Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 5 (2015), pp. 2877-2888

http://www.ripublication.com

Weighted Ensemble Hybrid In Spatial Autocorrelation Model's For Predicting The HDI In Java

Sitti Masyitah Meliyana R*[1], Hari Wijayanto[*2], Farit Mochamad Afendi[*3]

** Departement of Statistcs, Bogor Agricultural University16680, IPB Dramaga, Bogor, Indonesia*
[1] marek.girls@gmail.com
[2] hari_ipb@yahoo.com
[3] fmafendi@ipb.ac.id

ABSTRACT

The data modeling is often found an observation at a location having a relationship or influence with other nearby locations. One of the causes is the existance of spatial autocorrelation in the data. Problem of spatial autocorrelation consists of two they areexistance ofobservation dependence between locations (spatial autoregressive) that can be overcome with SAR and existance of error dependence inter-observation (error spatial autocorrelation) that can be overcome by SEM. However, in practice these two problems sometimes occur in one observation, and so we need a technique to combine both of the information. Technique of ensemble present as a solution to combine one or several models and provide a stronger prediction accuracy. The methodology will be applied to predict the value of theHuman Depelopment Index (HDI) in Java.Best estimation is done by selecting the smallest RMSEA of several estimation methods. The best estimation method is techniquen of spatial autocorrelation ensemblewith weighted regression (Wreg-EAS) with RMSEA of 1.817.

Keywords: EAS,Ensemble, Human Depelopment Index, SAR, SEM, Weighted Ensemble

1 INTODUCTION

Regression analysis is a statistical method used to describe the relationship between the response variable and the predictor variables. The resulting model is known as regression models that help researchers determine the causal relationship between

two or more variables. Problems arise when occured the offense assumptions relating to correlated residual problems and heterogeneityproblems in error. For example, because the observations at a location have a strong influence in other nearby locations. The condition known as a spatial effects, which can be divided into two parts, namely spatial autocorrelation and spatial diversity (Anselin [1] in 1988). When ignoring the information of the existance of spatial effects on the data, then the observations will result a formed model that is not feasible (Lesage [8] in 1997). Models that can overcome spatial autocorrelation namely Spatial Autoregressive Model (SAR), Spatial Error Model (SEM) and General Spatial Model (GSM).

SAR is a model that contains information for observation dependence between locations (autoregression). SEM is a model that contains information for error autocorrelation between locations. McMillen [9] in 1992 explains that the SAR and SEM methods usedmore appropriate in models that have a spatial autocorrelation. When autoregression and error autocorrelation information between locations is contained in one model, then this model is called the GSM model. Therefore GSM known as merging SAR and SEM. Philip [10] in 2010 states that GSM is not widely used in practice due to the absence of guidelines or theory when used the same weighted matrix (W = W1 = W2) which resulted in the identification problem. So the researchers suggest the Ensemble techniques to combine autoregression and autocorrelation error information, which will be called the Ensemble-hybrid Spatial Autocorrelation (EAS).

Ensemble present as a technique that can combine one or several models and provide a stronger prediction accuracy. In principle ensemble technique is to combine the prediction results of many models and then make predictions from the best model selected. One of the cases that are affected by the proximity of the area is the Human Development Index (HDI). Some of the factors that affect the level of IPM is the level of income, education and health. Where these factors are also influenced by the proximity of the region. In the case of HDI, an area near to the big cities tend to have a high HDI value, otherwise remote regions with large cities tend to have a lower HDI value. This is due to an interaction between the two regions spatial linkages. So that the data in this study using the data HDI BPS in 2012 in 117 districts / cities in Java.

2 RESEARCHMETHODS

Data

The data used is a secondary data obtained from the publication of the Central Bureau of Statistics Indonesia in 2012 and the data potential of the village 2011. Overall the data used covers 117 districts in Java. The parameters used in this study are as follows:

1. Response variable (Y) isHuman Development Index
2. Predictor variables (X) are:

a. The number of state universities (X1)
b. Percentage of Skills Institute Population Permille (X2)
c. Percentage of Health FacilityPopulation Permille (X3)

d. Percentage ofModern MarketPopulation Permille (X4)
e. Percentage ofCooperativePopulationPermille (X5)
f. Percentage of Resto Population Permille (X6)
g. Percentage of Hotel Population Permille (X7)
h. Percentage of Traditional MarketPopulation Permille (X8)
i. Number of School Participationage 19-24 Years (X9)

Analysis Method

1. Conductdata exploration
2. ConductClassical Regression Analysis

a. The explanatory variables used are explanatory variables that have a value of Variance Inflation Factor (VIF) < 10 andsignificance level$\propto = 5\%$.
b. Arrange Classical Regression Model$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ andmeet the assumptions of the classical regression.

3. CheckingSpatial Effect

a. Createspatial weight matrix (**W**) with queen contiguity method.
b. Create*Moran Scatter Plot*whichfunctionas avisual exploratory analysistodetectspatial effectonthe data and checks accompanied by Moran index valueobtained by$I = \frac{N\sum_i\sum_j \boldsymbol{w}_{i,j}(x_i-\bar{x})(x_j-\bar{x})}{(\sum_i\sum_j \boldsymbol{w}_{i,j})\sum_i(x_i-\bar{x})^2}$
c. Conduct LM Test

- LM test forcheckingspatial autoregressive is $LM_{lag} = \left[\varepsilon^T W y/\left((\varepsilon^T\varepsilon)/n\right)\right]^2/D$, if $LM_{lag} > \chi^2_{(1)}$ or P-value <αthen spatial autoregressive exist.
- LM test for checking error spatial autocorrelation is $LM_{err} = \frac{\left[\varepsilon^T W\varepsilon/\left((\varepsilon^T\varepsilon)/n\right)\right]^2}{tr\,(W^2+W^TW)}$, if $LM_{err} > \chi^2_{(1)}$ or P-value <αthenerror spatial autocorrelation exist.

4. Conductspatial regression analysis

a. Arrange SARmodelwithequation$\boldsymbol{y} = \rho\boldsymbol{W_1 y} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$where$\rho$is parameter ofspatial lag coefficient ofresponse variable.
b. Arrange SEM model withequaion$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \boldsymbol{u} = \lambda\boldsymbol{W u} + \boldsymbol{\varepsilon}$where$\lambda$is parameter of spatial lag error coefficient.
c. The explanatory variablesusedarethe explonatory varablesthat signicance level$\propto = 5\%$.

5. Conductpredictionswith ensemble technique

a. Add white noise on data.
b. Analyze data thathas been given white noise with SAR or SEM method
c. Repeatstep 1-2 as many as N times, butwith different white noise ineach iteration.
d. Calculatepredictionof Ensemble SAR (ESAR) andprediction of Ensemble SEM (ESEM) by using*simple averaging* as follow, $\hat{y}_i = \frac{1}{N}\sum_{j=1}^{N}\hat{y}_{ij}$; $i = 1, 2, \ldots, 117$ (many of observation)and j =many ofiteration (N)

6. Conductpredictionwithtechnique of Ensemble Autocorrelation Spatial (EAS)

a. Arrange theensemble membeship fromcombinationof ESAR and ESEM prediction.
b. Predict EAS byusing simple averaging that is$\hat{y}_{i\,EAS} = \frac{1}{2}(\hat{y}_{i\,SAR} + \hat{y}_{i\,SEM})$where$i = 1, 2, \dots, 117$.
7. ConductpredictionwithEAS techniquewithproportional weighted (Wpro-EAS)
a. Arrangeensemble membershipfromcombination ESAR and ESEM prediction bygivingweights proportionally.
b. Determinevalue ofweight b1 fromaverageof ratio y and$\hat{y}_{ESAR}$, andvalue ofweight b2 fomaverage of y and$\hat{y}_{ESEM}$.
c. CalculateestimationofWpro-EAS withformula $\hat{y}_{i\,Wpro-EAS} = \frac{1}{2}[(\hat{y}_{i\,ESAR} + b1\hat{y}_{i\,ESAR}) + (\hat{y}_{i\,ESEM} + b2\hat{y}_{i\,ESEM})]$where$i = 1,2, \dots, 117$
8. ConductpredictwithEAS techniqueby weightingregression (Wreg-EAS)
a. Arrangethe ensemble membershipfromcombination ESAR and ESEM predictionbyweightingin regression.
b. Regreting y, $\hat{y}_{ESAR}$, and$\hat{y}_{ESEM}$toget b0 from intercept, b1 fromcoefficient$\hat{y}_{ESAR}$and b2 from$\hat{y}_{ESEM}$.
c. Calculatethe estimationfrom Wreg-EAS withformulation $\hat{y}_{i\,Wreg-EAS} = b0 + b1\hat{y}_{i\,ESAR} + b2\hat{y}_{i\,ESEM}$.
9. ConductpredictionwithEAS techniquewithweightingcorrelation (Wcorr-EAS)
a. Arrange ensemble membershipfromcombination ESAR and ESEM predictionby weighting with correlation.
b. Determinevalue ofweight b1 fromcorrelation y and$\hat{y}_{ESAR}$, andvalue ofweight b2 fromcorrelation y and$\hat{y}_{ESEM}$.
c. CalculateestimationfromWpro-EAS withformulation $\hat{y}_{i\,Wcorr-EAS} = \frac{1}{2}[(\hat{y}_{i\,ESAR} + b1\hat{y}_{i\,ESAR}) + (\hat{y}_{i\,ESEM} + b2\hat{y}_{i\,ESEM})]$wherre$i = 1,2, \dots, 117$
10. Comparevalue of RMSEA amongclassical regression, SAR, SEM, E-SAR, E-SEM, EAS, Wreg-EAS, Wpro-EAS and Wcorr-EAS byusing statistic of *root mean square error aproctimation* (RMSEA), smallest valuemeansthe predictionis better.$RMSEA = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-1}}$

3 RESULT DAN DISCUSSION

Exploration of Data

Exploration of data useful for studying the characteristics of the data to make it easier to determine the appropriate statistical analysis model (or improvement of statistical analysis that has been planned).

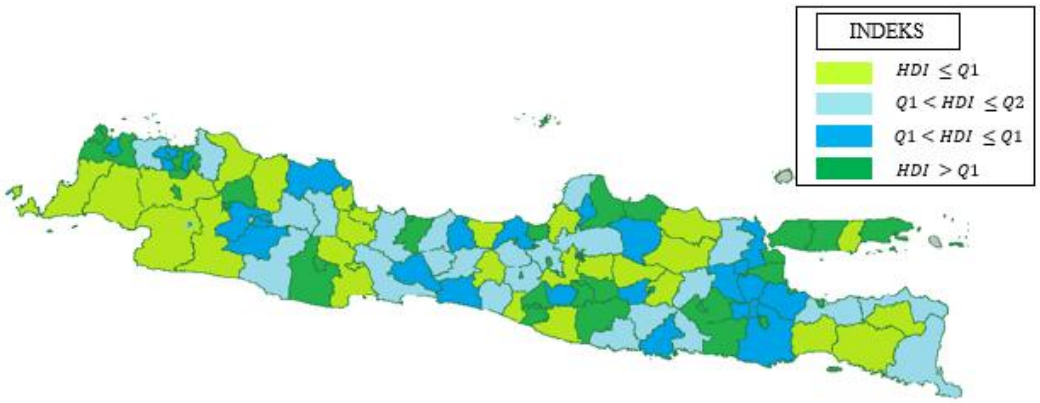


Figure 1 Mapping of HDI based ongroup value ofits quantile.

Data of IPM in Java in 2012 mapped based on the value of its quantile that seenexistance grouping of data distribution. It is worthy of suspicion of spatial effects in the data, it is seen areas with a high HDI value nearto areas that have a high HDI value anyway. Where the value of Q1, Q2, and Q3 are respectively 63.79, 64.16, and 65.04.

Classical Reggression Analysis

Modeling using classical regression analysis resulted in four real variables at the level of α = 5%. The variable is the number of universities (X1), the percentage of health facilities (X3), the percentage of modern market (X4) and the percentage of traditional markets (X8). To check multicolinearity can be seen from its VIF value. VIF value that less than 10 means that there is multicolinearity. VIF value presented by Table 2 indicates the absence of problems in its multicolinearity. So it can proceed with the classical regression analysis using the least squares method (MKT). Parameter estimation using the least squares method (MKT) are presented in Table 1.

Table 1 Parameter estimationof MKT regression

Variable	Coefficient	Standard Error	P-Value	VIF
Coefficient	68.336	0.749	0.000	***
X1	0.045	0.018	0.015	1.792*
X3	0.069	0.023	0.004	2.731**
X4	0.087	0.039	0.027	1.310*
X8	-0.258	0.128	0.046	1.232*

*Note:***) significancelevel 0.001, **) significancelevel 0.01, *) sinificance level 0.05*

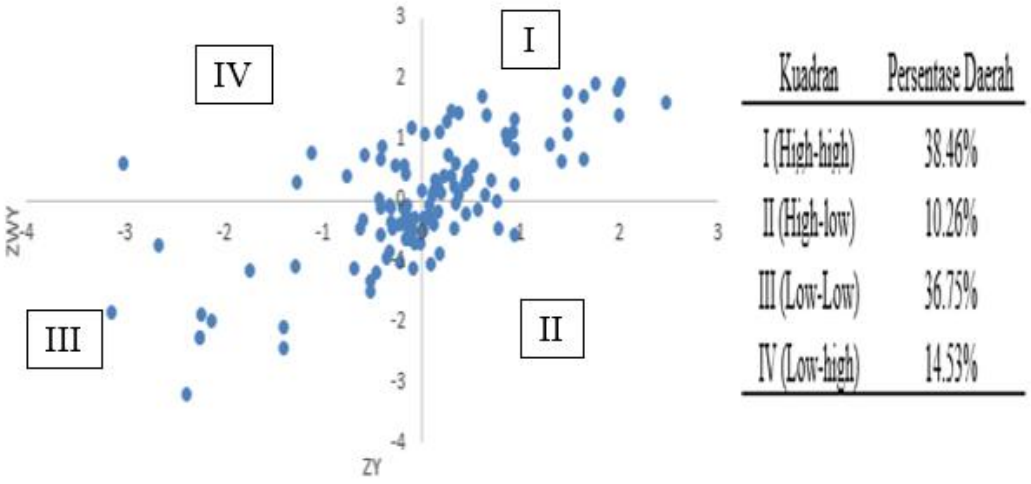
Classical regression equations formed using the least squares method (MKT) are as follows:

$$HDI = 68.336 + 0.045X1 + 0.069X3 + 0.087X4 - 0.258X8$$

Classical regression equation formed has a value of R-Square 55.1% which means the classic regression model can explain the diversity of the human development index by 55.1%. While the rest can be explained by other variables outside the model. Furthermore, test to check the existence of spatial autocorrelation using the Moran index.

Checking Spatial Effect

Moran index test conducted by making the weighting matrix appopriate with queen contiguity concept. Moran index test is required to check the presence or absence of spatial autocorrelation in the data to avoid mistakes that led to the model prediction is not feasible.



Kuadran	Persentase Daerah
I (High-high)	38.46%
II (High-low)	10.26%
III (Low-Low)	36.75%
IV (Low-high)	14.53%

Figure2Moran Scatter Plot

Based on the results of Moran scatter plot above shows the plot spreads in several quadrants. Quadrant I is located in the upper right called High-High quadrant, means that it has a positive autocorrelation because location observation value is high and surrounded by an area that is high too. Quadrant II is located right under called High-Low, means that ithas a negative autocorrelation because location observation value is highand surrounded by an area that has a low value. Quadrant III is located in the lower left called Low-Low quadrant, means thatit has a positive autocorrelation, because location observation value is low and surrounded by an area that is low too. Quadrant IV is located in the upper left called Low-High quadrant, it means thatit has a negative autocorrelation, because location observation value is low and surrounded

by a high area. Resultof Moran index is $I = 0.457$with a p-value = $1.654e^{-15}$ ($< \alpha = 5\%$). This concludes the existance of spatial autocorrelation in OLS residual.

Table 2 Lagrange Multiplier Test

Model	Parameter	Nilai-P
SAR	4.97	0.03*
SEM	55.47	9.50E-14***

*Note: ***) significance level 0.001, **) sinificance level 0.01, *) significance level 0.05*

Based on LM test showed spatial dependence on observation and its error. So these two models SAR and SEM are used to estimate the value of the HDI.

Analysis of Spasial Auto-Regressive (SAR) and Spatial Error Model (SEM)
SAR is a process to solve the spatial autocorrelation in its spatial lag. SEM is a process to solve the spatial autocorrelation in its spatial error.Parameter significance test of SAR and SEM partially on the z test statistic can be seen in Table 3briefly.

Table 3Parameter Estimation SAR

SAR		
Variable	Coeffisient	P-value
Intercept	65.632	0.000***
X1	0.066	0.000***
X3	0.122	0.000***
X4	0.109	0.002**
Rho	0.004	0.01**

*Note: ***) significance level 0.001, **) significance level 0.01, *) significance level 0.05*

Table 4Parameter Estimation SEM

SEM		
Variable	Coeffisient	P-value
Intercept	68.396	0.000***
X1	0.037	0.003**
X3	0.102	0.000***
X4	0.067	0.016*
Lambda	0.145	0.000***

*Note: ***) significance level 0.001, **) significance level 0.01, *) significance level 0.05*

Table 3 and 4 shows that the coefficient rho (ρ) and lambda (λ) significance with p-values of <0.05 (α), means that there are influence of spatial lag from a nearby location. Similarly, The number of state universities variable (X1), percentage of health facility variable (X3), and percentage of modern market variable (X4) statistically significant, means that the variables give a significant effect on the big changes in the Human Development Index in Java.

The equation obtained from the SAR analysis is as follows

$$Y = -81.54 - 0.0008Wy + 0.12X1 + 0.33X2 + 0.64X3$$

The equation obtained from the result of SAR analysis is as follows:

$$Y = -72.23 + 0.11X1 + 0.35X2 + 0.03X3 + u$$

$$u = -0.15Wu + \varepsilon$$

These results are still not enough to obtain the best prediction model, it will be arranged a model Ensemble SAR in further discussion.

Addition of Noise

On the addition of noise in the IPM (Y) data will provide the variety scale with a similar frame. As a result of the addition of different noise in each iteration will produce a noise result or dirty result. But by calculating average ofall the ensemble result it will give effect "clean"each other.

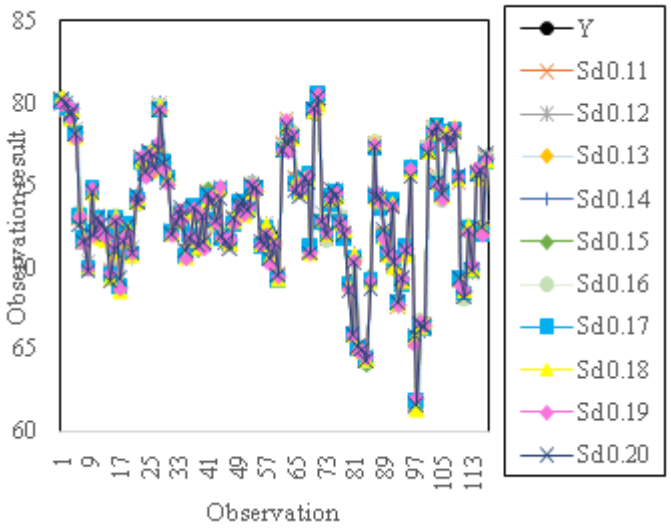


Figure 3 Data Plot Y andData Plot Y+*noise*

Zhang et al. [14] 2008 conducted experiments of noise with standard deviation between 0.1 or 0.2 as many as 100 trials. The value of σ was tested is 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19 and 0.20 resulting frame that the data is similar to the data on variable Y. Figure 6 shows that with the addition of noise with ceain value ofσresult similar fluctuations with a data on variable Y. Thus, for the addition of noise $(\varepsilon \sim N(0,\sigma))$ in the ensemble process to form the ensemblemembershipwill be used the value of σ above.

Ensemble Prediction

Estimation results are used to predict the value of HDI is the best estimation. Best estimation is done by selecting the smallest RMSEA of several estimation methods which have been doneabove. Based on the table above the best estimation method is the technique of spatial autocorrelation ensemble weighted regression (Wreg-EAS) with RMSEA 1.817335 as shown in Table 5 below:

Table5 RMSEA eachensemble estimation mehod

Prediction Method	Weight	RMSEA
SAR	-	2.482583
SEM	-	1.978192
ESAR	-	2.482516
ESEM	-	1.975315
EAS	-	2.189715
Wpro-EAS	b1 = -0.001	2.191464
	b2 = -0.0006	
Wreg-EAS	b0 = -10.727	1.817335
	b1 = -0.635	
	b2 = 1.782	
Wcorr-EAS	b1 = 0.269	15.53866
	b2 = 0.150	

Ensemble is a technique in predicting by combining several models that resulted from one method or several methods. This technique does not take one of the best models of a number of models generated from an analysis and do not do estimation fromthe best models. Estimation is done by combining the results of estimation of the various existing models. There are two main steps to make an ensemble. The first step is to make ensemblemembership and the second step is to determine the right combination of results from ensemble members to yield a single ensemble. Based on Table above for weighted ensemble technique seen much difference between the value of RMSEA for Wcorr-EAS and the value of RMSEA Wreg-EAS. This is because the weighting of Wreg-EAS involves the intercept between the response variable and the predicted results of the SAR and SEM while in Wcor-EAS using only the correlation between the predicted SAR and SEM with responsevariable. Such that Wreg-EAS is used to

combine the results of the predictions of the model SAR and SEM. Where regression models were formed in the human development index that uses the SEM models are:

$$Y = 0.145Wu + 0.036X1 + 0.102X3 + 0.067X4$$

From the SEM model that formedit is obtained R^2 of 70.44%, which means that the model formed can explain the diversity of HDIvariables by 70.44%, while 29.56% is explained by other variables outside the model. Models are formed to produce variables at the significance level of α = 0:05 namely number of universities variable (X1), percentage of health facilities variable (X3), and percentage of the modern market variable (X4). Significance of λcoefficients indicate that if an area surrounded by other regions of n, then the influence of each area surrounding them can be measured at 0.145 multiplied by the average error surroundings. Coefficient of the number of universities variable is 0.036 shows that any increase in the number of universities in Java by one percent then increase the Indonesian human development index for 0.036 points, with assumtion that other variables held constant, so too Inversely. It shows that to improve the human development index by one point, then each district should increase the number of university by 1 / 0.036 = 27.78%, with assumtionthat other variables held constant.

On health indicators in Java can be explained by the percentage of health facilities variable that provide a positive effect to the HDI. Any increase in the percentage of health facilities by one percent, the increase IPM in Java os 0.102 points, with assumption that other variables held constant, and vice versa. It shows that to increase the HDI by one point, then each district in Java should increase the percentage of health facilities by 9.804%, with assumptiont that other variables held constant.

Economic indicators in Java can be explained by a percentage of the modern market variablethat provide a positive effect on the HDI. This variable is oneof a positive indicator from the condition of the welfare of society. Therefore, if the value of these variable becoming more increased then the value of the HDI is also increased. Based on the obtained spatial models can be defined that, any increase in the percentage of modern market by one percent then the increase of HDI in Java is 0.067 points.

Prediction results of SEM above ensembled with the prediction results of SAR model that only have $R^2$is 53.44%. Similarly for regression model formed by using SAR models, namely:

$$Y = 0.004Wy + 0.066X1 + 0.122X3 + 0.109X4$$

by coefficients of ρ is significanceit indicates that if an area surrounded by other areaas many as n, then the influence of each area that surround them can be measured at 0.004 multiplied by the average of HDI variables in the surrounding area. Where the explanatory variables significance level at α = 0.05 namely number of universities variable (X1), percentage of health facilities variable (X3), and percentage of the modern market variable (X4).

WregEAS ensemble techniques are then used to combine the results of estimating the two models above and obtained an increase in R^2 become 75.05% with a RMSEA is

1.817. This is one of Ensemble Hybrid study that can still be developed for data in other studies.

CONCLUSION

The data in this study have satisfied the classical regression assumptions. From the spatial autocorrelation model can be stated that the SEM predicts better than the SAR. It indicatesonthe data of HDI,there are explanatory variables that are not included in a linear regression model such that counted as an error and the variables are correlated with the error in the other locations. In order to autoregressive information and error autocorrelation combined in one model then do not done the selection of the best model. Later models were formed, were combined to obtain a better prediction results. Then the spatial autocorrelation models developed using Ensemble hybrid and nonhybrid obtained that Hybrid method by using the regression concept to the formation of the weight (Wreg-EAS) which produce better predictions than other methods.

BIBLIOGRAPHY

[1] Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Academic Publishers.

[2] [BPS] Badan Pusat Statistik. 2008. *Indeks Pembangunan Manusia (IPM).* Publikasi IPM Badan Pusat Statistik: Jakarta-Indonesia.

[3] De Bock, K.W., Coussement, K., Van den Poel, D. 2010. *Ensemble Classification based on General Additive Models*. Computational Statistics & Data Analysis 54(6): 1535-1546)

[4] Dubin R. 2009. *Spatial Weight*. Fotheringham AS, PA Rogerson, editor, Handbook of Spatial Analysis. London: Sage Publication.

[5] Friedman, J.H. & Popescu, B.E. 2008. Predictive learning vi rule ensemble. *The Annals of Applied Statistics* 2(3): 916-954.

[6] Fransiska, H. 2014. Metode Dekomposisi Ensemble untuk Memprediksi Harga Beras DKI Jakarta [tesis]. Bogor (ID): Institut Pertanian Bogor.

[7] Griffith, D. 2000. A linear regression solution to the spatial autocorrelation problem.*Journal of Geographical System* 2, 141-156.

[8] LeSage, James P. 1997. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20, nos 1 dan 2, pp 113-129.

[9] McMillen, Daniel P. 1992. Probit with spatial autocorrelation. *Journal of Regional Science*, 32, No.3, pp, 335-348.

[10] Philip A. Viton. 2010. *Notes on Spatial Econometrics Model.* National Oceanic and Atmospheric Administration, Washington, DC.

[11] Rohmawati, N. 2015. Aplikasi Analisis Regresi Spasial Ensemble Pada Data Kemiskinan Di Pulau Jawa [tesis]. Bogor (ID): Institut Pertanian Bogor.

[12] Walpole, R.E. and Myers, R.H. 1995. *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan*. Terjemahan Bambang Sumantri, Edisi ke-4. Bandung: Penerbit ITB.

[13] Zaier, I., Shu, C., Quarda, T.B.M.J., Seidou, O. and Chebana, F. 2010. Estimation of ice thickness on lakes using artificial neural network ensembles. *Journal of Hydrology*, (383), pp. 330-340.

[14] Zhang, X., Lai, KK., and Wang SY.2008. A New Approach for Crude Oil Price Analysis on Empirical Mode Decomposition. *Energy Economics**30***, 905-918.

[15] Zhou, Z, H. 2012. Ensemble Methods Foundation and Algorithms. Camridge (UK) : CRC Press.

[16] Zhu, M. 2008. Kernel and ensembles: Perspectives on statistical learning. *The American Statistician* 62 (2): 97-101.