International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

The Study of Small Area Estimation Using Oversampling and M-Quantile Robust Regression Approach

Nia Aprillyana[a*], Kusman Sadik[b], Indahwati[c]

[a,b,c] *Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agriculture University, Bogor, Indonesia*

[a] *Email: nia.aprillyana@gmail.com*

[b] *Email: kusmansadik@gmail.com*

[c] *Email: indahwati_43@yahoo.co.id*

Abstract

Statistics Indonesia (BPS) calculates poverty indicators (Head Count Ratio, Poverty Gap, and Poverty Severity) using National Socio-Economic Survey (Susenas). Susenas is only designed to estimate province and municipality/regency area level, whereas the government requires estimation until smaller area level (sub-district and village). Estimating poverty indicators directly from Susenas for the smaller area often leads to inaccurate estimates. To solve this problem, BPS usually conduct additional survey called Regional Socio-Economic Survey (Suseda) by increasing number to the original sample (called oversampling) but with the very high cost. Therefore, we proposed small area estimation technique which based on the unit level model using Population Census 2010 (SP2010) as the population auxiliary variables and household per-capita expenditure (Susenas 2015) as the response variable. We utilized robust M-quantile regression model which robust to the outlier using three weight functions (Huber, Hampel, and Tukey Bisquare). Our results provide evidence that M-quantile model is more accurate than direct estimates with oversampling.

Keywords: M-quantile; oversampling; small area estimation; weight function.

---

* Corresponding author.

1. Introduction

The national survey has the important role in country development. The main objective of the national survey is to estimate a parameter of the population. The national survey only provides limited information and only estimates the large population or sub-population as stated by [1]. The national survey has large variance and low accuracy in estimating smaller area level (such as sub-district or village). This is due to the insufficient of the sample in estimating the small area level. The solution to overcoming the low-accuracy problem toward direct estimation is increasing the number of samples (called by oversampling) or utilizing the small area estimation technique (SAE) with the original sample only [2].

Oversampling has two benefits i.e to provides more accurate estimates and to overcomes the non-response [3]. But oversampling is inefficient, it needs extra time, high human resource, and high cost. To overcome the inefficient of oversampling, we can use the small area estimation technique with utilize the original sample only. We also called it “indirect estimation” or model-based estimation. Rao stated that indirect estimation can be implemented by "borrowing information" or “utilizing additional variables (auxiliary/explanatory variables)” obtained from other areas in the same survey, from the same area in the previous survey, or other variables related to variables that are of concern to small areas [4].

In Indonesia country, the poverty indicators are officially calculated through the National Socio-Economic Survey (Susenas) which be conducted by the national statistical office (BPS-Statistics Indonesia) for estimation of national, province, and district/regency level only [5]. Because of the decentralization of regional development, the government requires the indicators until the smaller area (sub-districts and villages level). So that, some BPS branch/representative office conducted an additional survey (Regional Socio-Economic Survey/ Suseda).

This study used indirect estimation technique with M-quantile models which do not depend on strong distributional assumption and automatically provide outlier robust inference [6,7,8]. M-quantile uses M-estimator that gives small weight to the outliers (and the residual) and ignores the assumptions in the least squares method (OLS). Some popular M-estimators which able to be utilized in M-quantile model are Huber, Tukey Bisquare, and Hampel weight function. Some M-quantile studies more often use Huber weight function than the other such as the researches in [2,9]. This our study will evaluate the accuracy of estimate using M-quantile with three weight functions and direct estimation (with original sample and oversample) at the sub-district area level.

This study also has some constraints and limitations such as (1) The interest area of this study is only for Musi Rawas regency (at South Sumatera province). We don’t estimate the national level of Indonesia. (2) The analysis uses the data sources which come from Susenas (March 2015), Suseda (September 2015) and Population Census 2010. (3) This study estimates poverty indicator until sub-district only. (4) Estimation technique uses the assumption that the sampling design is based on simple random sampling (SRS). It is different with national statistical office (BPS-Statistics Indonesia) which use multistage random sampling as the sampling design of Susenas officially. (5) Oversampling is a term that means increasing the number of samples.

(6) The original sample is a term of the sample without oversampling. (7) This study is the applied statistical theory based on the real data. We don't discuss the statistical formula in detail.

2. Methodology

2.1. Data Sources

We use three main data sources from census and survey in Musi Rawas Regency, these are Susenas (March 2015), Suseda (September 2015), and Population Census 2010 (SP2010). We determined household per-capita expenditure as the response variable (Y), while the auxiliary variables (X) in the indirect estimation using three variables from SP2010. Initially, there are eight selected variables, then using stepwise method, we filter three most significant variables, these are: (1) the number of household members (X1), (2) working field of head of household (X2), and (3) working status of head of household (X3). We get household per-capita expenditure data from Susenas 2015 (in rupiah unit), while the number of household members data, we get from Susenas 2015 and SP2010 (in individual unit). Working field (divided into 19 categories) and working status of head of household (divided into 6 categories) are also gotten from Susenas 2015 and SP2010.

2.2. Direct Estimation

The initial step of this research is the exploration of the response variable and its distribution pattern. Furthermore, we determine the poverty line based on official release from Statistics Indonesia (BPS) publication. In the final step, we calculate sub-district area-level poverty indicators estimate and its RRMSE. We estimate the poverty indicator using direct estimation either original sample or oversampling through Foster formula [10] below:

$$\hat{P}_{\alpha,d}^{Dir} = \frac{1}{n_d}\sum_{j=1}^{n_d}\left(\frac{t-\mathbf{y}_{jd}}{t}\right)^{\alpha} I(\mathbf{y}_{jd} \leq t) \quad (1)$$

Where d is sub-district (d = 1,2,….,14); α is Sensitivity parameter (0 for HCR 1 for PG, 2 for PS); $\mathbf{y}_{jd}$ is Household per-capita expenditure unit j in area (sub-district) d; t is poverty line (in rupiah); n_d is number of samples in sub-district d; $I(\mathbf{y}_{jd} \leq t)$ is indicator function for each unit under poverty line. $I(\mathbf{y}_{jd} \leq t) = 1$ if $\mathbf{y}_{jd} \leq t$ and $I(\mathbf{y}_{jd} \leq t) = 0$ if $\mathbf{y}_{jd} > t$.

RMSE ($P_{\alpha,d}^{Dir}$) of the direct estimate is calculated using the bootstrap method on resampling with 100 replacement (B = 100). The bootstrap method produces $\hat{P}_{\alpha,d}^{*B}$ which is used to calculate the RMSE of direct estimate by a formula:

$$\widehat{RMSE}\left(P_{\alpha,d}^{Dir}\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{P}_{\alpha,d}^{*B} - \hat{P}_{\alpha,d}^{Dir}\right)^2} \quad (2)$$

2.3. M-quantile Model

The next step is estimating the poverty indicator and its RRMSE using indirect estimation method (M-quantile model). The M-quantile has the general equation $\hat{\mathbf{y}}_{\mathrm{jd}} = \mathbf{X}_{\mathrm{jd}}^{\mathrm{T}}\boldsymbol{\beta}_{\psi}(\mathrm{q}_{\mathrm{jd}})$ with $\mathbf{y}_{\mathrm{jd}}, \mathbf{X}_{\mathrm{jd}} \in \mathbf{s}_{\mathrm{d}}$ (from Susenas data) in the quantile value q_{jd} = (0,01, … 0,99) where $\mathbf{X}_{\mathrm{jd}}$ is auxiliary variable matrices, include vector 1.

$\widehat{\boldsymbol{\beta}}_{\psi}(\mathrm{q}_{\mathrm{jd}})$ is the regression coefficient estimate in a q_{jd} value. The first step in modeling is getting q_{jd} value using interpolation. The second step is to calculate the area M-quantile coefficient ($\hat{\theta}_{\mathrm{d}}$). Then, calculate estimates $\widehat{\boldsymbol{\beta}}_{\psi}$ based on the area M-quantile coefficient ($\hat{\theta}_{\mathrm{d}}$) using $\hat{\mathbf{y}}_{\mathrm{jd}} = \mathbf{X}_{\mathrm{jd}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_{\mathrm{d}})$ formula through Iterative Reweighted Least Squares (IRLS) procedure. In iteration-0, we calculate initial estimate $\hat{\beta}^{(0)}$ using OLS formula $\hat{\mathbf{y}}_{\mathrm{jd}} = \mathbf{X}_{\mathrm{jd}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. The value of $\hat{\beta}^{(0)}$ is used to calculate residual $\mathbf{r}_{\mathrm{jd}}{}^{(0)}$ and initial scaled value $\mathrm{s}^{(0)}$. The next step, we use weight function w(u) which will be implemented toward scaled residual $\mathrm{u} = \frac{\mathbf{r}_{\mathrm{jd}}{}^{(0)}}{\mathrm{s}^{(0)}}$ to get initial weight $\mathrm{w}_{\mathrm{jd}}{}^{(0)}$ and matrix $\mathbf{W}^{(0)}$.

In the next step, $\mathrm{w}_{\mathrm{jd}}{}^{(0)}$ was used in first iteration (get new estimate $\widehat{\boldsymbol{\beta}}^{(1)}$) until get matrix $\mathbf{W}^{(\mathbf{1})}$. Finally, we get general formula for regression estimate from iteration process in the following way:

$$\widehat{\mathrm{B}}^{(\mathrm{iter})} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(\mathrm{iter}-\mathbf{1})}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(\mathrm{iter}-\mathbf{1})}\mathbf{Y} \qquad (3)$$

The IRLS procedure was running until there is convergence state. The convergence is stated when the percentage of change of the residual is less than ε value. When it convergence, we get the estimate as M-quantile regression coefficients. Convergence criteria is formulated as $\sqrt{\frac{\sum_{jd=1}^{n_d}\left(\mathbf{e}_{jd}{}^{(iter-1)}-\mathbf{e}_{jd}{}^{(iter)}\right)^2}{\sum_{jd=1}^{n_d}\mathbf{e}_{jd}{}^{(iter-1)^2}}} < \varepsilon$. Where ε value is the positive small number, usually is 0,0001.

After getting $\widehat{\boldsymbol{\beta}}_{\psi}$, we calculate the residual $\mathbf{e}_{\mathrm{jd}} = \mathbf{y}_{\mathrm{jd}} - \hat{\mathbf{y}}_{\mathrm{jd}}$ and construct a model $\hat{\mathbf{y}}_{\mathrm{kd}} = \mathbf{X}_{\mathrm{kd}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_{\mathrm{d}}) + \mathbf{e}_{\mathrm{jd}}^{*}$ where $\mathbf{e}_{\mathrm{jd}}^{*}$ is the random sample with replacement of size n_{d} in residual $\mathbf{e}_{\mathrm{jd}}$ and $\mathbf{X}_{\mathrm{kd}}$ is the matrix of auxiliary variables from population census 2010. We used Monte Carlo simulation with 50 iterations, and each iteration we calculate HCR, PG, and PS estimates based on the poverty formula by Foster and his colleagues [10]:

$$\widehat{\mathrm{P}}_{\alpha,\mathrm{d}}^{\mathrm{L}} = \frac{1}{\mathrm{N}_{\mathrm{d}}}\sum_{\mathrm{i}=1}^{\mathrm{N}_{\mathrm{d}}}\left(\frac{\mathrm{t}-\hat{\mathrm{y}}_{\mathrm{kd}}}{\mathrm{t}}\right)^{\alpha}\mathrm{I}(\hat{\mathbf{y}}_{\mathrm{kd}} \leq \mathrm{t}) \qquad (4)$$

So we get final HCR, PG, and PS estimates from the average of the previous simulation

$$\widehat{\mathrm{P}}_{\alpha,\mathrm{d}} = \frac{1}{\mathrm{L}}\sum_{\mathrm{l}=1}^{\mathrm{L}}\widehat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{L}} \qquad (5)$$

In calculating RMSE, we initially generate population bootstrap $\mathrm{U}^{*\mathrm{b}}$, where each population bootstrap $\mathrm{U}^{*\mathrm{b}}$ we calculate $\widehat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{b}}$ estimate. Then each $\mathrm{U}^{*\mathrm{b}}$we took 100 $\mathrm{U}^{*\mathrm{r}}$ sample bootstrap with simple random sampling without replacement as much small area sample number $\mathrm{n}_{\mathrm{d}}^{*} = \mathrm{n}_{\mathrm{d}}$ to calculate $\widehat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{br}}$ estimate. Bias and variance estimate are defined as:

$$\hat{\mathrm{B}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right) = \mathrm{B}^{-1}\mathrm{R}^{-1}\sum_{\mathrm{b}=1}^{\mathrm{B}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{br}} - \hat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{b}}\right) \quad (6)$$

$$\hat{\mathrm{V}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right) = \mathrm{B}^{-1}\mathrm{R}^{-1}\sum_{\mathrm{b}=1}^{\mathrm{B}}\sum_{\mathrm{r}=1}^{\mathrm{R}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}^{*\mathrm{br}} - \bar{\hat{\mathrm{P}}}_{\alpha,\mathrm{d}}^{*\mathrm{br}}\right)^2 \quad (7)$$

So the RMSE is defined as:

$$\widehat{\mathrm{RMSE}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right) = \sqrt{\hat{\mathrm{B}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right)^2 + \hat{\mathrm{V}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right)} \quad (8)$$

2.4. Accuracy Comparison

Then, we compare all estimation technique using RRMSE which can be computed as

$$\mathrm{R}\widehat{\mathrm{RMSE}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right) = \frac{\widehat{\mathrm{RMSE}}\left(\hat{\mathrm{P}}_{\alpha,\mathrm{d}}\right)}{\hat{\mathrm{P}}_{\alpha,\mathrm{d}}} \mathrm{x}100\% \quad (9)$$

3. Result and Discussion

3.1. Data Exploration

Musi Rawas is one of the regencies in South Sumatera Province in Indonesia. In 2015, Musi Rawas Regency consists of 14 sub-districts and 199 villages with 97.313 households.

In the Susenas March 2015, there were 504 household samples which were distributed in 48 villages. In the Suseda September 2015, there were 1.199 households which were distributed in 131 villages. The combination of Susenas and Suseda had the total of the sample is 1.703 households.

Poverty line 2015 of Musi Rawas is 342.956 rupiah. In this study, the oversampling has a meaning the combination of Susenas and Suseda samples and oversampling is only used in direct estimation. Table 1 shows the descriptive statistics of household per-capita expenditure variables.

Table 1: Household per-capita expenditure (rupiah)

Data	Mean	Median	Std. Dev	Min	Max
Without versampling	647 751	547 516	407 918	191 097	3 507 652
Oversampling	640 368	573 953	313 455	169 313	3 507 651

According to table 1, the mean and median of oversampling data are only slightly different with the nonoversampling data. The standard deviation of oversampling data is smaller than without oversampling data, meaning that the variance of oversampling data is smaller. Boxplot 1 shows the differences in data distribution.

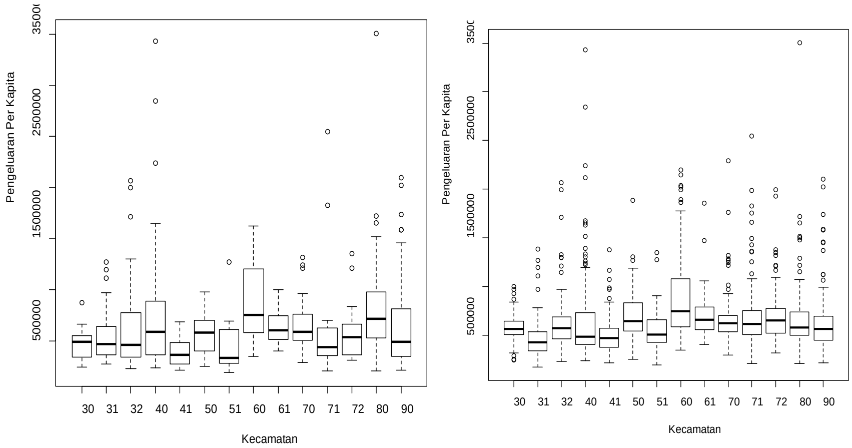

Figure 1: Household per-capita expenditure (left: original sample, right: oversample

The distribution pattern of household per-capita expenditure in Musi Rawas Regency is not symmetric and skew to the right (positive skewness). Outliers are above the boxplot, with the longer top whisker. The outliers will produce bad estimation if we utilize the least square method. So that, the using of robust estimation is necessary as the best alternative method.

3.2. Comparison of Estimation Technique

In direct estimation, the calculation of the HCR, PG, and PS estimate in each sub-district d only used $\mathbf{y}_{jd}$ from the sample group. In M-quantile model, Susenas 2015 data was used as the response variables. The calculation of poverty indicator estimator HCR, PG, and PS and its RRMSE using M-quantile required auxiliary variables. In this study, we used Population Census 2010 data (SP2010) as the auxiliary variables in the household unit level. This study also used three weight functions for M-quantile model. The three functions are Huber, Hampel, and Tukey Bisquare. The three functions have an effect to the estimate of the unit M-quantile coefficient (q_{jd}), the sub-district area M-quantile coefficient ($\hat{\theta}_d$), and sub-district regression coefficient ($\widehat{\boldsymbol{\beta}}_{\psi,d}$).

According to the table 2, there are two sub-districts (Jayaloka and Suka Karya) which have HCR estimate is 0%. It is potentially not represented the reality of poverty in general, maybe because the sample has no incidence of poor category ($\mathbf{y}_{jd} \leq t$). Most HCR direct estimates (with oversampling) produce lower percentage than direct estimates (without oversampling) and indirect M-quantile model.

Three weight functions in M-quantile model produce estimates which are close each other. The mean HCR estimates using Huber, Hampel, and Tukey Bisquare weight functions are 15.58%, 15.70%, and 15.91%, respectively. M-quantile produced greater mean HCR estimate than the direct estimation with oversampling, but smaller than direct estimation without oversampling. The advantage of M-quantile is the ability to estimate HCR which is 0% if using direct estimation (ie Jayaloka and Suka Karya sub-districts). The mean, standard deviation, and median HCR estimate using M-quantile are smaller than direct estimation without oversampling.

Table 2: Head count ratio (HCR) by sub-distrct (%)

Sub-district	M-quantil model			Direct estimation	
	Huber	Hampel	Tukey Bisquare	Without Oversampling	Over-Sampling
STL Ulu	25.12	25.89	25.28	20.00	5.00
Selangit	21.34	21.49	21.19	20.00	27.27
Sumber Harta	15.84	15.78	16.68	26.67	7.50
Tugumulyo	9.03	9.04	8.91	19.32	11.90
Purwodadi	23.48	23.93	25.18	43.33	15.97
Muara Beliti	16.85	17.07	16.84	10.00	1.82
TP Kepungut	19.07	18.81	20.83	60.00	12.00
Jayaloka	4.35	4.35	3.55	0.00	0.00
Suka Karya	7.29	7.56	6.79	0.00	0.00
Muara Kelingi	9.34	9.44	8.74	6.67	1.82
BTS Ulu	15.61	15.80	16.1	16.67	5.17
Tuah Negeri	12.89	12.73	13.03	13.33	3.64
Muara Lakitan	4.68	4.89	4.18	6.35	2.80
Megang Sakti	13.73	13.56	13.91	23.88	10.88
Mean[a]	15.58	15.70	15.91	22.18	8.81
Median[a]	15.73	15.79	16.39	19.66	6.34
Regency Level	13.63	13.67	13.44	17.46	7.63

[a] Not involving jayaloka and suka karya sub-district

Statistics Indonesia has released PG estimate is 2.06%. This value is under of M-quantile estimate or direct estimate without oversampling. The poverty gap in direct estimation with oversampling is smaller because may be the mostly household samples which be taken in Suseda have per-capita expenditure above the poverty line. Poverty gap is the mean shortfalls of the total population from the poverty line (counting the nonpoor as having zero shortfall), expressed as a percentage of the poverty line.

STL Ulu Sub-district has the highest poverty gap if calculated using M-quantile, while Selangit sub-district has the highest poverty gap using direct estimation with oversampling. Table 3 shows that M-quantile model was able to produce poverty gap which is not 0% for Jayaloka and Suka Karya sub-districts.

The poverty gap in direct estimation is always smaller than indirect estimation result. The mean poverty gap using direct estimation with oversampling is 1.33% and without oversampling is 3.65%. Three weight function on M-quantil gives poverty gap estimate close to each other. There are 5.13%, 5.14%, and 4.87% for Huber, Hampel, and Tukey Bisquare weight function, respectively.

Table 3: Poverty gap (PG) by sub-district (%)

Sub-district	M-quantil model			Direct estimation	
	Huber	Hampel	Tukey Bisquare	Without Over-sampling	Over-Sampling
STL Ulu	8.63	8.84	8.11	2.73	0.68
Selangit	7.93	7.87	7.28	2.18	3.82
Sumber Harta	4.89	4.86	4.79	5.19	1.38
Tugumulyo	2.82	2.80	2.53	2.66	1.77
Purwodadi	7.22	7.34	7.29	9.59	3.36
Muara Beliti	5.52	5.55	5.07	2.14	0.39
TP Kepungut	6.79	6.59	6.84	10.63	1.20
Jayaloka	1.44	1.41	1.15	0.00	0.00
Suka Karya	2.18	2.25	1.90	0.00	0.00
Muara Kelingi	3.10	3.12	2.70	0.69	0.19
BTS Ulu	4.91	4.96	4.73	2.36	0.73
Tuah Negeri	4.11	4.03	3.82	0.59	0.16
Muara Lakitan	1.57	1.64	1.37	1.58	0.70
Megang Sakti	4.09	4.02	3.90	3.50	1.59
Mean[a]	5.13	5.14	4.87	3.65	1.33
Median[a]	4.90	4.91	4.76	2.51	0.97
Regency Level	4.31	4.31	3.91	2.80	1.16

[a] Not involving jayaloka and suka karya sub-district

The largest poverty severity (PS) estimate is Purwodadi sub-district in direct estimation with oversampling and STL Ulu in M-quantile. This high poverty severity indicates the high imbalances of household per-capita expenditure among the poor population. The smallest poverty severity is Muara Kelingi sub-district in direct estimation with oversampling and Muara Lakitan Sub-district in M-quantil.

There are two sub-districts which have PS 0% ie Jayaloka and Suka Karya sub-districts. That value is due to the inexistence of poverty incidence (HCR = 0%). The advantages of M-quantil is the ability to produce PS estimate which in the direct estimate is 0% (Jayaloka and Suka Karya sub-districts).

The mean PS estimate from all sub-districts in direct estimation was 0.90 percent for non-oversampling, 0.32 for oversampling, and M-quantil using Huber, Hampel and Tukey Bisquare function are 2.85%, 2.83%, and 2.56% respectively.

Table 5 summarizes the mean RRMSE for HCR, PG, and PSE estimates in direct estimation without oversampling, direct estimation with oversampling, and M-quantile model.

Table 4: Poverty severity (PS) by sub-district (%)

Sub-district	M-quantil model			Direct estimation	
	Huber	Hampel	Tukey Bisquare	Without Over-sampling	Over-Sampling
STL Ulu	4.88	4.97	4.35	0.62	0.16
Selangit	4.78	4.70	4.17	0.41	1.00
Sumber Harta	2.59	2.56	2.41	1.36	0.35
Tugumulyo	1.52	1.50	1.30	0.51	0.36
Purwodadi	3.73	3.77	3.56	2.41	0.82
Muara Beliti	3.04	3.05	2.63	0.48	0.09
TP Kepungut	3.96	3.81	3.75	2.99	0.31
Jayaloka	0.82	0.79	0.64	0.00	0.00
Suka Karya	1.16	1.20	0.97	0.00	0.00
Muara Kelingi	1.76	1.76	1.47	0.09	0.02
BTS Ulu	2.65	2.66	2.41	0.62	0.19
Tuah Negeri	2.25	2.18	1.98	0.05	0.01
Muara Lakitan	0.91	0.94	0.78	0.54	0.24
Megang Sakti	2.10	2.05	1.90	0.73	0.33
Mean[a]	2.85	2.83	2.56	0.90	0.32
Median[a]	2.62	2.61	2.41	0.58	0.27
Regency Level	2.30	2.30	1.95	0.66	0.28

[a] Not involving jayaloka and suka karya sub-district

Table 5: Mean RRMSE of Poverty Indicator Estimates (%)

Estimation technique	HCR	PG	PS
Direct estimation			
Without versampling	39.20	48.84	57.21
With Oversampling	38.29	47.20	55.94
Indirect estimation (M-quantile)			
Huber	32.87	43.44	51.44
Hampel	32.08	42.77	51.63
Tukey Bisquare	36.30	51.82	64.00

In general, M-quantil model produces smallest mean RRMSE (HCR, PG, PS) than direct estimation either with or without oversampling. The weight function which produces smallest mean RRMSE (HCR) is Hampel as well as the mean RRMSE (PG). However, the weight function that produces the smallest mean the RRMSE (PS) is Huber function. This empirical study has shown that M-quantil is more feasible to use. The oversampling in Suseda is able in increasing the accuracy using direct estimates, but it has lower accuracy than M-quantile approach.

4. Conclusion

In direct estimation result, there are two sub-districts which have 0% for poverty indicator estimates (HCR, PG, and PS) either without or with oversampling. It indicates that oversampling only increase the accuracy of the estimation. On the contrary, M-quantile is able to calculate the poverty indicator estimate of some sub-district which have 0% if we use direct estimation. Based on mean RRMSE, M-quantil shows more accurate estimates than direct estimation either without or with oversampling. We should also select appropriate weight function to produce more accurate estimate. Our result shows that Huber and Hampel weight functions were better than tukey bisquare weight function because they give lower mean RRMSE. Based on this research, we conclude that the ability of m-quantile can overcome the additional survey (with oversample data). Oversampling is not necessary to be conducted again because of its inefficiency in terms of human resources, cost, and time.

5. Recommendation

Further research is necessary for estimating poverty indicator until village area level, even household level. Moreover, we can utilize the household and individual weight value to increase the accuracy of estimates. We also recommend to further researcher, to utilize estimation technique based on multistage random sampling as Statistics Indonesia (BPS) designed officially. Beside that, we recommend to implement this M-quantile model to other area which has data characteristics similarity so the model will give strong evidence in term of the accuracy and benefit.

Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the Editors. These led to a considerable improvement in the paper. This work is financially supported by National Statistical Office (BPS) scholarship program in collaboration with Department of Statistics at Bogor Agricultural University.

References

[1] I. Molina, J.N.K. Rao. (2010). “Small area estimation of poverty indicators”, The Canadian Journal of Statistics. .38(3),pp.369-385.

[2] C. Giusti, S. Marchetti, M. Pratesi, N. Salvati. 2012. “Robust small area estimation and oversampling in the estimation of poverty indicators”, Survey Research Methods. 6(3),pp.155-163.

[3] A.B. Kennickell, “The Role of Over-sampling of the Wealthy in the Survey of Consumer Finances”. In ISI 56th conference. 2007.

[4] J.N.K. Rao. Small Area Estimation. New York, US: John Wiley and Sons. 2003.

[5] Badan Pusat Statistik. Data dan Informasi Kemiskinan Ka-bupaten/ Kota 2015. Jakarta: Badan Pusat

Statistik. 2016.

[6] R. Chambers, N. Tzavidis. 2006. “M-Quantile Models for small area estimation”, Biometrika. 93(2),pp.255-268.

[7] N. Tzavidis, S. Marchetti, R. Chambers. 2010. “Robust Estimation of Small Area Means and Quantiles”. Australian and New Zealand Journal of Statistics. 52(2),pp.167-186.

[8] S. Marchetti, N. Tzavidis, M. Pratesi. 2012. ”Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators”, Computational Statstics and Data Analysis, 56,pp.2889-2902.

[9] Girinoto. “Kajian pendugaan area kecil untuk indikator kem-iskinan melalui pendekatan regresi kekar M-kuantil (studi kasus: tingkat kecamatan pada Kabupaten Bogor) “ M.Si thesis. Institut Pertanian Bogor, Bogor. 2017.

[10] J. Foster, J. Greer, E. Thorbecke. 1984 “A class of decomposable poverty measures”, Econometrica, 52(3), pp.761-766.