International Journal of Scientific & Engineering Research Volume 8, Issue 7, July-2017
ISSN 2229-5518

GSTARIMA Modeling of Area Extent of Rice Stem Borer Attack Using Spatial Weights Clustering

Niki Ariski, I Made Sumertajaya, Muhammad Nur Aidi

Abstract— patial weights based on distance and neighborhood are the geographics spatial approach that always been used as the weight in space time analysis. Geographics spatial approach is depending on the position between the locations. In this research, weights used were based on the approach of distribution path. The distribution path approach used was by grouping the location (clustering spatial) on the area extensive data of rice stem borer attack based on similarity of pest habitat using fuzzy c-means cluster analysis. Data of similarity of pest habitat used were climate data and also rice productivity data of each location. Intergroup centers obtained from cluster analysis were then used to calculate the distance matrix between locations. The model used was seasonal GSTARIMA (2,0,0;1)(0,0,1;1)6. The weights used was the weight of the queen contiguity obtained based on geographical approach and the clustering spatial weights obtained based on the distribution path approach. Seasonal model of GSTARIMA (2,0,0;1)(0,0,1;1)6 with spatial clustering weights was the best model based on the smallest RMSE and MAPE values i.e RMSE value of 0.450 and MAPE value of 14.898%

Index Terms— rice stem borer pest, fuzzy c-means, space time, seasonal.

1 INTRODUCTION

Rice (*Oryza Sativa L.*) is the main staple food for people in Indonesia. Rice production is always expected to meet the food needs of the community. The factors that affect the decline of rice production include the presence of pests, the disturbing organisms that cause damage to rice plants. Pest attack on rice plants leads to decrease harvested land area which also affects the amount of rice yield. There are several types of pests for rice plants, one of which is a stem borer (Sundep / Beluk) which attacked rice plants in whole phases of rice plant growth from seedbed to harvest.

The rice stem borer is one of the major pests of rice crops in Indonesia with the distribution of being influenced by the climate. The distribution of rice stem borer is particularly prevalent in the tropics, while in the sub-tropics it limited by the temperatures above 10 °C with a rainfall above 1,000 mm (Pathak, 1967 in Asikin and Thamrin, 2001). In Indonesia, there are 6 types of rice stem borer of which 4 species are most commonly found, namely yellow rice stem borer (*Scirpophaga incertulas*), white rice stem borer (*Scirpophaga innotata*), stripped rice stem borer (*Chilo sup-pressalis*), and pink rice stem borer (*Sesamia inferens*).

The extent of stem borer attack in some areas for a certain period of time is one of the multivariate time series data in which it includes space time factor. The space time model of autoregressive integrated moving average (STARIMA) was first introduced by Pfeifer and Deutch (1980). The STARIMA model assumes the same parameter values for each location so that the model is suitable for homogeneous location characteristics. The generalized space time model of auto-regressive integrated moving average (GSTARIMA) was introduced by Borovkova, et al (2002) as an extension of the STARIMA model which assumes different parameter values for each location that are commonly found in practice.

STAR and GSTAR model simulation has been carried out by Wijaya (2015). The result of modeling with the same Autoregressive (AR) time order for each location produced the best STAR model in forecasting, whereas the different AR time order modeling for each location produced the best GSTAR model.

Spatial weights on the space time model is a matrix that describes the relationship between a location and another location. The weighted matrix is obtained based on distance or neighborhood information and relies on the position between locations. This research used weights that referred to the weight obtained through the approach of distribution flow. The distribution path approach of stem borer distribution which is the basis of this research is based on similarity of habitat, in this case, using climate condition data and also rice productivity data at each location. Locations with similar climatic conditions will be grouped into one group. Ruchjana, et al (2013) classified the location according to the oil absorption rate of the new oil well replacement data and then used the GSTAR-Kriging model in each group. This study used the intergroup distance obtained by fuzzy c-means cluster analysis which will then be used as weights in the space time analy-

- *Niki Ariski is currently pursuing masters degree program in statistics in Bogor Agricultural University, Indonesia, PH +6285255926771. E-mail: niki.ariski91@gmail.com*
- *I Made Sumertajaya is Lecturer, Departement of Statistics, Bogor Agricu tural University, Bogor, Indonesia. E-mail: imsjaya@yahoo.com*
- *Muhammad Nur Aidi is Lecturer, Departement of Statistics, Bogor Agricu tural University, Bogor, Indonesia. E-mail: nuraidi@yahoo.com*

sis so that the model for each location can still be obtained. The objective is to compare GSTARIMA modeling using weights based on geographical approach and distribution path approach on the data of rice stem borer attack on 8 provinces in Sumatera Island.

2 RESEARCH METHOD

2.1 Data

The data used in this study were data obtained from the Directorate General of Food Crops Ministry of Agriculture that are the data of planting area of rice (hectares), and area extensive data of stem borer attack on 8 provinces in Sumatera island within last 7 years (January 2010-November 2016). The data time period was monthly. Data analysis was conducted using data of proportion of the area of pest attack in divided by planted land area at each location.

The cluster analysis for weights used the data from the Meteorology, Climatology and Geophysics Agency i.e temperature (°C), humidity (%), rainfall (mm), and duration of surveillance (hour) for each province by 2015 and data from the Central Statistics Agency was the data of rice crops productivity for each province in 2015.

2.2 Methods of Data Analysis

Steps in analyzing data:

1) Modeling and Forecasting of GSTARIMA model
   a. Dividing data into 2 parts, 2010-2015 data used for modeling (training data) and data year of 2016 used to test the model (testing data).
   b. Checking the stationary of data in the variance and mean, if the data is not stationary in the variance it will be transformed and if the data is not stationary in the mean it will be differenced.
   c. Determining the time series data structure using matrix of autocorrelation function (MACF) plot and partial matrix autocorrelation function (MPACF).
   d. Determining weights of locations using queen contiguity weights and weights from fuzzy c-means cluster analysis.
   e. Modeling the GSTARIMA.
   f. Estimating parameters by using the ordinary least squares method (OLS).
   g. Testing the goodness of the model to test whether the residual was in white noise condition.
   h. Forecasting with GSTARIMA model.
   i. Comparing the accuracy of the GSTARIMA model using the mean absolute percentage error (MAPE), and root mean of square error (RMSE).

2) Determining the Cluster Spatial Weights. Fuzzy C-Means Algorithm (Bezdek, 1981):
   a. Determining the number of clusters to be created. The criteria for the number of groups are given by the smallest value of XB index.

$$XB(c)=\frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^{m}\, d_{ik}^{2}(x_k,v_i)}{N \min_{i,j}\|v_i,v_j\|^{2}} \quad (1)$$

   b. Determining the weights (m)
   c. Initialization of initial matrix $U^{(0)}$ with conditions:

$$u_{ik} \in [0,1],\ \sum_{i=1}^{c} u_{ik} = 1,\ 0 < \sum_{k=1}^{n} u_{ik} < n \quad \text{untuk } \forall\, k \in \{1,2,\ldots,n\}$$

   Initial matrix is randomly selected.
   d. Counting the center of the clusters (v_{ij}) with the equation:

$$V_{ij}=\frac{\sum_{k=1}^{n} u_{ik}^{m} x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (2)$$

   e. Updating the U^{r+1} matrix by using the equation:

$$u_{ik}=\left[\sum_{j=1}^{c}\left(\frac{d_{ik}^{(2)}}{d_{jk}^{(2)}}\right)^{1/(m-1)}\right]^{-1} \quad (3)$$

$$d_{ik}^{2}=(x_k,v_i)=\|x_k - v_i\|^{2}=(x_k - v_i)^{T}(x_k - v_i) \quad (4)$$

   f. Compare the membership value in the U matrix. If $\Delta < \varepsilon$ then the algorithm is already convergent and iteration is stopped. If $\Delta \geq \varepsilon$ then go back to d step.

$$\Delta = \text{abs}\,(U^{r+1} - U^{r}) \quad (5)$$

   ε is a very small positive value and r is an iteration process, $r = 1,2,\ldots$

3 RESULT AND DISCUSSION

3.1 Rice Stem Borer Attack in Indonesia

Rice stem borer is one of the main pests of rice crops in Indonesia. Data from the Directorate General of Food Crops Ministry of Agriculture shows that the average area of rice stem borer attack is 2% of the total farmers' land area from 2010 to 2015. This amount is very large compared to the average area of other rice plant pests such as Brown Planthopper Ropes, Mice, Blasts, BLB/ Crackle, and Tungro which only range below 1% of the total planted area. As a result, this pest attack affects the lack of harvested area which leads to the decrease of production and harvest failure (*puso*).

Table 1. Descriptive statistics of the proportion of rice stem borer attack on 8 provinces in the Sumatera island

Province	Min	Max	Average	STD
Aceh	0.00032	0.09635	0.01897	0.01927
North Sumatera	0.00019	0.00333	0.00101	0.00064
West Sumatera	0.00002	0.00234	0.00027	0.00036
Riau	0.00027	0.07263	0.00909	0.01198
Jambi	0.00027	0.01393	0.00282	0.00231
South Sumatera	0.00031	0.04281	0.00794	0.00883
Bengkulu	0.00059	0.02041	0.00466	0.00369
Lampung	0.00020	0.06817	0.01153	0.01254

Descriptive statistics on the proportion data showed the highly significan different of the attack area, it can be seen in the minimum and maximum value of each province, then the log transformation was coducted.

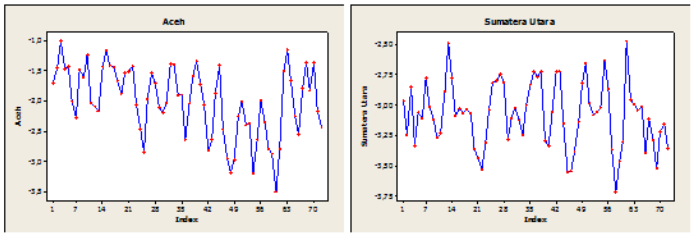

Figure 1. Plot of time series of log data of the proportion of rice plant stem borer attack in Aceh, North Sumatera and South Sumatera province.

3.2 Identification of the Data Stationary

Time series data requires that the data be stationary in average and variance. Formal test that can be conducted was Augmented Dickey-Fuller (ADF) test.

Table 2 Results of Augmented Dickey-Fuller test

Province	Rho	p value	Tau	p value
Aceh	-57.44	0.0007	-5.27	0.0001
North Sumatera	-66.18	0.0007	-5.59	0.0001
West Sumatera	-33.42	0.0007	-4.00	0.0024
Riau	-58.33	0.0007	-5.30	0.0001
Jambi	-42.62	0.0007	-4.48	0.0005
South Sumatera	-190.85	0.0001	-8.90	0.0001
Bengkulu	-81.47	0.0007	-6.27	0.0001
Lampung	-94.08	0.0007	-5.92	0.0251

The data has met the assumption of stationary. The p value was always smaller than the value of α (0.05) for each province in both Rho and Tau statistics

3.3 Model Identification

The MACF plot shows the MA model of the data while the MPACF Plot shows the AR model of the data. The symbol denoted by the sign (+) is defined as positive and significant correlation, the sign (-) means a negative and significant cross correlation, the sign (.) means of cross correlation is not significant.

Variable/Lag	1	2	3	4	5	6
Aceh	.-...-..	.....---	.....-..	..-.....	+.-.....	+.--+..+
North Sumatera	....-+..	+....+..	+-.....	......-.	-.....-.	..-.....
West Sumatera	...-..+.	..+.....	.....	.-..-...	.-..-...	....-..-
Riau	.....-..	.....+	.....+.+	....-...	.-.....-	--..--.-
Jambi	+.....-	-.....+	.....--.	......-.	..-.....	..--....
South Sumatera	+.....	.-.....	.-....--	...+...-	-.....	...-.+++
Bengkulu	...-..+.	..+.....	.....-..	.....	.....	.....+.+
Lampung	...-...+	.-....-.	.-.....-	-....-..	..--....	.+.-.+.+

Figure 2. Matrix Autocorrelation Function (MACF)

Variable/Lag	1	2	3	4	5	6
Aceh	+....-..	.....	.....	.....	.....	.....
North Sumatera	.+.....	.....	...+...+	.....	.....	.....
West Sumatera	...+....	.-.....	.....	.....	.....	.....
Riau	...+.+..	.....+	...-....	.....	.....	.....
Jambi	....+...	.....-.+	.....	.....	.....	.....
South Sumatera	+-...+..	..+..-..	..+.....	.....	.....	.....
Bengkulu	.....	.....	.....	.....	.....	.....
Lampung	...-...+	.-.....	.....	.....	.....	.....

Figure 3. Matrix Parsial Autocorrelation Function (MPACF)

The MACF plot in Figure 3 shows signs (+) and (-) which means significant correlated scattered in all lag, the sign (.) means that not significant correlation also dominating. However, in lag 6 there are still a lot of (-) and (+) which indicate the possibility of a seasonal pattern of 6. The MPACF plot in figure 4 shows that in lag 1 there are signs (+) and (-), then in lag 2 it is still many signs (+) and (-) but in lag 3 signs (+) and (-) are only 4 and the rest is a sign (.) so it can be concluded to start trimmed after the 3rd lag, however, in this study the time order is only limited to lag 2 as more orders are more difficult in interpretation. Thus the data structure is derived from the VARIMA model (2,0,0) $(0,0,1)_6$ then the space time model to be built is GSTARIMA (2,0,0;1) $(0,0,1;1)_6$. Generally, the spatial order is restricted to order 1.

3.3 Formation of the Weights Matrix

3.3.1 Queen Contiguty Weights Matrix

The queen contiguty weighted matrix is derived based on the neighborhood. A value of 1 is given if location-i is directly adjacent to the j-location, whereas a value of 0 is given if the location is not directly adjacent to the j-location.

	Aceh	North Sumatera	West Sumatera	Riau	Jambi	South Sumatera	Bengkulu	Lampung
Aceh	0	1	0	0	0	0	0	0
North Sumatera	0.333	0	0.333	0.333	0	0	0	0
West Sumatera	0	0.250	0	0.250	0.250	0	0.250	0
Riau	0	0.333	0.333	0	0.333	0	0	0
Jambi	0	0	0.250	0.250	0	0.250	0.250	0
South Sumatera	0	0	0	0	0.333	0	0.333	0.333
Bengkulu	0	0	0.250	0	0.250	0.250	0	0.250
Lampung	0	0	0	0	0	0.500	0.500	0

Figure 4 Queen Contiguty Weights Matrix

3.3.2 Weights Matrix of Cluster Spatial

The number of fuzzy c-means (FCM) cluster was determined by the smallest Xie Beni (XB) index. The size of

cluster 4 was a cluster with the smallest index XB of 0.016 so that the selected cluster was 4.

The result of FCM analysis using climate condition data and rice plant productivity at each location was obtained 66 iterations with an error of 0.650 (using software R). From the value of cluster center value obtained then calculated the distance between clusters using Euclidean distance with the formula:

$$\Delta(x, y) = \|x_i - y_i\|^2 \quad (6)$$

$\Delta(x, y)$ is the distance between the center clusters x and y.

	Aceh	North Sumatera	West Sumatera	Riau	Jambi	South Sumatera	Bengkulu	Lampung
Aceh	0	0.178	0.182	0.143	0.143	0.178	0	0.178
North Sumatera	0.203	0	0.281	0.156	0.156	0	0.203	0
West Sumatera	0.111	0.149	0	0.166	0.166	0.149	0.111	0.149
Riau	0.148	0.141	0.282	0	0	0.141	0.148	0.141
Jambi	0.148	0.141	0.282	0	0	0.141	0.148	0.141
South Sumatera	0.203	0	0.281	0.156	0.156	0	0.203	0
Bengkulu	0	0.178	0.182	0.143	0.143	0.178	0	0.178
Lampung	0.203	0	0.281	0.156	0.156	0	0.203	0

Figure 5. Weights Matrix of Cluster spatial

3.4 Seasonal Generalized Space Time Autoregressive Moving Average (Seasonal GSTARIMA)

3.4.1 Parameter Estimation

The GSTARIMA model (2,0,0;1) $(0,0,1;1)_6$ for one location is written as follows:

$$\mathbf{z_i}(t\text{-}1) = \emptyset_{10}^{(i)}\, \mathbf{z_i}(t\text{-}1) + \emptyset_{11}^{(i)} \sum_{j=1}^{8} \mathbf{W_{ij}}\, \mathbf{z_j}(t\text{-}1) + \emptyset_{20}^{(i)}\, \mathbf{z_i}(t\text{-}2) + \emptyset_{21}^{(i)} \sum_{j=1}^{8} \mathbf{W_{ij}}\, \mathbf{z_j}(t\text{-}2) - \theta_{60}^{(i)}\, \mathbf{e_i}(t\text{-}6) + \theta_{61}^{(i)} \sum_{j=1}^{8} \mathbf{W_{ij}}\, \mathbf{e_j}(t\text{-}6)$$

The parameter estimation using Ordinary Least Square (OLS) method obtained the parameter value for each weights as follows:

Table 3 Estimation of GSTARIMA Parameters (2,0,0;1) $(0,0,1)_6$

Province (i)	Parameter	Weights			
		LR	P-value	CS	P-value
1. Aceh	$\emptyset_{10}$	0.580	**<.0001**	0.482	**0.0003**
	$\emptyset_{11}$	-0.079	0.6667	0.011	0.2763
	$\emptyset_{20}$	-0.154	0.1620	-0.103	0.3823
	$\emptyset_{21}$	0.453	**0.0280**	0.016	0.1687
	Θ_{60}	0.444	**0.0005**	0.569	**<.0001**
	Θ_{61}	-0.240	0.1708	-0.024	**0.0263**
2. Nort Sematra	$\emptyset_{10}$	1.022	**<.0001**	0.948	**<.0001**
	$\emptyset_{11}$	0.097	0.4359	0.008	0.4435
	$\emptyset_{20}$	-0.139	0.3274	-0.030	0.8436
	$\emptyset_{21}$	0.036	0.7549	-0.002	0.8326
	Θ_{60}	0.250	0.0601	0.309	**0.0203**
	Θ_{61}	0.087	0.5209	0.003	0.8146

Table 3 Estimation of GSTARIMA Parameters (2,0,0;1) $(0,0,1)_6$ (*continued…*)

Province (i)	Parameter	Weights			
		LR	P-value	CS	P-value
3. West Sumatera	$\emptyset_{10}$	0.235	**0.0562**	0.352	**0.0035**
	$\emptyset_{11}$	0.285	0.2544	0.002	0.8597
	$\emptyset_{20}$	0.163	0.2057	0.246	0.0573
	$\emptyset_{21}$	0.582	**0.0150**	0.021	**0.0226**
	Θ_{60}	0.176	0.2280	0.128	0.3960
	Θ_{61}	-0.581	0.0356	-0.030	**0.0187**
4. Riau	$\emptyset_{10}$	0.474	**0.0008**	0.565	**<.0001**
	$\emptyset_{11}$	0.680	**0.0106**	0.015	0.3725
	$\emptyset_{20}$	-0.201	0.1197	-0.232	0.0880
	$\emptyset_{21}$	-0.147	0.5553	0.020	0.2062
	Θ_{60}	0.244	0.0946	0.213	0.1482
	Θ_{61}	0.309	0.3387	0.022	0.3003
5. Jambi	$\emptyset_{10}$	0.428	**0.0030**	0.526	**0.0006**
	$\emptyset_{11}$	0.467	**0.0053**	0.037	**0.0034**
	$\emptyset_{20}$	-0.005	0.9735	-0.034	0.8256
	$\emptyset_{21}$	0.099	0.5845	-0.003	0.7625
	Θ_{60}	0.170	0.2270	0.025	0.8628
	Θ_{61}	-0.389	**0.0205**	0.028	0.0562
6. South Sumatera	$\emptyset_{10}$	0.855	**<.0001**	0.973	**<.0001**
	$\emptyset_{11}$	0.624	**0.0039**	0.015	0.2834
	$\emptyset_{20}$	-0.492	**0.0012**	-0.732	**<.0001**
	$\emptyset_{21}$	-0.014	0.9429	0.030	**0.0256**
	Θ_{60}	0.025	0.8670	0.023	0.8726
	Θ_{61}	0.558	**0.0150**	0.006	0.6958
7. Bengkulu	$\emptyset_{10}$	0.419	**0.0025**	0.337	**0.0209**
	$\emptyset_{11}$	0.650	**<.0001**	0.052	**<.0001**
	$\emptyset_{20}$	-0.058	0.6744	0.002	0.9856
	$\emptyset_{21}$	-0.076	0.6230	-0.016	0.1752
	Θ_{60}	-0.318	**0.0223**	-0.346	**0.0148**
	Θ_{61}	0.731	**0.0002**	0.031	**0.0143**
8. Lampung	$\emptyset_{10}$	0.598	**0.0002**	0.677	**<.0001**
	$\emptyset_{11}$	0.394	**0.0274**	0.001	0.9756
	$\emptyset_{20}$	-0.239	0.1011	-0.404	**0.0020**
	$\emptyset_{21}$	0.178	0.3708	0.039	**0.0171**
	Θ_{60}	0.343	**0.0170**	0.301	**0.0305**
	Θ_{61}	-0.224	0.2121	-0.009	0.5727

If p value > α (0.05) then parameter was not significant while if p value < α (0.05) then parameter was significant.

3.4.2 Forecasting

Forecasting the same GSTARIMA model will be done as much as 11 periods ahead then will be calculated the value of accuracy using the testing data.

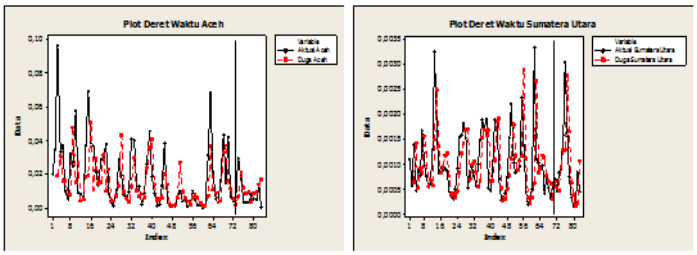

Figure 6 Plots of actual values and estimation values of provinces of Aceh and North Sumatera.

Figure 6 showed the pattern of forecast values following the actual value pattern so it could be concluded that the model used was good enough.

3.4.2 Accuration Comparison of Forecasting

The accuracy of forecasting was seen based on the RMSE value to see the magnitude of forecast deviation and MAPE to see how much forecasting errors.

Table 4 RMSE and MAPE values of GSTARIMA model

Provinsi	RMSE		MAPE	
	LR Weights	CS Weights	LR Weights	CS Weights
Aceh	0.484	0.493	17.146	20.665
North Sumatera	0.313	0.508	9.059	9.326
West Sumatera	0.492	0.283	13.103	11.673
Riau	0.502	0.563	15.030	17.040
Jambi	0.406	0.511	12.310	16.035
South Sumatera	0.613	0.539	27.051	17.580
Bengkulu	0.484	0.447	12.325	12.213
Lampung	0.381	0.349	14.494	14.651
Pulau Sumatera	0.467	**0.450**	15.065	**14.898**

Table 4 showed the RMSE and MAPE values for the GSTARIMA model with the CS weights was the smallest.

4 Conclusion

Results of GSTARIMA modeling of area extent data of rice stem borer attack on 8 provinces in Sumatera Island obtained the best model for modeling and forecasting area of rice stem borer attack was model of GSTARIMA by CS weights with RMSE value obtained was 0.450 and MAPE value was 14.898%.

References

[1] Asikin, S. dan M. Thamrin. 2001. Bionomi Penggerek Batang Padi dan Alternatif Pengendaliannya. *Dalam* Hama dan Penyakit Utama Padi di Lahan Pasang Surut. Departemen Pertanian Badan Penelitian dan Pengembangan Pertanian.

[2] Bezdek J. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York: Plenum Press.

[3] Borovkova, S. A., H. P. Lopuhaä and B. Nurani. 2002. Generalized STAR models with experimental weights. *Proceedings of the 17th International Workshop on Statistical Modelling 2002*, Chania. Greece.

[4] Pfeifer PE, Deutch SJ. 1980. A Three Stage Iterative Procedure for Space-Time Modeling. *Technometrics*. 22(1) : 35-47.

[5] Ruchjana, B. N., Abdullah A. S., Toharuddin, T., and Mindra Jaya I. G. N. 2013. Clustering Spatial on the GSTAR Model for Replacement New Oil Well. *AIP Conference Proceedings*. 1554 205.

[6] Wijaya, Ferdian Bangkit. 2015. Pendekatan Space Time Autoregressive (STAR) dan Generalized Space Time Autoregressive (GSTAR) melalui Metode Autoregressive dan Vector Autoregressive [Tesis]. Bogor (ID) : Institut Pertanian Bogor.