© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET21841116

Comparison of EBLUP and EBLUP Modification in Estimating Small Areas (Study : Percentages of Poverty in Bogor District)

Hary Merdeka, Kusman Sadik, Indahwati
Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT

A small area of the sample occurs when the sample size is very small. A large error will get if the parameters estimation is done with small the sample. One method to overcome it using a small area estimation (SAE) method. A small area estimator is a statistical technique to estimate the parameters of a sub-population with a small sample size. Estimates in the small area estimator method is based on the model and are indirect estimates. In this study the indirect method used is the EBLUP method and the modification of EBLUP estimator. The results of the alleged percentage of poverty in the Bogor district show that the EBLUP modification method is better compared to the expected method directly. This is based on the average of the RRMSE obtained.

Keywords: Small Area Estimation, EBLUP, Fixed-Effect, Random-Effect, EBLUP Modification, Percentage of Poverty

I. INTRODUCTION

The Statistic is a numerical quantity which is calculated from the sample. the sample is a collection of data collected from the population using the relevant method. One method for collecting data obtained by taking the sample from the population is a survey. In the process of collecting data for a small level, usually using a sample size that is relatively small, even in certain areas, it may not be sampled

Small data occur because the available data is not sufficient for estimating. Suppose that data on poverty in Indonesia is only adequate at the provincial, city or district level. However, at the sub-district or village level, the available data is very small, so to do statistical analysis with that data will produce a very large error. a very small sample size would have a large variety and could not even estimate when the area was not selected as an example unit

Small area estimation (SAE) can be regarded as a method for estimating parameters in a relatively small area in a pilot survey by utilizing information from outside the area, from within the area itself, and from outside the survey. The use of this information is an auxiliary variable which has a correlation with the variables observed.

The small area estimation method is based on the model and the small area estimation method is an indirect estimation. Therefore, additional information is needed from variables that have a relationship with the variable being observed which is called the accompanying variable. Small area estimators have several approaches, including Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB), and Hierarchical Bayes (HB). The EBLUP method is a technique for solving mixed-effect models that minimize the mean square error (MSE) generated by assuming a known variant component.

EBLUP is used to estimate linear area parameters. This model assumes that the regression parameters are constant but random intercepts. The random linear intercept model establishes a regression line to each area with the same slope but with a different intercept.

EBLUP estimates are negatively affected if intercept in some areas is much higher than other areas. To improve the accuracy, it is necessary to modify the EBLUP estimator based on the linear mixture model with one factor that has a fixed and random-effect. The EBLUP method and modification of EBLUP estimator is expected to be able to overcome the survey problems that have been explained so that they can accurately describe the poverty indicators P_0(percentage of poverty)

II. MATERIAL AND METHODS

A. MATERIAL

The data used in this application study is secondary data from the Central Statistics Agency (BPS), which is based on the 2013 National Socio-Economic Survey (Susenas) and the 2014 Village Potential Data (PODES). Bogor District consists of 40. There are 3 sub-districts in Bogor District who are not surveyed. The three sub-districts are Megamendung, Tanjungsari, and Parung Panjang sub-districts.

The response variable used in this study is the poverty indicator P_0 (percentage of poverty). The response variable used is the 2013 Susenas data, which is per capita expenditure data. Whereas the auxiliary variable in this study was obtained from PODES 2014 data. The accompanying variables used which were assumed to illustrate per capita expenditure in Bogor District were 17 variables with the description in Table 1 below.

Table 1 : The independent variables used are based on 2014 PODES data

Variable	Information
X_1	The proportion of villages with the main source of income in agriculture
X_2	The proportion of the number of villages with the main income sources in the processing industry
X_3	The proportion of the number of villages with the main source of income in the fields of big retail and restaurants
X_4	The proportion of villages with the main source of income in services
X_5	The proportion of state TA / RA / BA education levels
X_6	The proportion of private TA / RA / BA education levels
X_7	The proportion of the level of elementary school education
X_8	The proportion of the level of private elementary school education
X_9	The proportion of education in the senior high school
X_{10}	The proportion of education at the private senior high school
X_{11}	The proportion of levels of state vocational education
X_{12}	The proportion of the level of private vocational education
X_{13}	The proportion of doctor's office
X_{14}	The proportion of polyclinic and treatment center
X_{15}	The proportion of midwife's practice place
X_{16}	The proportion of posyandu
X_{17}	The proportion of pharmacies

B. METHOD

The step as follows:

1. Estimating the poverty indicator P_0 (percentage of poverty). It uses the direct estimation method

$$P_{\alpha i} = \frac{1}{n_i}\sum_{j=1}^{n_i} P_{\alpha ij}$$

Which:

$\alpha = 0$

$P_{\alpha ij} = \left(\frac{z-E_{ij}}{z}\right)^{\alpha} I\left(E_{ij} < z\right)$

$I\left(E_{ij} < z\right) = \begin{cases} 1; if\ E_{ij} < z\ or\ poor\ people \\ 0; if\ \ E_{ij} \geq z\ or\ not\ poor\ people \end{cases}$

z = The poverty line from BPS

E_{ij} = population per capita expenditure

i = $(1, \ldots, m)$, i = sub-district

j = $(1, \ldots, n)$, j = village

2. Variable selection with stepwise procces

3. Estimating percentage of poverty based on an estimation of EBLUP values. Estimating the EBLUP value using the model:

$$\hat{y}_i^{EBLUP} = \hat{\gamma}_i\, \bar{y}_i + (1-\hat{\gamma}_i)\ \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}$$

Which:

$\hat{\boldsymbol{\beta}}$: Vector of unknown regression parameters

$\bar{\mathbf{x}}_i^T$: The vector of the explanatory variable

$\hat{\gamma}_i$: Estimate of $\sigma_u^2/(\sigma_u^2 + \sigma_e^2)$

$\bar{y}_i$: Mean of sub-district sampling

4. Modify the EBLUP estimator with the process:

a. Identify the outlier of EBLUP with graphic

b. Divide the model into 2 parts, namely the fixed-effect model and random-effect. The area of the EBLUP estimator value that has outliers is assumed to be a fixed-effect (F) model and the remaining area as a random-effect (R) model. The models:

$$(F)\ \hat{Y}_i^{EBLUP} = \bar{x}_i^T\hat{\beta} + f_i\left(\hat{\bar{y}}_i - \hat{\bar{x}}_i^T\hat{\beta}\right)$$

$$(R)\ \hat{Y}_i^{BLUP} = (1-f_i)\left[\,\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}} + \gamma_i^w\left(\hat{\bar{y}}_i - \hat{\bar{\mathbf{x}}}_i^T\hat{\boldsymbol{\beta}}\right)\right] + f_i[\hat{\bar{y}}_i + (\bar{\mathbf{x}}_i^T - \hat{\bar{\mathbf{x}}}_i^T)\hat{\boldsymbol{\beta}}]$$

Which :

$f_i = n_i/N_i$

5. Evaluate the based on the measurement of accuracy estimation through the Relative Root Mean Squared Error (RRMSE) percentage

c. RRMSE Direct Estimation

$$\text{RRMSE}\ (\hat{Y}_i) = \frac{\sqrt{MSE(\hat{Y}_i)}}{\hat{Y}_i}$$

d. RRMSE EBLUP and EBLUP modification

$$\text{RRMSE}\ (\hat{Y}_i^{eblup}) = \frac{\sqrt{MSE(\hat{Y}_i^{eblup})}}{\hat{Y}_i^{eblup}}$$

III.RESULTS AND DISCUSSION

A. Direct Estimation

The total sample of households is in the Susenas data in Bogor District as many as 1,806 households while the total households in Bogor District are as many as 1245963 households. An example is available in the Susenas data to estimate the percentage of poverty is very small compared to the total number of households in Bogor District. Based on Susenas data the percentage of poverty in Bogor District is 7.944% using a simple random withdrawal method. This means that out of 1,245,963 households in Bogor District there are 98,984 households that are below the poverty line

B. Variable Selection

A very small example if you use the direct estimation method will produce a large error. Method to overcome this by using a small area estimation. One of the methods is Empirical Best Linear Unbiased Prediction (EBLUP). The basic assumption in estimating small areas is the diversity of response variables can be explained by the diversity of permanent influences or additional information on the accompanying variables. There were 17 candidates for the accompanying variables tested in the model. The selection of variables that will be used using the stepwise regression method.

Stepwise regression is one method to get the best model from a regression analysis. By definition is a combination of forward and backward methods, the first variable entered is the variable with the highest and significant correlation with the dependent variable, the second incoming variable is the variable with the highest correlation and still significant after certain variables enter the model then other variables the model is evaluated, if there is a variable that is not significant then the variable is issued. The used in selecting variables uses a stepwise regression method with $\propto$ =0.05. it indicates that the variable is significant

The selection of accompanying variables was carried out for each poverty indicator so that the accompanying variables that influenced the response variable were obtained. The accompanying variables that have been selected based on the stepwise regression method will add information to the response variable. The variable response is the percentage of poverty obtained based on direct estimates. The accompanying variables obtained using stepwise regression can be seen in Table 2

Table 2 : Stepwise Regression Percentage of Poverty

Predictor	Coefisien	P-Value	VIF
Constant	-5.97	0.058	
X_1	147.6	0.001	1.14
X_8	67.3	0.003	1.1
X_{12}	-90.6	0.022	1.05
X_{14}	21.65	0.016	1.2

The accompanying variables are used to add information to the direct estimates (response variables) of the percentage of poverty as follows:

1. X_1 = The proportion of villages with the main source of income in agriculture
2. X_8 = The proportion of the level of elementary school education
3. X_{12} = The proportion of the level of private vocational education.
4. X_{14} = The proportion of polyclinic and treatment center

Selected variables that are selected based on variables that have a significant effect on the response variable are four variables. The accompanying variable is used for the EBLUP method and EBLUP modification.

The four variables that have been selected will be used to be the accompanying variables. Of the four variables, one variable that reduces the percentage of poverty is the variable X_{12} (the proportion of levels of private vocational education). This is based on the value of the variable coefficient X_{12}which is negative. The value is interpreted if more and more residents with private SMK graduates in Bogor District, the percentage of poverty in Bogor District will also decrease and vice versa. Whereas there are 3 accompanying variables that have positive coefficient values, namely variable X_1 (proportion of villages with the main source of income in agriculture), X_8 (proportion of levels of private elementary), and X_{14} (proportion of polyclinics and treatment centers)

The X_1 variable based on the coefficient value (Table 2) which has a positive value can be interpreted if the proportion of villages with the main income in agriculture is increasing in Bogor District, then the percentage of poverty in Bogor District also increases. Variable X_8 can be interpreted if the proportion of the education level of the population in Bogor District with private SD / MI graduates increases, then the percentage of poverty in Bogor District also increases. The X_{14} variable is interpreted if the proportion of polyclinics / medical centers increases, the percentage of poverty also increases and vice versa.

The accompanying variables chosen were also in line with the research conducted by Erwan (2007). The number of poor people in Indonesia is mostly in rural areas where the majority of the population is from poverty as farmers. Most poor people have low education. Nearly 50% of poor people do not pass elementary school so that more and more residents who work as farmers and education levels are only elementary schools, the higher the percentage of poverty in the area. While the population with a level of private vocational education has a low percentage of poverty so that it can be interpreted that the poverty rate will decline if more and more graduates from private vocational schools.

A number of studies have shown a link between poverty and health. Health is a condition that is able to create the potential that exists in society to be more optimal, both physically and socially. Poverty is a factor that greatly inhibits efforts to create such conditions. Therefore, the higher the poverty rate, the worse the health conditions (Sunyoto et al., 2007). Based on research from Sunyoto, it can be identified that if an area of poverty is very high, the health level is also low so that there will be more polyclinics / medical centers built by the government.

C. Direct Estimation Method and EBLUP

The results of the estimation of the percentage of poverty in Bogor District using the direct estimation method and the EBLUP method can be seen in Table 3

Table 3: Estimation of indicators poverty in Bogor District using EBLUP method

No	Sub-District	Hous-ehold	Direct (%)	EBLUP (%)
1	Nanggung	20	45.000	21.203
2	Leuwiliang	39	12.821	12.504
3	Leuwisadeng	30	26.667	17.667
4	Pamijahan	58	5.172	12.469
5	Cibungbulang	28	0.000	8.635
6	Ciampea	20	0.000	6.694
7	Tenjolaya	9	0.000	12.158
8	Dramaga	19	10.526	5.515
9	Ciomas	28	3.571	3.972
10	Tamansari	29	6.897	7.580
11	Cijeruk	8	25.000	12.678
12	Cigombong	17	0.000	7.269
13	Caringin	28	3.571	13.443
14	Ciawi	25	12.000	8.918
15	Cisarua	28	0.000	4.759
16	Sukaraja	44	2.273	9.625
17	Babakan Madang	28	0.000	5.514
18	Sukamakmur	16	0.000	11.740
19	Cariu	10	10.000	14.088
20	Jonggol	8	0.000	11.199
21	Cileungsi	44	0.000	4.269
22	Kelapa Nunggal	30	3.333	10.560
23	Gunung Putri	48	2.083	9.359
24	Citeureup	66	7.576	10.772
25	Cibinong	89	0.000	1.428
26	Bojong Gede	55	0.000	1.000
27	Tajur Halang	46	6.522	3.882
28	Kemang	20	20.000	10.161
29	Ranca Bungur	8	12.500	11.921
30	Parung	27	0.000	5.081
31	Ciseeng	8	12.500	10.875
32	Gunung Sindur	19	0.000	5.602
33	Rumpin	47	14.894	12.042
34	Cigudeg	36	25.000	16.993
35	Sukajaya	24	16.667	19.391
36	Jasinga	30	6.667	10.259
37	Tenjo	37	2.703	10.191
	Mean		7.944	9.768

There are sub-districts where the examples are very small and there are no examples at all so that to predict the poverty indicators in some of these sub-districts produces an estimated value of 0 using the direct estimation method. This indicates that there is no poverty in the sub-district. Allegations with direct estimation methods are worth 0 as many as 13 sub-districts, including Cibungbulang District, Ciampea, Tenjolaya, Cigombong, Cisarua, Babakan Madang, Sukamakmur, Jonggol, Cileungsi, Cibinong, Bojong Gede, Parung, and Gunung Sindur.

The percentage of poverty in Bogor District with the direct estimation method is 7.944%, which means that from 1245963 households in Bogor District there are 98984 households that are below the poverty line. Whereas by using the EBLUP method the percentage of poverty is 9,768% or 121,706 households that are below the poverty line. There is a difference of 1.823% estimated percentage of poverty in Bogor District or equal to 22722 households.

The biggest percentage of poverty using the direct estimation method and the EBLUP method is in Nanggung District. The direct estimation method produces a percentage of poverty in Nanggung Subdistrict by 45% while the EBLUP method produces an estimate of 21.203%

Alleged percentages of poverty in each sub-district have resulted in outliers. Therefore the EBLUP modification is used to overcome the outlier of the alleged EBLUP.

D. EBLUP modification

Modifications to EBLUP are based on outliers of alleged poverty indicators. Districts which are outliers will be used as new datasets in modeling. The suspected EBLUP is detected using a plot. Outliers for each sub-district in Bogor District can be seen in Figure 1.

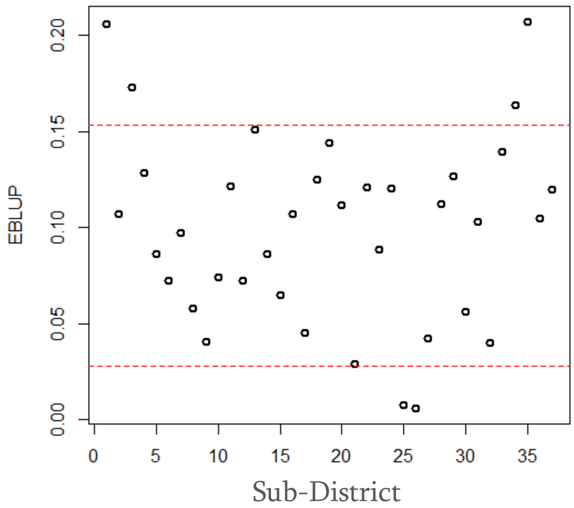


Figure 1

The outage of poverty is outside the dashed red line. Based on Figure 1 there are 6 sub-districts assumed to be outliers, namely Nanggung, Leuwisadeng, Cibinong, Bojong Gede, Cigudeg, and Sukajaya Subdistricts.

The EBLUP modification process will be divided into 2 models, Model-1 and Model-2.

1. Model-1

Model-1 assumes all sub-districts in Bogor District as fixed effects. Model-1 does not assume there is an outlier in the EBLUP estimates so that outliers in the alleged EBLUP have no effect on Model-1

2. Model-2

Each sub-district in Bogor District is divided into 2 parts, namely sub-districts which are outliers and sub-districts that are not outliers. Districts that are outliers are assumed to be fixed effects and sub-districts that are not outliers are assumed to be random. Districts which are outliers of the percentage of poverty are Nanggung, Leuwisadeng, Cibinong, Bojong Gede, Cigudeg, and Sukajaya Subdistricts. While the subdistrict dataset is not outliers using the EBLUP method. The expected results of the two datasets are accumulated into one dataset.

E. Direct Estimation, EBLUP, and Modification of EBLUP

Comparison of Alleged Direct, EBLUP, and Modified EBLUP Methods in estimating the percentage of poverty can be seen in Table 3

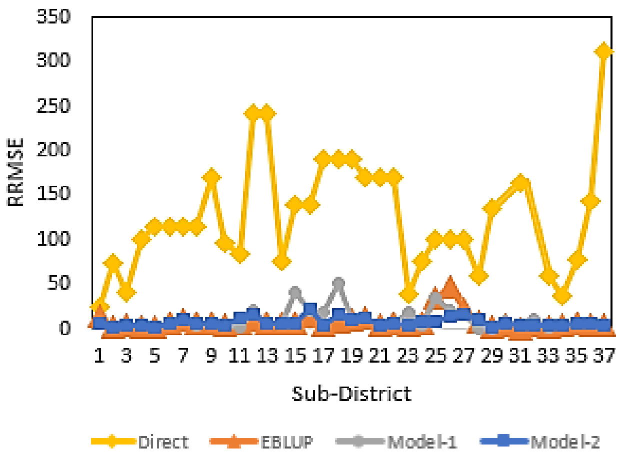
Table 3: Alleged Poverty Percentages with Direct Estimation Method, EBLUP, Model-1 and Model-2

No	Sub-District	Direct (%)	EBLUP (%)	Model-1 (%)	Model-2 (%)
1	Nanggung	45.000	21.203	40.927	36.498
2	Leuwiliang	12.821	12.504	13.269	9.278
3	Leuwisadeng	26.667	17.667	27.276	25.495
4	Pamijahan	5.172	12.469	7.673	7.101
5	Cibungbulang	0.000	8.635	2.874	6.412
6	Ciampea	0.000	6.694	1.923	5.219
7	Tenjolaya	0.000	12.158	2.391	6.991
8	Dramaga	10.526	5.515	8.482	6.611
9	Ciomas	3.571	3.972	3.638	4.702
10	Tamansari	6.897	7.580	7.899	6.665
11	Cijeruk	25.000	12.678	23.820	9.899
12	Cigombong	0.000	7.269	4.532	5.479
13	Caringin	3.571	13.443	8.038	7.325
14	Ciawi	12.000	8.918	8.430	6.463
15	Cisarua	0.000	4.759	0.084	4.265
16	Sukaraja	2.273	9.625	7.367	6.377
17	Babakan	0.000	5.514	1.057	4.127
18	Sukamakmur	0.000	11.740	2.482	6.775
19	Cariu	10.000	14.088	7.742	9.538
20	Jonggol	0.000	11.199	5.990	7.400
21	Cileungsi	0.000	4.269	2.774	3.972
22	Kelapa N	3.333	10.560	5.785	6.117
23	Gunung Putri	2.083	9.359	1.114	10.298
24	Citeureup	7.576	10.772	11.206	7.649
25	Cibinong	0.000	1.428	0.145	5.307
26	Bojong Gede	0.000	1.000	2.083	3.009
27	Tajur Halang	6.522	3.882	5.086	5.406
28	Kemang	20.000	10.161	20.595	8.560
29	Ranca B	12.500	11.921	11.338	8.709
30	Parung	0.000	5.081	3.440	4.241
31	Ciseeng	12.500	10.875	14.923	7.811
32	Gunung S	0.000	5.602	1.301	5.198
33	Rumpin	14.894	12.042	14.033	9.708
34	Cigudeg	25.000	16.993	23.563	24.968
35	Sukajaya	16.667	19.391	19.580	15.078
36	Jasinga	6.667	10.259	7.343	8.133
37	Tenjo	2.703	10.191	5.471	6.627
	Rataan	7.944	9.768	9.072	8.741

The lowest average percentage of poverty is generated using the direct estimation method of 7.944%. The highest percentage of poverty uses the EBLUP method of 9,768%. The difference in the estimated results between the direct estimation method and the EBLUP method is 1.824%, between direct estimates with Model-1 of 1.555%, and between direct estimates with Model-2 of 0.797%. The method that produces the closest value is the direct estimation method using Model-2

F. Empirical Evaluation

Empirical evaluation is used to compare the measure of goodness from 3 methods, namely direct estimation, EBLUP, and EBLUP modification using RRMSE. The comparison of the ARRMSE value of the estimated percentage of poverty using can be seen in Figure 2



The highest RRMSE direct estimation method value compared to other methods (Figure 2). RRMSE in the direct estimation is missing in some sub-districts because there is a sub-district with a poverty percentage of 0. While the value of Model-1 and Model-2 RRMSE is relatively the same as the EBLUP method because the graph shows lines from EBLUP, Model-1, and Model-2

To see a comparison of methods that are more accurate, it will be compared through the average RRMSE (ARRMSE). Calculation and comparison of ARRMSE values for each method can be seen in Table 4.

Table 4 : ARRMSE Results of Alleged Poverty Percentage

No	Method	ARRMSE
1	Direct	78.466
2	EBLUP	6.769
3	Model-1	7.000
4	Model-2	6.736

The best method used for percentage of poverty based on the average value is using Model-2 (Table 4). While using the direct RRMSE estimation method that is produced is very large when compared with other methods.

IV.CONCLUSION

The method used is the direct estimation method, EBLUP method, and EBLUP modification. Modifications to EBLUP are carried out through two methods namely Model-1 and Model-2.

The nested regression model has a random and fixed intercept for estimating linear parameters from a small area. This is the basis for modifications to the EBLUP method. The EBLUP method assumes the effect of random areas is random. Modifications to EBLUP are carried out through two methods namely Model-1 and Model-2. Model-1 assumes that all areas have a fixed influence and Model-2 assumes that the area is the outlier as a fixed influence while the area that is not the outlier is assumed to be a random influence. Model-1 uses a regression model with the sub-district as a dummy variable. After that, the calculation is done based on the model.

Based on the ARRMSE values obtained there are differences in the results of the three methods. After modifications to EBLUP, Model-2 is better at predicting poverty indicators. In general, it can be concluded that the modification of the EBLUP estimator results in a lower RRMSE value than the direct estimation method and the EBLUP.

V. REFERENCES

[1] [BPS] Badan Pusat Statistik. 2013. Data dan informasi kemiskinan kabupaten/kota tahun 2013. Jakarta (ID): BPS.

[2] [BPS] Badan Pusat Statistik. 2016. Perhitungan dan analisis kemiskinan makro indonesia tahun 2016. Jakarta (ID): BPS.

[3] Anisa R, Kurnia A, Indahwati. 2014. Cluster information non-sampled area in small area

estimation. IOSR Journal of Mathematics. 10(1): 15-19.

[4] Ferreti C, Molin, I. 2012. Fast EB for estimating complex poverty indicators in large populations. Journal of the Indian Society of Agricultural Statistics, 66 (1): 105 -120.

[5] Chandra H, Sud UC, Gharde Y. 2015. Small Area Estimation Using Estimated Population Level Auxiliary Data. J Communications in Statistics-Simulation and Computation, 44:5, 1197-1209. DOI: 10.1080/03610918. 2013.810255

[6] M.Herrador, M. D Esteban, T. Hobza, 2013, D. Morales. A Modified Nested-Error Regression Model for Small Area Estimation, Vol.47, No.2,258-273, http://dx.doi.org/

[7] Liu, H., Shah, S., Jiang, W. (2004), "On-line pencilan detection and data cleaning," Computers and Chemical Engineering, 28, 1635–1647.

[8] Menteiga-Gonzales. (2008) Bootstrap Mean Squared Error of a Small-Area EBLUP, J Communications in Statistics-Simulation and Computation, 78:5, 443-462. DOI: 10.1080/030949650601141811.

[9] McCulloch CE, Searle SR. 2001. Generalized, Linear and Mixed Models. New York: John Wiley & Sons, Inc.

[10] Namazi-Rad MR, Steel D. 2015. What Level of Statistical Model Should We Use in Small Area Estimation.Australian & New Zealand Journal of Statistics 57 (2) :275-298

[11] Rao JNK, Molina I. 2015. Small Area Estimation second edition. New York: Wiley.

[12] Sadik K. 2009. The best linear unbiased prediction method and hierarchical bayes for estimating small areas based on state space models [dissertation]. Bogor (ID): Institut Pertanian Bogor.

[13] V.Y Sundara, Sadik K, Kurnia A, Cluster information of non-sampled area in small area estimation of poverty indicators using Empirical Bayes. AIP Conference Proceedings 1827, 020026 (2017); doi: 10.1063/1.4979442

[14] Ybarra LMR, Lohr SL. 2008. Small Area Estimation when Auxiliary Information Measured with Error. Biometrika 95(4): 919-931