IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 1 Ver. II. (Feb. 2014), PP 15-19
www.iosrjournals.org

Cluster Information of Non-Sampled Area In Small Area Estimation

Rahma Anisa, Anang Kurnia, Indahwati
Department of Statistics, Bogor Agricultural University, Indonesia

ABSTRACT: *Empirical Best Linear Unbiased Predictor (EBLUP) has been widely used to predict parameters in area with small or even zero sample size. The problem is when this model should be used to predict the parameters of non-sampled area. Ordinary EBLUP predicted the parameters using synthetic model which ignore the area random effects because lack of non-sampled area information. Thus, those prediction will be distorted based on a single line of the synthetic model. One of idea that developed in this paper is to modify the prediction model by adding cluster information by assuming that there are similiarities among particular areas. These information will be added into the model to modify the intercept of prediction model. Another approach is by adding random effects of auxiliary variable into the previous model in order to modify both intercept and slope of the prediction model. In this paper, simulation process is carried out to study the performance of the proposed models compared with ordinary EBLUP. All models evaluated based on the value of Relative Bias (RB) and Relative Root Mean Squares Error (RRMSE). The results show that the addition of cluster information can improve the ability of the model to predict on non-sampled areas.*

Keywords – *Clustering, EBLUP, Linear Mixed Model*

I. INTRODUCTION

Small area estimation (SAE) is used as an alternative approach to estimate the parameters of the areas with a very small or even zero sample size. This demand appears when we need an estimator with good accuracy in a smaller subpopulation level, such as at the level of regency/municipality, subdistrict, or even village [1]. Indirect estimator is prefered to estimate parameters in a small area because it is able to overcome the weakness of direct estimator which can produce a large standard error [2]. One of the approach is Empirical Best Linear Unbiased Prediction (EBLUP) method. The problem is ordinary EBLUP perform parameter estimation for non-sampled area using synthetic model which will ignore the area random effects because of the lack of information of non-sampled area [3]. As a consequence, the resulting predictive values will be distorted into a single line of the synthetic model and may caused considerable bias.

One of the ideas developed in this paper is to assume that there are similiarities among particular areas which can be analyzed using clustering technique. Cluster information is expected to be able to improve the estimation for non-sampled area by modifying intercept of the prediction model. Another approach is by adding the average of random effects estimator of area and auxiliary variables in each cluster to modify both intercept and slope of the prediction model.

In this paper, simulation process carried out to compare the performance of the proposed models with ordinary EBLUP. The proposed models with an intercept and or slope that has been modified are expected to yield a better precision in estimating the parameters.

II. EMPIRICAL BEST LINEAR UNBIASED PREDICTOR

Consider special case of linear mixed model for i-th area and j-th unit:

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij} \qquad (1)$$

with y_{ij} denotes sample observation, x_{ij} denotes auxiliary variable whose values is known for all units in population, v_i denotes area random effects which distributed $v_i \sim iid \quad N(0,\sigma_v^2)$ and e_{ij} is error term that $e_{ij} \sim iid \quad N(0,\sigma_e^2)$, which depends on parameter σ called variance component, if defined that $\sigma = \left(\sigma_e^2, \sigma_v^2\right)$.

EBLUP is a two-stage estimator of parameter $t(\sigma)$ which depends on unknown parameter σ [4]. This approach replace the parameter σ with its estimator, $\hat{\sigma}$, so estimation will be carried out on parameter $t(\hat{\sigma})$. Note that the variance components σ have to be estimated before we can estimate the parameter of interest.

If it is defined that parameter of interest is the *i*-th small area mean, EBLUP estimator for the sampled area mean can be written as:

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} \right) \tag{2}$$

with s_i denotes sampled units and r_i denotes non-sampled units in the i -th area . Thus, $\hat{y}_{ij}$ is estimated value for non-sampled units which calculated with following formula:

$$\begin{aligned} \hat{y}_{ij} &= x'_{ir}\hat{\beta} + \hat{\gamma}_i(\bar{y}_{is} - \bar{x}_{is}\hat{\beta}) \\ &= x'_{ir}\hat{\beta} + \hat{v}_i \end{aligned} \tag{3}$$

where $\hat{\beta} = \hat{\beta}(\hat{\sigma})$ is generalized least squares estimator of β , and $\hat{\gamma}_{ij} = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + n_i^{-1}\hat{\sigma}_e^2 \right)$. EBLUP estimator for non-sampled area mean is:

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j* \in r_i} \hat{y}_{ij*} \right) \tag{4}$$

with $\hat{y}_{ij*}$ is an estimated value which calculated by the following formula:

$$\hat{y}_{ij*} = x'_{ij*}\hat{\beta} \ . \tag{5}$$

III. CLUSTER ANALYSIS

Cluster analysis is a multivariate technique to classify objects based on its similiarities [5]. Characteristic similarities between objects can be measured by Euclidean distance, Mahalanobis distance, and others. There are two approaches in clustering method, hierarchical and non-hierarchical [6]. Hierarchical clustering method consists of agglomerative and divisive approach. This method is used when the number of clusters is unknown. While non-hierarchical clustering method is used when the number of clusters is known.

The problem that might occured in the process of clustering is a violation of the assumption, such the presence of multicollinearity and outliers. Hierarchical cluster analysis method k-medoid can be an alternative to overcome outliers. Another issue that might be found is when clustering analysis is based on categorical variables, or a mix of categorical and numerical variables. One approach that can deal with these problems is two-step cluster method. In addition, this method is also capable of handling large data sets.

IV. PROPOSED MODELS

There are five models proposed which developed from the basic model EBLUP (Model-0) by modifying its intercept and or slope. Model-1, modification of EBLUP model by adding average of the area random effects on each clusters into prediction model for non-sampled area. If it is defined that average of area random effects is:

$$\hat{\hat{v}}_{i(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k} \hat{v}_i \tag{6}$$

with m_k is the number of sample area on the *k*-th cluster, then Model-1 can be written as:

a. model for population:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + v_i + e_{ijk} \tag{7}$$

b. prediction model for sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{v}_i \tag{8}$$

c. prediction model for non-sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{\hat{v}}_{i(k)} \ . \tag{9}$$

Model-2, modification of EBLUP model by adding cluster effects into prediction models. If it is defined that $\hat{C}_k$ is a simplified form of dummy variable coefficients estimator for a number of k cluster, then it can be described that $\hat{C}_k = \hat{\alpha}_1 d_1 + ... + \hat{\alpha}_{k-1} d_{k-1}$, with $d_1, ..., d_{k-1}$ are dummy variables for cluster. This model can be written as follow:

a. model for population:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + C_k + v_i + e_{ijk} \tag{10}$$

b. prediction model for sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{C}_k + \hat{v}_i \tag{11}$$

c. prediction model for non-sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{C}_k . \tag{12}$$

Model-3, that is combination of Model-1 and Model-2. This model is obtained by adding average of the area random effects on each clusters and also cluster effects into prediction model for non-sampled area. It should be noted that for sampled area, the prediction model of Model-3 is actually the same as Model-2. This model can be written as follow:

a. model for population:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + C_k + v_i + e_{ijk} \tag{13}$$

b. prediction model for sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{C}_k + \hat{v}_i \tag{14}$$

c. prediction model for non-sampled area:

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{C}_k + \hat{\hat{v}}_i \ . \tag{15}$$

Model-4, this model is not only having random effects for area but also for auxiliary variables on each area. In other words, this model is modification of Model-1 by adding the average of auxiliary variable random effects on each area of each cluster. This model can be written as follow:

a. model for population:

$$\begin{aligned} y_{ijk} &= \beta_0 + \beta_1 x_{ijk} + \gamma_{0i} + \gamma_{1i} x_{ijk} + e_{ijk} \\ &= \left(\beta_0 + \gamma_{0i}\right) + \left(\beta_1 + \gamma_{1i}\right) x_{ijk} + e_{ijk} \end{aligned} \tag{16}$$

b. prediction model for sampled area:

$$\hat{y}_{ijk} = \left(\hat{\beta}_0 + \hat{\gamma}_{0i}\right) + \left(\hat{\beta}_1 + \hat{\gamma}_{1i}\right) x_{ijk} \tag{17}$$

c. prediction model for non-sampled area:

$$\hat{y}_{ijk} = \left(\hat{\beta}_0 + \hat{\hat{\gamma}}_{0(k)}\right) + \left(\hat{\beta}_1 + \hat{\hat{\gamma}}_{1(k)}\right) x_{ijk} \qquad (18$$

with $\hat{\hat{\gamma}}_{0(k)}$ is average of the area random effects on *k*-th cluster and $\hat{\hat{\gamma}}_{1(k)}$ is average of auxiliary variable random effects on each area of *k*-th cluster. Both of them can be calculated respectively using following formulas:

$$\hat{\hat{\gamma}}_{0(k)} = \frac{1}{m_k} \sum_{i=1}^{m_k} \hat{\gamma}_{0i} \tag{19}$$

$$\hat{\hat{\gamma}}_{1(k)} = \frac{1}{m_k} \sum_{i=1}^{m_k} \hat{\gamma}_{1i} \ . \tag{20}$$

Model-5, modification of Model-4 by adding cluster effects. This model can be written as:

a. model for population:

$$y_{ijk} = \left(\beta_0 + \gamma_{0i}\right) + \left(\beta_1 + \gamma_{1i}\right) x_{ijk} + C_k + e_{ijk} \tag{21}$$

b. prediction model for sampled area:

$$\hat{y}_{ijk} = \left(\hat{\beta}_0 + \hat{\gamma}_{0i}\right) + \left(\hat{\beta}_1 + \hat{\gamma}_{1i}\right) x_{ijk} + \hat{C}_k \tag{22}$$

c. prediction model for non-sampled area:

$$\hat{y}_{ijk} = \left(\hat{\beta}_0 + \hat{\hat{\gamma}}_{0(k)}\right) + \left(\hat{\beta}_1 + \hat{\hat{\gamma}}_{1(k)}\right) x_{ijk} + \hat{C}_k . \tag{23}$$

V. Empirical Study

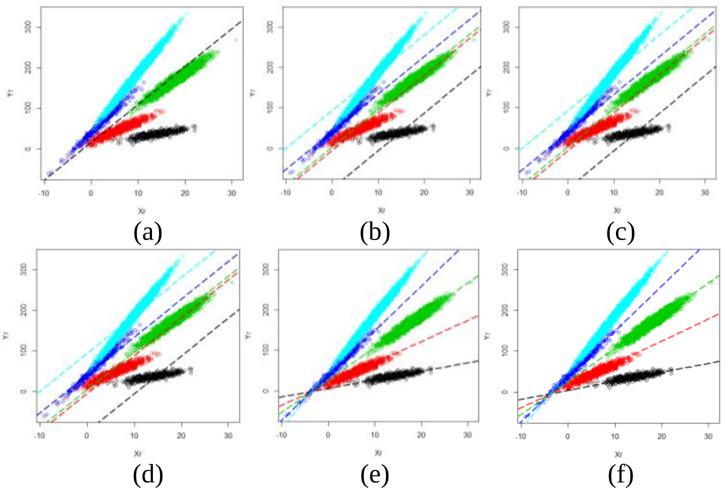
The data in simulation process are obtained by generating population consisting of 40 areas, with a population size of each area ranged from 100 to 1500 units. The population is assumed to consist of 5 clusters. Auxiliary variable is normally distributed $X \sim iid \quad N(\mu_k, \sigma_x^2)$ with different mean for each cluster. If μ is a vector containing μ_k then it can be written as $\mu' = (14, 6, 18, 2, 10)'$, and $\sigma_x^2 = 3$. Random effects *v* and *ε* are also normally distributed with zero means and each variance are $\sigma_v^2 = 3$ and $\sigma_\varepsilon^2 = 5$. Thus, this would generate a population where the data are completely separate between clusters. Simulation data sets are constructed through the following population model:

$$y_{ijk} = \beta_{0k} + \beta_{1k} x_{ijk} + v_i + e_{ijk} \tag{24}$$

with β_{0k} and β_{1k} are coefficients which have different value for each cluster.

Sample size which used in this simulation is 3% of the total observations in the population. Data of sampled areas is used to build the models, EBLUP models (Model-0) and the proposed models (Model-1 to Model-5). In non-sampled area cases, Model-0 yield a global model prediction while the proposed models yield local model predictions.

In Figure 1(a), it can be seen that Model-0 has only one line of prediction model to estimate observations on non-sampled areas. Model-1, Model-2 and Model-3 has an intercept that had been modified by adding cluster effects and or average of area random effects estimator in each cluster. Thus, the prediction models are more capable to fit the actual observed values of non-sampled areas compared to Model-0 (ordinary EBLUP). Model-4 and Model-5 are model that has been modified both in its intercept and slope. It can be seen in Figure 1(e) and 1(f) that both models are even better than the previous models in fitting the actual values of non-sampled areas.



(a) (b) (c)

(d) (e) (f)

Fig. 1. Plots of non-sampled area data and prediction models based on (a) Model-0, (b) Model-1, (c) Model-2, (d) Model-3, (e) Model-4, and (f) Model-5.

The resulting prediction from resampling process and the modeling that has been repeated 1000 times is used to to calculate Relative Bias (RB) and Relative Root Mean Squares Error (RRMSE) for each area in the population. Calculation results for the sampled area can be seen in Table 1, and for the non-sampled area can be seen in Table 2 and 3.

Table 1 Median of RB and RRMSE over sampled areas

	Model-0	Model-1	Model-2	Model-3	Model-4	Model-5
Median of RB	7.9419	7.9419	0.0450	0.0450	8.1700	-3.6028
Median of RRMSE	37.0732	37.0732	2.0388	2.0388	34.3445	13.9384

The simulations shows that the estimates in Model-2 and Model-3 are same. This occurs because the value $\hat{\bar{v}}_{i(k)}$ on Model-3 is close to zero. As a result, the values of RB and RRMSE calculated from both models are also equal. This could be due to area random effects are mostly already covered within the cluster effects. Note that for the sampled area, prediction model of the Model-1 is the same as the Model-0. Therefore, RB and RRMSE values of both models are also equal one another.

Model-2 and Model-3 have the smallest absolute median of RB and RRMSE than other models for the sampled area. Other models tend to have relatively larger median of RRMSE, including Model-0 or EBLUP models. This is quite interesting because the EBLUP model has considered the area random effects within predicted values for the sampled area. The addition of cluster effects was able to minimize median of RRMSE for sampled areas, of course assuming that the clustering for all area in the population is carried out perfectly.

Table 2 Median of RB and RRMSE over non-sampled areas

	Model-0	Model-1	Model-2	Model-3	Model-4	Model-5
Median of RB	-15.1701	-0.3475	-0.0836	-0.0836	-0.0637	-0.0211
Median of RRMSE	40.7609	2.9050	2.3392	2.3392	2.4500	2.4655

Calculation results in Table 2 shows that the model with intercept and slope modification simultaneously has a smaller absolute median of RB than other models for prediction of non-sampled area. The smallest absolute median of RB resulted by Model-5, but the value only slightly different with Model-4. The smallest absolute median of RRMSE for non-sampled area is resulted by Model-2 and Model-3. This means that models with intercept and slope modification have a better performance when evaluated based on the value of RB, but models with simply intercept modification have better performance based on the value of RRMSE. Note that resulting median of RRMSE values of the last two models are slightly larger than Model-2 and Model-3, the models with intercept modifications by adding the cluster fixed effects. It is probably caused by variance component of Model-4 and Model-5 which more complex because both models assume that the auxiliary variables are random.

VI. Conclusion

The addition of cluster effects on the prediction model for non-sampled areas can improve the capability of the model so that the resulting Relative Bias (RB) and Relative Root Mean Squares Error (RRMSE) is smaller than the ordinary EBLUP model. If it is assumed that the effects of auxiliary variables are fixed, the addition of average area random effects estimator of each cluster is less effective to used in the model that already contains cluster effects. Modification of the intercept and slope of the prediction models simultaneously was able to reduce the bias of prediction values for non-sampled areas, but the resulting RRMSE is larger than the model with only intercept modifications. It should be note that this would work with proper clustering technique and variable selection which are most capable to describe variations of the observed variables of interest.

References

[1] A. Kurnia. “Prediksi Terbaik Empirik untuk Model Transformasi Logaritma di Dalam Pendugaan Area Kecil dengan Penerapan pada Data Susenas". Unpublished doctoral dissertation, Department of Statistics FMIPA-IPB, Indonesia, 2009.

[2] J.N.K. Rao. *Small Area Estimation*. New York: John Wiley & Sons, 2003.

[3] A. Saei and R. Chambers. “Empirical Best Linear Unbiased Prediction for Out of Sample Area”. S3RI Methodology Working Papers, Southampton Statistical Sciences Research Institute, March 2005.

[4] K. Das, J. Jiang, and J.N.K. Rao. “Mean Square Error of Empirical Predictor”. *The Annals of Statistics*, vol. 32, no. 2, pp. 818-840, June 2004.

[5] A.A. Mattjik and I.M. Sumertajaya. *Sidik Peubah Ganda*. Bogor: Department of Statistics FMIPA-IPB, 2011.

[6] R.A. Johnson and D.W. Wichern, D.W. *Applied Multivariate Statistical Analysis 6th Edition*. London: Prentice-Hall, 2007.