Aldi Rochman Nulkarim, Ika Yuni Wulansari

M-Quantile Small Area Estimation for Household Per Capita Income: A Case Study of Sub-Districts in Special Region Yogyakarta, Indonesia

Aldi Rochman Nulkarim¹, Ika Yuni Wulansari²

¹ STIS Polytechnic of Statistics, Jakarta, Indonesia, 211709522@stis.ac.id

Abstract:
As the needs for local area indicators are growing, Statistics Indonesia has begun to produce small area indicators using popular Small Area Estimation (SAE) methods such as Empirical Best Linear Unbiased Predictor under the Fay-Heriot model (EBLUP-FH) since 2019. One of the important economic welfare indicators that need to be measured on a micro-scale as a compass to the local area development is the household income per capita. It has two natural characteristics: the data tend to be right-skewed and contain lots of outliers. However, The EBLUP-FH, as a parametric approach depends on strong distributional assumptions and can easily be influenced by outliers. Hence, this study attempts to implement an alternative approach in estimating household income per capita at a sub-districts level using M-quantile small area model. The M-quantile model avoids specifying the random effect and allows for robust inference against outliers. The data used are the Socio-Economic Survey and Potential Village Data. The Relative Root Mean Squared Error (RRMSE) produced by the M-quantile model is smaller than both in the EBLUP-FH model and direct estimation. It can be said that the M-quantile model gives a more precise estimation than the EBLUP-FH model and direct estimation.

Keywords:
Small Area Estimation, M-quantile model, Household Per Capita Income

1. Introduction
Over the last few decades, there has been increasing demands on small area statistics. The Indonesian government system's change from centralized to decentralized has pushed the growth of local area policy. The local policymakers need to know statistical indicators in their respective areas to make the formulated policies suitable in their area circumstances. Unfortunately, a regular survey does not satisfy local statistics needs. The cost constrains in survey design can lead to a small sample size within small areas. This lack of sample size within direct estimation using survey data will cause unacceptable precision. One reliable approach used to solve this problem is using Small Area Estimation (SAE) methods. SAE yields more efficient estimates as it borrows strength from auxiliary information obtained from censuses or administrative records.

Since 2019, Statistics Indonesia - Badan Pusat Statistik (BPS) has begun to produce small area indicators using popular SAE method such as Empirical Best Linear Unbiased Predictor under the Fay-Heriot model (EBLUP-FH). EBLUP-FH model is the random effect model where its random area effect explaining between area variation beyond that explained by covariates (Fay & Herriot, 1979; Rao, 2003). Such a model is classified as a parametric approach where its random parts depend on strong distributional assumption and can easily be influenced by outliers. In practice, however, this distributional assumption is difficult to satisfy, and the real data usually have outliers. Chambers & Tzavidis (2006) have offered an alternative solution to overcome these disadvantages using M-quantile small area model. Unlike the EBLUP-FH model, the M-quantile model does not require distributional assumptions and robust inference against outliers. It can be done because the M-quantile model does not require formal specification of random effect but using the variation of area-specific M-quantile coefficients to explain inter-domain differences instead.
Household per capita income often becomes attention in SAE. It is one of the crucial economic welfare indicators that need to be measured on a micro-scale as a compass to the local area development. Therefore, the accuracy of household per capita income estimation is a crucial issue. BPS obtained household per capita income data from a survey named Socio-Economic Survey (SUSENAS). Its sample size is designed to estimate the parameter at the district level. As a result, its small area estimates yield unacceptable levels of precision. The EBLUP-FH model can be an alternative method to estimate household per capita income at small areas level. However, according to Barigozzi et al. (2009), household per capita income data tend to be right-skewed. Battistin et al. (2009) also report that household per capita income data naturally have outliers. These two natural characteristics of household per capita income data can lead to difficulties in satisfying distributional assumptions and formal specification in random parts of the EBLUP-FH model. Hence, the M-quantile small area model, which avoids problem-related to the specification of random effect, is expected to improve the precision of EBLUP-FH in estimating small area level household per capita income.

Based on that, this study inspects the estimated mean of household income per capita at the sub-district level, Yogyakarta Province in 2018, by using direct estimation, EBLUP-FH, and M-quantile small area model. According to BPS, Yogyakarta had the highest Gini ratio in all provinces in 2018. Hence, it expected to have the highest variation of household income per capita. So, the data are expected to contain lots of outliers and be right-skewed distribution. Based on that results, the following analysis compares the performance of EBLUP-FH and M-quantile, evaluated by Root Mean Square Error (RMSE) and Relative Root Mean Square Error (RRMSE). Hence, we can determine the most reliable methods to produce household income estimates with the highest precision at the small area level.

2. Methodology
Data Used in the Study
The data in this study is secondary data from BPS microdata. The data come from two sources. The first part is household per capita income as the variable of interest extracted from SUSENAS 2018. The second part is SAE covariates obtained from village potential data (PODES) 2018 as administrative records. Furthermore, SUSENAS data are at the household level with 3667 samples household and 233 samples village, while PODES data are at the village level with 416 villages. Both data will be used to estimate the average household income per capita of each 74 sub-districts.

Covariates used in the SAE model are as follows: number of senior high schools (X1), number of auxiliary health centres (X2), and number of doctor’s offices (X3). These covariates are selected based on Pearson correlation values higher than 0.5 and optimal Akaike Information Criterion (AIC) value.

M-quantile Small Area Model
Chambers & Tzavidis (2006) have introduced M-quantile as an alternative approach in small area estimation. Its framework is based on the use of M-quantile regression. Moreover, M-quantile regression is a generalization of quantile regression by integrating the influence function to the M-estimation. Its framework is based on the use of M-quantile approach in estimating small area parameter.

The general linear mixed model in SAE as follows:

\[ y_{ij} = x_{ij}^T \beta + v_i + \epsilon_{ij}, \quad j = 1, ..., N, \ i = 1, ..., m \]

Where \( y_{ij} \) is variable of interest, \( x_{ij} \) are covariates, \( v_i \) denotes random area effect of each domain and \( \epsilon_{ij} \) represents random individual effect. Both of them are assumed to be normally distributed. Suppose \( \hat{m}_j \) is small area statistics estimated by EBLUP.

\[
\hat{m}_j = N_j^{-1} \left[ \sum y_i + \sum x_i^T \hat{\beta} + z_i^T \hat{\theta}_j \right]
\]

\( v_i \) in the model has a role to characterize the differences of the conditional distribution \( y \) given \( x \) among small areas. On the other hand, by using the M-quantile model, Chambers & Tzavidis (2006) show \( m_j \) can be estimated as follows.

\[
\hat{m}_j = N_j^{-1} \left[ \sum y_i + \sum x_i^T \hat{\beta}_{q} \right]
\]
Where \( \psi \) denotes the influence function related to the M-quantile model. In contrast, M-quantile small area model can avoid specifying random effect by using variations of area-level coefficient \( \hat{\theta} \) to characterize inter-domain differences.

The Mean Squared Error (MSE) of the M-quantile small area model is described in Chambers & Tzavidis (2006). Furthermore, the form of MSE estimator for area \( j \) can be expressed as follows.

\[
\hat{m}_j = \hat{V}(\hat{m}_j) + \left[ \hat{B}(\hat{m}_j) \right]^2
\]

Where \( \hat{V}(\hat{m}_j) \) is variance estimator for area \( j \) and can be written as

\[
\hat{V}(\hat{m}_j) = N_j^{-2} \sum_g \sum_{i \in s_g} \left\{ \left( w_i - 1 \right)^2 + \frac{N_j - n_j}{n_j - 1} \right\} I(g = j) + w_i^2 I(g \neq j) \cdot \{ y_i - x_i^T \hat{\beta}_g(\hat{\theta}_g) \}^2
\]

With \( w_i \) is a set of weight produced by the iteratively reweighted least squares algorithm and \( \hat{B}(\hat{m}_j) \) is the form of bias estimator for area \( j \) and can be written as

\[
\hat{B}(\hat{m}_j) = N_j^{-1} \left[ \sum_g \sum_{i \in s_g} w_i x_i^T \hat{\beta}_g(\hat{\theta}_g) - \sum_{i \in j} x_i^T \hat{\beta}_g(\hat{\theta}_g) \right]
\]

3. Result

Estimation of Average Household Per Capita Income

The direct estimation in this study uses a design-based method where it refers to SUSENAS design, which is two stages one phase stratified sampling. It estimates average household per capita income at the sub-district level using the SUSENAS sampling design, then calculating the Standard Error (SE) and Relative Standard Error (RSE) to examine levels of precision. Next, the estimating process of the EBLUP-FH model uses average household per capita income resulted from previous direct estimation. M-quantile small area model is a unit-level model approach. The village is used as unit-level in the M-quantile model, so it has the same level as its covariates. Descriptive statistics in the three tables summarize the results.

| Table 1. Descriptive Statistics on Average Household Per Capita Income Based on Direct Estimation |
|---------------------------------------------------|-----------------|-----------------|
| Statistics | Estimates | SE | RSE |
| Minimum | IDR 477,237 | IDR 3,541 | 0.47 % |
| Median | IDR 1,069,362 | IDR 131,898 | 12.68 % |
| Mean | IDR 1,322,941 | IDR 191,801 | 13.24 % |
| Maximum | IDR 3,060,216 | IDR 840,470 | 42.16 % |
| Std. Deviation | IDR 640,864 | IDR 180,488 | 8.15 % |

| Table 2. Descriptive Statistics on Average Household Per Capita Income Based on EBLUP-FH model |
|---------------------------------------------------|-----------------|-----------------|
| Statistics | Estimates | RMSE | RRMSE |
| Minimum | IDR 505,629 | IDR 3,541 | 0.47 % |
| Median | IDR 1,063,822 | IDR 118,109 | 10.61 % |
| Mean | IDR 1,223,023 | IDR 126,459 | 10.31 % |
| Maximum | IDR 3,172,364 | IDR 254,312 | 20.69 % |
| Std. Deviation | IDR 504,684 | IDR 67,431 | 4.49 % |

| Table 3. Descriptive Statistics on Average Household Per Capita Income Based on M-quantile model |
|---------------------------------------------------|-----------------|-----------------|
| Statistics | Estimates | RMSE | RRMSE |
| Minimum | IDR 467,374 | IDR 5,514 | 0.68 % |
| Median | IDR 1,028,234 | IDR 34,647 | 1.92 % |
| Mean | IDR 1,147,823 | IDR 55,768 | 3.04 % |
| Maximum | IDR 2,528,080 | IDR 374,878 | 23.13 % |
| Std. Deviation | IDR 477,848 | IDR 65,835 | 4.90 % |

Based on the results, Table 1 shows the maximum RSE produced by direct estimation is at an unacceptable level of precision. So, this result suggests the use of SAE to improve the level of precision. Next, Table 2 shows that the EBLUP-FH model's estimates have a better RRMSE value than the RSE of direct estimation. Hence, the EBLUP-FH model is immensely improving levels of precision. Then it
can be seen in Table 3, estimates produced by the M-quantile model yield the lowest RRMSE value on average. It means, the M-quantile small area model does improve the levels of precision for both direct estimation and the EBLUP-FH model.

Comparison of RRMSE
The comparison of the RRMSE value examines which estimation methods provide the best estimation result and see the RRMSE value's stability among the three methods. This part presents each district’s RRMSE value where their sample size sorts the districts in descending order.

Figure 1. The Comparison of RRMSE on Average Household Per Capita Income

Based on Figure 1, it can be shown that the RRMSE of the M-quantile model is the best and slightly more stable, especially from the left side to the middle when the sample size within sub-districts is quite large. Unlike the EBLUP-FH model, the M-quantile model is not seen to be influenced by direct estimation. Therefore, it can be said that the M-quantile small area model is the best method in estimating the average household per capita income at the sub-districts level in Yogyakarta Province.

4. Discussion and Conclusion
This study attempts to implement the M-quantile small area model as an alternative approach in small area estimations in Indonesia. The study estimates household income per capita at the sub-district level in Yogyakarta Province as the region with the highest Gini ratio. So, the data are expected to contain lots of outliers and violate the normality assumption. The results show that the M-quantile model produces estimates with the best levels of precision. Furthermore, The RRMSE of the M-quantile model offers higher stability when the sample size within sub-districts is quite large.

On the contrary, the RRMSE of EBLUP-FH is quite unstable regardless of its sample size. So far, based on the findings in this study, it can be concluded that the M-quantile small area model gives a more precise estimation than the EBLUP-FH model and direct estimation. However, we should be careful to use the M-quantile model when the sample size within small areas is very small.

References: