International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET231063

A Comparison Small Area Estimation for Skewed Data with EBLUP and Hierarchical Bayes Approaching using Rao-Yu Model

Titin Yuniarty[1], Indahwati[2], Aji Hamim Wigena[3]

[1,2,3]Department of Statistics, IPB University, Bogor, Central Java, Indonesia

yuniarty_titin@apps.ipb.ac.id[1]

ARTICLEINFO

Article History :

Accepted: 02 Nov 2023
Published: 20 Nov 2023

Publication Issue :

Volume 10, Issue 6
November-December-2023

Page Number :

132-143

ABSTRACT

Small Area Estimation (SAE) is a method based on modeling for estimating small area parameters, that applies the Linear Mixed Model (LMM) as its basic. It is conventionally solved with Empirical Best Linear Unbiased Prediction (EBLUP). The main requirement for LMM to produce high precision estimates is normally distributed for its sampling error. However, the researching data namely per capita food expenditure for food crop farmers' households in Southeast Sulawesi Province has been positively skewed. Applying EBLUP for positively skewed data will produce less accurate estimates. Meanwhile, doing a transformation process will potentially produce biased estimates. The Hierarchical Bayes (HB) approaching often known as Full Bayesian is more flexible regarding normality assumptions and can determine distribution based on data. Because this research will carry out estimates at the district/city level throughout Southeast Sulawesi Province, totaling 17 regions, the efficiency of the estimates can be increased by using the SAE Rao-Yu model which includes area and time random effects. For this reason, this study compares SAE modeling with the EBLUP and HB approaches which are assumed to follow normal, log-normal, and skew-normal distributions. By comparing the Relative Root Mean Square Error (RRMSE), Deviance Information Criterion (DIC), and Coefficient of Variation (CV) values, it is concluded that estimates from the log-normal and skew-normal SAE HB models are more efficient than the SAE EBLUP and normal HB models. However, the SAE estimate that is closer to the direct estimation results is the skew-normal SAE HB.

Keywords: SAE, EBLUP, Hierarchical Bayes, Skewed Data, Rao-Yu Model

I. INTRODUCTION

Nowadays, the demand for parameter estimation at the area or small domain level is increasing as the need for microdata becomes more diverse. A small area is subset of the population where a variable is of interest. Two methods can be used to produce estimates in small areas or domains. The first is an estimate that is only based on a certain sampling design or what is known as a direct estimate. This method produces unbiased estimates, but has low precision due to large variance, as a result of inadequate sample size. Second, indirect estimation, or what is usually called Small Area Estimation (SAE), borrows the power of information from other variables in adjacent domains to increase the effectiveness of the sample size.

SAE is a model-based estimation method that applies a linear mixed model (LMM) as the basic model, which can be solved using Empirical Best Linear Unbiased Prediction (EBLUP). This approach assumes normal distributed random influences to be able to obtain more precise estimates [1]. However, in reality, there is a lot of data with a positive trend pattern. This contradicts the normality assumption required in conventional SAE modelling. Another approach to solving the SAE model, which is relatively flexible with normality assumptions, is Hierarchical Bayes known as full Bayes, which can adaptively determine the distribution of data based on survey results [2]. In Hierarchical Bayes, all unknown parameters are considered as variables and have a distribution [3]. Estimation of small area parameters is carried out on the posterior distribution, which is the result of multiplying the prior distribution and likelihood function of the observed data.

This research uses monthly food per capita expenditure data for food crop farmers' households in Southeast Sulawesi Province. This data was obtained from the of National Socio-Economic Survey (Susenas) on March period, which produces estimates at the district/city level throughout Indonesia. Because Susenas sampling unit is general households, but food crop farmers households per district/city are included a sub-domain of households in Susenas, SAE is used to estimate the mean of food expenditure per capita. The results of initial exploration of data on per capita expenditure on food commodities for food crop farmers' households in Southeast Sulawesi Province showed that there was a positive trend pattern with a spread value of more than 0.

Several SAE studies on continuous-scale panhandle data such as per capita expenditure, using a full Bayes approach, including those carried out by Ferraz and Moura (2012), Moura et al (2017), and Fabrizi et al (2017) [4]–[6]. This study aims to produce estimates of small area parameters in panhandle data at the scale of the original data because if it is carried out using a transformation approach it has a potential source of bias when back-transformation is carried out. The data used in these studies is cross-section type data, so it uses LMM which was first introduced by Fay and Herriot (1979). In 1994, Rao-Yu introduced the development of the Fay-Herriot model for combined cross-section and time series data [7]. SAE modeling on a combination of cross-section and time series data can increase the efficiency of estimation results [8]–[11]. Neves et al (2020) carried out SAE modeling on a combination of cross-section and time series data with a positive trend pattern and obtained the results that SAE using the Rao-Yu model with random walk effects was more efficient than the Fay-Herriot model [12].

Based on several previous studies, this research will use the area-level Rao-Yu model to obtain a more effective and efficient estimate of the mean per capita food expenditure in food crop farming households at the district/city level of Southeast Sulawesi Province. Because this research focuses on generating SAE estimates on the scale of the original data, modeling will be compared between the EBLUP and Hierarchical Bayes approaches for data that follows the

assumptions of normal, log-normal, and skew-normal distribution.

II. METHODS AND MATERIAL

This research uses secondary data from BPS-Statistics Indonesia for the 2018-2021 period. In SAE modeling, data is divided into two types of variables, namely response variables and accompanying variables. The response variable (*Y*) is variable that is the SAE target, which in this study is food expenditure per capita in food crop farming households at the district/city level in Southeast Sulawesi Province. The *Y* variable data comes from Susenas for the March period.

The auaxiliaries variable (*X*) is a variable that functions as an explanation, just like simple regression analysis. The criteria for a good auxiliary variable are those that can explain variation between small areas and are not susceptible to small sample sizes obtained from census data or administrative compilations. Therefore, data on auxiliary variables in this research was obtained from the Village Potential Data Collection (Podes) and the Southeast Sulawesi Publication in Figures (DDA). There are 12 accompanying variables used in this research, namely: the amount of rainfall (X_1), the proportion of villages/sub-districts with food crop centers that have irrigation canals or dams or reservoirs or embungs for irrigating agricultural land (X_2), the proportion of villages/sub-districts with crop centers food crops that have good condition farming roads (X_3), population dependency ratio (X_4), proportion of natural disasters occurring in food crop center villages/sub-districts (X_5), proportion of villages/sub-districts Unit Cooperatives (KUD) with active status in villages/sub-districts food crop center (X_6), the proportion of farmer groups in the villages/sub-districts food crops center (X_7), the ratio of poor letters (SKTM) that is issued by government from villages/sub-districts of food crops center per 100 population (X_8), proportion of residents suffering from malnutrition in the village/centra village food crops (X_9), proportion of slum settlements families in villages/sub-districts food crop center (X_{10}), the ratio of active Junior High School (SMP/MTS) that is located in villages/sub-districts of food crop center in 100 population (X_{11}), and the ratio of health workers living in villages/sub-districts with food crop centers per 100 population (X_{12}).

EBLUP Rao-Yu Model

SAE is a method of estimating parameters in a small area based on modelling by borrowing the power of information from accompanying variables in adjacent domains to increase the precision of the estimated results and the effectiveness of the sample size. There are two basic SAE models based on availability of auxiliary variables, namely the area-level model and the unit-level model. Because the auxiliary variable data in this study is available at the area level, namely district/city which is an aggregation of the village/sub-district level of food crop centers, the SAE model used is the area level model.

The area-level model was first developed by Fay-Herriot in 1979 as SAE-based model. It is only including area random effects to minimize the MSE value. However, in surveys that are carried out periodically, such as Susenas, the efficiency of SAE estimation can be increased by including random effects of area and time. This model was popularly introduced by Rao and Yu (1994), as a development of Fay-Herriot model on combined cross-section and time series data, which consists of sampling error model:

$$\hat{Y}_{it} = \theta_{it} + e_{it};\ t = 1, \dots, T; i = 1, \dots, m \qquad (1)$$

and connecting model:

$$\theta_{it} = \boldsymbol{x}_{it}^{T}\boldsymbol{\beta} + v_i + u_{it} \qquad (2)$$

If model (1) dan (2) combined, will become:

$$\hat{Y}_{it} = \boldsymbol{x}_{it}^{T}\boldsymbol{\beta} + v_i + u_{it} + e_{it} \qquad (3)$$

where $\boldsymbol{\beta}$ is coefficients variable parameter of $\boldsymbol{x}_{it}^{T}$ that is assumed to be constantly, $\boldsymbol{x}_{it}^{T}$ is the vector of auxiliary

variable for area i-th and time t-th, $\hat{Y}_{it}$ is direct estimate for area i-th time t-th, θ_{it} is mean function of area i-th time t-th, and e_{it} is sampling error that is normally distributed with expected value 0 and matrices variance covariance $\boldsymbol{\psi_{it}} = blokdiag(\boldsymbol{\psi_{1t}}, \dots, \boldsymbol{\psi_{mt}})$ that might be changed at the time $v_i \underset{\sim}{iid} N(0, \sigma_v^2)$ and u_{it} followed this process:

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1 \tag{4}$$

dimana $\varepsilon_{it} \underset{\sim}{iid} N(0, {\sigma_\varepsilon}^2)$, dengan $\{e_{it}\}$, $\{v_i\}$, dan $\{\varepsilon_{it}\}$ diasumsikan saling bebas. Dari model tersebut, diketahui θ_{it} tergantung pada efek area spesifik v_i dan area oleh waktu spesifik u_{it} yang berkorelasi antar waktu.

Rao-Yu analogizes model (3) to a Generalized Linear Mixed Model (GLMM) because the combination of fixed effects and random effects. There are several approaches that can be used to estimate GLMM parameters, including Best Linear Unbiased Prediction (BLUP). However, BLUP assumes that the variance of the random effects (σ_v^2dan ${\sigma_\varepsilon}^2$) is known in the GLMM, which is hereinafter called the variance component. However, in fact, the variance components are unknown so they are estimated based on data. For this reason, Harville (1977) studied methods for estimating variance components, namely by incorporating the Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods [13]. The BLUP obtained when the variance components are substituted from the estimator is called Empirical Best Linear Unbiased Prediction (EBLUP). EBLUP works under assumption that random effects have a normal distribution.

Hierarchical Bayes Rao-Yu Model

In SAE modelling with Hierarchical Bayes, all unknown parameters are considered as random variables and each has certain prior distribution, thus creating a hierarchical arrangement of unknown model parameters. Parameter estimation is carried out on the posterior distribution, which is obtained by multiplying the prior distribution and the likelihood function of the observed data. The general shape of the posterior distribution is as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{a}) f(\boldsymbol{a}|\boldsymbol{b}) \tag{5}$$

where $\boldsymbol{\theta}$ is parameter which will be estimated in the model so the prior distribution will first be determined, namely $f(\boldsymbol{\theta}|\boldsymbol{a})$. Furthermore, the distribution parameter has a prior distribution $f(\boldsymbol{a}|\boldsymbol{b})$and $f(\boldsymbol{\theta}|\boldsymbol{y})$ is posterior distribution. To obtain posterior distribution in equation (5), a high-level integration process is required, the results of which are often not close-form. For this reason, it is solved using Markov Chain Monte Carlo (MCMC) approach with the Gibbs Sampling algorithm [14].

The SAE model with random effects of area and time for food expenditure data per capita of food crop farming households in Southeast Sulawesi Province is expressed in the following three Hierarchical Bayes (HB) models:

1) Model HB normal

Level 1 : $\widehat{\boldsymbol{Y}}_{\boldsymbol{it}}|\boldsymbol{\theta_{it}} \sim N(\boldsymbol{\theta_{it}}, \boldsymbol{\Psi}_{it})$; $\boldsymbol{\Psi}_{it}$ is known

Level 2 : $\boldsymbol{\theta_{it}}|\boldsymbol{\beta}, u_{it}, \tau_v \sim N\left(\boldsymbol{x_{it}^T\beta} + u_{it}, \tau_v\right)$

Level 3 : $u_{it}|u_{i,t-1}, \tau_\varepsilon \sim N\left(\rho u_{i,t-1}, \tau_\varepsilon\right)$

Level 4 : $f(\boldsymbol{\beta}, \tau_v, \tau_\varepsilon) = f(\boldsymbol{\beta}) f(\tau_v)\, f(\tau_\varepsilon)$

$$f(\boldsymbol{\beta}) \sim N\left(\hat{\beta}_k, \frac{1}{\left(se(\hat{\beta}_k)\right)^2}\right), \tau_v \sim G(a_v, b_v),$$

$$\tau_\varepsilon \sim G(a_\varepsilon, b_\varepsilon) \tag{6}$$

2) Model HB log-normal

Level 1 : $\widehat{\boldsymbol{Y}}_{\boldsymbol{it}}|\boldsymbol{\theta_{it}} \sim LN(\boldsymbol{\theta_{it}}, \boldsymbol{\Psi}_{it})$; $\boldsymbol{\Psi}_{it}$ is known

Level 2 : $\boldsymbol{\theta_{it}}|\boldsymbol{\beta}, u_{it}, \tau_v \sim N\left(\boldsymbol{x_{it}^T\beta} + u_{it}, \tau_v\right)$

Level 3 : $u_{it}|u_{i,t-1}, \tau_\varepsilon \sim N\left(\rho u_{i,t-1}, \tau_\varepsilon\right)$

Level 4 : $f(\boldsymbol{\beta}, \tau_v, \tau_\varepsilon) = f(\boldsymbol{\beta}) f(\tau_v)\, f(\tau_\varepsilon)$

$$f(\boldsymbol{\beta}) \sim N\left(\hat{\beta}_k, \frac{1}{\left(se(\hat{\beta}_k)\right)^2}\right), \tau_v \sim G(a_v, b_v),$$

$$\tau_\varepsilon \sim G(a_\varepsilon, b_\varepsilon) \quad (7)$$

3) Model HB skew-normal

Level 1: $\hat{\boldsymbol{Y}}_{\boldsymbol{it}}|\boldsymbol{\theta}_{\boldsymbol{it}}, \lambda, n_{it}, \phi_i \sim SN\left(\boldsymbol{\theta}_{\boldsymbol{it}}, \sqrt{\psi_{it}}, \lambda'/\sqrt{n_{it}}\right)$

$\psi_{it}|n_{it}, \phi_i \sim G\left(\frac{1}{2}(n_{it}-1), \frac{1}{2}(n_{it}-1)\phi_i^{-1}\right)$

$\phi_i^{-1}|a_\phi, b_\phi \sim G(a_\phi, b_\phi) \rightarrow$ parameter *scale*

$\lambda \sim N(0,100) \rightarrow$ parameter skewness

Level 2: $\boldsymbol{\theta}_{\boldsymbol{it}}|\boldsymbol{\beta}, u_{it}, \tau_v \sim N\left(\boldsymbol{x}_{\boldsymbol{it}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_{it}, \tau_v\right)$

Level 3: $u_{it}|u_{i,t-1}, \tau_\varepsilon \sim N\left(\rho u_{i,t-1}, \tau_\varepsilon\right)$

Level 4: $f(\boldsymbol{\beta}, \tau_v, \tau_\varepsilon) = f(\boldsymbol{\beta}) f(\tau_v)\, f(\tau_\varepsilon)$

$$f(\boldsymbol{\beta}) \sim N\left(\hat{\beta}_k, \frac{1}{\left(se(\hat{\beta}_k)\right)^2}\right), \tau_v \sim G(a_v, b_v)$$

$$\tau_\varepsilon \sim G(a_\varepsilon, b_\varepsilon) \quad (8)$$

(7)

$a_\phi \sim G(0.01,0.01)$ and $b_\phi \sim G(0.01,0.01)$, $\tau_v = 1/\sigma_v^2$ and $\tau_\varepsilon = 1/{\sigma_\varepsilon}^2$, while parameter prior value for τ_v and τ_ε are $a_v = b_v = c_v = d_v = 0{,}01$, n_{it} is sampling size for area *i*-th and time *t*-th from N_{it} population.

Framework of The Research

This research will compare the goodness of estimation of the SAE model on data that does not meet the normal assumption because the pattern extends to the right. The approaches used are EBLUP and Hierarchical Bayes (HB). For HB approach, SAE modeling is carried out on three data distribution assumptions, namely normal, log-normal, and skew-normal. For next, we will review the differences in estimation results obtained by comparing the Deviance Information Criterion (DIC) and Relative Root Mean Square Error (RRMSE) values as a measure to determine the goodness of the model.

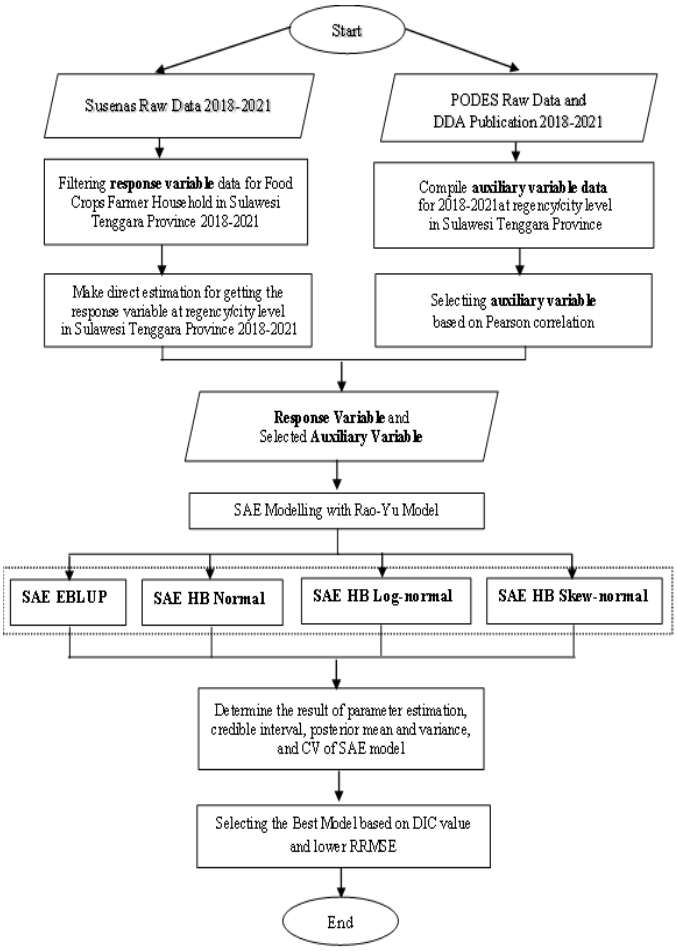

Figure 1: Flowchart of the research

III. RESULTS AND DISCUSSION

Overview of Direct Estimation

After calculating the direct estimator of the per capita food expenditure of food crop farmer households in Southeast Sulawesi Province, the following bar diagram is presented to provide an overview of the per capita food expenditure of food crop farmer households in all districts/cities in Southeast Sulawesi Province.

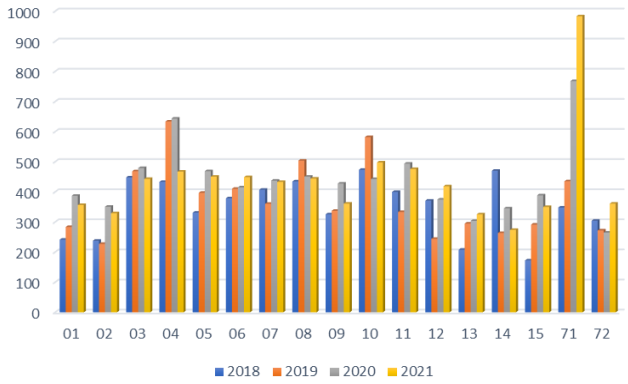

Figure 2: Direct estimation of Per Capita Food Expenditure from Food Crops Farmers' Household at Regency/City Level in Sulawesi Tenggara Province, 2018-2021

Figure 2 above shows that Kendari City (71), Kolaka Regency (04) and North Konawe (10) are three regions with relatively higher per capita food expenditure than other regions in Southeast Sulawesi Province during 2018-2021.

Table 1. Correlation between Auxiliary Variable and Direct Estimation of Per Capita Food Expenditure

Variable	Pearson Correlation	Sig. (2-tailed)	N
Y_1	1		68
X_1	**.262***	**0.031**	68
X_2	**.343****	**0.004**	68
X_3	0.166	0.177	68
X_4	**-.525****	**0.000**	68
X_5	**-.319****	**0.008**	68
X_6	0.024	0.847	68
X_7	0.042	0.737	68
X_8	**-.353****	**0.003**	68
X_9	**-.290***	**0.016**	68
X_{10}	**-.325****	**0.007**	68
X_{11}	-0.133	0.281	68
X_{12}	0.071	0.566	68

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Based on **Table 1** above, it is shown from 12 auxiliary variables, there are seven variables that are significantly correlated with the direct predictor of per capita food expenditure of food crop farmer households in Southeast Sulawesi Province, as seen from the Pearson correlation value and its significance. These variables are X_1, X_2, X_4, X_5, X_8, X_9, and X_{10}.

Table 2. Summary of VIF and Significantly Correlated between Auxiliaries Variable

Variable	VIF	Significantly Correlated
X_1	1.4795854	X_4, X_7
X_2	2.5701967	X_3, X_4, X_7, X_8, X_9, X_{10}
X_3	1.7878755	X_2, X_4, X_{10}, X_{12}
X_4	2.4111556	X_1, X_2, X_3, X_5, X_8, X_9, X_{10}
X_5	1.7203957	X_4, X_8, X_9
X_6	1.7161165	X_7, X_{12}
X_7	2.1407962	X_1, X_2, X_6, X_8, X_{10}, X_{12}
X_8	2.7609296	X_2, X_4, X_5, X_7, X_9, X_{10}, X_{11}
X_9	2.2309659	X_2, X_4, X_5, X_8, X_{10}
X_{10}	1.6506743	X_2, X_3, X_4, X_7, X_8, X_9
X_{11}	1.5932032	X_8
X_{12}	1.5167777	X_3, X_6, X_7

In this research, all auxiliary variables will be used because we look at the correlation values between accompanying variables and test the multicollinearity assumption through the Variance Inflation Factor (VIF). Observing correlation between auxiliary variables and testing multicollinearity assumptions is an important process before entering SAE modelling, to find out whether the inclusion of a variable can be represented by the presence of other variables in the model [14]. In **Table 2**, all accompanying variables have a VIF value of less than 5. VIF less than 5 indicates that there is no multicollinearity between the accompanying variables in the simple linear regression model [15]. Meanwhile, if we look correlation between auxiliary variables, all accompanying variables are

correlated with each other. Therefore, in this study, 12 candidate auxiliary variables were used.

Indirect Estimation with SAE Rao-Yu Modeling

After obtaining the auxiliary variables, the SAE indirect estimation process will be carried out in the Rao-Yu model with the EBLUP and Hierarchical Bayes (HB) approaches. Below is a table of estimated coefficients using the SAE EBLUP method.

Table 3. Summary of Hyperparameter - Model SAE EBLUP

Parameter	Mean	Std.error	t-value	p-value
β_0	499.289	152.170	3.281	**0.001**
β_1	34.523	16.428	2.101	**0.036**
β_2	122.757	49.111	2.500	**0.012**
β_3	-126.662	91.382	-1.386	0.166
β_4	-272.930	197.939	-1.379	0.168
β_5	0.573	42.064	0.014	0.989
β_6	-460.586	847.817	-0.543	0.587
β_7	-2.249	14.830	-0.152	0.880
β_8	33.596	59.436	0.565	0.572
β_9	-6.781	40.418	-0.168	0.867
β_{10}	-31.163	32.272	-0.966	0.334
β_{11}	-297.089	380.880	-0.780	0.435
β_{12}	88.852	62.938	1.412	0.158
τ_ε	4599.312	1308.246	-	-
τ_v	340.292	1017.239	-	-
ρ	0.000	0.280	-	-

Based on **Table 3** above, it can be concluded that with a 95% confidence level, there are only two variables that significantly influence the per capita food expenditure of food crop farming households at the district/city level in Southeast Sulawesi Province, namely X_1 and X_2. If we look at the estimation results per district/city per year, it can be seen that Kolaka District (04), North Kolaka (08), North Konawe (10), and East Kolaka (11) are the four districts with the highest per capita food expenditure for farming households. food crops in Southeast Sulawesi Province. The results of the Rao-Yu EBLUP estimation are slightly different from the results of direct estimation. This can be seen in Figure 3 below.

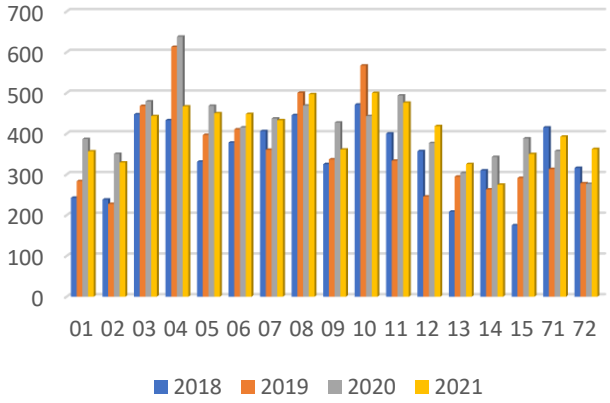

Figure 3: EBLUP Rao-Yu estimation of Food Per Capita Expenditure from Food Crops Farmers' Household at Regency/City Level in Sulawesi Tenggara Province, 2018-2021

Estimating small area parameters using EBLUP is less precise because the data from direct estimation is not distributed normally, so that it is not in line with conventional SAE assumptions. Below are presented the results of the data normality test resulting from direct estimation in **Table 4**.

Table 4. Shapiro Wilk test statistics for normality of Per Capita Food Expenditure Direct Estimation, 2018-2019

Direct Estimation	Statistik Uji (W)	p-value
Y_{2018}	0.89289	2.20E-16
Y_{2019}	0.88388	2.20E-16
Y_{2020}	0.84498	2.20E-16
Y_{2021}	0.83933	2.20E-16

With 95% confidence level, it is concluded that direct estimation data not normally distributed. Therefore, in

this research modelling using Hierarchical Bayes (HB) approach was carried out. SAE HB modelling process begins with determining prior distribution for unknown parameters in the model such as equations (6), (7), and (8). Completion of the hierarchical model is carried out through Markov Chain Monte Carlo (MCMC) process to obtain the posterior distribution.

Table 5. Summary of Hyperparameter Posterior Distribution from Model SAE HB Normal

Parameter	Mean	Std.error	Percentile	
			2.5%	97.5%
β_0	623.700	86.160	**477.800**	**820.400**
β_1	-7.138	11.140	-28.750	14.860
β_2	46.530	34.250	-22.510	112.700
β_3	-26.930	62.490	-151.400	100.400
β_4	-484.100	118.400	**-714.100**	**-245.700**
β_5	-21.840	25.100	-72.460	27.210
β_6	11.740	517.300	-1001.000	1006.000
β_7	10.040	10.380	-10.230	30.190
β_8	-42.760	46.070	-136.000	45.620
β_9	10.170	23.790	-37.030	57.110
β_{10}	-9.948	19.110	-47.950	26.650
β_{11}	-219.400	242.500	-710.000	246.400
β_{12}	119.500	44.510	**34.280**	**207.800**
τ_ε	2103.000	275.900	1610.000	2695.000
τ_v	2.074	18.820	0.000	14.030
ρ	0.585	0.107	0.368	0.790

In forming the model, whether SAE HB normal, SAE HB log-normal, or SAE HB skew-normal, the number of total iterations, iterations for the burn-in period, and thin must be determined first by trial and error, starting from a small value until convergence is achieved. algorithm for all parameters in the model. After going through the simulation process, the best model combination was obtained when the number of Markov chains was 3 with the number of iterations for each chain being 200,000, where 50,000 iterations were for the burn-in period and the remaining 150,000 iterations were for posterior analysis. The chain was thinned by taking every 10-th sample value to reduce autocorrelation between samples generated.

Table 5 shows the results of estimating the posterior distribution of the normal SAE HB model. Based on the 95% credible posterior distribution interval for the model hyperparameters in the table above, there are only two accompanying variables that have a significant effect because they do not contain the value 0 in the interval, namely X_4 and X_{12}. If you look at the estimation results per district/city per year based on **Figure 4**, Kolaka District (04), North Konawe (10), North Kolaka (08), and East Kolaka (11) are the four districts that have the highest level of food expenditure per capita for food crop farming households in Southeast Sulawesi Province. The HB Rao-Yu estimation results for the normal distribution are no different from the estimation results from the EBLUP Rao-Yu. Another thing that is in line with the Rao-Yu EBLUP is that the coefficient of the random effect of time ($1/\sigma_\varepsilon^2$) less than domain random effect ($1/\sigma_v^2$).

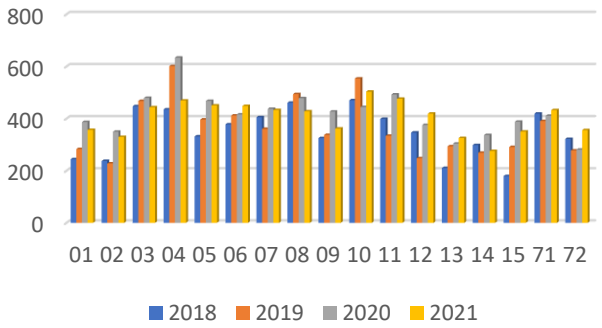

Figure 4: SAE HB Normal Rao-Yu estimation of Food Per Capita Expenditure from Food Crops Farmers' Household at Regency/City Level in Sulawesi Tenggara Province, 2018-2021

Estimation of small area parameters with Rao-Yu model via the HB log-normal approach is presented in **Table 6**.

Table 6. Summary of Hyperparameter Posterior Distribution from Model SAE HB Log-normal

Parameter	Mean	Std.error	Percentile

			2.5%	97.5%
β_0	6.724	0.369	**6.017**	**7.395**
β_1	-0.049	0.049	-0.143	0.052
β_2	0.045	0.149	-0.255	0.340
β_3	0.028	0.254	-0.507	0.543
β_4	-1.680	0.394	**-2.509**	**-0.972**
β_5	-0.060	0.109	-0.267	0.153
β_6	2.029	1.182	-0.260	4.362
β_7	0.061	0.048	-0.028	0.163
β_8	-0.085	0.185	-0.448	0.266
β_9	-0.002	0.107	-0.207	0.215
β_{10}	-0.025	0.080	-0.182	0.135
β_{11}	-0.964	0.690	-2.309	0.417
β_{12}	0.322	0.182	-0.014	0.702
τ_ε	0.026	0.021	0.002	0.080
τ_v	0.031	0.008	0.018	0.049
ρ	0.367	0.220	-0.082	0.776

Based on the 95% credible interval of the posterior distribution of the SAE HB log-normal model, only one variable have a significant effect on the direct estimate of food expenditure per capita of food crop farming households in Southeast Sulawesi province, namely X_4. The direction of the relationship between the significant auxiliary variables and the response variable is in line with the direction of the correlation coefficient. The mean, std. error and percentile values in **Table 6** are expressed in natural logarithms (ln). It can be seen that the parameter coefficients of log-normal SAE HB model after back-transformation (exp) tend to be smaller than parameter coefficients of the SAE EBLUP and SAE HB normal model. As with SAE HB normal model, the time random effect ($\exp(\tau_\varepsilon) = 1/exp(\sigma_\varepsilon^2)$) is less than ($\exp(\tau_v) = 1/exp(\sigma_v^2)$) as domain random effect.

Parameter estimates by district/city per year, based on SAE HB log-normal model are presented in **Figure 5**. In contrast to the estimated results from the SAE EBLUP and SAE HB Normal models, small areas are estimated to have a higher level of food expenditure per capita for food crop farming households in in Southeast Sulawesi provinces namely Kendari City and Regency of Kolaka, North Konawe, and North Kolaka.

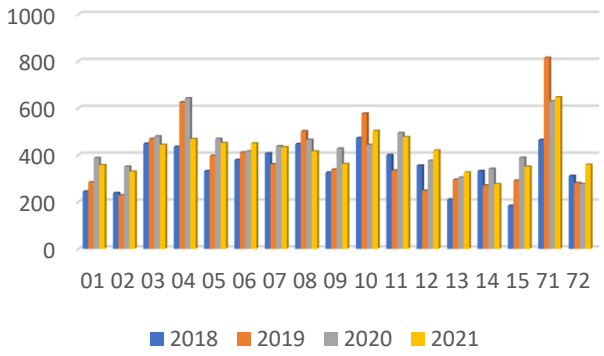

Figure 5: SAE HB Log-normal Rao-Yu estimation of Food Per Capita Expenditure from Food Crops Farmers' Household at Regency/City Level in Sulawesi Tenggara Province, 2018-2021

Table 7 shows the results of parameter estimation from the skew-normal SAE HB model. Based on the 95% credible posterior distribution interval for the model of hyperparameters, there are two significantly variables that influence the amount of food expenditure per capita of food crop farming households in Southeast Sulawesi Province, namely X_4 and X_{12}. This is in line with the expected results of the normal SAE HB model. The direction of relationship between auxiliary variables that have a significant effect on the response variable is in line with the direction of the correlation coefficient. Likewise with coefficient of time random effect ($1/\sigma_\varepsilon^2$) less than domain random effect ($1/\sigma_v^2$), which is in line with the estimated results of the SAE EBLUP, SAE HB normal, and SAE HB log-normal models.

Table 7. Summary of Hyperparameter Posterior Distribution from Model SAE HB Skew-normal

Parameter	Mean	Std.error	Percentile	
			2.5%	97.5%
β_0	637.000	88.430	**463.400**	**819.500**
β_1	-7.399	10.790	-28.430	13.470
β_2	40.650	33.530	-26.250	105.300
β_3	-31.270	66.750	-166.300	102.400

β_4	-519.000	119.400	**-740.500**	**-293.700**
β_5	-16.970	26.360	-68.670	34.240
β_6	61.720	508.700	-929.000	1036.000
β_7	11.630	10.340	-7.530	32.270
β_8	-32.230	41.650	-115.800	50.510
β_9	8.907	23.850	-36.780	56.570
β_{10}	-8.503	18.910	-46.390	28.100
β_{11}	-208.600	232.800	-658.000	254.600
β_{12}	118.300	43.440	**28.300**	**201.700**
τ_ε	1.006	1.005	0.000	3.739
τ_v	2125.000	277.400	1628.000	2718.000
ρ	0.587	0.101	0.379	0.778
λ	1.342	0.9698	-0.4984	3.324

Meanwhile, if we look at the estimated results according to the small area in **Figure 6**, there are similarities with the estimated results of the SAE HB log-normal model, there are four regions that have the highest level of food expenditure per capita for food crop farming households in Southeast Sulawesi Province namely Kolaka Regency, Kendari City, North Konawe, and North Kolaka.

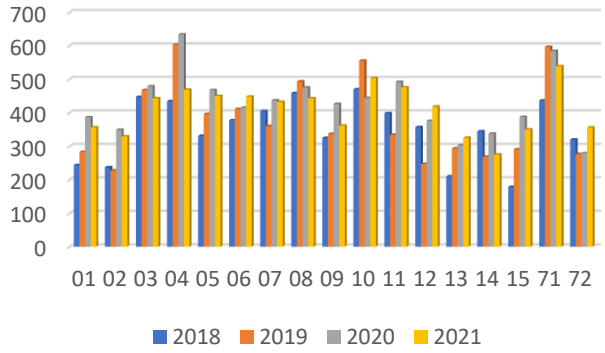

Figure 6: SAE HB Skew-normal Rao-Yu estimation of Food Per Capita Expenditure from Food Crops Farmers' Household at Regency/City Level in Sulawesi Tenggara Province, 2018-2021

Model Comparison

To compare the estimator of direct and SAE EBLUP, along with the SAE HB normal, log-normal, and skew-normal, the following boxplot is presented in **Figure 7**.

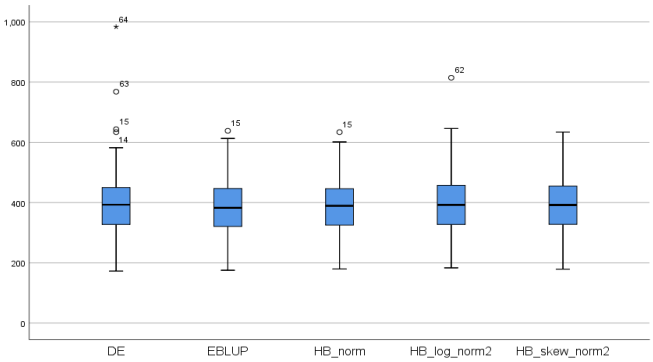

Figure 7: Boxplot of Per Capita Food Expenditure Estimation between Direct Estimate (DE), SAE EBLUP, SAE HB normal, SAE HB log-normal, and SAE HB skew-normal

Based on **Figure 7**, it can be seen that direct estimator is more similar to the skew-normal HB estimator. One thing that is different is that the direct estimator has outliers, as do the SAE EBLUP, SAE HB normal and SAE HB log-normal estimators. To confirm statistical similarity, below is presented the Pearson correlation coefficient and its significance.

Table 8. Correlation between Direct Estimation and SAE

Model	Corr.	p-value	Description
EBLUP	0.702	0.000	Signifikan
HB normal	0.761	0.000	Signifikan
HB log-normal	0.845	0.000	Signifikan
HB skew-normal	**0.864**	**0.000**	**Signifikan**

Based on Table 8, it is known that the similarity between the direct estimator and the skew-normal SAE HB estimator is the highest, namely 86.4%. The correlation value is in accordance with the similarity shown in **Figure 7**.

To make it easier to compare between these models, below are presented the mean and maximum values of RSE/RRMSE for each observation resulting from direct estimation, SAE EBLUP, normal HB, log-normal HB, and skew-normal HB.d

Table 9. RSE/RRMSE (%) of Direct Estimation and SAE

Model	Mean	Maximum	RRMSE
Direct Estimation	8.371	99.711	-
EBLUP	4.694	35.608	26.521
HB normal	**3.793**	**19.335**	24.150
HB log-normal	4.400	25.457	19.173
HB skew-normal	3.831	21.891	**18.324**

Based on **Table 9**, it can be seen that the RRMSE of the SAE estimator for HB has a maximum value of less than 25%, but for EBLUP less over 25%. This means that the three SAE HB methods have succeeded in improving the precision of direct estimates for all districts/cities. While EBLUP is caused of the normality assumption is not required. When viewed from the mean, the SAE HB normal has the smallest value as well ase its maximum. However, if the RRMSE is calculated for model, the skew-normal HB estimator has the smallest value than to the others. This is also confirmed when comparing the distribution of RRMSE for each small area between the five estimators for 2021, which is presented in **Figure 8**.

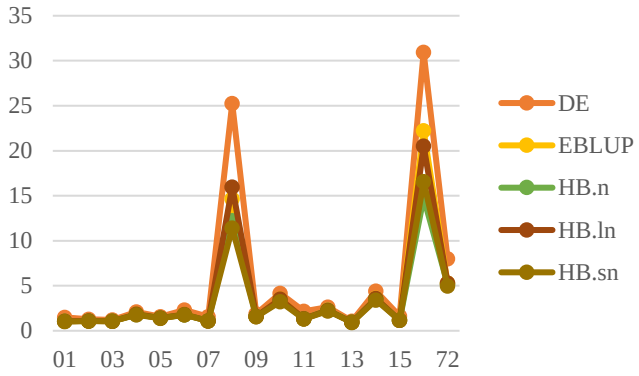

Figure 8: RRMSE Plot between Direct Estimate (DE), SAE EBLUP, SAE HB normal, SAE HB log-normal, and SAE HB skew-normal of Regency/City in Sulawesi Tenggara Province, 2021

Based on **Figure 8**, it is shown that the RRMSE of the skew-normal HB estimator is the lowest compared to other estimators for each district/cities level in Southeast Sulawesi Province. This indicates that the skew-normal HB estimator is more efficient than other estimators for positive trend data conditions. Phenomenon is further strengthened by looking at the boxplot of the Coefficient of Variation of each estimator in **Figure 9**.

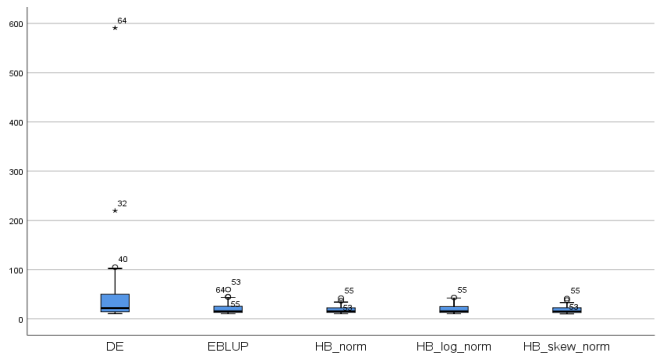

Figure 9: Coeficient of variation (CV) between DE, SAE EBLUP, SAE HB normal, SAE HB log-normal, and SAE HB skew-normal

Figure 9 shows that SAE HB skew-normal is the models with the best precision values, compared to other models. The magnitude of the reduction in CV to the direct estimator was largest for the skew-normal HB, namely 58.49% and normal HB 57.90%. Meanwhile, in the HB log-normal model, there was a reduction in CV of 54.75% and 53.47% in the EBLUP model. Even though the largest decrease in CV comes from the skew-normal SAE HB model, if we seen from the Deviance Information Criterion (DIC), as measure of the goodness of the model in the Bayes approach, the log-normal SAE HB model is the model with the smallest DIC compared to the other models, as presented in **Table 10** below.

Table 9. Model Selection by DIC

Model	$\bar{D}$	pD	DIC
EBLUP	640.51	32.00	672.51
HB normal	527.00	57.63	584.63
HB log-normal	**522.30**	**53.55**	**575.90**
HB skew-normal	521.10	58.00	579.10

The model with the smallest DIC is the best model, which in this study is the skew-normal SAE HB model.

IV. CONCLUSION

Small area modeling using the Empirical Best Linear Unbiased Prediction (EBLUP) and Hierarchical Bayes (HB) approaches is able to increase the efficiency of estimation, compared it to direct estimation. However, among the four SAE models, HB approach on positive panhandle data for data that is assumed to follow a skew-normal and log-normal distribution is proven to be more efficient than EBLUP and HB normal approaches. This can be seen from the RRMSE, DIC and CV values which are relatively smaller for skew-normal and log-normal SAE HB models than for SAE EBLUP and SAE HB normal. However, if look at the data distribution pattern, what is more like a direct estimator is the skew-normal SAE HB model. This is shown by the box plot and the Pearson correlation coefficient value of the direct estimator which is the largest compared to the other three SAE models.

V. REFERENCES

[1]. S. K. Sinha and J. N. K. Rao, "Robust small area estimation," Can. J. Stat. / La Rev. Can. Stat., vol. 37, no. 3, pp. 381–399, Jun. 2009.

[2]. B. Liu, Hierarchial Bayes Estimation and Empirical Best Prediction of Small Area Proportions. 2009.

[3]. W. M. Boldstad, Bayesian Statistics, vol. 2. 2007. doi: 10.2307/2682504.

[4]. V. R. S. Ferraz and F. A. S. Moura, "Small area estimation using skew normal models," Comput. Stat. Data Anal., vol. 56, no. 10, pp. 2864–2874, 2012.

[5]. E. Fabrizi, C. Trivisano, and M. R. Ferrante, "Bayesian small area estimation for skewed business survey variables Bayesian small area estimation for skewed business survey variables," R. Stat. Soc., pp. 1–18, 2017, doi: 10.1111/rssc.12254.

[6]. F. A. S. Moura, A. F. Neves, and D. B. do N. Silva, "Small area models for skewed Brazilian business survey data," J. R. Stat. Soc. Ser. A Stat. Soc., vol. 180, no. 4, pp. 1039–1055, 2017, doi: 10.1111/rssa.12301.

[7]. J. N. . Rao and I. Molina, Small Area Estimation. New Jersey: John Wiley & Sons, 2015.

[8]. K. Sadik and K. A. Notodiputro, "Hierarchial Bayes Estimation Using Time Series and Cross-setional Data: A Case of Per-capita Expenditure in Indonesia," in Conference on Small Area Estimation in Spain, 2009, pp. 1–4.

[9]. F. E. Supriatin, B. Susetyo, and K. Sadik, "EBLUP Method of Time Series and Cross-Section Data for Estimating Education Index in District Purwakarta," Indones. J. Stat., vol. 20, no. 7, pp. 34–38, 2015.

[10]. Y. Susianto, K. A. Notodiputro, A. Kurnia, and H. Wijayanto, "Modifikasi Model Rao-Yu untuk Pendugaan Area Kecil Musiman dengan Penerapan Data Susenas," Institut Pertanian Bogor, 2017.

[11]. H. J. Boonstra, "Time-Series Small Area Estimation for Unemployment Based on a Rotating Panel Survey," Stat. Netherlands, vol. 17, no. June, pp. 1–39, 2014.

[12]. A. Neves, D. Britz, and F. Ant, "Skew normal small area time models for the Brazilian annual service sector survey," Stat. Transit., vol. 21, no. 4, pp. 84–102, 2020, doi: 10.21307/stattrans-2020-032.

[13]. S. R. Patel, Statistical Inference: Classical and Bayesian Inference. I.K. International Publishing House Pvt. Limited, 2022.

[14]. T. Purwa, A. T. Rumiati, and I. Zain, "Small Area Estimation dengan Pendekatan Bivariate Hiearchical Bayes (HB) untuk Estimasi Rata-Rata Pengeluaran Per Kapita per Bulan Komoditi Makanan dan Non Makanan di Provinsi Bali

Tahun 2019," Institut Teknologi Sepuluh Nopember Surabaya, 2019.

[15]. B. W. Y. Priambodo and I. Irhamah, "Pemetaan Jumlah Property Crime di Provinsi Jawa Timur Menggunakan Metode Geographically Weighted Negative Binomial Regression (GWNBR) dan Geographically Weighted Poisson Regression (GWPR)," Inferensi, vol. 2, no. 2, p. 53, 2019, doi: 10.12962/j27213862.v2i2.6818.

Cite this article as :

Titin Yuniarty, Indahwati, Aji Hamim Wigena, "A Comparison Small Area Estimation for Skewed Data with EBLUP and Hierarchical Bayes Approaching using Rao-Yu Model", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 6, pp. 132-143, November-December 2023. Available at doi : https://doi.org/10.32628/IJSRSET231063
Journal URL : https://ijsrset.com/IJSRSET231063