Mean Square Error of Non-Sampled Area in Small Area Estimation

Faisal Haris\(^1\), Azka Ubaidillah\(^2\)
{15.8610@stis.ac.id\(^1\), Azka@stis.ac.id\(^2\)}

Computational Statistic, Polytechnic of Statistics STIS 13330, Indonesia\(^1\)\(^2\)

Abstract. Small area estimation (SAE) is a statistical technique to predict the parameter of subpopulation with small or even zero sample size. An area with zero sample size can be estimated with the support of cluster information. The area random effect assumed has a similarity between region and can be analyzed by clustering the auxiliary variables. In SAE, Mean square error (MSE) is used to compare the precision of parameter estimates. But, there is no study that discuss the MSE of non-sampled area in SAE. The main idea of this research is to modify the existing statistical method by adding the cluster information with the assumption that there are similar characteristics of similar areas. The new method was evaluated by data simulation and case study to check the performance. The data simulation show that all modified methods produce a relatively similar MSE of non-sampled area.

Keywords: Clustering non sampled area, Mean Square Error, Small Area Estimation.

1 Introduction

According to Rao, Small Area Estimation (SAE) is a method to estimate the parameters of subpopulation with small sample size\(^9\). In this case, small area means an area that cannot be directly estimated because it can produce a very large standard error.

One of the most widely used SAE method is Empirical Best Linear Unbiased Prediction (EBLUP). The model was first applied by Fay and Herriot to estimate per capita income in small places based on survey data from American Census Bureau\(^8\). The EBLUP estimator used by them is a weighted average of the direct estimation and a regression estimation that obtained by fitting linear regression equation to the data. Fay-Herriot combine two main ideas of SAE in EBLUP, that is the combination of diversity of target variable can be explained by auxiliary variable and the specific area random effect that cannot be explained by auxiliary variable. Therefore, Fay-Herriot model is also commonly known as linear mixed model.

Because of the small sample size, usually there are several areas that do not have sample. Because of that, the direct estimation of that area cannot be estimated. Non sampled area can be estimated with only synthetic estimation or using cluster information that recently introduced by rahma annisa\(^1\). Gonzales stated that synthetic estimation is an indirect estimation that using variable characteristic of large sample area to estimate the variable of small sample size\(^5\). Synthetic estimation uses an assumption that an area with small sample size have a similar characteristic with large sample size area. However, synthetic estimation doesn’t take into account effect of random area. Non sampled area doesn’t have area random effect, so there will be bias in the estimation.
To overcome the bias in synthetic estimation, Rahma Annisa uses cluster information of the same characteristic with non sampled area to improve the precision of the estimation\(^1\). Rahma Annisa stated that random area effect have a similarity between areas and can be analyzed by clustering the auxiliary variable. The cluster information can be added to the model to improve the estimated result of non sampled area estimation. One of the proposed model by Rahma Annisa is to add the average of random area effect of the sampled area to the non sampled area synthetic estimation with the same cluster. By using this technique, the estimation has a smaller MSE and bias compared to synthetic estimation.

In research study, error can be very important because it can determine the validity of the model. One way to see the precision of the estimator is using Mean square error (MSE)\(^7\). MSE is the average difference in estimation and true value. Therefore, low MSE mean the estimation is close to true value of the variable. In SAE there are 3 approach that usually used to calculated MSE, there are Prasad-Rao Estimator\(^8\), Using the Jackknife concept by Jiang and Lahiri\(^6\) and with Bootstrap approad used in Butar and Lahiri\(^2\). The Prasad-Rao estimator is the most widely used method because it can be easily applied and usually the fastest to calculate. Meanwhile, Jackknife and Bootstrap approach uses resampling method and usually take longer to calculate. The resampling method is a nonparametric method in statistic that calculate the MSE by looking at the statistical change from the subsample that drawn from the sample itself. The most basic difference between bootstrap and jackknife is that bootstrap using resampling with replacement and jackknife using resampling without replacement. Therefore, the value of MSE from bootstrap tends to change every time a calculation is made. However, for a large repetition these change tend to shrink to a certain point. Meanwhile, for jackknife approach, the estimated value of MSE will be the same.

Until now, there is still no research that discussing the estimation of MSE for non-sampled area in SAE. Whereas, the calculation of MSE is very important in parameter estimation, because it can determine the precision of the model and the feasibility of the estimation result to be published. Therefore in this study we propose a calculation method to estimate the MSE of non sampled area in SAE using cluster information. The estimation of MSE will be carried out by utilizing the cluster information using three approach , namely Prasad-Rao Estimator, Bootstrap and Jackknife.

To evaluate the proposed model, we conduct a data simulation based on a reference research data that use cluster information to estimate non sampled area in SAE by Rahma Annisa\(^1\). After that, the developed method will be applied to estimate the average expenditure per capita of west java resident at the sub-district level based on data from March 2018 SUSENAS and 2014 podes that obtained from Badan Pusat Statistik (BPS).

## 2 Materials

We use 2 type of data in this research. The first data are simulation data that generated by computer. The generated data is based on the research study by rahma annisa\(^1\). But there are small modification in generating auxiliary variable for cluster and generating variance of direct estimation. For easy instruction in simulation study, you can see the flowchart below.
The second data is a case study from a real-world dataset. The direct estimation of sub-districts was obtained from March 2018 SUSENAS and only covered a portion of sub-districts in West Java. Meanwhile, PODES are used as auxiliary variables and as the basis for clustering information analysis.

3Method

The SAE model used in this paper is the EBLUP area level. This method is the development of Best Linear Estimation (BLUP) that introduced by Henderson. This model assumes that area-specific random effect and sampling error are identically and independent, or we can say that \( v_i \sim N(0, \sigma_v^2) \) and \( e_i \sim N(0, \Psi_i) \). \( \sigma_v^2 \) is area random effect variance and \( \Psi_i \) is direct estimation sampling variance. BLUP model that Henderson \((\hat{\theta}^H)\) proposed is:

\[
\hat{\theta}^H = z_i^T \hat{\beta} + \gamma_i (\hat{\theta}_i - z_i^T \hat{\beta})
\]  

Or

\[
\hat{\theta}^H = \gamma_i \hat{\theta}_i + (1 - \gamma_i) z_i^T \hat{\beta}
\]  

**Fig. 1.** Simulation Study Flowchart.
Where:

\[ \gamma_i = \frac{\sigma^2 \hat{b}_i^2}{(\Psi_i + \sigma^2 \hat{b}_i^2)} \]  

(3)

\[ \tilde{\beta} : \text{is Best Linear Unbiased Estimator (BLUE) from } \beta \text{ that can be estimated with} \]

\[ \tilde{\beta} = \tilde{\beta}(\sigma_v^2) = \left[ \sum_{i=1}^{m} \frac{z_i^2}{(\Psi_i + \sigma^2 \hat{b}_i^2)} \right]^{-1} \left[ \sum_{i=1}^{m} \frac{z_i \hat{\theta}_i}{(\Psi_i + \sigma^2 \hat{b}_i^2)} \right] \]  

(4)

\( \hat{\theta}_i \): direct estimation from area-i

\( z_i \): area-specific auxiliary variable

\( b_i \): random effect constants

\( v_i \): area-specific random effect

\( \Psi_i \): direct estimation variance

\( \sigma_v^2 \): area random effect

\( m \): number of area

From we can see that \( \tilde{\theta}_i^H \) is weighted average of direct estimation and synthetic estimation with \( \gamma_i \) as a weight. \( \gamma_i \) indicating how much the model affecting direct estimation. It was measured by comparing model variance random effect (\( \sigma_v^2 \hat{b}_i^2 \)) and total variance (\( \Psi_i + \sigma^2 \hat{b}_i^2 \)).

BLUP still using the assumption that variance random effect from the area are known. In real life data variances random effect is nearly impossible to compute, so we use estimation of variance random effect from samples data. By estimating variance random effect component from sample, it became Empirical Best Linear Unbiased Prediction (EBLUP). To estimate \( \sigma_v^2 \) there are several method such as Moment, Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML). In this study we use REML to estimate the variance of area random effect.

To estimate the estimation of non sampled area in small area estimation we use the first model of Rahma Annissa\(^1\). In this model we add the average area random effect to the synthetic estimation of non sampled area of the same cluster. The model is:

**Population model**:

\[ y_{ijk} = \beta x_{ijk} + v_i + e_{ijk} \]  

(5)

**Sampled area model**:

\[ \hat{y}_{ijk} = \beta x_{ijk} + \hat{v}_i \]  

(6)

**Non sampled area model**:
\[ y_{ijk} = \hat{\beta} x_{ijk} + \tilde{v}_{(k)} \]  

where:

\[ \tilde{v}_{(k)} = \frac{\sum_{i=1}^{m_k} \tilde{v}_i}{m_k} \]  

with \( m_k \) is number of area in cluster \( k \).

\( \hat{\beta} \) : model koefisien

\( \tilde{v}_i \) : random area effect from sample

\( x_{ijk} \) : auxiliary variabel

To estimate the MSE of non sampled area we use three method of MSE estimation. There is Prasad-Rao Estimator, Jackknife approach and Bootstrap approach. For each method we modify the formula to use cluster information in case there are no data in non sampled area.

The first and mainly used MSE estimation is Prasad-Rao Estimator. It was designed according to taylor series expansion. The formula for Prasad-Rao Estimator is:

\[ \text{MSE}(\hat{y}_{EBLP}) = g_{1i}(\tilde{\sigma}_v^2) + g_{2i}(\tilde{\sigma}_v^2) + 2g_{3i}(\tilde{\sigma}_v^2) \]  

Where:

\[ g_{1i}(\tilde{\sigma}_v^2) = \frac{\tilde{\sigma}_v^2 \psi_i}{(\psi_i + \tilde{\sigma}_v^2)} = \hat{\gamma}_i \psi_i \]  

\[ g_{2i}(\tilde{\sigma}_v^2) = (1 - \hat{\gamma}_i)^2 x_i^T \left( \sum_{i=1}^{m} \frac{x_i x_i^T}{(\psi_i + \tilde{\sigma}_v^2)} \right)^{-1} x_i \]  

\[ g_{3i}(\tilde{\sigma}_v^2) = \psi_i^2 (\psi_i + \tilde{\sigma}_v^2)^{-3} \mathcal{V}(\tilde{\sigma}_v^2) \]  

\[ \mathcal{V}(\tilde{\sigma}_v^2) = \text{asymptotic variance of } (\tilde{\sigma}_v^2) = 2m^{-2} \sum_{i=1}^{m} (\tilde{\sigma}_v^2 + \psi_i)^2 \]

Another approach to estimate MSE is using resampling method. There are 2 method that used to estimate MSE using resampling, namely using jackknife and bootstrap. The main different between bootstrap and jackknife is how to generate subsample from sample data. jackknife generate subsample with replacement, so the number of subsample is finite. Meanwhile bootstrap generate subsample without replacement, therefore the number of subsample is infinite and can be defined by researcher.
Jiang, Lahiri and Wan give an explanation on how to estimate MSE using jackknife. To estimate MSE jiang only use \(g_1(.)\) in Prasad-Rao estimator without \(g_2(.)\) and \(g_3(.)\). The method to estimate MSE using jackknife is:

**Step 1:** estimate \(\hat{\theta}_i\) and \(g_1(\hat{\sigma}_v^2)\) from sample data with the same formula from Prasad-Rao estimator.

**Step 2:** create subsample with taking out one sample and estimate \(\hat{\beta}_{-l}\) and \(\hat{\sigma}_v^2\) from every subsample.

**Step 3:** estimate \(\hat{\theta}_{i,-l}\) and \(g_1(\hat{\sigma}_{v,-l}^2)\) using parameter in step 2.

**Step 4:** estimate \(\hat{M}_{1i}\) to correcting bias \(g_1(\hat{\sigma}_v^2)\), with

\[
\hat{M}_{1i} = g_1(\hat{\sigma}_v^2) - \frac{m-1}{m} \sum_{l=1}^{m}(g_1(\hat{\sigma}_{v,-l}^2) - g_1(\hat{\sigma}_v^2))^2
\]

(14)

**Step 5:** estimate \(\hat{M}_{2i}\) with

\[
\hat{M}_{2i} = \frac{m-1}{m} \sum_{l=1}^{m}(\hat{\theta}_{i,-l} - \hat{\theta}_i)^2
\]

(15)

**Step 6:** estimate jackknife MSE for every area with

\[
MSE_{\text{jack}}(\hat{\theta}_i) = \hat{M}_{1i} + \hat{M}_{2i}
\]

(16)

Meanwhile, Butar and Lahiri modify jackknife method with bootstrap approach [2]. With an assumption asume \(v_i\) and \(e_i\) are normal and \(\sigma_v^2 > 0\), the method to estimate MSE using bootstrap is:

**Step 1:** create independent parameter subsample for every area with generating data :

\[
\hat{\theta}_{i,*} \sim N(\hat{\theta}_i, \hat{\sigma}_v^2)
\]

(17)

**Step 2:** generate \(\hat{\theta}_{i,*}\) as direct estimation of subsample with:

\[
\hat{\theta}_{i,*} \sim N(\hat{\theta}_{i,*}, \psi_i)
\]

(18)

**Step 3:** estimate \(\hat{\theta}_{i,*}^{\text{EBLUP}}\) using EBLUP with the same auxiliary information \((z_i^T)\) and varians \((\psi_i)\).

**Step 4:** repeat step 1 – 4. The more repetition, the estimation of MSE are more convincing.

**Step 5:** estimate MSE of Bootstrap with :

\[
MSE_{\text{Bootstrap}}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B}(\hat{\theta}_{i,*}^{\text{EBLUP}} - \hat{\theta}_i)^2
\]

(19)

**4 Discussion**

To estimate MSE of non sampled area we modify Prasad-Rao estimator, Jackknife and Bootstrap approach to cover non sampled area. For Prasad-Rao estimator we propose using
\( \tilde{\psi}_i(k), \tilde{g}_3(\tilde{\sigma}_v^2)(k) \) and \( \tilde{y}_i(k) \) with cluster information. So the Prasad-Rao estimator to estimate non sampled area is:

\[
MSE_{\text{prasad}}(\hat{\theta}_i) = g_{1i}(\tilde{\sigma}_v^2)(k) + g_{2i}(\tilde{\sigma}_v^2)(k) + 2g_{3i}(\tilde{\sigma}_v^2)(k)
\]  

(20)

where:

\[
g_{1i}(\tilde{\sigma}_v^2)(k) = \frac{\tilde{\sigma}_v^2\tilde{\psi}_i(k)}{(\tilde{\psi}_i(k) + \tilde{\sigma}_v^2)}
\]  

(21)

\[
g_{2i}(\tilde{\sigma}_v^2)(k) = (1 - \tilde{y}_i(k))^2x^T_iQx_i
\]  

(22)

\( \tilde{\psi}_i(k), \tilde{g}_3(\tilde{\sigma}_v^2)(k), \tilde{y}_i(k) \) is average from sampled area in the same cluster.

\( Q \): sampled area component \((\sum_{i=1}^m \frac{x_i^T}{(\tilde{\psi}_i + \tilde{\sigma}_v^2)})^{-1})\)

Meanwhile for Jackknife approach we use \( \tilde{\beta}_{\nu}^{2i-1} \) and \( \tilde{\sigma}_v^{2i-1} \) from sampled area to estimate \( \tilde{\theta}_{\nu} \) and \( g_{1i}(\tilde{\sigma}_v^{2i-1}) \) and the rest is the same with sampled area. The estimation of MSE with jackknife is:

\[
MSE_j(\hat{\theta}_i) = \tilde{M}_{1i} + \tilde{M}_{2i}
\]  

(23)

Where:

\[
\tilde{M}_{1i} = g_{1i}(\tilde{\sigma}_v^2) - \frac{m-1}{m} \sum_{i=1}^m (g_{1i}(\tilde{\sigma}_v^{2i-1}) - g_{1i}(\tilde{\sigma}_v^2))
\]  

(24)

\[
\tilde{M}_{2i} = \frac{m-1}{m} \sum_{i=1}^m (\tilde{\theta}_{\nu} - \tilde{\theta}_i)^2
\]  

(25)

For Bootstrap we modify \( \tilde{\psi}_i(k) \) so it was the average \( \tilde{\psi}_i \) of the sampled area. Meanwhile we use \( (\tilde{\sigma}_v^2) \) with the model from sampled area. To estimate MSE of non sampled area we use:

\[
MSE_b(\hat{\theta}_i) = B^{-1} \sum_{b=1}^B (\hat{\theta}_i^b - \theta_i)^2
\]  

(26)

Where:

\[
\theta_i \sim N(x_i^T\tilde{\beta}, \tilde{\sigma}_v^2)
\]  

(27)

\[
\hat{\theta}_i \sim N(\theta_i, \tilde{\psi}_i(k))
\]  

(28)

5 Result and Conclusion

5.1 Simulation Study

From simulation study we see the stability of our model to estimate MSE of non-sampled area. To see the similarity between each method we see the average of relative root
mean square error (RRMSE). From three generated non sampled domain, the average RRMSE are similar.

**Table 1.** Table captions should always be positioned above the tables.

<table>
<thead>
<tr>
<th>No Area</th>
<th>MSE sampled</th>
<th>Prasad-Rao Estimator</th>
<th>MSE with Jackknife</th>
<th>MSE with Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>15</td>
<td>1.043413</td>
<td>1.052440288</td>
<td>1.039885707</td>
<td>1.044885773</td>
</tr>
<tr>
<td>30</td>
<td>1.069324</td>
<td>1.077951128</td>
<td>1.064557369</td>
<td>1.069120909</td>
</tr>
<tr>
<td>45</td>
<td>1.166019</td>
<td>1.176403722</td>
<td>1.169906514</td>
<td>1.168643182</td>
</tr>
</tbody>
</table>

To see the similarity in each repetition we can see the plot of MSE. The plot for every non sampled MS are as below.

**Fig. 2.** RRMSE plot for Non-Sampled Area 1.

**Fig. 3.** RRMSE plot for Non-Sampled Area 2.
Fig. 4. RRMSE plot for Non-Sampled Area 3.

From every non-sampled area and repetition, every MSE method have a similar plot and every line are overlapping. From that graphic we can see that every repetition have a similar distribution. To be more convincing, we use a formal statistical test to test the similarity of MSE from its variance using one way ANOVA. From that test we get the p-value of the three RRMSE of non sampled area as 0.968, 0.914 and 0.922. From that result we can conclude that every method to estimate MSE for non sampled area are statistically similar to EBLUP MSE estimator if that area are sampled.

5.2 Case Study

We perform case study to see the compatibility of the model with real life data. To perform case study we estimate expenditure per capita from March 2018 SUSENAS data in subdistrict level in West Java. We choose this data because there are 19 non sampled area and 607 sampled area in this data. For auxiliary variable we use PODES 2014 data. PODES data is in village level, so we aggregate the data to subdistrict level.

For sampled area, there are 11 auxiliary variable that statistically significant for EBLUP model. and the estimation are as below.

<table>
<thead>
<tr>
<th>Subdistrict code</th>
<th>Expenditure pre capita</th>
<th>Prasad-Rao Estimator</th>
<th>MSE with Jackknife</th>
<th>MSE with Bootstrap</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>3202190</td>
<td>Rp 790,514.35</td>
<td>13395114081</td>
<td>13402255773</td>
<td>13734735710</td>
<td>1</td>
</tr>
<tr>
<td>3211080</td>
<td>Rp 905,511.76</td>
<td>19115033364</td>
<td>1913878226</td>
<td>18951811154</td>
<td>3</td>
</tr>
<tr>
<td>3202070</td>
<td>Rp 794,400.48</td>
<td>19119393348</td>
<td>19147932140</td>
<td>19048797402</td>
<td>3</td>
</tr>
<tr>
<td>3206211</td>
<td>Rp 753,198.95</td>
<td>19120100995</td>
<td>19134702882</td>
<td>18794678264</td>
<td>3</td>
</tr>
<tr>
<td>3204101</td>
<td>Rp 978,417.70</td>
<td>19121461423</td>
<td>19170674159</td>
<td>20205085319</td>
<td>3</td>
</tr>
<tr>
<td>3205111</td>
<td>Rp 851,626.83</td>
<td>19121575338</td>
<td>19153118463</td>
<td>19193686346</td>
<td>3</td>
</tr>
<tr>
<td>3205040</td>
<td>Rp 848,922.69</td>
<td>19122069993</td>
<td>19149418129</td>
<td>19872309725</td>
<td>3</td>
</tr>
<tr>
<td>3202172</td>
<td>Rp 889,735.21</td>
<td>19122696239</td>
<td>19161603845</td>
<td>18976891940</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 2. Non Sampled Estimation for Case Study (cont).

<table>
<thead>
<tr>
<th>Subdistrict code</th>
<th>Expenditure per capita</th>
<th>Prasad-Rao Estimator</th>
<th>MSE with Jackknife</th>
<th>MSE with Bootstrap</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>3215020</td>
<td>Rp 919,667.35</td>
<td>19123004208</td>
<td>19156091219</td>
<td>20355100030</td>
<td>3</td>
</tr>
<tr>
<td>3210091</td>
<td>Rp 917,150.88</td>
<td>19124542178</td>
<td>19167834947</td>
<td>18462639720</td>
<td>3</td>
</tr>
<tr>
<td>3202250</td>
<td>Rp 897,371.90</td>
<td>19125015909</td>
<td>19170041556</td>
<td>19013331238</td>
<td>3</td>
</tr>
<tr>
<td>3208051</td>
<td>Rp 884,793.36</td>
<td>19125406694</td>
<td>19163300009</td>
<td>17753384709</td>
<td>3</td>
</tr>
<tr>
<td>3205221</td>
<td>Rp 878,086.36</td>
<td>19127021461</td>
<td>19166702806</td>
<td>19312505357</td>
<td>3</td>
</tr>
<tr>
<td>3209191</td>
<td>Rp 976,594.53</td>
<td>19127636480</td>
<td>19179093627</td>
<td>19249803016</td>
<td>3</td>
</tr>
<tr>
<td>3208021</td>
<td>Rp 914,817.75</td>
<td>19127831613</td>
<td>19177453207</td>
<td>19404232539</td>
<td>3</td>
</tr>
<tr>
<td>3212162</td>
<td>Rp 915,476.67</td>
<td>19128116122</td>
<td>19177315367</td>
<td>19646260763</td>
<td>3</td>
</tr>
<tr>
<td>3206161</td>
<td>Rp 729,632.07</td>
<td>19133212871</td>
<td>19157604544</td>
<td>19684617982</td>
<td>3</td>
</tr>
<tr>
<td>3205161</td>
<td>Rp 810,902.42</td>
<td>19133986648</td>
<td>19169120918</td>
<td>20115871057</td>
<td>3</td>
</tr>
<tr>
<td>3214011</td>
<td>Rp 821,765.53</td>
<td>19138126464</td>
<td>19173534692</td>
<td>19257283193</td>
<td>3</td>
</tr>
</tbody>
</table>

Source: SUSENAS 2018 (Processed)

From that data we can see that the estimation of MSE of non sampled area are stable and similar between the three proposed method. The graphic to see the similarity of MSE can be seen as below.

![RRMSE Plot for Non Sampled Area in Case Study](image)

**Fig. 5.** RRMSE Plot for Non Sampled Area in Case Study.

To be more convincing, we test the similarity between every RRMSE with one way anova. From that test the p-value is 0.833. With that we conclude that the RRMSE of non-sampled area from every method in case study are similar.
5.3 Conclusion

From the discussion we can conclude that the three estimation method can be used to estimate MSE of estimator in non sampled area. The three estimation method that used in this study (Prasad Rao Estimator, Jackknife approach and Bootstrap approach) produce a similar result in simulation study and case study. But, Prasad Rao Estimator has an advantage in calculation time because it is not using resampling.
References