Small Area Estimation of Child Poverty on Java Island In 2021 (Comparison of EBLUP and Hierarchical Bayes)

Nofita Istiana[1*], Erwin Tanur[2], Azka Ubaidillah[1], Yuliana Ria Uli Sitanggang[1], Rosalinda Nainggolan[2]
[1]Politeknik Statistika STIS, Indonesia
[2]Badan Pusat Statistik, Indonesia
*Corresponding author: nofita@stis.ac.id

Received: 2 June 2025
Revised: 4 November 2025
Accepted: 10 November 2025

ABSTRACT – Information about child poverty is very important to ensure that children get their rights. Indonesia's decentralized system requires child poverty data in each district/city. Data provision at this level is constrained by a non-specific sample design used for certain age groups, so the sample age group for children is not always sufficient for each district/city. Therefore, direct estimation produces a high relative standard error (RSE), so it requires small area estimation (SAE). SAE that is often used is EBLUP, which assumes that the variable of interest is normally distributed. Child poverty data does not meet the normality assumption, so SAE with Hierarchical Bayes with Beta distribution (HB Beta) is proposed in this study. The result is direct estimation, EBLUP, and HB Beta produce relatively similar estimated values, but HB Beta has the lowest RSE.

Keywords– Child poverty; HB Beta; RSE

I. INTRODUCTION

One of the 2020-2024 National Medium-Term Development Plan (RPJMN) strategies is to improve the quality of children through the realization that Indonesia is suitable for children (Idola). This can be achieved by strengthening the child protection system to ensure that children obtain their rights. One of the rights of children is to live in prosperity. However, welfare cannot be enjoyed by all children in Indonesia. According to data from Statistics Indonesia (BPS), in 2016 there were 10.86 percent of the poor, and 40.22 percent of them were children [1].

Poverty makes children lose their ability to survive. In addition, poor children are more vulnerable to exploitation, harassment, discrimination, and stigmatization [2]. Therefore, appropriate policies are needed to eradicate child poverty in Indonesia. The decentralized system in Indonesia results in the need for child poverty data in every district/city. So that policies between districts/cities can differ according to the level of child poverty in the region.

Information on child poverty is available at national and provincial levels [3]. The data was obtained from National Social and Economic Survey (Susenas) conducted by BPS. Data provision at smaller level (district/city) has constraints. The problem is that the sample design used is not specific for certain age groups, so that samples for the child age group are not always sufficient for each district/city. According to [4], if there is small sample size in sub-population, a very large error can be produced by direct estimation. In addition, direct estimation cannot be done for sub-populations that do not have sample. Therefore, an indirect estimation technique is needed that can increase the effectiveness of sample size so that it can reduce standard error. One of the techniques that can be used to solve this problem is small area estimation (SAE).

Research on estimating poverty data in small areas has been carried out by several researchers, including [5], [6], [7], [8], [9], dan [10]. Research on estimating child poverty in small areas is still very limited. [10] estimated child poverty in Indonesia in 2020 using Empirical Best Linear Unbiased Predictor (EBLUP). EBLUP assumes that direct estimators have normal distribution. The use of normality assumption on proportion variable can produce estimated values outside the range of 0 and 1. In addition, this method also produces Relative Standard Error (RSE) which is not good enough because there are still RSE values that are more than 50 percent. According to Australian Bureau of Statistics (ABS) in [11], RSE that exceeds 50 percent must be suppressed because of doubtful reliability. Therefore, another method is needed to handle proportion variable in order to obtain a better RSE.

One method that can be used to handle variables in the range 0 and 1 is Hierarchical Bayes (HB) with beta distribution [12]. Therefore, this study aims to use HB method with beta distribution to estimate child poverty on Java Island in 2021 at district/city level. The results of this method are then compared with EBLUP and direct estimation.

II. LITERATURE REVIEW

A. Small Area Estimation

Small area estimation is a method for estimating parameters in small areas by utilizing information from all areas through a liaison model based on additional information derived from censuses or administrative records [4]. Small area estimation is divided into two, namely design based and model based [13]. Design based uses survey weighing and survey design in the inference process. This method is also known as direct estimation. Meanwhile, model based or indirect estimation uses additional information in the inference process. Indirect estimation is done by adding information from surrounding areas and auxiliary variables derived from censuses or administrative records. The auxiliary variables used have relationship with variables of interest so as to increase the effectiveness of sample size. Indirect estimation in this study was carried out using EBLUP and Hierarchical Bayes.

B. Empirical Best Linear Unbiased Predictor (EBLUP)

Suppose auxiliary variable $x_i = \left(x_{1i}.x_{2i}.,,,.x_{pi}\right)^T$, and parameters to be estimated are θ_i for each area i $(i = 1.2.,,,m)$ then it is assumed that both have sufficient correlation, so that it can be expressed as linear model

$$\theta_i = x_i^T\beta + b_i u_i, \tag{1}$$

where b_i is positive random number, $\beta = \left(\beta_1.,,,.\beta_p\right)$ is regression coefficient vector of size $p \times 1$ and area random effect $u_i \sim N(0, \sigma_u^2)$ which is iid (independent and identically distributed). When estimating the value of $\underline{Y}$, it is assumed that direct estimator $(\underline{\hat{Y}})$ available expressed as

$$\hat{\theta}_i = g(\underline{\hat{Y}}) = \theta_i + e_i. \tag{2}$$

Sampling error $e_i \sim N(0, \psi_i)$ which *iid* and ψ_i is assumed to be known. Then, equations (1) and (2) are combined to linear mixed model,

$$\hat{\theta}_i = x_i^T\beta + b_i u_i + e_i. \tag{3}$$

It should be noted that model from equation (3) contains two errors, namely sampling error (e_i) and area random effect (u_i) which are assumed to be independent of each other. The model of equation (3) is known as Fay-Herriot model.

BLUP model assumes that variance component of area random effect is known. However, it is difficult to know variance of area random effect, so it must be estimated from existing samples. Replacing variance random effect of population area with variance random effect of sample area, a new model is obtained, namely Empirical BLUP (EBLUP) [13].

EBLUP model estimates σ_u^2 from sample using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) approaches. EBLUP model is stated as follows

$$\hat{\mu}_i^H = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) x_i^T \hat{\beta}, \quad i = 1.,,,.m, \tag{4}$$

which

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_i}. \tag{5}$$

The mean square error (MSE) of EBLUP estimator is as follows

$$MSE\left(\hat{\mu}_i^H\right) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2), \tag{6}$$

which

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i, \tag{7}$$

$$g_{2i}(\sigma_v^2) = (1 - \gamma_i)^2 x_i^T \left[\sum_{i=1}^m \frac{x_i x_i^t}{(\sigma_v^2 + \varphi_i)}\right]^{-1} x_i. \tag{8}$$

$$g_{3i}(\sigma_v^2) = \psi_i^2 (\psi_i + \sigma_v^2)^{-3} V(\sigma_v^2). \tag{9}$$

Furthermore, σ_v^2 is estimated by $\hat{\sigma}_v^2$ so that the equation turns into

$$MSE\left(\hat{\mu}_i^H\right) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2). \tag{10}$$

In the EBLUP approach, the classical Fay–Herriot model was employed directly on the area-level estimates of poverty proportion. Although proportion data are bounded between 0 and 1 and may not fully satisfy the normality assumption, several studies (e.g.,[13]) have shown that the Fay–Herriot model can still provide reliable results, particularly when the sampling variances are moderate and the direct estimates are not close to the boundaries.

Therefore, no transformation was applied to the response variable. Instead, model diagnostics were conducted to verify that the normality assumption was not severely violated. This approach also facilitates a direct comparison with the hierarchical Bayesian Beta model, which explicitly accounts for the bounded nature of the data.

C. Hierarchical Bayes (HB)

Different from classical statistical theory (frequentist), bayesian analysis treats all unknown parameters as random variables and has distribution [14]. The Bayesian theorem is based on posterior distribution which is combination of prior distribution (past information prior to observation) and observational data used to construct likelihood function. The relationship between posterior, prior and likelihood distributions can be written as follows

Posterior ditribution ∝ likelihood × prior distribution. (11)

According to Bayes' theorem, if there is parameter θ given by observational data y, then probability distribution for posterior θ will be proportional to multiplication between prior distribution θ and likelihood function θ given by the data y. Mathematically it can be written as follows.

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \tag{12}$$

$$f(\theta|y) \propto f(y|\theta)f(\theta) \tag{13}$$

where $f(\theta|y)$ is posterior distribution that is proportional to multiplication between likelihood functions $f(y|\theta)$ and prior distribution $f(\theta)$.

D. Distribusi Prior

In Bayesian analysis, a prior distribution reflects the information or belief about unknown parameters before any data are observed. The choice of a prior mainly depends on how much prior knowledge is available and on the expected properties of the resulting posterior distribution. In practice, several types of priors can be considered [15]:

1. Conjugate prior or non-conjugate prior
   A conjugate prior is selected when it shares the same functional form as the likelihood, making the posterior distribution easier to derive. Otherwise, a non-conjugate prior can be used, although it typically requires simulation-based inference.
2. Proper prior or improper prior
   A proper prior assigns a finite total probability to the entire parameter space, while an improper prior does not. The latter is often used when little prior information is available, as long as the resulting posterior remains proper.
3. Informative prior or non-informative prior
   Informative priors are based on previous studies or expert knowledge, whereas non-informative priors are intended to have minimal influence on the inference results.
4. Pseudo prior
   These are priors derived from frequentist estimates, often used to help stabilize convergence in complex hierarchical models.

E. Hierarchical Bayesian Model with Beta Distribution (HB Beta)

Let $\hat{\theta}_i$ denote the direct estimate for the true parameter θ_i in area i ($i = 1, \dots, m$). Referring to [13], the Hierarchical Bayesian model with a Beta distribution can be formulated as follows:

1. $\hat{\theta}_i|\theta_i, \beta, \sigma_u^2 \sim Beta(a_i, b_i)$
2. $a_i = \theta_i v$
3. $b_i = (1 - \theta_i)v$
4. $v \sim Gamma(g_1, g_2)$
5. $Logit(\theta_i) \sim N(x_i^T \beta, \sigma_u^2)$
6. $\beta_j \sim N\left(\mu_{\beta_j}, \sigma_{\beta_j}^2\right)$
7. $\sigma_u^2 \sim IG(t_1, t_2)$

where IG(t_1, t_2) is inverse gamma distribution with parameters t_1 dan t_2. In this case the values g_1, g_2, μ_{β_j}, $\sigma_{\beta_j}^2$, t_1 and t_2 are fixed at predetermined values. Combining the likelihood and prior distributions, the posterior distribution is expressed as:

$$f(\theta_i|\hat{\theta}_i) \propto f(\hat{\theta}_i|\theta_i, \beta, \sigma_u^2)f(\beta_j)f(\sigma_u^2)f(v). \tag{14}$$

In the Hierarchical Bayesian (HB) model with Beta distribution, prior distributions were assigned to the model parameters at the third hierarchical level. Specifically, the regression coefficients β_j were assigned Normal priors $N\left(\mu_{\beta_j}, \sigma_{\beta_j}^2\right)$, the area-level variance σ_u^2 was assigned an Inverse-Gamma prior $IG(t_1, t_2)$, and the precision parameter v followed a Gamma prior $Gamma(g_1, g_2)$. These priors were specified as weakly informative to ensure that the posterior inference was mainly driven by the observed data while maintaining numerical stability in small areas with limited samples. Since the Beta likelihood and Normal–Inverse Gamma priors are non-conjugate, the resulting posterior distributions have no closed-form expression and were estimated using the Markov Chain Monte Carlo (MCMC) method.

Estimator for HB Beta model and variance of HB Beta estimator are obtained from posterior formula as follows

$$\hat{\theta}_i^{HB} = E(\theta_i|\hat{\theta}_i) \tag{15}$$

$$V(\hat{\theta}_i^{HB}) = V(\theta_i|\hat{\theta}_i). \tag{16}$$

The values of the HB Beta estimator and its variance were obtained using the Markov Chain Monte Carlo (MCMC) simulation method. The MCMC approach is highly effective in reducing the computational burden when solving complex integration problems. [16] suggests that this method enables simulation by generating random samples from complex stochastic models. The basic idea of MCMC is to iteratively generate sample data from the posterior distribution according to a Markov chain process using Monte Carlo simulations until convergence to the posterior distribution is achieved [15]. This convergence state is known as the stationary condition or equilibrium. Once the chain has reached this state, parameter samples are drawn, and these samples are guaranteed to represent the posterior distribution of the corresponding parameters.

F. Child Poverty

Data on child poverty in Indonesia is presented in terms of percentage of poor children, so this study also presents calculation results in percentage form. According to BPS [3], percentage of poor children is percentage of children aged 0-17 years who live in households below poverty line. The calculation of this indicator uses the following formula.

$$percentage\ of\ poor\ children = \frac{number\ of\ children\ aged\ 0\text{–}17\ years\ in\ poor\ households}{number\ of\ children\ aged\ 0\text{–}17\ years} \times 100 \tag{17}$$

III. METHODOLOGY

A. Data Sources

The data used in this study came from National Socioeconomic Survey (Susenas) and village potential data collection (Podes). Direct estimation of child poverty was obtained from March 2021 Susenas. Meanwhile, auxiliary variables were obtained from 2019 Podes.

B. Data Analysis

The process of data analysis in this study is as follows.

1. Identify the data used to produce child poverty and data that produce auxiliary variables.
2. Estimating child poverty at district/city level with direct estimation method and its RSE. Direct estimation is based on Susenas survey design.
3. Exploring data on direct estimation of child poverty.
4. Choose auxiliary variables from Podes data. The first step in selecting auxiliary variables is examine variables that have high correlation with child poverty. High correlation is indicated by significant Pearson correlation value. Then select auxiliary variables using stepwise method.
5. Estimating child poverty at district/city level with EBLUP and its RSE.
6. Estimating child poverty at district/city level with HB Beta and its RSE.
7. Comparing RSE of direct estimation, EBLUP, and HB Beta to get the best model. The best model is model with the lowest RSE.

IV. RESULTS AND DISCUSSIONS

A. Data Exploration

Data exploration was carried out on the results of direct estimation of child poverty at district/city level on Java Island in 2021. The direct estimation was carried out based on survey design from Susenas. Based on Table 1, the lowest child poverty is 0.391 percent and the highest is 80.191 percent, while the average child poverty by district/city on Java is 15.625 percent.

RSE for direct estimation has the lowest value of 9.313 percent and the highest of 70.002 percent (see Table 2). There are 4 districts/cities with RSE above 50 percent. This shows that direct estimation produces an estimate that is not good enough. Therefore, another method is needed to estimate child poverty which aims to reduce RSE. The method that can be used to improve direct estimation are EBLUP and HB with beta distribution.

Table 1 Summary Statistics of Child Poverty Direct Estimation on Java Island

Summary Statistics	Child Poverty	RSE
Minimum	0.391	9.313
First quartile	4.982	15.106
Median	10.573	19.458
Mean	11.801	23.107
Third quartile	16.378	27.044
Maximum	39.019	70.002

Table 2 RSE Recapitulation of Child Poverty Direct Estimation on Java Island

RSE	<= 25	(25,50)	>=50
Number of districts/cities	81	34	4

The SAE method that is often used is EBLUP. EBLUP assumes that the variable of interest is normally distributed. Therefore, direct estimation of child poverty is identified by its distribution. Histogram and boxplot direct estimation of child poverty shows that the data is skewed to the right (positively skewed) (Figures 1 and 2). This condition occurs because there is more data at high value; besides that, it can also occur because of an upper outlier. This indicates that child poverty data is not normally distributed. The same thing can be seen from the Shapiro-Wilk normality test, which produces a p-value of 5.9×10^{-5}, which means that child poverty data on Java Island is not normally distributed.

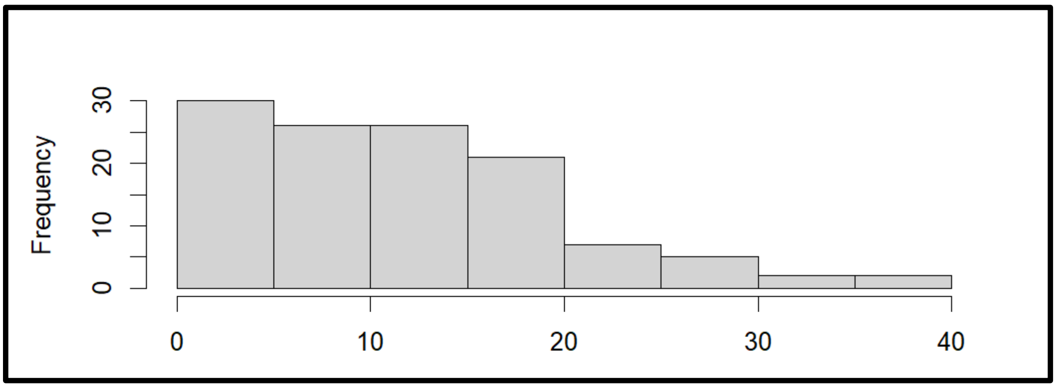

Figure 1 Histogram of Child Poverty Direct Estimation by District/City on Java Island

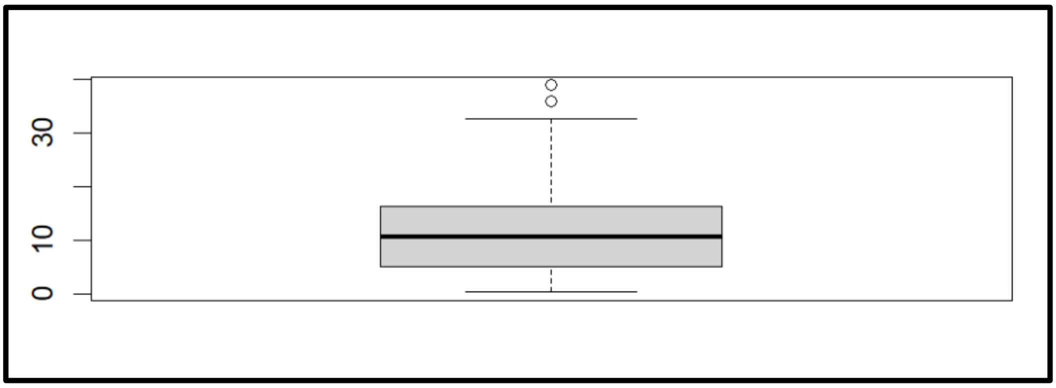

Figure 2 Boxplot of Child Poverty Direct Estimation by District/City on Java Island

There are many variables that can be used as auxiliary variables from Podes data as shown in Table 3. Therefore, variable selection is carried out before doing SAE. Variable selection begins by looking at correlation between variables in Podes and child poverty. Furthermore, variables with significant Pearson correlation value are selected for regression modeling. After that, stepwise regression was carried out to select variables that really affect child poverty.

Table 3 Auxiliary Variables for Estimating Child Poverty at District/City Level on Java Island

Variable	Description
X_1	Percentage of villages that have a pawnshop
X_2	Percentage of villages that have a hospital
X_3	Percentage of villages that have academies/universities
X_4	Average number of resident doctors
X_5	Percentage of villages that have private commercial banks
X_6	Percentage of villages that have SMA/MA/SMK/SMALB
X_7	Percentage of villages with flat land
X_8	Average number of PAUD/TK/RA/BA educational facilities
X_9	Percentage of villages that have pubs/discotheques/karaoke places
X_{10}	Average number of government commercial banks
X_{11}	Average number of practicing midwives
X_{12}	Number of restaurants
X_{13}	Average number of sufferers of the COVID-19 outbreak
X_{14}	Average distance to the nearest poskesdes
X_{15}	Number of doctors' offices
X_{16}	Average number of families
X_{17}	Average number of families using PLN electricity
X_{18}	Percentage of villages that have food stalls
X_{19}	Average number of SD/MI/SDLB educational facilities
X_{20}	Percentage of villages that have hotels
X_{21}	Number of sufferers of the COVID-19 outbreak
X_{22}	Percentage of villages experiencing floods

Variable	Description
X_{23}	Number of villages that have furniture industry
X_{24}	Number of villages that have disabled people
X_{25}	Number of deaf people
X_{26}	Number of industrial centers
X_{27}	Poly coverage

B. Small Area Estimation of Child Poverty

Based on the summary statistics of estimated values using three methods (direct, EBLUP, and HB Beta) in Table 4, the lowest minimum of child poverty is produced by EBLUP, while the highest maximum is produced by direct estimation. All three methods yield an average child poverty around 11 percent. Estimating child poverty with these three methods produces relatively similar values. This can be seen in Figure 3 where the graph almost coincides.

Table 4 Summary Statistics of Child Poverty Estimation at District/City Level on Java Island

Summary Statistics	Direct	EBLUP	HB Beta
Minimum	0.391	0.3873	0.5133
First quartile	4.982	5.2586	5.2145
Median	10.573	11.0324	10.8233
Mean	11.801	11.4334	11.8094
Third quartile	16.378	15.9095	16.0031
Maximum	39.019	31.2892	37.7436

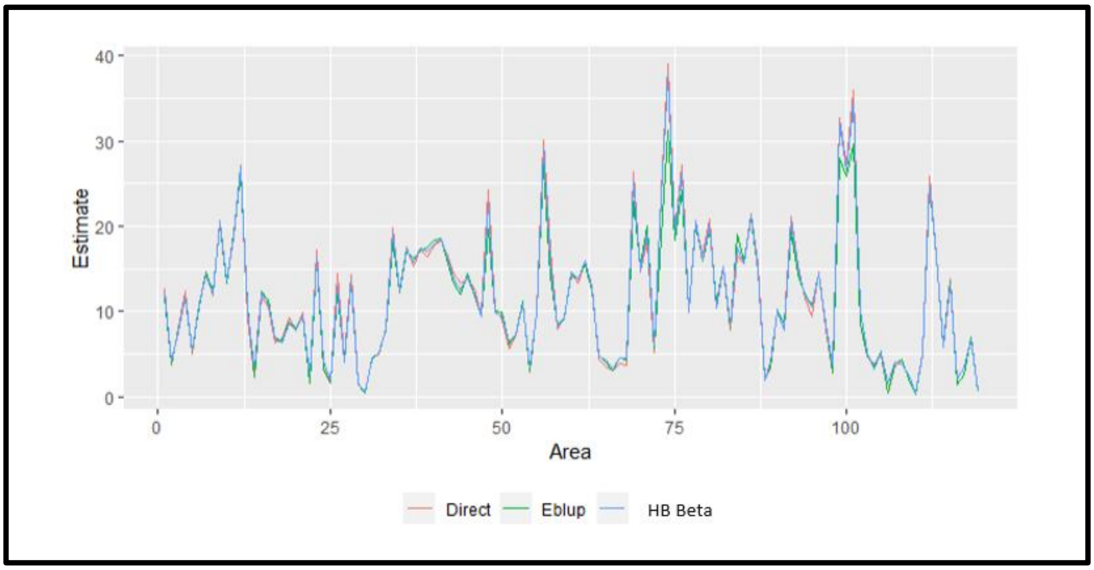

Figure 3 Graph of Child Poverty Estimated Value at District/City Level by Direct Estimation, EBLUP, and HB Beta

Although the estimated values using direct estimation, EBLUP, and HB Beta produce relatively similar values, estimation using HB Beta produces a smaller RSE than direct estimation and EBLUP (Figure 4). This can be seen from the RSE HB Beta graph, which is located under the RSE graph of direct estimation and EBLUP. In addition, it appears that RSE of EBLUP is still close to RSE of direct estimation. So that EBLUP has not been able to significantly reduce RSE.

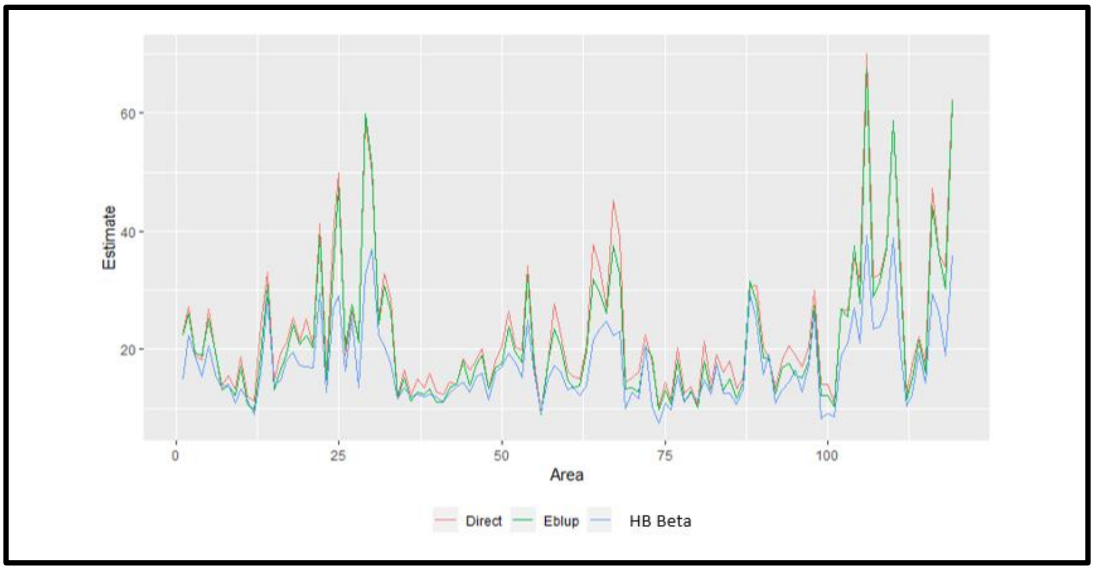

Figure 4 RSE Graph of Child Poverty Estimated Value at District/City Level by Direct Estimation, EBLUP, and HB Beta

Table 5 RSE Summary Statistics of Child Poverty Estimation at District/City Level on Java Island

Summary Statistics	Direct	EBLUP	HB Beta
Minimum	9.313	8.995	7.519
First quartile	15.106	13.429	12.516
Median	19.458	18.078	15.580
Mean	23.107	21.566	17.318
Third quartile	27.044	26.324	20.700
Maximum	70.002	67.627	39.193

The same thing can also be seen through summary statistics of RSE (see Table 5). EBLUP and HB Beta were able to reduce the mean and median of RSE when compared to direct estimation. However, the RSE maximum value of EBLUP is still close to direct estimation. It can be concluded that HB Beta tends to produce lower RSE than direct estimation and EBLUP. Based on the recapitulation in Table 6, none of the RSE HB Beta is above 50 percent. This shows that HB Beta has been able to improve direct estimation results.

Table 6 RSE Recapitulation on Small Area Estimation of Child Poverty at District/City Level on Java Island

RSE	<= 25	(25,50)	>=50
Direct	81	34	4
Eblup	85	29	5
HB Beta	103	16	

V. CONCLUSIONS AND SUGGESTIONS

Direct estimation, EBLUP, and HB Beta produce relatively similar estimated values but quite different RSE. The use of EBLUP is not good enough to estimate child poverty at the district/city level on Java Island in 2021. This can be seen from the presence of RSE, which is more than 50 percent, although more areas have RSE below 25 percent compared to direct estimation. The use of HB Beta is able to produce more accurate estimated values because RSE for all regions is below 50 percent. Child poverty data that is not normally distributed can be the cause of inappropriate use of EBLUP. Therefore, the selection of the SAE method should consider the distribution form of interest variable. Furthermore, SAE research with various types of data is important to develop.

REFERENCES

[1] BPS and UNICEF, *Analisis Kemiskinan Anak dan Deprivasi Hak-Hak Dasar Anak di Indonesia*. 2017.

[2] I. K. Kertayana, "Kondisi Kemiskinan Anak Dan Harapan Sustainable Development Goals (Sdgs) Di Indonesia," 2017.

[3] KPPPA and BPS, "Indeks Perlindungan Anak Indonesia 2020," 2020.

[4] I. Molina and J. N. K. Rao, "Small Area Estimation of Poverty Indicators," vol. 38, no. 3, pp. 369–385, 2010, doi: 10.1002/cjs.

[5] M. D. Esteban, D. Morales, A. Pérez, and L. Santamaría, "Small Area Estimation Of Poverty Proportions Under Area-Level Time Models," *Comput. Stat. Data Anal.*, vol. 56, no. 10, pp. 2840–2855, 2012, doi: 10.1016/j.csda.2011.10.015.

[6] M. Y. Darsyah and R. Wasono, "Pendugaan Tingkat Kemiskinan Di Kabupaten Sumenep Dengan Pendekatan SAE," 2013.

[7] T. Hobza, D. Morales, and L. Santamaría, "Small Area Estimation Of Poverty Proportions Under Unit-Level Temporal Binomial-Logit Mixed Models," *TEST*, 2017, doi: 10.1007/s11749-017-0545-3.

[8] R. Ghaniswati, A. Saefuddin, and A. Kurnia, "Nonstationary EBLUP on Prediction of Poverty Rate at Village Level in Lembata Regency," vol. 6, no. 1, pp. 124–130, 2019.

[9] D. Gartina and L. Khikmah, "Pendugaan Kemiskinan Menggunakan Small Area Estimation Dengan Pendekatan Emperical

Best Linier Unbiased Prediction (EBLUP),” vol. 8, no. 2, 2020.
[10] J. Wulansari, N. Permatasari, and A. Ubaidillah, “Pendugaan Area Kecil Persentase Anak-anak Usia Kurang dari 18 Tahun yang Hidup di Bawah Garis Kemiskinan Tingkat Kabupaten / Kota di Indonesia Tahun 2020,” pp. 383–394, 2022.
[11] S. Horn, “Hilda Project Technical Paper Series (Guide to Standard Errors for Cross Section Estimates),” 2004.
[12] P. Anjoy, H. Chandra, and P. Basak, “Estimation of Disaggregate-Level Poverty Incidence in Odisha Under Area-Level Hierarchical Bayes Small Area Model,” *Soc. Indic. Res.*, vol. 144, no. 1, pp. 251–273, 2018, doi: 10.1007/s11205-018-2050-9.
[13] J. N. K. Rao and I. Molina, *Small Area Estimation*. Hoboken, New Jersey: John Wiley &Sons, Inc., 2015.
[14] W. M. Bolstad, *Introduction to Bayesian Statistics*. Hamilton, New Zealand: John Wiley & Sons, Inc, 2007. doi: 10.1137/1.9781611970654.
[15] I. Ntzoufras, *Bayesian Modeling Using WinBUGS*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2009.
[16] D. P. M. Scollnik, “An Introduction to Markov Chain Monte Carlo Methods And Their Actuarial Applications,” pp. 114–165.