Scientific Journal of Informatics
Vol. 12, No. 2, May 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

The Empirical Best Linear Unbiased Prediction and The Emperical Best Predictor Unit-Level Approaches in Estimating Per Capita Expenditure at the Subdistrict Level

Ghina Fauziah[1*], Anang Kurnia[2], Anik Djuraidah[3]
[1, 2, 3]Department of Statistics, IPB University, Indonesia

Abstract.
Purpose: This study aims to estimate and evaluate per capita expenditure at the subdistrict level in Garut Regency by employing unit-level Small Area Estimation (SAE) techniques, specifically utilizing the Empirical Best Linear Unbiased Predictor (EBLUP) and the Empirical Best Predictor (EBP) methods.
Methods: The data used in this study are socio-economic data, specifically per capita household expenditure in Garut Regency. Socio-economic data generally skew positively rather than the normal distribution, so a method that can approximate or come close to the normal distribution is needed, for example, log-normal transformation. To improve the performance of EBLUP, which may lead to inefficient estimators because of violation of the assumption of normality, this study proposes the Empirical Best Predictor (EBP) method. It handles positively skewed data by applying log-normal transformation to sample data so that it more closely conforms to the desired distribution.
Result: The EBP results are more stable than EBLUP since EBLUP is highly sensitive to outliers, and in cases where the normality assumption is violated, it produces a significant mean square error and inefficient estimators. Evaluating the estimates with both EBLUP and EBP shows Relative Root Mean Squared Error (RRMSE) values above 25%, especially in the subdistricts of Pamulihan, Sukaresmi, and Kersamanah. This is probably due to the household samples being taken in these three subdistricts being comparatively small compared to the other.
Novelty: In this research, we use EBP to improve the performance of EBLUP, which produces inefficient estimators when the normality assumption is violated.

Keywords: Small area estimation unit level, EBLUP, EBP
Received May 2025 / **Revised** June 2025 / **Accepted** June 2025

INTRODUCTION

Small Area Estimation (SAE) is a statistical methodology for the production of reliable and accurate estimates of characteristics of populations, i.e., socio-economic indicators, in geographic areas where data are available from only a limited number of or no direct survey sample observations [1], [2]. This method proves extremely useful when small sample sizes do not allow conventional estimation techniques. SAE offers a practical alternative approach for producing estimates for areas that cannot be surveyed adequately through direct survey [3].

SAE models can be further categorized into area-level and unit-level models. Additional information at the area level makes the area-level SAE model available. Meanwhile, data at the unit level should ideally use an error components regression model [1], [4] which is simple and widely used. Furthermore, normality is assumed by the SAE method for both sampling errors and random area effects, irrespective of whether it is an area-level or a unit-level model [5]. These assumptions are typically not met in practice. For instance, the response variable is commonly distributed as non-normal in socio-economic data with a right skew, and thus, the normality assumption for sampling error is violated [6], [7].

To estimate the mean or total, the Empirical Best Linear Unbiased Predictor (EBLUP) obtained by minimizing the Mean Square Error (MSE) and calculating the unknown quantities with the use of the Restricted Maximum Likelihood (REML) method is widely applied [8], [9]. EBLUP assumes that the functional form (relationship) between the response variable and some additional information is linear.

*Corresponding author.
Email addresses: ghinafauziah57@gmail.com (Fauziah)*, anangk@apps.ipb.ac.id (Kurnia), anikdjuraidah@apps.ipb.ac.id (Djuraidah)
DOI: 10.15294/sji.v12i2.25037

Furthermore, EBLUP assumes normality for the random effects of small areas [10]. However, in practice, these assumptions are often not met. For example, in socio-economic data, the response variable often follows a non-normal distribution, skewed to the right, thus violating the normality assumption for the sampling error [11].

In socio-economic research, particularly regarding per capita expenditure variables, it is often found that these variables have a positively skewed distribution. Small area models involving welfare variables such as expenditure or income are often specified using logarithmic transformation applied to the sample data [12]. EBLUP is not suitable for adjusting skewed data or handling outliers. It can be highly influenced by outliers or violations of model assumptions [13]. If the normality assumption cannot be met due to outliers, EBLUP will inflate the mean square error and result in inefficient estimators [14].

In this study, to improve the performance of EBLUP, which results in inefficient estimators due to violations of the normality assumption, the Empirical Best Predictor (EBP) method is proposed to correct positively skewed data by applying a log-normal transformation to the sample data of the response variable, which will then approximate the desired distribution [15]. This study uses a unit-level model, so the variance of per capita expenditure across subdistricts is quite diverse. After obtaining the estimation results using the EBLUP and EBP models, a measurement of the goodness of fit of the model will be carried out to determine the performance of the two models. In this study, the Relative Root Mean Squared Error (RRMSE) will be used to provide an overview of the magnitude of the relative error rate of the estimation results that are standardized to eliminate the unit factor or relative value of the average root mean square error [12], [16].

Garut Regency is one of three regencies with the lowest Human Development Index (HDI) in West Java. The dimensions of the HDI are based on three things: longevity and good health, knowledge, and standard of living. Per capita expenditure is one of the components used to measure the dimensions of the standard of living [7], [17]. Garut Regency is one of the areas with the lowest average per capita expenditure in West Java, IDR 11,694,000 for men and IDR 4,523,000 for women. Socio-economic data, including per capita expenditure, has a right-skewed distribution, so modeling using small area estimation will violate the assumption. Therefore, EBP is used to handle distributed data skewed to the right. Therefore, in this study, Garut Regency's per capita expenditure data is used as the object to be estimated.

METHODS

Data description

The data used in this study are secondary data obtained from the 2022 National Socio-Economic Survey (SUSENAS KOR) and the 2021 Village Potential (PODES) data for Garut Regency, sourced from BPS Indonesia, which provides information at the household level. The SUSENAS KOR data are used for parameter estimation at the subdistrict level using SAE. To apply SAE, additional variables are required as supplementary information. The PODES data serve this purpose, providing information on all villages within each subdistrict, such as geographical characteristics, the number of households with telephone access, the number of poor households, the number of BPJS users, the number of PLN users, types of drinking water used, education, health services, and more. A complete list of the variables used in this study is presented in Table 1 below.

Table 1. Research Variables

Variable	Description	Data Source
$\mathbf{Y}$	Per capita expenditure	SUSENAS KOR 2022
$\mathbf{X_1}$	Number of PLN user households	PODES 2021
$\mathbf{X_2}$	Number of Certificates of Inability	PODES 2021
$\mathbf{X_3}$	Number of households subscribed to landline phones	PODES 2021
$\mathbf{X_4}$	Number of village markets	PODES 2021
$\mathbf{X_5}$	Number of active integrated health service posts (posyandu)	PODES 2021
$\mathbf{X_6}$	Number of Village Cooperative Units	PODES 2021
$\mathbf{X_7}$	Number of BUMDes business units	PODES 2021
$\mathbf{X_8}$	Number of beneficiary households	PODES 2021
$\mathbf{X_9}$	Number of health centers	PODES 2021

Small area estimation

Small Area Estimation (SAE) is a statistical technique used to estimate parameters for small sub-populations, typically when the target sub-population is included in a broader survey. The term "small area"

refers to small geographic units such as districts, subdistricts, or villages. In national surveys that cover the entire population, the sample size within these small areas is often too limited to produce accurate direct estimates. To address this limitation, additional data sources, such as census data, can be used to improve the reliability of the estimates for these small areas. In this study, small area estimation will be used using a unit level model.

This refers to a model in which the data for the response variable and the supplementary variables must correspond on an individual basis. This allows for the formation of a nested regression model [18]–[21]:

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij} \quad i = 1, \dots, m, j = 1, \dots, n_i \tag{1}$$

where β is the vector of regression coefficients $(\beta_0, \dots, \beta_p)$ of size $(px1)$, v_i is the random area effect, and e_{ij} is the error of unit j in small area i. The random area effect u_i and e_{ij} are independently distributed as $v_i \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.

Empirical best linear unbiased predictor (EBLUP)

In the best linear unbiased predictor (BLUP) linear model, which has the smallest variance within its class of unbiased predictors, the predictor often relies on unknown variance and covariance parameters. When these unknown parameters are replaced with estimates, BLUP becomes the empirical best linear unbiased predictor (EBLUP) [22]–[25]. Below is the nested error regression model:

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij} \tag{2}$$

where y_{ij} is the response variable for unit *j* in area *i*, x_{ij} is the vector of covariates for unit *j* in area *i*, β is the vector of regression parameters, and v_i and e_{ij} are the random area effects and random sampling errors, respectively, with $v_i \sim N(o, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, and $y_{ij} \sim N(x_{ij}^T\beta, \sigma_v^2 + \sigma_e^2)$.

Model (2) is a special case of the General Linear Mixed Model (GLMM) which can be written in the form:

$$y = X\beta + Zv + e \tag{3}$$

where *y* is the observation vector (*N* x *1*), X is the covariate matrix (*N* x *p*), β is the regression parameter vector (*p* x *1*), *Z* is the random effects matrix (*N* x *M*) with $v \sim N(0, G)$ being the vector of random area effects, and $e \sim N(0, R)$ is the residual error vector. Therefore, the distribution of *y* is $y \sim N(X\beta, V)$, where $V = ZGZ^T + R$.

Therefore, the EBLUP for μ_i under the nested error regression model is given as:

$$\hat{\mu}_i^{EBLUP} = \frac{1}{N_i}\left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{EBLUP}\right] \tag{4}$$

where $\hat{y}_{ij}^{EBLUP} = x_{ij}^T\hat{\beta} + \hat{v}_i$ with $\hat{\beta}$ is the estimated regression parameter $\hat{\beta} = (\sum_{i=1}^M X_i^T V_i^{-1} X_i)^{-1}(\sum_{i=1}^M X_i^T V_i^{-1} y_i)$, and $\hat{v}_i$ is the estimated area effect obtained with the shrinkage factor $\hat{v}_i = \hat{\gamma}_i(\bar{y}_{is} - \bar{x}_i^T\hat{\beta})$, where the shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{n_i}}$ where $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ are variance estimates obtained using the maximum likelihood (ML) method, restricted maximum likelihood (REML), or method of moments.

According to Prasad and Rao (1990), the mean squared error (MSE) of EBLUP can be approximated as [12]:

$$MSE(\hat{\mu}_i^{EBLUP}) \approx g_{1i}(\sigma_v^2, \sigma_e^2) + g_{2i}(\sigma_v^2, \sigma_e^2) + g_{3i}(\sigma_v^2, \sigma_e^2) \tag{5}$$

Where:

$$g_{1i}(\sigma_v^2, \sigma_e^2) = \gamma_i\left({\sigma_e^2}/{n_i}\right)$$

$$g_{2i}(\sigma_v^2, \sigma_e^2) = (\bar{X}_i - \gamma_i\bar{x}_i)^T(X^TV^{-1}X)^{-1}(\bar{X}_i - \gamma_i\bar{x}_i)$$

$$g_{3i}(\sigma_v^2, \sigma_e^2) = n_i^{-2}\left(\sigma_v^2 + {\sigma_e^2}/{n_i}\right)^{-3}[\sigma_e^4 var(\tilde{\sigma}_v^2) + \sigma_v^4 var(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_v^2 cov(\tilde{\sigma}_v^2, \tilde{\sigma}_v^2)]$$

where $var(\tilde{\sigma}_e^2) = 2(n - M - p + \lambda)^{-1}\sigma_e^4$; $\lambda = 0$ if model (3) does not have an intercept, and $\lambda = 1$ otherwise. $var(\tilde{\sigma}_v^2) = 2n_*^2(n - M - p + \lambda)^{-1}x(M - \lambda)(n - p)\sigma_e^4 + 2n_*^2\sigma_v^2\sigma_e^2 + n_{**}\sigma_v^4$. Thus, the estimator for MSE($\hat{\mu}_i^{EBLUP}$) is as follows:

$$MSE(\hat{\mu}_i^{EBLUP}) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \tag{6}$$

where g_{1i}, g_{2i}, dan g_{3i} are calculated using variance estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$.

Empirical best predictor (EBP)

In many socio-economic analyses, particularly those involving income or expenditure, the response variable often exhibits a positive skew. In this case, EBLUP is an inaccurate small area parameter estimator because it violates the normality assumption. The Empirical Best Predictor (EBP) has been developed, which assumes that the response variable follows a lognormal distribution. The log transformation is modeled using the nested error regression framework as follows [26]–[28]:

$$log y_{ij} = y_{ij}^{\#} = x_{ij}^T \beta + v_i + e_{ij} \quad (7)$$

where v_i and e_{ij} are independent; with $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, and $y_{ij}^{\#}$ follows a normal distribution with a mean of $x_{ij}^T \beta$ and variance $\sigma_v^2 + \sigma_e^2$, or $y_{ij}^{\#} \sim N(x_{ij}^T \beta, \sigma_v^2 + \sigma_e^2)$.

If the linear parameter of interest to be estimated is the small area mean, then the Best Predictor (BP) for μ_i is as follows:

$\hat{\mu}_i^{BP} = \frac{1}{N_i}\left[\sum_{j \epsilon s_i} y_{ij} + \sum_{j \epsilon r_i} \hat{y}_{ij}^{BP}\right]$

where $\hat{y}_{ij}^{BP} = E\left[y_{ij} | (y, x)\right] = exp\left(x_{ij}^T \hat{\beta} + \gamma_i\left(\bar{y}_{is}^* - \bar{x}_{is}^T \hat{\beta}\right) + 0.5\sigma_e^2(\gamma_i n_i^{-1} + 1)\right)$; $\gamma_i = \sigma_v^2(\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1}$; $\bar{y}_{is}^* = \frac{1}{n_i}\sum_{j \epsilon s_i} y_{ij}^*$; $\bar{x}_{is}^T = \frac{1}{n_i}\sum_{j \epsilon s_i} x_{ij}^T$.

In practice, $\hat{\mu}_i^{EBP}$ is rarely known because the predictor is a function of the variance parameters $\theta = (\sigma_v^2, \sigma_e^2)$ which are unknown. Therefore, these parameters are replaced by their estimates $\hat{\theta} = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. Thus, $\hat{\mu}_i^{EBP}$ is obtained as follows [29]:

$$\hat{\mu}_i^{EBP} = E\left(y_{ij} | v_i\right) = \frac{1}{N_i}\left[\sum_{j \epsilon s_i} y_{ij} + \sum_{j \epsilon r_i} \hat{y}_{ij}^{EBP}\right] \quad (8)$$

where $\hat{y}_{ij}^{EBP} = \left(c_{ij}\right)^{-1} \exp\left(x_{ij}^T \hat{\beta} + \hat{v}_i + 0.5\sigma_e^2(\gamma_i n_i^{-1} + 1)\right)$ with $\hat{v}_i = \hat{\gamma}_i\left(\bar{y}_{is}^* - \bar{x}_{is}^T \hat{\beta}\right)$, bias correction factor $c_{ij} = 1 + \frac{1}{2} x_{ij} V\left(\hat{\beta}\right) x_{ij}^T + \frac{1}{8}\nabla(\hat{\sigma}_v^2 + \hat{\sigma}_e^2)$ and $V\left(\hat{\beta}\right) = [(X^T V^{-1} X)^T]^{-1}$ is the variance of $\hat{\beta}$.

The MSE of EBP for μ_i is as follows [30], [31]:

$$MSE(\hat{\mu}_i^{EBP}) = M_{1i}(\theta) + M_{2i}(\theta) \quad (9)$$

where $M_{1i}(\theta) = E[(\hat{\mu}_i^{BP}(\theta) - \mu_i)^2]$ and $M_{2i}(\theta) = E\left[\left(\hat{\mu}_i^{EBP}\left(\hat{\theta}\right) - \hat{\mu}_i^{BP}(\theta)\right)^2\right]$.

$$\begin{aligned} M_{1i}(\theta) &= E[(\hat{\mu}_i^{BP}(\theta) - \mu_i)^2] \\ &= \frac{k_i}{N_I^2}\left\{\left(\sum_{j \epsilon r_i} exp\left(x_{ij}^T \beta\right)\right)^2 \xi_i + \left(\sum_{j \epsilon r_i} exp\left(2x_{ij}^T \beta\right)\right)\psi_i\right\} \end{aligned} \quad (10)$$

where $\xi_i = exp(\gamma_i \eta_i^{-1}\sigma_e^2) - 1$, $\psi_i = exp(\gamma_i \eta_i^{-1}\sigma_e^2 + \sigma_e^2) - exp(\gamma_i \eta_i^{-1}\sigma_e^2)$ and $k_i = exp\{2\beta_0 + 2\gamma_i^2(\sigma_v^2 + n_i^{-1}\sigma_e^2) + \gamma_i \eta_i^{-1}\sigma_e^2 + \sigma_e^2\}$

$$\begin{aligned} M_{2i}(\theta) \approx N_i^{-2}\Big[&\sum_{j,k \epsilon r_i} exp\{\mu_{ij} + \mu_{ik} + \gamma_i^2(\sigma_v^2 + n_i^{-1}\sigma_e^2) + 0.5\delta_{ijk} I[j \neq k] + 2v_{ij} I[j = k]\} - 2\sum_{j,k \epsilon r_i} exp\{\mu_{ij} + \mu_{ik} + 2\gamma_i^2(\sigma_v^2 + n_i^{-1}\sigma_e^2) + 0.5v_{ij}\} + \sum_{j,k \epsilon r_i} exp\{\mu_{ij} + \mu_{ik} + \\ &2\gamma_i^2(\sigma_v^2 + n_i^{-1}\sigma_e^2)\}\Big] = \bar{M}_{2i}(\theta). \end{aligned} \quad (11)$$

where:

$$\delta_{ijk} = \left(b_{ij} + b_{ik}\right)^T V\{\hat{\beta}\}\left(b_{ij} + b_{ik}\right) + 4tr(E[f_i f_i^T] V\{\hat{\sigma}^2\}) \quad (12)$$

$$v_{ij} = b_{ij}^T V\{\hat{\beta}\} b_{ij} + tr(E[f_i f_i^T] V\{\hat{\sigma}^2\}) \quad (13)$$

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right)^T;\ \hat{\sigma} = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)^T; f_i = \left(f_{i,1} f_{i,2}\right)^T \quad (14)$$

$$b_{ij}^T = (1 - \gamma_i, x_{ij}' - \gamma_i \bar{x}_{is}^T) \quad (15)$$

$$f_{i,1} = (\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1}(1 - \gamma_i)(\bar{y}_{is}^* - \beta_0 - \bar{x}_{is}^T \beta_1) + 0.5(1 - \gamma_i)^2 \quad (16)$$

$$f_{i,2} = -\gamma_i(\sigma_v^2 + n_i^{-1}\sigma_e^2)^{-1} n_i^{-1}(\bar{y}_{is}^* - \beta_0 - \bar{x}_{is}^T \beta_1) + 0.5\gamma_i^2 n_i^{-1} + 0.5 \quad (17)$$

Model fit measure

Model fit is measured using the Relative Root Mean Squared Error (RRMSE). The RRMSE provides an indication of the magnitude of relative error in standardized predictions, helping to eliminate the unit factor [32]. It effectively expresses the relative value of the root mean square error. The best-fitting model is the one with the smallest RRMSE value. Estimation results can be categorized based on the RRMSE as follows:

a. RRMSE ≤ 25%: considered accurate;
b. 25%< RRMSE ≤ 50%: caution is needed if it is to be used;

c. RRMSE > 50%: considered very inaccurate.

$$\text{RRMSE}(\hat{\mu}_i) = \frac{\sqrt{MSE(\hat{\mu}_i)}}{\hat{\mu}_i} \quad (18)$$

Analysis stages

The sequential steps adopted in this research are shown in Figure 1.

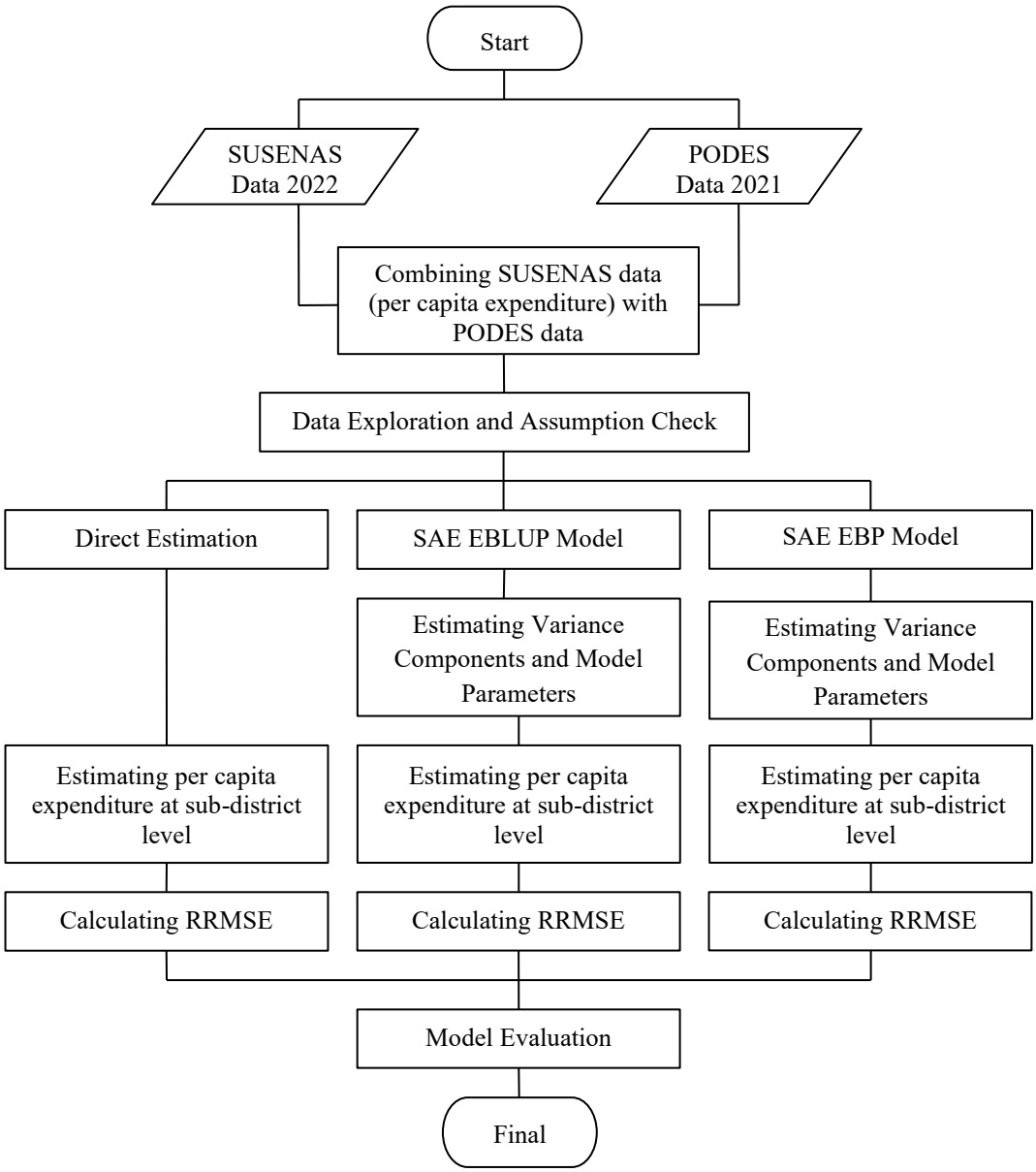


Figure 1. Research flow

RESULT AND DISCUSSIONS

Data exploration

The response variable used in this study is per capita expenditure, sourced from the 2022 SUSENAS data. It is calculated as the total monthly expenditure on food and non-food items divided by the number of household members. This variable serves as the unit-level response variable to be estimated. The supplementary variables are derived from the 2021 PODES data. Several of these variables are included based on the assumption that they have a statistical relationship with per capita expenditure.

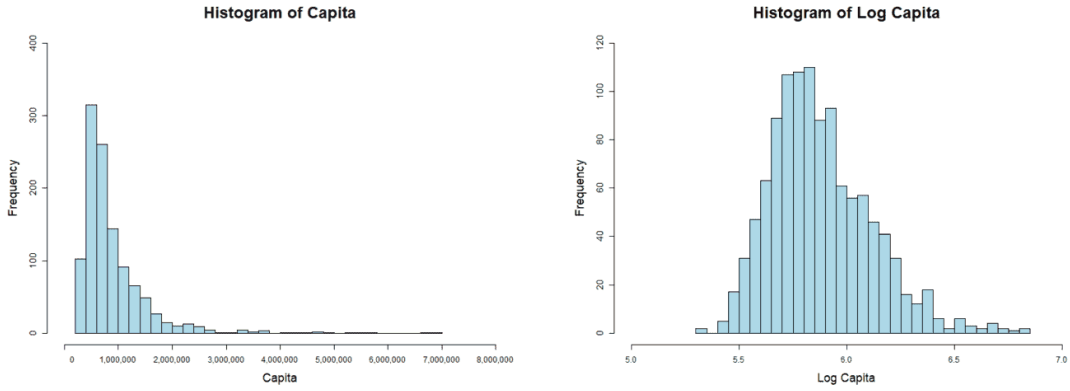
Table 2. Descriptive Statistics of the Response and Independent Variables

Variable	Minimum	Mean	Maximum	Standard Deviation
Per capita expenditure	213.904,8	898.02	6.838.91	675.298,8
Number of PLN user households	738	2.006,82	6.72	813,23
Number of Certificates of Inability	0	11,66	1.05	99,02
Number of households subscribed to landline phones	0	129,62	690	119,91
Number of village markets	1	135,31	375	70,02
Number of active integrated health service posts	0	5,567	6	1,45
Number of village cooperative units	0	1,178	5	0,75
Number of BUMDes business units	0	0,06	1	0,24
Number of beneficiary households	3	9,61	29	3,65
Number of health centers	0	0,54	2	0.55

Table 2 presents the descriptive statistics of the response and supplementary variables used in this study. Based on the 2022 SUSENAS sample data from Garut Regency, the average per capita expenditure at the subdistrict level is IDR 898,020. This value ranges from a minimum of IDR 213,904.80 to a maximum of IDR 6,838,910, with a standard deviation of IDR 675,298.80, indicating substantial variation in expenditure across subdistricts. Regarding sources of lighting, most households in Garut Regency use PLN electricity, with an average of 2,006.82 PLN user households per subdistrict. The number of PLN users ranges from 738 to 6,720 households, with a standard deviation of 813.23. Although smartphone use has become widespread in the digital era, some households in Garut Regency still subscribe to landline phones. This is evident from the average number of landline phone subscribers per subdistrict, which is 129.62, with a maximum of 690 and a standard deviation of 119.91.

In Garut Regency, many villages continue to issue Certificates of Inability (Surat Keterangan Tidak Mampu), with an average of 129,624 certificates issued. The highest number of Certificates of Inability issued was 690, with a standard deviation of 119,907. Meanwhile, the average number of households receiving Direct Cash Assistance (Bantuan Langsung Tunai) during the first three months in each subdistrict is 135,314. The number of recipients ranges from 1 to 375 households, with a standard deviation of 70,023.

In Garut Regency, the number of active village markets is limited, with an average of approximately 5.566 markets per village. The village with the highest number of markets has 6, and the data distribution is indicated by a standard deviation of 1.453. Meanwhile, the number of BUMDes business units per village is also low, with an average of 1.177 units. The village with the most BUMDes has 5 business units, and the standard deviation is 0.745. A similar situation occurs with village cooperative units (KUD), which are very few in number. The average number of KUDs per village is only 0.063, with a maximum of 1 KUD in a few villages and a data variation of 0.243. In terms of health facilities, each village in Garut Regency has an average of about 9.614 active integrated health service posts (posyandu). The number of integrated health service posts (posyandu) per village varies, ranging from a minimum of 3 to a maximum of 29, with a standard deviation of 3.647. In contrast, the number of health centers (puskesmas) per village is much lower than that of integrated health service posts (posyandu). This is reflected in the average number of puskesmas per village, which is only 0.541, with many villages lacking a puskesmas entirely. The village with the most puskesmas has only 2, with a standard deviation of 0.550.



(i) (ii)

Figure 2. (i) Per capita expenditure histogram; (ii) Log per capita expenditure histogram

Figure 2 (i) shows that the average per capita household expenditure has an asymmetric distribution, with a tendency to stretch to the right. This indicates that most households have low to moderate per capita expenditures, while a small portion has very high per capita expenditures. Therefore, the researcher assumes that applying a logarithmic transformation to the per capita expenditure variable will result in a normal distribution. As a result, the transformed per capita expenditure is expected to satisfy the assumptions of the small area estimation unit-level model. This can be seen in Figure 2 (ii), which shows that the logarithmically transformed per capita expenditure is assumed to follow a normal distribution. The average per capita expenditure without transformation is influenced by some large values (extremes or outliers). The extreme values for the supplementary variables can be seen in Figure 3, where the distribution of each variable is asymmetric. However, this study did not address these extreme values.

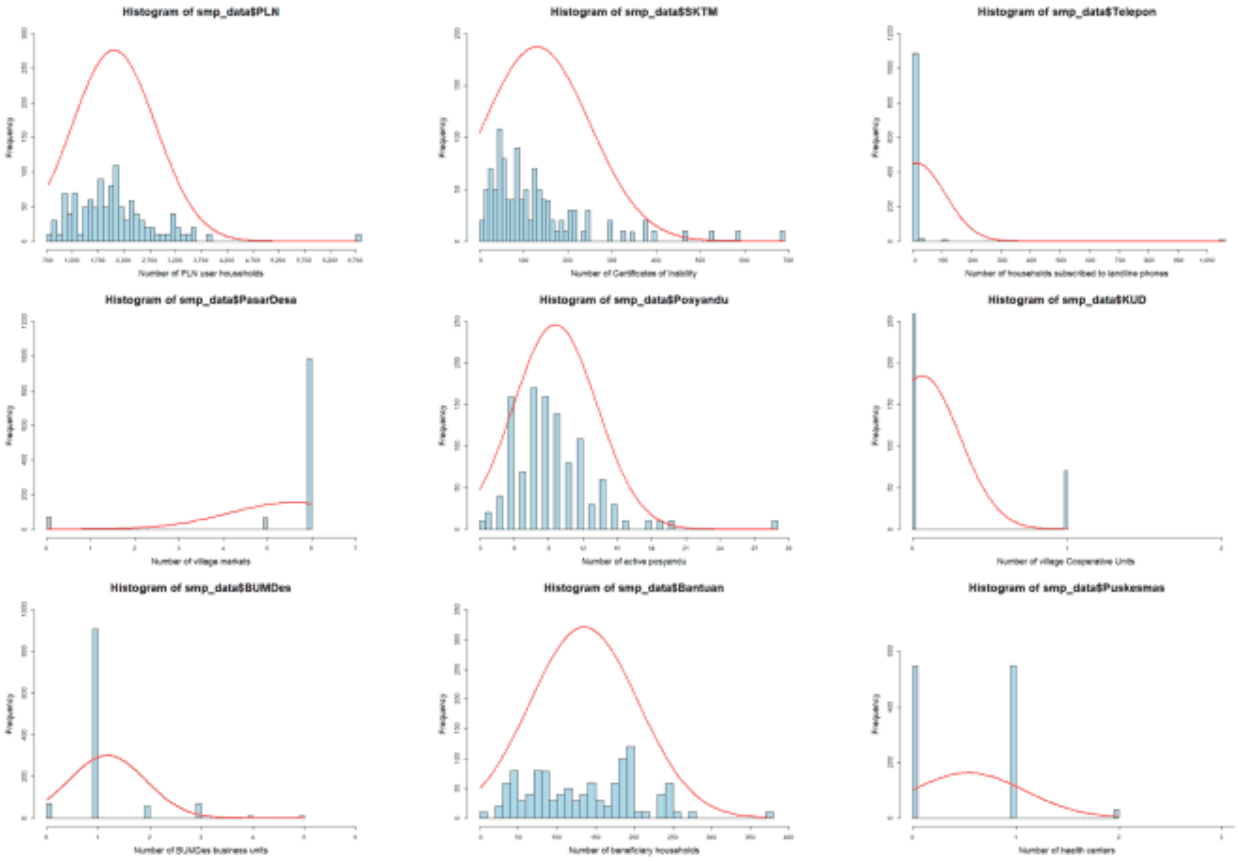


Figure 3. Distribution of independent variable

In small area estimation testing, there are several assumptions that must be met, one of which is the normality test. Normality checking is conducted using a Q–Q plot and the Kolmogorov–Smirnov test. Figure 4 (i) shows the distribution of residuals for per capita expenditure, and Figure 4 (ii) displays the results of the normality test on the log-transformed per capita expenditure, which shows that the residuals for per capita expenditure are asymmetric, while the residuals for the log-transformed per capita expenditure exhibit a symmetric pattern.

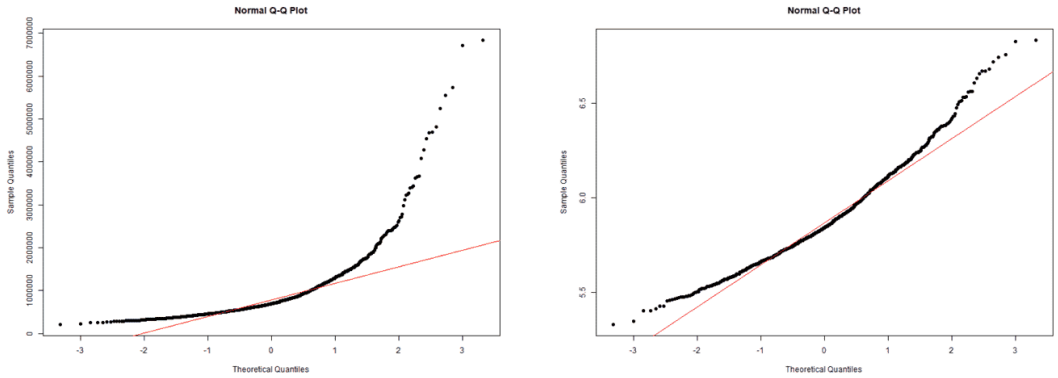


Figure 4. (i) Q-Q plot of per capita expenditure residuals distribution; (ii) Q-Q plot of residuals from log-transformed per capita expenditure distribution

Further testing required prior to small area estimation modeling involves examining correlations between variables. The results reveal a strong correlation (0.73) between the number of households using PLN electricity and the number of integrated health service posts (posyandu) in each village, indicating a substantial positive unidirectional relationship. This means that as the number of PLN subscriber households increases, the number of integrated health service posts (posyandu) in each village also tends to increase, and vice versa. When two or more variables in a regression model exhibit high correlation, it may potentially cause instability in regression coefficient estimates. Figure 5 illustrates the inter-variable correlations.

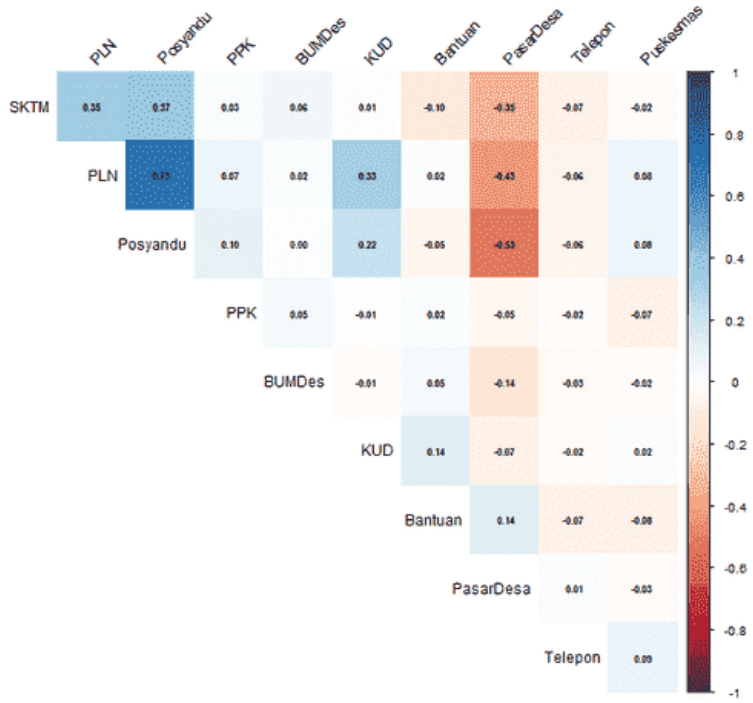


Figure 5. Correlation plot between variables

Multicollinearity check is one of the assumptions that must be fulfilled. Multicollinearity can be assessed from the Variance Inflation Factor (VIF) values of each independent variable. In Table 3, it can be seen that the calculated VIF values for the independent variables are all less than 10, indicating that no multicollinearity is detected among the independent variables.

Table 3. VIF Values for Each Covariate

Independent Variable	VIF
Number of PLN user households	2.3663
Number of Certificates of Inability	1.2504
Number of households subscribed to landline phones	1.5072
Number of village markets	1.0230
Number of active integrated health service posts (posyandu)	2.4817
Number of village Cooperative Units	1.1565
Number of BUMDes business units	1.0359
Number of beneficiary households	1.0693
Number of health centers	1.0248

Based on the exploratory data analysis of both the response and auxiliary variables, it was observed that the distribution of per capita expenditure exhibits right skewness, indicating a departure from symmetry. Nevertheless, other key model assumptions, such as the presence of correlation among variables and the absence of multicollinearity, are satisfactorily met. However, the deviation from normality in the response variable's distribution may pose a significant limitation for the application of standard SAE techniques, as these methods typically rely on the assumption of normally distributed sampling errors within the underlying model structure.

In the context of SAE, EBLUP is a widely applied method for socioeconomic data. However, the implementation of EBLUP faces limitations when applied to economic indicators, such as per capita expenditure, which empirically tend to exhibit right-skewed distributions. Under these conditions, the normality assumption required by EBLUP is violated, potentially leading to biased estimates. To address this issue, the EBP method has been proposed as a solution by incorporating a log-normal transformation. This approach enables the normalization of the data distribution prior to model estimation, while also including a bias correction procedure during the back-transformation to the original scale. Consequently, EBP provides more accurate estimates for data characterized by right-skewed distributions.

The SAE model employed for the EBP is a mixed model, combining fixed effects and random effects. In this study, the fixed effects are represented by the auxiliary variables included in the model, which aim to capture the relationship between the response variable and the covariates that is consistent across areas. The random effects, on the other hand, account for unobserved area-specific characteristics, such as those of subdistricts, that capture between-area variation not explained by the covariates. The EBP method includes a bias correction procedure to address distortions introduced by the transformation process. As a result, EBP yields more accurate estimates by accounting for the actual distribution of the data.

Table 4 presents the estimated average per capita expenditure at the subdistrict level, obtained using direct estimation, EBLUP, and EBP methods. As shown in the table, the direct estimation indicates that the RRMSE value in all sub-districts exceeds 25%, so the estimation results cannot be used. Meanwhile, in EBLUP and EBP, sub-districts with larger sample households tend to produce more stable estimates, while sub-districts with smaller sample sizes show greater variability. In some samples with limited households, the RRMSE exceeds 25%. Therefore, caution is advised when interpreting the estimation results for subdistricts with RRMSE values above 25%, particularly for Pamulihan, Sukaresmi, and Kersamanah. These results align with research conducted by Dian Handayani, where estimation results using EBP provide more stable results than direct estimation, which has a bias toward estimation results. Meanwhile, the EBLUP method cannot handle data that is skewed to the right because it violates the normality assumption in the response variable data.

The results of this study can help the Garut Regency government determine and evaluate socio-economic policies and determine which sub-districts will be given intervention first. This can be seen from the stability of the per capita expenditure estimation results, which are also strengthened by looking at the RRMSE value.

Table 4. Estimated Mean and RRMSE of Per Capita Expenditure by Subdistrict in Garut Regency Based on Direct Estimation, EBLUP, and EBP Methods (in thousand rupiah)

Area	N	n	Direct		EBLUP		EBP	
			Estimation	RRMSE	Estimation	RRMSE	Estimation	RRMSE
Cisewu	8365	30	827139.46	55.16	1135745.72	12.93	825511.69	13.51
Caringin	8065	30	894182.28	33.04	690124.65	11.43	958059.78	12.16
Talegong	7193	30	614133.25	35.66	703102.61	13.70	686987.33	19.67
Bungbulang	14588	30	1212835.26	37.41	621656.87	8.85	1172813.39	7.62
Mekarmukti	4608	20	563580.65	53.53	814566.52	18.47	599866.29	22.50
Pamulihan	4675	10	600025.55	30.89	1096475.38	21.90	719789.03	25.32
Pakenjeng	17133	40	611844.47	68.08	840735.85	14.28	604596.31	15.80
Cikelet	11213	30	1119325.83	60.95	751687.98	9.13	1032814.15	11.47
Pameungpeuk	10735	10	1278308.63	37.80	996742.15	14.07	1307561.18	16.10
Cibalong	11240	40	509598.28	27.47	660457.74	17.03	571586.64	16.83
Cisompet	13260	40	871574.67	49.48	611142.37	10.37	857239.57	12.61
Peundeuy	5973	20	710224.88	42.64	702955.72	15.59	800505.97	19.57
Singajaya	11875	30	773581.22	43.91	687940.20	9.49	769810.27	12.38
Cihurip	4693	10	622817.95	23.98	868575.17	21.63	677948.28	18.00
Cikajang	21135	50	1225649.25	80.05	797878.47	6.42	1047140.00	7.57
Banjarwangi	14488	30	578049.94	26.40	822289.77	17.76	664876.00	15.50
Cilawu	27130	49	1129344.24	111.72	675786.28	6.39	961061.64	9.66
Bayongbong	25925	20	604019.84	46.19	665192.47	16.38	635096.10	17.64
Cigedug	11008	10	790417.87	45.76	547019.02	16.97	795527.92	22.64
Cisurupan	24735	39	617260.33	62.01	1114458.99	14.45	610444.17	12.93
Sukaresmi	10033	10	1609503.70	70.86	667842.68	11.24	1336317.07	11.23
Samarang	19450	30	1293696.13	125.96	912035.15	9.01	884883.78	15.33
Pasirwangi	16308	40	1585577.84	67.93	1013222.32	6.25	1481418.78	6.03
Tarogong Kidul	28915	30	1612952.06	76.75	1092680.63	6.87	1305598.06	10.68
Tarogong Kaler	24275	30	955583.66	45.17	874490.33	11.91	987712.09	11.22
Garut Kota	31928	30	742438.46	56.99	649591.75	14.77	756239.39	13.82
Karangpawitan	34455	40	751269.55	56.13	637547.40	11.96	808247.31	12.16
Wanaraja	12180	30	1080701.22	58.36	878304.34	12.07	989103.59	9.05
Sucinaraja	7355	10	1210486.43	41.33	902122.24	12.52	1161215.08	14.38
Pangatikan	10603	10	999048.20	29.40	943498.57	14.33	1006640.86	13.08
Sukawening	14135	20	1030644.12	63.15	865640.06	15.19	1017123.91	12.45
Karangtengah	4653	20	584841.73	32.88	1351016.58	17.30	701760.74	21.32
Banyuresmi	22930	29	845703.82	52.92	860179.38	11.48	940113.30	9.72
Leles	21053	20	805985.24	55.36	1510417.99	16.83	762080.81	14.54
Leuwigoong	11728	20	700833.98	63.31	1461505.43	16.68	667334.21	22.04
Cibatu	18563	29	822184.46	69.62	1071331.66	12.23	839467.24	14.99
Kersamanah	9730	10	607500.70	37.72	1033878.15	25.24	663065.66	29.07
Cibiuk	8750	20	835731.98	45.76	1227602.37	14.31	847449.00	16.02
Kadungora	23205	20	637401.17	47.36	861093.70	20.01	700065.97	19.00
Bl. Limbangan	19705	30	839848.53	64.64	1115130.50	11.73	834215.82	10.90
Selaawi	10545	10	1261654.50	42.76	776473.03	12.07	1153076.39	14.75
Malangbong	32678	70	895137.39	40.58	966061.50	7.48	905547.16	7.95

CONCLUSION

This study estimates and tests the performance of the EBLUP and EBP methods, assuming that the logarithm-transformed variables follow a normal distribution within the framework of a small area estimation model at the unit level. Applying the EBP method to per capita expenditure data at the sub-district level in the Garut Regency provides satisfactory results, where the performance results of EBP provide more stable results compared to direct estimation and EBLUP. However, several sub-districts still show RRMSE values greater than 25%, especially in sub-districts with small samples, indicating areas that may require further refinement. It is estimated that other researchers can conduct similar studies by adding other methods or increasing the number of samples to individuals and not households so that it is possible to obtain better results.

REFERENCES

[1] Food and Agriculture Organization of United Nations, "Small Area Estimation with Unit Level Models," *Sustainable Development Goals*, 2022. https://openknowledge.fao.org/server/api/core/bitstreams/b0a12b78-7263-45ab-931d-f27f772057b0/content

[2] P. A. Parker, R. Janicki, and S. H. Holan, "Comparison of Unit-Level Small Area Estimation Modeling Approaches for Survey Data Under Informative Sampling," *J. Surv. Stat. Methodol.*, vol. 11, no. 4, pp. 858–872, Sep. 2023, doi: 10.1093/jssam/smad022.

[3] V. M. Santi, K. A. Notodiputro, I. Indahwati, and B. Sartono, "RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION FOR MULTIVARIATE LINEAR MIXED MODEL IN ANALYZING PISA DATA FOR INDONESIAN STUDENTS," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 2, pp. 607–614, Jun. 2022, doi: 10.30598/barekengvol16iss2pp607-614.

[4] R. J. N. K. and I. Molina, *Introduction to Small Area Estimation Techniques*, 2nd ed. Hoboken: John Wiley & Sons, Inc., 2020.

[5] L. Maestrini, F. K. C. Hui, and A. H. Welsh, "Restricted maximum likelihood estimation in generalized linear mixed models," *arXiv*, no. 1962, 2025, [Online]. Available: http://arxiv.org/abs/2402.12719

[6] S. Sugasawa and T. Kubokawa, "Small area estimation with mixed models: a review," *Japanese J. Stat. Data Sci.*, vol. 3, no. 2, pp. 693–720, Dec. 2020, doi: 10.1007/s42081-020-00076-x.

[7] Z. Lyu and A. H. Welsh, "Small Area Estimation using EBLUPs under the Nested Error Regression Model," *ArXiV*, pp. 1–35, 2022, [Online]. Available: https://arxiv.org/abs/2210.09502

[8] R. Anisa, A. Kurnia, and I. Indahwati, "Cluster Information of Non-Sampled Area In Small Area Estimation," *IOSR J. Math.*, vol. 10, no. 1, pp. 15–19, 2014, doi: 10.9790/5728-10121519.

[9] L. Mori and M. R. Ferrante, "Small Area Estimation of Household Economic Indicators under Unit-Level Generalized Additive Models for Location, Scale and Shape," *J. Surv. Stat. Methodol.*, vol. 13, no. 1, pp. 160–196, Feb. 2025, doi: 10.1093/jssam/smae038.

[10] Z. Lyu and A. H. Welsh, "Asymptotics for EBLUPs: Nested Error Regression Models," *J. Am. Stat. Assoc.*, vol. 117, no. 540, pp. 2028–2042, Oct. 2022, doi: 10.1080/01621459.2021.1895178.

[11] Z. Lyu and A. H. Welsh, "Increasing Cluster Size Asymptotics for Nested Error Regression Models," *ArXiv*, 2021, [Online]. Available: https://arxiv.org/abs/2101.08951

[12] M. Viljanen, L. Meijerink, L. Zwakhals, and J. van de Kassteele, "A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands," *Int. J. Health Geogr.*, vol. 21, no. 1, p. 4, Dec. 2022, doi: 10.1186/s12942-022-00304-5.

[13] I. Molina, N. Salvati, and M. Pratesi, "Bootstrap for estimating the MSE of the Spatial EBLUP," *Comput. Stat.*, vol. 24, no. 3, pp. 441–458, Aug. 2009, doi: 10.1007/s00180-008-0138-4.

[14] I. Molina and Y. Marhuenda, "SmallAreaEstimation: Package 'sae.'" 2020. [Online]. Available: https://cran.r-project.org/web/packages/sae/sae.pdf

[15] P. C. Rodas, I. Molina, and M. Nguyen, "Pull your small area estimates up by the bootstraps," *J. Stat. Comput. Simul.*, vol. 91, no. 16, pp. 3304–3357, Nov. 2021, doi: 10.1080/00949655.2021.1926460.

[16] J. Pinheiro and D. Bates, "Linear and Nonlinear Mixed Effects Models: Package 'nlme.'" 2025. [Online]. Available: https://cran.r-project.org/web/packages/nlme/nlme.pdf

[17] I. Molina and Y. Marhuenda, "Sae: An R package for small area estimation," *R J.*, vol. 7, no. 1, pp. 81–98, 2015, doi: 10.32614/rj-2015-007.

[18] H. Li, Y. Liu, and R. Zhang, "Small area estimation under transformed nested-error regression models," *Stat. Pap.*, vol. 60, no. 4, pp. 1397–1418, Aug. 2019, doi: 10.1007/s00362-017-0879-7.

[19] J. Breidenbach and R. Astrup, "Small area estimation of forest attributes in the Norwegian National Forest Inventory," *Eur. J. For. Res.*, vol. 131, no. 4, pp. 1255–1267, Jul. 2012, doi: 10.1007/s10342-012-0596-7.

[20] P. A. Parker, R. Janicki, and S. H. Holan, "A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling," *J. Surv. Stat. Methodol.*, vol. 11, no. 4, pp. 829–857, Sep. 2023, doi: 10.1093/jssam/smad020.

[21] N. Würz, "The R Package saeTrafo for Estimating unit-level Small Area Models under Transformations," no. September, 2023, [Online]. Available: https://www.researchgate.net/publication/374004978

[22] P. Corral and S. S. Juarez, "Small Area Estimation: Area Level Model," *World Bank*, 2022. https://documents1.worldbank.org/curated/en/099060524110616197/pdf/P1798581967c7e0919dc

0144d06fcfcab3.pdf
[23] N. Hasanah, K. A. Notodiputro, and B. Sartono, "Kajian Simulasi dan Empiris: Kinerja Model Copula dan Regresi Galat Tersarang dalam Menduga Pengeluaran per Kapita Kecamatan di Kabupaten Pidie," 2023. [Online]. Available: http://repository.ipb.ac.id/handle/123456789/124345
[24] T. Yuniarty, Indahwati, and A. H. Wigena, "Pendugaan Area Kecil dengan Log-normal HB dan Skew-normal HB untuk Pengeluaran Per Kapita RTP Tanaman Pangan di Provinsi Sulawesi Tenggara," 2024. [Online]. Available: http://repository.ipb.ac.id/handle/123456789/137376
[25] D. Handayani, K. A. Notodiputro, A. Saefuddin, I. W. Mangku, and A. Kurnia, "Spatial Empirical Best Predictor of Small Area Poverty Indicator," *Int. J. Adv. Soft Comput. its Appl.*, vol. 16, no. 2, pp. 103–122, 2024, doi: 10.15849/IJASCA.240730.07.
[26] E. Berg, "Empirical Best Prediction of Small Area Means Based on a Unit-Level Gamma-Poisson Model," *J. Surv. Stat. Methodol.*, vol. 11, no. 4, pp. 873–894, Sep. 2023, doi: 10.1093/jssam/smac026.
[27] M. Guadarrama, D. Morales, and I. Molina, "Time stable empirical best predictors under a unit-level model," *Comput. Stat. Data Anal.*, vol. 160, p. 107226, Aug. 2021, doi: 10.1016/j.csda.2021.107226.
[28] D. Handayani, K. A. Notodiputro, A. Saefuddin, I. Wayan Mangku, and A. Kurnia, "Empirical Best Predictor for Nested Error Regression Small Area Models," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 187, p. 012036, Nov. 2018, doi: 10.1088/1755-1315/187/1/012036.
[29] I. Molina and N. Martín, "Empirical best prediction under a nested error model with log transformation," *Ann. Stat.*, vol. 46, no. 5, Oct. 2018, doi: 10.1214/17-AOS1608.
[30] T. Hobza and D. Morales, "Empirical Best Prediction Under Unit-Level Logit Mixed Models," *J. Off. Stat.*, vol. 32, no. 3, pp. 661–692, Sep. 2016, doi: 10.1515/jos-2016-0034.
[31] M. Guadarrama, I. Molina, and Y. Tille, "Small area estimation methods under cut-off sampling," 2019. [Online]. Available: https://liser.elsevierpure.com/ws/portalfiles/portal/11419510/WP_n_2019_01.pdf
[32] A. Chwila and T. Żądło, "On properties of empirical best predictors," *Commun. Stat. - Simul. Comput.*, vol. 51, no. 1, pp. 220–253, Jan. 2022, doi: 10.1080/03610918.2019.1649422.