International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 830~846
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1621

Performance Evaluation of the XGBoost Method for Measuring Uncertainty in Small Area Estimation

Fida Fariha Amatullah[1], Khairil Anwar Notodiputro[2], Anwar Fitrianto[3]
[1,2,3]Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:
Received May 18, 2026
Revised July 11, 2026
Accepted July 18, 2026

Keywords:
EBLUP-FH
Bootstrap
Small Area Estimation
Winsorization
XGBoost

ABSTRACT

This study looks at using Small Area Estimation (SAE) to assess per capita spending in subdistricts of Jambi Province. Direct estimators in small areas usually show high variance because of small sample sizes, so model-based methods are needed. The Fay–Herriot model based on EBLUP–FH was used as the primary indirect estimation method, while XGBoost served as a flexible machine learning benchmark. These approaches show fundamentally different methods. EBLUP-FH acts as a model-based estimator with a clear sampling-error framework, while XGBoost functions as a prediction model. Its uncertainty is evaluated through bootstrap resampling, so their standard errors can't be directly compared. Before modeling, data was cleaned and winsorized to lessen outlier impact. The normality assumption for the sampling error in EBLUP-FH was breached, and a logarithmic transformation was used which improved the model components' distribution. XGBoost achieved the smallest bootstrap-based standard error, indicating reduced variability in predictions across resamples rather than greater estimation accuracy. Subdistrict-level analysis revealed that XGBoost tends to produce homogeneous estimates that fail to capture contextual socioeconomic variation, confirming that a small bootstrap standard error does not necessarily indicate accurate small-area estimates.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Fida Fariha Amatullah,
Statistics and Data Science,
IPB University,
Darmaga Street, Babakan, Dramaga, Bogor, West Java, Indonesia.
Email: fidafarihaafida@apps.ipb.ac.id

1. INTRODUCTION

Sample surveys are an efficient and cost-effective method for obtaining information about a population, including its various subpopulations, or what are known as domains. However, not all domains have an adequate sample size. When the sample size for a domain is too small, the resulting direct estimates tend to be imprecise and have low accuracy [1], [2]. This condition is the background for the development of Small Area Estimation (SAE), a statistical approach that utilizes additional information from other variables to produce more reliable estimates in domains with sample limitations [1], [3], [4].

One of the most commonly used models at the SAE area level is the Fay-Herriot model. This model is widely chosen because of its relatively simple structure and ease of application [5]. However, the Fay-Herriot model requires data to meet the assumptions of normality and linearity, which are often not met in real data, especially when the data contains extreme values or the relationship between variables is non-linear [1], [6]. Efforts to mitigate these shortcomings have included the use of the partially linear mixed-t model, which integrates the EM algorithm with B-spline functions to account for non-linearity [6]. Still, the model's flexibility remains constrained,

as it considers only one explanatory variable, limiting the non-linear effects that can be incorporated.

The constraints of traditional models have driven the use of other options, such as machine learning, which lacks distribution assumptions and can better account for complex interdependencies among variables [7]. Multiple studies report that methods based on gradient boosting perform better than traditional methods for SAE [8], [9]. Currently, XGBoost is one of the most popular implementations of gradient boosting. It optimizes the treatment of missing data and can account for non-linearities and complex interactions among variables without the need for user-defined processes [10], [11]. Several studies, including one by Viljanen et al. 2022, show that XGBoost significantly outperforms random forests, SVMs, and MLPs, in terms of prediction accuracy [12], [13]. In Viljanen et al.'s study on SAE, XGBoost was able to produce better accuracy with a significantly shorter computation time compared to the STAR model. However, the authors of the STAR model only used mean squared error (MSE) in their evaluation, indicating that the description of XGBoost's performance is still rather limited.

Whatever the merits of the chosen method, why the quality of estimates in SAE is affected by two additional properties: sample size and the presence of multicollinearity among the covariates. A small sample size results in high model bias and consequently high estimation variance [14]. In the presence of multicollinearity among the covariates, the estimates may become unstable, which is similar to the aforementioned case [15]. Thus, the uncertainty related to the estimates is an important concern. Hence, one of the methods to assess the uncertainty of the estimates is the bootstrap method, which is a simple resampling method and not based on particular distributional assumptions, and has proven to be effective in the context of estimating the uncertainty of statistical estimates [16].

These problems grow complicated when focusing on secondary data in Indonesia. Secondary data in Indonesia do not escape many challenges due to various reasons, including the infrequency in the updating of data, the fragmentation of data among many institutions, and the mistakes made in the collection and recording of data [17]. One relevant, and to some extent, critical, variable for further consideration is per capita expenditure as the monthly total of the expenditure on consumption by members of a household after adjustment for purchasing power parity [18]. This parameter adjusts for the economic ability of a household and for the setting of the poverty line and social protection program target beneficiaries. Thus, it warrants the greatest attention when attempting to capture it. Per capita spending data, though, is highly variable across zones in spatio-temporal dimensions, and many of the socio-economic variables/factors are often inextricably linked and therefore, in many cases, complexity cannot be captured by conventional approaches.

Given the previous discussion, this study applies XGBoost as a flexible, assumption-free prediction model in the SAE context specifically as a benchmark prediction model for per capita expenditure, not as a substitute for the Fay–Herriot linking model. The EBLUP–FH method serves as the primary model-based indirect estimator, with XGBoost included to evaluate whether a machine learning approach can serve as a complementary prediction tool. It is important to mention that the two methods work under very different systems. EBLUP-FH gives estimates based on explicit design-based sampling theory with a clear sampling-variance structure. XGBoost, on the other hand, is a prediction model based on data, and its uncertainty is evaluated empirically through bootstrap resampling. Since these methods generate standard errors in distinct ways, their standard errors represent different types of uncertainty and should not be seen as directly comparable. MSE and MAPE assess prediction performance compared to direct estimates, while bootstrap resampling measures estimation uncertainty for XGBoost. The study also examines the impact of sample size and multicollinearity among covariates, aiming to provide a more complete picture to guide more precise and effective policy formulation.

2. RESEARCH METHOD

Data

The data used in this study are secondary data sourced from the Jambi Provincial Central Statistics Agency, specifically the 2021 National Socioeconomic Survey (SUSENAS) data as the

response variable and the 2021 Village Potential Survey (PODES) data as the explanatory variable. The explanatory variables used are based on the explanatory variables found in [19].

Table 1. Variables used

Variable	Description	Data Types
$\mathbf{Y}$	Average Per Capita Expenditure by Subdistrict in Jambi Province (IDR 000)	Continuous
$\mathbf{X_1}$	Percentage of Households Living in Slums	Continuous
$\mathbf{X_2}$	Percentage of Households Using PLN Electricity	Continuous
$\mathbf{X_3}$	Percentage of Households Using Non-PLN Electricity	Continuous
$\mathbf{X_4}$	Ratio of Households with Educational Facilities	Continuous
$\mathbf{X_5}$	Ratio of Households with Health Facilities	Continuous
$\mathbf{X_6}$	Percentage of Households Using Landline Telephones	Continuous
$\mathbf{X_7}$	Ratio of Households with Microenterprises	Continuous
$\mathbf{X_8}$	Ratio of Households with Economic Facilities	Continuous
$\mathbf{X_9}$	Ratio of Households to the Number of Cell Phone Towers or Base Transceiver Stations	Continuous
$\mathbf{X_{10}}$	Percentage of Households Living in Settlements Under SUTET/SUTT/SUTTAS	Continuous
$\mathbf{X_{11}}$	Percentage of Households Located in Settlements Along Riverbanks	Continuous
$\mathbf{X_{12}}$	Ratio of Households to Community Organizations	Continuous
$\mathbf{X_{13}}$	Percentage of Households Receiving Direct Cash Assistance in the First 3 Months of 2020	Continuous
$\mathbf{X_{14}}$	Ratio of Households to BUMDes Business Units	Continuous
$\mathbf{X_{15}}$	Ratio of Households with Active Cooperatives	Continuous
$\mathbf{X_{16}}$	Ratio of Households with Operating Financial Institutions	Continuous
$\mathbf{X_{17}}$	Ratio of Households with Places of Worship	Continuous
$\mathbf{X_{18}}$	Ratio of Households with UKBM	Continuous
$\mathbf{X_{19}}$	Average Value of Direct Cash Assistance Received by Each Household (first 3 months)	Continuous

Direct Estimation

Parameter estimation using direct estimators is based solely on sample data obtained from the area in question. Direct estimators use values of the variable of interest only for the time period and sample units within the area of interest [1]. The direct estimation in this study is the per capita expenditure in the i-th subdistrict, calculated as the weighted average of household per capita expenditure, with weights equal to the number of household members. Mathematically:

$$\hat{\mu}_i^D = \frac{\sum_{j=1}^{n_i} (y_{ij} \times a_{ij})}{\sum_{j=1}^{n_i} a_{ij}} \quad (1)$$

y_{ij}: per capita household expenditure of household j in subdistrict i

a_{ij}: number of members in household j in subdistrict i

n_i: number of sample households in subdistrict i.

Empirical Best Linear Unbiased Prediction Fay-Herriot (EBLUP-FH)

Indirect estimation in SAE uses the EBLUP method with the Fay–Herriot model, which is an area-level model that operates at the small-area level. This model consists of two main components: the sampling model and the linking model [1]. The sampling model states that the direct estimator y_i contains a sampling error e_i, while the linking model relates the true parameters to the explanatory variables via linear regression with area random effects. The combination of the two forms the model:

$$y_i = X_i'\beta + v_i + e_i \quad (2)$$

The objective of this model is to optimally estimate the area parameters θ_i by combining information from survey data and the regression model. The fixed-effect parameters β are estimated using Generalized Least Squares (GLS):

$$\left(\hat{\beta}\right) = (X'V^{-1}X)^{-1}X'V^{-1}y \quad (3)$$

with the total covariance matrix $V = \sigma_v^2 + \sigma_{e_i}^2$. Next, the variance of the random effect σ_v^2 is estimated using Restricted Maximum Likelihood (REML), which is more stable for small samples. The area random effect is then estimated using:

$$\hat{v}_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{\bar{\varepsilon}_i}^2}\left(y_i - X_i^1\hat{\beta}\right) \quad (4)$$

and is used to obtain the EBLUP estimator:

$$\hat{\theta}_i^{EBLUP} = X_i'\hat{\beta} + \hat{v}_i \quad (5)$$

Equivalently, EBLUP is a weighted combination of direct estimators and model-based estimators:

$$\hat{\theta}_i^{EBLUP} = \gamma_i y_i + (1 - \gamma_i) x_i^T \hat{\beta} \quad (6)$$

where $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{e_i}^2}$. These weights are adaptive; if the information from the direct data is strong, the estimate relies more on y_i, whereas if it is weak, it relies more on the model [20]. Thus, EBLUP produces more stable and efficient estimates because it optimally combines the two sources of information to minimize error [1].

eXtreme Gradient Boosting (XGBoost)

XGBoost implements a specific form of Gradient Boosting. First, Chen and Guestrin's framework are cited for their efficiency and flexibility. Chen and Guestrin's framework also allows for first and second order gradients, implementation of regularization, and sampling, which contributes to controlling against overfitting [21]. Within Social Automation Engineering (SAE), XGBoost can be implemented for non-linear data with a larger caliber of predictors. Generally, XGBoost predictions are obtained by aggregating multiple decision trees [22]:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in \mathcal{F} \quad (7)$$

The model learning is structured through minimizing a loss function. Commonly, the loss function of choice is the mean squared error:

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (8)$$

with each iteration, the learning process adjusts with new residuals:

$$r_i = y_i - \hat{y}_i \quad (9)$$

Predictive results are updated with the inclusion of a new model:

$$\hat{y}_i^{\text{new}} = \hat{y}_i^{\text{old}} + \eta f_k(X_i) \quad (10)$$

where η signifies the learning rate, thus, determines the new model's contribution. This learning process endures as defined by iterations, and stops when the model loss reaches the minimum:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum w_j^2 \quad (11)$$

Regularization functions control the complexity of the model. Within XGBoost, regularization can be optimized through the following function of; T, which is the number of leaves, w_j stands for the weight at leaf j, γ is the penalty term for the number of leaves, which regulates the number of trees within the model, and λ, which is the L2 regularization on the leaf weights that is used to regulate complexity within the model.

Analysis Procedure

The analysis was performed in R on the Kaggle Notebook platform. The steps of the analysis include the following:

1. Non-recurring expenditure items (e.g., vehicle purchases, wedding expenditures, and incidental expenditures) are adjusted for or subtracted in order to yield a consumption measure that is more stable and representative.
2. Data exploration was conducted to determine the number of household samples in each subdistrict in Jambi Province and to identify the distribution and presence of outliers in the per capita expenditure data.
3. Any extreme values were treated with Winsorization at the 5th and 95th percentiles so that no observations were removed [23].
4. For direct estimation, weighted averages of per capita expenditure for each subdistrict were obtained, as shown in equation 1. The measure of uncertainty for the direct estimate is derived from the sampling variance of each area, which is expressed as $D_i = \frac{S_i^2}{n_i}$. This estimate is used to determine the standard error of the direct estimate.
5. The estimation results were analyzed to evaluate the data distribution and the degree of skewness. Assessment of the initial regression assumptions including linearity and multicollinearity of the explanatory variables were also conducted.
6. Indirect estimation was performed using the EBLUP-FH method through a series of steps, namely calculating the area variance ψ_i, applying a log transformation to the response variables if they were not normally distributed, selecting variables to avoid multicollinearity, building the

model using the eblupFH function, and estimating the components $\hat{\beta}$ dan $\hat{\sigma}_u^2$. Assumptions of normality and independence among components were then evaluated to ensure the model's validity. The measure of uncertainty for the EBLUP-FH method is obtained from the MSE value calculated using the mseFH function in R, which is based on the Prasad-Rao approach as described in [1]. The square root of this MSE value is used as the standard error of the EBLUP-FH estimate.

7. Indirect estimation was also performed using the XGBoost method, following these steps. The data, at the subdistrict level (n = 141 subdistricts), was split into 80% training data (approximately 113 subdistricts) and 20% test data (approximately 28 subdistricts) [24], then converted into the xgb.DMatrix format. Hyperparameter tuning was performed using Bayesian optimization, yielding the following optimal parameters: learning rate (eta) = 0.05, maximum tree depth (max_depth) = 4, subsampling ratio (subsample) = 0.8, column subsampling (colsample_bytree) = 0.8, and minimum child weight (min_child_weight) = 3. The model was then trained using the xgb.train function with a maximum of 300 rounds and early stopping after 20 consecutive rounds without improvement. Model evaluation was performed by calculating the MSE and MAPE of the difference between the direct estimate and the XGBoost estimate, for both the training and test data. Direct estimates were used as the reference for the XGBoost evaluation. This practice was done at the owner's risk and with utmost caution. Direct estimates of small areas are imprecise and therefore, the computed MSE and MAPE values of direct estimates demonstrate the degree of error of two estimators rather than the true prediction error of the small area estimate of the unknown parameter. This limitation is acknowledged when interpreting evaluation results in Section 3.5. To assess the model's consistency, the bootstrap approach was applied to the residuals (the difference between the direct prediction and the XGBoost prediction) over 100 iterations. The resulting bootstrap residual distributions were then compared between the training and test data to evaluate whether the XGBoost model consistently learns data patterns without overfitting or underfitting.
8. Model evaluation was performed using MSE (Mean Squared Error), which is the average of the squared differences between the estimates and the reference values, and MAPE (Mean Absolute Percentage Error), with the following accuracy thresholds: < 10%: Highly Accurate Prediction; 10 – 20%: Good Prediction; 20 – 50%: Reasonable Prediction; >50%: Inaccurate Prediction [25].
9. Uncertainty estimates were performed for all three methods to obtain standard error values that could then be compared with one another. For direct estimation, the standard error is obtained from the square root of the sampling variance for each subdistrict, namely , which was calculated during the direct estimation stage. For the EBLUP-FH method, the standard error is obtained from the square root of the MSE value calculated using the Prasad-Rao approach via the mseFH function in R [1]. If the EBLUP-FH model is built on a log scale, the MSE value needs to be transformed back to the original scale using the Delta method, so that the EBLUP-FH standard error is obtained via , where is the EBLUP-FH estimate on the log scale. Unlike direct estimation and EBLUP-FH, which employ analytical approaches to measure uncertainty, XGBoost lacks a built-in analytical mechanism for this purpose. Therefore, uncertainty in the XGBoost method is measured using a bootstrap approach in which subdistrict-level observations are resampled with replacement for B = 100 iterations. The bootstrap procedure is carried out in the following steps:$\widehat{SE}_{Direct} = \sqrt{\hat{V}(\bar{y}_i)} = \sqrt{s_i^2/n_i}\widehat{SE}_{EBLUP} = \sqrt{MSE_i \times (\exp(\hat{\mu}_i))^2}\hat{\mu}_i$
   - First, the subdistrict-level training data is randomly resampled with replacement B = 100 times to generate bootstrap samples.
   - Second, the XGBoost model is retrained on each bootstrap sample using the optimal hyperparameters obtained previously.
   - Third, each bootstrap model is used to predict per capita expenditure across all target subdistricts, and the prediction results are then averaged by subdistrict. This provides the

consumption measure along with the $n \times B$ prediction matrix where n is the number of subdistricts.

- The standard error of XGBoost for subdistrict i is the standard deviation of the bootstrap prediction results, namely $\widehat{SE}_{XGBoost,i} = SD(\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(B)})$. This approach reflects the degree of variation in XGBoost estimates when the model is trained on different data samples, thereby providing an empirical illustration of XGBoost's ability to capture data variation.

10. The standard errors for each method, namely , , and , are presented and compared below as measures of uncertainty to assess the stability and reliability of each method in producing per capita expenditure estimates.$\widehat{SE}_{Direct}\widehat{SE}_{EBLUP}\widehat{SE}_{XGBoost}$
11. Conclusions were drawn based on a comparison of the performance and stability of the two methods in producing estimates for small areas, as well as the level of uncertainty derived from the bootstrap results.

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics

The population of Jambi Province in 2021 reached 3.5 million, with 8.09% of residents having an average monthly per capita expenditure below the poverty line. Average monthly per capita expenditure in Jambi Province is divided into two categories: food at Rp. 613,753 and non-food at Rp. 566,507 [26].

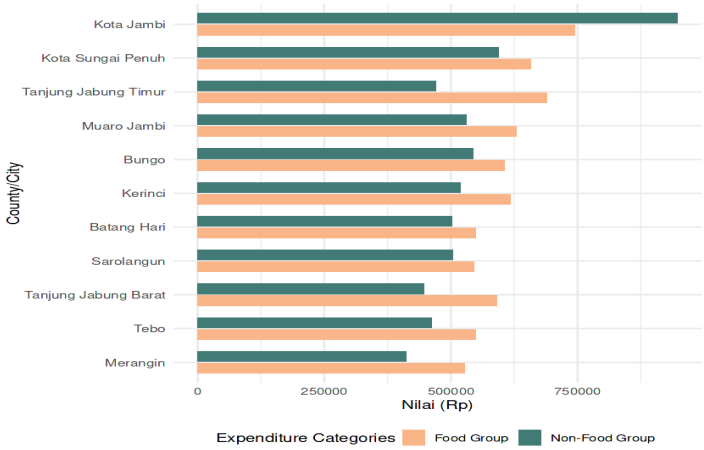


Figure 1. Average per capita expenditure in Jambi Province
Source: BPS Jambi Province

The figure shows the average monthly per capita expenditure in the regencies/cities of Jambi Province in 2021. Average per capita expenditure is dominated by the food category; however, in Jambi City, a larger portion of expenditure is allocated to the non-food category. This difference in spending patterns reflects the socioeconomic characteristics between urban and rural areas. In urban areas such as Jambi City, relatively higher household income, higher non-food living costs (including rent, education, transportation, communication, and recreation), and a more modern lifestyle cause the proportion of expenditure on the non-food category to be more dominant compared to the food category.

Per capita expenditure not only reflects consumption patterns but also serves as the basis for determining the poverty line and identifying the poor. Therefore, the differences in expenditure patterns across regions, as described earlier, are closely linked to poverty conditions in each area. Table 1 shows the percentage of the poor population in each regency and city in Jambi Province in 2021.

Table 2. Percentage of the Poor Population by Regency/City in Jambi Province in 2021

City/Regency	Poor people (%)
Kerinci	7.71
Merangin	9.11
Sarolangun	8.87
Batang Hari	10.05

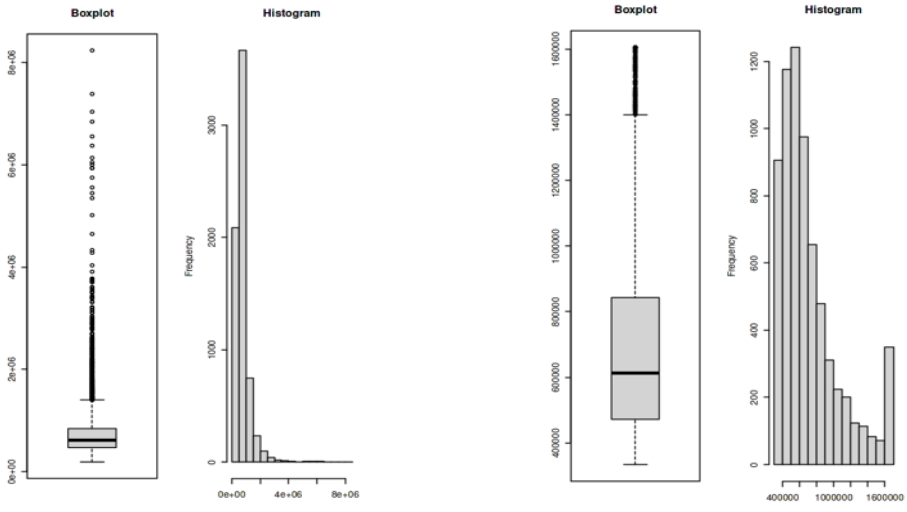
City/Regency	Poor people (%)
Muaro Jambi	4.53
Tanjung Jabung Timur	11.39
Tanjung Jabung Barat	10.75
Tebo	6.68
Bungo	6.23
Kota Jambi	9.02
Kota Sungai Penuh	3.41

In general, poverty rates vary significantly across regions in Jambi Province, ranging from 3.41% to 11.39%. Sungai Penuh City recorded the lowest percentage of the poor population at 3.41%, followed by Muaro Jambi Regency at 4.53%. Conversely, Tanjung Jabung Timur Regency had the highest percentage of the poor population at 11.39%, followed by Tanjung Jabung Barat Regency at 10.75% and Batang Hari Regency at 10.05%.

When analyzing urban and rural areas, some interesting trends can observe. From the results, the urban areas in Jambi, namely, the City of Jambi and the City of Sungai Penuh, appear to have lower poverty rates as compared to the majority of Jambi's regencies, which include urban areas of relatively high economic development. This is also reflected in the urban areas' spending data. In contrast, the eastern coastal districts of Jambi, particularly East and West Tanjung Jabung, which are rich in natural resources, experience remarkably high poverty rates. The aforementioned situation exemplifies the importance of estimating per capita expenditures at the subdistrict level to allow for better planning and targeting of poverty reduction programs in order to address the glaring disparities in poverty and welfare among the various regions of Jambi.

3.2. Household Data Preprocessing and Exploration

The response variable includes household food and non-food expenditures obtained from the 2021 National Socio-Economic Survey (SUSENAS). The data first underwent preprocessing by removing non-routine expenditures, converting expenditures from annual to monthly, and calculating the monthly expenditures of each member of the household. Using this method, 6,913 households from Jambi Province were obtained. The household expenditure data is visualized in Figure 2 with additional preprocessing.



(a) (b)

Figure 2. Boxplot and histogram of household expenditure per capita before winsorization (a) and after winsorization (b)

In the early stages of preprocessing this data, the distribution of per capita household expenditure is right-skewed and has high outlier values. The presence of outliers is likely to threaten the stability of estimations, especially in dealing with inferential analyses like SAE.

To mitigate the issue of outliers, a winsorization technique which sets the upper 95th percentile to the value of the 95th percentile, and the lower 5th percentile to the value of the 5th

percentile, was adopted. The effect of winsorization was that the maximum value decreased to Rp1,605,183, but the median and other quartile values did not change.

Table 3. Descriptive statistics on per capita household expenditures

Statistics	Before Winsor	After Winsor
Minimum	191.133	334.438
Q1	472.258	472.258
Median	613.248	613.248
Mean	748.409	712.340
Q3	843.188	843.188
Maximum	8.232.886	1.605.183
Std. Deviation	514.460,3	336.114,5

As a result, the average per capita expenditure decreased due to winsorization from Rp748,409 to Rp712,340 (approximately a 4.8% decrease), and the standard deviation decreased from Rp514,460.3 to Rp336,114.5 (approximately a 34.6% decrease). The large decrease in the standard deviation indicates the variance in the data has become more stable. The small increase in the mean of the winsorized data indicates the distribution of the data has retained its overall shape. Thus, the winsorization process produces more robust data.

3.3. Analysis of Direct Estimation Results

The quality of the direct estimates is strongly influenced by the sample size in each subdistrict. Figure 3 shows how the number of sample households impacts the variance of the direct estimates for each subdistrict. Subdistricts with sample households fewer than 30 tend to show larger, more variable direct estimates. Some of these subdistricts even have extremely high variance close to 1.2×10^{10}. The more sample households a subdistrict has, the less variable the direct estimate is likely to be. This shows that, for a number of subdistricts, a small sample size results in direct estimates that are more imprecise [27].

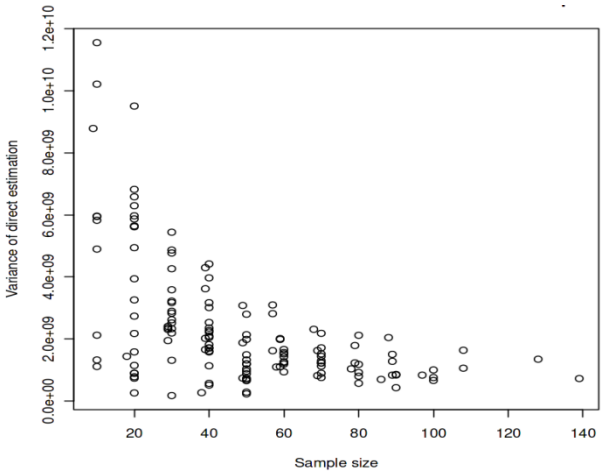


Figure 3. Correlation between sample size and direct estimation

The direct estimation results of per capita expenditure by subdistrict show that this data is more evenly distributed compared to the data that was used earlier. Figure 4 shows the direct per capita expenditure estimates of a number of selected subdistricts from Jambi Province. The results indicate that the most developed subdistricts of Jambi City (15071), with the most developed urban economy, have the highest per capita expenditure which is over Rp1,000,000, while the subdistricts of East Tanjung Jabung Regency (1507), which have the highest poverty rate in Jambi Province of 11.39%, have the lowest per capita expenditure.

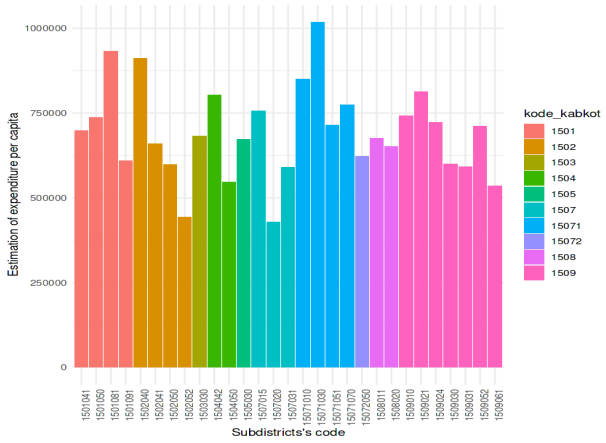
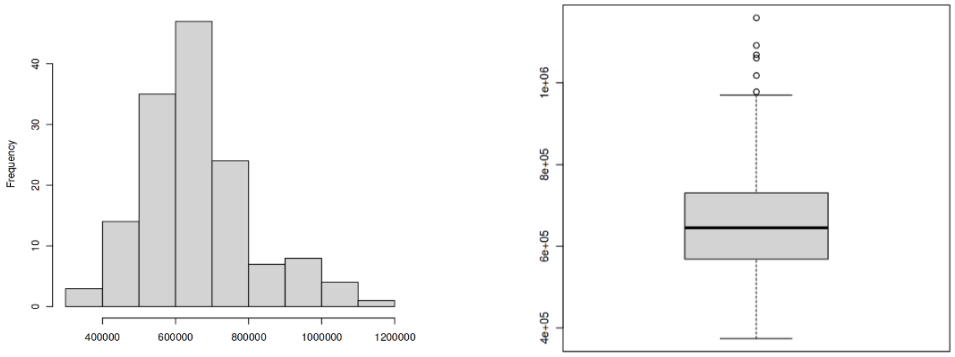


Figure 4. The results of direct estimates of per capita expenditure in a number of selected subdistricts in Jambi Province

Descriptively, the minimum value recorded was Rp373,643.9 and the maximum was Rp1,159,540, with a median of Rp645,939.4 and a mean of Rp662,822.2. The first and third quartiles were Rp568,510.2 and Rp730,734.1, respectively, and the standard deviation was Rp146,709.1. The relatively small difference between the mean and the median indicates that the distribution of the direct estimation results tends to be closer to symmetrical, although there is still a slight skew toward the right.



(a) (b)

Figure 5. Histogram (a) and box-plot (b) direct estimation

Based on Figure 5, the distribution of the direct estimators of per capita expenditure by subdistrict shows a right-skewed pattern with the presence of several extreme values in the upper tail of the distribution. This finding is supported by the boxplot visualization, which indicates that approximately 6 subdistricts were identified as outliers, with direct estimator values that are significantly higher than those of the majority of other subdistricts. Such characteristics are commonly found in small-area socioeconomic data, particularly when the sample size per area is relatively small and inter-household variability is high, thereby increasing the likelihood of unstable estimates [1], [28].

Next, the estimation results were evaluated through a series of assumption tests to ensure the model's suitability as a basis for further analysis. Based on the normality test using the Kolmogorov-Smirnov test, both the residuals and the response variable showed p-values of 0.4252 and 0.138, respectively (p-value > 0.05), so there was insufficient evidence to reject the assumption of normality. However, the results of the Jarque-Bera test on the residuals yielded a very small p-value of 4.492×10^{-6}, indicating a deviation from normality, particularly regarding skewness and kurtosis. This suggests that while the distribution generally approximates a normal distribution, there are still deviations in the tails of the distribution.

The results of the Ramsey RESET test show a p-value of 0.3292, indicating the absence of model specification errors. Additionally, the Breusch-Pagan test yields a p-value of 0.6801, indicating that the residual variances are homogeneous. Regarding multicollinearity, most explanatory variables have low Variance Inflation Factor (VIF) values (<5), but there are several variables with very high VIF values (>10), indicating the potential for strong multicollinearity, which must be taken into account when interpreting the model. In a parametric linear model, multicollinearity can cause the estimates of the model's parameters to be unstable and the model's coefficients to have a large variance. However, it does not bias the estimates [29].

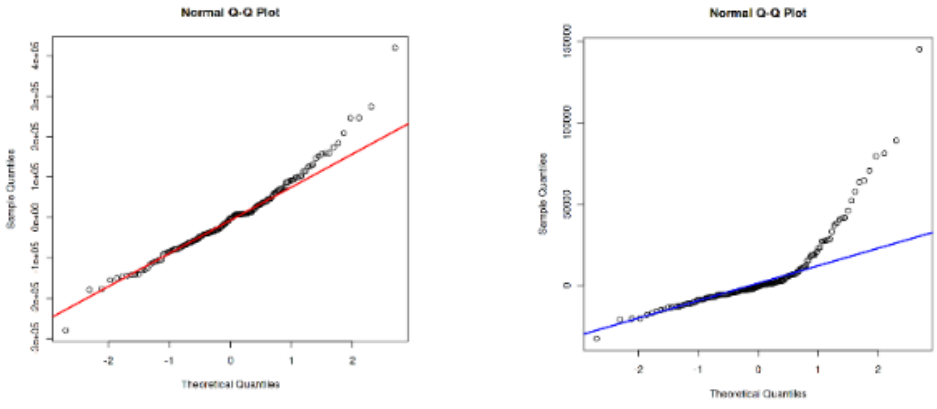
3.4. EBLUP-FH Method

It was determined via exploratory analysis and preliminary tests that the variables per capita expenditure and residuals of the regression model both observed the normality assumption. Hence, no transformations were applied in the first phase, and the analysis was directed to the selection of the explanatory variables. All 19 candidate variables listed in Table 1 (X1–X19) were considered in the selection process. Variable selection was performed in two stages. First, those that were statistically significant at the 5% significance level (p-value < 0.05) in the ordinary least squares regression of log-transformed per capita expenditure were retained. Second, variables with high multicollinearity were excluded based on the Variance Inflation Factor (VIF) criterion, retaining only those with VIF < 5 [29]. This sequential procedure yielded four variables X2, X4, X13, and X17 as the final set of explanatory variables, all of which were statistically significant and had low VIF values, indicating no evidence of multicollinearity [30].

Table 4. Selected explanatory variable

Variable	P-Value	VIF
X2	0,01386	1,119
X4	0,00764	1,788
X13	0,00129	1,292
X17	0,04075	1,478

The resulting regression model has a fairly good explanatory power with an adjusted R-squared value of 0.3239. However, when this model is applied within the EBLUP-FH framework, it is found that the area random effect component (u_i) meets the normality assumption, with a p-value of 0.2113, whereas the sampling error component (e_i) is not normally distributed (p-value < 0.05). Furthermore, there was a fairly strong correlation between the two components, namely 0.8058, indicating a potential violation of the independence assumption.

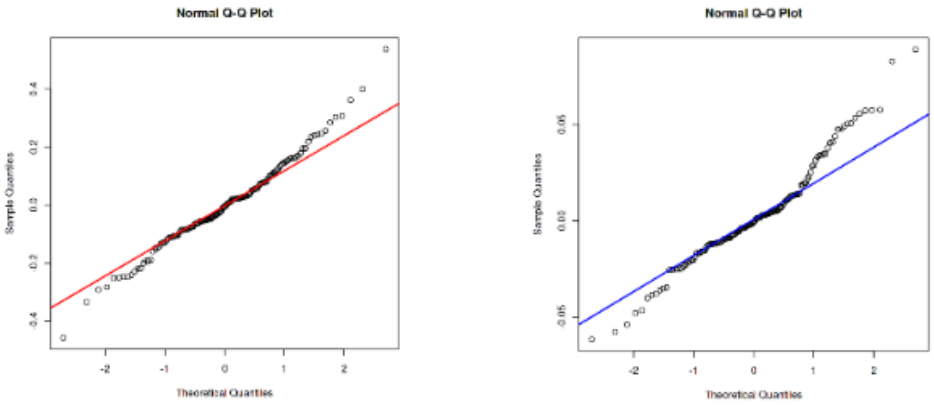


(a) (b)

Figure 6. Q-Q plot components u_i (a) and e_i (b) in the EBLUP-FH model before transformation

To address these issues, a logarithmic transformation was applied to the response variable. The results of the transformation showed an improvement in the distribution of the model components, with the area random effect and sampling error becoming closer to normal, with p-

values of 0.5263 and 0.05009, respectively. This represents an improvement compared to the pre-transformation results.



(a) (b)

Figure 7. Q-Q plot components u_i (a) and e_i (b) in the EBLUP-FH model after transformation

The EBLUP-FH estimates for a number of selected subdistricts following the transformation are presented in Figure 8. Compared to the previous direct estimates, the EBLUP-FH results show a pattern that is more consistent with the poverty rates of each regency/city. Districts in Jambi City (15071) continue to record the highest per capita expenditure, while districts in East Tanjung Jabung Regency (1507) remain in a lower range, consistent with their highest poverty rates in Jambi Province. More importantly, the anomalies previously observed in the direct estimates specifically, subdistricts with per capita expenditure that did not align with their regency's conditions appear significantly reduced in the EBLUP-FH results. This occurs because the EBLUP-FH method utilizes information from covariates, resulting in more stable estimates that are less dependent on the limited sample size of subdistricts. Nevertheless, the stability of these estimates requires further examination through an uncertainty analysis to ensure the reliability of the obtained results.

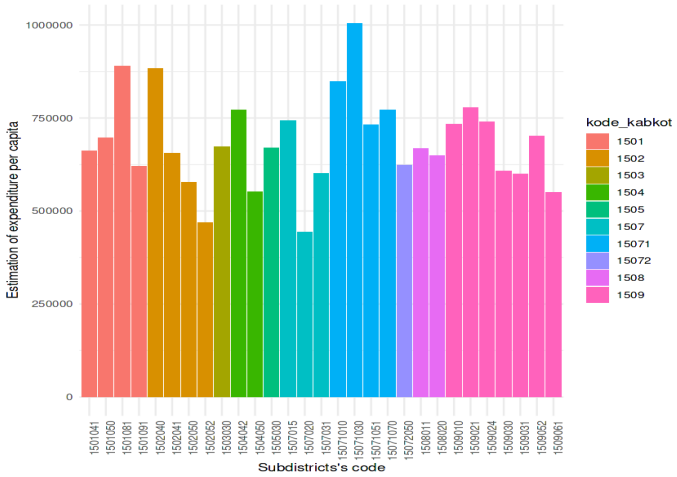


Figure 8. The results of EBLUP-FH estimate of per capita expenditure in a number of selected subdistricts in Jambi Province

Further evaluation using the bootstrap approach yielded results consistent with previous findings. The estimated results using the transformed data yield an MSE of 1808992868.57, which increased to 2143865120.68 after bootstrapping, indicates the presence of estimation variability.

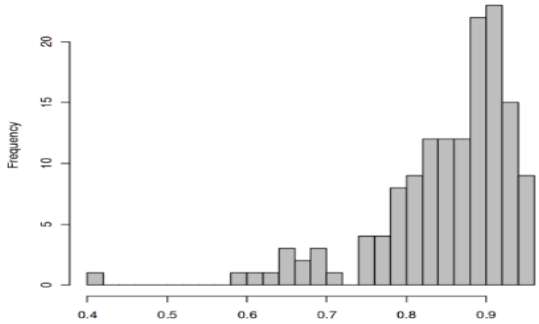


Figure 9. Shrinkage factor value

The model also does not satisfy the independence assumption, with a correlation of 0.8457, and the shrinkage factor is close to 1, indicating that EBLUP tends to rely on direct estimates [20]. This is indicated by the low MAPE value of 1.901% since MAPE is calculated using the difference between the direct estimate and the EBLUP-FH estimate.

This indicates that, although the transformation is effective in improving the distribution, there remains a dependency structure that must be taken into account when interpreting the results.

3.5. XGBoost Method

The results of the indirect estimation using the XGBoost method show that the model's performance on the training data yielded an MSE of 8,552,375,734 and a MAPE of 11.44%, while on the testing data, an MSE of 12,278,344,618 and a MAPE of 13.20% were obtained. The difference in error values between the training and testing data indicates a decline in model performance on untrained data, suggesting overfitting, although it remains within reasonable limits for a machine learning-based model.

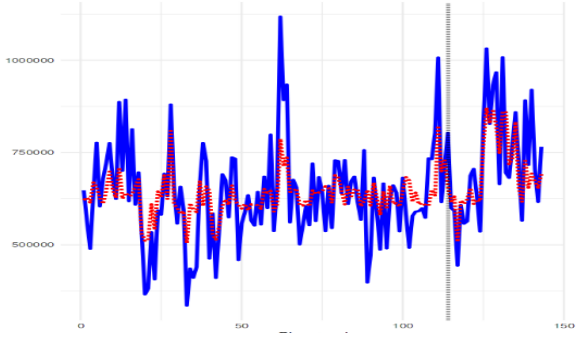


Figure 10. Comparison of results from the simple model and XGBoost

The image illustrates how the values of direct estimator are compared to values of indirect estimator, while the vertical line indicates the division of the training from the testing data. XGBoost (in red) appears to match the trend of the direct estimator (in blue) in both the training and testing data. Thus, the model is capable of recognizing the structure of the data.

Results of the analysis of feature importance indicate that in particular variable X12 was the most important XGBoost variable in the model, and accounted for a gain of 21.91%, followed by variables X4, X2 and X17 with gains of 12.46%, 9.88% and 9.08%, respectively. These variables accounted for over 50% of the total gain of the model. These results show that unlike the linear regression model, XGBoost is able to capture complex and non-linear relationships between variables.

Table 5. Important variables in the XGBoost model

Variable	Gain	Cumulative Gain
X12	0,219	0,219
X4	0,125	0,344
X2	0,099	0,443
X17	0,091	0,534

Variable	Gain	Cumulative Gain
X5	0,086	0,620
X13	0,083	0,703
X7	0,064	0,767
X14	0,048	0,815
X16	0,037	0,852
X8	0,036	0,888
X18	0,024	0,912
X11	0,023	0,935

As a result from the bootstrap analysis, by evaluating the XGBoost model, a large error was observed. Using the training data, the MSE loss was 17,206,945,347 and the MAPE was 16.39%, while using the testing data MSE loss was 23,929,363,748 and the MAPE was 18.60%. XGBoost was chosen because it is relatively new, unlike the more classical statistical methods (due to the large error) and XGBoost has been proven to perform better with High dimensional sparse data. This indicates that while XGBoost is highly adaptable to complex patterns, the model is less likely to be more robust to changes in the data. As a result, XGBoost is flexible in that it can function as a benchmark model without requiring a great deal of parametrization.

In addition to model performance, Figure 11 shows the XGBoost projected per capita expenditure for a sample of subdistricts within the test data. In general, the estimates provided by XGBoost remain aligned with the level of poverty for each regency/city. Jambi City (15071) continues to have the highest recorded average per capita expenditure. Nonetheless, results provided by XGBoost for each regency tend to be more aligned in comparison to results provided by EBLUP-FH and direct estimates. This characteristic of XGBoost is due to the nature of decision trees and is likely to be a result of general patterns observed within the training data. Local variations within subdistricts are less likely to be captured and, as is the case in subdistrict 1502040, where a value of per capita expenditure significantly higher than the value of all other subdistricts of the regency, resulted, XGBoost is likely to produce outlier results in the presence of training data outliers. As we advance to the next subsection, the focus will be the exploration of the estimation of uncertainty through the use of XGBoost.

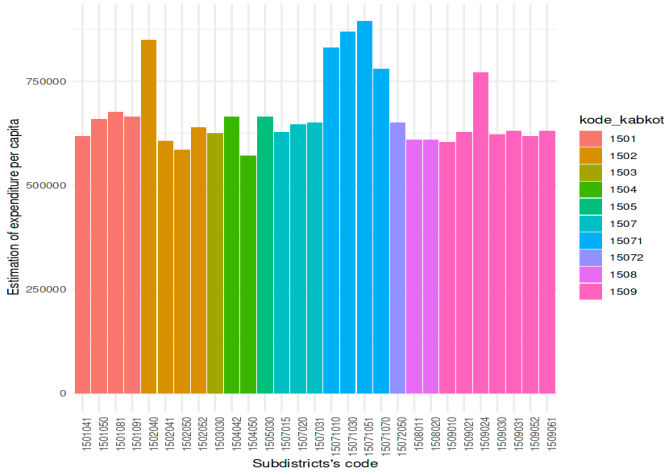


Figure 11. The results of XGBoost estimate of per capita expenditure in a number of selected subdistricts in Jambi Province

3.6. Measure of Uncertainty

Let the measured standard errors of three different estimation techniques (direct estimation, EBLUP–FH, and XGBoost) be the focus for an uncertainty analysis. Before analyzing, mention that deriving these standard errors is done from different approaches. For example, the standard error of direct estimation is derived by analytically examining the subdistrict's sampling variance. For EBLUP–FH, the Prasad–Rao MSE estimator is used, providing an estimate of EBLUP–FH accuracy that considers the sampling error and the random area effect. The standard error of XGBoost, by contrast, is an empirical bootstrap standard deviation reflecting variability in

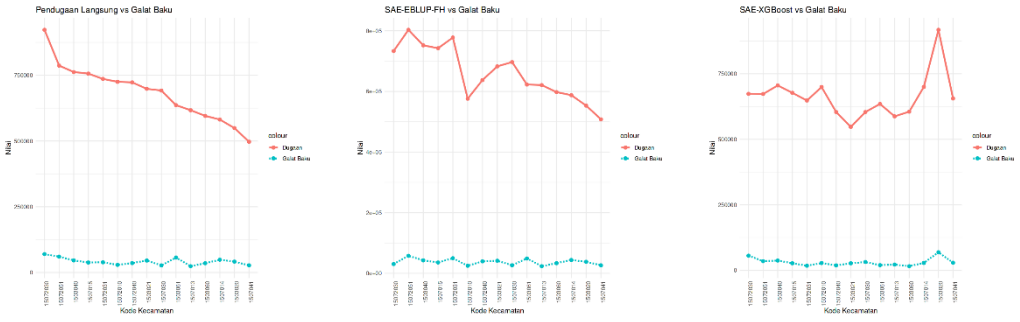
predictions across resampled training datasets it does not carry the design-based or model-based interpretation of the other two. Consequently, these three standard errors represent different types of uncertainty and should not be interpreted as perfectly comparable quantities; the comparison below is intended as a descriptive benchmark rather than a formal equivalence test. With this caveat in mind, a comparison of the standard errors reveals an interesting pattern. The average standard error of EBLUP–FH, at 40.404, is smaller than that of direct estimation, at 44.772, reflecting improved estimation precision as a result of the shrinkage mechanism inherent in EBLUP–FH. This mechanism has also proven effective in addressing the anomalies found in the data presented in Table 6.

Table 6. Estimates of 3 methods by subdistrict in regencies/cities with high, medium, and low poverty rates

Area	City/Regency's Code	Direct Estimation	EBLUP-FH Estimation	XGBoost Estimation
1503040	1503	762467.8	752044.5	705314.1
1503021	1503	698445.8	682649.1	547088.0
1503061	1503	636479.4	622750.8	634740.3
1503060	1503	595626.0	597707.5	605257.8
1503020	1503	549173.7	553185.6	917396.6
1507015	1507	756445.3	742559.2	677086.1
1507030	1507	692225.3	696801.5	603899.4
1507013	1507	617279.8	620981.4	587312.5
1507014	1507	581723.9	587625.5	699719.4
1507041	1507	497289.5	508083.1	655599.4
15072030	15072	922996.7	733308.9	673254.8
15072051	15072	786967.7	803170.6	672537.9
15072031	15072	736208.3	777731.5	647615.0
15072010	15072	725294.5	575920.5	699192.2
15072040	15072	723174.2	637504.7	603816.2

Several subdistricts in Tanjung Jabung Timur Regency (1507) with the highest poverty rates actually recorded per capita expenditures that exceeded those of some subdistricts in the regency with the lowest poverty rates (15072). EBLUP-FH was able to correct this anomaly by producing estimates that are more contextually plausible while also yielding a smaller standard error than direct estimates.

Meanwhile, XGBoost recorded the smallest average standard error of 34.261, but this result needs to be interpreted with caution. In subdistricts with the highest estimated values, a small standard error is not always accompanied by the accuracy of the estimated values, indicating that the model is overconfident in its predictions.



(a) (b) (c)

Figure 12. Comparison of SE and the estimation results of the method direct (a) EBLUP-FH (b) XGBoost (c)

Furthermore, XGBoost tends to produce homogeneous estimates across poverty groups, thereby failing to meaningfully distinguish between the poorest and richest subdistricts within a regency. As mentioned, the sudden spike that is observed in subdistrict 1503020 in Figure 12 accentuates XGBoost's disadvantage of being unable to adequately differentiate socio-economic disparities in the given regions. Thus, although XGBoost outperforms EBLUP-FH numerically in terms of standard error, EBLUP-FH is considered more reliable in producing consistent and contextually relevant estimates, particularly in regions with varying socio-economic conditions.

4. CONCLUSION

Based on the analysis results, the per capita expenditure data, which skews to the right and contains outliers, was successfully stabilized through winsorization, resulting in more robust data without losing the main structure of the distribution. In indirect estimation, the logarithmic transformation in the EBLUP-FH model was able to improve some of the distribution assumptions of the model components and produce relatively stable estimates based on MSE, MAPE, and bootstrap evaluations. However, the EBLUP-FH model still exhibits significant limitations, particularly a violation of the assumption of independence among error components with a correlation of 0.8457, as well as a tendency for the model to be highly dependent on direct estimates, as indicated by a shrinkage factor approaching 1.

XGBoost demonstrated superiority in handling nonlinear relationships and complex data structures without relying on parametric assumptions, producing the smallest average bootstrap-based standard error (34.261) compared to EBLUP–FH (40.404) and direct estimation (44.772). It is important to clarify that these values are average standard errors, not MAPE values; the MAPE results reported in Section 3.5 are separate metrics. However, in this study, XGBoost exhibited high variation in bootstrap results and produced homogeneous subdistrict-level estimates, reflecting lower model stability under conditions of limited sample size and high variability. Furthermore, XGBoost's small bootstrap standard error does not always correspond to accurate estimates, indicating that the model tends to be overconfident under certain data conditions. This example shows that in the context of SAE, the standard error and degree of estimation accuracy (how closely the estimation is to the actual value) are two distinct evaluation metrics that cannot be confused or mixed.

Consequently, in the context of the distinct methodologies involved, the evaluation results illustrate that both EBLUP-FH and XGBoost have their own advantages and disadvantages. EBLUP-FH is a model-based estimator that is grounded on the sampling theory. It produces estimates that are in general (but not always, as it is influenced by certain violations of a model's assumptions) more analytically precise than XGBoost standard errors. In this case, XGBoost is a flexible benchmark predictive model that estimates sampling uncertainty through bootstrap resampling. Because EBLUP-FH and direct estimators are fundamentally different, any comparison with XGBoost for standard errors will describe rather than quantify. In SAE, precision, which can be signified by a small standard error, must be evaluated separately from accuracy, which pertains to closeness to the true parameter. A small standard error from any method does not reflect contextually accurate or unbiased small-area estimates. This must be addressed, as standard error alone does not denote accuracy and is contextually biased.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The data used in this study were obtained directly from BPS (Statistics Indonesia) Provinsi Jambi under a data request for academic/research purposes.

AUTHOR CONTRIBUTIONS

Fida Fariha Amatullah: Conceptualization; Methodology; Formal Analysis; Software; Data Curation; Writing – Original Draft; Visualization. **Khairil Anwar Notodiputro**: Conceptualization; Validation; Supervision; Writing – Review & Editing. **Anwar Fitrianto**: Methodology; Validation; Supervision; Writing – Review & Editing.

REFERENCES

[1] J. N. K. Rao and I. Molina, *Small Area Estimation*, 2nd ed. United States of America: John Wiley & Sons, Inc., 2015.
[2] A. O. Fengki, K. A. Notodiputro, and K. Sadik, "Bisakah Memperoleh Statistik Indeks Harga

Konsumen Tingkat Provinsi Di Indonesia Dengan Ketelitian Yang Lebih Baik?," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, pp. 297–306, 2020, doi: 10.34123/semnasoffstat.v2019i1.178.
[3] S. M. Mah, M. Brown, R. C. Colley, L. C. Rosella, G. Schellenberg, and C. Sanmartin, "Exploring the use of experimental small area estimates to examine the relationship between individual-level and area-level community belonging and self-rated health," *Heal. Reports*, vol. 35, no. 3, 2024, doi: 10.25318/82-003-x202400300001-eng.
[4] A. Ubaidillah *et al.*, "Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia," *J. Appl. Stat.*, vol. 0, no. 0, pp. 1–17, 2019, doi: 10.1080/02664763.2019.1615420.
[5] S. Sugasawa and T. Kubokawa, *Mixed-Effects Models and Small Area Estimation*. 2023.
[6] S. Elahe, D. Shahsavani, and M. Reza, "Small area estimation with partially linear mixed-t model with measurement error," *J. Comput. Appl. Math.*, vol. 446, pp. 1–12, 2024, doi: 10.1016/j.cam.2024.115871.
[7] T. C. Nokeri, *Data Science Solutions with Python*. New York: Apress, 2022.
[8] C. Robinson, B. Dilkina, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala, "Machine learning approaches for estimating commercial building energy consumption," *Appl. Energy*, no. May, pp. 889–904, 2017, doi: 10.1016/j.apenergy.2017.09.060.
[9] C. E. Kontokosta, B. Hong, N. E. Johnson, and D. Starobin, "Computers , Environment and Urban Systems Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities," *Comput. Environ. Urban Syst.*, no. March, pp. 151–162, 2018, doi: 10.1016/j.compenvurbsys.2018.03.004.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
[11] 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.*, pp. 1–18, 2022, doi: 10.1186/s12942-022-00304-5.
[12] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geosci. Front.*, vol. 12, no. 1, pp. 469–477, 2021, doi: 10.1016/j.gsf.2020.03.007.
[13] Z. Shao, M. N. Ahmad, and A. Javed, "Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface," *Remote Sens.*, vol. 16, no. 4, 2024, doi: 10.3390/rs16040665.
[14] J. Frank E. Harrell, *Regression Modeling Strategies*, 2nd ed. New York: Springer, 2015.
[15] C. F. Dormann *et al.*, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," *Ecography (Cop.).*, vol. 36, no. 1, pp. 27–46, 2013, doi: 10.1111/j.1600-0587.2012.07348.x.
[16] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*. New York: Cambridge University Press, 1997.
[17] S. Hermawan and Amirullah, *Metode Penelitian Pendekatan Kuantitatif Kualitatif*. Sidoarjo: Universitas Muhammadiyah Sidoarjo, 2016.
[18] W. A. Nurizza, "Penerapan Model Fay-Herriot pada Small Area Estimation Studi Simulasi Pengeluaran Per Kapita Level Kabupaten/Kota Provinsi Kalimantan Timur 2020," *BESTARI Bul. Statisitika dan Apl. Terkini*, vol. 1, no. 1, pp. 1–7, 2021.
[19] R. Ananda, K. A. Notodiputro, and M. N. Aidi, "Modified Mixed Effects Random Forest in Small Area Estimation Using PCA and Rotation Forest with Correlated Auxiliary Variables," *Sci. J. Informatics*, vol. 11, no. 3, pp. 705–720, 2024, doi: 10.15294/sji.v11i3.10633.
[20] S. L. Lohr, "Sample Surveys : Model-Based Approaches," *Int. Encycl. Soc. Behav. Sci.*, vol. 20, pp. 886–889, 2015, doi: 10.1016/B978-0-08-097086-8.42166-5.
[22] M. Sagming, R. Heymann, and M. V. Visaya, "Using topological data analysis and machine learning to predict customer churn," *J. Big Data*, vol. 11, pp. 1–31, 2024, doi: 10.1186/s40537-024-01020-6.
[23] L. Yang, X. Zhang, and J. Chen, "Winsorization greatly reduces false positives by popular

differential expression methods when analyzing human population samples," *Genome Biol.*, vol. 25, pp. 1–4, 2024, doi: 10.1186/s13059-024-03230-w.
[24] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.
[25] E. Vivas, H. Allende-Cid, and R. Salas, "A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score," *Entropy*, vol. 22, no. 12, pp. 1–24, 2020, doi: 10.3390/e22121412.
[26] BPS Jambi, *Provinsi Jambi Dalam Angka 2022*. Jambi: BPS Provinsi Jambi, 2022.
[27] A. Moretti and A. Whitworth, "Estimating the Uncertainty of a Small Area Estimator Based on a Microsimulation Approach," *Sociol. Methods Res.*, vol. 52, no. 4, pp. 1785–1815, 2023, doi: 10.1177/0049124120986199.
[28] N. Tzavidis, L. Zhang, and A. Luna, "From start to finish : a framework for the production of small area official statistics," *J. R. Stat. Soc. Ser. A*, vol. 4, pp. 927–979, 2018, doi: https://doi.org/10.1111/rssa.12364.
[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York: Springer Nature, 2021.
[30] P. Corral, I. Molina, A. Cojocaru, and S. Segovia, *GUIDELINES TO SMALL AREA ESTIMATION FOR*. Washington DC: The World Bank, 2022.