International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Spatial Autoregressive Regression Modeling with Heteroskedasticity Using Bayesian Approach on GRDP of Java

Fitri Ramadhini[a]*, Anik Djuraidah[b], Aji Hamim Wigena[c]

[a,b,c]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Dramaga Campus, Bogor 16680, Indonesia*

[a]*Email: ramadhini_fitri@apps.ipb.ac.id*

[b]*Email: anikdjuraidah@apps.ipb.ac.id*

[c]*Email: ajiwigena@ymail.com*

Abstract

Area data is aggregation data according to the area and location information. Modeling the area data needs to concern to the dependency and heteroskedasticity between areas. Heteroskedasticity occurs because units in the area generally differ in size and characteristics. The spatial autoregressive (SAR) regression models only consider dependence on the response variable. Most of SAR estimators are valid if there is no violation in the error assumption. In the condition of heteroskedasticity, the SAR parameter estimator with the maximum likelihood (ML) method becomes invalid. An alternative method that can be used is Bayesian method, that solves the problem of heteroskedasticity by modeling the structure of variance-covariance matrix. In this study, the Bayesian method was applied to Java's GRDP in 2017. This data contains spatial dependence and heteroskedasticity so the ML method is not suitable to be applied. The explanatory variables were used in this study are number of workers, regional revenue, regional minimum wage, and human development index. The result shows that number of workers, regional revenue, and regional minimum wage are statistically significant affecting Java's GRDP in 2017. This model provides a pseudo R^2 value of 74%, which means it is good enough to illustrate the diversity of Java's GRDP in 2017.

Keywords: Bayesian; Spatial autoregressive; Spatial dependence; Heteroskedasticity; GRDP.

---

* Corresponding author.

1. Introduction

Indonesia is a developing country that development progress is one of the important things in realizing progress in Indonesia. In the implementation of development, if a country's economic growth is good, then it can be said that its development is also in good condition. Economic growth of a country can not be separated from economic growth in each region. Economic growth in the regional area is closely related to the increase in goods and services produced by a region. This increase can be seen from the value of gross regional domestic product (GRDP). GRDP is one of the important indicators to determine the economic conditions in a region for a certain period, both at current prices and at constant prices.

GRDP data is area data obtained from aggregation of several observed values in each region. This data usually influenced by a variety of measurement problems, namely spatial dependence and spatial heterogeneity [1] called spatial effects. Classical regression modeling by ignoring the spatial effects of this data becomes inappropriate. In the spatial regression model, spatial dependence can be described in autoregressive on response, error, or combination of the two [2,3]. The models with dependence on the response are called the spatial autoregressive models (SAR) while in spatial heterogeneity can be modeled with geographically weighted regression (GWR) [4].

In general, spatial regression considers only one of the spatial effects. In its application, it is often found the problem of heteroskedasticity that comes from the process of averaging data with many different observations at the time of aggregation [5,6]. The estimation method of spatial regression parameters with maximum likelihood (ML) [7] depends on the assumption that the errors are normal distribution and homogeneous. This causes the ML estimator for spatial data with heteroskedasticity problem to be invalid. Several ways can be used to overcome this problem, including modification of the estimator of the generalized method of moments (GMM) [8] and Bayesian approach [9,1]. The Bayesian method solves the problem of heteroskedasticity in spatial model by modeling the structure of variance-covariance matrix that allows flexible models for each distribution of spatial data with high accuracy [10].

The data in this study is Java's GRDP in 2017 from the Indonesian Central Bureau of Statistics. Java Island is still the central of development in Indonesia with a wide variety of GRDP distribution so that the problem of heteroskedasticity is very likely to occur. Several studies on GRDP have been carried out using GWR and SAR model [11, 12]. This study aims to model the Java's GRDP in 2017 including spatial effects. The model is SAR with heteroskedasticity with Bayesian approach.

2. Materials and methods

2.1. SAR model with Bayesian approach

The SAR model can be expressed as follows [1]:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,;\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \qquad (1)$$

$\boldsymbol{y}$ is an n×1 vector of response variables, $\boldsymbol{X}$ is an n×p matrix of explanatory variable, $\boldsymbol{\beta}$ is a p×1 vector of

coefficient regression parameter, $\boldsymbol{W}$ is an n×n spatial weight matrix , $\boldsymbol{\varepsilon}$ is an n×1 vector of no autocorrelation error, ρ is a spatial lag coefficient, and $\boldsymbol{I}$ is an identity matrix.

Estimation using Bayesian method assumes that

$\boldsymbol{\Sigma} = \sigma_0^2\, \boldsymbol{V}$, with $\boldsymbol{V} = diag(v_1, v_2, \dots, v_n)$, and $v_i = \frac{\sigma_i^2}{\sigma_0^2}$ for $i = 1, \dots, n$

This assumption indicates that the heteroskedasticity specification has two components: (i) the constant component σ_0^2, and (ii) a component v_i that varies between observations [13]. SAR model in equation (1) with $\boldsymbol{\varepsilon}$ has a heteroskedastic pattern can be written as:

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X\beta} + \boldsymbol{\varepsilon}^* \qquad (2)$$

with $\boldsymbol{\varepsilon}^* = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}$, and , $\boldsymbol{\varepsilon}^* \sim MVN(\boldsymbol{0}, \sigma_0^2\boldsymbol{V}((\boldsymbol{I} - \rho\boldsymbol{W}')^{-1}(\boldsymbol{I} - \rho\boldsymbol{W})^{-1})$. Let $\boldsymbol{v} = (v_1, \dots, v_n)'$ is an n×1 vector and the likelihood function of this model can be written as:

$$L(\boldsymbol{y}|\boldsymbol{\theta}, \sigma_0^2, \boldsymbol{v}) = (2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}} \prod_{i=1}^{n} v_i^{-\frac{1}{2}} |\boldsymbol{S}(\rho)|$$

$$\times exp\left[-\frac{1}{2}(\boldsymbol{S}(\rho)\boldsymbol{y} - \boldsymbol{X\beta})'(\sigma_0^2\boldsymbol{V})^{-1}(\boldsymbol{S}(\rho)\boldsymbol{y} - \boldsymbol{X\beta})\right] \qquad (3)$$

with $\boldsymbol{S}(\rho) = (\boldsymbol{I} - \rho\boldsymbol{W})$ and $\boldsymbol{\theta}$ is hyperparameter for ρ and $\boldsymbol{\beta}$. The posterior function for the Bayesian method with prior distribution assumptions can be stated as [14]:

$$p(\boldsymbol{\theta}, \sigma_0^2, \boldsymbol{v}, r|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\theta}, \sigma_0^2, \boldsymbol{v})\, \pi(\boldsymbol{\beta})\, \pi(\sigma_0^2)\, \pi(\boldsymbol{v})\, \pi(r)\, \pi(\rho)$$

$$\propto (2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}} \prod_{i=1}^{n} v_i^{-\frac{1}{2}} |\boldsymbol{S}(\rho)|$$

$$\times exp\left[-\frac{1}{2}(\boldsymbol{S}(\rho)\boldsymbol{y} - \boldsymbol{X\beta})'(\sigma_0^2\boldsymbol{V})^{-1}(\boldsymbol{S}(\rho)\boldsymbol{y} - \boldsymbol{X\beta})\right]$$

$$\times |\boldsymbol{T}|^{-\frac{1}{2}} exp(-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{T}^{-1}\boldsymbol{\beta}))$$

$$\times (\sigma^2)^{-(a+1)} exp\left(\frac{-b}{\sigma^2}\right)$$

$$\times \left(\frac{r}{2}\right)^{\frac{nr}{2}} \left[\Gamma\left(\frac{r}{2}\right)\right]^{-n} \prod_{i=1}^{n} v_i^{-\frac{r+2}{2}} exp\left\{-\frac{r}{2v_i}\right\}$$

$$\times\ r^{m-1}\, exp\left\{-\frac{r}{k}\right\} \times p(\rho) \qquad (4)$$

2.2. Data

The data is in Table 1. It's from the Indonesian Central Bureau of Statistics. GRDP as a response variable is the total value of final goods and services produced by all economic units in a region. The explanatory variables are the number of workers, regional revenue, regional minimum wage, and human development index. The scope of this study includes all regencies / cities in Java in 2017.

Table 1: The data in the study

Variable	Unit
GRDP at the constant price of the base year 2010	Trillion Rupiahs
Number of workers	Thousand Peoples
Regional revenue	Million Rupiahs
Regional minimum wage	Million Rupiahs
Human development index	Percent

2.3. Data analysis procedure

The steps of analysis are as follows:

1. Data exploration

Data exploration is used to see data distribution and correlation between regions by summarizing data description and GRDP distribution map.

2. Making a spatial weight matrix $\boldsymbol{W}$

The spatial weight matrix used is k-nearest neighbor weight matrix with $k = 2$. This matrix sorts the central distance (d) in each spatial unit (i) with all spatial units $j \neq i$ on n observations,

$$d_{ij(1)} \leq d_{ij(2)} \leq \cdots \leq d_{ij(n-1)}$$

For each size of the closest distance $k = 1, \dots, n-1$, the closest set of spatial units $N_k(i) = \{j(1), j(2), \dots, j(k)\}$ contains the closest k to i unit so that the k-nearest neighbor weight matrix, $\boldsymbol{W}$, has spatial weights with normalized elements as follows:

$$w_{ij} = \begin{cases} 1/k, & j \in N_k(i) \\ 0, & \text{others} \end{cases} \quad (5)$$

3. Testing the spatial effects

Spatial effect tests include spatial dependence and spatial heterogeneity or heteroskedasticity problem. The spatial heteroskedasticity was tested using the Breusch Pagan (BP) test, while the spatial dependence was tested using the Moran's index [1].

4. Identifying the spatial model

The Lagrange multiplier (LM) test was conducted to test the effect of spatial dependence on responses and errors. LM tests include spatial dependence on response (SAR), error (spatial error model / SEM), and a combination of both (generalized spatial model / GSM) [1,14].

5. Estimating parameters with Bayesian method, which is defining the distribution of priors for each hyperparameter then finding the posterior distribution using Markov chain Monte Carlo (MCMC). The priors in the SAR model assumed [14]:

- β from normal distribution $\beta \sim N(c, T)$, with T is a large value of variance.
- $\sigma_0^2 \sim IG(a, b)$ where $IG(a, b)$ is the inverse gamma distribution with shape $a > 0$ and scale $b > 0$
- $\frac{r}{v_i} \sim \chi^2(r)$, for $i = 1, \ldots, n$, with degree of freedom r is from Gamma distribution, $r \sim G(m, k)$.
- $\rho \sim U(\frac{1}{\lambda_{min}}, \frac{1}{\lambda_{max}})$ with $U(.)$ is uniform distribution where λ_{min} and λ_{max} are minimum and maximum eigen values of $\boldsymbol{W}$.

3. Results

3.1. Data exploration

Java is the 13th largest island in the world with an administrative area of 138,793.6 km² with a population density of around 1317 peoples/km². Java consists of 6 provinces, namely Banten, the Special Capital Region of Jakarta, West Java, Central Java, Special Region of Yogyakarta, and East Java with total of 119 districts / cities. The average of Java's GRDP in each province is presented in Figure 1.

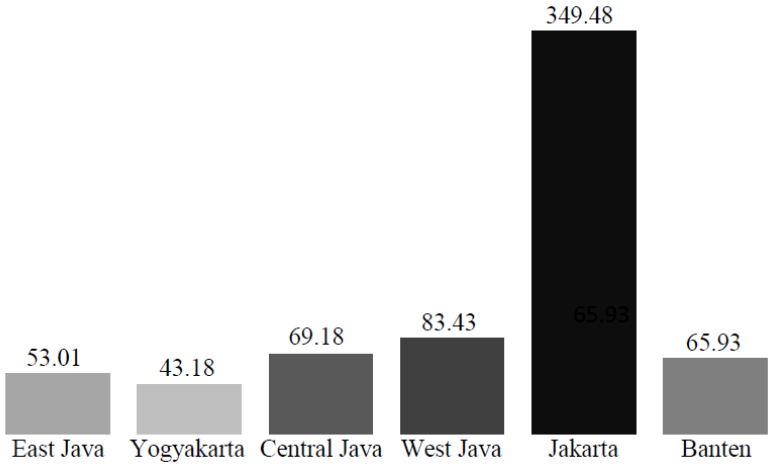

Figure 1: Average of Java's GRDP in 2017 in each province (trirllion Rupiah)

Figure 1 shows that Jakarta Province has the highest average of GRDP in Java which indicates a high level of economic growth. Meanwhile, the lowest average of GRDP occurred in Yogyakarta Province. Table 2 shows that district / city with the lowest GRDP value of 3.909 trillion rupiahs is Banjar City, while district / city with the highest GRDP value of 590.650 trillion rupiahs is West Jakarta. The average value of Java's GRDP is 63.300 trillion rupiahs with standard deviation of 108.050 trillion rupiahs. These show the occurrence of inequity in GRDP values in Java.

Table 2: Data description for Java's GRDP in 2017

Statistics	GRDP (Trillion Rupiah)
Mean	63.300
Minimum	3.909 (Banjar City)
Maximum	590.650 (West Jakarta)
Standard deviation	108.050

The amount of GRDP produced in an area cannot be separated from the geographical conditions, population, and social conditions and welfare of the people that can differ between regencies / cities. These differences can cause the GRDP values to vary. Regions with a central government and economy tend to produce high GRDP values such as Jakarta and Surabaya. Figure 2 shows the distribution of GRDP values in districts / cities in Java. The distribution of GRDP tends to be similar to the surrounding regencies / cities, indicating a spatial relationship between regencies / cities. This can be seen from the thick colored area (high GRDP) around the area that has high or moderate GRDP. Whereas more faded areas (low GRDP) tend to be around areas that have the same color. Regencies / cities with high GRDP values (> 200 trillion rupiahs) are Jakarta, Surabaya, Bekasi, Bandung and Sukabumi. Regencies / cities that have low GRDP values (<10 trillion rupiahs) are Probolinggo, Mojokerto, Blitar, Banjar City, and several other regencies / cities.

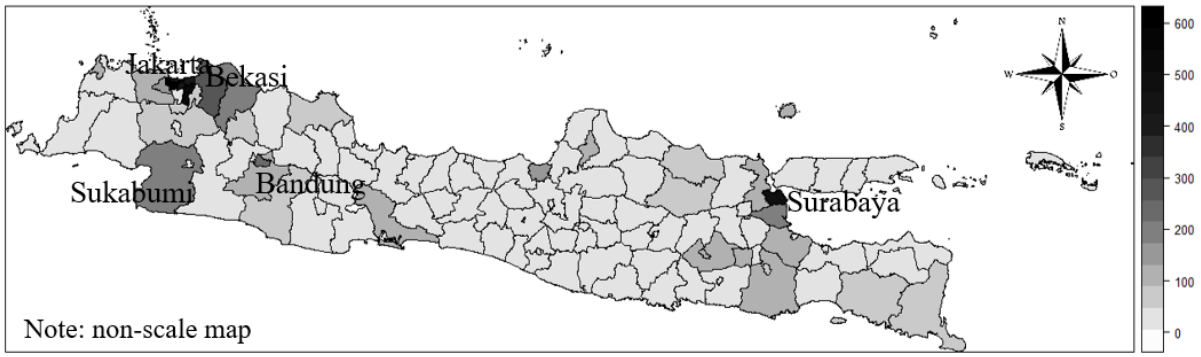

Figure 2: Map of Java's GRDP in 2017 distribution

3.2. Spatial effect tests

The p value of BP test is 3.184×10^{-7}. This shows that there is a problem with heteroskedasticity in the data. The Moran's index is 0.395 with p value 2.525×10^{-7} indicates that there is a positive spatial autocorrelation ($I > 0$) between GRDP in each regency / city in Java. This means that Java's GRDP in 2017 tends to be similar to the surrounding area.

3.3. Lagrange multiplier test

Table 3 shows the results of the spatial dependence test with the LM test. The p value of LM test for spatial dependence on response (SAR) is 0.010 indicates there is spatial dependence in the response at 5% significant level. In addition, the p value of LM test for spatial dependence in the error (SEM) and in the response and error (GSM) results are less than 0.05 so that it can also be modeled with SEM or GSM.

Table 3: LM test for spatial dependence

Model	LM Test Statistics	P Value
LM lag (SAR)	6.678	0.010*
LM error (SEM)	3.960	0.047*
SARMA (GSM)	6.749	0.034*

*significant at α=5%

3.4. Parameter estimation using Bayesian method on SAR model

The parameter estimation $\boldsymbol{\rho}$ gives a positive value of 0.017. This indicates that there is a positive spatial dependence or there are similarities in adjacent regencies / cities. The results of parameter estimation $\boldsymbol{\beta}$ in SAR model using Bayesian method are presented in Table 4.

Table 4: Predicted parameters from SAR model with Bayesian method

Variable	Parameter	Coefficient
(Intercept)	c	-15.037
Number of workers	β_1	0.026*
Regional minimum wage	β_2	7.326*
Regional revenue	β_3	0.073*
Human development index	β_4	-0.043
$\hat{\rho}$		0.017
Pseudo-R^2		0.7404

*significant at α=5%

Based on Table 4, the explanatory variables that influence GRDP in Java 2017 are number of workers, regional minimum wage, and regional revenue. This model with Bayesian approach gives pseudo-R^2 value of 74.04%. It means, this model explain GRDP diversity in Java by 74.04% and the remaining 25.96% is explained by other factors outside the model.

4. Conclusion

Bayesian method in SAR modeling can be used to analyze data containing heteroskedasticity. Java's GRDP in 2017 contains spatial dependence and spatial heterogeneity that makes the ML method not suitable to be applied. The application of Bayesian method for GRDP provides three significant variables that influence Java's GRDP in 2017 with positive effect. Those variables are number of workers, regional revenue, and regional minimum wage. This model provides a pseudo R^2 value of 74%, which means it is good enough to illustrate the diversity of Java's GRDP in 2017.

5. Recommendations

Socio-economic data such as GRDP tend to have two spatial effects which require analysis that takes into account these effects. The other methods can be used for modeling GRDP and the results can be compared with the Bayesian method. In addition, the other spatial weight matrices can be used, for example exponential distance or contiguity weight matrix.

Acknowledgments

This work is fully supported by Kemenristek DIKTI (Kementerian Riset Teknologi dan Pendidikan Tinggi) of Indonesia.

References

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

[2] J. P. LeSage. Spatial Econometrics. Toledo: Department of Economics University of Toledo, 1998.

[3] L. Anselin. Spatial Econometrics. Dallas: University of Texas, 1999.

[4] A. S. Fotheringham, C. Brunsdon, and M. Charlton. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: John Wiley & Sons, 2002.

[5] W. E. Griffiths. A Companion to Theoretical Econometrics. New Jersey: Blackwell Publishing, 2007.

[6] L. Lee and X. Liu. "Efficient Gmm Estimation of High Order Spatial Autoregressive Models with Autoregressive Disturbances." Econom. Theory, vol. 26, no. 1, pp. 187–230, 2010.

[7] K. Ord. “Estimation Methods for Models of Spatial Interaction.” J. Am. Stat. Assoc., vol. 70, no. 349, pp. 120–126, 1975.

[8] H. H. Kelejian and I. R. Prucha. “Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances.” J. Econom., vol. 157, no. 1, pp. 53–67, 2010.

[9] L. Anselin. “A Note on Small Sample Properties of Estimators in a First-Order Spatial Autoregressive Model.” Environ. Plan. A Econ. Sp., vol. 14, no. 8, pp. 1023–1030, 1982.

[10] G. Koop. Bayesian Econometrics. Chichester: John Wiley & Sons, 2003.

[11] I. Yulita, A. Djuraidah, and A. H. Wigena. “Geographically Weighted Regression Include the Data Containing Multicollinearity.” Indones. J. Stat., vol. 20, no. 2, pp. 5–8, 2015.

[12] N. Hikmah. "Produk Domestik Regional Bruto Kabupaten / Kota Jawa Barat Dengan Spasial Data Panel." B.S. thesis, IPB University, Bogor, 2012.

[13] O. Doğan and S. Taşpinar. “Spatial Autoregressive Models with Unknown Heteroskedasticity: A Comparison of Bayesian and Robust GMM Approach.” Reg. Sci. Urban Econ., vol. 45, no. 1, pp. 1–21, 2014.

[14] J. P. LeSage and R. K. Pace. Introduction to Spatial Econometrics. New York: Taylor & Francis Group, 2009.