IJCSN - International Journal of Computer Science and Network, Volume 6, Issue 6, December 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

The Estimation of The Total Number of Agricultural Families in Ogan Komering Ilir Regency of South Sumatra Province Under Incomplete Sampling Frame

[1]Asih Maulida, [2]Farit Mochamad Afendi , [3]Kusman Sadik

[1] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

[2] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

[3] Department of Statistics, Bogor Agricultural University, IPB
Bogor, 16680, Indonesia

Abstract - The various geographic and topography condition in Indonesia makes several areas in Indonesia have limited access. It needs a high cost and spends a long time on collecting data onto this area so some researchers tend to exclude this area from the sampling frame. Incomplete sampling frames influence the inclusion probabilities of the non-included unit in sampling frame and arises bias. Several approaches could be used to reduce bias, one of them is Predecessor-Successor method. We used a direct estimation of the total number of agricultural families in Ogan Komering Ilir regency by classical sampling theory and Predecessor-Successor method then evaluated their estimators. The results showed Predecessor-Successor method could reduce bias more effectively than classical sampling theory on a large sample size. Using an appropriate estimation method of a complete frame, the best estimator will be gotten. If it is unattainable, Predecessor-Successor method can be used to direct estimates of population quantity.
.

Keywords - *Coverage Error, Predecessor-Successor, Remote Area.*

1. Introduction

Indonesia has various geographic and topography conditions which make several areas has limited access (remote areas). This limitation causes higher budget and longer time was needed to hold a survey in these areas. The limited time and budget to hold survey encourages researchers to exclude those areas from sampling frames. This condition causes an incomplete sampling frame. The incomplete sampling frame ignores the principle of randomization so the estimator has a low level of validity and accuracy. The randomization gives the same opportunities for units in the sampling frame to be selected as a sample.

An incomplete sampling frame is one of the five main blocks of possible coverage errors [1]. This condition is very common in the implementation of the survey because it is unattainable to get a perfect sample frame, such as: complete, accurate and up to date especially in household surveys [2]. Some approaches are already being developed to deal with incomplete sampling frames, i.e. considering the inclusion probabilities as a function of the covariate and model it with logistic regression for estimating population total [3], combining multiple frames which are not complete and independent for estimating population size and totals [4], using the capture-recapture techniques for estimating population size and totals [5], using calibration weighting for estimating population size and totals [6], using Predecessor-Successor method for estimating population totals ([7], [8], [9]).

This research focuses on remote areas which cause incomplete sampling frames, so we use the Predecessor-method which was designed for estimating the population total of incomplete sampling frames. Besides these methods, we will also use the classical sampling theory for estimating it as a comparison of the estimator from the Predecessor-Successor method. From this research, we expect to know the performance of those methods so we can decide the appropriate sampling method of the

incomplete sampling frame caused by the existence of the remote areas.

2. Materials and Methods

2.1 Data

We took a case study in Ogan Komering Ilir regency of South Sumatra Province. South Sumatra Province is chosen because it has a higher percentage of the disadvantages village than the other provinces on the island of Sumatra [10]. We also chose Ogan Komering Ilir Regency because it has a higher percentage of the disadvantages village than the other regency in South Sumatra Province and it has the most number of remote areas in South Sumatra Province based on the results of updating the Master File of the Village of the year 2016 (MFD2016) 1st quarter. Ogan Komering Ilir Regency consists 18 districts and 327 villages.

This research will simulate some sampling techniques on data set that the result of merge between the 2014 Village Potential Census (Podes2014) data and the Master Files of the village of the year 2016 (MFD2016) quarter I data of South Sumatra Province. This research will direct estimate the number of agricultural families in the Ogan Komering Ilir regency because agriculture is the leading sector, which gives a significant contribution to the original local government revenue (PAD) of Ogan Komering Ilir Regency [11].

2.2 Methods of Analysis

First, the number of agricultural families' data was explored to know its characteristics either remote area or not remote area. Second, simulating data with the population total estimation used the classical sampling theory and Predecessor-Successor method. The classical sampling theory assumed that a complete frame exists. It included simple random sampling and stratified random sampling with two strata. The sample is randomly selected without replacement and it has five types of sample size. The simulation process was run 1 000 times under R. Third, their estimators were compared with their true value. Finally, the performance of their estimators was evaluated.

a. Simple Random Sampling of Simulation

This method simulated the population total estimation on an incomplete frame which has been already excluded remote areas. From the remaining list of villages (Table 1) sample is randomly taken with five types of sample sizes, such as: 8, 23, 73, 123, and 243 that represent the five of sample size levels, such as small, rather medium, medium, rather large and large.

Table 1 The Scenario Simulation of a Simple Random Sampling and a Stratified Random Sampling in Ogan Komering Ilir Regency

The Condition of Village	Rural/urban Classification		The Existence of Base Transceiver Station (BTS)		Total
	Rural	Urban	Available	Not Available	
Non Remote Area	283	21	104	200	304
Remote Area	22	1	4	19	23

Denote $\bar{y}$ is an unbiased estimator for the mean number of agricultural families in Ogan Komering Ilir Regency (Y), $\hat{T}$ is an unbiased estimator for the total number of agricultural families in Ogan Komering Ilir Regency (T) so to estimate the total number of agricultural families in Ogan Komering Ilir Regency, we use an unbiased estimator ($\hat{T}$) [12]:

$$\hat{T} = \frac{N}{n}\sum_{i=1}^{n} y_i = N\bar{y} \tag{1}$$

from Equation 1 with the sample variance

$$s_y^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$$

so an unbiased estimator of the variance of $\hat{T}$ is:

$$\hat{v}\left(\hat{T}\right) = N^2\left(\frac{N-n}{N}\right)\left(\frac{s_y^2}{n}\right) \tag{2}$$

:

b. Stratified Random Sampling of Simulation

Similar to the simple random sampling simulation, this method simulated the population total estimation on an incomplete frame which had been already excluded remote area. However, this simulation is still assumed that a complete frame exists. There are two strata will be used in this simulation, i.e: the urban/rural classification and the existence of Base Transceiver Station (BTS).

The population of N sampling unit are divided into H layers with N_h sampling units in stratum H, and n_h samples are randomly selected from N_h population units in

stratum h so we can estimate the population total in stratum h by [12]:

$$\widehat{T}_h = N_h \overline{y}_h, \overline{y}_h = \frac{1}{n_h} \sum_{j \grave{o} S_h} y_{hj} \quad (3)$$

from Equation 3, we estimate the population total by:

$$\hat{T} = \sum_{h=1}^{L} \widehat{T}_h \quad (4)$$

with $s_h^2 = \frac{\sum_{j \in s_h} \left(y_{hj} - \overline{y}_h \right)^2}{n_h - 1}$ so we can obtain an

unbiased estimator of variance of $\hat{T}$ by

$$\hat{v}\left(\hat{T}\right) = \sum_{h=1}^{L} N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{s_h^2}{n_h} \right) \quad (5)$$

- Stratified Random Sampling with Urban/Rural Classification as Stratum

The remaining of the villages was divided into two layers based on urban/rural classification. One layer included rural villages and the other ones included urban villages with their number of villages can be seen in Table 1. Samples were randomly selected from included villages on frame. The total of sample size allocated the number of sampled units in each stratum is proportional to the size of the stratum. There are five types of the total sample size, such as: 8, 23, 73, 123, and 243. They represent the fifth sample size of level, such as: small, rather medium, medium, rather large and large. The total number of agricultural families in Ogan Komering Ilir Regency were estimated by calculating the total number of agricultural families from sampled units are selected on each stratum.

- Stratified Random Sampling with The Existence of Base Transceiver Station (BTS) as Stratum

Similar to the simulation of stratified random sampling with urban/rural classification as stratum. The remaining of the villages was divided into two layers based on the existence of a base transceiver station. One layer included villages have a base transceiver station and the other ones included villages don't have a base transceiver station with their number of villages can be seen in Table 1. The total of sample size allocated the number of sampled units in each stratum is proportional to the size of the stratum. There are five types of the total sample size, such as: 8, 23, 73, 123, and 243. They represent the fifth sample size of the level, such as: small, rather medium, medium, rather large and large. Samples were randomly selected from included villages on frame. The total number of agricultural families in Ogan Komering Ilir Regency was estimated by calculating the total number of agricultural families from sampled units are selected on each stratum.

c. Predecessor-Successor Method of Simulation

This method was first proposed by Hansen, Hurwitz and Jabine [7] to obtain the missing information in frames as the impact of various problems in sampling from the incomplete frame (was cited in Singh [8]), while its mathematical formulation was proposed by Singh ([8], [13]) which was applied to illustrations by Agarwal and Gupta ([9]). This procedure treats the included units and the non-included units in a frame as two separate strata. This method provides the same probability of inclusion for either the included units or the non-included units in a frame by assuming the ordering of sampled units followed its geographical ordering or the rules of ordering are already determined. This method determines the sample of the non-included units in a frame as a successor and the sample of the included units in a frame as a predecessor [8].

There is no prior information on the non-included units so we must consider two conditions in the calculation of it's estimator.i.e. [13]:

a. The non-included units have the same characters as the included units.
b. The non-included units have the different characters as the included units.

Unlike the classical sampling theory of simulation, this method simulated the total population of estimator on a complete frame, but it is assumed incomplete frame based on the remote area classification. Let remote areas denote non-included unit (successor), while non remote areas denote included unit (predecessor).

Both of them are established in a geographical ordering on the sampling frame. The ordering of the 2014 village potential census data as sampling frames of this research followed the sequence of administrative regions which refers to geographical location.

- Two Groups with the Same Characteristics

Assuming the number of remote areas is known and both of the remote areas and non remote area behave similar, this method only observes the value of the character under study for the selected included units and note the number of non-included units between the selected included units and the next included unit in the frame.

Let N_1 the included units in the frame which belong to the target population and let M the non-included units in the frame which belong to the target population so the target population consists of T units ($N_1 + M$). Select a sample of size n_1 from N_1 units randomly without replacement. Denote y_i is the value of the character under study for the ith included unit. All M_i units are automatically selected in the sample because M_i is the number of non-included unit between two the included units. Let m_i is the number of non-included unit between the ith included unit and (i+1)th included units. Therefore, an unbiased estimator for the target population total (T) is [9]:

$$\widehat{T} = N_1 \overline{y}_{n_1} + \overline{m} N_1 \overline{y}_{n_1} = N_1 (1+\overline{m}) \overline{y}_{n_1} \quad (6)$$

with $\overline{y}_{n_1} = n_1^{-1} \sum_{i=1}^{n_1} y_i$ and $\overline{m} = n_1^{-1} \sum_{i=1}^{n_1} m_i$ so an unbiased estimator for the population non-included unit (M) is $\hat{M} = N_1 \overline{m}$. From Equation 6 with $\hat{v}(\overline{m}) = \left(\frac{1}{n_1} - \frac{1}{N_1} \right) s_m^2$, we can get an unbiased estimator of the variance of $\widehat{T}$ by:

$$\hat{v}(\hat{T}) = N_1^2 \left[(1+\overline{m})^2 \hat{v}(\overline{y}_{n_1}) + \hat{v}(\overline{m}) \left(\overline{y}_{n_1}^2 + \hat{v}(\overline{y}_{n_1}) \right) \right] \quad (7)$$

.

- Two Groups with the Different Characteristics

Assuming the number of remote areas is known and both of the remote areas and non remote areas behave differently, this method observes the character under study from either included units or non-included units in the frame. The character under study (y) from non-included units was observed by tracing them by the Predecessor-Successor method. As the two groups of the same characteristics, suppose y_i is the value of the character under study for the included unit ith. Denote m_i is the number of non-included unit between the included unit ith and included units (i+1)th then $m = \sum_{i=1}^{n_1} m_i$.

Let z_i is the sum of the values of y from the included unit ith and all the m_i units following it. In this case an unbiased estimator for the target population total (T) is [13]:

$$\hat{T} = \frac{N_1}{n_1} \sum_{i=1}^{n_1} z_i \quad (8)$$

with $z_i = y_i + \sum_{j=1}^{m_i} y_{ij}$. From Equation 8 and $s_z^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (z_i - \overline{z})^2$, so an unbiased estimator of the variance of $\hat{T}$ is:

$$\hat{v}(\hat{T}) = \frac{N_1 (N_1 - n_1)}{n_1} s_z^2 \quad (9)$$

- Estimation of The Total Number of Agricultural Families

Ogan Komering Ilir Regency has 23 remote areas (M) and 304 non-remote areas (N_1) and 327 total villages. From non-remote areas are randomly selected sample of size n_1, i.e., n_1= 8, 23, 73, 123, and 243. The fifth sample size represents the five levels of sample size are small, rather medium, medium, rather large, and large. Before simulating the estimation, we must explore the characteristics of two groups of the village in Ogan Komering Ilir Regency to know the similarity between them. After knowing the similarity between them, we can determine the case of Predecessor-Successor method will be used. Based on sample have selected by Predecessor-Successor method, we can estimate the total number of agricultural families in Ogan Komering Ilir Regency.

d. Evaluation of The Performance of Several Sampling Techniques

This research used absolute relative bias (ARB) and relative root mean square error (RRMSE) to evaluate the performance of several sampling techniques on their estimators. The absolute relative bias (ARB) of an

estimator of the target population total T is given by (Rao 2003):

$$\text{ARB} = \left| \frac{1}{1000} \sum_{r=1}^{1000} \left(\frac{\hat{T}_r - T}{T} \right) \right| \quad (10)$$

with $\hat{T}_r$ is the value of an estimator of the target population total T for the rth simulation run (r=1,.....,1 000) and the relative root mean square error (RRMSE) of an estimator of the target population target T is given by [14]:

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{1000} \sum_{r=1}^{1000} \left(\hat{T}_r - T \right)^2}}{T} \quad (11)$$

3. Results and Discussion

3.1 The Characteristics of Remote Areas

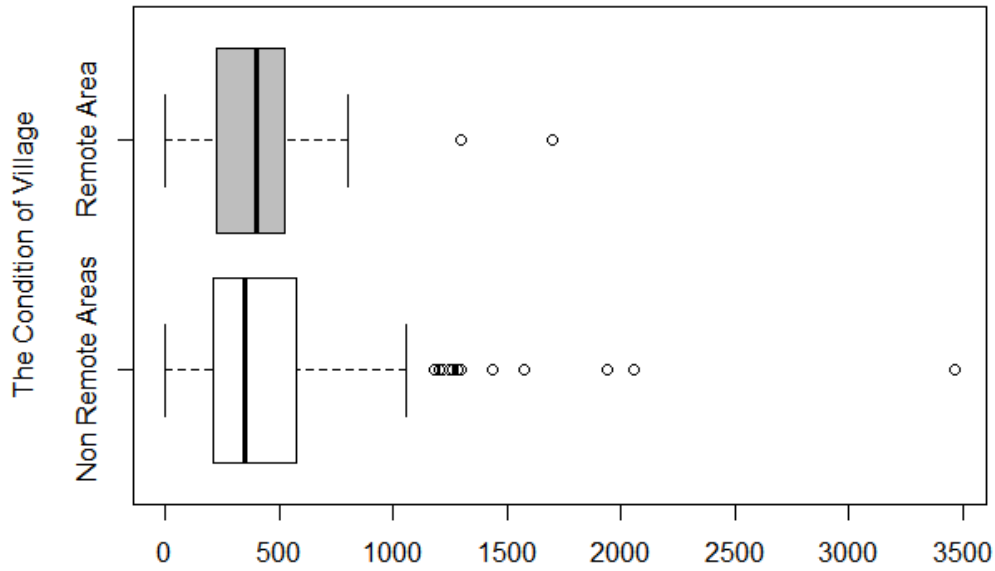

Figure 1 Boxplot of Number of Agricultural Families based on The Condition of Village in Ogan Komering Ilir Regency

Figure 1 showed the similarity between the distributions of the number of agricultural families between remote area and non remote area in Ogan Komering Ilir Regency. Their shape of a distribution was skewed to the right. Their variability of a number of agricultural families was similar, 50% of villages in both groups had a number of agricultural families were not much different. Similary, 25% of the villages have the number of the lowest and highest agricultural families in both groups were similar. In addition, both groups also have several villages that the number of agricultural families as potential outliers.

Table 2 Characteristics of Two Village Groups Based on Descriptive and Inferential Analysis in Ogan Komering Ilir Regency

The Condition of Village	Number of Villages	Number of Agricultural Families		
		Average	T-Test	
			T Value	P-Value
Non Remote Area	304	436.12	-0.45	0.65
Remote Area	23	471.35		

The description of the similarity between the characteristics belonging to both groups was reinforced by the result of testing for comparing two population means (see Table 2).

The result of testing showed there was not enough evidence to conclude that the mean number of agricultural families of remote area differed from the mean number of agricultural families of non remote area in Ogan Komering Ilir Regency with a significance level of α=0.05. Based on the boxplot of a number of agricultural families according to the condition of the village and the result of testing for comparing two population means, so we can conclude that the characteristics of two groups of villages were similar. Therefore, we used the Predecessor-Successor method of the same characteristics of two groups to estimate the total number of agricultural families in Ogan Komering Ilir Regency.

3.2 Evaluation of Estimation Methods under Incomplete Sampling Frame

Table 3 Evaluation of Estimation Results of Number of Agricultural Families of Ogan Komering Ilir Regency from Several Sampling Techniques in The Incomplete Sampling Frame

Sampling Techniques	Sample Size									
	8		23		73		123		243	
	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE	ARB	RRMSE
Simple Random Sampling	0.217	103.314	0.144	66.225	0.090	40.689	0.080	35.156	0.077	30.198
Stratified Random Sampling1[a]	0.223	106.393	0.146	66.112	0.090	40.782	0.080	35.103	0.076	29.854
Stratified Random Sampling2[b]	0.225	107.449	0.179	86.220	0.090	40.611	0.081	35.515	0.076	29.953
Predecessor-Successor Method	0.244	123.790	0.150	72.020	0.077	36.028	0.053	24.928	0.021	10.201

[a]Stratified Random Sampling with Urban/Rural Classification as Stratum
[b]Stratified Random Sampling with The Existence of Base Transceiver Station (BTS) as Stratum

Table 3 showed ARB values of three classical sampling theories of estimators were smaller than the Predecessor-Successor method of estimator for the small and rather medium sample size. On the other hand, the Predecessor-Successor method of estimator has a smaller ARB value than three classical sampling theories of estimators for the medium, rather large and large sample size.

Similar to the pattern of ARB values, RRMSE values of three classical sampling theories of estimators were smaller than the Predecessor-Successor method of estimator for the small and the rather medium sample size (Table 3). However, RRMSE value of the Predecessor-Successor method of estimator was smaller than three classical sampling theories of estimators for medium, rather large and large sample size.

From the simulation results were seen that the fifth of the sample size level had different performance of biased. The simulation result showed that the validity of level was higher for a larger sample size than the smaller sample size. This was seen that the increasing of the sample size

decreased the biased value. This simulation result was similar with Levi and Lemeshow [15] that the validity of the estimator will be improved with an increasing sample size.

This simulation results also showed that there was the difference performance of two kinds of sampling techniques. Overall predecessor-successor method was more effective in reducing bias than classical sampling theories. This simulation result was similar to the illustrations results which were proposed by Agarwal and Gupta [9] that the predecessor-successor method of estimator was better than classical sampling theories of estimators.

In addition, the results of the simulations also showed that the influence of the interaction between the sample size and sampling techniques. The increasing sample size increased the performance predecessor-successor method effectiveness in reducing bias. As can be seen in Figure 2 that the decreasing bias of the predecessor-successor method‘s estimator was a very sharp than classical sampling theories’s estimators for larger sample. On the other hand, the decreasing biased of the classical sampling theories’s estimators were more significant than the predecessor-successor method‘s estimator decreased to smaller sample.

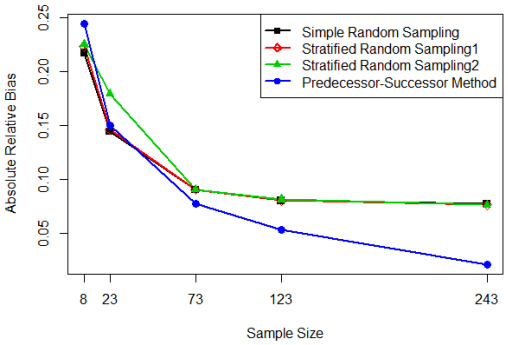

Figure 2 Interaction Plot between Sample Size and Sampling Techniques

Overall the biased value of the simple random sampling estimator was smaller than the biased value of the stratified random sampling estimator in either urban/rural classification or the existence of Base Transceiver Station as a stratum. Nevertheless, the difference in the biased values of the three methods estimators is insignificant, so two strata used are appropriate for estimating the number of agricultural families. The best stratum to estimate the total number of agricultural families in Ogan Komering Ilir District is urban/rural classification. It was seen that both ARB value and RRMSE value of the estimator of stratified random sampling with urban/rural classification as stratum were smaller than the estimator of stratified random sampling with the existence of Base Transceiver Station (BTS) as stratum. In addition, some simulation results showed the RRMSE value of the estimator of stratified random sampling was smaller than the estimator of simple random sampling. It was one of several advantages of stratified random sampling. This was equal to Lohr [12] that explained one of benefit using a stratified random sampling is the precise of estimator will increase.

The selection case of Predecessor-Successor method will be used on an incomplete frame based on the characteristics of both groups. If the characteristics of two groups are the same, we use Predecessor-Successor method that is designed for the same characteristics of two groups. On the other hand, if the characteristics of two groups are different, so we use Predecessor-Successor method that is designed for the different characteristics of two groups. The two cases of Predecessor-Successor method have the different way of estimating. Because of the same characteristics of two groups of villages in Ogan Komering Ilir Regency, this research used Predecessor-Successor method with the same characteristics of two groups to estimate the total number of agricultural families in Ogan Komering Ilir Regency.

The Predecessor-Successor method of the same characteristics of two groups required a faster time to run its simulation. In addition, the calculation of estimating parameters was simpler than the Predecessor-Successor method of the different characteristics of two groups.

Different to the illustration was presented by Agarwal and Gupta [9], this research used three kinds of replication i.e. 100, 1 000 and 10 000. Because of the same pattern of three estimators, this study only showed the estimators of 1 000 replication. Another consideration is the efficiency of time and memory for calculation.

4. Conclusions

A biased value of the estimator which is caused coverage error due to the incomplete sampling frame will be significantly minimized by the Predecessor-Successor method on medium and large sample size. On the other hand, classical sampling theories include either simple sampling method or stratified sampling method will be more efficient than the Predecessor-Successor method for small and rather medium sample size.

Using the appropriate estimation method can minimize the biased problems. The results of the estimator would be better if all of the units were the target of the population and included in a frame. However, if a complete frame cannot be obtained especially region, which has remote area, Predecessor-Successor method can be chosen as an alternative method to estimate a population quantity.

References

[1] [UN] United Nations, Handbook on Economic Tendency Surveys (ETS), New York: UN, 2015.

[2] [UN] United Nations, Designing Household Survey Samples: Practical Guidelines, New York: UN. 2008.

[3] D.E. Haines, K.H. Pollock, and S.G. Pantula, “Population Size and Total Estimation When Sampling From Incomplete List Frames With Heterogeneous Inclusion Probabilities”, Survey Methodology,Vol.26, No.2, 2000, pp. 121-129.

[4] D.E. Haines, and K.H. Pollock, “Combining Multiple Frames, to Estimate Population Size and Totals”, Survey Methodology, Vol.24, No.1, 1998, pp. 79-88.

[5] K.H. Pollock, S.C. Turner and C.A. Brown, “Use of Capture-Recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable”, Survey Methodology, Vol.20, No.2, 1994, pp. 117-124.

[6] P. S. Kott, “Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors”, Survey Methodology, Vol.32, No.2, 2006, pp. 133-142.

[7] M.H. Hansen, W.N. Hurwitz, and T.B. Jabine, “The Use of Imperfect Lists for Probability Sampling at The U.S. Bureau of the Census”, Bulletin of the International Statistical Institute, Vol.40, Book 1, 1963, pp. 497-517. (Was cited in R.Singh, “Methods of Estimation for the from Incomplete Sampling Frames”, Austral J Statist, Vol.31, No.2, 1989, pp.269-276).

[8] R.Singh,“Methods of Estimation for the from Incomplete Sampling Frames”, Austral J Statist, Vol.31, No.2, 1989, pp.269-276.

[9] B. Agarwal, and P.C. Gupta, "Estimation from Incomplete Sampling Frames in Case of Simple Random Sampling", Model Assisted Statistics and Applications, Vol.3, 2008, pp.113-117.

[10] [Bappenas] Ministry for National Development Planning/National Development Planning Agency of The Republic of Indonesia, 2014 The Village Development Index, Jakarta: Bappenas, 2015.

[11] [BPS OKI Regency] Statistics of Ogan Komering Ilir Regency, Ogan Komering Ilir Regency in Figures 2016, Kayu Agung: BPS OKI Regency, 2016.

[12] S.L. Lohr, Sampling: Design and Analysis, 2nd edition, Boston: Brooks/Cole, 2010.

[13] R.Singh, "Predecessor-Successor Method", In Encyclopedia of Statistical Sciences, 2006, DOI: 10.1002/0471667196. ess2044.

[14] J.N.K. Rao, Small Area Estimation, New Jersey: J. Wiley, 2003.

[15] P.S. Levy, and S. Lemeshow, Sampling of Population: Methods and Applications, New York: J. Wiley, 1999.

[1]**First Author** Student of masters degree programs in applied statistics at Bogor Agricultural University, Indonesia and attained bachelor degree from the Gajahmada University in 2006.

[2]**Second Author,** Lecturer of Department of Statistics, Bogor Agricultural University, Indonesia. His research interest in Geoinformatics. His doctorate degree was attained from Nara Institute of Science and Technology, Japan.

[3]**Third Author,** Lecturer of Department of Statistics, Bogor Agricultural University, Indonesia. His research interest in statistical modelling, small area estimation, robust statistics. His doctorate degree was attained from Bogor Agricultural University, Indonesia. A part of his courses and research of doctoral was taken at the University of Maryland, United States.