Spatial hierarchical Bayes estimation of mean years of schooling

Dwi A S Wahyuni[1]**, Sutarman Wage**[2] **and Open Darnius**[3]

Department of Mathematics, Universitas Sumatera Utara, Medan 20155 Indonesia

E-mail: [1]dwiasihseptiwahyuni@gmail.com; [2]sutarman@usu.ac.id; [3]open@usu.ac.id

Abstract. A spatial hierarchical bayes for estimating mean years of schooling district level is proposed. We developed spatial hierarchical bayes within a Monte Carlo simulation study with R software. The simulation generated posterior distribution invers gamma. The spatial correlation used rook contiguity for each district. Hierarchical bayes method with spatial weighted provides smaller relative bias and relative root mean square.

1. Introduction

Mean years of schooling is one of the key indicators of social performance. Statistics Indonesia annually publishes mean years of schooling to regency level. Regency level divided into some small area. The small area such as district level, village level, etc. Local government are needed mean years of schooling in small area level for regional planning and fund allocation. Mean years of schooling collected by national socio economic survey. The sample size of national socio economic survey is not sufficient to provide mean years of schooling to smaller level. The bayes method is applied in small area estimations by borrowing information from other small areas that have similar characteristics or values in the past and the values of variables that are related to the variables being observed [5].

The application of bayes methods in small area parameters consists of four kinds methods of Best Linear Unbiased Prediction (BLUP), Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB), and Hierarchical Bayes (HB). Hierarchical bayes method can solve very complex models for the data based on very simple models as building blocks [3]. HB methods also can give solution for complex small area model by markov chain monte carlo methods [4]. Model parameters in hierarchical bayes are treated as random variables and assigned a prior distribution [2]. The use of sampling weights in bayesian hierarchical models for small area estimation is a review of Chen et al [1] presenting bayesian spatial smoothing models that result in less Mean Square Error (MSE) values than the general bayesian method approach without spatial weights.

This study aims to obtain the estimation of the mean years of schooling parameter in district level in Tapanuli Tengah Regency using spatial hierarchical bayes method. The hierarchical bayes method has several advantages, such as: can be applied to estimate general indicators defined as functions of the response variable model, hierarchical bayes estimators are unbiased and optimal models when minimizing posterior variance, the bootstrap method for MSE estimation is not required so that the total timing less than empirical bayes methods [5]. The estimation parameter of mean years of schooling district level is done by utilizing National socio economic survey 2015 and Village potential 2014 data as the auxiliary variables.

2. Spatial Hierarchical Bayes Model

Let y_i be the value of the modeling estimation in the i-th region where i is the number of small areas in-sample $(i = 1, ..., m)$. Auxiliary variables $x_{ij} = (x_{i1}, ..., x_{in})^T$ available for a small area $i = 1, ..., m$ and variable $j = 1, ..., n$. Sampling model is the first stage in hierarchical bayes written as follows:

$$y_i|\theta_i, \sigma_v^2\, N_p(\theta_i, \psi_i)$$

where ψ_i is a sampling variance that is assumed to be equal for all areas to form a model $\theta_i = \alpha + x_i\beta + e_i + v_i$ where θ_i is the actual parameter value of the target variable in the i-th area, v_i describes the spatial correlation of random effects [5].

Secondstage model or population model spatial hierarchical bayes is:

$$\theta_i|\beta, \sigma_v^2 \;\sim N(x_i^T\beta, \sigma_v^2 D^{-1})$$

where $D = \lambda R + (1-\lambda)\,I$ and R is a spatial weighted matrix with members r_{ii} is the number of neighboring area in the i-th and $r_{il} = 1$ if l is neighboring area i and $r_{il} = 0$ if others , $i \neq l$. Value of λ explained spatial autocorrelation parameters whose value ranges from 0 to 1. The spatial weighted matrix (R) used in this study uses a rook contiguity type defined as starting 1 for the adjacent areas in the north, south, west and east or called the side of the common side while 0 (zero) for others. This method will give a value of 1 if the i-region coincides with the j-region and 0 (zero) if the i-region does not coincide with j-region.

Both stages of spatial hierarchical bayes are then explained into the conditional distribution of Gibbs Sampling to be processed through Markov Chain Monte Carlo (MCMC). The established markov chain is $\eta = (\mu^T, \lambda^T)$ where $\mu = (\theta_1, \ldots, \theta_m)^T$. θ is small area parameter and $\lambda = \left(\beta^T, \sigma_v^2\right)^T$ is model parameters. Partition Gibbs Sampling are:

(i) $[\theta_i|\beta, \lambda, \sigma_v^2, y_i] \sim MVN[\Lambda y, (1-\Lambda)X\beta, \Lambda E]$ dimana $\Lambda = (E^{-1} + D/\sigma_V^2)^{-1}E^{-1}$ dengan $E = \{\widetilde{\sigma}_1^2, ..., \widetilde{\sigma}_1^2\}$ dan $X = (x_1, ..., x_m)^T$

(ii) $[\beta|\theta, \lambda, \sigma_v^2, y_i] \sim MVN[(X'DX)^{-1}X'D\theta, \sigma_v^2(X'DX)^{-1}]$

(iii) $[\lambda|\theta, \beta, \sigma_v^2] \sim |[\lambda R + (1-\lambda)I]^{-1}|^{\frac{1}{2}} x \exp\{-\frac{1}{2\sigma_v^2}(\theta - X\beta)'[\lambda R + (1-\lambda)I](\theta - X\beta)\}$

(iv) $[\sigma_v^2|\theta, \beta, \lambda] \sim IG[\frac{m}{2} + a, \frac{1}{2}[(\theta - X\beta)'D(\theta - X\beta) + b]$

Based on the form of Gibbs Sampling, parameter estimation θ,β,σ_v^2 can be generated directly from (i), (ii), and (iv) via the Gibbs Sampling algorithm because these three parameters have a clear, normal and Inverse Gamma (IG) distribution.

3. Simulated Models

Methods based on simulated models using sample data proposed for small area estimation [4]. Data analysis was performed to obtain the best method of spatial hierarchical bayes through simulation. Simulation is done by R software. Simulation is generating data from the posterior distribution $\sigma_v^2 \sim IG(a, b)$ where averages and variance are assumed to be equal $(a = b)$ raised with values $0.0001; 0.001; 0.01; 0.1; 1$ on hierarchical bayes and spatial hierarchical bayes models. Simulation which gives the Relative Bias (RB) and the smallest Relative Root Mean Square Error $(RRMSE)$ values used to simulate the sampling variance. Here is the formula for calculating RB and $RRMSE$. $RB_i = \frac{1}{K}\frac{\sum_{k=1}^{K}(\widehat{\overline{Y}}_i^{(k)} - \overline{Y}_i)}{\overline{Y}_i}$ and $RRMSE_i = \frac{\sqrt{\frac{1}{k}\sum_{k=1}^{K}(\widehat{\overline{Y}}_i^{(k)} - \overline{Y}_i)^2}}{\overline{Y}_i}$

The value of sampling variance ψ_i raised with value 0,001;0,01;1.The simulation steps are as follows:

(i) Building a population consist of 18 small areas in this study is the district area. The independent variable is generated from the normal distribution . The population is a random component derived from the normal distribution $x_{ij} \overset{iid}{\sim} N(0,1)$ so that the linear mixed model built in this scenario will satisfy the normalized assumption. In the spatial hierarchical bayes model is given an additional treatment of spatial weights. Spatial weights are formed based on rook contiguity type which is 1 if area i is squashed side with region j and 0 (zero) if region i does not coincide with j.

(ii) Generating independent variables x_{ij} as many as N units, value x_{ij} used to spatial hierarchical bayes model in the simulation process.

(iii) Generating data with inverse gamma distribution parameters (a, b) set to $a = b = 0.0001$, $a = b = 0.001$, $a = b = 0.01$, $a = b = 0.1$, $a = b = 1$.

(iv) Calculating RB and RRMSE from the simulation $IG\,(\,0.0001; 0.0001)$, $IG \sim (0.001; 0.001)$, $IG \sim (0.01; 0.01)$, $IG \sim (0.1; 01)$, $IG \sim (1; 1)$

(v) Generating data with ψ_i amount 0.001; 0.01; and 1 and value θ from direct estimate.

(vi) Calculating RB dan RRMSE from simulation $\psi_i = 0.001$, $\psi_i = 0.01$, and $\psi_i = 1$.

(vii) Calculating the response variable value y_{ij} depend on coefficient value β which has been specified for each small area.

In simulation, every iteration process is done, the new k-algorithm is obtained for the parameter β,σ^2, Y from district in the first until eighteenth. Algorithm is done as much as 5.000 iterations. The value of the parameter estimation is derived from the stationary markov mean value. The following is the result of iteration of parameter estimation of spatial hierarchical bayes beta parameter and response variable.

In the process of spatial hierarchical bayes modelling using spatial weights rook contiguity. Figure 1 simulation process stops when the burn in process has been completed on the trace plot. So that the posterior distribution has converged. Data comes from the target distribution and produces a fairly stable value because it has not formed a certain pattern. The density plot image shows the distribution pattern of the observed parameter observers tending to be symmetrical. The autocorrelation plot shows that the autocorrelation values in the first lag are close to one and then the values continue to decrease to zero, so it can be said that in the chain there is a correlation of estimation of parameters between iterations. The correlation indicates that the algorithm is already within the target distribution area.

Table 1 and 2 show that the mean and distribution variance the inverse gamma that gives the smallest RB and RRMSE value is the value of $a = b = 1$ where a is the average and b is the variance of inverse gamma distribution. Table 1 shows that 12 of the 18 district have the smallest RB values on $a = b = 1$. Table 2 shows that the smallest $RRMSE$ values spread almost evenly distributed on the average value and variance of inverse gamma distributions. However, the smallest $RRMSE$ values are mostly on average and variance 1. By therefore it can be concluded that the posterior distribution with the mean value and variance 1 gives the smallest RB and $RRMSE$ values.

Table 3 shows RB and RRMSE values of spatial hierarchical bayes with generated distribution data posterior inverse gamma (1,1) and variance sampling 1. The simulation result show that the RB reaches a sufficiently variable minimum value on the spatial HB model. However, the minimum RRMSE value is maximized on the model spatial hierarchical bayes. The simulation results show that hierarchical spatial bayes gives the least RRMSE value when variance sampling 1 and generated by inverse gamma distribution with average and variance 1.

Figure 1. *Trace Plot, Density Plot, Autocorrelation Plot* Beta 1

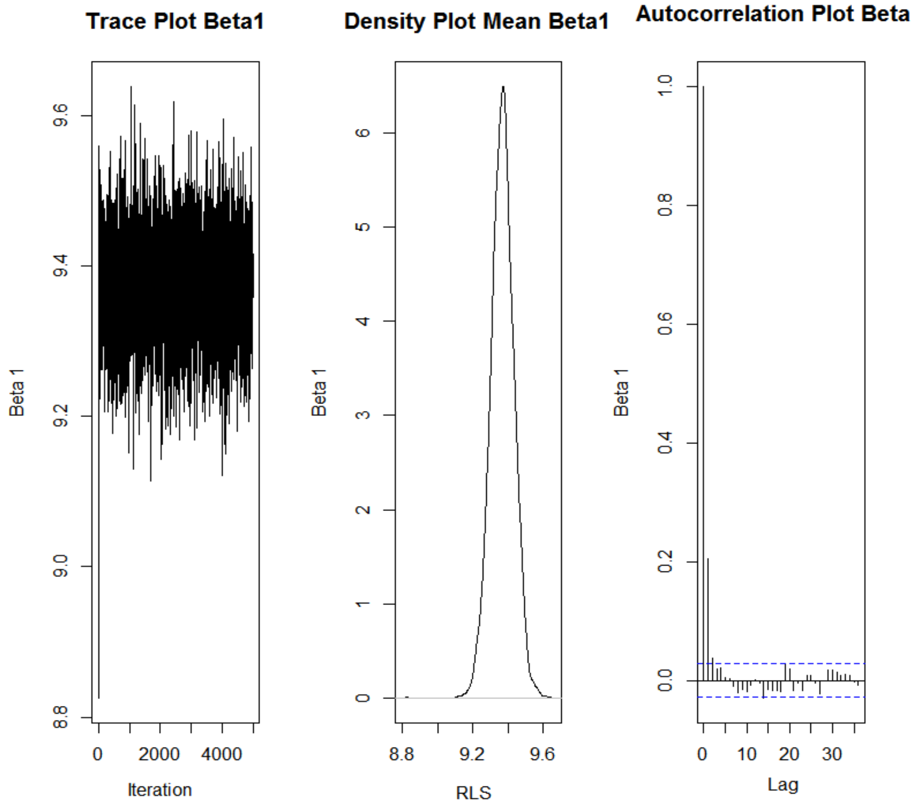

Table 1. Comparison of relative biased of spatial hierarchical bayes methods with a simulated posterior inverse gamma distribution

District	$a = b = 0.0001$	$a = b = 0.001$	$a = b = 0.01$	$a = b = 0.1$	$a = b = 1$
Pinang Sori	-0.062	0.058	-0.066	-0.089	-0.107
Badiri	-0.020	-0.020	-0.018	-0.019	-0.022
Sibabangun	-0.039	-0.029	-0.047	-0.085	-0.101
Lumut	-0.229	-0.211	-0.259	-0.354	-0.401
Sukabangun	-0.049	-0.048	-0.058	-0.079	-0.091
Pandan	0.241	0.253	0.237	0.199	0.184
Tukka	-0.148	-0.138	-0.163	-0.209	-0.235
Sarudik	-0.038	-0.049	-0.030	-0.026	-0.051
Tapian Nauli	0.034	0.041	0.031	0.006	-0.007
Kolang	0.041	0.043	0.038	0.044	0.047
Sorkam	0.070	0.067	0.066	0.073	0.061
Sorkam Barat	0.129	0.128	0.136	0.168	0.186
Barus	0.160	0.159	0.164	0.188	0.194
Sosorgadong	0.093	0.096	0.087	0.081	0.067
Andam Dewi	-0.052	-0.068	-0.022	-0.091	-0.142
Barus Utara	0.130	0.132	0.141	0.182	0.206
Manduamas	-0.228	-0.218	-0.237	-0.248	-0.263
Sirandorung	0.084	0.081	0.095	0.137	0.151

Table 2. Comparison of relative root mean square error of spatial hierarchical bayes method with a simulated posterior inverse gamma distribution

District	$a = b = 0.0001$	$a = b = 0.001$	$a = b = 0.01$	$a = b = 0.1$	$a = b = 1$
Pinang Sori	0.209	0.216	0.204	0.206	0.212
Badiri	0.235	0.217	0.220	0.190	0.184
Sibabangun	0.172	0.192	0.173	0.185	0.191
Lumut	0.322	0.353	0.329	0.372	0.397
Sukabangun	0.210	0.211	0.201	0.202	0.203
Pandan	0.227	0.240	0.216	0.185	0.172
Tukka	0.250	0.273	0.258	0.284	0.299
Sarudik	0.174	0.193	0.162	0.155	0.155
Tapian Nauli	0.196	0.195	0.177	0.165	0.160
Kolang	0.184	0.179	0.172	0.160	0.156
Sorkam	0.203	0.193	0.188	0.167	0.161
Sorkam Barat	0.230	0.213	0.216	0.210	0.217
Barus	0.217	0.206	0.209	0.209	0.210
Sosorgadong	0.198	0.203	0.181	0.163	0.154
Andam Dewi	0.650	0.557	0.561	0.347	0.350
Barus Utara	0.254	0.236	0.232	0.227	0.238
Manduamas	0.387	0.391	0.376	0.377	0.385
Sirandorung	0.251	0.221	0.234	0.214	0.215

Table 3. Relative bias and relative root mean square error of spatial hierarchical bayes modelling of Tapanuli Tengah Regency 2015

District	RB	RRMSE
Pinang Sori	-0.062	0.2177
Badiri	-0.021	0.2302
Sibabangun	-0.035	0.183
Lumut	-0.224	0.341
Sukabangun	-0.051	0.210
Pandan	0.243	0.230
Tukka	-0.146	0.262
Sarudik	0.045	0.183
Tapian Nauli	0.034	0.195
Kolang	0.037	0.183
Sorkam	0.067	0.202
Sorkam Barat	0.127	0.222
Barus	0.157	0.210
Sosorgadong	0.089	0.199
Andam Dewi	-0.062	0.603
Barus Utara	0.126	0.244
Manduamas	-0.225	0.388
Sirandorung	0.080	0.235

The modeling results show that the simulation process of the spatial hierarchical bayes method is constructed by generating a posterior distribution of inverse gamma with a mean value and variance of 0.0001; 0.001; 0.01; 0.1; 1 then the best RB and RRMSE value raised the value of the sampling variance of 0.001; 0.01; 1. The spatial hierarchical bayes model which gives the smallest RB and RRMSE values is the posterior inverse gamma distribution with averages and variance 1 and the sampling variance 1.

4. Conclusion

Small area estimation modelling use spatial hierarchical bayes with iteration process as much as 5,000 times. Some conclusions as follows:

(i) The process of generating data can be done by varying the mean value and variance of a significant posterior gamma distribution so that the convergence of the markov chain is achieved more quickly and results in smaller RRMSE values.

(ii) The spatial hierarchical bayes model which gives the smallest RB and RRMSE values is the posterior inverse gamma distribution with averages and variance 1 and the sampling variance 1.

5. References

[1] Chen C, Wakefield J, and Lumely T. (2014). The Use of Sampling Weights in Bayesian Hierarchical Models For Small Area Estimation. *Spatial and Spatio-temporal Epidemiology*, Vol. 11, 33–43.

[2] Guadarrama M, Molina I, dan Rao J.N.K. (2015). A Comparison of Small Area Estimation Methods for Poverty Mapping. *Statistics and Econometrics*, Vol. 15, No. 5, 1–24.

[3] Molina, Isabel, Nandram B, and Rao J.N.K. (2014). Small Area Estimation of General Parameters With Application to Poverty Indicator: A Hierarchical Bayes Approach. *The Annuals of Applied Statistics*, Vol. 8, No. 2, 852–885.

[4] Rao, J.N.K (2008). Some Methods For Small Area Estimation. *Rivista Internazionale in Science Sociali*, No.4, 387–406.

[5] Rao, J.N.K and Molina I (2015). *Small Area Estimation* . (New Jersey: John Wiley & Sons)