Efficiency of Several Complex Survey Design using EBLUP in Small Area Estimation

Nadra Yudelsa Ratu[1], Ika Yuni Wulansari[2]
{nadra.ratu@bps.go.id[1], ikayuni@stis.ac.id[2]}

Statistic Indonesia, Jakarta[1], Polytechnic of Statistics STIS, Jakarta[2]

Abstract. Thedissemination of data from the survey was carried out by estimating the parameter of the survey results. The implementation of the survey at BPS now is getting more complex where direct estimation results are presented into small areas. However, the sample size of direct estimation in small area has a relatively small size so that it is not reliable enough, not efficient and has low precision. Therefore, other statistics are needed that can accommodate the dissemination from total household expenditure data in the small area. In this study of small area, it was carried out by applying the Small Area Estimation (SAE) method, which is Empirical Best Linear Unbiased Prediction (EBLUP) by involving a complex survey design. The sampling method in complex survey design that used are Simple Random Sampling Without Replacement called SRSWOR, One Stage Cluster (SRSWOR), Two Stage Cluster (SRSWOR-SRSWOR) and Two Stage Cluster (Probability Proportional to Size called PPSWR-SRSWOR). The efficiency of estimation result is evaluated based on MSE and RRMSE values that obtained in each method of the complex survey design. According to the calculation results, the largest MSE and RRMSE value of the estimation was obtained from Two Stage Cluster (SRSWOR-SRSWOR) sampling method. Besides, the smallest MSE and RRMSE value was obtained from the SRSWOR sampling method that seem to have distinct advantage over the other sampling method.

Keywords: Complex Survey, EBLUP, Efficiency, SAE

1 Introduction

Data is a piece of information that provides an overview of situation interests that happened such as in a particular unit or area. The accuracy and correctness of data and information are needed to obtain the actual situation that has occurred in the area. It was getting from the real data source by conducting such as the survey activity. Survey is an activity of collecting data and information from elements or members of the population. Sample survey data are extensively used to disseminate the reliable direct estimation of totals or means for the whole population and area [10]. The resulted of the survey have long been recognized as cost-effective means to get information of area by using the sampling method from a specific domain that derived from the survey. It was referred to direct estimation by using the design-based approach. Design-based estimation involves the weighting values from sample survey and inferencing values such as standard error based on the distribution of probability from the sample designed by using the appropriate sampling techniques [9].

Nowadays, the implementation of the survey at BPS is getting more complex where direct estimation results are presented into small areas. However, the sample size of direct estimation

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290476

in small area has a relatively small size so that it is not reliable enough, not efficient and has low precision[11]. It also can't present the small domain estimation to estimate the value of parameter and produce the large variance. If the domain sample size tends to increase, it will cause the implementation of survey cost become more expensive and non-sampling error will be larger so that the survey becomes inefficient. These direct estimates frequently fail to provide reliable estimates for small areas due to very small sample sizes in the areas [15].

Small Area Estimation (SAE) is one of the technique of statistics that used to estimate the result of estimation of parameter with relatively small variance and reliable [9]. SAE used the model that determines the other related data included in the process of estimation. The presentation of result from SAE is done by combining the result from direct estimation and other related data to produce indirect estimation with a model-based approach. This indirect estimation is more representative because of the increasing precision, efficiency and decreasing the number of variance and provides the more accurate parameter estimation [9]. Therefore, the problem of direct estimation can be overcome by developing the method to estimate the parameter of estimation using SAE method [6].

The estimation of the SAE method is done by borrowing the strength from auxiliary variable from a large domain such as census data or administrative data to predict variables that become interest in a small domain. The estimation using SAE has a higher level of precision because the sample that used in the estimation drawn from a population randomly[2]. In additionally, the adequacy of precision in SAE can be determined by performing a bootstrap resampling simulation on different sample sizes [3]. Unfortunately, the application of SAE still focused on the use of assumptions with SRS sampling [1]. It assumed that each selected unit has the same opportunities and having no consideration of sampling weight in a survey complex design.

In this study of small area, it was carried out by applying the Small Area Estimation (SAE) method. The method of SAE that used in this research is Empirical Best Linear Unbiased Prediction (EBLUP) by involving the complex survey design. The sampling method in complex survey that used are Simple Random Sampling Without Replacement called SRSWOR, One Stage Cluster (SRSWOR), Two Stage Cluster (SRSWOR-SRSWOR) and Two Stage Cluster (Probability Proportional to Size called PPSWR-SRSWOR). This method is used to study the efficiency of SAE method such as EBLUP in a complex survey design by using the bootstrap resampling simulation on different sample sizes. The methods were applied on total household expenditure in East Java Province in 2018 as one of the most important indicators in measuring poverty.

2 Materials

The method of complex survey design by using EBLUP SAEwere simulated at the raw data from the expenditure of households in 2018 from Survei Sensus Ekonomi Nasional (SUSENAS) 2018 and a few variables from Potensi Desa (PODES) 2014 at East Java Province hold by BPS.The basic raw of data consist of distribution data that can be seen in **Figure 1**on histogram format.

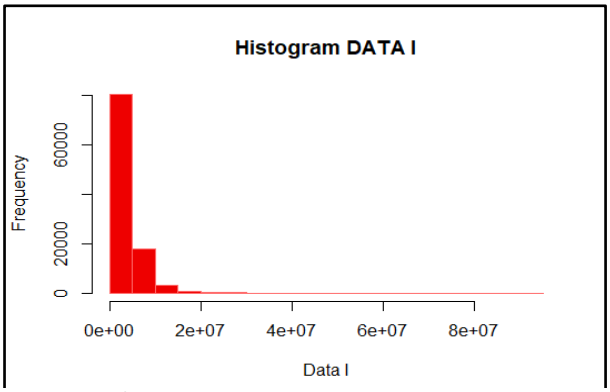

Fig.1. Histogram of raw data household expenditure in East Java Province

The expenditure of households is the cost incurred for the consumption of all household members from purchases, gifts, and self-production for a month[3]. The household consumption is divided into food and non-food consumption without regarding the origin of the goods and it's limited to expense for household needs that not include consumption or expenditure for business purposes or given to other parties. The expenditure of food consumption was calculated during the past week, while for non-food was calculated during the past month and 12 months. Both of food consumption and non-food consumption are subsequently converted into an average expenditure of one month.

The analysis of complex survey design is constructed under-sampling design by following three steps. First, calculate the estimation of expenditure household using EBLUP method. In this step, it compared the result between direct estimation that publishing by Statistics Indonesia and EBLUP method. Then, perform the parameter estimation by applying the EBLUP method and the four existing survey designs with the application of resampling method. After that, compare the obtaining results to see the efficiency of result estimation.

2.1 Direct Estimation

2.1.1Simple Survey Design

Direct estimation is an estimation obtained from the calculation of data from the implementation of survey in the area from the observation object [10]. It was generally used as a design-based approach. Direct estimation is obtained by estimating the total sample used by the survey design. This estimation is done by involving the weight or sampling fraction from the survey design to calculate the parameter values of the population. One of direct estimation method that used is Simple Random Sampling (SRS). SRS is a sample selection method by using the theory of probabilistic where the opportunities of each population unit have equal probability of being selected [12]. Direct estimation that used is total estimation of SRS Without Replacement (SRSWOR). It can be calculated by using this following formula:

$$\hat{Y} = N \times \frac{\sum_{i=1}^{n} y_i}{n}. \quad (1)$$

$$v(\hat{Y}) = N^2 \times \frac{(1-f)}{n} \times \frac{1}{n-1} \times \sum_{i=1}^{n} (y_i - \bar{y}_0)^2 . \quad (2)$$

where n is the total of sample, N is the total of population, f is the sampling fraction, y_i is the value of observation unit at element i, and $\bar{y}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$.

2.1.2Complex Survey Design

Complex survey design is a sample survey that carried out by using the disproportionate sampling probabilities and divided into sub-population such as strata or cluster. The advantage of using the complex survey design is reducing the cost and make the time of collecting data more efficient. Besides, complex survey design allowed to access the sample framework that difficult to obtain while using simple survey design [4]. Sampling cluster is one of the method of sampling where the selected sampling unit is a collection of elements that come from the group called cluster. Generally, the elements of same cluster are heterogeneous while in different cluster are homogeneous. Cluster sampling consists of two types that are one stage cluster and multistage cluster. In the two-stage cluster, the first cluster stage is Primary Sampling Unit (PSU) that selected the -n sampling cluster from the -N population cluster randomly. The second stage is called as Secondary Sampling Unit (SSU) which is selected as -m random elements from the -M population elements in each cluster [13].

One Stage Cluster (SRSWOR)

The estimation of total and variance from one stage cluster (SRSWOR) can be formulated:

$$\hat{Y} = N\bar{M} \times \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i}. \quad (3)$$

$$v(\hat{Y}) = N^2 \bar{M}^2 \times \frac{(1-f)}{n} \times \frac{1}{n-1} \times \frac{\sum_{i=1}^{n} {M_i}^2 (\bar{y}_{i.} - \bar{y}_n)^2}{\bar{M}^2}. \quad (4)$$

where $\bar{y}_i.$ is the average of unit observation in every cluster -i, M_i is the total of element from observation unit in every cluster -i, $\bar{M} = \frac{\sum_{i=1}^{n} M_i}{n}$ and $\bar{y}_n = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i}$.

Two Stage Cluster (SRSWOR-SRSWOR)

The estimation of total and variance from two stage cluster (SRSWOR-SRSWOR) can be formulated as:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i. \quad (5)$$

$$v(\hat{Y}) = N^2 (1 - f_1) \frac{s_b^2}{n} + \frac{N}{n} \sum_{i=1}^{n} {M_i}^2 (1 - f_{2i}) \frac{s_{wi}^2}{m_i}. \quad (6)$$

where

$\bar{y}_i. = \frac{1}{m_i} \sum_{i=1}^{m_i} y_{ij}$, $s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (M_i \bar{y}_i - \frac{1}{n} \sum_{i=1}^{n} M_i \bar{y}_i)^2$ and $s_{wi}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$.

Two Stage Cluster (PPSWR-SRSWOR)

The estimation of total and variance from two stage cluster (PPSWR-SRSWOR) can beformulated as:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{y}_i}{p_i}. \quad (7)$$

$$v(\hat{Y}) = \frac{\sum_{i=1}^{n}(\frac{M_i \bar{y}_i}{p_i} - \hat{Y})^2}{n(n-1)}.(8)$$

where $\bar{y}_i. = \frac{1}{m_i}\sum_{i=1}^{m_i} y_{ij}$, $p_i = \frac{M_i}{M_0}$ and $M_0 = \sum_{i=1}^{N} M_i$.

The understanding of variability averagein the result of data, comes from the number ofsampling process and non-sampling processcalled mean square error. The mean square error (MSE) is obtained from the number of quadratic bias andthe variance of eachsampling design. The value focuses on the variability of sampling design to see the power and strength of sampling design.It takes that the result of direct estimation is getting under the unbiased assumptionwhere the value of MSE ($\hat{Y}_i$)will be same as $v(\hat{Y}_i)$[14].

2.2Small Area Estimation

2.2.1Small Area Model (Basic Area Level)

Let $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \dots, \boldsymbol{x}_{pi})^T$ is the auxiliary variable with the estimated parameter that is assumed to be a function of response variable y and assumed to have a relationship with $\boldsymbol{x}_i$ [10]. The auxiliary variable has been used to build a linear model that is:

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_i v_i, \quad i = 1, \dots, m.(9)$$

where $b_i$is known as positive constants and β is the regression coefficients in $p \times 1$. Then, v_i is a random area effect that assumed to be independent and identically distributed. The model from equation (9) is combining with the model that comes from sampling process in each area, that is:

$$\hat{Y}_i = g(\bar{Y}_i) = Y_i + e_i, i = 1, .., m . \quad (10)$$

The sampling error e_i are independent with $E_p(e_i|Y_i) = 0$and $V_p(e_i|Y_i) = \psi_i$. From the two models, it overcomes the mixed model known as Fay Herriot model. Fay Herriot model has the following equation:

$$\hat{Y}_i = \boldsymbol{x}_i{}^T \boldsymbol{\beta} + b_i v_i + e_{i,} \; i = 1, \dots, m . \quad (11)$$

From the model,$v_i$istherandom area effect and$e_i$is sampling error from sampling process wherev_i $iid \sim N(0, \sigma_v^2)ande_i \sim N(0, \psi_i)$.

2.2.2Empirical Best Linear Unbiased Prediction (EBLUP)

BLUP (Best Linear Unbiased Prediction) is an estimation of parameter that minimizes the mean square error (MSE) in the other class of unbiased parameter by assuming the variance component of random effect is known [9]. However, the values from the variance component of random effect are difficult to determine so that the estimation of variance component is estimated by using the data from result of sampling [8]. The method that used to estimate the unknown of variance component called Empirical Best Linear Unbiased Prediction (EBLUP) method. The estimation of EBLUP in the area level model is written by:

$$\hat{Y}_{iEBLUP} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i(\hat{Y}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}).(12)$$

where:
x_i = auxiliary variable in area$i$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{MSE(\hat{Y}_i) + \hat{\sigma}_v^2}$$

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\sigma}_v^2) = \left[\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{(MSE(\hat{Y}_i) + \hat{\sigma}_v^2)}\right]^{-1} \left[\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \hat{Y}_i}{(MSE(\hat{Y}_i) + \hat{\sigma}_v^2)}\right]$$

The calculation of $\hat{\gamma}_i$ in EBLUP is getting from the value of MSE in direct estimation and random effect variance ($\hat{\sigma}_v^2$). If the value from $\hat{\sigma}_v^2$ is very big, the values of $\hat{\gamma}_i$ close to 1 so the estimation tends to direct estimation. On otherwise, if the value from $\hat{\sigma}_v^2$ is very small, the values of $\hat{\gamma}_i$ close to 0 so the estimation tends to synthetic estimation. In addition, if the value of MSE ($\hat{Y}_i$) is very big, the values of $\hat{\gamma}_i$ close to 0 so the estimation tends to synthetic estimation. On otherwise, if the value from MSE ($\hat{Y}_i$) is very small, the values of $\hat{\gamma}_i$ close to 1 so the estimation tends to direct estimation.

In EBLUP studying and literature, the random effect variance ($\hat{\sigma}_v^2$) can be approximate with used the restricted maximum likelihood (REML) method by applying the iteration from fisher scoring by setting:

$$\hat{\sigma}_v^{2(a+1)} = \hat{\sigma}_v^{2(a)} + \left[\mathfrak{I}\hat{\sigma}_v^{2(a)}\right]^{-1} s(\tilde{\boldsymbol{\beta}}^{(a)}, \hat{\sigma}_v^{2(a)}) \quad (13)$$

where:

$$\mathfrak{I}(\hat{\sigma}_v^2) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{(\hat{\sigma}_v^2 + MSE(\hat{Y}_i))^2}$$

$$s(\tilde{\boldsymbol{\beta}}^{(a)}, \hat{\sigma}_v^{2(a)}) = -\frac{1}{2}\sum_{i=1}^{m} \frac{1}{\left(\hat{\sigma}_v^2 + MSE(\hat{Y}_i)\right)^2} + \frac{1}{2}\sum_{i=1}^{m} \frac{(\hat{Y}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2}{(\hat{\sigma}_v^2 + MSE(\hat{Y}_i))^2}$$

The best model from EBLUP in basic area level can be determined from the result of estimation by using the estimation of MSE with the following formula:

$$MSE\left(\hat{Y}_{iEBLUP}\right) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \quad (14)$$

where:

$$g_{1i}(\hat{\sigma}_v^2) = \hat{\gamma}_i MSE(\hat{Y}_i)$$

$$g_{2i}(\hat{\sigma}_v^2) = (1 - \hat{\gamma}_i)^2 \boldsymbol{x}_i^T \left[\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{(MSE(\hat{Y}_i) + \hat{\sigma}_v^2)}\right]^{-1} \boldsymbol{x}_i$$

$$g_{3i}(\hat{\sigma}_v^2) = \left[MSE(\hat{Y}_i)\right]^2 \left(MSE(\hat{Y}_i) + \hat{\sigma}_v^2\right)^{-3} \bar{V}(\hat{\sigma}_v^2)$$

Besides that, the estimation of MSE in EBLUP involved the asymptotic variance from random effect variance that symbolized with $\bar{V}(\hat{\sigma}_v^2)$. The formulation for this asymptotic variance is given by $\bar{V}(\hat{\sigma}_v^2) = 2m^{-2}\sum_{i=1}^{m}(MSE(\hat{Y}_i) + \hat{\sigma}_v^2)^2$.

2.3Relative Root Mean Square Error (RRMSE)

RRMSE is a measure of variability that is related to direct and indirect estimation where the method of calculation comes from the value of standard error relative to the estimation. This measurement of variability usually used to evaluate the performance of estimation from different types of data. The formulation of RRMSE can be written by:

$$RRMSE = \frac{\sqrt{MSE\ (\hat{Y})}}{\hat{Y}} \times 100 (15)$$

3Result and Discussion

3.1The Result of Household Expenditure inDirect and EBLUP Estimation

The result from direct estimation was modelled with auxiliary variable in basic area level. The selection of auxiliary variables is based on the value from correlation between direct estimation and auxiliary variables. This selected auxiliary variables also can get from the value of non-multicollinearity and coefficient determination in linear models. The next step is we do the estimation of SAE by using EBLUP with basic area level. Figure 2 performed the power of EBLUP estimation wherefrom both figures show that the value of MSE in EBLUP estimation is smaller than the value of MSE from direct estimation. The decreasing of MSE estimation from the data also as usual decrease the number of relative standard error from data.

Table 1. The Result of Auxiliary Selection

Methods	Correlation						R^2	Multicollinearity
	X1	X2	X3	X4	X5	X6		
SRSWOR	0.303	0.551	0.305	0.542	0.609	0.362	0.585	no multicollinearity
One Stage Cluster (SRSWOR)	0.314	0.552	0.306	0.543	0.611	0.407	0.571	no multicollinearity
Two Stage Cluster (SRSWOR-SRSWOR)	0.305	0.506	0.308	0.545	0.618	0.364	0.534	no multicollinearity
Two Stage Cluster (PPSWR-SRSWOR)	0.307	0.559	0.399	0.531	0.615	0.369	0.536	no multicollinearity

Source: proceeded

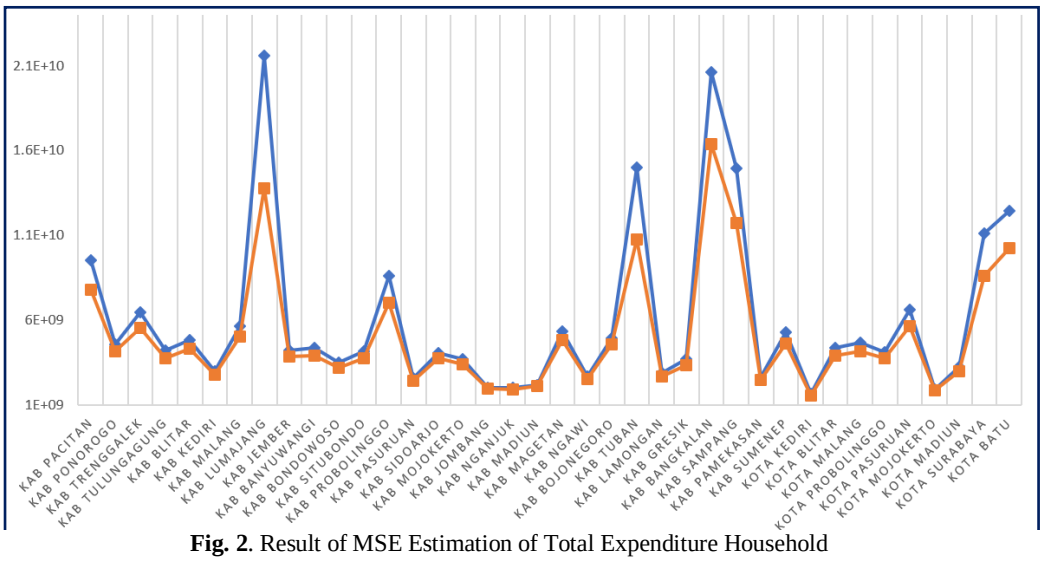

Fig. 2. Result of MSE Estimation of Total Expenditure Household

Based on the results of the selection of variables in table 1, the auxiliary variables obtained from the selected variables are:

X1 = Number of SLS units under the villages
X2 = Number of families who using electricity of PLN
X3 = The amount of gas/LPG/oil fuel for cooking
X4 = Number of liquid waste disposal sites in the form of infiltration/drainage holes
X5 = Number of households that have studied in educational/skills institutions
X6 = Number of SKTM issued by the village

3.2The Simulation of Complex Survey Designin EBLUP Estimation

In this research, a simulation by using a complex survey design is estimated directly and indirectly that involves the method of SAE. The estimation using the EBLUP method utilizes the output of direct estimation in the form of total estimation, variance or MSE and RRMSE estimated using a complex survey design. In addition, the information from the method to estimate the components of the variance is needed, in this case, we use the REML method. Some of the methods that applied are Simple Random Sampling Without Replacement (SRSWOR), SRSWOR One Stage Cluster, SRSWOR Two Stage Cluster in both stages and Proportional to Size With Replacement (PPSWR) in the first stage and SRSWOR in the second stage. The calculation of each direct estimation sampling method is simulated by using different sampling fractions. In addition, it is also carried out in each estimation of the sampling method so that the population of each area is more represented.

The data are considered as population data by involving complex survey designs with different sampling fractions. Furthermore, the selected sample is estimated to get the total estimation value and variance by involving a process of resampling with repetitions of 100 times so that the sample obtained can be representative. The level of estimation area that are carried out in this research is regency. Furthermore, the direct estimation results which are assumed to be unbiased in different sampling fractions are estimated using the EBLUP method. The example code of complex survey design by using resampling process in 100 times repetitionabove by replacing the parameter iteration become 100. The manual code function representing in R language :

```
...
samp_bs <- sample(1:length(bs_u$bs_u), size = round(prop*length(bs_u$bs_u)), replace
= FALSE)
sampi <- bs_u[samp_bs,]
...
est_ybar_1cs <- function(data,f){
      y_bar_i <- aggregate(hs_sam$pop, by=list(hs_sam$bs_o), FUN=mean)
      y_bar_n <- (sum(y_bar_i$x*Mi$Freq))/sum(Mi$Freq)
    var_y_bar_n<-(1-f)*( sum((Mi$Freq^2)*((y_bar_i$x-
    rep(c(y_bar_n),nrow(Mi)))^2))/sum((Mi$Freq^2)*((y_bar_i$x-
    rep(c(y_bar_n),nrow(Mi)))^2)))
      Y_cap_est <- (nrow(Mi)/f)*M_bar*y_bar_n
      Var_cap_est <- ((nrow(Mi)/f)*M_bar)^2*var_y_bar_n
    }
    result <- est_ybar_1cs(hsl_smpl, prop)
  }
...
```

```
  for(b in r102){
   data[[b]] <- subset(data, kab %in% b)
  }
  try_ye <- list()
  for(c in r102){
   try_ye[[c]] <- cb_sampling(data = data[[c]], prop = prop)
}
 }
 dataset = list()
 for( j in 1:iterasi){
  set.seed(j)
  dataset[[j]] = propinsi(data, prop)
 }
...
```

The result of direct estimation then was estimated by using EBLUP methods on area regency level. The average of result estimation from direct and EBLUP method above:

Table 2. The Average of Total Direct and EBLUP Estimation

Sampling Methods	**Sampling Fraction**	**Estimation Method**	**Average of Total Estimation**	**Average of Total MSE**	**Average of Total RRMSE**
SRSWOR	90 percent	Direct	2351561,37	755741135,77	1,1255
		EBLUP	2350956,10	753916723,95	1,1257
	60 percent	Direct	2350537,32	4523667703,49	2,7410
		EBLUP	2346918,52	4457326414,74	2,7456
	40 percent	Direct	2352785,54	10261467911,10	4,0979
		EBLUP	2344627,92	9913490683,42	4,1085
	20 percent	Direct	2347124,30	27044689650,78	6,5157
		EBLUP	2324913,07	24515048133,12	6,5393
One Stage Cluster (SRSWOR)	90 percent	Direct	2353262,53	2086892525,75	1,7107
		EBLUP	2351025,04	2068161260,22	1,7101
	60 percent	Direct	2356656,34	12433963620,09	4,1528
		EBLUP	2343301,04	11730267817,00	4,1412
	40 percent	Direct	2351630,76	28011315859,78	6,1627
		EBLUP	2320839,98	24264735515,43	6,1026
	20 percent	Direct	2348437,62	73692870722,15	9,7128
		EBLUP	2262471,88	46743577353,12	9,1009
Two Stage Cluster (SRSWOR-SRSWOR)	90 percent	Direct	2349831,30	2755883254,54	2,0451
		EBLUP	2347211,58	2726761169,76	2,0443
	60 percent	Direct	2352494,68	18531506319,39	5,2709
		EBLUP	2335153,20	17165252855,65	5,2482
	40 percent	Direct	2366709,93	49822469879,07	8,4304
		EBLUP	2318887,78	40556735057,56	8,3023
	20 percent	Direct	2337543,92	172415078891,43	15,0981
		EBLUP	2183967,34	59359373429,71	11,1037

Two Stage Cluster (PPSWR-SRSWOR)	90 percent	Direct	2349835,06	21615370152,89	5,4192
		EBLUP	2324222,37	19291032561,83	5,3722
	60 percent	Direct	2346347,08	36809161972,33	7,2445
		EBLUP	2307360,35	30639793566,98	7,0945
	40 percent	Direct	2362880,48	70962752491,62	9,7850
		EBLUP	2284296,02	47661790427,98	9,2191
	20 percent	Direct	2320831,24	201296091486,59	16,0541
		EBLUP	2149059,22	56450939469,32	10,7611

Source: proceeded

3.3The Efficiency Result

According to the result in table 2, it shows that the average total estimation of the direct and EBLUP estimation methods is not too different when the sampling fraction used is large for each method of the survey design. The average results of the total estimation of the estimated EBLUP are closest to the average direct estimate, which is at the sampling fraction of 90 percent using the SRSWOR method. Furthermore, the average total estimation results are the most different, namely in the Two Stage Cluster (PPSWR-SRSWOR) method with a sampling fraction of 20 percent. In table 2, it is also seen that the greater the percentage of the sampling fraction used causes the results of the average total estimation of the EBLUP method to be equal to the average total estimation of the direct estimation method for each complex survey design.

One of the methods to determine the goodness of the model and the best estimation is based on the values of MSE and RRMSE that we obtained. From the estimation results in each area, it is known that the smaller the sampling fraction used causes the average total MSE results and the total RRMSE average of the direct method to be even greater. Furthermore, in the EBLUP method, the smaller the sampling fraction used also caused the total mean MSE and total RRMSE average to be large but did not change significantly such as the total MSE and total mean results. In addition, in table 2 it is also seen that the more complex the survey method used causes the average MSE results and the average RRMSE in the small sampling fraction has a value that is increasingly different compared to the direct estimation results. However, for a very large sampling fraction such as at a sampling fraction of 90 percent, the total results of total MSE and total RRMSE from the direct and EBLUP estimation methods are not much different. Therefore, from the use of different survey designs, the largest MSE value obtained from the estimated total household expenditure of each regenciy in East Java Province was obtained from the Two Stage Cluster (SRSWOR-SRSWOR) sampling method while the smallest MSE value was obtained from the method SRSWOR sampling.

Furthermore, the estimation of the EBLUP method by involving a complex survey design can be seen in detail the progress of the estimation results in each area that will be described in the plot form of the MSE obtained. From the figure 3 until 18, it shows that when the sampling fraction used is large and the method of the survey design used is simple, it causes the estimation results from EBLUP to have estimates that are as good as direct estimates. However, when the result of estimation is carried out on a complex survey design and on a small sampling fraction, it caused the MSE value especially from areas that have outlier MSE have a smaller MSE value. The pattern of changes in the MSE results of each method and the survey design used can be clearly seen where the more complex the method of the survey design is used, the estimation results using the EBLUP method are better to do.

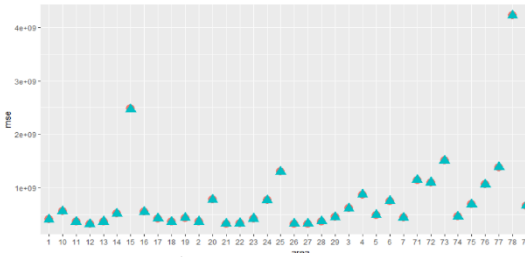

Fig. 3. Plot of MSE Direct and EBLUP in SRSWOR at sampling fraction 90 percent

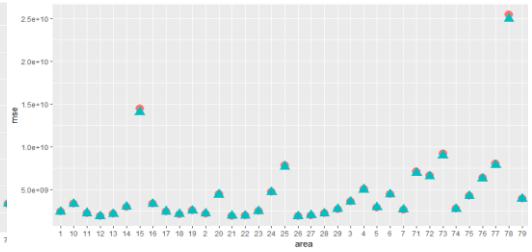

Fig.4. Plot of MSE Direct and EBLUP in SRSWOR at sampling fraction 60 percent

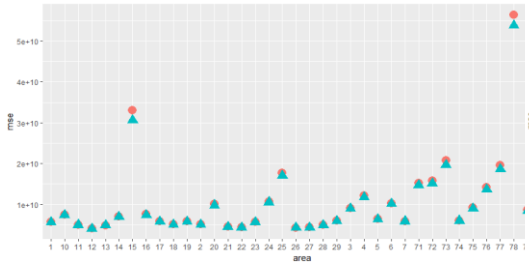

Fig.5. Plot MSE Direct and EBLUP in SRSWOR at sampling fraction 40 percent

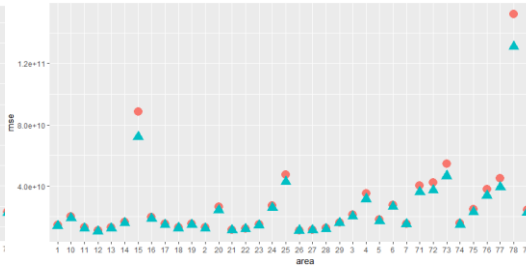

Fig.6. Plot MSE Direct and EBLUP in SRSWOR at sampling fraction 20 percent

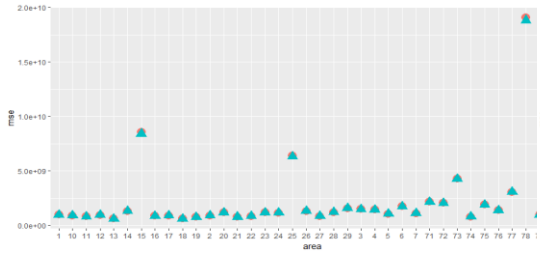

Fig.7. Plot MSE Direct and EBLUP in One Stage Cluster (SRSWOR) at sampling fraction 90 percent

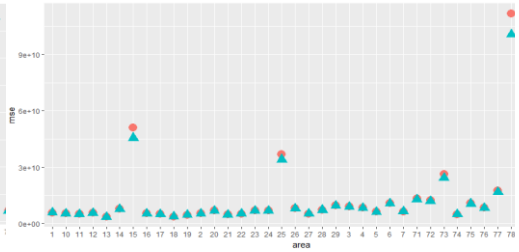

Fig.8. Plot MSE Direct and EBLUP in One Stage Cluster (SRSWOR) at sampling fraction 60 percent

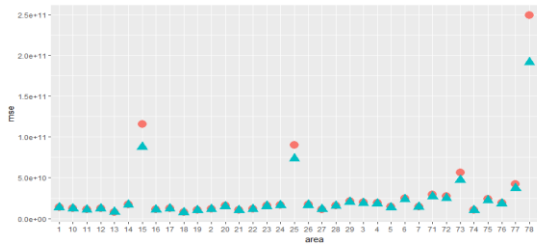

Fig.9. Plot MSE Direct and EBLUP in One Stage Cluster (SRSWOR) at sampling fraction 40 percent

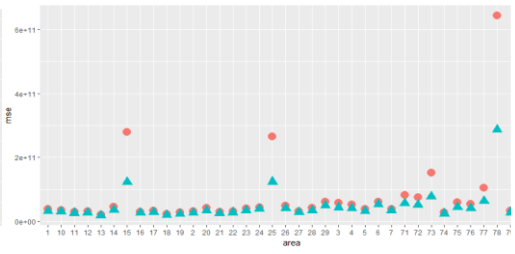

Fig.10. Plot MSE Direct and EBLUP in One Stage Cluster (SRSWOR) at sampling fraction 20 percent

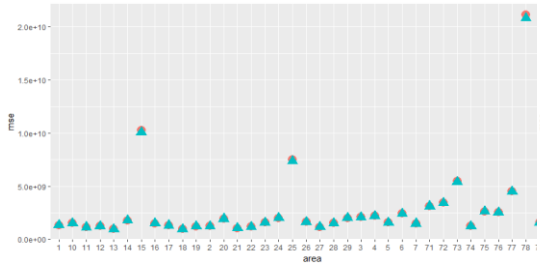

Fig.11. Plot MSE Direct and EBLUP in Two Stage Cluster (SRSWOR-SRSWOR) at sampling fraction 90 percent

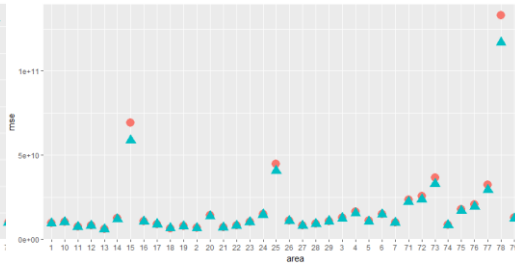

Fig.12. Plot MSE Direct and EBLUP in Two Stage Cluster (SRSWOR-SRSWOR) at sampling fraction 60 percent

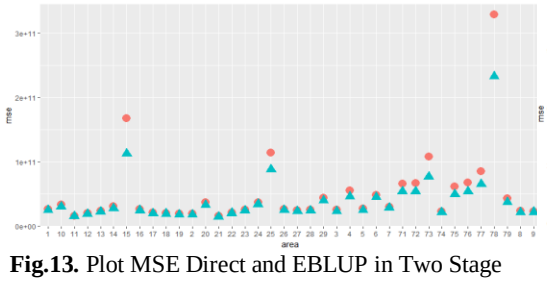

Fig.13. Plot MSE Direct and EBLUP in Two Stage Cluster (SRSWOR-SRSWOR) at sampling fraction 40 percent

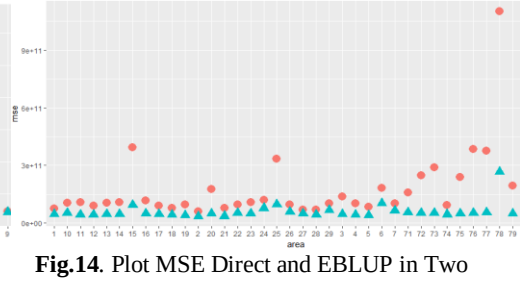

Fig.14. Plot MSE Direct and EBLUP in Two Stage Cluster (SRSWOR-SRSWOR) at sampling fraction 20 percent

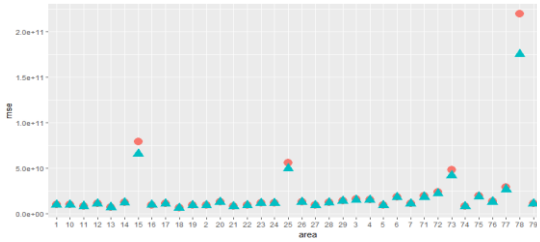

Fig.15. Plot MSE Direct and EBLUP in Two Stage Cluster (PPSWR-SRSWOR) at sampling fraction 90 percent

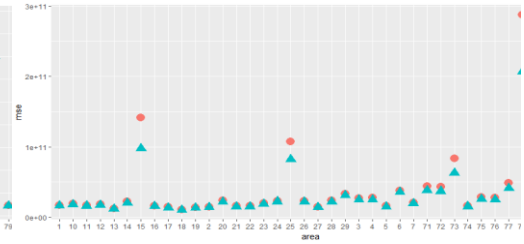

Fig.16. Plot MSE Direct and EBLUP in Two Stage Cluster (PPSWR-SRSWOR) at sampling fraction 60 percent

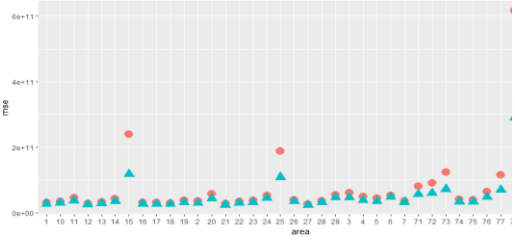

Fig.17. Plot MSE Direct and EBLUP in Two Stage Cluster (PPSWR-SRSWOR) at sampling fraction 40 percent

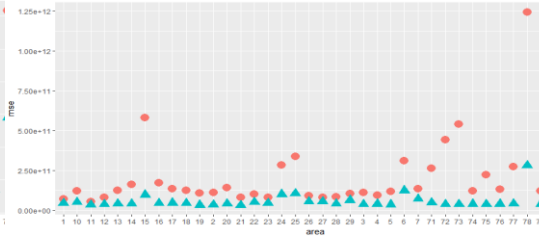

Fig.18. Plot MSE Direct and EBLUP in Two Stage Cluster (PPSWR-SRSWOR) at sampling fraction 20 percent

4. Conclusion

Calculation of estimates from SAE is carried out by using the area-based EBLUP method in complex survey designs using the SRSWOR sampling method, one stage cluster (SRSWOR), two stage clusters (SRSWOR-SRSWOR) and two stage clusters (PPSWR-SRSWOR) by generating estimates in the total estimation, MSE and RRMSE. The more complex the survey method used causes the average MSE results and the average RRMSE in the small sampling fraction has a value that is increasingly different compared to the estimated results from direct. However, at a very large sampling fraction, the average total MSE and total RRMSE results from the direct and EBLUP estimation methods are not much different. Therefore, from the use of different survey designs, the largest MSE value obtained from the estimated total household expenditure of each district / city in East Java Province was obtained from the Two Stage Cluster (SRSWOR-SRSWOR) sampling method while the smallest MSE value was obtained from the method SRSWOR sampling. The pattern of changes in the MSE results of each method and the survey design used can be clearly seen where the more complex the method of the survey design is used, the estimation results using the EBLUP method are better to do.

References

[1] Dedianto, D., & Wulansari, I. Y. (2019). Application of Small Area Estimation (SAE) Method Pseudo-EBLUP in Official Statistics of Indonesia. ASKS Journal, 33-38.

[2] Herrador, M., Morales, D., & Esteban, M. D. (2008). Sampling Design Variance Estimation of Small Area Estimators in The Spanish Labour Force Survey. SORT, 177.

[3] Ikhsan, E., Hidayat, C. A., & Nurizza, W. A. (2019). The Efficiency of EBLUP in Small Area EstimationCase Study: Estimation of Poor PogpulationPercentage in East Nusa Tenggara Province in 2017. ASKS Journal, 1-12.

[4] Lehtonen, R., & Pahkinen, E. (2004). Practical Method for Design and Analysis of Complex Survey Second Edition. Chichester: John Wiley & Sons.

[5] Lilja, D. (2017). An Implementation of the Linpack Benchmark for the R Programming Language. LinpackR.

[6] Matualagi, D. (2012). Metode Prediksi Tak Bias Linier Terbaik Empiris Spasial pada Area Kecil untuk Pendugaan Pengeluaran Perkapita. *Paper Tesis IPB*, 1.

[7] Niashinta, B. R. (2016). Pendugaan Area Kecil terhadap Pengeluaran Perkapita di Kabupaten Sragen dengan Pendekatan Kernel. Jurnal Gaussian, Volume 5, Nomor 1, 71-80.

[8] Ningtyas, R., Rahmawati, R., & Wilandari, Y. (2015). Penerapan EBLUP pada Model Penduga Area Kecil dalam Pendugaan Pengeluaran di Kabupaten Brebes. Jurnal Gaussian Volume 4, Nomor 4, 977-986.

[9] Rao, J. N. (2003). Small Area Estimation. Canada: John Willey & Sons Inc.

[10] Rao, J. N., & Molina, I. (2015). Small Area Estimation Second Edition. Canada: John Willey & Sons, Inc. Hoboken, New Jersey.

[11] Rumiati, A. T. (2012). Thesis: Model Bayes untuk Pendugaan Area Kecil dengan Penarikan Contoh Berpeluang Tidak Sama. Bogor: Institut Pertanian Bogor.

[12] Shalabh, I. K. (2011). Sampling Theory. Chapter 2 Simple Random Sampling, hal. 10.

[13] Singh, D., & Chaudry, F. (1986). Theory and Analysis of Sample Survey Designs. John Wiley & Sons.

[14] Wang, J., & Fuller, W. A. (2011). The Mean Squared Error of Small Area Predictors Constructed With Estimated Area Variances. Journal of American Statistical Association, 716-723.

[15] You, Y., & Chapman, B. (2006). Small Area Estimation Using Area Level Models and Estimated Sampling Variances. Survey Methodology Statistics Canada, 97-103.