Estimated North Sumatra Province Poor Population Percentage Using Penalized Spline Semiparametric Approach and Small Area Estimation

Jihan Adelia Nasution*, Rina Widya Sari, Ismail Husein
Department of Mathematics, UIN Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Keywords:

Poor People, Small Area Estimation, Semiparametric, Penalized Spline

ABSTRACT

Poverty is one of the many problems that have not been completely resolved by the government in Indonesia, one of which is poverty in the province of North Sumatra. To estimate the percentage of poor people, data is needed in each area using the Small Area Estimation method. Small Area Estimation is used to estimate the parameters of a subpopulation that has a small scope. However, to get a better estimate, you can use an indirect estimation method, one of which is the semiparametric Penalized Spline approach. This method can be used in conjunction with small area estimation because it can connect the two components in the model between the response variable and the predictor variable which is linear and the relationship between the response variable and the predictor variable is non-linear. Based on the small area estimation model with a semiparametric penalized spline approach, the best is found in model 4 with a coefficient of determination value of 0.645 where the value is close to 1, which means the results are good to use. The average poor population in North Sumatra province is estimated at 15.38%, the highest poor population is in Pakpak Bharat at 22.66% and the lowest estimated poor population is in Deli Serdang at 7.51%.

Corresponding Author:

Jihan Adelia Nasution,
Department of Mathematics,
UIN Sumatera Utara, Medan, Indonesia
Email: Jihanadelianasution@gmail.com

1. INTRODUCTION

Poverty in the province of North Sumatra is just one of the numerous issues that the Indonesian government has not been able to fully resolve[1]. Data shows that while the number of impoverished individuals is still erratic, it has declined by 172.91 thousand between 2016 and 2019[2]. There are still a large number of citizens who fall into the poverty category, despite the fact that the rate has dropped. The government of Indonesia has developed initiatives aimed at reducing poverty; these include the Indonesian Health Card (KIS), direct cash support (BLT), rice subsidies for the poor (Raskin), and the poor student assistance program (BSM)[3]-[5].

A poor person is someone who is typically defined as having experienced some form of deprivation, such as economic hardship, substandard housing, a lack of education, or a lack of socialization. As a result, poor people often struggle to obtain employment, end up unemployed, and rely solely on government assistance [6]. Researchers raised this title in order to determine the average number of impoverished individuals in North Sumatra province and its cities/districts, as well as the highest and lowest percentages of impoverished individuals and the province's average level of education. Next, you can attempt to estimate a small area through data optimization utilizing Small Area Estimation (SAE) in order to determine the total number of impoverished people in each location [7]-[9].

2. RESEARCH METHODE

A statistical method called small area estimation is used to estimate subpopulation parameters for which there is either no sample at all or a limited sample size.

Area-Based Model

Models exploring the interaction between supporting data from various domains within an area and direct estimates of tiny areas are examples of area-based models. Because the research employs district/city area level data, this is the model that the author uses. The model's general form is:

$$\hat{\theta} = x_{ij}^{T}\beta + b_i v_i + e_i, i = 1,2,\dots,m \quad (1)$$

Model Based on Units

Unit-based models are those that have the ability to integrate supporting variable unit values with corresponding direct estimation variables. It is assumed that the unit-based model has the following general shape.

$$y_{ij} = x_{ij}^{T}\alpha + v_i + e_{ij}, j = 1,2,\dots,n, i, j = 1,\dots,m, v_i \sim N(0,\sigma_{ei}^{2}) \quad (2)$$

Regression of Splines

The generalized additive model is the semiparametric regression tool that researchers employ (GAM). For instance, paired data, such as (yi, xi, ti), are assumed in the formula.

$$y_i = X_i\beta + f(t_i) + \varepsilon_i, i = 1,2,\dots,n \quad (3)$$

Nonparametric regression, on the other hand, generally takes the following form.

$$y_i = m(x_i) + e_i, i = 1,2,\dots,n \quad (4)$$

Since $m(x_i)$ has a nonlinear structure, its definition is knots (k) where $k_1,\dots,k_n$ and its basis function is a discontinuous polynomial. then the model that follows is produced.

$$m(x_i) = \beta_0 + \beta_i x_i + \dots + \beta_p x_i^p + \sum_{j=1}^{k} y_j (x_i - k_j)_+^{\,p} \quad (5)$$

Thus, it can be expressed as:

$$y_i = \beta_0 + \beta_i x_i + \dots + \beta_p x_i^p + \sum_{j=1}^{k} y_j (x_i - k_j)_+^{\,p} e_i \quad (6)$$

Regresi Penalized Spline

Choosing the smoothing character is the first step, and choosing the knots and location is the second. overall form

$$y = X\beta + Z\gamma + e \quad (7)$$

The penalized least squares (PLS) function's minimum procedure yields the penalized spline estimator, namely

$$L = \left\|y - X\beta - Z\gamma\right\|^2 + \lambda\gamma^T\gamma \quad (8)$$

by example $c = [X, Z]$ and $\ddot{\theta} = \begin{bmatrix}\beta \\ \gamma\end{bmatrix}$ so that the example in the equation

$$L = \left\|y - C\ddot{\theta}\right\|^2 + \lambda\ddot{\theta}^T D\ddot{\theta} \quad (9)$$

Where D is the penalty matrix

$$\begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \vdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{(p+1)x2} & \mathbf{0}_{(p+1)xK} \\ \mathbf{0}_{kx(P+1)} & \boldsymbol{I}_{KxK} \end{bmatrix}$$

D Using the smoothing parameter λ, where $\lambda \leq 0$. The roughness penalty is the second term, and the sum of the squared errors is the first (Apriani, 2017). Thus, the resulting penalized spline estimator is:

$$\hat{\ddot{\theta}} = (C^T C + \lambda \mathrm{D})^{-1} C^T y \quad (10)$$

then it is obtained $\hat{y} = C\hat{\ddot{\theta}}$

$$\hat{y} = C(C^T C + \lambda \mathrm{D})^{-1} C^T y \quad (11)$$

Selection of the Optimal Number of Knots (K).

The following is the generalized cross-validation form:

$$GCV(K) = \frac{MSE(K)}{\left[n^{-1}trace\left(I - A(K)\right)\right]^2} \quad (12)$$

Meanwhile, fixed selection typically takes the following form:

$$K = \min\left(\frac{1}{4} * banyak\ x_i yang\ berbeda, 35\right) \quad (13)$$

Fixed selection is often calculated using quantiles for Kk at various x_i which are expressed as follows.

$$K_k = \left(\frac{k+1}{k+2}\right), k = 1,2, \dots, K \quad (14)$$

Small Area Estimation with the Semiparametric Penalized Spline Approach

$$K = \min\left(\frac{1}{4} * banyak\ x_i yang\ berbeda, 35\right) \quad (15)$$

[10] The small area model can be written as follows:

$$\theta_i = x_{ij}^T\alpha + b_i v_i + e_i, i = 1,2, \dots, n;\ v_i \sim N(0, \sigma_v^2) \quad (16)$$

The spline function for a semiparametric model with a single response, x_1 is expressed as $m(x_1)$ where the value of m(.) is assumed to be sufficiently good but not known.

$$m(x_i) = \beta_0 + \beta_i x_i + \cdots + \beta_p x_i^p + \sum\nolimits_{j=1}^{k} \gamma_j \left(x_1 - k_j\right)_+^{\ p} \quad (17)$$

The formula to get the Mean can be used.

$$\bar{X} = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i \quad (18)$$

Tthe formula to get the variance can be used.

$$s^2 = \frac{1}{(n-1)}\sum\nolimits_{i=1}^{n} (X_i - \bar{X})^2 \quad (19)$$

The formula for standard deviation can be used.

$$\sigma = \sqrt{\sigma^2} \quad (20)$$

Calculating the standard deviation of the bad product proportion (Y)

$$\sigma = \sqrt{22{,}17} = 4{,}71 \quad (21)$$

The variables included in the investigation were subjected to descriptive analysis. This is to give data-related information. There are minimum and maximum values, variance, standard deviation, and mean value for the data exploration. An examination of the data shown in Table 1 is provided below.

Table 1 Summary of Estimated Values in North Sumatra

Variable	Mean	Variance	S. Deviation	Min	Max
Y	10,80	22,17	22,17	22,17	22,17
X_1	56,24	4,71	4,71	4,71	4,71
X_2	5,45	3,88	3,88	3,88	3,88
X_3	3,03	25,69	25,69	25,69	25,69
X_4	38,20	16,49	16,49	16,49	16,49

Figure 1 illustrates that in 2020, West Nias district had the largest percentage of impoverished individuals in North Sumatra Province (25.69%), while Deli Serdang district had the lowest percentage (3.88%). Figure illustrates the differences for further information.

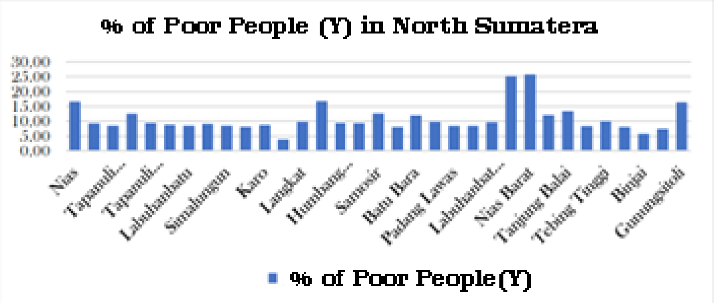

Figure 1. Graph of the Percentage of Poor Population in North Sumatra Province in 2020

3. RESULT AND ANALYSIS

Use small area estimation and the penalized spline semiparametric technique while conducting research to estimate the percentage of the population that is impoverished. This can be accomplished by searching
Using the formula, the correlation coefficient value indicates whether or not the linear link is weak.

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2\,][n\sum y^2 - (\sum y^2)]}} \tag{22}$$

$$r_{x1y} = \frac{668574{,}92 - 661349{,}22}{\sqrt{[3357054{,}44 - 104907{,}95][145829{,}41]}} = \frac{7225{,}70}{699683{,}70} = 0{,}01$$

$$r_{x2y} = \frac{59984{,}496 - 64119{,}82}{\sqrt{[39800{,}9984 - 1243{,}78][145829{,}41]}} = \frac{-4135{,}324}{76185{,}014} = -0{,}05428$$

$$r_{x3y} = \frac{178819{,}86 - 35634}{\sqrt{[21779{,}93 - 680{,}62][145829{,}41]}} = \frac{143185{,}86}{56357{,}38} = 2{,}54$$

$$r_{x4y} = \frac{499201{,}55 - \ 449148{,}75}{\sqrt{[1787812{,}12 - 55869{,}13][145829{,}41]}} = \frac{50052{,}79}{510603{,}17} = 0{,}10$$

Tabel 2 Value of Correlation between Variables

Correlation	Correation Coefficient
Y with X_1	0,01
Y with X_2	-0,05428
Y twith X_3	2,54
Y with X_4	0,10

As can be observed, there appears to be a higher correlation between these variables and the other variables because the coefficient value of y on X3 is greater than that of the other variables.

Estimated North Sumatra Province Poor Population Share Using a Semiparametric Method Penalized Spline

Using the formula, you can determine the knot point value.

$$y_j = \sum_{j=0}^{m} \beta_{j+m}\left(x_1 - k_j\right)^m_{\ +} + \cdots + \sum_{j=1}^{k} \beta_{j+m}\left(x_4 - k_j\right)^m_{\ +} + \varepsilon_j\,[1, (x_1 - k_1), (x_2 - k_2), (x_3 - k_3), (x_4 - k_4)] \tag{23}$$

Finding the Values of Generalized Cross Validation (GCV)
The generalized cross validation formula's general form is

$$GCV_{(k1,k2,\ldots,kx)} = \frac{MSE_{(k1,k2,\ldots,kx)}}{\left[n^{-1}trace\left(I - A(k1,k2,\ldots,kx)\right)\right]^2}$$

Table 1: Estimated Values Summary for North Sumatra

Optimal Knot Point	MSE	GCV
X1 = 48.29	18.68922	21.17853
X2 = 2.84	19.53223	22.13381
X3 = 1.35	18.98774	21.51681
X4 = 18.2	12.47989	14.14214

Model formation in subsequent study can be separated into two phases. The Generalized Additive Model approach, or GAM for short, can be used to build this. By employing smoothing techniques, GAM can also be utilized to mitigate the nonlinear impact of each predictor variable on the response variable. The game's model is

$$y = f(X) + m(t_i) + u$$

This study looked at four different model types to find the optimal gaming model. Model 1 of the models uses $X1$, $X2$, and $X3$ to estimate its parametric components, whilst model 4 purports to use a nonparametric penalized spline with a model shape.

$$y = f(X_1, X_2, X_3) + f(X_4) + u$$

For model 2, i.e., the model where the parametric components are estimated using X_1, X_2, and X_4, the nonparametric penalized spline with model form is used to estimate X_3.

$$y = f(X_1, X_2, X_4) + f(X_3) + u$$

Next, for model 3, in which the parametric components are estimated using variables X1, X2, and X3, and the nonparametric penalized spline with model form is used to estimate X2.

$$y = f(X_1, X_3, X_4) + f(X_2) + u$$

Finally, model 4 is the case when variables $X2$, $X3$, and $X4$ are used to estimate the parametric components, while $X1$ is evaluated using a nonparametric penalized spline with a model form.

$$y = f(X_2, X_3, X_4) + f(X_1) + u$$

Choosing the Best Model

Following the acquisition of five models, one model—small area estimation—was achieved parametrically, while the remaining four—the semiparametric penalized spline approach—were obtained semiparametrically. The next stage is to use the coefficient of determination to identify which of the five models is the best.

Model	R^2
SAE Model Estimate	0.467
Model 1	0.445
Model 2	0,470
Model 3	0,489
Model 4	0,474

We may compare the coefficient of determination values for the five models 2 after glancing at the table. Subsequently, the maximum value of R was found, which is 0.489. This value is in close proximity to 1, indicating that the model is sufficiently accurate in estimating the proportion of impoverished individuals in North Sumatra Province and has a positive correlation with all variables.

4. CONCLUSION

The best model, according to the semiparametric penalized spline approach for small area estimation, is model 3, which has a coefficient of determination value of 0.489, which is close to 1, indicating that the results are suitable for usage.

Using the third model, the estimated impoverished population in North Sumatra province is 18.98% on average; Pematang Siantar has the most estimated impoverished population (25.38%), while Pakpak Barat has the lowest estimated impoverished population (11.92%).

REFERENCES

[1] Mulyono, Edy, S. (2017). Kemiskinan dan Pemberdayaan Masyarakat. Yogyakarta: Penerbit Ombak.

[2] Badan Pusat Statistik Provinsi Sumatera Utara. (2020, Maret). Profil Kemiskinan Provinsi Sumatera Utara Maret 2020. Dipetik Desember 30, 2020, dari https://sumut.bps.go.id/.

[3] Beik I, Arsyianti L. (2017). Ekonomi Pembangunan Syariah. Jakarta: Rajawali Pers.

[4] Soleh, A. (2018). Analisis dan Strategi Pengentasan Kemiskinan di Provinsi Jambi. Jurnal Ilmiah

[5] Riyaldi, H. (2017). Kedudukan dan Prinsip Pembagian Zakat dalam Mengatasi Permasalahan

[6] Alhudori, M. (2017). Pengaruh IPM, PRDB, dan Jumlah Pengangguran Terhadap Penduduk Miskin

[7] Aristi, *et al.* (2018). Small Area Estimation Terhadap Kemiskinan di Ketapang dengan Pendekatan

[8] Azmi, S. (2019). Small Area Estimation Terhadap Angka Kelahiran di Kota Medan dengan Menggunakan Pendekatan Kernel-Bootstrap. Skripsi. Universitas Islam Negeri Sumatera Utara. Medan.

[9] Febriani, E., Yozza, H., Rahmi, I. (2019). Pendugaan Jumlah Penduduk Miskin di Indonesia Pada Suatu Area Kecil dengan Pendekatan Kernel-Bootstrap. Jurnal Matematika Unand, 8(3),39-46.

[10] Ningrum, M., Satyahadewi, N., Debataraja, N. (2020). Pemodelan Faktor-Faktor yang Mempengaruhi Indeks Pembangunan Manusia di Indonesia Menggunakan Regresi Semiparametrik Spline.