Logit-Normal Spatial Model for Small Area Estimation: Case Study of Poverty in Bengkulu

E Sunandi*

Department of Mathematics, The University of Bengkulu, Bengkulu, 38125, Indonesia
E-mail: esunandi@unib.ac.id

Abstract. The main objective of this research is to develop a Logit-Normal model with spatial effects. Spatial effects are characterized by a spatial weighted matrix. The weighted matrix used is the distance of the amount of auxiliary variables of area. It was used to estimate the proportion/parameters using Hierarchical Bayesian method. The design of the research is a case study. Data used in the case study is poverty data in Bengkulu Province. The results show that the spatial effect can improve the value of the Root Mean Square Error.

Introduction
Small Area Estimation (SAE) is an indirect estimation of parameter in a relatively small area in the survey sampling. This estimation is applied when the direct estimation is not able to provide sufficient accuracy when the sample size in the small area, so generated statistic will have a large variance or the predictions cannot be made because they are not represented in the survey [1]. The estimation of parameters of SAE is based on the model of a small area that requires additional information that has a relationship with the variables being observed, also known as auxiliary variables. The auxiliary variable can be obtained from other surveys and is expected to correlate with the observed variables [2].

Models in SAE assume that the random effects of errors area are independent. However, in some cases, this assumption is often violated. The reason is the diversity in an area influenced by the surrounding area so that the spatial effects can be incorporated into a random effect. Spatial effects are common place between one area to another area, this means that one area affects other areas. In statistics, a spatial model can explain the relationship between an area with the surrounding area.

Spatial analysis has been developed by several researchers. In 2010, [3] in his dissertation discussed the model of small area estimate used is a unit-based Normal model introduced by Battase-Harter-Fuller (1988). Mercer et al. [4] examined the properties of spatial smoothing model with a various weighted matrix. Other studies were conducted by Matualage et al. [5] using Spatial Empirical Best Linear Unbiased Prediction (SEBLUP) method to estimate outcome per capita in Jember East Java. The result is a SEBLUP method better than direct predictor methods and EBLUP. In 2015, [6] conducted a logarithmic transformation in the Spatial EBLUP method.

Contrast to previous studies, this study applies the Logit-Normal model to add spatial effects and response variable Binomial distribution. Spatial effects are characterized by spatial-weighted matrix. The weighting matrix used is the nearest neighbor. The estimation of parameters used in this model is a Hierarchical Bayes (HB) Method which was specially developed for the data of binary response variable by adding a spatial effect. Case Study of this study is poverty data in Bengkulu.
Materials and Methods
The case study was performed on poverty data in Bengkulu Province. The data were sourced from data Susenas and Podes 2016. The model used in this study is Hierarchical Bayes Logit-Normal with weighted spatial correlation. The data used is the proportion of poor families ($Y$). The auxiliary variables are assumed to affect and to describe poverty are Number of Families Without Electricity ($X_1$), the number of Education Facilities ($X_2$), the number of health Facilities ($X_3$), Number of Recipients JAMKESNAS ($X_4$), and the number of SKTM ($X_5$).

Conditional Autoregressive (CAR) Spatial Model
According to [7], CAR is defined as a spatial model that is normally distributed. Random effect of CAR model $\nu_i$ depends on the influence of other areas. Joint distribution of $\nu = (\nu_1, ..., \nu_m)$ is obtained by $\nu \sim N_m(0, (I - \rho Q)^{-1}M)$ with $\nu$ is a vector of random effects area. $Q$ is a spatial weighting matrix showing the relationship with the i-th area and l-th area, and M is diagonal matrix area of variance random effects i-th area $= diag(\sigma^2_{\nu,1}, ..., \sigma^2_{\nu,m})$. Matrix $(I - \rho Q)^{-1}$ is a symmetric and definite positive matrix. Meanwhile, the spatial correlation ($\rho$) is obtained by $\rho = \frac{1}{\lambda_1}$; $-1 \leq \rho \leq 1$. According to [8], $\lambda_1$ is the root of the biggest Eigen value of spatial weighting matrix Q. If $\rho = 0$, then the random effects of model CAR assumed are independent. According to [9], several types of formation of a spatial weighted matrix can be used. The study used the nearest neighbor weighted matrix. This matrix has the following rules:

$$Q_{il} = \begin{cases} 1, & r_{il} \geq 0.5 \\ 0, & r_{il} < 0.5 \end{cases}, \quad i = l = 1, 2, ..., m$$

with $r_{il}$ correlation between auxiliary variables i-th area and l-th area.

Logit-Normal Model
According to [8] defined a HB version of the Logit-Normal as follows:

$$y_i | p_i \sim Binomial(n_i, p_i)$$

$$\theta = logit(p) = X\beta + \nu, \nu \sim N_m(0, I\sigma^2_{\nu})$$

$\beta$ and $\sigma^2_{\nu}$ are independent

$$f(\beta) \propto 1$$

$$\frac{1}{\sigma^2_{\nu}} \sim Gamma(a, b) \quad ; a > 0, b > 0$$

The estimation of parameters using HB cannot be resolved analytically. To make it easier, it takes a numerical approach. MCMC model could be the solution. The famous MCMC procedure is Gibb conditionals. According to [8], Gibb conditionals corresponding to the HB model are given by

i. $[\beta | p, \sigma^2_{\nu}, y] \sim N_p(\beta^*, \sigma^2_{\nu}(\Sigma_{i=1}^m x_i x'_i)^{-1})$

ii. $[\sigma^2_{\nu} | \beta, p, y] \sim Gamma(\frac{m}{2} + a, \frac{1}{2} \Sigma_{i=1}^m (\theta_i - x'_i \beta)^2 + b)$

iii. $f(p_i | \beta, \sigma^2_{\nu}, y) \propto h(p_i | \beta, \sigma^2_{\nu}) k(p_i)$

Estimation of parameters $\beta$ and $\sigma^2_{\nu}$ raised directly from (i) and (ii). Parameter $\beta^*$ in part (i) of equation (1) is expressed by $\beta^* = (\Sigma_{i=1}^m x'_i x_i)^{-1} (\Sigma_{i=1}^m x'_i \theta_i)$. Meanwhile, part (iii) of equation (1) is expressed as

a) $f(p_i | \beta, \sigma^2_{\nu}, y) \propto h(p_i | \beta, \sigma^2_{\nu}) k(p_i)$

(2)
b) $h(p_i | \beta, \sigma^2_v) = \frac{\partial \theta_i}{\partial p_i} \exp \left\{ -\frac{(\theta_i - x_i^v \beta)^2}{2\sigma^2_v} \right\}$

c) $k(p_i) = p_i^{y_i} (1 - p_i)^{n-y_i}$

Values of Hierarchical Bayes Proportions $(p_{HB}^{(i)})$ are determined by simulation Gibbs Sampling Metropolis-Hastings (MH). Gibb samples can be directly generated from (c) in equation (2). The MH algorithm is as follows:

1. Taken $p_i^*$ from a uniform distribution $(0, 1)$.
2. Generated $\theta_j \sim ind N(X^v \beta, 1\sigma^2_v)$, then searched for value $p_i^{(0)} = g^{-1}(\theta_i)$
3. Calculated $r(p_i^{(k)}, p_i^*) = \min \left\{ \frac{k(p_i^*)}{k(p_i^{(k)})}, 1 \right\}; k = 0, 1, ..., D$
4. Generated $u$ from a uniform distribution $(0,1)$.
5. Selected $p_i^{(k+1)} = p_i^*$ if $u \leq r(p_i^{(k)}, p_i^*)$.
6. Repeat step 3 until D samples are obtained.

After the MH simulation, the following proportional estimates are obtained $\{p_1^{(k)}, ..., p_m^{(k)}; k = 1, ..., D\}$. Then, the posterior being observed can be calculated. The estimator of HB proportion is

$$(p_{HB}^{(i)}) \approx \frac{1}{D} \sum_{k=1}^{D} p_i^{(k)} = p_i^{(1)}$$

(3)

While the proportion of variance of estimators hierarchical Bayes ($\hat{p}_{HB}^{(i)}$) is

$$V(p_i^{(i)}) = \frac{1}{D - 1} \sum_{k=1}^{D} (p_i^{(k)} - p_i^{(1)})^2$$

(4)

On the other hand, characteristics of $(p_{HB}^{(i)})$ are MSE and bias. MSE is a scale to measure the variance of small area estimators. Meanwhile, the bias is the difference between the expectation of the estimator and the parameter. Smaller the MSE and bias, then the parameters estimator will be more valid and more accurate. Mathematically written as follows:

$$Bias(p_i^{(i)}) = \frac{1}{D} \sum_{q=1}^{D} \left[ p_i^{(q)} - p_i^{(i)} \right]$$

$$MSE(p_i^{(i)}) = \frac{1}{D} \sum_{q=1}^{D} \left[ p_i^{(q)} - p_i^{(BH)} \right]^2$$

A measurement of accuracy and validation in the Hierarchical Bayes proportion estimator is Root Mean Square Error (RMSE). RMSE is defined as root difference average amount quadratic actual proportions and estimator. Thus, estimating the most accurate will lead to the smallest RMSE value to zero. RMSE is given by

$$RMSE = \sqrt{MSE}$$
Results and Discussion

Spatial Logit-Normal Model

The fundamental difference between the Logit-Normal model and the Spatial Logit-Normal model is in the matrix variance-covariance of area effect. The spatial Logit-Normal is as follows:

\[ y_i \sim \text{Binomial}(n_i, p_i) \]
\[ \mathbf{\theta} = \text{logit}(p) = X'\beta + \nu \]
\[ \nu \sim N_m(0, (I - \rho Q)^{-1}M) \]
\[ \beta \text{ and } \sigma_\nu^2 \text{ are independent} \]
\[ f(\beta) \propto 1 \]
\[ \frac{1}{\sigma_\nu^2} \sim \text{gamma}(a, b) ; a \geq 0, b > 0 \]

Similar to the Logit-Normal Model, the Spatial Logit-Normal model is estimated through MCMC sampling generation. The famous MCMC procedure is conditional Gibbs. According to [10], the model is

1. \[ [\beta | p, \sigma_\nu^2, y] \sim N_p(\beta^*, \sigma_\nu^2(\sum_{i=1}^m x_i x'_i)^{-1}) \]
2. \[ [\sigma_\nu^2 | p, \beta, y] \sim \text{Gamma} \left[ \frac{m}{2}, \frac{1}{2} \left[ (\mathbf{\Theta} - X'\beta)'(\mathbf{\Theta} - X'\beta) \right] + b \right] \]
3. \[ f(p_i | \beta, \sigma_\nu^2, y) \propto h(p_i | \beta, \sigma_\nu^2)k(p_i) \]

Estimation of parameters \( \beta \) and \( \sigma_\nu^2 \) raised directly from (i) and (ii). Parameter \( \beta^* \) in part (i) of equation (5) is expressed by \( \beta^* = (\sum_{i=1}^m x_i x'_i)^{-1}(\sum_{i=1}^m x_i \theta_i) \). Meanwhile, part (iii) of equation (5) is expressed as

\[ a) \quad f(p_i | \beta, \sigma_\nu^2, y) \propto h(p_i | \beta, \sigma_\nu^2)k(p_i) \]
\[ b) \quad h(p_i | \beta, \sigma_\nu^2) = \frac{\partial^2}{\partial p_i^2} \exp \left\{ -\frac{1}{2d_{ii}} \left[ \theta_i - x'_i \beta \right]^2 \right\} \]
\[ c) \quad k(p_i) = p_i^{y_i}(1 - p_i)^{n - y_i} \]

Where \( d_{ii} \) is diagonal matrix of \((I - \rho Q)^{-1}M)\). The value of Hierarchic Bayes proportions will be estimated through a Gibbs Sampling Metropolis-Hasting (M-H) simulation. The MCMC sample gibbs can be generated directly from (c) in equation (6). The M-H algorithm is as follows:

1. Take \( p_i^0 \) from a uniform distribution \((0, 1)\).
2. Generate \( \mathbf{\theta} \sim \text{ind } N(X'\beta, ((I - \rho Q)^{-1}M)) \), then searched for value \( p_i^{(0)} = g^{-1}(\theta_i) \)
3. Calculated \( r(p_i^{(k)}, p_i^*) = \min \left\{ \frac{k(p_i^*)}{k(p_i^{(k)})}, 1 \right\} \); \( k = 0, 1, ..., D \)
4. Generate \( u \) from a uniform distribution \((0,1)\).
5. Selected \( p_i^{(k+1)} = p_i^* \) if \( u \leq r(p_i^{(k)}, p_i^*) \).
6. Repeat step 3 until D samples are obtained.

Further estimator of the HB proportion of i-th area and its variance can be calculated using equations (3) and (4).

Poverty Modeling in Bengkulu using Spatial Logit-Normal Model

This study applied spatial Logit-Normal model into poverty data in Bengkulu. Gibbs samples are generated 500 samples and are replicated 100 times. The results are obtained from these estimations of the HB estimator have trend (tendency) which is equal to the direct proportion (Figure 1). It means that both methods indicate the estimation produced consistent
estimators. The maximum value of proportion of poor families per district can be seen also in Figure 1

**Figure 1.** Comparison of Direct Estimators and HB Estimator

The results of this HB Estimator also show that 5.88% (13 villages) of the sample villages had a proportion of poverty greater than 50%. Padang Nangka Village Singaran patih Subdistrict Bengkulu town has the highest proportion of poverty (67%). Meanwhile, Gunung Alam Village North Bengkulu Subdistrict has the lowest proportion of poverty (7%).

In addition, HB method also estimated the variance and MSE of proportion estimators. In terms of RMSE of HB estimators per sample villages (Figure 1), majority villages have the variance of direct estimator value greater than the variance of HB estimators. However, in general, the estimator of the HB variance is smaller than the direct estimator. So, it can be concluded that the estimation of proportion using Spatial Logit-Normal model HB better than direct estimation.

**Conclusion**

From the results of the estimation can be concluded that the proportion prediction through Hierarchical Bayes (HB) is better than direct estimator. This is proved by most of the RMSE of the HB estimation is smaller than of the RMSE Direct estimator.

The results of this HB Estimator also show that 5.88% (13 villages) of the sample villages had a proportion of poverty higher than 50%. Padang Nangka Village Kec. Singaran Patih Bengkulu town has the highest proportion of poverty (67%). Meanwhile, Gunung Alam Village North Bengkulu has the lowest proportion of poverty (7%).
References

[1] Prasad NGN and Rao JNK 1990 The Estimation of The Mean Squared Error of The Small Area Estimators *Journal of American Statistical Association*, 85, p 163-171.

[2] Rumiati, A.T 2012 Model Bayes Untuk Pendugaan Area Kecil dengan Penaksiran Contoh Berpeluang Tidak Sama pada Kasus Respon Binomial dan Multinomial (Doctoral dissertation Institut Pertanian Bogor)

[3] Martin V 2010 Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011 (thesis University of Trier)

[4] Mercer L, Wakefield, Chen C, Lumley T 2014 A Comparison of Spatial Smoothing Methods for Small Area Estimation with Sampling Weights. *J Spat Stat.* doi:10.1016/j.spasta.2013.12.001 p 69–85.

[5] Matualage D, Saefuddin A, Wigena AH 2011 Pendekatan Small Area Estimation Untuk Menduga Pengeluaran Per Kapita Rumah Tangga Tiap Desa Dengan Empirical Best Linear Unbiased Prediction (Studi kasus: Kabupaten Jember Provinsi Jawa Timur). *Proceedings Seminar Nasional Statistika Universitas Diponegoro*, ISBN: 978-979-097-142-4, p 655-668.

[6] Zainuddi HA, Notodiputro KA, Sadik K 2015 A Study of Logarithmic Transformation Model in Spatial Empirical Best Linear Unbiased Prediction (SEBLUP) Method of Small Area Estimation, *Proceedings of "The 7Th SEAMS-UGM Conference 2015"*. Topic, p 67-74.

[7] Banerjee, BP Carlin, AE. Gelfan 2004 Hierarchical Modeling and Analysis for Spatial Data (Boca Raton: Chapman & Hall / CRC Press Company)

[8] Rao, JNK and Molina, I 2015 Small Area Estimation. 2nd Edition (New York: Wiley) [https://doi.org/10.1002/9781118735855](https://doi.org/10.1002/9781118735855)

[9] Fotheringham AS, Rogerson PA 2009 *The SAGE Handbook of Spatial Analysis* (Los Angeles: SAGE)

[10] Etis S, Khairil AN, Anik D 2011 Model Spasial Bayes dalam Pendugaan Area Kecil dengan Peubah Respon Biner (kasus: pendugaan proporsi keluarga miskin Di Kabupaten Jember Jawa Timur) (Thesis Institut Pertanian Bogor)