Scientific Journal of Informatics
Vol. 11, No. 3, Aug 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Modified Mixed Effects Random Forest in Small Area Estimation Using PCA and Rotation Forest with Correlated Auxiliary Variables

Rizki Ananda[1], Khairil Anwar Notodiputro[2*], Muhammad Nur Aidi[3]
[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences,
IPB University, Indonesia

Abstract.
Purpose: The per capita expenditure data in Jambi Province, Indonesia have been plagued with severe multicollinearity problems. To address the issue, this study developed an effective small area estimation (SAE) method, which is essential for formulating comprehensive regional development policies in Jambi Province. By modifying the mixed effects random forest (MERF) method, we introduced PCA-MERF (which applies principal component analysis prior to MERF) and MERoF (which replaces the standard random forest with rotation forest) to handle multicollinearity more effectively. Data from the National Socioeconomic Survey (Susenas) in March 2021 and Village Potential (PODES) in 2021 were utilized. The methods were evaluated using metrics such as root mean square error (RMSE), relative root mean square error (RRMSE), coefficient of variation (CV), and their ability to capture random area effects. The random effect block (REB) bootstrap approach was employed to obtain MSE estimates for evaluating area-level estimate quality.
Result: The results showed that MERoF outperformed both MERF and PCA-MERF, particularly in unit-level (village) estimation. Additionally, MERoF demonstrated superior capability in capturing variation between subdistricts compared to MERF and PCA-MERF. PCA-MERF performed better than MERF and MERoF at the area level (subdistrict). All three methods showed acceptable performance with RRMSE and CV values ranging between 8% and 10%, indicating precise and reliable predictions for per capita expenditure in small areas. These modifications to MERF prove effective and advantageous for small-area estimation in datasets with significant multicollinearity.
Novelty: This research introduces a novel semi-parametric, tree-based SAE approach, enhancing the precision of per capita expenditure estimates and supporting more informative regional policy decisions, thus filling a gap in current SAE methodologies.

Keywords: Tree-based method, Generalized linear mixed models, Multicollinearity, Poverty, Statistics Indonesia, Jambi Province
Received July 2024 / **Revised** August 2024 / **Accepted** August 2024

INTRODUCTION

The issue of multicollinearity is often encountered in empirical data, including per capita expenditure data. Multicollinearity is a phenomenon where two or more independent variables in a regression model are strongly correlated, which can lead to problems in interpreting and estimating the reliability of model parameters [1]. Variables related to per capita expenditure, such as household demographic conditions, availability of infrastructure facilities, or the economic conditions of a region, often have high correlations with each other. This issue presents a significant challenge in modeling the estimation of per capita expenditure. On the other hand, per capita expenditure data in Indonesia are only officially available at the province and regency/municipality levels as the data at these levels are considered to be more precise and accountable. However, to achieve comprehensive regional development goals, the government urgently needs per capita expenditure data at smaller area levels, such as villages and subdistricts. Estimating parameters at the village and subdistrict levels remains challenging due to the extremely small sample size of survey data sources at these levels. Therefore, small area estimation (SAE) is often used in survey data analysis to estimate parameters in areas with relatively small or even non-existent sample sizes by utilizing additional information from different data sources.

*Corresponding author.
Email addresses: rizkiananda@apps.ipb.ac.id (Ananda), khairil@apps.ipb.ac.id (Notodiputro)*, muhammadai@apps.ipb.ac.id (Aidi)
DOI: 10.15294/sji.v11i3.10633

Estimating parameters in small areas can be carried out through the utilization of additional information from outside the area, within the area itself, and outside the survey. Indirect estimation can be carried out using values from observed variables from connected areas, thereby increasing the effectiveness of the sample size [2]. This is in line with Kurnia [3] and Kurnia et al. [4], who stated that small area estimation methods essentially leverage the strength of surrounding areas and data sources outside the area for which statistics are desired.

Fay [5] proposed a mixed model approach to achieve more accurate estimates at smaller area levels with limited sample data. This approach improves precision by utilizing information from different data sources. Standard SAE methods generally use linear mixed models (LMM), which consist of fixed effects and random effects. Fixed effects represent constant parameters across the population or group, derived from the same covariate variables for all small areas. Random effects account for the variability of observed variables among small areas that the fixed effects from additional information cannot explain. These random effects provide varying contributions for each small area, helping to produce more precise estimates for those areas.

Hajjem et al. [6] developed the mixed effects random forest (MERF) method to address clustered or structured data. This method combines the predictive power of random forest (RF) with mixed effects models, which excel in handling clustered or hierarchically structured data. This approach enhances prediction accuracy in situations where observations are not independent but are related within specific clusters. This research is extended by adapting the MERF method into the domain of small area estimation [7], [8], where SAE is closely associated with mixed effects models to handle clustered or hierarchically structured data. By considering both random and fixed effects, the MERF method provides more reliable estimates when observational data are related within geographic or administrative clusters.

The predictive performance in SAE depends heavily on the validity of model assumptions. One advantage of design-based estimation is its independence from model assumptions, rendering it unaffected by model misspecification [9]. In terms of assumptions, Krennmair and Schmid [7] only demonstrated the reliability of MERF in non-linear and non-symmetric problems. However, their research did not explicitly explore MERF's reliability under other conditions, such as multicollinearity. Therefore, this study aims to assess the reliability of MERF in the presence of multicollinearity within the SAE framework to determine whether MERF is a viable solution or if new methods are necessary for improved outcomes.

One approach for addressing multicollinearity is through principal component analysis (PCA) [10]. PCA identifies patterns and relationships among variables in complex datasets by transforming a set of correlated variables into a set of uncorrelated components. These principal components are linear combinations of the original variables, ordered by the variance they explain. PCA effectively reduces dimensionality without significant information loss and helps identify correlations among variables, potentially eliminating redundant variables.

Rodriguez et al. [11] applied the concept of PCA in constructing classifier ensembles by rotating the axes of variables used to build decision trees, a method known as rotation forest (RoF). Despite the use of PCA, all principal components are retained to construct decision trees, preserving the completeness of the data. Many studies have shown that the performance of rotation forest (RoF) is significantly better than that of random forest (RF). Rotation forest (RoF) builds decision trees that are independent of each other. The resulting trees are more diverse, and the models produced by the RoF algorithm are more stable and accurate [12], [13], [14]. In addition to classification, Pardo et al. [15] also developed RoF for regression estimation. In regression, RoF's performance remains excellent, similar to its performance in classification problems. RoF also provides the best performance with a smaller average RRMSE value in multi-target regression compared to bagging and random forest [16].

Based on the above, this study aims to modify the forest method in MERF in two ways. First, the study applies PCA to the variables used in MERF (hereinafter referred to as “PCA-MERF”) to produce more independent trees. Additionally, the study modifies MERF by replacing the standard random forest with a rotation forest, which preserves the completeness of all data information as fixed effects in the mixed model (hereinafter referred to as “MERoF”). These modifications aim to create more robust and reliable methods for SAE, particularly in handling multicollinearity issues.

Many studies in Indonesia have conducted small area estimations for per capita expenditure using various models [17], [18], [19], [20], [21], [22]. Nevertheless, no study has yet performed estimation using a tree-based semi-parametric approach that efficiently addresses non-linearity, non-parametric, and multicollinearity issues simultaneously. Based on these considerations, this research is expected to provide an effective small area estimation method for various empirical data problems and serve as a foundation for more informed and effective regional policy decisions.

METHODS

Data description

The data utilized in this study were sourced from Statistics Indonesia (Badan Pusat Statistik), specifically the National Socioeconomic Survey (Susenas) in March 2021 and the 2021 Village Potential (Podes) data for Jambi Province. The 2021 data were selected to represent the most recent dataset available for Podes. Susenas data served as the survey data, characterized by a small sample size, covering only 0.76% of the household population across 40% of the villages. Using Susenas data, small area estimation was performed on the variable of interest, namely the average per capita expenditure. The Podes data were utilized as registry data, providing population data as additional auxiliary variables in the estimation process. The variables are detailed in Table 1 below.

Table 1. Research variables

Variable	Description	Data Scale	Data Source
Avg_PCE	Average Per Capita Expenditure of Villages (000 IDR)	Continuous Ratio	Susenas
Num_Families	Number of Families in the Village	Discrete Ratio	Podes
Pct_Fam_PLN	Percentage of Families Using PLN Electricity in the Village	Continuous Ratio	Podes
Pct_Fam_NonPLN	Percentage of Families Using Non-PLN Electricity in the Village	Continuous Ratio	Podes
Pct_Fam_SlumAreas	Percentage of Families Living in Slum Areas in the Village	Continuous Ratio	Podes
Edu_Fac_Ratio	Ratio of Educational Facilities to the Number of Families in the Village	Continuous Ratio	Podes
Health_Fac_Ratio	Ratio of Health Facilities to the Number of Families in the Village	Continuous Ratio	Podes
Poverty_Cert_Ratio	Ratio of Poverty Certificates to the Number of Families in the Village	Continuous Ratio	Podes
Pct_Fam_Landline	Percentage of Families with Landline Phones in the Village	Continuous Ratio	Podes
IMK_Ratio	Ratio of Micro and Small Industries (IMK) to the Number of Families in the Village	Continuous Ratio	Podes
Distance_to_SubDistrict	Distance from the Village to the Subdistrict (km)	Continuous Ratio	Podes
Transport_Cost	Transportation Costs from the Village to the Subdistrict (000 IDR)	Continuous Ratio	Podes
Eco_Fac_Ratio	Ratio of Economic Facilities to the Number of Families in the Village	Continuous Ratio	Podes

Mixed effects random forest (MERF)

Krennmair and Schmid [6] adopted the MERF method in small area estimation (SAE), resulting in excellent predictive performance for small areas. In general, MERF is similar to a linear mixed model (LMM) but the fixed effect $X_i\beta$ is replaced with the random forest function $f(X)$ to estimate the fixed effect coefficients.

$$y_{ij} = f(X_{ij}) + Z_i v_i + \epsilon_{ij} \quad (1)$$

In this study, $y_{ij} = \left[y_{i1}, \dots, y_{in_i}\right]^T$ is the target variable vector representing the average per capita expenditure in the j-th village of the i-th subdistrict, with a size of n_i x 1, where n_i is the number of village observation in the i-th subdistrict ($j = 1, \dots, n_i$); $X_{ij} = \left[x_{i1}, \dots, x_{in_i}\right]^T$ is the covariate matrix from the j-th village, with a size of n_i x p, where p is the number of covariates and $f(X_{ij})$ is the fixed effect estimated by the random forest; $Z_i = \left[z_{i1}, \dots, z_{in_i}\right]^T$ is the random effect covariate matrix with a size of n_i x q, where q is the dimension of the random effect, usually consisting of a subset of the covariates X_{ij}; $v_i =$

$[v_{i1}, \ldots, v_{iq_i}]^T$ is the random effect vector of the i-th subdistrict with a size of q x 1. The random effect $Z_i v_i$ was assumed to be linear. This study used only a random intercept, and therefore Z_i became a (n_i x 1) vector of 1s. $\epsilon_{ij} = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^T$ is the error vector of village unit observations with a size of n_i x 1. Observations between subdistrict areas were assumed to be independent, and v_i and ϵ_{ij} were mutually independent and normally distributed, with variance-covariance matrices H_i for the random effects of each i-th subdistrict area and R_{ij} or the error of the j-th village unit observation in the i-th subdistrict. The covariance matrix of the observations y is $Cov(y) = V = diag_{1 \le i \le D}(V_i)$, with

$$V_i = Z_i H_i Z_i^T + R_{ij} \tag{2}$$

To model equation (1), the expectation-maximization (EM) algorithm approach performed by Hajjem et al. [6] was used. MERF iteratively estimates the function of the forest by assuming that the random effect components are correct and then estimate the random effect components by assuming that the out-of-bag (OOB) predictions of the forest are correct. OOB predictions use observations not included in the construction of each tree in the forest [23], [24] and can demonstrate the classification ability, feature importance, and other dataset patterns of the random forest [25], [26].

The MERF algorithm was outlined as follows [6, Sec 2.2]:

1) Iteration $b = 0$ was set and the random component $\hat{v}_{(0)} = 0$ was initialized.
2) For each iteration $b = b + 1$, $\hat{f}(X)_{(b)}$ and $\hat{v}_{(b)}$ were updated as follows:
   (a) The new target variable $y^*_{(b)}$ was computed by subtracting the contribution of the random effects from the previous iteration from the original target variable y.

$$y^*_{(b)} = y - Z\hat{v}_{(b-1)} \tag{3}$$

   (b) The function $\hat{f}()_{(b)}$ was estimated by training a random forest model with the new target variable $y^*_{(b)}$ and covariates X. Note that $\hat{f}()_{(b)}$ was the same function for all areas i.
   (c) Out-of-bag (OOB) predictions: $\hat{f}(X)^{OOB}_{(b)}$ was obtained, namely the predictions from $\hat{f}()_{(b)}$ on units not used to train the random forest.
   (d) LMM modeling was performed using the OOB predictions:

$$y = \hat{f}(X)^{OOB}_{(b)} + Z\hat{v}_{(b)} + \epsilon \tag{4}$$

   This model assumed $\hat{f}(X)^{OOB}_{(b)}$ as the fixed effect, $Z\hat{v}_{(b)}$ as the random effect, and ϵ as the residual error from the observations.
   (e) The variance components from the trained LMM model were obtained, namely the residual error ($\hat{\sigma}^2_{\epsilon,(b)}$) and the covariance matrix of the random effects ($\hat{H}_{(b)}$) and estimate the random effects with:

$$\hat{v}_{(b)} = \hat{H}_{(b)} Z^T \hat{V}^{-1}_{(b)} (y - \hat{f}(X)^{OOB}_{(b)}) \tag{5}$$

3) Step 2 was repeated until the change in model parameters between successive iterations was below a certain threshold, indicating that the algorithm had converged. In this study, the convergence criterion is a marginal change in the generalized log-likelihood (GLL) of less than $1e^{-4}$.

The estimation of $\hat{\sigma}^2_{\epsilon}$ was naive and thus could not be considered a valid estimator for the variance σ^2_{ϵ} of the unit-level error ϵ. Consequently, this estimator was corrected for bias to obtain the residual variance σ^2_{ϵ} from the random forest model through a bootstrap approach with the following steps [7], [27]:

1) Out-of-bag (OOB) predictions $\hat{f}(X_{ij})^{OOB}$ was utilized from the model that had achieved convergence.
2) Bootstrap samples $y^*_{(b)} = \hat{f}(X)^{OOB} + \epsilon^*_{(b)}$ were generated using $\epsilon^*_{(b)}$ where $\epsilon^*_{(b)}$ was sampled with replacement from the centered marginal residuals $\hat{e} = y - \hat{f}(X)^{OOB}$.
3) For each bootstrap sample, new OOB predictions $\hat{f}(X)^{OOB}_{(b)}$ were computed by training a random forest model using $y^*_{(b)}$ as the dependent variable.
4) Bias correction $K(\hat{f})$ was estimated using:

$$K(\hat{f}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{f}(X)^{OOB} - \hat{f}(X)^{OOB}_{(b)}\right)^2 \quad (6)$$

This correction represented the average squared difference between the convergent OOB predictions and the OOB predictions from the bootstrap samples.
5) The bias-corrected estimate of the residual variance was given by:

$$\hat{\sigma}^2_{bc,\epsilon} = \hat{\sigma}^2_{\epsilon} - K(\hat{f}) \quad (7)$$

where $\hat{\sigma}^2_{\epsilon}$ is the naïve estimate of the residual variance, and $K(\hat{f})$ is the bias correction term. This procedure provided a more accurate estimate of the residual variance by correcting for bias introduced by the random forest model.

After obtaining the unit-level estimates, these estimates were used to estimate the averages for each area by leveraging additional information from census or administrative data. The average estimate for each area i was computed as follows:

$$\bar{\hat{f}}(X_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{f}(X_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{f}(x_{ij}) \quad (8)$$

where N_i is the number of units in area i, and $\hat{f}(x_{ij})$ represents the predicted value for the j-th unit in area i.

The final value of the best linear unbiased predictor (BLUP) $\hat{v}_i$ obtained from equation (5) was used to estimate the contribution of random effects for each area i. This component added a correction for random effects to the average fixed predictions to provide a more accurate estimate for the area average:

$$\hat{\mu}_i = \bar{\hat{f}}(X_i) + Z_i\hat{v}_i \quad (9)$$

For areas without sample data, the estimate was based solely on the fixed effect component of the random forest model, which is the average value of the fixed predictions:

$$\hat{\mu}_i = \bar{\hat{f}}(X_i) \quad (10)$$

This ensured that estimates could still be made even in the absence of sample data from the area.

To evaluate the area-level estimates, this method measures the quality of the estimates by calculating the uncertainty of the indicator at the area level through a bootstrap scheme. Chambers and Chandra [28] introduced random effect block (REB) scheme to estimate the mean squared error (MSE) of the area-level estimates. The steps for the REB bootstrap for this study were as follows [6, Sec 3]:
1) The vector of marginal residuals was calculated: $\widehat{e_{ij}} = y_{ij} - \hat{f}(X_{ij})$
2) Level-2 residuals for each area were computed based on the marginal residuals $\widehat{e_{ij}}$:

$$\overline{r_i} = \frac{1}{n_i}\sum_{j=i}^{n_i}\widehat{e_{ij}} \quad (11)$$

where $\bar{r} = [\overline{r_i}, \ldots, \overline{r_D}]'$ is a vector D x 1 of level-2 residuals.
3) Marginal residuals were used to compute level-1 residual vectors: $\widehat{r_{ij}} = \widehat{e_{ij}} - 1_{n_i}\overline{r_i}$

The residuals $\widehat{r_{ij}} = \left[\widehat{r_{1j}}', \ldots, \widehat{r_{Dj}}'\right]'$ were scaled to the bias-corrected residual variance and centered, denoted as $\widehat{r_{ij}}^{c} = \left[\widehat{r_{1j}}^{c'}, \ldots, \widehat{r_{Dj}}^{c'}\right]'$. The level-2 residuals $\bar{r}_i$ were also scaled to the estimated variance $\widehat{H}_i = \hat{\sigma}_v^2$ and centered, denoted as $\bar{r}_i^{c} = [\bar{r}_1^{c}, \ldots, \bar{r}_D^{c}]'$

4) Bootstrap was performed: For $b = 1, \ldots, B$.
   (a) Randomly sampled with replacement from the level-1 and level-2 residuals:

$$r_{ij}^{(b)} = srswr\left(\widehat{r_{ij}}^{c}, n_i\right) \text{ and } \bar{r}_i^{(b)} = srswr(\bar{r}_i^{c}, D)$$

   (b) The bootstrap population was calculated as: $y_{ij}^{(b)} = \hat{f}\left(X_{ij}\right) + Z\bar{r}_i^{(b)} + r_{ij}^{(b)}$ and the area-level true mean of the bootstrap population was computed as: $\mu_i{}^{(b)} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}^{(b)}$ for all areas i.
   (c) Each bootstrap population was sampled with size n_i equal to the original sample size. Then, the sample was used to estimate $\hat{f}^{(b)}()$ and $\hat{v}^{(b)}$ as described in the MERF algorithm.
   (d) The area-level mean from the bootstrap sample was calculated as follows:

$$\hat{\mu}_i{}^{(b)} = \bar{\hat{f}}^{(b)}(X_i) + Z_i\hat{v}_i^{(b)} \tag{12}$$

   These steps helped assess the uncertainty in the area-level estimates by using a robust bootstrap approach that accounts for the hierarchical structure of the data.
   (e) Using B bootstrap samples, the mean squared error (MSE) estimate for the area-level prediction was obtained as follows:

$$\widehat{MSE}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\mu_i^{(b)} - \hat{\mu}_i^{(b)}\right)^2 \tag{13}$$

Principal component analysis-mixed effects random forest (PCA-MERF)

Principal component analysis (PCA) is a statistical technique used to reduce the dimensionality of data by transforming a large set of variables into a smaller set of principal components that retain most of the variation present in the original data. When combined with random forest, PCA is applied as a preprocessing step before training the random forest model. This approach helps eliminate multicollinearity by converting the original features into orthogonal (uncorrelated) principal components. By reducing dimensionality and addressing multicollinearity, the random forest model can concentrate on the most informative features, potentially enhancing the model's performance.

The algorithm for PCA-MERF for this study was outlined as follows:

1) Data were standardized to ensure that each covariate had a mean of 0 and a variance of 1.
2) Principal component analysis (PCA) was performed on the standardized covariates to obtain principal components that are orthogonal to each other.
3) The original covariates were projected into the space defined by the principal components.
4) The projected covariates were used to train the random forest model as the fixed effects component, which was subsequently combined with linear mixed model (LMM) to estimate the random effects components. This estimation was carried out using the expectation-maximization (EM) approach, similar to the method used by [7] in MERF.
5) The steps for estimating unit-level values, area-level means, and mean squared error (MSE) using the REB bootstrap approach, as outlined by [7].

Mixed effects rotation forest (MERoF)

Rotation forest (RoF) is an ensemble method that utilizes principal component analysis (PCA) to rotate the feature axes upon which decision trees are built. Decision trees are used as the basis for classification due to their sensitivity to feature axis rotation while maintaining accuracy. Although PCA is employed, all principal components are used to construct the decision trees, ensuring that the completeness of the data information is preserved [11].

Pardo et al. [15] developed rotation forest (RoF) for regression estimation, where previous methods were only for classification. The adapted RoF algorithm for regression was as follows:

1) We let $x = [x_1, \dots, x_p]^T$ be the observation vector with p variables, X be the dataset consisting of vectors x size n x p, and F be the set of p variables. We also let $y = [y_1, \dots, y_n]^T$ be the response vector of size n x 1.
2) F was randomly divided into k disjoint groups of variables, each with approximately the same number of varibles (m) . $F_{i,j}$ is the group of variables used to build tree D_i with $i = 1, \dots, U$ and m_j variables, for $j = 1, \dots, k$. $X_{i,j}$ is the dataset X with variables $F_{i,j}$.
3) Bootstrap sampling was performed on the dataset $X_{i,j}$. The bootstrap sample of the dataset was denoted as $X^*_{i,j}$.
4) Principal component analysis (PCA) was performed on $X^*_{i,j}$ and the principal component coefficient was saved as $a_{i,j}^{(1)}, a_{i,j}^{(2)}, \dots, a_{i,j}^{(m_j)}$.
5) The principal component coefficient vectors were arranged into a rotation matrix R_i as follows:

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, \dots, a_{i,1}^{(m_1)} & [0] & \dots & [0] \\ [0] & a_{i,2}^{(1)}, \dots, a_{i,2}^{(m_2)} & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & a_{i,k}^{(1)}, \dots, a_{i,k}^{(m_k)} \end{bmatrix} \quad (14)$$

6) The columns of the matrix R_i were reordered to match the arrangement of the variable groups. The reordered rotation matrix was denoted as R_i^a which is p x p.
7) The i-th regression tree (D_i) was built using (XR_i^a, y).
8) Steps 2 through 7 were repeated until U regression trees were obtained.

The steps outlined for RoF were consistently applied in the estimation of forest functions when modeling mixed effects rotation forest (MERoF). This approach ultimately enabled the estimation of fixed and random effects components, unit-level and area-level mean, and mean squared error (MSE) as a measure of estimation uncertainty. All steps were conducted according to the procedures described in [7].

The modeling and data analysis procedure in this study is presented in Figure 1.

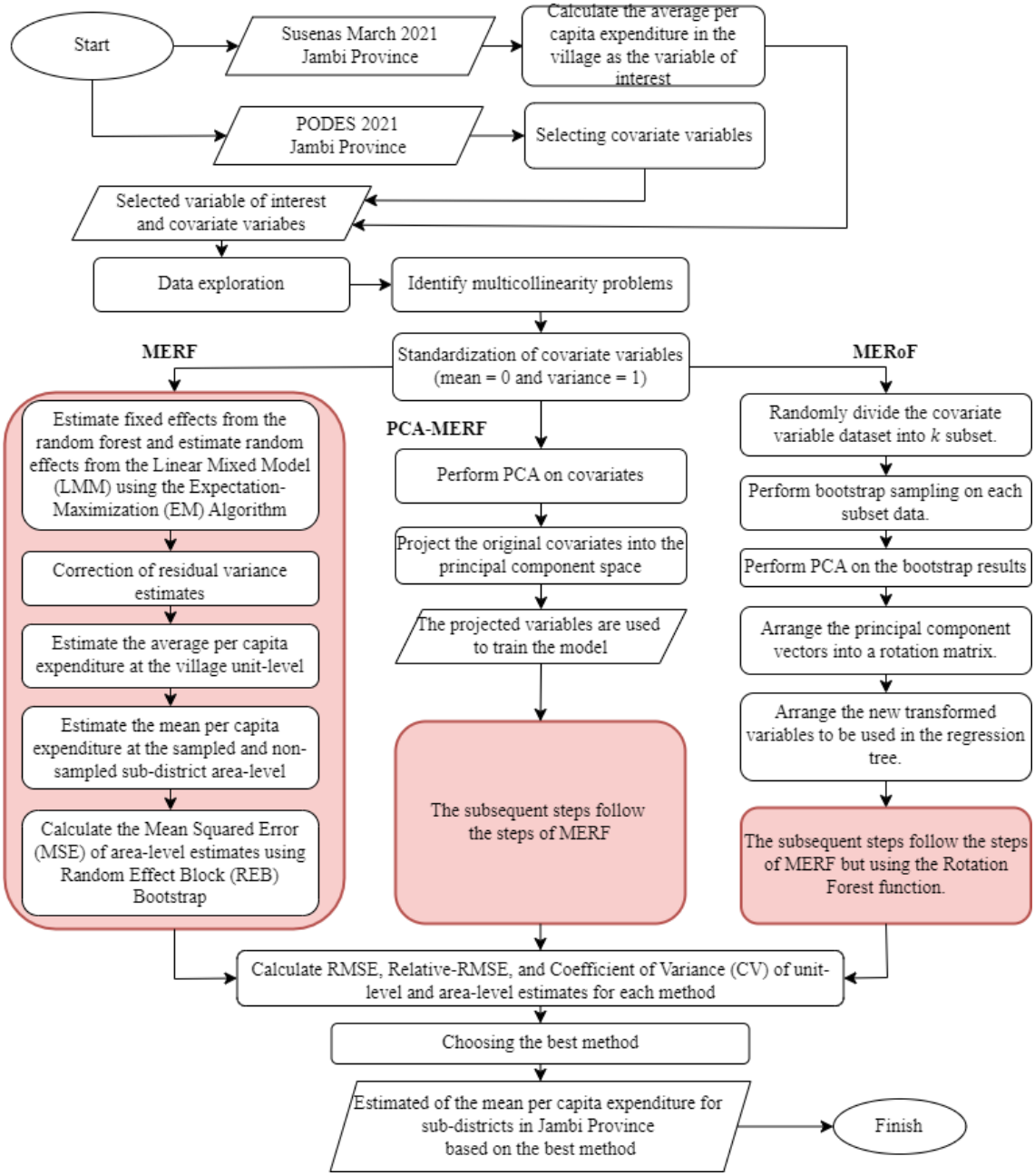

Figure 1. Flowchart of modeling and data analysis process

RESULTS AND DISCUSSIONS

Jambi Province consists of 11 regencies/municipalities, 144 subdistricts, and 1,562 villages. The March 2021 Susenas of Jambi Province featured a sample size of 6,913 households. This sample covered only 142 subdistricts from 625 villages in Jambi Province. Table 2 summarizes the March 2021 Susenas sample distribution in Jambi Province by regency/municipality. Table 2 shows that certain areas (villages and even subdistricts) are without sample representation. The variables used from the Susenas data included the total monthly expenditure on food and non-food items and the number of household members. The average per capita expenditure in a village was calculated by dividing the total expenditure by the total number of household members in the village.

Table 2. Summary of march 2021 susenas sample in jambi province

No	Regency/Municipality	Number of sampled villages	Number of non-sampled villages	Number of sampled subdistricts	Number of non-sampled subdistricts	Number of household samples
1	Kerinci Regency	60	227	18	-	608
2	Merangin Regency	63	215	23	1	653
3	Sarolangun Regency	61	97	10	1	615
4	Batang Hari Regency	58	66	8	-	616
5	Muaro Jambi Regency	64	91	11	-	656
6	Tanjung Jabung Timur Regency	55	38	11	-	618
7	Tanjung Jabung Barat Regency	55	79	13	-	611
8	Tebo Regency	56	56	12	-	660
9	Bungo Regency	65	88	17	-	659
10	Jambi Municipality	44	18	11	-	687
11	Sungai Penuh Municipality	44	25	8	-	530
	Total	625	937	142	2	6913

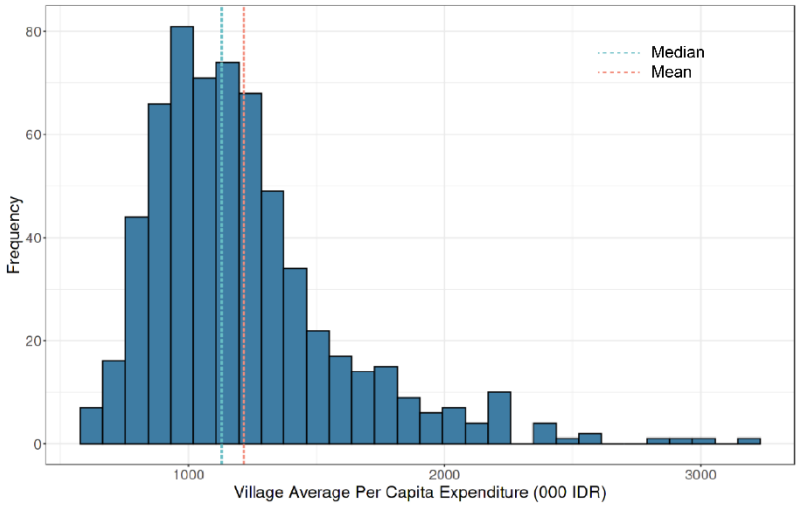

Figure 2. Distribution of village average per capita expenditure

Figure 2 illustrates that villages' average per capita expenditure does not exhibit a symmetric distribution. The right-skewed data indicate that most observations have relatively low or moderate per capita expenditures, while a small proportion of observations exhibit very high per capita expenditures. The median value being lower than the mean suggests that extreme (outlier) values influence the mean on the right side, which causes the average per capita expenditure to appear higher than it is for most observations.

Additionally, the correlation between the target variable and the covariates in the dataset was analyzed. As shown in Figure 3, there is one particularly strong correlation between covariates, specifically between the variables Pct_Fam_PLN and Pct_Fam_NonPLN, with a correlation coefficient of -0.97. This indicates a very strong and negative relationship between Pct_Fam_PLN and Pct_Fam_NonPLN, meaning that as the value of one variable increases, the value of the other tends to decrease consistently. Such strong relationships among covariates might indicate potential issues. Multicollinearity occurs when two or more variables in a regression model are highly correlated, which can lead to difficulties in interpreting the regression coefficients.

Figure 3 only shows pairwise correlations between two variables at a time and does not capture more complex multicollinearity interactions that may involve more than two variables. To assess multicollinearity, the variance inflation factor (VIF) was used [1] to evaluate the level of multicollinearity among the independent variables in a regression model. A VIF value below 5 typically indicates that multicollinearity is not a significant issue. Meanwhile. a VIF value between 5 and 10 suggests a considerable level of multicollinearity that could impact the regression coefficient estimates [29]. This serves as a warning that some variables may be too highly correlated within the regression model. A VIF value exceeding 10 generally indicates a severe multicollinearity problem, where the estimation of regression coefficients may become highly unstable due to strong correlations among the variables

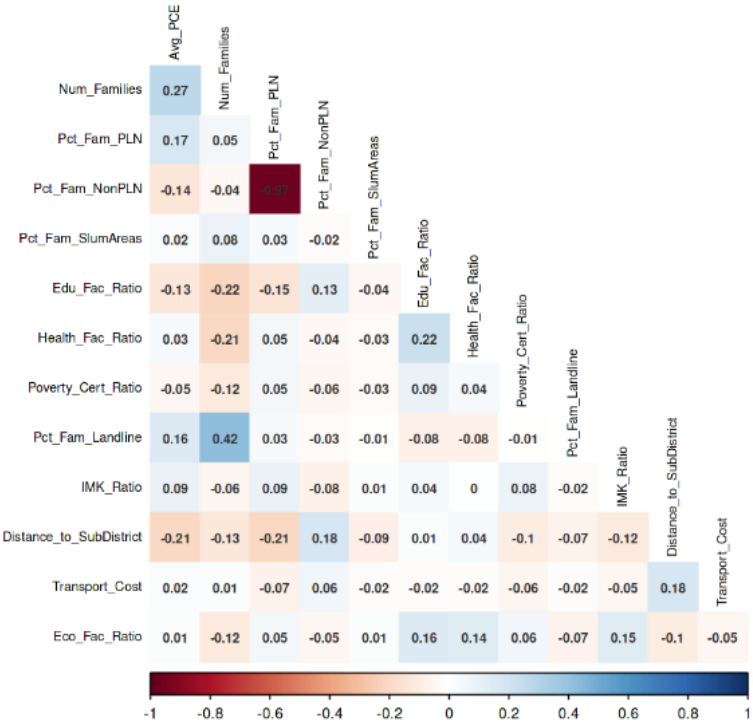

Figure 3. Correlation plot between variables

Table 3 presents the VIF values for each covariate. The variables Pct_Fam_PLN and Pct_Fam_NonPLN have VIF values of 18.84 and 18.47, respectively. These values confirm the presence of severe multicollinearity, indicating that the inclusion of these variables in the model is likely to cause estimation issues. The other variables have small VIF values of approximately 1, suggesting that multicollinearity is not a concern for these variables.

Table 3. VIF values for each covariate

Variable	VIF
Num_Families	1.35
Pct_Fam_PLN	18.84
Pct_Fam_NonPLN	18.47
Pct_Fam_SlumAreas	1.02
Edu_Fac_Ratio	1.14
Health_Fac_Ratio	1.11
Poverty_Cert_Ratio	1.05
Pct_Fam_Landline	1.22
IMK_Ratio	1.05
Distance_to_SubDistrict	1.15
Transport_Cost	1.04
Eco_Fac_Ratio	1.08

Estimation results of per capita expenditure for small area using MERF, PCA-MERF, and MERoF

Based on the results of empirical data exploration, it is evident that the dataset used in this study exhibits characteristics of a non-symmetric distribution. Additionally, several variables have been found to have very high correlations, indicating serious multicollinearity issues, as confirmed by the variance inflation factor (VIF) calculations. A dataset with these issues presents significant challenges for small area estimation using standard SAE methods. Such issues must be addressed beforehand to ensure that these methods' assumptions are met. MERF (mixed effects random forest) offers a potential solution by leveraging machine learning techniques, providing advantages for non-linear and non-parametric data. MERF combines the strengths of mixed models with the predictive power of random forest.

We acknowledge that the most effective estimation method is one that achieves the highest precision. Given the identified severe multicollinearity in the dataset, this study modified MERF in two ways: first, by

applying PCA (principal component analysis) during the preprocessing stage before MERF modeling, and second, by replacing the forest in the mixed model with rotation forest to specifically address multicollinearity issues. The estimation outcomes revealed the performance of per capita expenditure estimation for small areas with problematic multicollinearity data through the methods MERF, PCA-MERF, and MERoF.

In mixed models, the principal components consist of fixed effects, which are variables that consistently affect the entire dataset. In MERF, these fixed effects are derived from the original covariates. In PCA-MERF and MERoF, the fixed effects are based on the original variables transformed into new, linearly independent components. The distinction between PCA-MERF and MERoF lies in how the forest is trained: PCA-MERF employs a random forest that uses only a subset of variables to construct the trees, whereas MERoF incorporates all variables in tree construction to preserve complete information. In addition to fixed effects, random effects capture variation among groups within the data—an essential aspect of small area estimation (SAE). These mixed models can capture variation among subdistricts as random effects and variation within subdistricts that originate from covariate variables as fixed effects.

To evaluate the quality of the estimations, this study examined the quality of estimations at the unit level (village) and the quality of estimations at the area level (subdistrict). The quality of estimations at the village level was assessed by comparing the estimated values with the actual observed values. Meanwhile, the quality of estimations at the subdistrict level was evaluated by estimating the uncertainty of the MSE using the REB bootstrap scheme, which compared the estimated average values for subdistricts with those derived from bootstrap results. The estimation results using MERF, PCA-MERF, and MERoF are presented in Table 4.

Table 4. Estimation results using MERF, PCA-MERF, and MERoF

Component	MERF	PCA-MERF	MERoF	Unit
Evaluation Metric for Unit-level Estimation				
RMSE	287,89	283,05	279,57	000 IDR
Relative-RMSE	23,68	23,28	22,99	%
Coefficient of Variation (CV)	17,23	17,34	16,45	%
Evaluation Metric for Area-level Estimation				
RMSE	111,81	110,64	123,47	000 IDR
Relative-RMSE	9,04	8,76	10,31	%
Coefficient of Variation (CV)	8,84	8,53	10,07	%
Random Effects from LMM (Linear Mixed Models)				
SD of Area	182,19	200,21	243,71	000 IDR
SD of Residual	308,43	305,13	304,71	000 IDR
ICC	0,26	0,30	0,39	-

Note: Best Value

Overall, the evaluation of village-level predictions using RMSE, RRMSE, and CV values, as shown in Table 4, indicates that MERoF has the smallest values compared to the other two methods. This suggests that the dispersion or variability of errors for the MERoF method is lower than that for the other methods in village-level estimation. However, the RMSE, RRMSE, and CV values for village-level predictions are higher than those at the subdistrict level. This is likely due to the larger sample sizes at the subdistrict level, which result in more stable estimates. In contrast, the evaluation of subdistrict level predictions revealed that PCA-MERF outperforms MERF and MERoF in predicting the average per capita expenditure at the subdistrict level, as evidenced by its smaller RMSE, RRMSE, and CV values. This is more clearly illustrated in Figure 4. Nevertheless, all three methods can still be categorized as good, reliable, and stable for predicting small areas, with values falling within the 8-10% range.

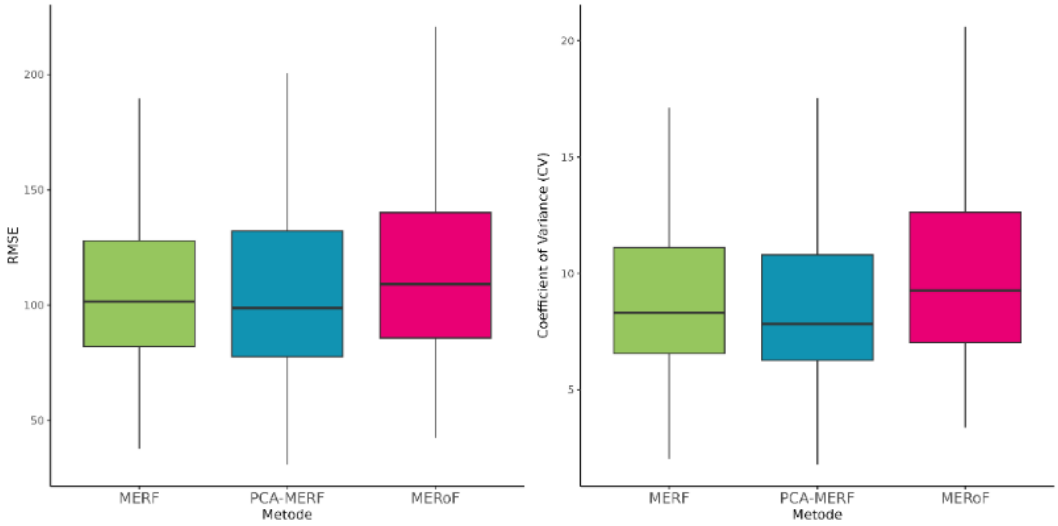

Figure 4. Boxplot of RMSE and CV values of three methods

Subsequently, we examined the estimation results for the random effects of each model. The standard deviation (SD) of area effects measured the variation between subdistricts, while the SD of residuals captured the variation in unexplained errors by the model. The intraclass correlation coefficient (ICC) measured the proportion of total variability attributed to variation between subdistricts. For the SD of area effects, MERF yielded a value of 182.19, which was smaller compared to PCA-MERF's value of 200.21. MERoF, on the other hand, resulted in a larger SD of area effects at 243.71, indicating more significant variation in the target variable, average per capita expenditure, among different subdistricts. As for the ICC, MERoF also had a higher value at 0.39. This suggests that the variation between subdistricts contributes more significantly to the total variability in the data. MERoF demonstrates a better ability to identify and capture the variation between areas than MERF and PCA-MERF. A higher ICC value implies that average per capita expenditure varies significantly between different subdistricts, likely due to differences in demographic, economic, or socio-ecological characteristics among the subdistricts. In terms of the standard deviation of residuals, MERoF had a value of 304.71, slightly lower than PCA-MERF's value of 305.13. MERF had the highest residual variance among the three methods at 308.43. A higher residual variance indicates greater variability in the data that the model has not captured, possibly due to local factors such as geographical differences, spatial variability, or measurement uncertainties. These results suggest that MERoF controls unexplained variation slightly better than the other two methods, though it is still not entirely conclusive that MERoF is definitively superior.

Table 5. Estimation of mean of per capita expenditure per subdistrict in several subdistricts based on MERF, PCA-MERF, and MERoF methods (000 IDR)

Id_Subdistrict	MERF	PCA-MERF	MERoF	Number of Village Samples
0440	1211.45	1251.86	1178.23	10
0911	1302.41	1382.22	1305.36	10
0510	1084.00	1094.32	1074.74	10
0730	1341.59	1406.63	1259.47	9
0632	1279.88	1254.53	1256.68	9
...	...	...	...	...
0230	1151.73	1131.59	1079.65	5
0350	993.13	970.06	914.20	5
7120	1464.54	1450.04	1333.80	4
0650	1225.80	1168.12	1207.02	4
7150	1882.76	1935.77	1823.02	4
...	...	...	...	...
7160	1311.29	1353.86	1209.10	1
0222	1104.78	1099.19	1004.18	1
0265	1328.88	1195.79	1158.34	-
0361	1204.48	1177.68	1180.38	-

In addition to estimating areas covered by the sample, MERF, PCA-MERF, and MERoF were also able to estimate areas not included in the sample. In the March 2021 Susenas data, 142 subdistricts in Jambi Province were included in the sample, excluding 2 subdistricts. Table 5 and Figure 5 illustrate the estimation of average per capita expenditure for subdistricts using MERF, PCA-MERF, and MERoF, sorted by the number of village samples from largest to smallest. Subdistricts with a larger number of samples tended to have more consistent estimates, with smaller differences among methods. This suggests that a greater number of samples provides more stable estimates. Generally, PCA-MERF tended to yield higher estimates compared to the other methods, especially in subdistricts with a larger number of samples. Conversely, MERoF tended to provide lower estimates. In subdistricts without sample representation, the differences among methods became more pronounced, with MERF providing significantly higher estimates compared to the other two methods. This suggests that estimates made using methods that do not address multicollinearity and exclude random effects can result in much higher estimates.

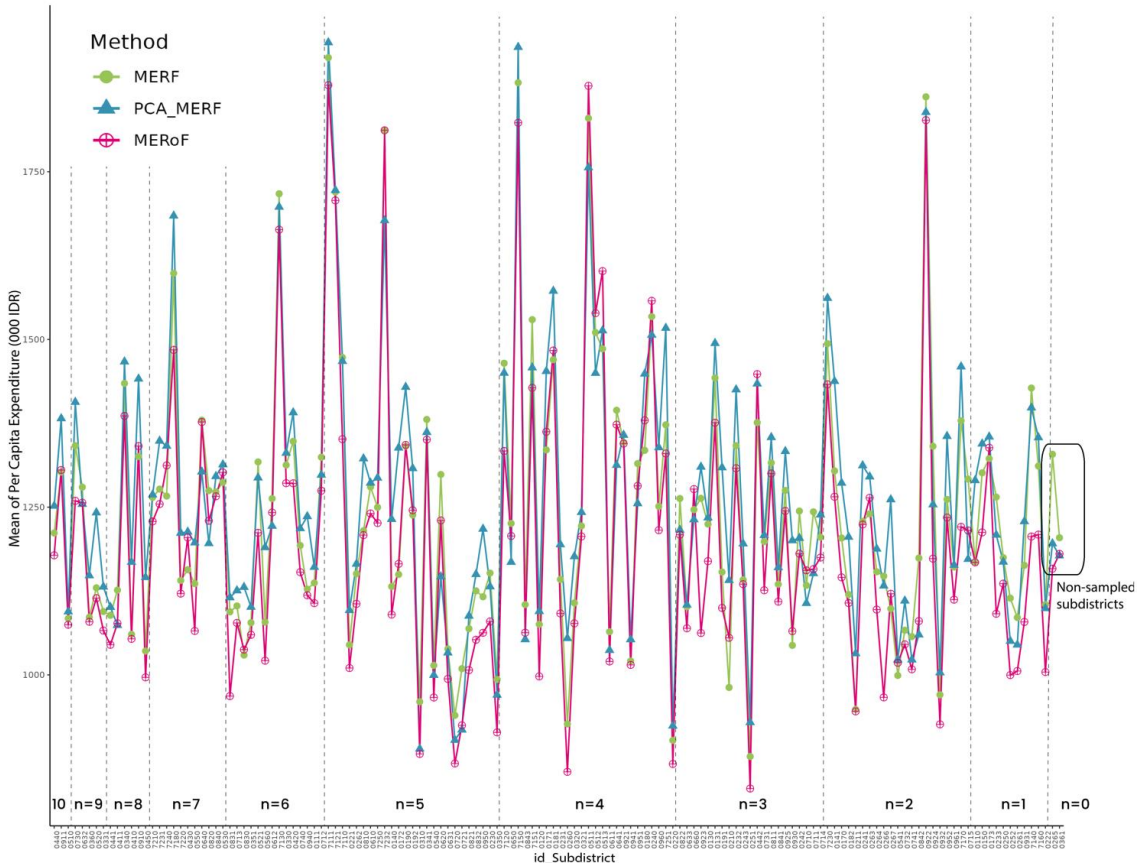

Figure 5 Estimates from MERF, PCA-MERF, and MERoF Methods, sorted by village samples size from largest to smallest

The findings of this study highlight the effectiveness of modifying the mixed effects random forest (MERF) method to address the critical issue of multicollinearity in small area estimation (SAE). Both PCA-MERF and MERoF demonstrate superior performance compared to the original MERF method, with PCA-MERF excelling in subdistrict level estimation and MERoF showing significant advantages in village level estimation. This suggests that tailored approaches to handling multicollinearity can yield more accurate and reliable estimates, depending on the specific level of analysis.

The research also underscores the importance of considering the hierarchical structure of data, particularly when dealing with geographically clustered areas such as subdistricts and villages. The higher intraclass correlation coefficient (ICC) observed in MERoF suggests that this method better captures the inherent variability between subdistricts, which is crucial for understanding regional disparities in per capita expenditure. Moreover, the lower residual variance in MERoF indicates that it is slightly more effective in controlling unexplained variability.

However, while PCA-MERF and MERoF provide notable improvements, it is evident that no single method is universally superior across all levels of analysis. The choice of method should therefore be guided by the specific requirements of the estimation level—whether it is at the village or subdistrict level—and by the nature of the underlying data. This study contributes to the ongoing development of more robust SAE methods by demonstrating the value of incorporating advanced techniques such as PCA-random forest and rotation forest in mixed models.

Given the limitations related to sample size and data availability at smaller area levels, the findings of this study also emphasize the need for continued innovation in SAE methodologies. Future research could further explore the integration of additional machine learning techniques or hybrid models to enhance the robustness and flexibility of SAE methods, particularly in regions with limited data. By addressing these challenges, researchers and policymakers can obtain more accurate and actionable insights, ultimately supporting more effective and targeted regional development policies.

CONCLUSION

The per capita expenditure data in Jambi Province exhibit significant multicollinearity issues. To address this critical problem, modifications to MERF were implemented using two alternative approaches: PCA-MERF and MERoF, both designed to handle multicollinearity. The results showed that all three methods demonstrate precise and reliable performance in estimating average per capita expenditure at the subdistrict level, with RRMSE and CV values ranging from 8-10%. This indicates that MERF effectively manages multicollinearity issues, particularly in subdistrict level estimation. However, the modified methods provide even better performance compared to MERF. MERoF excels in village-level estimation and in capturing variation among subdistricts, making it a strong candidate for estimating average per capita expenditure at the village level. Meanwhile, PCA-MERF outperforms MERF and MERoF in estimating average per capita expenditure at the subdistrict level, making PCA-MERF the preferable choice for subdistrict-level estimates. This research indicates that modifying MERF is a sound and beneficial approach for estimating per capita expenditure in small areas in Jambi Province.

However, this study is not without its limitations and challenges. This study utilizes 2021 Podes as the population data, and therefore the unit of the March 2021 Susenas survey data used in this research is the village, consistent with the unit in the Podes data. This choice was made due to the difficulty in obtaining the most recent household population data. A weakness in this study includes the lack of household-level data, which necessitated the use of village-level data. Using village-level data may not fully capture individual variability, potentially affecting the precision of the estimates. Another difficulty encountered was the limited sample representation in some subdistricts and villages, leading to instability in the estimates for areas not represented in the survey sample. This challenge highlights the need for more adaptive and flexible methods to handle very limited or incomplete data.

For future research, it is recommended that household-level data be used, whenever available, to enhance the precision of the estimates. Further exploration of integrating machine learning techniques or hybrid models into SAE is also necessary to overcome the challenges faced in this study such as data incompleteness.

REFERENCES

[1] R. K. Paul, "Multicollinearity: Causes, effects and remedies," *IASRI, New Delhi*, vol. 1, no. 1, pp. 58–65, 2006.

[2] J. N. Rao and I. Molina, *Small area estimation*. John Wiley & Sons, 2015. [Online]. Available: 10.1002/9781118735855

[3] A. Kurnia, "Prediksi Terbaik Empirik untuk Model Transformasi Logaritma di dalam Pendugaan Area Kecil dengan Penerapan pada Data Susenas," *Disertasi, IPB Bogor*, 2009.

[4] A. Kurnia, D. Kusumaningrum, A. M. Soleh, D. Handayani, and R. Anisa, "Small area estimation with winsorization method for poverty alleviation at a sub-district level," *International Journal of Applied Mathematics & Statistics*, vol. 53, no. 6, pp. 77–84, 2015.

[5] R. E. Fay III and R. A. Herriot, "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 269–277, 1979, doi: 10.1080/01621459.1979.10482505.

[6] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," *Journal of Statistical Computation and Simulation*, vol. 84, no. 6, pp. 1313–1328, 2014, doi: 10.1080/00949655.2012.741599.
[7] P. Krennmair and T. Schmid, "Flexible domain prediction using mixed effects random forests," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 71, no. 5, pp. 1865–1894, 2022, doi: 10.1111/rssc.12600.
[8] P. Krennmair, N. Würz, and T. Schmid, "Analysing opportunity cost of care work using mixed effects random forests under aggregated census data," *arXiv preprint arXiv:2204.10736*, 2022, doi: 10.48550/arXiv.2204.10736.
[9] J. Jiang and J. S. Rao, "Robust small area estimation: An overview," *Annual review of statistics and its application*, vol. 7, pp. 337–360, 2020, doi: 10.1146/annurev-statistics-031219-041212.
[10] A. S. Gwelo, "Principal components to overcome multicollinearity problem," *Oradea Journal of Business and Economics*, vol. 4, no. 1, pp. 79–91, 2019, doi: 10.47535/1991ojbe062.
[11] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006, doi: 10.1109/TPAMI.2006.211.
[12] N. Poona, A. Van Niekerk, and R. Ismail, "Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data," *Sensors*, vol. 16, no. 11, p. 1918, 2016, doi: 10.3390/s16111918.
[13] A. Bagnall, M. Flynn, J. Large, J. Line, A. Bostrom, and G. Cawley, "Is rotation forest the best classifier for problems with continuous features?," *arXiv preprint arXiv:1809.06705*, 2018, doi: 10.48550/arXiv.1809.06705.
[14] A. S. Bukhari, K. A. Notodiputro, and B. Sartono, "Comparing Rotation Forest Model And Enhanced Random Forest Model On Imbalanced Data (Application To Classification Of Poverty Households In Sampang Regency, 2019)," *Jurnal Ekonomi Pertanian dan Agribisnis*, vol. 7, no. 2, pp. 933–943, 2023, doi: 10.21776/ub.jepa.2023.007.02.44.
[15] C. Pardo, J. F. Diez-Pastor, C. García-Osorio, and J. J. Rodríguez, "Rotation Forests for regression," *Applied Mathematics and Computation*, vol. 219, no. 19, pp. 9914–9924, 2013, doi: 10.1016/j.amc.2013.03.139.
[16] J. J. Rodrıguez, M. Juez-Gil, C. López-Nozal, and A. Arnaiz-González, "Rotation Forest for Multi-target Regression", *International Journal of Machine Learning and Cybernetics*, 13(2):523–548, 2022, doi: 10.1007/s13042-021-01354-0.
[17] A. Salma, K. Sadik, and K. A. Notodiputro, "Small area estimation of per capita expenditures using robust empirical best linear unbiased prediction (REBLUP)," in *AIP Conference Proceedings*, AIP Publishing, 2017, doi: 10.1063/1.4979443.
[18] Y. Susianto, K. A. Notodiputro, A. Kurnia, and H. Wijayanto, "Small area estimation models with time factor effects for repeated measurement data," *Applied Mathematical Sciences*, vol. 11, no. 41, pp. 1995–2010, 2017, doi: 10.12988/ams.2017.74142.
[19] C. Sumarni, K. Sadik, K. A. Notodiputro, and B. Sartono, "Estimation of per capita household expenditure: A likelihood approach of robust extension of small area estimation," *Journal of Applied Probability and Statistics*, vol. 14, no. 3, pp. 75–93, 2019.
[20] A. Ubaidillah, K. A. Notodiputro, A. Kurnia, and I. W. Mangku, "Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia," *Journal of Applied Statistics*, vol. 46, no. 15, pp. 2845–2861, Nov. 2019, doi: 10.1080/02664763.2019.1615420.
[21] N. H. Pusponegoro, A. Kurnia, K. A. Notodiputro, A. M. Soleh, and E. T. Astuti, "Small area estimation of sub-district's per capita expenditure through area effects selection using LASSO method," *Procedia Computer Science*, vol. 179, pp. 754–761, 2021, doi: 10.1016/j.procs.2021.01.064.
[22] N. Hasanah, K. A. Notodiputro, and B. Sartono, "Performance of copula and nested error regression models in estimating per capita expenditure of sub-district in Pidie Regency," *Jurnal Natural*, vol. 23, no. 2, pp. 64–71, 2023, doi: doi: 10.24815/jn.v23i2.31095.
[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[24] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[25] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.
[26] S. Truex, L. Liu, M. E. Gursoy, and L. Yu, "Privacy-preserving inductive learning with decision trees," in *2017 IEEE International Congress on Big Data (BigData Congress)*, IEEE, 2017, pp. 57–64. doi: 10.1109/BigDataCongress.2017.17.
[27] G. Mendez and S. Lohr, "Estimating residual variance in random forest regression," *Computational statistics & data analysis*, vol. 55, no. 11, pp. 2937–2950, 2011, doi: 10.1016/j.csda.2011.04.022.
[28] R. Chambers and H. Chandra, "A random effect block bootstrap for clustered data," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 452–470, 2013, doi: 10.1080/10618600.2012.681216.
[29] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. in Wiley Series in Probability and Statistics. Wiley, 2021. [Online]. Available: https://books.google.co.id/books?id=tCIgEAAAQBAJ