dependent variable while Lagrange error multiplier testing was used to see error spatial dependency model. The statistic of lag Lagrange multiplier testing is

$$LM_p = (1/T)(e'Wy/\sigma^2)^2 \sim X^2(1)$$

where
$T = \text{trace } (W + W').*W$
$\sigma^2 = e'e/n$
e: residual value of MKT
n: number of observation
C: Wqueen standard matrix

The decision making here is H_0 is rejected if value $LM_\rho > X^2(\alpha;1)$ or p-value $< \alpha$. Meanwhile, statistic of error Lagrange multiplier testing is

$$LM\lambda = (e'We/\sigma^2)[T_{22} - (T_{21})^2 var(\rho)]^{-1} \sim X^2(1)$$

where

$$\begin{aligned} T_{22} &= trace(W*W + W'W), \\ T_{21} &= trace(W*CA^{-1} + W'CA^{-1}), \\ A &= (In - \rho C) \end{aligned}$$

with decision making that H_0 is rejected if value $LM\rho > X^2(\alpha;1)$ or p-value $< \alpha$.

3.1 Parameter Estimate of Spatial Regression Model

In equation (1) if value $\rho \neq 0$ and $\lambda = 0$, spatial regression model would be lag spatial regression model (Anselin, [1]) with equation:

$$y = \rho W_y + X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2 I)$.

Likelihood maximum method was used to estimate lag spatial regression model parameter. The likelihood function for $\varepsilon \sim N(0, \sigma^2 I)$ could be formulated as

$$L(\varepsilon,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\Big(-\frac{\varepsilon^T\varepsilon}{2\sigma^2}\Big)$$

where $\varepsilon = y - \rho W_y - X\beta$.

Partial derivative of likelihood function from lag spatial regression model for dependent variable (y) is

$$\Big|\frac{\partial_\varepsilon}{\partial_y}\Big| = \Big|\frac{\partial_{y-\rho W_y - X\beta}}{\partial_y}\Big| = |I - \rho W|$$

Thus, likelihood function for dependent variable can be written as

$$L(\lambda,\beta,\sigma^2) = \frac{|I-\rho W|}{(2\pi)^{n/2}\sigma^n}\times \exp\left(-\frac{(y-\rho Wy - X\beta)^T(y-\rho Wy - X\beta)}{2\sigma^2}\right)$$

with its log likelihood function as

$$\ln\big(L(\lambda,\beta,\sigma^2)\big) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}(\sigma^2) + \ln|I-\rho W| - \frac{(y-\rho Wy - X\beta)^T(y-\rho Wy - X\beta)}{2\sigma^2} \tag{2}$$

To obtain the estimation of σ^2, β and λ could be conducted by maximizing log likelihood function in equation (2). Thus, the estimation for σ^2 obtained was as follows:

$$\widehat{\sigma^2} = \frac{(y - \rho Wy - X\beta)^T(y - \rho Wy - X\beta)}{2\sigma^2}$$

and estimation for β is

$$\hat{\beta} = (X^TX)^{-1}X^T(I - \rho W)y$$

Parameter estimation for λ could be done through an approach with numeric method. Next, based on equation (1), if value $W_1 = 0$ or $\rho = 0$, spatial regression model would be error spatial model with equation of

$$y = X\beta + \lambda W_2 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2 I)$. Maximum likelihood method was used to estimate error spatial regression model parameter. Likelihood function for $\varepsilon \sim N(0, \sigma^2 I)$ in equation above is

$$L(\varepsilon, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\Big(-\frac{\varepsilon^T\varepsilon}{2\sigma^2}\Big)$$

where $\varepsilon = (I - \lambda W)(y - X\beta)$. Likelihood function from error spatial regression model for dependent variable (y) is

$$\Big|\frac{\partial_\varepsilon}{\partial_y}\Big| = \Big|\frac{\partial_{(I-\lambda W)(y-X\beta)}}{\partial_y}\Big| = |I - \lambda W|$$

Thus, likelihood function for dependent variable is

$$L(\lambda, \beta, \sigma^2) = \frac{|I - \lambda W|}{(2\pi)^{n/2}\sigma^n} \times$$
$$\exp\left(-\frac{(I - \lambda W)^T(I - \lambda W)(y - X\beta)^T(y - X\beta)}{2\sigma^2}\right)$$

Therefore, likelihood log function is

$$\ln\big(L(\lambda, \beta, \sigma^2)\big) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}(\sigma^2) + \ln|I - \lambda W| - \frac{(I - \lambda W)^T(I - \lambda W)(y - X\beta)^T(y - X\beta)}{2\sigma^2} \tag{3}$$

To obtain the estimator of λ, β, σ^2 the derivation of (3) was conducted to each parameter to obtain

$$\frac{\partial \ln\big(L(\lambda, \beta, \sigma^2)\big)}{\partial y} = \Sigma_{i=1}^n \frac{-\omega_t}{1 - \lambda\omega_t} -$$
$$\frac{(y - X\beta)^T(I - \lambda W)(y - X\beta)}{2\sigma^2} \frac{\partial \ln\big(L(\lambda, \beta, \sigma^2)\big)}{\partial \beta}$$
$$= \frac{-X^T(I - \lambda W)^T(I - \lambda W)(y - X\beta)}{2\sigma^2} \frac{\partial \ln\big(L(\lambda, \beta, \sigma^2)\big)}{\partial \sigma^2}$$
$$= \frac{-n}{2\sigma^2} + \frac{(y - X\beta)^T(I - \lambda W)^T(I - \lambda W)(y - X\beta)}{2\big(\sigma^2\big)^2}$$

so that

$$\hat{\beta} = [(X - \lambda W^r X)^T(X - \lambda WX)]^{-1}(X - \lambda Wy)T(y - \lambda W)y$$

According to Anselin [1], determination coefficient (R^2) in regression model could be stated as

$$R^2 = 1 - \frac{e^T(I - \lambda W)^T(I - \lambda W)e}{(-Ay_w)^T(I - \lambda W)^T(I - \lambda W)(y - Ay_w)}$$

where A is unit vector $(nx1)$, $y_w = n(\lambda - 1)^2$. Determination coefficient is nonnegative with $0 \leq R^2 \leq 1$. The closer the value R^2 with 1, the more suitable model with the data.

3.2 Significance Test for Spatial Regression Model

Significance testing of parameter model was conducted using likelihood ratio test. It was done to see which parameter influenced model significantly. The steps to test parameter include:

(i) $H_0 : \rho, \lambda = 0$ (spatial parameter was not significantly influenced)
$H_1 : \rho, \lambda \neq 0$ (spatial parameter was significantly influenced)

(ii) Significance level (α)

(iii) Critical area: rejects H_0 if $LRT_{\rho,\lambda} > \chi^2_{(\alpha;1)}$

(iv) Statistic of likelihood ratio testing for lag spatial regression model is

$$\begin{aligned} LRT_\rho &= -2[\ln\big(L(\rho, \beta, \sigma^2)\big) - \ln\big(L(\beta, \sigma^2)\big)] \\ &= -2[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}(\sigma^2) + \ln|I - \rho W| \\ &\quad - \frac{(y - \rho W_y - X\beta)^T(y - \rho W_y - X\beta)}{2\sigma^2} \\ &\quad + \frac{n}{2}\ln(2\pi) + \frac{n}{2}(\sigma^2) + \frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}] \\ &= -2\ln|I - \rho W| \\ &\quad + \frac{(y - \rho W_y - X\beta)^T(y - \rho W_y - X\beta)}{2\sigma^2} \\ &\quad - \frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2} \end{aligned}$$

Meanwhile, the statistic of likelihood ratio testing for error spatial regression model is

$$\begin{aligned}
LRT_\lambda &= -2[\ln\big(L(\lambda,\beta,\sigma^2)\big) - \ln\big(L(\beta,\sigma^2)\big)] \\
&= -2[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}(\sigma^2) + \ln|I-\lambda W| \\
&- \frac{(y-X\beta)^T(I-\lambda W)^T(I-\lambda W)(y-X\beta)}{2\sigma^2} \\
&+ \frac{n}{2}\ln(2\pi) + \frac{n}{2}(\sigma^2) + \frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}] \\
&= -2\ln|I-\lambda W| \\
&+ \frac{(y-X\beta)^T(I-\lambda W)^T(I-\lambda W)(y-X\beta)}{2\sigma^2} \\
&- \frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}
\end{aligned}$$

(v) Conclusion: If $LRT > \chi^2_{(\alpha;1)}$, spatial parameter is significantly affected to the model.

4 RESULTS AND DISCUSSION

Linear regression model of paddy production in Indonesia in 2011 by using ILS method can be written by

$$\begin{aligned}
Y_i &= 8563420 + 4.894764X_{1i} - 453.7847X_{2i} \\
&- 24534.28X_{3i} - 250119.5X_{4i} - 64.33128X_{5i} \\
&+ 15.44982X_{6i} + 0.4266961X_{7i}
\end{aligned} \tag{4}$$

In linear regression model (4), there are some variables which are not significantly affected. For this, it was necessary to conduct the best regression model using step method. In a paddy plantation production case in 2011, the best regression model that was obtained was a model by adding independent variables such as the paddy harvested area, average monthly temperature and numbers of farming workers in sub sector food plantation. The linear regression model was conducted through a step method as follows:

$$Y_i = 5747827 + 5.032749X_{1i} - 220472.5X_{4i} + 0.4093232X_{7i}$$

with value R^2 was 0.994140. It means that 99.4 % paddy production in Indonesia in 2011 could be explained by the paddy harvested area, average monthly temperature and numbers of farming workers in sub sector food plantation. Meanwhile, the rest at 0.586 % was explained by other unobserved factors in this study. In the entire test on the best regression model, it was obtained the value of $4.5515 \times 10^{-33} < \alpha = 0.05$; thus, it can be concluded that at least there was one regression parameter significantly influential for the paddy production in Indonesia in 2011. The values of parameter estimation and p-value to each parameter could be observed in Table 1. As seen in Table 1, three parameters in regression model have a significant influence because each of p-value $< \alpha = 0.05$. Then, regression assumption model testing was done with the result showing the fulfilled normality, homogeneity and non multi collinearity assumptions.

Table 1: Parameter estimation value and p-value to each parameter

Independent variable	Parameter estimation value	p-value
Constants	5747827	0.0002957
X_1	5.032749	0.0000000
X_4	-220472.5	0.0002068
X_7	0.4093232	0.0062885

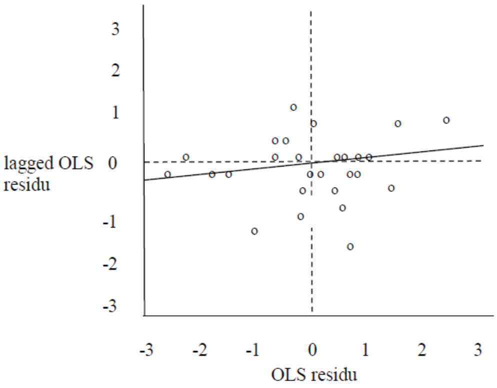

Figure 1: Scatter Diagram of Moran Index for Paddy Production

However, error on closed-relation area may connect or was spatially correlated. The indicators of any auto spatial correlation effects could be observed from Moran index in scatter diagram as seen in Figure 1. Diagonal lines were closed to zero, meaning no spatial autocorrelation in the model by using spatial regression model. Furthermore, steps to construct spatial regression model were conducted by testing spatial effects with Lagrange multiplier, maximum likelihood estimation, parameter significance testing, assumptions testing.

4.1 Lag Spatial Model

The Moran's scatter diagram can only be applied to identify the existence of spatial autocorrelation in a certain area but cannot discover spatial autocorrelation in model. Hence Lagrange lag and error testing were highly needed. The Lagrange lag multiplier testing included:

1. H_0: there was no lag spatial dependency ($\rho = 0$)
   H_1: there was lag spatial dependency ($\rho \neq 0$).
2. Significance level $(\alpha) = 0.05$.
3. Critical area
   H_0 was rejected if $LM_p > \chi^2(0.05, 1) = 3.841$.
4. Statistic value, $LM_p = 5.3173358$.
5. Conclusion
   Since $LM_p = 5.3173358 > 3.841$ so we concluded

that H_0 was rejected meaning that there was lag spatial dependency in the model.

Therefore, regression model had spatial autocorrelation in lag; thus, it could be presented with lag spatial regression model. The value of parameter estimation of lag spatial regression model and z could be calculated as seen in Table 2. Model of lag spatial discussed can be seen as follows:

$$Y = 0.03951494W_y + 4.996965X_1 - 252877.3X_4 + 0.3585264X_7. \tag{5}$$

Based on Table 2 it could be seen that absolute value of z calculated in each independent variable was greater than $z_{0.025} = 1.96$; thus, it can be concluded that paddy harvested area, average monthly temperature and numbers of farming workers in sub sector food plantation had significant influence on Indonesia paddy production in 2011. After obtaining the estimation of parameter model, the significance testing of parameter model was conducted using likelihood ratio testing that included:

Table 2: Parameter estimation value and z value calculated in lag spatial regression model

Independent variable	Parameter estimation	z value
W_y	0.03951494	2.49251
Constants	6577387	0.0000001
X_1	4.996965	26.33706
X_4	252877.3	-5.425352
X_7	0.3585264	2.963367

1. H_0: $\rho = 0$ (spatial parameter was not significantly influenced)
   H_1: $\rho \neq 0$ (spatial parameter was significantly influenced).
2. Significance level $(\alpha) = 0.05$.
3. Critical area
   H_0 was rejected if $LRT_\rho > \chi^2(0.05, 4) = 3.841$.
4. Statistic value, $LRT_\rho = 5.716419$.
5. Conclusion
   Since $LRT_\rho = 5.716419 > 3.841$, it could be concluded that H_0 was rejected meaning that lag spatial parameter was significantly influenced.

Furthermore, homoscedastic testing was conducted to determine whether a spatial lag regression models met the assumptions. Homoscedastic test performed by Breusch-Pagan could be seen as follows:

1. H_0: There was no heteroscedasticity H_1: There was heteroscedasticity
2. Significance level $(\alpha) = 0.05$.
3. Critical area
   H_0 was rejected if $BP > \chi^2_{(0.05,2)} = 9.488$.
4. Statistic value, $BP = 1.453798$.
5. Conclusion
   Since $BP = 1.453798 < 9.488$, it then could be concluded that H_0 was not rejected meaning that there was no heteroscedasticity in the model.

4.2 Error Spatial Model

The initial step in presenting error spatial model of Indonesia paddy production data in 2011 is Lagrange multiplier testing:

1. H_0: $\lambda = 0$ (spatial parameter was not significantly influenced)
   H_1: $\lambda \neq 0$ (spatial parameter was significantly influenced).
2. Significance level $(\alpha) = 0.05$.
3. Critical area
   H_0 was rejected if $LRT_\lambda > \chi^2_{(0.05,4)} = 3.841$.
4. Statistic value, $LRT_\lambda = 0.2367617$.
5. Conclusion
   Since $LRT_\lambda = 0.2367617 < 3.841$, it could be concluded that H_0 was not rejected meaning that error spatial parameter was not significantly influenced.

Based on Lagrange error multiplier testing, it could be concluded that this model did not have any spatial autocorrelation in error. Thus, given lag spatial dependency and no error spatial dependency causes Indonesia paddy production in 2011 could be presented as lag spatial regression model. Model (5) shows that an increase of one hectare of paddy harvested area would increase paddy production as 4.996965 tons. Each increasing of 1^0C average monthly temperature would reduce paddy production as 252877.3 tons and each of increase of one agricultural labor in food crop sub sector would increase paddy production as 0.3585264 ton. Spatial autocorrelation at lag in the model (5) is shown by the spatial lag parameter value of 0.03951494.

Based on lag spatial model, paddy production in all provinces was estimated so the production map in Indonesia could be created (Figure 2). There are 21 provinces that are included in the low category and 9 provinces included in the moderate category. Provinces with highest paddy production are Central Java, East Java and West Java.

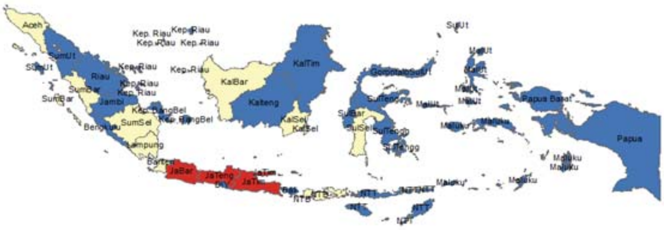

Figure 2: Prediction of Paddy Production in Indonesia

References

[1] Anselin, L., *Spatial Econometrics: Methods and Models*, Kluwer Academic Press, London, 1988.

[2] Badan Pusat Statistik, *Production of Paddy, Maizze, and Soybeans*, www.bps.go.id/releases/ Production_of_Paddy_Maizze_ and_Soybeans, 2012.

[3] Lee, J., and Wong, D.W.S, *Statistical Analysis with ArcView GIS*, John Wiley and Sons, Inc., New York, 2001.

[4] Lesage, J.P., *The Theory and Practice of Spatial Econometrics*, Dept. of Economics, University of Toledo, Ohio, 1999.

[5] Susanti, Y. and Pratiwi, H., Robust Regression Model for Predicting the Soybean Production in Indonesia, *Canadian Journal on Scientific and Industrial Research*, Vol. 2, No. 9, pp. 318-328, 2001.

[6] Yohai, V.J., High Breakdown Point and High Efficiency Robust Estimates for Regression, *The Annals of Statistics*, Vol. 15, No. 20, pp. 642-656, 1987.

[7] Zhang, Z., *Robust Estimations*, http: //www.sop. Inria.tr/robust/personel/zzhang, 1996.