A comparative study of multivariate Fay-Herriot model for small area estimation in various sample sizes

W A Nurizza\textsuperscript{1*} and A Ubaidillah\textsuperscript{1}

\textsuperscript{1}Department of Statistics, STIS Polytechnic of Statistics, East Jakarta, 13330, Indonesia

*E-mail: 14.8432@stis.ac.id

Abstract. In general, surveys are designed for large areas with sufficient sample size. If the survey is used for small areas in which the sample sizes are not sufficient, the results of estimates may not be reliable due to the large standard error. Therefore, a Small Area Estimation (SAE) method was developed, to increase the effectiveness of the sample size by borrowing the strength of the neighboring region and information from the auxiliary variables that have a strong relationship with the observational variable. This study aims to analyze the SAE using Multivariate Fay-Herriot (MFH) model and Univariate Fay-Herriot (UFH) model for a variety of sample sizes. Simulations were conducted by using household expenditure per capita of food group and non-food group data from Susenas on March 2017. The simulation results showed that the average Root Mean Square Errors (RMSEs) using the MFH models in various sample size are smaller than the UFH model and the direct estimation.

1. Introduction
The accuracy of small area statistics is a problem being faced by statisticians in the last few decades [1]. In the current era of regional autonomy, the government also needs accurate data and information that coverage to the smallest area. The government uses these data as a basis for policy formulation. Data accuracy is expected to make government policies more targeted. So, the accuracy of small area statistics is needed by many users.

In general, the survey was designed for large areas. If the survey design is used for small area estimates, the result of an estimate is not reliable because it will produce a large standard error. This problem can be overcome by Small Area Estimation (SAE). SAE was first applied by Fay and Herriot to estimate capita income. According to Kurnia and Notodiputro [2] in accordance with Rao and Molina [3] that SAE is a statistical technique to estimate the parameters of subpopulations with small sample size. SAE can increase the effectiveness of sample size by borrowing the strength of neighboring areas and information from the auxiliary variables that have a strong relationship with observational variables [4]. Therefore, the result of estimates by SAE is more accurate for estimations in small areas.

One indicator that is still being a focus by the government is the household expenditure per capita. The small area estimation method that is suitable for use is Empirical Best Linear Unbiased Prediction (EBLUP) because the household expenditure per capita has a continuous response. Fay-Herriot in Benavente and Morales [5] also uses EBLUP method to estimate the capita income. It is known that the calculation of capita income in Indonesia is also approached by calculating capital expenditure.

The household expenditure per capita can be distinguished based on the expenditure group, which is a food group and non-food group. Based on Susenas [6], provinces with high household expenditure per
capita of food group also tend to have a high household expenditure per capita of a non-food group. This indicates a strong correlation between the variables of the household expenditure per capita of food group and the non-food group.

Datta et al. in Benavent and Morales [5], mentioning that by utilizing the advantages of the correlation between observed variables, the Multivariate Fay-Herriot (MFH) model produces more efficient parameter estimates compared with the Univariate Fay-Herriot (UFH) model. Still, in Benavent and Morales [5], estimates based on the MFH model have a Root Mean Square Error (RMSEs) smaller compared to the UFH model. Therefore, MFH model is recommended to be used in data which has a strong correlation.

The estimated RMSEs for each model will be compared by using a simulation study. Such simulations have been carried out in other studies, including by Ghosh and Rao [7]. This study compared the estimated RMSEs for direct estimation, UFH model, and MFH model through some simulations, as have been carried out in other works, including by Ghosh and Rao [7].

BPS has carried out estimates up to the district/city level with a sufficient number of samples using the National Socio-Economic Survey (Susenas) data in March 2017. Therefore, the estimation will be carried out at the district/city level with simulations on different sample size. So, this study aims to analyze the SAE using the MFH model and the UFH model for a variety of sample sizes. Simulations were conducted by using household expenditure per capita of food group and non-food group data from Susenas on March 2017. The simulation results showed that the average RMSEs using the MFH models in various sample size are smaller than the UFH model and the direct estimation.

2. Materials
This study using secondary data from Susenas March 2017 obtained from Statistics Indonesia or BPS. There are two dependent variables, the household expenditure per capita of food group and non-food group in 511 districts/cities in Indonesia. The household expenditure per capita of food group as \( y_1 \) and the household expenditure per capita of the non-food group as \( y_2 \). In addition, Potensi Desa (Podes) data from BPS will be used as an auxiliary variable in indirect estimation. The following are the details of the auxiliary variables: the family of PLN electricity users (\( x_1 \)), families of non-PLN electricity users (\( x_2 \)), settlement locations on river banks (\( x_3 \)), Indonesian workers/TKI (\( x_4 \)), SD/MI education institutions (\( x_5 \)), vocational education institutions (\( x_6 \)), academy/college education institutions (\( x_7 \)), health centers without hospitalization (\( x_8 \)), Puskesmas Pembantu (\( x_9 \)), polyclinics (\( x_{10} \)), doctor’s practice (\( x_{11} \)), midwife practice (\( x_{12} \)), Polindes (\( x_{13} \)), Posyandu (\( x_{14} \)), pharmacy (\( x_{15} \)), micro and small industry (\( x_{16} \)), restaurant (\( x_{17} \)), and lodging (\( x_{18} \)).

3. Methods
In this study, the estimation method used is Empirical Best Linear Unbiased Prediction (EBLUP) with UFH and MFH models. The parameters to be estimated are the household expenditure per capita of food group and non-food group. Before carrying out the analysis, the preparation of the data will be used first. Calculation of the household expenditure per capita of food group and non-food group by dividing each the household expenditure by the number of the household members. For Podes data, the calculation is done by rowing each variable value with the population (\( x \times 10,000 \)).

The analysis steps used are as follows: Direct estimation of parameters and estimated variance and covariance data on the household expenditure per capita of food group and non-food group for each district/city using a formula that corresponds to Benavent and Morales [5]. The direct estimation of the household expenditure per capita in each district/city (Formula 1) with \( \hat{Y}_{dir} \) is the household expenditure per capita in each district/city for population, \( w_i \) is household weighted, \( \hat{y}_{dir} \) is household expenditure per capita, and \( N_{dir} \) is a number of the household population.
\[
\hat{y}_\text{dir} = \frac{\hat{y}_\text{dir}}{\hat{N}_\text{dir}},
\]

where \( \hat{y}_\text{dir} = \sum_{i=1}^{n} w_i y_{i\text{dir}} \) and \( \hat{N}_\text{dir} = \sum_{i=1}^{n} w_i \).

And here are variants of the household expenditure per capita of food group (Formula 2), variants of the household expenditure per capita of a non-food group (Formula 3), and covariances of the household expenditure per capita of food group and non-food group (Formula 4) (Benavent and Morales, 2015). The code "1" for the household expenditure per capita of food group and the code "2" for the household expenditure per capita of a non-food group.

\[
\text{cov}(\hat{y}_{\text{dir}1}, \hat{y}_{\text{dir}1}) = \sum_{i=1}^{n} w_i (w_i - 1) (y_{i\text{dir}1} - \bar{y}_{\text{dir}1})(y_{i\text{dir}1} - \bar{y}_{\text{dir}1}),
\]

\[
\text{cov}(\hat{y}_{\text{dir}2}, \hat{y}_{\text{dir}1}) = \sum_{i=1}^{n} w_i (w_i - 1) (y_{i\text{dir}2} - \bar{y}_{\text{dir}2})(y_{i\text{dir}1} - \bar{y}_{\text{dir}1}),
\]

\[
\text{cov}(\hat{y}_{\text{dir}2}, \hat{y}_{\text{dir}2}) = \sum_{i=1}^{n} w_i (w_i - 1) (y_{i\text{dir}2} - \bar{y}_{\text{dir}2})(y_{i\text{dir}2} - \bar{y}_{\text{dir}2}),
\]

Fay-Herriot Multivariate in General Linear Mixed Model (GLMM) written as follows [3]:

\[
y = X\beta + Zu + e, \quad u \sim N(0, G), \quad e \sim N(0, R),
\]

where \( u = \text{col}_{1 \leq d \leq D}(u_d) \) and \( e = \text{col}_{1 \leq d \leq D}(e_d) \) is independent with \( D \) is the amount of interest variable. \( y = \text{col}_{1 \leq d \leq D}(y_d) \) is vector of interest variable with \( y_d = (y_{1d}, ..., y_{md})^T \), \( X = \text{col}_{1 \leq d \leq D}(X_d) \), \( X_d = \text{col}_{1 \leq d \leq D}(x_{id}) \) is matrix of auxiliary variable with \( x_{id} = (x_{1i}, ..., x_{pi})^T \). \( G = V_u \otimes I_m \), is covariance matrix from area effect and \( R \) is covariant sampling matrix with size \( Dm \times Dm \) which is assumed to be known from sampling error survey in general.

EBLUP estimator \( \hat{\mu} \) can be written with this formula with \( X \) is matrix of auxiliary variables \( (m \times p) \), \( \beta \) is vector regression coefficient \( (p \times 1) \), \( Z \) is positive constant matrix, and \( \Omega \) is a covariance matrix of \( y \) (Rao dan Molina [3] in Ubaidillah [4]):

\[
\hat{\mu} = X\hat{\beta} + Z\hat{\Omega}^{-1}(y - X\hat{\beta}),
\]

\[
\hat{\beta} = (X^T\hat{\Omega}^{-1}X)^{-1}X^T\hat{\Omega}^{-1}y,
\]

Multivariate MSE (\( \text{MSE}(\hat{\mu}) \)) and covariance (\( \text{cov}(\hat{\mu}) \)) of EBLUP written as follows [4]:

\[
\text{MSE}(\hat{\mu}) = \text{col}_{1 \leq j \leq Dm}[\text{cov}(\hat{\mu})], \quad j = 1, ..., Dm,
\]

\[
\text{cov}(\hat{\mu}) \approx \varphi_1(\hat{\sigma}_u^2) + \varphi_2(\hat{\sigma}_u^2) + 2\varphi_3(\hat{\sigma}_u^2)
\]
with
\[
\varphi_1(\hat{\sigma}_u^2) = \Gamma R, \\
\varphi_2(\hat{\sigma}_u^2) = (1 - \Gamma)^2 X (X^T \hat{\Omega}^{-1} X)^{-1} X^T (1 - \Gamma)^T, \\
\varphi_3(\hat{\sigma}_u^2) \approx \sum_{k=1}^{q} \sum_{l=1}^{q} \text{cov}(\hat{\sigma}_{uk}, \hat{\sigma}_{ul}) \Gamma_{(k)} \Omega \Gamma^T_{(l)},
\]
(9)

where \( \Gamma_{(k)} = \frac{\partial \psi}{\partial \sigma_u^2} \) with \( \Gamma = ZGZ^T \hat{\Omega}^{-1} \), and \( k, l = 1, ..., q \). \( \text{cov}(\hat{\sigma}_{uk}, \hat{\sigma}_{ul}) \) is element from Fisher information in REML estimator. \( g_t(\hat{\sigma}_u^2) = \text{col}_{15} \text{sdm}([\varphi_t(\hat{\sigma}_u^2)]_j \text{ and } [\varphi_t(\hat{\sigma}_u^2)]_j \) is diagonal of \( j \)-th element diagonal matrix of \( \varphi_t(\hat{\sigma}_u^2) \) with \( t = 1, 2, 3 \).

Then, resampling is done with Simple Random Sampling (SRS) method. The steps resampling taken were as follows:

1. First, sampling was carried out with different sample sizes, namely 10 percent, 25 percent, 40 percent, 60 percent, 75 percent, and 90 percent of the total sample of Susenas in each district/city. Sampling is carried out repeatedly or iterated 100 times so that the results of representative sampling.

2. Second, direct estimation of parameters, variance, and covariance on household expenditure per capita of food group and non-food group data for each district/city.

3. Next, the selection of auxiliary variables. The auxiliary variable that used is the best auxiliary variable results of regression between parameters with variables that do not have an error. So, in this study used auxiliary variables from Potensi Desa data. Selection is done by backward elimination.

4. Next, indirect estimation by EBLUP method on household expenditure per capita of food group and non-food group data for each district/city. And the estimation of random effect variance using the Restricted Maximum Likelihood (REML) method. (Indirect estimation is an estimation method that borrows strength from a sample domain and includes an auxiliary variable that comes from a census or registration [8].

5. Calculate the average of estimated parameters and the RMSEs of average per capita food and non-food household expenditure.

6. Comparing the results of the average RMSEs using the Fay-Herriot Multivariate model, the Fay-Herriot Univariate model, and direct estimates.

4. Results and Discussion

Data simulation results show the household expenditure per capita of food group and non-food group using direct estimation, UFH model, and MFH model as follows.

In indirect estimation using UFH and MFH model, almost all of the auxiliary variables used 5 percent significance level. The auxiliary variables that influence each the data on average household expenditure per capita of food group and non-food group. As many as 14 auxiliary variables were obtained for data on the household expenditure per capita of food group. These variables are the auxiliary variables with codes \( x_2, x_3, x_4, x_5, x_6, x_7, x_9, x_{10}, x_{11}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18} \), and \( x_{18} \). And obtained as many as 13 auxiliary variables for the household expenditure per capita of non-food group, namely \( x_1, x_2, x_4, x_5, x_6, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, \) and \( x_{17} \).

Variables resulting from the selection are variables that have a joint influence on the household expenditure per capita of food group and non-food group. Some auxiliary variables are obtained for the
household expenditure per capita of food group, and some others are obtained for the household expenditure per capita of a non-food group. There are some similar variables on the two data because each variable has a different influence on the household expenditure per capita of food group and non-food group. The selected auxiliary variables are also free from multicollinearity.

The auxiliary variables aside obtained from the Podes data, 5 dummy variables based on the economic corridors of the Indonesian region were added in the model. The addition of dummy variables is done because the districts/cities to be studied are spread across several islands that have different geographical characteristics. The dummy variables used are $d_1$, $d_2$, $d_3$, $d_4$, and $d_5$ for districts/cities in succession on the islands of Sumatra, Java, Kalimantan, Sulawesi, and Bali and Nusa Tenggara.

The auxiliary variables $x_1$ to $x_{18}$ and the dummy variables $d_1$ to $d_5$ stands for increasing the effectiveness of the sample size by borrowing the strength of information from these variables. The variables used are attempted to not contain error so that the SAE model has a minimum error. So, this is the advantage of SAE with a small number of a sample can produce a model with minimum error.

First formed UFH model that provides a value of variance random effect (ref var) of 0.0278 for the household expenditure per capita of food group and 0.0263 for the household expenditure per capita of a non-food group. While the use of the MFH model provides a ref var value of 0.0438 for the household expenditure per capita of food group and 0.0413 for the household expenditure per capita of a non-food group. A ref var value which is not 0 means that there is an effect of random on the results of the estimation that has been done. Therefore, modeling using small area estimation is feasible.

The following are given Figures 1 and 2 which illustrate the estimated RMSEs values between the direct estimation method, UFH model, and MFH model with different sample sizes, namely on the sample size of 10 percent, 25 percent, 40 percent, 60 percent 75 percent and 90 percent. Due to the process resampling, an iteration process is carried out so that the estimated value produced can be generalized. The displayed RMSEs value is the average result of RMSEs for all district/cities in Indonesia and carried out as many as 100 iterations.

![Figure 1. The RMSEs of $y_i$ boxplot](image-url)
In Figure 1 is seen that in each sample size, MFH model have RMSEs value lower than RMSEs value using UFH model and direct estimation. It can also be seen that the RMSEs values produced in each sample size have different ranges (see the RMSEs value on the vertical axis on each sample size). The larger the sample size used, the smaller the range of RMSEs values generated. For example, the average RMSEs value generated at 90 percent sample size will have a smaller average RMSEs value than the larger sample size.

The comparison of RMSEs value can also be seen through the comparison in different sample sizes. The RMSEs value using the direct estimation method on a 10 percent sample size has a greater value than the RMSEs value generated by the direct estimation method on larger sample size. In the use of UFH model and MFH model, the same pattern was also found. The RMSEs values generated using UFH model and MFH model on a 10 percent sample size also have a greater value than the RMSEs values generated with the model on larger sample size. The number of sample sizes used, the smaller the RMSEs value.

![Figure 2](image.png)

The results shown in the household expenditure per capita of the non-food group are also in line with the results of the household expenditure per capita of food group. In Figure 2, it can be seen that in each sample size, the average estimation of MFH model produces a lower RMSEs value than the RMSEs value with UFH model and direct estimation. The larger the sample size used, the smaller the range of RMSEs values produced.

The same pattern is also shown by the household expenditure per capita of a non-food group. That is showed if the same model is compared to each sample size used. The result of RMSEs values using direct estimation method, UFH model, and MFH model on 10 percent sample size has a greater value than the RMSEs value generated by the same estimation method on larger sample size. The resulting RMSEs value gets smaller as the number of sample sizes is used.

When viewed in more detail, the RMSEs values generated with MFH model have shorter boxplot ranges than boxplots produced with UFH model and direct estimates. This shows that the RMSEs values generated by MFH model tend to be more homogeneous than UFH model and direct estimates for all the iterations performed.
In Figure 3, the average RMSEs description of average capita household expenditure on different sample sizes is obtained. The lines formed for each method are sloping as the sample size increases. The larger the sample size, the smaller the average RMSEs produced. It is clear that the difference in the average RMSEs in the 10 percent sample size ranges from 0.1 to the direct estimation method, while the 100 percent sample size is only around 0.03.

In Figure 4, the average RMSEs of $y_2$ are shown for different sample sizes. The trend is similar to that of $y_1$, with the average RMSEs decreasing as the sample size increases. The difference in RMSEs is also evident, with a range of 0.1 to 0.03 for the 10 percent sample size and around 0.03 for the 100 percent sample size.
In Figure 4, the average RMSEs description of the household expenditure per capita of the non-food group is obtained on different sample sizes. As with Figure 3, the lines formed for each method are sloping as the sample size increases. The larger the sample size, the smaller the average RMSEs produced. It was clear that the difference in the average RMSEs in the 10 percent sample size ranged from 0.15 to the direct estimation method, while the 100 percent sample size was only around 0.045.

When compared between methods, the average RMSEs using MFH model always has a smaller average value than using UFH model and direct estimation. This applies to each different sample size. This shows that using the MFH model is more efficient than using UFH model and the direct estimation method. Besides that, based on Figure 3 and Figure 4, it can be seen that the UFH model as a general model has provided a smaller RMSEs value compared to direct estimation. So when the MFH model is applied, the RMSEs value generated is much smaller. And the gap that occurs between the three methods will widen when the sample used is less.

The MFH model has more advantages compared with the UFH model and direct model. That is because the MFH model besides receiving benefits from using the SAE model as the UFH model, the MFH model also receiving benefits from utilizing the correlation between the observational variables. So that the MFH model can produce more efficient parameter estimates when the observational variables have a strong correlation.

5. Conclusion
Based on the results and previous discussions, some conclusions were drawn on the household expenditure per capita of food group and non-food group by district/city level in Indonesia in 2017:

1. The simulation results show that the average RMSEs by using the MFH model is smaller than the average RMSEs by using the UFH model and direct estimation.

2. Based on the conclusion of point 1 it can be concluded in general that the MFH model is better than using the UFH model and direct estimation.

This paper still has many weaknesses that can be developed in future research, such as: apply on other data that have a strong relationship or increase the number of observational variables, apply with a more complex estimation method, add time and spatial aspects, etc.

References
[1] Dewi L 2006 Penerapan Metode Empirical Bayes Pada Model Small Area Estimation Dalam Pendugaan Pengeluaran Per Kapita di Kota Batam, Bogor: IPB
[2] Kurnia A and Notodiputro KA 2006 EB-EBLUP Estimator on Small Area Estimation with Application to BPS Data Accessed on April 2018 through http://web.ipb.ac.id/~anangk/home/uploads/file/(AK+KAN)ICoMS-1.pdf
[3] Rao JNK and Molina 2015 Small Area Estimation 2nd Edition New York: John Wiley and Sons Inc.
[4] Ubaidillah A 2017 Simultaneous Equation Models for Small Area Estimation, Bogor, IPB.
[5] Benavent R and Morales D 2016 Computational Statistics and Data Analysis 94 372.
[6] Badan Pusat Statistik 2017 The household expenditure per capita in Indonesia in 2017 Accessed on Nov 2017 through www.bps.ac.id.
[7] Ghosh M and Rao JNK 1994 Stat. Science 9(1) 55.
[8] Putri, MFL 2016 Penerapan Metode Empirical Best Linear Unbiased Prediction (EBLUP) pada Small Area Estimation Jakarta: STIS.