Int. J. Advance Soft Compu. Appl, Vol. 18, No. 1, March 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

A Spatial Approach to Correlated High-Dimensional Stunting Data in Indonesia Using a Modified Generalized Lasso

Septian Rahardiantoro, Aida Darajati, Hari Wijayanto and Anang Kurnia*

Program in Statistics and Data Science
School of Data Science, Mathematics, and Informatics
IPB University - Indonesia
*e-mail: anangk@apps.ipb.ac.id

Abstract

Stunting remains a significant public health issue in Indonesia. Although the national prevalence declined by 6.1% in 2024, several provinces continue to exhibit alarmingly high rates. This study aims to explore the spatial patterns of stunting across Indonesia, evaluate the performance of the generalized lasso model in identifying potential regional coefficient groupings based on various neighborhood structures, and determine the most influential factors contributing to stunting. The data, sourced from Statistics Indonesia and the Ministry of Home Affairs in 2024, cover 38 provinces and include ten predictor variables. The analysis employs a modified elastic net approach within the generalized lasso framework by incorporating a custom penalty matrix into the L_2 regularization term to mitigate multicollinearity among predictors. The proposed models were evaluated against the Spatial Autoregressive (SAR) model, standard elastic net, and standard generalized lasso using specified neighborhood adjacency methods and tuning parameters. Optimal tuning parameters were selected using the Approximate Leave-One-Covariate-Out Cross-Validation (ALOCV) method. The best-performing model was identified as the K-Nearest Neighbors (KNN) model with k=3 and the custom penalty matrix, based on an optimal balance of degrees of freedom, lower AIC and RMSE, and optimum sensitivity criteria. The results reveal that the most influential factors associated with stunting prevalence in 2024 include the poverty rate—particularly in southern Kalimantan, several provinces in Sumatra, and East Nusa Tenggara—and child health insurance coverage, which spans provinces across Indonesia.

Keywords: *ALOCV, high-dimensional data, generalized LASSO, KNN, stunting*

1 Introduction

Stunting is a serious global health issue caused by chronic malnutrition, recurrent infections, and lack of stimulation during the first 1,000 days of life, leading to impaired growth, cognitive development, and future productivity. According to WHO, the global stunting prevalence among children under five rose to 23.2% in 2024, affecting approximately 150.2 million children, with the highest rates found in low- and middle-

Received 26 October 2025; Accepted 21 February 2026

income countries like Indonesia. In response, the Indonesian government has prioritized stunting reduction through national development plans, achieving a 6.1% decrease in prevalence by 2024, which supports both improved child well-being and national socioeconomic progress [1].

Although the prevalence of stunting in Indonesia has declined annually, several provinces still report alarmingly high rates. This disparity exists because each region has distinct contributing factors that influence stunting prevalence. Previous studies have identified poverty, limited education, lack of health insurance, and inadequate sanitation as common causes of stunting; however, these factors are generally analyzed at the national level rather than in a region-specific context. As a result, many interventions fail to effectively target the areas most in need, allowing high stunting rates to persist in certain provinces. Therefore, stunting prevention efforts must be strengthened by comprehensively examining the contributing factors across all provinces. Moreover, only a limited number of studies have incorporated spatial considerations into stunting research, despite evidence that stunting prevalence in a given region can be influenced by neighboring areas due to shared social, economic, and geographic characteristics [2].

Generalized LASSO is highly suitable for this study as it accommodates spatial variation by allowing region-specific coefficients and grouping similar effects among neighboring areas [3]. To further address multicollinearity in high-dimensional data, this study proposes an extension by incorporating a custom $\boldsymbol{D}_2$ penalty into the L_2 regularization to capture shared influences among highly correlated predictors. Given the limitations of previous research and the spatially diverse nature of stunting determinants, there is a pressing need for more regionally sensitive analytical approaches. This study seeks to address that gap by employing a spatial modeling framework tailored to high-dimensional data. Specifically, the objectives of this research are threefold. First, to explore the spatial patterns of stunting prevalence across all provinces in Indonesia. Second, to evaluate the performance of the generalized lasso model—both with and without a customized penalty matrix ($\boldsymbol{D}_2$)—in identifying potentially grouped coefficients among neighboring regions under various neighborhood structures. Third, to identify and interpret the most influential predictors associated with stunting prevalence in Indonesia in 2024. By incorporating spatial aspects and applying a modified regularization approach, this study aims to provide more accurate and region-specific insights for targeted stunting interventions and policy development.

Previous studies have applied various Bayesian spatial and spatio-temporal Conditional Autoregressive (CAR) models to identify key risk factors driving stunting across Indonesia [4-5]. These analyses, which evaluated over 1,200 models in total, consistently found that poverty rate and low birth weight were strongly associated with higher stunting risk, while child dietary diversity acted as a protective factor. Spatial clustering revealed that provinces such as Sulawesi Barat and Nusa Tenggara Timur have persistently high risks, whereas DKI Jakarta remains the lowest, providing an important foundation for further region-specific analyses using methods such as the generalized lasso to refine targeted intervention strategies.

2 Related Work

Conventional methods such as Ordinary Least Squares (OLS) are limited in analyzing regional stunting data, as they produce a single coefficient per variable and fail to account for spatial heterogeneity. Moreover, OLS performs poorly with high-dimensional data due to instability and sensitivity to small changes, particularly when strong correlations exist among predictors [6]. While regularization techniques such as lasso [6] and elastic net [7] address multicollinearity by adding penalty terms, they too do not incorporate spatial structures and produce uniform coefficients across all regions. To overcome these limitations, Tibshirani and Taylor (2011) [3] proposed the generalized lasso, which integrates a penalty matrix $\boldsymbol{D}$ into the L_1 regularization to accommodate spatial dependencies between neighboring regions. This method allows for region-specific coefficient estimation and spatial grouping without relying on strong assumptions, making it superior to traditional spatial regression models.

The generalized lasso [3] is able to cluster spatial objects by incorporating spatial structure into the penalty matrix of L_1 regularization. Several studies in spatial clustering demonstrated its application [8-11]. Furthermore, the application of generalized lasso in spatial modeling could be found in [12-14]. In these cases, the generalized lasso is applied in region-based modeling to reveal the similar associations between predictors and the response variable. The study in [14], they successfully applied generalized lasso to detect spatial-temporal patterns of stunting across 34 provinces in Indonesia using nine predictors.

However, as stunting remains a critical and evolving issue, further investigation using updated 2024 data from 38 provinces and a broader set of variables is necessary. High-dimensional settings present additional challenges, such as high inter-variable correlation and coefficient instability. Tibshirani and Taylor (2011) [3] addressed these by incorporating L_2 regularization into the generalized lasso function, stabilizing estimates through a ridge-like adjustment. In this study, we propose a modification to generalized lasso by introducing a custom penalty component $\boldsymbol{D}_2$ into the L_2 regularization. This matrix, built from strongly correlated variable pairs, enables the model to retain these predictors while capturing their shared influence on stunting outcomes across provinces. Additionally, the Approximate Leave-One-Out Cross Validation (ALOCV) method is employed to determine the optimal tuning parameter λ, as it provides accurate and computationally efficient error estimates for high-dimensional, spatially structured data [15]. The contribution of this paper lies not in proposing a fundamentally new estimator, but in integrating spatial correlation awareness into generalized lasso for small-area, high-dimensional public health data, and demonstrating its practical advantages over classical spatial and regularized models in the Indonesian stunting context.

3 The Proposed Method: The Modified Generalized Lasso for Correlated High-Dimensional Data

Linear regression is widely used to predict a response variable based on one or more predictor variables, expressed as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the response vector, $\boldsymbol{X}$ is the predictor matrix, $\boldsymbol{\beta}$ represents the regression coefficients, and $\boldsymbol{\varepsilon}$ denotes the error term [16]. Ordinary Least Squares (OLS) estimates $\boldsymbol{\beta}$ by minimizing the sum of squared residuals, but this method tends to overfit when predictors are highly correlated, leading to models with low bias but high variance. To mitigate overfitting, regularization techniques add penalty terms to constrain model complexity. Elastic net, proposed by [7], combines the benefits of lasso (L_1 penalty) and ridge regression (L_2 penalty) by minimizing the sum of squared errors along with penalties on the absolute and squared values of coefficients:

$$\boldsymbol{\beta}^{EN} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\}, \tag{1}$$

where λ_1 and λ_2 are tuning parameters. The L_1 penalty encourages sparsity by shrinking some coefficients to zero, effectively performing variable selection, while the L_2 penalty stabilizes estimates and reduces overfitting. Elastic net is especially suitable for high-dimensional data with highly correlated variables; however, it does not account for spatial dependencies within the data.

To overcome the limitation of ignoring spatial structure in traditional regularization methods, the generalized lasso [3] extends lasso by incorporating a penalty matrix $\boldsymbol{D}_1$ representing spatial relationships among regression coefficients. The generalized lasso estimator is formulated as:

$$\boldsymbol{\beta}^{GL} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1\|\boldsymbol{D}_1\boldsymbol{\beta}\|_1\}, \tag{2}$$

where $\boldsymbol{D}_1$ encodes spatial adjacency or similarity between regions. If $\boldsymbol{D}_1$ equals the identity matrix, this reduces to standard lasso. This approach allows coefficients to vary across spatial units while encouraging similar values for neighboring regions, thus capturing spatial heterogeneity. Nevertheless, when applied to high-dimensional data, generalized lasso faces challenges such as coefficient instability and non-unique solutions due to rank deficiencies in $\boldsymbol{X}$ [3].

To address these issues, Tibshirani and Taylor (2011) [3] proposed incorporating an L_2 ridge penalty into the generalized lasso framework, enhancing estimation stability in high-dimensional settings:

$$\boldsymbol{\beta}^{GLHD} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1\|\boldsymbol{D}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\}, \tag{3}$$

Building on this, a further modification introduces a special penalty matrix $\boldsymbol{D}_2$, constructed from pairs of highly correlated predictors, to reduce multicollinearity effects by decorrelating their impacts. This modified objective resembles elastic net but incorporates spatial structure through $\boldsymbol{D}_1$ and the correlation-aware penalty $\boldsymbol{D}_2$, which the modified generalized lasso estimator is formulated as:

$$\boldsymbol{\beta}^{MGLHD} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1\|\boldsymbol{D}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{D}_2\boldsymbol{\beta}\|_2^2\}, \tag{4}$$

Equations (3) and (4) resemble the elastic net with the addition of matrix ($\boldsymbol{D}_1$ and $\boldsymbol{D}_2$). However, both equations can be reformulated into a generalized lasso. This reformulation is carried out to enable modeling using existing programs, thereby simplifying the computational process. After reformulation, the generalized lasso equation is as follows (Tibshirani and Taylor, 2011):

$$\boldsymbol{\beta}^{MGLHD} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y}^* - (\boldsymbol{X}^*)\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{D}_1\boldsymbol{\beta}\|_1\}, \tag{5}$$

where $\boldsymbol{y}^* = (\boldsymbol{y}, 0)^T$, $\lambda_2 > 0$, and $\boldsymbol{X}^* = \begin{bmatrix}\boldsymbol{X}\\ \varepsilon.I\end{bmatrix}$ for the equation without matrix $\boldsymbol{D}_2$, and $\boldsymbol{X}^* = \begin{bmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}.\boldsymbol{D}_2\end{bmatrix}$ for the equation with matrix $\boldsymbol{D}_2$. The model (5) can be equivalently reformulated to fit within standard generalized lasso computational frameworks, preserving ridge stability without complicating optimization. This allows more accurate parameter estimation in high-dimensional spatial data, improving interpretability and predictive performance [3]. This study also includes a sensitivity analysis on the addition of the correlation-aware penalty $\boldsymbol{D}_2$ to evaluate its robustness.

4 Methodology

This section outlines the methodology employed in this study, consisting of two main parts. The first part describes the data used, while the second details the analytical procedures carried out to achieve the research objectives.

4.1 Data

Table 1: List of data variables

Code	Variable Name	Reference
Y	Prevalence of Stunting in Indonesia	-
X_1	Percentage of the Population Living in Poverty	[26]
X_2	Gini ratio	[27]
X_3	Percentage of Mothers with Health Insurance Coverage	[28]
X_4	Percentage of Children with Health Insurance Coverage	[28]
X_5	Percentage of Unmet Need for Healthcare Services	[14]
X_6	Percentage of Exclusive Breastfeeding	[29]
X_7	Percentage of Access to Improved Sanitation	[30]
X_8	Percentage of Access to Improved Drinking Water	[30]
X_9	Percentage of Mothers Who Smoke	[28]
X_{10}	Average Years of Schooling	[26]

This study uses secondary data obtained from the official website of Statistics Indonesia [17-25] and publications by the Ministry of Home Affairs [1]. The dataset includes one response variable and ten predictor variables, collected from 38 provinces in Indonesia for the year 2024. The list of variables used in this study is presented in Table 1.

4.2 Analysis Procedure

The data analysis was conducted using the RStudio software. The following steps outline the analytical procedure using the modified generalized lasso method.

a. An exploratory analysis was conducted by mapping the distribution of the response variable to visualize stunting prevalence across all provinces in Indonesia.

b. The data preprocessing involved assessing and transforming the response variable for symmetry, standardizing predictors using Z-score normalization, examining correlations with Pearson correlation, and restructuring the predictor matrix into a block-diagonal form to support region-specific spatial modeling [12].

c. Construction of Penalty Matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$

   i. The penalty matrix $\boldsymbol{D_1}$ was constructed to represent spatial proximity among provinces by first defining a base matrix $\boldsymbol{D_0}$ using queen contiguity and k-Nearest Neighbors (KNN) with $k = 2$ and $k = 3$. Queen contiguity [31] defines neighbors based on shared boundaries or vertices, while KNN [32] defines neighbors based on geographical distances between regional centroids. The resulting $\boldsymbol{D_0}$ matrix was replicated p times along the diagonal to form the block-diagonal matrix $\boldsymbol{D_1}$, yielding three variants: $\boldsymbol{D}_{1,Queen}$, $\boldsymbol{D}_{1,KNN(2)}$, and $\boldsymbol{D}_{1,KNN(3)}$.

   ii. The matrix $\boldsymbol{D}_2 \in \mathbb{R}^{(n\times b)\times(n\times p)}$ was constructed to penalize strongly correlated predictor variables (absolute Pearson correlation ≥ 0.5) by assigning -1 and 1 values to each correlated variable pair across regions (b), thereby reducing redundancy in the model. To incorporate $\boldsymbol{D}_2$, the design matrix was extended to $\boldsymbol{X}^* \in \mathbb{R}^{(n+n\times b)\times(n\times p)}$ and the response vector to $\boldsymbol{y}^* \in \mathbb{R}^{(n+n\times b)\times 1}$ by appending zeros, while models without $\boldsymbol{D_2}$ used the original $\boldsymbol{X} \in \mathbb{R}^{(n)\times(n\times p)}$ and $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$.

d. Modeling and sensitivity analysis were conducted using spatial Auto-regressive (SAR), Elastic Net, generalized lasso without $\boldsymbol{D}_2$, and proposed generalized lasso with $\boldsymbol{D}_2$ for $\lambda_2 = (0.01, 0.1, 1, 10)$. The SAR and generalized lasso models were fitted under three spatial neighborhood definitions (queen contiguity, KNN with $k = 2$, and KNN with $k = 3$), while the Elastic Net model followed the specified λ_2 values. The optimal penalty parameter λ_1 for each model (except SAR) was selected using the Approximate Leave-One-Out Cross-Validation (ALOCV) method [15]. Sensitivity analysis was then performed by comparing generalized lasso with $\boldsymbol{D}_2$ against generalized lasso without $\boldsymbol{D}_2$. Model performance was evaluated using AIC [33], RMSE, sensitivity values, and the correlation between fitted values of the specified model and fitted values of the baseline model.

e. The selection of the most appropriate model was accompanied by a bootstrap procedure [34] to assess the stability of the parameter estimates, after which the estimated parameters were visualized using heatmaps to examine the magnitude of each predictor's effect at specific locations on the response variable. Furthermore, in this step will also identify and interpret the most influential predictors associated with stunting prevalence in Indonesia in 2024.

5 Results, Analysis, and Discussions

5.1 Data Exploration and Preprocessing

The spatial distribution of stunting prevalence in Indonesia in 2024 is illustrated in Figure 2. The map uses color gradients to represent different prevalence levels: green for low (<5%), yellow to orange for moderate (10–15%), and red for high (>20%). The eastern and central regions of Indonesia—such as Sulawesi, Nusa Tenggara, and parts of Papua—tend to have higher stunting rates than the western regions like Sumatera and Java. West Sulawesi recorded the highest stunting prevalence at 23.9%, while provinces such as South Kalimantan, South Sumatera, and Jakarta reported the lowest rates, each below 2%. These findings emphasize regional disparities in stunting that require targeted policy interventions.

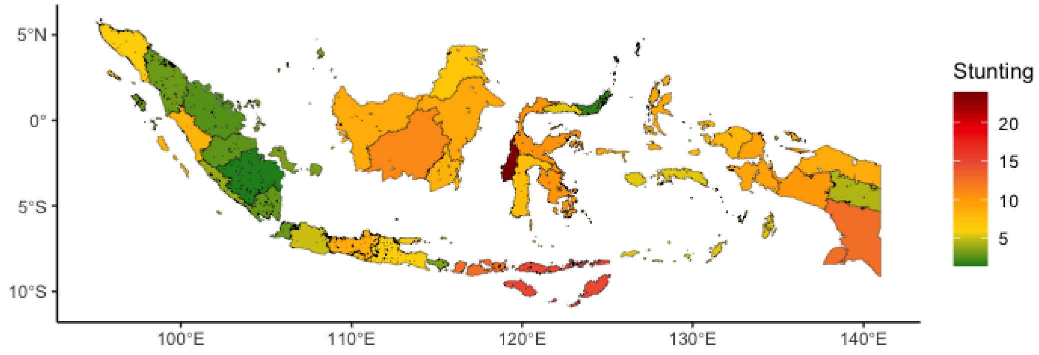
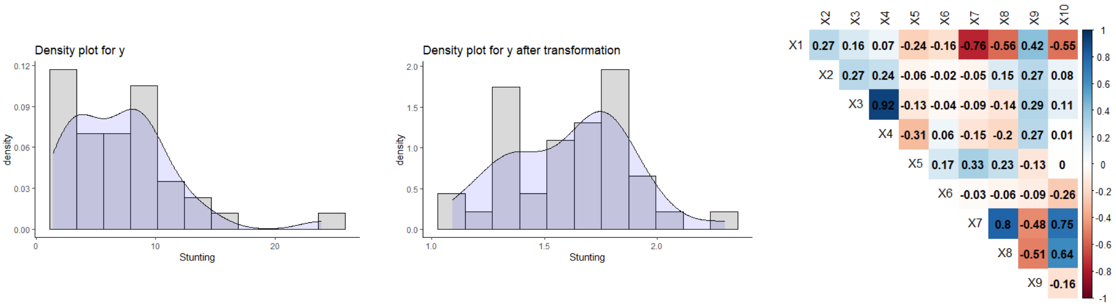


Figure 2: Map of Stunting Prevalence in Indonesia in 2024

To prepare the data for modeling, the symmetry of the response variable (stunting) was examined using density plots. The results in Figure 3(a), showed a right-skewed distribution with an extreme outlier (West Sulawesi). A Box-Cox transformation was applied to address this asymmetry, resulting in a more symmetric distribution while retaining important outlier information (Figure 3(b)). Subsequently, predictor variables were standardized using Z-score transformation to ensure comparability across variables with different units. The response variable was also centered to remove the intercept effect caused by the standardization process. Pearson correlation analysis (Figure 3(c)) revealed eight pairs of predictor variables with strong correlations ($|r| \geq 0.5$), which were later used to construct a special penalty matrix $\boldsymbol{D}_2$ to minimize multicollinearity effects in the modeling process. The following is a list of 8 variable pairs with $|r| \geq 0.5$: X_3 & X_4; X_7 & X_8; X_7 & X_{10}; X_8 & X_{10}; X_1 & X_7; X_1 & X_8; X_1 & X_{10}; and X_8 & X_9.



(a) (b) (c)

Figure 3: (a) Density plot for $\boldsymbol{y}$; (b) Density plot for $\boldsymbol{y}$ after Box-Cox transformation; (c) Pearson correlation matrix among predictor variables

To enable the application of the generalized lasso model, the original predictor variable matrix $\boldsymbol{X}$, with dimensions 38×10, was converted into a higher-dimensional matrix of 38×380. This restructuring allows each predictor variable to be uniquely associated with each province, supporting region-specific coefficient estimation. The high dimensionality ($p >> n$) enables the model to detect spatial patterns and groupings more effectively. The inclusion of the $\boldsymbol{D}_2$ penalty matrix ensures that highly correlated variables are jointly penalized, enhancing model accuracy and interpretability by reducing bias and preventing redundant parameter estimation.

The penalty matrix $\boldsymbol{D}_1$ was constructed by first forming the neighborhood matrix $\boldsymbol{D}_0$, which captures spatial adjacency relationships among provinces using three proximity measures: queen contiguity, KNN with $k = 2$, and KNN with $k = 3$. In the queen contiguity approach, provinces are considered neighbors if they share a border or vertex, resulting in 38 adjacent pairs—mainly within the same island—leading to isolated island provinces without neighbors. In contrast, the KNN approach forms neighbor relationships regardless of geographic separation by sea, yielding 47 and 72 adjacent pairs for $k = 2$ and $k = 3$, respectively. Each $\boldsymbol{D}_0$ matrix was extended into a block-diagonal structure to form $\boldsymbol{D}_1$, with dimensions scaled according to the 10 predictor variables used in the study. Additionally, the penalty matrix $\boldsymbol{D}_2$, with dimensions 304×380, was built from 8 strongly correlated variable pairs ($|r| \geq 0.5$), and integrated into the generalized lasso model to form an augmented design matrix $\boldsymbol{X}^*$ of size 342×380. To match the dimensions for regression analysis, the response vector $\boldsymbol{y}$ was extended by appending 304 zeros, resulting in $\boldsymbol{y}^*$ of size 342×1. For models without $\boldsymbol{D}_2$, the original matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ with dimensions 38×380 and 38×1, respectively, were retained.

5.2 Modeling and sensitivity analysis

The baseline modeling was conducted using the Spatial Autoregressive (SAR) model, with spatial neighborhood structures defined by queen contiguity and k-nearest neighbors (KNN) with $k = 2$ and $k = 3$. Model performance was evaluated using the AIC and RMSE. Subsequently, elastic net modeling was performed, where the technical specification was simplified into a generalized lasso model with the penalty matrix defined as the identity matrix. Several values of the tuning parameter were considered, namely $\lambda_2 = (0, 0.01, 0.1, 1, 10)$. Furthermore, the proposed generalized lasso model was compared with several alternative specifications, including spatial neighborhood definitions based on queen contiguity and KNN with $k = 2$ and $k = 3$, while also accounting for the presence of the penalty matrix $\boldsymbol{D_2}$ and several values of $\lambda_2 = (0.01, 0.1, 1, 10)$. For both the elastic net and generalized lasso models, the evaluation metrics included the selected value of λ_1, the degrees of freedom (df), the sensitivity value based on the baseline model, and the correlation between the fitted and actual values (where the actual values were those obtained from the baseline model). The baseline model used for the elastic net was the elastic net model with $\lambda_2 = 0$. Meanwhile, the baseline model used for the generalized lasso was the generalized lasso model without the $\boldsymbol{D}_2$ penalty matrix. In total, 14 models were examined in this study. The generalized lasso and elastic net modeling was carried out by determining the optimal λ_1 for each model using the ALOCV method.

The modeling results of SAR and generalized lasso as shown in Table 2, indicate that the generalized lasso approach substantially improved model performance compared to the SAR model across all spatial neighborhood specifications. For the queen contiguity structure, increasing λ_2 led to progressively lower AIC and RMSE values, with relatively stable degrees of freedom, while sensitivity and correlation showed moderate performance. For KNN with $k = 2$, the generalized lasso also reduced AIC and RMSE compared to SAR, although the sensitivity and correlation values varied considerably across different λ_2 values, indicating some instability in model fit. In contrast, for KNN with $k = 3$, the generalized lasso produced a substantial improvement in goodness-of-fit measures, particularly at $\lambda_2 = 0.01$, where the model achieved a markedly lower AIC and RMSE, a perfect correlation (1.000) with the baseline generalized lasso model (without the $\boldsymbol{D_2}$ penalty matrix), and a small sensitivity value. Importantly, in this study, a smaller sensitivity value indicates a more desirable result, as it implies that the model is not overly sensitive to the inclusion of the $\boldsymbol{D_2}$ penalty matrix and is therefore more robust. At the same time, the selected model must still demonstrate better goodness-of-fit than its baseline counterpart. Under this interpretation, the generalized lasso with KNN ($k = 3$) and $\lambda_2 = 0.01$ achieves a desirable balance: it remains sufficiently robust (low sensitivity) while providing substantial improvements in AIC and RMSE compared to the baseline model.

Table 2: Summary of SAR and generalized lasso model performances

Model	Specification	Selected λ_1	DF	AIC	RMSE	Sensitivity	Correlation
Queen contiguity							
SAR	-	-	-	19.168	0.220975	-	-
Generalized Lasso	without $\boldsymbol{D}_2$*	9.420E-07	380	-132.769	0.000008	-	-
	$\lambda_2 =$ 0.01	1.334E-05	379	-5319.527	0.000138	0.551	0.851

	$\lambda_2 = 0.1$	1.334E-05	379	-6457.927	0.000026	0.736	0.767
	$\lambda_2 = 1$	1.334E-05	379	-7630.541	0.000005	0.740	0.765
	$\lambda_2 = 10$	1.334E-05	379	-7755.206	0.000004	0.740	0.765
KNN ($k = 2$)							
SAR	-	-	-	18.134	0.215	-	-
Generalized Lasso	without $\boldsymbol{D}_2$*	0.183	34	-113.246	0.092	-	-
	$\lambda_2 = 0.01$	0.332	26	-2057.538	0.046	0.296	0.962
	$\lambda_2 = 0.1$	0.332	29	-1992.185	0.050	0.681	0.796
	$\lambda_2 = 1$	0.299	31	-1897.309	0.057	1.013	0.144
	$\lambda_2 = 10$	0.350	49	-1853.231	0.058	1.035	0.090
KNN ($k = 3$)							
SAR	-	-	-	18.435	0.216	-	-
Generalized Lasso	without $\boldsymbol{D}_2$*	0.052	30	-131.216	0.081	-	-
	$\lambda_2 = 0.01$	0.052	30	-2404.352	0.027	0.028	1.000
	$\lambda_2 = 0.1$	0.085	32	-2060.124	0.045	0.687	0.843
	$\lambda_2 = 1$	0.092	48	-1911.924	0.053	1.004	0.131
	$\lambda_2 = 10$	0.002	112	-4495.735	0.001	1.178	0.018

Note: *the generalized lasso model without the $\boldsymbol{D}_2$ penalty matrix as a baseline model for sensitivity and correlation calculations

When compared with the elastic net results (Table 3), several specifications also exhibited reasonably small sensitivity values, indicating adequate robustness to changes in λ_2. However, the elastic net models did not achieve AIC and RMSE values as low as those obtained by the generalized lasso under the KNN ($k = 3$) and $\lambda_2 = 0.01$ specification. Therefore, although the Elastic Net models were relatively robust, they were not selected because their overall goodness-of-fit was inferior to that of the best-performing generalized lasso model. Overall, considering the combined criteria of goodness-of-fit (AIC and RMSE), stability of degrees of freedom, sensitivity relative to the baseline model, and correlation between fitted and baseline values, the best-performing model remains the generalized lasso with KNN $k = 3$ and with a correlation-aware penalty $\boldsymbol{D}_2$ for $\lambda_2 = 0.01$.

Table 3: Model performances summary of elastic net as a generalized lasso simplification

Model	λ_2	Selected λ_1	DF	AIC	RMSE	Sensitivity	Correlation
Elastic Net	0*	0.136	30	-113.592	0.102	-	-
	0.01	0.427	10	-2255.392	0.066	0.816	0.635
	0.1	0.137	32	-2825.861	0.032	0.053	0.999
	1	0.481	12	-2186.386	0.071	0.894	0.597
	10	0.500	23	-2062.892	0.080	0.995	0.341

Note: *the elastic net model with $\lambda_2 = 0$ as a baseline model for sensitivity and correlation calculations

Then, bootstrap resampling was performed at the spatial-unit level using a case-weighted formulation [34] to preserve the original spatial structure encoded in the generalized lasso penalty for the best model selected. The results of bootstrap resampling based on the best model are shown in Figure 4. The boxplots show that the bootstrap results for each evaluation metric—selected λ_1 , Degrees of Freedom, AIC, and RMSE—are quite consistent, with relatively narrow ranges and no apparent outliers. This consistency indicates that the model's performance is stable across different bootstrap samples, supporting the conclusion that the selected model is reliable and well-chosen.

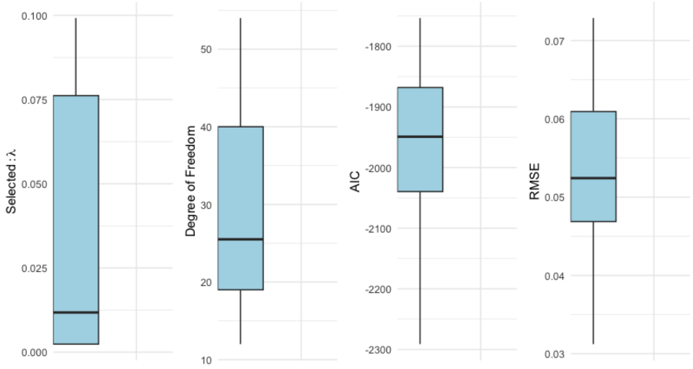


Figure 4: Boxplots of selected model performances based on bootstrap resampling

The heatmap of selected model estimates shown in Figure 5. In this study, the generalized lasso model interpreted only the effects that aligned with the general correlation between each explanatory variable and stunting. Following the principle of parsimony [35], opposing effects were considered weak and set to zero, likely due to unmeasured factors not included in the model. This selected model identified several explanatory variables that are positively correlated with stunting, including the poverty rate (X_1), Gini ratio (X_2), unmet health needs (X_5), and maternal smoking (X_9). There is a strong association between the poverty rate (X_1) and stunting in the cluster of provinces located in southern parts of Kalimantan, several provinces in Sumatra, and East Nusa Tenggara. This cluster also shows a close association with the variable unmet health needs (X_5). Additionally, there is another cluster of provinces in Java, Sulawesi, and Kalimantan that is associated with the poverty variable (X_1). Regarding the clusters formed by the association between the Gini ratio (X_2) and maternal smoking (X_9) with stunting, the cluster patterns are quite similar, with maternal smoking showing a particularly strong association. This cluster includes provinces from Sumatra, Papua, Java, and Sulawesi.

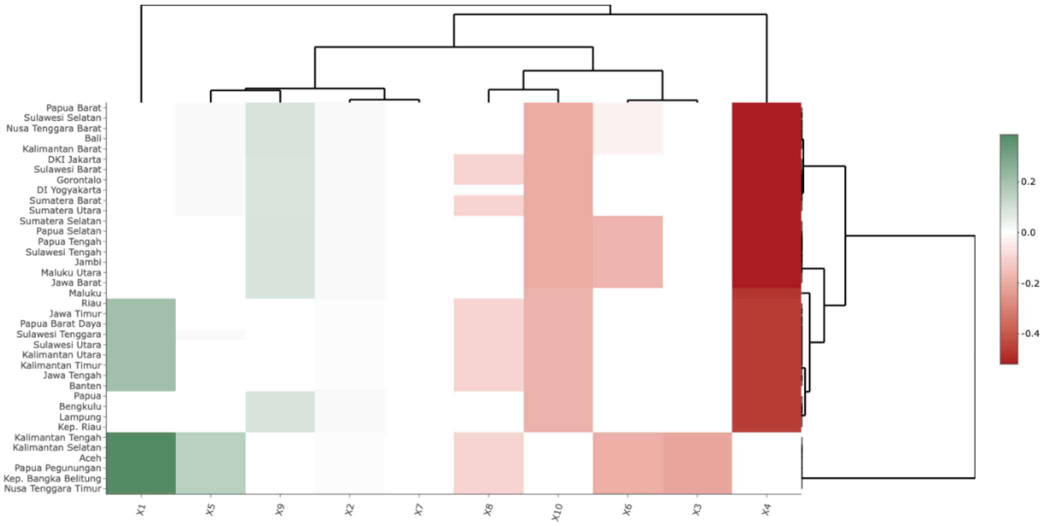


Figure 5: Heatmap of selected model estimates with agglomerative clustering

Conversely, several variables were found to have a negative correlation with stunting, indicating a protective effect. These include maternal health insurance (X_3), child health insurance (X_4), exclusive breastfeeding (X_6), access to proper sanitation (X_7), access to safe drinking water (X_8), and average years of schooling (X_{10}). The variable access to proper sanitation (X_7) tends to show no association with stunting in any region. For the variables maternal health insurance (X_3), exclusive breastfeeding (X_6), and access to safe drinking water (X_8), the clusters of associated regions are similar, primarily including southern parts of Kalimantan, several provinces in Sumatra, and East Nusa Tenggara, although other regional clusters also exist. In contrast, the variables child health insurance (X_4) and average years of schooling (X_{10}) form different clusters beyond these provinces, with the association between child health insurance (X_4) and stunting being notably stronger.

Based on the result above, the model indicates that stunting is positively associated with poverty, income inequality, unmet health needs, and maternal smoking, particularly in provinces in southern Kalimantan, several areas of Sumatra, East Nusa Tenggara, Java, Sulawesi, and Papua. The strongest positive association with stunting is observed for the variable poverty rate (X_1), particularly in southern Kalimantan, several provinces in Sumatra, and East Nusa Tenggara, with an association value of 0.386. Conversely, maternal and child health insurance, exclusive breastfeeding, access to safe drinking water, and higher average years of schooling show protective effects, with clusters in southern Kalimantan, Sumatra, and East Nusa Tenggara. The strongest negative association with stunting is observed for child health insurance (X_4), spanning provinces in Sumatra, Java, Kalimantan, Sulawesi, Bali, and Papua, with an association value of -0.518. These findings suggest that reducing economic inequality and improving access to healthcare by registering the insurance are essential strategies to effectively lower stunting prevalence across Indonesia.

6 Conclusion

In conclusion, the central and eastern regions of Indonesia (Sulawesi, Nusa Tenggara, and parts of Papua) tend to have a higher stunting prevalence compared to the western regions (Sumatra and Java) in 2024, with West Sulawesi recording the highest prevalence. The modified generalized lasso method proved to be well-suited for high-dimensional spatial data, as it effectively grouped the influence of ten explanatory variables in neighboring areas on stunting prevalence. In this study, the model incorporating the additional $\boldsymbol{D}_2$ penalty consistently yielded lower AIC and RMSE value by identifying similar effects among pairs of highly correlated explanatory variables. The KNN ($k = 3$) adjacency for the generalized lasso model with the $\boldsymbol{D}_2$, especially for $\lambda_2 = 0.01$ with sufficiently robust (low sensitivity) while providing substantial improvements in AIC and RMSE, the matrix was selected as the best-performing model. The analysis revealed that the poverty rate and child health insurance were the most influential factors affecting stunting prevalence in Indonesia in 2024, which spanned over several provinces in the main islands of Indonesia.

The government should give special attention to provinces with a high prevalence of stunting, particularly West Sulawesi, which recorded the highest stunting rate in 2024. Intervention strategies should take into account the most influential factors identified in each province to formulate more targeted and effective policies for reducing stunting prevalence. Although this study did not primarily focus on variable selection, future researchers interested in identifying the most dominant factors [36]. Furthermore, if accessibility factors—such as the ease of interregional connectivity—are to be considered, neighborhood matrices can be constructed based on inter-provincial accessibility across Indonesia [11]. This approach would allow for a more spatially sensitive analysis, which is crucial for designing context-specific interventions.

ACKNOWLEDGEMENTS
This research is supported by The Directorate General of Higher Education, Research and Technology of the Ministry of Education, Culture, Research and Technology the Republic of Indonesia based on contract number 027/E5/PG.02.00.PL/2024 for the Fundamental Research schema.

References

[1] Ministry of Home Affairs. (2024). Stunting Prevalence in Indonesia. https://konvergensi.bangda.kemendagri.go.id/emonev/DashPrev [accessed on January 23, 2025, at 16:30 WIB]

[2] Ghazali, M. F., Aqzela, A., Gracia, C., Febrianingtyas, R. S., & Wijayanti, D. (2022). Analisis Geospasial Kasus Stunting menggunakan Artificial Neural Network (ANN) di Kecamatan Gadingrejo, Pringsewu-Lampung. *Jurnal Majalah Geografi Indonesia*. 37(1):1-11. doi: 10.22146/mgi.70474

[3] Tibshirani, R. J., & Taylor, J. (2011). The Solution Path of The Generalized LASSO. *The Annals of Statistics*. 39(3): 1335-1371. doi: 10.1214/11-AOS878.

[4] Aswi, A. Rahardiantoro, S., Kurnia, A., Sartono, B., Handayani, D., Nurwan, N., & Cramb, S. (2024). Childhood stunting in Indonesia: assessing the performance of Bayesian spatial conditional autoregressive models. *Geospatial Health*, 19(2):1321. doi: 10.4081/gh.2024.1321

[5] Aswi, A. Rahardiantoro, S., Kurnia, A., Sartono, B., Handayani, D., & Nurwan, N. (2025). Bayesian spatio-temporal conditional autoregressive localized modeling techniques for socioeconomic factors and stunting in Indonesia. *MethodsX*. 15:103464. doi: https://doi.org/10.1016/j.mex.2025.103464

[6] Tibshirani. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistics Society Series B*. 58: 267-288.

[7] Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 67(2): 301-320. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] Arnold, T. B., & Tibshirani, R. J., 2016. Efficient implementations of the generalized lasso dual path algorithm. *J. Comput. Graph. Stat*. 25 (1), 1–27. https://doi.org/10.1080/10618600.2015.1008638.

[9] Rahardiantoro, S., & Sakamoto, W. (2022). Optimum tuning parameter selection in generalized lasso for clustering with spatially varying coefficient models. *IOP Conf. Series: Earth Environ. Sci*. 950 (1), 012093. doi: https://doi.org/10.1088/1755-1315/950/1/012093.

[10] Rahardiantoro, S., & Sakamoto, W. (2024). Spatio-temporal clustering analysis using generalized lasso with an application to reveal the spread of Covid-19 cases in Japan. *Comput. Stat*. 39, 1513–1537. doi: https://doi.org/10.1007/s00180-023-01331-x.

[11] Rahardiantoro, S., Oktarina, S.D., Kurnia, A., Maharani, N.S., & Juhanda, A.R.N. (2024). Spatio-temporal clustering using generalized lasso to identify the spread of Covid-19 in Indonesia according to provincial flight route-based connections. *Spat. Stat*. 63 (100857). doi: https://doi.org/10.1016/j.spasta.2024.100857.

[12] Rahardiantoro, S., & Sakamoto, W. (2021). Clustering regions based on socio-economic factors which affected the number of COVID-19 Cases in Java Island. *J. Phys.: Conference Series 1863* (1), 012014. doi: https://doi.org/10.1088/1742-6596/1863/1/012014.

[13] Kurnia, A., Rahardiantoro, S., Oktarina, S. D., Anisa, R., Rahman, N. A. N., & Handayani, D. (2024). Modified Generalized Lasso for Variable Selection in Lag Distributed Modeling of Fresh Fruit Bunch Production from Oil Palm Plantations in Riau-Indonesia. *Int. J. Advance Soft Compu. Appl*. 16 (1), 1–17. doi: https://doi.org/10.15849/ IJASCA.240330.01.

[14] Rahardiantoro, S., Juhanda, A. R. N., Kurnia, A., Aswi, A., Sartono, B., Handayani, D., Soleh, A. M., Yanti, Y., & Cramb, S. (2024). Spatio-temporal modeling to identify factors associated with stunting in Indonesia using a Modified Generalized Lasso. *Spatial and Spatio-temporal Epidemiology*. 51 (100694). doi: https://doi.org/10.1016/j.sste.2024.100694

[15] Wang, S., Zhou, W., Maleki, A., Lu, H., & Mirrokni, V. (2018). Approximate Leave-One-Out for High- Dimensional Non-Differentiable Learning Problems. *arXiv:1810.02716*

[16] Montgomery, D. C., Peck, E. A, & Vining, G. G. (2012). *Introduction to Linear Regression Analysis 5th Edition*. Hoboken (NJ): John Wiley & Sons Publication.

[17] Statistics Indonesia. (2024). Unmet Need Pelayanan Kesehatan 2024. [accessed on 2025 May 28]. https://kepri.bps.go.id/en/statistics-table/2/Mjk3IzI=/unmet-need-pelayanan-kesehatan.html.

[18] Statistics Indonesia. (2024). Persentase Bayi Usia Kurang Dari 6 Bulan Yang Mendapatkan Asi Eksklusif Menurut Provinsi (Persen), 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/MTM0MCMy/persentase-bayi-usia-kurang-dari-6-bulan-yang-mendapatkan-asi-eksklusif-menurut-provinsi.html.

[19] Statistics Indonesia. (2024). Persentase Rumah Tangga menurut Provinsi dan Memiliki Akses terhadap Sanitasi Layak (Persen), 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/ODQ3IzI=/persentase-rumah-tangga-menurut-provinsi-dan-memiliki-akses-terhadap-sanitasi-layak.html.

[20] Statistics Indonesia. (2024). Persentase Rumah Tangga yang Memiliki Akses terhadap Sumber Air Minum Layak Menurut Provinsi (Persen), 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/ODQ1IzI=/persentase-rumah-tangga-menurut-provinsi-dan-sumber-air-minum-layak--persen-.html.

[21] Statistics Indonesia. (2024). Persentase Unmet Need Pelayanan Kesehatan Menurut Provinsi (Persen), 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/MTQwMiMy/unmet-need-pelayanan-kesehatan-menurut-provinsi.html.

[22] Statistics Indonesia. (2024). Profil Kesehatan Ibu dan Anak 2024 Vol. 10. [accessed on 2025 Jan 22] https://www.bps.go.id/id/publication/2024/12/31/a919c55a72b74e33d011b0dc/profil-kesehatan-ibu-dan-anak-2024.html.

[23] Statistics Indonesia. (2024). Rata-Rata Lama Sekolah Penduduk Umur 15 Tahun ke Atas Menurut Provinsi, 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/MTQyOSMy/rata-rata-lama-sekolah-penduduk-umur-15-tahun-ke-atas-menurut-provinsi.html.

[24] Statistics Indonesia. (2025). Gini Ratio Menurut Provinsi dan Daerah, 2024. [accessed on 2025 Jan 22]. Tersedia dari: https://www.bps.go.id/id/statistics-table/2/OTgjMg==/gini-ratio-menurut-provinsi-dan-daerah.html.

[25] Statistics Indonesia. (2025). Persentase Penduduk Miskin (P0) Menurut Provinsi dan Daerah (Persen), 2024. [accessed on 2025 Jan 22]. https://www.bps.go.id/id/statistics-table/2/MTkyIzI=/persentase-penduduk-miskin--september-2024.html.

[26] Karyati. (2021). Pengaruh Jumlah Penduduk Miskin, Laju Pertumbuhan Ekonomi, dan Tingkat Pendidikan Terhadap Jumlah *Stunting* di 10 Wilayah Tertinggi Indonesia Tahun 2010-2019. *Journal Riset Ilmu Ekonomi dan Bisnis*. 1(2). doi:10.29313/jrieb.v1i2.401.

[27] Resfaliza, Kamami, N., & Purwasutrisno. (2024). Kausalitas Antara Variabel Makro dan *Stunting* di Sumatera Barat Periode 2021-2023. *Jurnal Informatika Ekonomi Bisnis.* 6(3):658-668. doi:10.37034/infeb.v6i3.964.

[28] Husain, H., Dewi, A. F., & Wardani, A. E. (2024). Pemodelan Prevalensi *Stunting* Indonesia Menggunakan Regresi Nonparametric Spline Truncated. 2024. *Journal of Analytical Research, Statsitics and Computation*. 3(1). doi:10.4590/jarsic.v3i1.26.

[29] Hikmahrachim, H. G., Rohsiswatmo, R., & Ronoatmodjo, S. (2020). Efek ASI Eksklusif Terhadap *Stunting* pada Anak Usia 6-59 Bulan di Kabupaten Bogor Tahun 2019. *Jurnal Epidemiologi Kesehatan Indonesia*. 3(2). doi:10.7454/epidkes.v3i2.3425.

[30] Astuti, Y. R. (2022). Pengaruh Sanitasi dan Air Minum Terhadap *Stunting* di Papua dan Papua Barat. *Poltekita: Jurnal Ilmu Kesehatan*. 16(3):261-267. doi:10.33860/jik.v16i3.1470.

[31] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers. Dordrecht: Boston.

[32] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*. 46(3): 175–85. doi:10.2307/2685209

[33] Flynn, C. J., Hurvich, C. M., & Simonoff, J. S. (2013). Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models. *Journal of the American Statistical Association, 108*(503), 1031–1043. https://doi.org/10.1080/01621459.2013.801775

[34] Abram, S. V., Helwig, N. E., Moodie, C. A., DeYoung, C. G., MacDonald, A. W. III., & Waller, N. G. (2016). Bootstrap Enhanced Penalized Regression for Variable Selection with Neuroimaging Data. *Front. Neurosci*. 10:344. doi: 10.3389/fnins.2016.00344

[35] Goloboff, P.A., Torres, A., & Arias, J.S., (2018). Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*. 34 (4), 407–437. doi: https://doi.org/10.1111/cla.12205.

[36] Jiang, Y., He, Y., & Zhang, H. (2017). Variable Selection with Prior Information for Generalized Linear Models via the Prior LASSO Method. *J Am Stat Assoc*. 111(513):355–376. doi: 10.1080/01621459.2015.1008363

Notes on contributors

Septian Rahardiantoro is a lecturer at the Study Program in Statistics and Data Science, IPB University, Bogor, Indonesia. His main teaching and research interests include Data Science, Statistical Machine Learning, and Statistical Modeling in Environmental Science.

Aida Darajati is a graduate of the Statistics and Data Science Program at IPB University. Her research interests focus on statistical modeling related to stunting.

Hari Wijayanto is a Professor at the Study Program in Statistics and Data Science, IPB University, Bogor, Indonesia. His main teaching and research interests include Sampling and Survey Methodology, Statistical Modeling, and Machine Learning. He has published several research articles in international journals of statistics, data science, and its applications.

Anang Kurnia is a Professor at the Study Program in Statistics and Data Science, IPB University, Bogor, Indonesia. His main teaching and research interests include Data Science, Statistical Machine Learning, Statistical Inference, Generalized Linear Mixed Model, and Small Area Estimation. He has published several research articles in international journals of statistics and data science.