Small Area Estimation of Extreme Poverty Using Zero-Inflated Binomial GLMM: A District-Level Case Study in North Sumatra 2024

M D F Br L Gaol[1,*], B S Iryani[2], and E Lestariningsih[3]

[1] BPS-Statistics Indonesia of Padang Lawas Regency, Jl. Karya Pembangunan Lingkungan VI Pasar Sibuhuan, Padang Lawas Regency, 22763, Indonesia
[2,3] BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, 10710, Indonesia

*Corresponding author's email: martadesnafitria@gmail.com

Abstract. Eradicating extreme poverty is a key objective of SDG 1, with a global benchmark of reducing the proportion of people living below the US$1.90 PPP poverty line. However, in 2024, Indonesia—particularly North Sumatra Province—continues to face persistent challenges in achieving this target. Direct estimation based on the Foster-Greer-Thorbecke (FGT) formula using Susenas microdata suffers from large sampling errors (RSE > 25 percent) and zero estimates in multiple districts due to small or absent samples, indicating serious issues of zero inflation and overdispersion. Overdispersion occurs when the observed variability in poverty data exceeds what is expected under a standard binomial distribution, leading to underestimated standard errors and unreliable inferences. To overcome these limitations, this study applies a model-based SAE approach using the ZIB-GLMM. This method incorporates auxiliary variables from the 2024 Podes dataset. The SAE-ZIB GLMM model showed the best overall performance, effectively capturing the high zero prevalence and area-level variability in North Sumatra's extreme poverty data. It achieved the lowest RMSE and MAE and an optimal AIC value, indicating superior prediction accuracy and model fit compared to other approaches. Thus delivering essential evidence to inform data-driven policy decisions and the design of targeted poverty alleviation strategies.
Keyword: extreme poverty, overdispersion, small area estimation, zero-inflated binomial GLMM, zero inflation

1. Introduction
The eradication of extreme poverty is a global priority under the Sustainable Development Goals (SDGs), specifically Goal 1: No Poverty, with Indicator 1.1.1 measuring the proportion of the population living below the international poverty line of US$1.90 PPP per day [1]. The ambitious target to eliminate extreme poverty by 2030 has encouraged countries, including Indonesia, to design more comprehensive and data-driven strategies. However, as of 2024, Indonesia has not yet achieved the zero percent target for extreme poverty [2], prompting intensified efforts for poverty alleviation as mandated in Presidential Instruction No. 8 of 2025 [3].

Although the poverty rate in North Sumatra Province in 2023 (8.15 percent) was lower than the national poverty rate of 9.36 percent, pockets of extreme poverty remain a critical issue that must be addressed [2]. This condition is reflected in the high number of households and individuals in the first income decile (classified as poor) in the province, which ranks among the highest in Indonesia in 2023. In response, thegovernment has implemented various poverty alleviation programs focusing on ensuring access to basic needs, improving access to health services, and providing disaster assistance, with the aim of reducing extreme poverty more effectively and equitably.

In Indonesia, the persistence of extreme poverty remains a significant challenge, particularly in resource-constrained regions. North Sumatra Province, with its diverse socioeconomic landscapes, continues to face issues in measuring extreme poverty due to limited survey sample coverage. This limitation often leads to imprecise and biased direct estimates, and in some districts, estimates cannot be generated at all due to the absence of sample data.

To overcome these challenges, Small Area Estimation (SAE) approaches have been widely adopted in recent studies and official practices to produce reliable estimates in small domains. SAE leverages auxiliary data to improve estimation precision through statistical modeling. However, most SAE methods such as Empirical Best Linear Unbiased Prediction (EBLUP) or standard GLMM do not explicitly address the issues of zero inflation and overdispersion, which are common in extreme poverty microdata [4].

Overdispersion occurs when the variance of the observed data substantially exceeds the variance assumed by a standard binomial model, often caused by unobserved heterogeneity, clustering effects, or uneven distribution of events across small areas [5]. This phenomenon inflates type I errors, biases parameter estimates, and leads to underestimation of standard errors, ultimately reducing the precision and reliability of small area estimates. In the context of extreme poverty estimation, overdispersion commonly arises due to the rarity of events in many districts and high variability in areas with small sample sizes. Conventional SAE models such as EBLUP, standard GLMM, or basic INLA are not specifically designed to address this issue, resulting in poor model fit and unstable predictions in zero-heavy datasets.

Previous studies have demonstrated that extreme poverty data often contains excessive zeros—not necessarily indicating a true absence of extreme poor in the population, but rather a result of limited sampling [4]. Moreover, overdispersion arises when data variance exceeds that of a standard binomial model, leading to inaccuracies in parameter estimation and inference. These mismatches highlight the need for models that better align with the data’s characteristics.

To address these issues, this study adopts the Zero-Inflated Binomial Generalized Linear Mixed Model (ZIB-GLMM) within the SAE framework. This model is capable of simultaneously addressing zero inflation and overdispersion, and has shown improved accuracy in similar contexts, as demonstrated by previous researchers [4], [6], [7], [8], [9]. The Zero-Inflated Binomial Generalized Linear Mixed Model (ZIB-GLMM) addresses zero inflation by modeling two underlying processes: (1) a structural zero component that captures areas with zero extreme poverty cases due to survey limitations or actual absence, and (2) a binomial component that models the probability of extreme poverty in areas with non-zero cases. This dual structure allows the model to properly distinguish between true zeros and sampling zeros, improving estimation accuracy in zero-inflated datasets.

In addition, ZIB-GLMM incorporates random effects at the area level, which allows the model to capture unobserved heterogeneity and account for overdispersion. By including random intercepts, the model absorbs extra variability across areas, leading to more realistic variance estimates and more robust inferences. This combination of zero-inflation handling and random effects makes ZIB-GLMM particularly well-suited for small area estimation of rare events such as extreme poverty.

The ZIB-GLMM approach within the SAE framework produces more accurate, reliable, and realistic estimates than conventional SAE methods because it directly addresses key challenges commonly found in small area poverty data, namely zero inflation and inter-area variability. Unlike conventional models that assume standard distributional structures, ZIB-GLMM introduces a zero-inflation component that distinguishes between structural zeros and sampling zeros. This allows the model to handle areas with no observed cases more appropriately, avoiding the severe overestimation often produced by traditional methods. In addition, the inclusion of random effects captures differences between districts, leading to more stable estimates in areas with limited or no samples.

Furthermore, the model’s flexible structure enables a better fit to the true distribution of the data, resulting in lower prediction errors and improved model performance indicators such as RMSE, MAE, and AIC. By simultaneously accounting for excess zeros and area-level heterogeneity, ZIB-GLMM generates estimates that are closer to real conditions on the ground and more robust for policy use. Previous studies have shown the effectiveness of this approach in improving the quality of small area estimates, including applications in

Indonesia's unemployment data [8] and broader SAE methodological development [10]. This provides a strong justification for using ZIB-GLMM over conventional SAE methods in extreme poverty estimation. Therefore, statistical modeling through SAE is employed to address these data gaps and produce reliable district-level estimates.

Previous studies have explored various approaches in applying SAE methods for poverty estimation. Ramadhan and Ubaidillah integrated SAE with Partitioning Around Medoids (PAM) clustering to estimate the percentage of people living in extreme poverty at the regency/city level in East Java Province [4]. This approach was effective in grouping similar areas to stabilize estimates, particularly in small sample areas. Additionally, Larasati and Permatasari compared SAE EBLUP and SAE Hierarchical Bayes (HB) methods in estimating poverty levels in East Java Province, showing that HB-based models can produce more stable estimates than EBLUP under certain small area conditions [11]. However, neither of these approaches explicitly addressed the issues of zero inflation and overdispersion commonly found in microdata on extreme poverty. Therefore, this study proposes an advanced methodology by integrating the ZIB-GLMM within the SAE framework, which simultaneously handles both challenges. This approach is expected to yield more accurate, reliable, and realistic small area estimates for extreme poverty in sparse areas characterized by excess zeros and high inter-area variability.

2. Research Method

This study employs secondary data sourced from Statistics Indonesia (Badan Pusat Statistik), specifically the raw dataset of the March 2024 National Socio-Economic Survey (Susenas) Core Module, supplemented with auxiliary variables obtained from the official publication of the 2024 Village Potential Statistics (Potensi Desa/Podes). The study focuses on North Sumatra Province, encompassing 25 districts and 8 cities. A total of 77,836 individual observations were utilized for direct estimation, while indirect estimation was conducted across 33 districts/cities. All data analyzed correspond to the 2024 reference period. The auxiliary variables used for the indirect estimation procedure are detailed in table 1. The selection of auxiliary variables was guided by previous empirical studies that identified socioeconomic and infrastructure-related indicators as strong predictors in small area estimation of poverty and related social indicators [4].

Table 1. Predictor variables of extreme poverty

Variables	Description
X1	Proportion of villages/districts in which the majority of the population is employed in the agricultural sector
X2	Proportion of villages/districts where the main source of drinking water for most households is derived from natural, unimproved sources such as springs, rivers, rainwater, and others
X3	Proportion of villages/districts where household sanitation facilities are predominantly non-hygienic or not classified as improved latrines
X4	Proportion of villages/districts that have received government-initiated labor-intensive public works programs (cash-for-work schemes)
X5	Proportion of villages/districts that have received direct cash transfer assistance from social protection programs
X6	Proportion of villages/districts reporting the existence of individuals suffering from malnutrition
X7	Ratio of educational infrastructure (from primary to senior secondary levels) per 1000 population
X8	Ratio of healthcare infrastructure (hospitals) per 1000 population

The estimation of extreme poverty at the district/city level in North Sumatra Province in 2024 was conducted using two approaches: direct estimation and indirect estimation. The direct estimation was performed using the Foster-Greer-Thorbecke (FGT) formula, while the indirect estimation applied Small Area Estimation (SAE) models, which were simulated using SAE binomial INLA, SAE Zero-Inflated

ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

Binomial (ZIB) INLA, SAE beta-binomial GLMM, SAE ZIB Generalized Linear Mixed Model (GLMM), and SAE beta-binomial GLMM models.

Equation (1) represents the general formula of the FGT index, where denotes the poverty line and y_I represents the average monthly per capita expenditure of individuals below the poverty line (<), while α is the poverty sensitivity parameter, taking values of 0, 1, or 2. When = 0, the measure corresponds to the proportion of the population living in poverty. Therefore, in this study, extreme poverty was directly estimated using the Poverty Headcount Index (), which reflects the percentage of the population classified as extremely poor. The extreme poverty line applied in this study refers to the World Bank threshold, which defines extreme poverty as living on less than US$1.90 Purchasing Power Parity (PPP) per day [1].

$$P_\alpha = \frac{1}{n}\sum_{i=1}^{q} \left[\frac{z - y_i}{z}\right]^\alpha \tag{1}$$

$$P_0 = \frac{1}{n}\sum_{i=1}^{q} 1 \tag{2}$$

Referring to Equation (2), represents the proportion of the population living below the poverty line or the extreme poverty rate; n is the total population; and $i, \ldots, q$ indicates the number of individuals living below the poverty line.

This study adopts a SAE framework using the ZIB model to address the presence of excessive zeros in binary response data. The SAE-ZIB framework is capable of estimating both surveyed and non-surveyed domains, thus enhancing the reliability of small area statistics. To account for uncertainty in model-based estimators, particularly those involving generalized linear mixed models with zero inflation, the standard error is approximated through a bootstrap method. This approach is based on the techniques recommended by some researchers [6], [7], [12]. Thus, the SAE-ZIB method is adapted in this study to estimate extreme poverty at the district and city levels. It accommodates both survey and auxiliary data, capturing overdispersion and structural zeros in the poverty status outcome.

The estimation of extreme poverty of this study was conducted in a series of structured steps below and can be seen in figure 1.

Direct estimation of extreme poverty was performed using data from the March 2024 round of the Susenas. Individuals were classified as extremely poor if their per capita expenditure fell below the international extreme poverty line of $1.90 per day, equivalent to IDR 322,170 per month. Direct estimates were generated without applying individual sampling weights, as the available weights were excessively large and led to overestimation. The weights were excluded because they produced unrealistic population totals exceeding census benchmarks.

The Relative Standard Error (RSE) values of the direct estimates of extreme poverty were calculated at the district/city level for North Sumatra Province in 2024. In five out of the seven areas with extreme poor populations, RSE values exceeded 25 percent—surpassing the commonly accepted threshold—indicating low precision and reliability [13]. The reference for this RSE cutoff value is based on the standards used by the Australian Bureau of Statistics and Statistics Indonesia. Consequently, SAE modeling was pursued to improve the accuracy of estimates in these domains.

The direct estimation results revealed the presence of zero-inflation in the data (i.e., a high frequency of zero values), necessitating an SAE model capable of addressing such data structure.

Microdata was compiled and merged with area-level auxiliary data.

An overdispersion test was conducted to assess the distributional characteristics of extreme poverty data. As the data exhibited overdispersion in North Sumatra Province, it was necessary to employ an SAE model that could appropriately account for this feature.

Explanatory variable selection was carried out through stepwise selection, correlation analysis, and multicollinearity diagnostics. This process ensured the identification of the best combination of predictors, minimized multicollinearity, and promoted a parsimonious final model.

Several SAE models were simulated to generate accurate and precise estimators that accommodate both overdispersion and zero-inflation in the data. The simulations included binomial SAE using INLA, ZIB-INLA, beta-binomial GLMM SAE, ZIB-GLMM SAE, and a hybrid beta-binomial GLMM SAE. Although negative binomial models are suitable for handling overdispersion, they are less appropriate for zero-inflated binary outcomes such as extreme poverty prevalence. Therefore, the Zero-Inflated Binomial GLMM was preferred, as it explicitly models excess zeros and accounts for area-level variability.

Lastly, the simulated models were compared by evaluating their predictive accuracy based on relevant performance metrics. The best-performing model was then selected based on these evaluation results for final estimation and reporting.

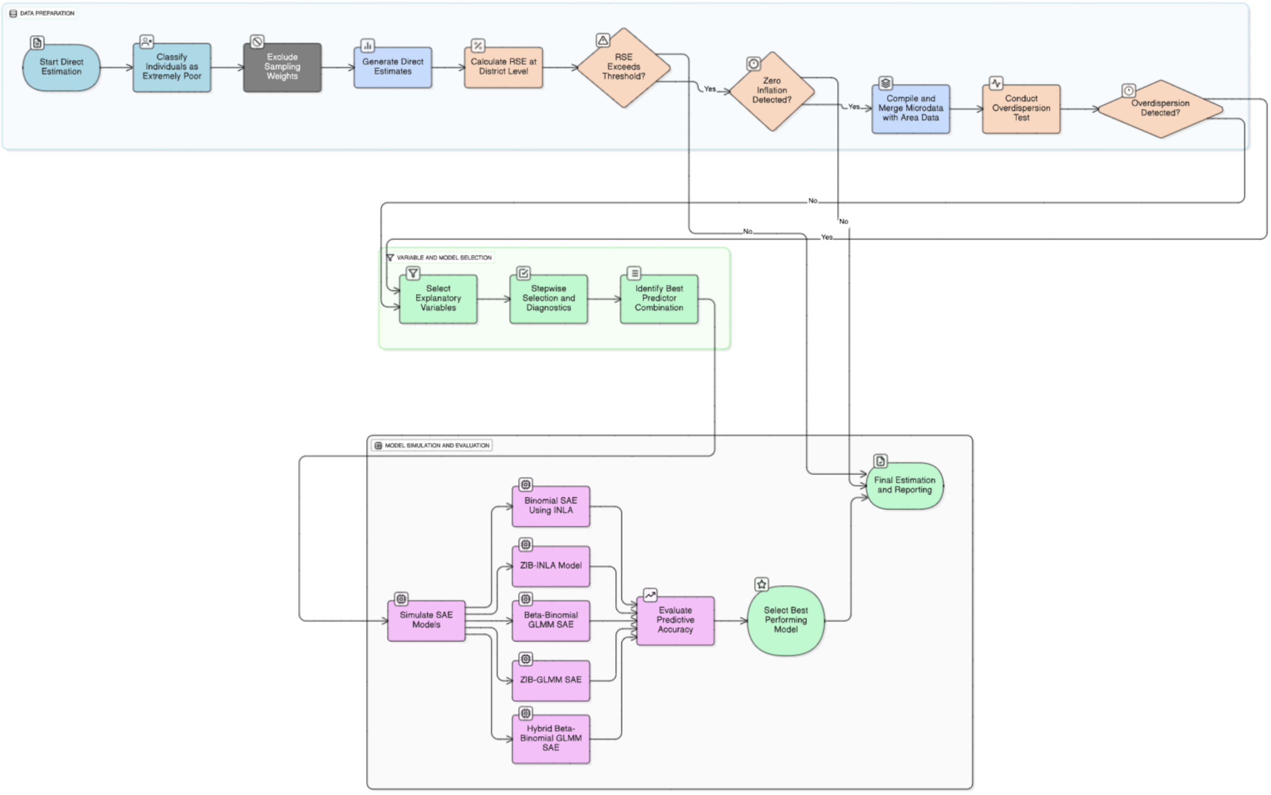

Figure 1. Flowchart of extreme poverty estimation using SAE ZIB-GLMM approach

3. Result and Discussion

3.1 Direct Estimate of Extreme Poverty

Extreme poverty in North Sumatra Province in 2024 shows substantial spatial disparities across districts and cities. Of the 33 districts/cities, only seven reported the presence of extremely poor individuals: Nias, Deli Serdang, Langkat, South Nias, North Nias, West Nias, and Gunung Sitoli City. The remaining 26 areas recorded zero observations of extreme poverty. Out of the 33 districts and cities analyzed, 26 (78.8%) recorded zero extreme poverty cases in the direct estimation, indicating severe zero inflation in the dataset.

The reliability of direct estimates was evaluated using the Relative Standard Error (RSE). Only South Nias and North Nias achieved RSE values below 25 percent, indicating acceptable precision (see table 2). The other five areas had RSE values exceeding 25 percent, rendering them unreliable and demonstrating the limitations of direct estimation in small-sample areas. This underscores the need for alternative methods, such as Small Area Estimation (SAE).

Table 2. Direct estimates and RSE of extreme poverty in North Sumatra, 2024

District/City	Extreme Poverty Proportion	Standard Error	RSE (percent)	Accuracy Category
Nias	0.0037	0.0012	33.27	Not Accurate (> 25 percent)
Deli Serdang	0.0019	0.0007	37.76	Not Accurate (> 25 percent)
Langkat	0.0025	0.0009	37.75	Not Accurate (> 25 percent)
South Nias	0.0133	0.0023	17.56	Accurate (≤ 25 percent)
North Nias	0.0231	0.0031	13.58	Accurate (≤ 25 percent)
West Nias	0.0025	0.0010	40.77	Not Accurate (> 25 percent)
Gunung Sitoli City	0.0038	00012	33.27	Not Accurate (> 25 percent)

Data source: Susenas March, 2024

Direct estimation often suffers from instability and large variances in areas with small samples, leading to high RSE and inconsistent poverty rates [14]. To address this, model-based SAE methods improve estimation by borrowing strength across areas. Due to the unreliability of direct estimates—characterized by high relative standard errors (RSE > 25 percent) [15] and missing estimates in unsampled areas—a SAE approach is required to improve accuracy and enable estimation in data-sparse regions [16]. Several other districts/cities yielded Not Available (NA) values for the RSE, due to direct estimates returning zero or the absence of extreme poor individuals in the survey sample [11]. However, this does not imply the complete absence of extreme poverty in the population of these areas [4]; rather, it reflects the inherent limitations of survey-based estimation methods in capturing rare events within small samples.

This study compares INLA, ZIB-INLA, Beta-Binomial GLMM, and ZIB-GLMM to evaluate their performance in handling zero inflation and overdispersion in extreme poverty data. While INLA provides a baseline Bayesian model, it struggles with zero-heavy data. ZIB-INLA addresses zeros but remains unstable with sparse samples. Beta-Binomial GLMM captures overdispersion but ignores zero inflation. ZIB-GLMM combines both strengths, offering a more flexible and robust framework. This comparative approach ensures that the model selection is evidence-based rather than assumption-driven.

3.2 Overdispersion and Zero Inflation of Extreme Poverty Data

Direct estimation of extreme poverty in North Sumatra Province reveals a substantial prevalence of zero values: 26 out of 33 districts/cities reported no individuals classified as extremely poor. This strongly indicates a zero-inflated data structure, warranting the application of a SAE model that accounts for excess zeros—particularly the Zero-Inflated Binomial (ZIB) SAE model. ZIB models are especially suitable for binomial outcomes dominated by zero counts, enhancing estimation accuracy at the small-area level. The zero inflation is likely due to survey-based sampling limitations, which may fail to capture rare events in small domains. Importantly, the absence of observed cases does not imply the actual absence of extreme poverty in the population [4].

Additionally, overdispersion tests confirm that the binomial variance significantly exceeds expected values (variance > 2), indicating overdispersion in the data. This issue often arises from unit-level heterogeneity aggregated at the area level [8]. In such contexts, standard binomial models within SAE frameworks are prone to inefficiency or bias, necessitating more flexible alternatives—such as the Beta-Binomial or Zero-Inflated Binomial Generalized Linear Mixed Model (ZIB-GLMM).

Empirical evidence further supports this choice. For instance, SAE-ZIB was successfully applied to unemployment data in Jambi Province, Indonesia, outperforming both direct estimation and synthetic ZIB approaches in terms of standard errors and model fit [8]. These results reinforce the relevance of ZIB-based SAE models when dealing with overdispersed, zero-inflated data structures, such as those often seen in extreme poverty indicators.

In light of these findings, this study investigates multiple SAE modeling strategies that explicitly account for overdispersion and excess zeros. Through simulation and model comparison, the aim is to determine the most appropriate and robust approach—one that improves estimation accuracy, incorporates hierarchical area-level variation via random effects, and provides valid estimates of extreme poverty across districts and cities.

3.3 Model Comparison and Justification of ZIB-GLMM SAE as the Best Fit

A binary logistic regression model developed using a stepwise selection approach yielded a combination of significant explanatory variables. Although the model including all variables was significant, the exclusion of X2 produced a lower Akaike Information Criterion (AIC), suggesting a better fit. The final model retained seven explanatory variables: X1, X3, X4, X5, X6, X7, and X8, showing a relatively low AIC and residual deviance. However, a strong correlation was detected among X1 and X5 ($r = 0.94$), X1 and X8 ($r = -0.85$), and X5 and X8 ($r = -0.78$), with Variance Inflation Factor (VIF) values exceeding 10, indicating severe multicollinearity. To address this issue, variables X1, X7, and X8 were excluded. The resulting model, excluding X1, X2, X7, and X8, maintained significance while reducing multicollinearity, as all VIF values were below 5. Multicollinearity was assessed and managed before proceeding to mixed model implementation.

The four explanatory variables retained in this study—related to sanitation, malnutrition, and the presence of social assistance programs—are conceptually aligned with key indicators used in multidimensional poverty measurement, particularly under the dimensions of health and standard of living [17]. The prevalence of non-sanitary defecation facilities corresponds to inadequate sanitation, while the presence of malnutrition reflects deprivation in health. Meanwhile, the inclusion of cash-for-work and direct cash transfer programs serves as proxies for state interventions targeting extremely poor populations. The fact that these four variables were statistically selected and retained in the final model highlights their empirical strength in explaining and estimating extreme poverty, thereby confirming their theoretical relevance and practical utility in the context of small area estimation.

The Bayesian INLA-based binomial SAE model, which assumed a standard binomial distribution for extreme poverty proportions, yielded high Deviance Information Criterion (DIC) and Watanabe-Akaike Information Criterion (WAIC) values. Moreover, predictions were unrealistic in areas with zero observations, indicating that this model was not optimal for handling zero-inflation, often leading to overestimation and lacked robustness in small areas with high RSE.

The SAE ZIB-INLA model incorporated zero-inflated binomial distribution to account for excessive zeros, offering some improvement, but it still produced overestimated predictions in zero-observation areas. This suggests the model remained overfit in certain regions despite producing lower Root Mean Square Error (RMSE) (0.4566) and Mean Absolute Error (MAE) (0.2337) than the standard binomial SAE, and the WAIC remained high.

A beta-binomial GLMM-based SAE model, applying a frequentist approach, successfully addressed overdispersion in proportion data and captured inter-area variation that the standard binomial model could not. With AIC = 84.8, this model was better suited for overdispersed data but did not explicitly address the issue of zero-inflation, limiting its accuracy in districts with zero extreme poverty.

A hybrid model combining ZIB and beta-binomial GLMM was also simulated. This extended version aimed to balance both zero-inflation and overdispersion. Although it addressed both issues and produced an AIC of 89.1, it did not significantly outperform the ZIB-GLMM model and introduced additional complexity.

The SAE-ZIB GLMM model, incorporating a binomial GLMM with a zero-inflation component, demonstrated superior performance. It was well-suited for the characteristics of extreme poverty data in North Sumatra, which includes many zero values and inter-area variability. The zero-inflation component was statistically significant (intercept = 1.18, $p = 0.0066$), supporting the model's appropriateness. Furthermore, the ZIB-GLMM model yielded better prediction accuracy compared to other models, as evidenced by the lowest RMSE (0.1088) and MAE (0.0559), particularly in districts with zero observed extreme poverty. Its AIC value (87.4) also reflects the model's balance between complexity and fit, surpassing alternative models tested in this study.

These model-based estimates provide a strategic statistical foundation for monitoring progress toward SDGs Target 1.1 on eradicating extreme poverty at a granular administrative level. The extreme sparsity in poverty counts across most districts—reflected in over 75 percent of areas reporting zero extreme poor—underscores the necessity of a model that appropriately handles zero inflation and avoids overfitting in zero-observation zones. ZIB-GLMM produced predicted proportions that closely aligned with direct estimates, particularly in low-prevalence districts such as North Nias and South Nias, while alternative models substantially overestimated the incidence. Recent applications of SAE in national statistical institutes, such as by ISTAT in Italy, have demonstrated the advantages of model-based estimation in producing efficient and coherent poverty indicators at subnational levels [18]. This reinforces the value of using SAE—particularly flexible models like ZIB-GLMM—in supporting official statistics for SDGs and addressing small sample limitations.

Not all models presented in table 3 include values for RMSE, MAE, and AIC/DIC/WAIC due to methodological and structural differences among estimation approaches. The direct estimation method does not involve a model fitting process, so goodness-of-fit and predictive performance metrics such as AIC, DIC, WAIC, RMSE, or MAE are not applicable. Similarly, the basic SAE binomial INLA model produced unstable estimates for many areas with zero observations, leading to unreliable error metrics. Since the model failed to converge consistently across all areas due to extreme zero inflation, RMSE and MAE were not computed. Additionally, AIC is not directly available for INLA-based models, and DIC values—although technically estimable—were not reliable due to poor model convergence. In contrast, GLMM-based models (Beta-Binomial GLMM, ZIB-GLMM, and ZIB Beta-Binomial GLMM) produced stable predictions across domains, allowing for the calculation of model performance and information criteria. This explains why RMSE, MAE, and AIC are only reported for these models.

ZI-GLMMs offer significant advantages when dealing with sparse count data that contain excess zeros and hierarchical structures [9], as commonly found in extreme poverty datasets. By modeling distinct zero-generating processes and incorporating random effects, Empirical applications of the Zero-Inflated Binomial SAE model for unemployment estimation in Indonesia's Jambi Province, have demonstrated its practical advantages over both direct estimation and synthetic alternatives [8]. The method yielded lower Relative Root Mean Square Error (RRMSE) and standard error, supporting its superior precision. This further underscores the suitability of ZIB-based SAE approaches for sparse outcome variables at small-area levels.

Table 3. Summary of SAE model simulation using multiple approaches

Model	RMSE	MAE	*Zero Inflation*	Overdispersion	AIC/DIC/ WAIC	Description
Direct Estimate	–	–	Not addressed	Not addressed	–	Many areas with RSE > 25 percent, resulting in unstable estimates for small areas
SAE Binomial INLA	–	–	Not addressed	Not addressed	DIC ~ 140–150	Many areas with RSE > 25 percent, resulting in unstable estimates for small areas
SAE ZIB-INLA	0.4566	0.2337	Yes	No	WAIC high	Performs better than Binomial, but still overfits in zero areas
SAE Beta-Binomial GLMM	–	–	No	Yes	AIC = 84.8	Suitable for overdispersed data, but does not explicitly address zero inflation
SAE ZIB-GLMM	0.1088	0.0559	Yes	Yes	AIC = 87.4	Best model; high accuracy, realistic, effectively addresses zero inflation
SAE ZIB Beta-Binomial GLMM	–	–	Yes	Yes	AIC = 89.1	Very complex model with slight improvement; does not outperform ZIB GLMM

Source: Author's estimation results, 2025

Therefore, the ZIB-GLMM is concluded to be the most appropriate model for estimating the proportion of extreme poverty at small area levels. Previous research highlights that small area estimation models incorporating zero inflation—whether Bayesian or frequentist—tend to yield higher accuracy than traditional Empirical Best Linear Unbiased Prediction (EBLUP) or design-based estimators, particularly in contexts characterized by data sparsity, such as poverty mapping. Although the magnitude of improvement may differ by domain, these zero-inflated approaches consistently outperform alternatives across most small-area applications [12], [19]. It offers the best trade-off between model complexity and data fit, as indicated by the lowest AIC.

The ZIB-GLMM model is robust in addressing both zero-inflation and overdispersion while simultaneously accounting for area-level heterogeneity (see table 3). Simulation-based evidence demonstrates that SAE approaches—particularly those incorporating zero-inflated models—outperform design-based methods when estimating area-level means in the presence of excessive zeros [19]. Although frequentist and Bayesian SAE models may exhibit higher bias than traditional methods, they consistently yield greater accuracy, especially under conditions of small sample sizes and zero-inflation [9], [19]. Notably, Bayesian SAE methods tend to be more reliable in populations where the proportion of zero values exceeds 50 percent, as is often the case in extreme poverty data.

Compared to other models (see table 3), the ZIB-GLMM SAE demonstrated superior accuracy, especially in zero-proportion districts, making it the most reliable approach for small area estimation of extreme poverty in North Sumatra Province. Altogether, the ZIB-GLMM model stands out not merely for its statistical performance, but also for its capacity to offer meaningful insights into spatial poverty dynamics in data-scarce environments. Thus, the use of ZIB-GLMM is not only methodologically justified but also essential for practical implementation. Ultimately, this makes ZIB-GLMM not only a statistically robust approach but also a critical tool for spatial poverty diagnostics and evidence-based policymaking in low-resource settings.

ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

3.4 ZIB-GLMM SAE for Estimate Extreme Poverty

The SAE ZIB-GLMM was estimated using the glmmTMB approach. This model integrates fixed effects of auxiliary variables (the proportion of villages/sub-districts with poor sanitation, participation in labor-intensive programs, receipt of direct cash assistance, and areas with malnutrition prevalence) and random effects at the area level to accommodate inter-area heterogeneity.

The results show that the labor-intensive program variable has a significant negative coefficient (-5.2659; $p < 0.001$), indicating that a higher proportion of participation in labor-intensive programs reduces the probability of extreme poverty in an area. This highlights the effectiveness of labor-intensive programs as a government intervention to reduce extreme poverty at the local level.

The direct cash assistance (BLT) variable also has a significant negative coefficient (-16.2789; $p < 0.01$), confirming the substantial role of BLT in reducing extreme poverty. The effectiveness of cash transfer programs as a social safety net is clearly supported by this finding.

In contrast, the malnutrition prevalence variable has a significant positive coefficient (3.2020; $p < 0.001$), indicating that areas with higher prevalence of malnutrition are associated with a higher likelihood of extreme poverty. This reinforces the notion that malnutrition is a key indicator of multidimensional deprivation linked to extreme poverty.

Table 4. Parameter estimates and significance testing results

Variable	Estimate	p-value	Interpretation
Intercept	11.1459	0.05903	Not significant
Toilet (Jamban)	0.4402	0.70259	Not significant
Cash for Work (Padat Karya)	-5.2659	4.42e-06 ***	Significant negative
Cash Assistance (BLT)	-16.2789	0.00783 **	Significant negative
Malnutrition	3.2020	6.14e-05 ***	Significant positive

Source: Author's estimation results, 2025

The poor sanitation variable does not show a statistically significant effect (0.4402; $p = 0.70259$), indicating that in the context of North Sumatra Province in 2024, poor sanitation access is not a primary determinant of extreme poverty, although theoretically it remains part of the multidimensional poverty index.

The model achieved an AIC of 87.4 and BIC of 97.9, the lowest among the tested models (see table 4), indicating good model fit. The standard deviation of the random area effect was 0.1988, reflecting unexplained variability at the area level. Moreover, the integrated zero-inflation component shows a significant intercept of 1.1838 ($p = 0.00662$), suggesting that a proportion of areas statistically report zero extreme poverty households, which does not necessarily reflect the actual absence of extreme poverty in the population [4].

Tzavidis et al. [10] emphasized that defining sensible geographies and estimation targets supported by available data are critical for successful SAE implementation. In addition, adopting parsimonious models, conducting proper model diagnostics, and minimizing the need for additional data sources enhance estimation accuracy.

To ensure robust modeling of count data with excess zeros and potential overdispersion, several best practices have been suggested in the literature [6], [7]. These include assessing the distributional characteristics of the data to detect signs of zero-inflation or overdispersion. Incorporating random effects, especially via random-intercept models in both components of a zero-inflated model, has been recommended to properly account for within-area correlation. Furthermore, model selection should consider multiple fit indices such as AIC and BIC, with preference given to the model with consistently lower values. The nature of zero values should also be evaluated to distinguish between structural and sampling zeros, guiding the decision to use either zero-inflated or hurdle models. Finally, it is crucial to interpret the model results in a subject-specific context, particularly when using random-effects models like GLMMs. These guidelines substantiate the relevance of applying a ZIB-GLMM for estimating extreme poverty in small areas, where zero-inflation and overdispersion are common and domain-level estimates require careful model calibration.

3.5 Evaluation of Model Predictive Performance

As shown in table 5, the comparative evaluation of ZIB model performance was conducted RMSE and MAE as key accuracy metrics. The ZIB-GLMM model substantially reduced prediction errors (RMSE = 0.1088; MAE = 0.0559) relative to the ZIB-INLA model (RMSE = 0.4566; MAE = 0.2336). The improved accuracy of ZIB-GLMM suggests that its random effects structure better accommodates area-level heterogeneity and the presence of excessive zeros, providing more robust and reliable estimates of extreme poverty rates across small areas. These findings are consistent with prior evidence supporting the advantages of mixed models over purely hierarchical Bayesian alternatives in settings with highly sparse count data.

Table 5. Comparative evaluation of model accuracy of ZIB models

Model ZIB	RMSE	MAE
ZIB-INLA	0.4566	0.2336
ZIB-GLMM	0.1088	0.0559

Source: Author's estimation results, 2025

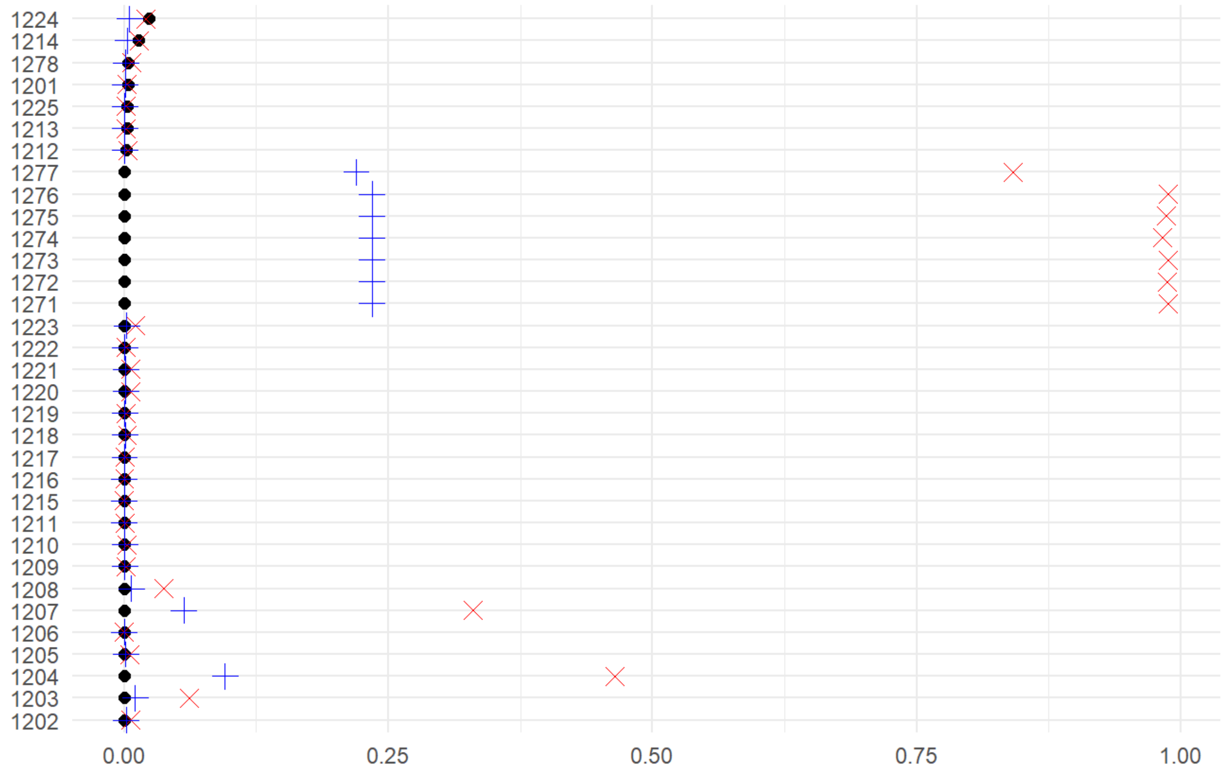

Figure 2. Prediction vs. observation plots of ZIB-INLA and ZIB-GLMM model results

Figure 2 compares the observed and predicted proportions of extreme poverty across 33 districts and cities in North Sumatra Province using the ZIB-INLA and ZIB-GLMM approaches. The x-axis represents the proportion of extreme poverty, while the y-axis lists the districts and cities. The black dots indicate the observed proportions derived from direct estimation, the red crosses represent the predictions from the ZIB-INLA model, and the blue plus symbols represent the predictions from the ZIB-GLMM model.

ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

The figure clearly shows that ZIB-INLA tends to overestimate the prevalence of extreme poverty in areas with zero observed cases, as indicated by the red crosses scattered toward higher values, reaching up to 1.0 in several urban districts such as Sibolga City (1271), Tanjungbalai City (1272), and Tebing Tinggi City (1274). In contrast, ZIB-GLMM predictions are more closely aligned with the observed values, particularly in low-prevalence and zero-observation areas, indicating its better ability to account for excessive zeros and area-level heterogeneity. This improvement in predictive accuracy is consistent with the results presented in table 6, where the ZIB-GLMM model achieved substantially lower prediction errors (RMSE = 0.1088; MAE = 0.0559) compared to the ZIB-INLA model (RMSE = 0.4566; MAE = 0.2337). These findings confirm that ZIB-GLMM provides more stable, realistic, and reliable estimates for small area estimation of extreme poverty, making it a stronger tool to support evidence-based poverty reduction policies at the district level.

The plot highlights the distinct behaviors of the two models. The ZIB-INLA model tends to produce overestimated predictions in areas with zero observed extreme poverty, evident from the high scattering of red crosses along the proportion axis, particularly toward the upper values (up to 1.0). This suggests that ZIB-INLA struggles to fully capture the zero-inflated structure in highly sparse data.

Conversely, the ZIB-GLMM model demonstrates a more consistent alignment with the observed values, especially in low-prevalence and zero-observation areas. The blue plus symbols are concentrated near the black dots, indicating improved prediction accuracy and reduced bias. The incorporation of random effects in the ZIB-GLMM framework allows it to better accommodate area-level heterogeneity and account for excessive zeros, thus reducing overestimation tendencies.

Overall, the visual comparison reinforces the superior performance of the ZIB-GLMM model over ZIB-INLA. The results support previous simulation findings, where ZIB-GLMM achieved lower RMSE and MAE metrics, making it the most both statistically reliable and practically robust model for small area estimation of extreme poverty in the presence of zero inflation and overdispersion.

The detailed comparison of predicted extreme poverty proportions across the 33 districts and cities in North Sumatra Province further confirms the superior performance of the ZIB-GLMM model. As shown in table 6, the majority of areas (26 out of 33) recorded zero observations of extreme poverty, which highlights the severe zero-inflation characteristic of the dataset. This presents a significant challenge for standard estimation methods.

In areas with zero observed values, the predictions from the ZIB-INLA model were highly unstable and prone to severe overestimation, often exceeding 0.9. For example, ZIB-INLA produced predicted values of 0.988, 0.987, and 0.982 for areas 1271 (Sibolga City), 1272 (Tanjungbalai City), and 1274 (Tebing Tinggi City), respectively, where no extreme poverty cases were observed. In contrast, the ZIB-GLMM model generated more reasonable estimates for these same areas, with predicted values consistently around 0.234, better capturing the underlying uncertainty while avoiding excessive overestimation

The improvement in predictive performance was also evident in areas where extreme poverty was actually observed. In these cases, the ZIB-GLMM model consistently produced estimates closer to the observed values than the ZIB-INLA model. For example, in area 1212 or Deli Sedang City (observed = 0.00193), ZIB-GLMM yielded a prediction of 0.0006, compared to 0.0032 from ZIB-INLA.

Table 6. Direct estimation vs. estimation of ZIB-INLA and ZIB-GLMM model results

Area	Direct Estimation	ZIB-INLA Estimation	ZIB-GLMM Estimation
1201/Nias	0.00367	0.0024	0.0007
1202/Mandailing Natal	0	0.0064	0.0017
1203/Tapanuli Selatan	0	0.0617	0.0147
1204/Tapanuli Tengah	0	0.4640	0.0954
1205/Tapanuli Utara	0	0.0054	0.0012
1206/Toba	0	0.0001	0
1207/Labuhan Batu	0	0.3300	0.0565

Area	Direct Estimation	ZIB-INLA Estimation	ZIB-GLMM Estimation
1208/Asahan	0	0.0371	0.0068
1209/Simalungun	0	0.0013	0.0003
1210/Dairi	0	0.0023	0.0006
1211Karo	0	0.0023	0.0006
1212/Deli Serdang	0.00193	0.0032	0.0006
1213/Langkat	0.00253	0.0017	0.0005
1214/Nias Selatan	0.0133	0.0144	0.0033
1215/Humbang Hasundutan	0	0.0001	0
1216/Pakpak Bharat	0	0.0002	0
1217/Samosir	0	0.0088	0.0023
1218/Serdang Bedagai	0	0.0029	0.0007
1219/Batu Bara	0	0.0013	0.0004
1220/North Padang Lawas	0	0.0058	0.0016
1221/Padang Lawas	0	0.0057	0.0012
1222/South Labuhanbatu	0	0.0012	0.0003
1223/North Labuhanbatu	0	0.0104	0.0022
1224/North Nias	0.0231	0.0201	0.0055
1225/West Nias	0.00255	0.0015	0.0004
1271/Sibolga City	0	0.9880	0.2340
1272/Tanjungbalai City	0	0.9870	0.2340
1273/Pematangsiantar City	0	0.9870	0.2340
1274/Tebing Tinggi City	0	0.9820	0.2340
1275/Medan City	0	0.9870	0.2340
1276/Binjai City	0	0.9880	0.2340
1277/Padangsidimpuan City	0	0.8420	0.2200
1278/Gunung Sitoli City	0.00376	0.0066	0.0013

Source: Author's estimation results, 2025

Table 6 presents a comparison between direct estimates and model-based estimates of extreme poverty proportions at the district/city level in North Sumatra Province for 2024. The direct estimation column represents values obtained from the *Foster-Greer-Thorbecke (FGT)* formula using Susenas microdata. These figures reflect survey-based calculations and serve as the observed baseline for each area. The ZIB-INLA estimation and ZIB-GLMM estimation columns display the model-generated estimations based on the SAE framework. These predictions incorporate auxiliary variables from Podes 2024 and are statistically adjusted to account for zero inflation and small sample sizes, thereby improving the stability of the estimates.

Overall, the ZIB-GLMM model achieved substantial error reduction, as reflected in its lower RMSE = 0.1088 and MAE = 0.0559, compared to the ZIB-INLA model (RMSE = 0.4566; MAE = 0.2337). This suggests that the random effects structure of ZIB-GLMM effectively accommodated area-level heterogeneity and the presence of excess zeros. The difference between these two is most visible in districts with zero observations from the survey. While the direct estimates are zero due to sampling limitations, the model-based estimates produce non-zero estimations by borrowing strength from auxiliary variables and other areas. This allows for more realistic and robust poverty estimation at the small area level.

These findings further reinforce the conclusion that ZIB-GLMM is the most reliable and robust model for small area estimation of extreme poverty in North Sumatra Province. Its ability to avoid overfitting in data-sparse zones while accurately capturing the variation in areas with observed extreme poverty cases makes it a critical tool for evidence-based poverty diagnostics and policy formulation. The district-level estimates produced by the ZIB-GLMM model can support targeted interventions, such as prioritizing social assistance programs in districts with high predicted extreme poverty rates and improving the allocation of poverty alleviation resources.

4. Conclusion

This study demonstrates the effectiveness of using the Zero-Inflated Binomial Generalized Linear Mixed Model (ZIB-GLMM) within the Small Area Estimation (SAE) framework to estimate extreme poverty at the district and city levels in North Sumatra Province for the year 2024. By addressing key statistical challenges—namely zero inflation due to sparse microdata and overdispersion in binary poverty indicators—the ZIB-GLMM approach outperformed both direct estimation and other SAE models, including binomial-INLA, ZIB-INLA, beta-binomial GLMM, and ZIB beta-binomial GLMM. The incorporation of contextual auxiliary variables derived from the 2024 Village Potential Statistics (PODES) significantly improved the model's predictive power, particularly in domains lacking sufficient survey samples. The findings confirm that ZIB-GLMM offers more accurate, stable, and realistic small area estimates of extreme poverty, thereby providing critical input for evidence-based policy-making and targeted poverty alleviation programs. This methodology not only contributes to the refinement of official poverty statistics but also supports Indonesia's commitment to eradicating extreme poverty, as emphasized in national development agendas and SDGs Target 1.1.

However, this research has several limitations. First, the estimation relies on cross-sectional survey data, which limits the ability to capture temporal dynamics of extreme poverty. Second, the auxiliary variables are constrained by the available PODES indicators and may not fully reflect multidimensional deprivation. Third, the analysis focuses on area-level estimation, without explicitly modeling spatial dependencies. Future research may address these limitations by incorporating panel or longitudinal data to capture poverty dynamics, expanding the set of auxiliary variables using administrative or geospatial sources, and applying spatially explicit SAE models to better account for geographic dependencies. Moreover, the approach could be extended to other regions or to multidimensional poverty indicators with similar data challenges, further strengthening its contribution to official poverty statistics and evidence-based policymaking.

References

[1] World Bank, "Poverty and Shared Prosperity 2016: Taking on Inequality," 2016. doi: 10.1596/978-1-4648-0958-3.

[2] Badan Amil Zakat Nasional, "Peta Kemiskinan Ekstrem Nasional," *Puskas BAZNAS*, pp. 1–33, 2024, [Online]. Available: https://www.puskasbaznas.com/publications/published/officialnews/1863-peta-kemiskinan-ekstrem-nasional

[3] *Instruksi Presiden Republik Indonesia Nomor 8 Tahun 2025 Tentang Ommalisasi Pelaksanaan Pengentasan Kemiskinan Dan Penghapusan Kemiskinan Ekstrem*. 2025.

[4] Y. Ramadhan and A. Ubaidillah, "Pendugaan Persentase Penduduk Miskin Ekstrem Menggunakan Small Area Estimation dengan Partitioning Around Medoids Clustering," *Semin. Nas. Off. Stat.*, vol. 2023, no. 1, pp. 557–568, 2023, doi: 10.34123/semnasoffstat.v2023i1.1717.

[5] C. B. Dean and E. R. Lundy, "Overdispersion," *Wiley StatsRef Stat. Ref. Online*, pp. 1–9, 2016, doi: 10.1002/9781118445112.stat06788.pub2.

[6] B. Pittman, E. Buta, K. Garrison, and R. Gueorguieva, "Models for Zero-Inflated and Overdispersed Correlated Count Data: An Application to Cigarette Use," *Nicotine Tob. Res.*, vol. 25, no. 5, pp. 996–1003, 2023, doi: 10.1093/ntr/ntac253.

[7] L. Li, L. Lin, J. C. Cappelleri, H. Chu, and Y. Chen, "ZIBGLMM: Zero-inflated bivariate generalized linear mixed model for meta-analysis with double-zero-event studies," *Res. Synth. Methods*, vol. 16, no. 2, pp. 251–271, 2025, doi: 10.1017/rsm.2024.4.

[8] B. Hartono, A. Kurnia, and Indahwati, "Zero Inflated Binomial Models in Small Area Estimation with Application to Unemployment Data in Indonesia," *IJCSN -International J. Comput. Sci. Netw.*, vol. 6, no. 65, pp. 746–752, 2017.

[9] L. P. Fávero, J. F. Hair, R. de F. Souza, M. Albergaria, and T. V. Brugni, "Zero-inflated generalized linear mixed models: A better way to understand data relationships," *Mathematics*, vol. 9, no. 10, pp. 1–28, 2021, doi: 10.3390/math9101100.

[10] N. Tzavidis, L. C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla, "From start to finish: a framework for the production of small area official statistics," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 181, no. 4, pp. 927–979, 2018, doi: 10.1111/rssa.12364.

[11] N. Permatasari and Widya Larasati, "Perbandingan Metode SAE EBLUP dan SAE HB Pada Pendugaan Area Kecil (Studi Kasus Pendugaan Kemiskinan di Provinsi Jawa Timur)," *J. Stat. dan Apl.*, vol. 6, no. 1, pp. 96–108, 2022, doi: 10.21009/jsa.06109.

[12] S. Krieg, H. J. Boonstra, and M. Smeets, "Small-area estimation with zero-inflated data – a simulation study," *J. Off. Stat.*, vol. 32, no. 4, pp. 963–986, 2016, doi: 10.1515/JOS-2016-0051.

[13] D. A. Puspitasari *et al.*, "Pendugaan Area Kecil Persentase Anak Usia 0-17 Tahun yang Hidup di Bawah Garis Kemiskinan Tingkat Kabupaten/Kota di Pulau Jawa Tahun 2023," *Semin. Nas. Off. Stat.*, vol. 2024, no. 1, pp. 31–38, 2024, doi: 10.34123/semnasoffstat.v2024i1.1967.

[14] A. Bikauskaite, I. Molina, and D. Morales, "Multivariate mixture model for small area estimation of poverty indicators," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 185, no. S2, pp. S724–S755, 2022, doi: 10.1111/rssa.12965.

[15] P. Priatmadani, P. P. Sari, E. N. Rahmat, P. B. Aji, F. A. Nafiis, and N. Istiana, "Small Area Estimation of Maluku and Papua Island Child Poverty Levels in 2023," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 46–61, 2024, doi: 10.20956/j.v21i1.35293.

[16] Z. O. Briantiko and R. Agustina, "Aplikasi Small Area Estimation Pada Penghitungan Nilai Estimasi Indikator Imunisasi Dasar Lengkap Di Pulau Jawa Dan Bali Tahun 2020," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 449–458, 2022, doi: 10.34123/semnasoffstat.v2022i1.1551.

[17] S. J. Adwendi, A. Saefuddin, and B. Susetyo, "a Study of Small Area Estimation To Measure Multidimensional Poverty With Mixed Model Poisson, Zip, and Zinb," *Barekeng*, vol. 17, no. 1, pp. 439–448, 2023, doi: 10.30598/barekengvol17iss1pp0439-0448.

[18] M. D'Alò, D. Filipponi, S. Gerosa, and F. Isidori, "Small Area Estimation of Poverty Indicators," *Riv. Ital. di Econ. Demogr. e Stat.*, vol. LXXIX, no. 2019, pp. 131–142, 2025, doi: 10.71014/sieds.v79i3.345.

[19] K. Sadik, R. Anisa, and E. Aqmaliyah, "Small area estimation on zero-inflated data using frequentist and bayesian approach," *J. Mod. Appl. Stat. Methods*, vol. 18, no. 1, pp. 1–22, 2019, doi: 10.22237/jmasm/1582727606.