© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

A Comparison of Cluster Method and Nearest Neighbor Method for Non-sample Area in the Small Area Estimation

Annastasia Nika Susanti[1], Kusman Sadik[2], Anang Kurnia[3]

[1,2,3]Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

Corresponding Author: Kusman Sadik (**kusmansadik@gmail.com**)

ABSTRACT

Small area estimation is an indirect method to estimate the parameter of a population by using the model approach. The problem that often arises in the small area estimation is non-sampled area then the area random effect of non-sampled areas is can not estimate because no sample units are available in these areas. This paper proposed a method to solve the non-sampled area problem by adding the cluster information and by using the nearest neighbor area to estimate the area random effect through the Fast Hierarchical Bayes (FHB) approach. These methods are compared by using the simulation study and the evaluation is based on the Absolute Relative Bias (ARB) and Relative Root Mean Square Error (RRMSE). The result shows that the estimation by using the cluster method has smaller ARB and RRMSE values than the estimation by using the nearest neighbor area in various sample sizes and various population sizes. Then it can be said that cluster method is better to provide the estimators of non-sampled area than the nearest neighbor area method.

Keywords : Small Area Estimation, Non-Sampled Area, Cluster Method, The Nearest Neighbor Area, Poverty Indicators.

I. INTRODUCTION

Small area estimation is one of parameter estimation methods that has been applied in many fields since it can provide estimation to a small area level. This method is used when the availability of sample size on an area is too small. Based on Kurnia (2009), small area estimation method uses the additional information in the form of auxiliary variables from outside the area, from within the area itself, and from outside the survey. The auxiliary variables can be obtained from a census or administrative data (Ghosh and Rao, 1994). This method basically uses the model approach to estimate the parameter so it is also called as indirect estimation method (Molina et al., 2014)

In the small area estimation method, the possible problems that might occur are some non-sample areas in the survey. This problem causes the area random effects of the non-sampled area cannot be estimated. The previous study used the cluster information to estimate the parameter in the non-sample area. Anisa et al. (2014) stated that the additional of cluster information in the model provides small Relative Bias (RB) and Relative Root Mean Square Error (RRMSE). Then, Wahyudi et al. (2014) compared some clustering methods and concluded that Ward method is better than others. Based on that study, Sundara et al. (2017) used the cluster information by using the Ward method to estimate the area random effect of non-sampled area with the empirical Bayes (EB) method which was first introduced by Molina et al. (2010). Then in 2014, Molina et al. proposed another method to estimate the FGT poverty indicators i.e hierarchical Bayes (HB) method. The study resulted that both EB and HB practically give the similar estimation. Yet, both HB and EB method require the completeness of auxiliary variables for each unit in the survey which is difficult to reach. Moreover, the

HB method is ineffective to use when the population size is very large since the number of samples to be generated in each Monte Carlo iteration is as much as $N_d - n_d$; $d = 1, \dots, D$. As an alternative to this problem, there is a faster version of the HB method which is called as fast hierarchical Bayes (FHB) method. It is called faster since the methods just needs as much as the original sample size in every Monte Carlo iteration. The FHB approach can be implemented analogously to the fast EB (FEB) approach (Molina et al. 2012).

Based on that background, the authors are interested to study deeply about the non-sampled area in the small area estimation model mainly by using the FHB method since the method is more simple to use. The authors try to provide another solution to resolve the non-sampled area problem by using the nearest neighbor (NN) area. Then the results will be compared with the estimation by using the cluster information. The goodness of estimation is measured by the Absolute Relative Bias (ARB) and Relative Root Mean Square Error (RRMSE).

The organization of this paper is as follows. Section 2 describes the estimation algorithm of the FHB method. Section 3 presents the result of the simulation study to compare the estimators between nearest neighbor area and cluster information. Finally, in Section 4 gives some concluding remarks of the research.

II. METHODS AND MATERIAL

This paper is done by using a simulation study with the design based simulation to estimate the poverty indicators, such as poverty incidence (P_0), poverty gap (P_1) , and poverty severity (P_2) index. The number of Monte Carlo repetition is $H = 100$ and the number of sampling repetition is $k = 100$. This simulation uses the various sample sizes (n_{d1}=5 , n_{d2}=10, and n_{d3}=80) and the various population area sizes (N_{d1}=1000 , N_{d2}=3000 , and N_{d3}=20000). The FHB method is done by forming the posterior distribution just as on the HB method by Molina et al. (2014) which is defined by

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \sigma^2, \rho | \boldsymbol{y}_s) = \pi_1(\boldsymbol{u}|\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{y}_s)\pi_2(\boldsymbol{\beta}|\sigma^2, \rho, \boldsymbol{y}_s)\pi_3(\sigma^2|\rho, \boldsymbol{y}_s)\pi_4(\rho|\boldsymbol{y}_s) \quad (1)$$

The simulation is started as follows :

1) Generated of the population, which the number of areas (D) is 40 areas where three of them are assumed non-sampled (area 16, 21, and 40), the coefficient regression β is $(3, 0.03, -0.04)'$, the random area effects σ_u^2 is 0.15^2, and the error variance σ^2 is 0.5^2.

2) Generated of variable of interest y_{di} based on the Nested Error Model (NER) as in Molina et al. (2014) below:

$$\boldsymbol{y}_{di} = \boldsymbol{x}'_{di}\boldsymbol{\beta} + u_d + e_{di}\,, d = 1, \dots, D; i = 1, \dots, N_d \quad (2)$$

with u_d is a random effect of area d which is distributed as $u_d|{\sigma_u}^2 \overset{iid}{\sim} N(0, {\sigma_u}^2)$, and e_{di} is errors which is distributed as $e_{di}|\sigma^2 \overset{iid}{\sim} N(0, \sigma^2 {w_{di}}^{-1})$. The auxiliary variable is generated by $x_1 \sim binom(N_d,\ p = 0.3 + 0.5x\frac{d}{D})$ and $x_2 \sim binom(N_d, p = 0.2)$.

3) Draw the sample unit without replacement and assumed that area 16, 21, and 40 are non-sampled then calculated the parameters and the direct estimators of the poverty indicators.
4) Generated the posterior distribution as follows :
   4.1) Generated the distribution of ρ , the intra-class corelation which makes a grid $R = 1000$ then the $\pi_4(\rho|\boldsymbol{y}_s)$ can be written as
   $$\pi_4(\rho_r) = \frac{k_4(\rho_r)}{\sum_{r=1}^{R} k_4(\rho_r)} \quad (3)$$
   4.2) Generated ρ as much H by discrete distribution $\{\rho_r, \pi_4(\rho_r)\}_{r=1}^{R-1}$ then add it to the uniform distribution in the interval (0, 1/R).

5) Generated the error variance distribution $\pi_3(\sigma^2|\rho, \boldsymbol{y}_s)$ as below

$$\sigma^{-2}|\rho, \boldsymbol{y}_s \sim Gamma\left(\frac{n-p}{2}, \frac{\gamma(\rho)}{2}\right) \quad (4)$$

Then take $\sigma^2 = 1/\sigma^{-2}$.

6) Generated the $\pi_2(\boldsymbol{\beta}|\sigma^2, \rho, \boldsymbol{y}_s)$ distribution i.e

$$\boldsymbol{\beta}|\sigma^2, \rho, \boldsymbol{y}_s \sim N\left(\widehat{\boldsymbol{\beta}}(\rho), \sigma^2(\boldsymbol{Q})^{-1}(\rho)\right) \quad (5)$$

7) Generated the area random effects from

$$u_d|\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{y}_s \overset{ind}{\sim} N\left[\lambda_d(\rho)(\bar{y}_d - \bar{x}'_d\boldsymbol{\beta}), \left(1 - \lambda_d(\rho)\right)\frac{\sigma^2\rho}{1-\rho}\right] \quad (6)$$

The area random effects of the non-sampled area are estimated by following this method.

7.1) The nearest neighbor area (NN)

The area random effects of non-sampled area is generated from the equation (6) with mean $\lambda_d(\rho)(\bar{y}_d - \bar{x}'_d\boldsymbol{\beta})$ and variance $\left[\lambda_d(\rho)(\bar{y}_d - \bar{x}'_d\boldsymbol{\beta}), \left(1 - \lambda_d(\rho)\right)\frac{\sigma^2\rho}{1-\rho}\right]$ where d is the nearest area from the non-sampled area.

7.2) The cluster information

The area random effect of the non-sampled area by using the cluster information is done by clustering all areas first, then the non-sampled area will be known lied on which cluster. Furthermore, the area random effects are generated by using (6) with mean and variance obtained by averaging the λ_d, $\bar{y}_d$, and $\bar{x}_d$ from others area in the same cluster with non-sampled area.

8) Finally the posterior distribution of $\pi(Y_{di}|\boldsymbol{y}_s)$ is

$$Y_{di}|\boldsymbol{y}_s, \boldsymbol{\theta} \sim N(x'_{di}\boldsymbol{\beta} + u_d, \sigma^2) \quad (7)$$

9) Generated the variable of interest based on (7) as much of the original sample size, then calculated the FHB estimator of the poverty indicators from H times Monte Carlo repetition as below

$$\hat{\delta}_d^{FHB} = \hat{P}_{\alpha d}^{FHB} = E(\delta_d|\boldsymbol{y}_s) \approx \frac{1}{H}\sum_{h=1}^{H}\hat{\delta}_d{}^{(h)} \approx \frac{1}{H}\sum_{h=1}^{H}\hat{P}_{\alpha d}{}^{(h)} \quad (8)$$

with

$$\delta_d{}^{(h)} = P_{\alpha d}^{(h)} = \frac{1}{N_d}\left[\sum_{i\in s_d}\left(\frac{z-E_{di}}{z}\right)^{\alpha} I(E_{di} < z) + \sum_{i\in r_d}\left(\frac{z-E_{di}{}^{(h)}}{z}\right) I\left(E_{di}{}^{(h)} < z\right)^{\alpha}\right] \quad (9)$$

z is the poverty line and

$$I(E_{di} < z) = \begin{cases} 1; \text{if } E_{di} < z \text{ (poor)} \\ 0; \text{if } E_{di} \geq z \text{ (not poor)} \end{cases}$$

10) The last step repeats the step (3) until step (9) as much of k times sampling repetition, then calculated the average of estimators which is obtained by direct estimators, HB method, and FHB method.

III. RESULTS AND DISCUSSION

The simulation study in this research is basically done to see how well both methods provide an estimator for the parameter of the population. The measure of goodness which is used in this research is ARB and RRMSE. The smaller the value indicates that the method is better than others. The simulation is done by using various sample sizes to see the effect of sample increasing towards the estimator. In this study, the non-sampled areas are assumed in area 16, 21, and 40. Then, based on the simulation study, the result of the ARB and RRMSE values for each non-sampled area in the various sample sizes is presented in the following tables.

Table 1: The ARB and RRMSE value of area 16 in the

Area		Sample Sizes	ARB		RRMSE	
			Cluster	NN	Cluster	NN
16	P_0	5	0.4040	0.3462	0.4250	0.3782
		10	0.3430	0.3511	0.3646	0.3910
		80	0.2960	0.4016	0.3026	0.4161
	P_1	5	0.5582	0.4378	0.5923	0.4872
		10	0.4522	0.4354	0.4897	0.4935
		80	0.3804	0.4533	0.3889	0.4766
	P_2	5	0.7436	0.5610	0.7931	0.6270
		10	0.5887	0.5390	0.6408	0.6220
		80	0.4852	0.5159	0.4943	0.5503

Based on the table it can be seen that the increasing in the number of sample sizes resulted in the ARB and RRMSE values which is obtained by the direct estimation, HB, and FHB method becoming smaller based on cluster method. On the other hand, the ARB and RRMSE values which obtained by the Nearest Neighbor (NN) method is tend to be fluctuation as the increasing of sample sizes. The ARB and RRMSE values from NN method generally can be said that it is getting bigger as the increasing of the sample sizes. If the ARB and RRMSE values which obtained by the cluster method and the NN method is compared, it can be concluded that based on the Table 1, cluster method produces the smaller value of ARB and RRMSE than the NN method. Then the results for area 21 are presented in the Table 2 below.

Table 2: The ARB and RRMSE value of area 21 in the various sample sizes

Area		Sample Sizes	ARB		RRMSE	
			Cluster	NN	Cluster	NN
21	P_0	5	0.7849	1.0080	0.8173	1.0869
		10	0.7335	1.0747	0.7593	1.1452
		80	0.7024	1.3112	0.7090	1.3382
	P_1	5	1.1207	1.4300	1.1784	1.5598
		10	1.0332	1.4964	1.0741	1.6199
		80	0.9508	1.7705	0.9609	1.8159
	P_2	5	1.4293	1.8173	1.5214	2.0075
		10	1.2953	1.8731	1.3551	2.0603
		80	1.1444	2.1667	1.1586	2.2325

Based on the Table 2, it can be seen that the increasing of sample sizes causes the value of ARB and RRMSE which is produced by cluster method for area 21 is getting smaller. Otherwise, the ARB and RRMSE values of NN method for area 21 are getting bigger. Table 2 shows that the cluster method provides a better estimator than the NN method since the value of the ARB and RRMSE is smaller than the NN method. The last result to know the effects of sample sizes is presented in the table below for area 40.

Table 3: The ARB and RRMSE value of area 40 in the various sample sizes

Area		Sample Sizes	ARB		RRMSE	
			Cluster	NN	Cluster	NN
40	P_0	5	0.2626	0.2475	0.3061	0.3237
		10	0.2290	0.3321	0.2658	0.4145
		80	0.2829	0.4394	0.2901	0.4691
	P_1	5	0.3932	0.3784	0.4570	0.5003
		10	0.3356	0.4780	0.3890	0.6037
		80	0.3937	0.6122	0.4038	0.6518
	P_2	5	0.5322	0.5222	0.6206	0.7037
		10	0.4474	0.6271	0.5164	0.7987
		80	0.5035	0.7781	0.5164	0.8273

Table 3 for area 40, presented the differences result from the previous tables. Based on the Table 3 it can be seen that the cluster method tends to produce the fluctuating value of ARB and RRMSE while the ARB and RRMSE values of NN method is getting bigger as the increasing of sample sizes. Yet, if we compared the ARB and RRMSE values of those methods, the cluster method is better than the NN method since it produces the smaller value. This shows that the nearest neighbor area does not always guarantee that they have the similar characteristics. Otherwise, the average of other areas which is laid on the same cluster can produce better estimators.

This simulation also uses some population area sizes i.e. N_{d1}=1000, N_{d2}=3000, and N_{d3}=20000. It aimed to see the effect of population sizes toward the ARB and RRMSE values of both methods. The result of each non-sampled area is presented in the following tables.

Table 4: The ARB and RRMSE value of area 16 in the

Area	Indicators	ARB		RRMSE	
		Cluster	NN	Cluster	NN
N_{d1}	P_0	0.0982	0.1822	0.1243	0.2374
	P_1	0.2630	0.2876	0.3056	0.3821
	P_2	0.5373	0.5030	0.5941	0.6449
N_{d2}	P_0	0.6765	0.6851	0.6773	0.6892
	P_1	0.7635	0.7688	0.7642	0.7720
	P_2	0.8085	0.8120	0.8092	0.8148
N_{d3}	P_0	0.2684	0.2315	0.2906	0.2587
	P_1	0.3643	0.2701	0.4011	0.3032
	P_2	0.4717	0.3009	0.5249	0.3396

Based on the Table 4, it can be seen that generally the increasing of population area sizes causes the ARB and RRMSE values also getting bigger in the cluster method and the NN method. It can be known from Table 4 that the ARB value which is obtained by a cluster method is smaller than the NN method in all population area sizes. The same result is given by area 21 and area 40 then in this case, those results no need to display.

The comparison of the ARB and RRMSE value of cluster method and the NN method is presented in the following figures for P_0 indicators as an example. The figures were obtained by using the biggest population area sizes since the FHB method is better to use in the large population and by using the sample sizes as much of 80 each area.

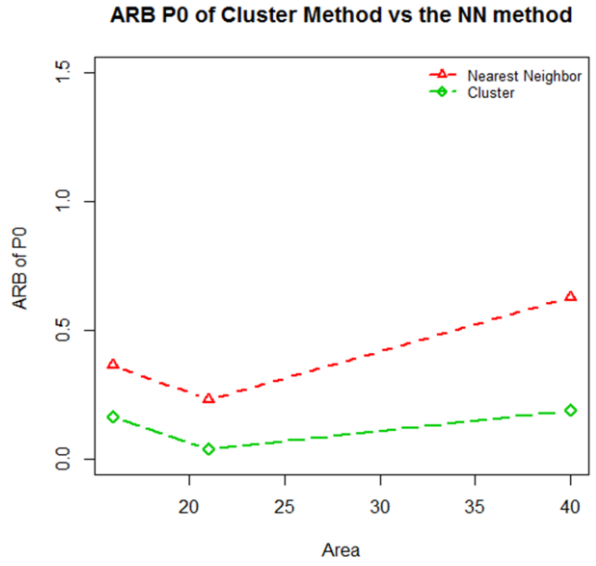

Fig 1: The ARB of poverty incidence by using cluster method and nearest neighbor method

Fig. 1 shows that the cluster method gives the smaller ARB and RRMSE value than the nearest neighbor method. It can be said that the cluster information can provide a better estimation of non-sampled area in terms of small area estimation. The next figure below shows the RRMSE values of P_0 indicators as an example by using the cluster method and the nearest neighbor method. The figure also shows the same result as the Fig. 1. The RRMSE values of the cluster method are smaller than the nearest neighbor method. Therefore, the cluster method is better to use than the nearest neighbor in the small area estimation when there are some non-sampled areas.

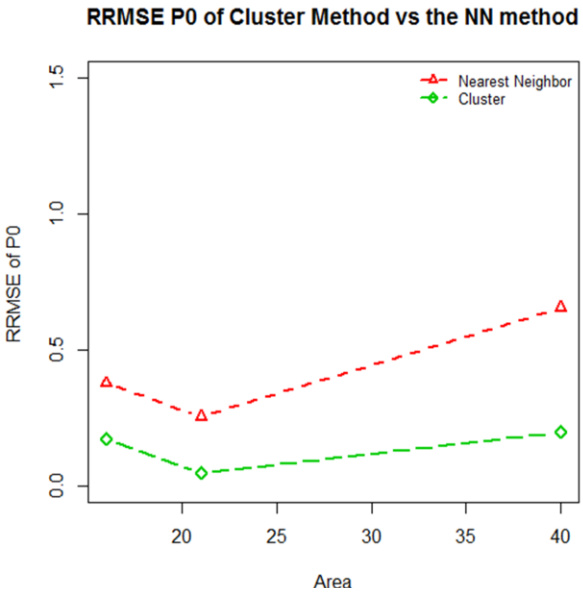

Fig 2: The RRMSE of poverty incidence by using cluster method and nearest neighbor method

IV.CONCLUSION

Based on the simulation study, it can be concluded that cluster method is better than the nearest neighbor method since it provides a smaller value of ARB and RRMSE than the nearest neighbor area method for non-sampled area. It can be said that the non-sampled area not always guarantee will have the similar characteristics with its nearest area .

V. REFERENCES

[1] Kurnia A. 2009."An empirical best prediction method for logarithmic transformation model in small area estimation with particular application to susenas data". [*doctoral dissertation*] Bogor Agricultural University, Indonesia.

[2] Ghos M and Rao JNK. 1994."Small Area Estimation : An Appraisal". Statistical Science. 9(1): 55 -76 DOI:10.1214/ss/1177010647

[3] Molina I, Nandram B, Rao JNK. 2014. "Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach". The Annals of Applied Statistics. 8 (2): 852-885 DOI: 10.1214/13-AOAS702

[4] Anisa R, Kurnia A, and Indahwati. 2014. "Cluster Information of Non-sampled Area in Small Area Estimation*"*. IOSR Journal of Mathematics (IOSR-JM).10(1): 15-19.

[5] Wahyudi, Notodiputro KA, Kurnia A, and Anisa R. 2016. "A Study of Area Clustering using Factor Analysis in Small Area Estimation (An Analysis of Per Capita Expenditures of Subdistricts Level in Regency and Municipality of Bogor)". AIP Conference Proceedings.1707(1). DOI : 10.1063/1.4940874

[6] Sundara VY, Sadik K, and Kurnia A. 2017. "Cluster Information of Non-sampled Area in Small Area Estimation of Poverty Indicators using Empirical Bayes", AIP Conference Proceedings. 1827(1). DOI : 10.1063/1.4979442

[7] Molina I and Rao JNK. 2010. "Small Area Estimation of Poverty Indicators".The Canadian Journal of Statistics. 38(3) : 369 – 385

[8] Ferreti C and Molina I. 2012. "Fast EB Method for Estimating Complex Poverty Indicators in Large Population". Journal of The Indian Society of Agricultural Statistics. 66(1) :105 -120.