IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 11, No. 8, August, 2024, pp. 2100-2108
Journal homepage: http://iieta.org/journals/mmep

Poverty Modeling in North Sumatera Province Considering County Location Using Geographical Weighted Regression and LASSO

Open Darnius[1*], Yuli Greace Cesilia Turnip[2], Sutarman[2], Enita Dewi Tarigan[1], Tulus Joseph Marpaung[1], Muhammad Romi Syahputra[2], Benar Surbakti[3], Israil Sitepu[4]

[1] Department of Statistics, Faculty of Vocational, Universitas Sumatera Utara, North Sumatera 20155, Indonesia
[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, North Sumatera 20155, Indonesia
[3] Diploma Program Mechanical Engineering, Politeknik Negeri Medan, North Sumatera 20155, Indonesia
[4] Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Katolik Santo Thomas, North Sumatera 20155, Indonesia

Corresponding Author Email: open@usu.ac.id

https://doi.org/10.18280/mmep.110811

Received: 1 May 2024
Revised: 16 July 2024
Accepted: 24 July 2024
Available online: 28 August 2024

Keywords:
spatial data, Geographically Weighted Regression, multicollinearity, Least Absolute Shrinkage and Selection Operator, Least Angle Regression

ABSTRACT

Spatial data is data that contains the influence of location with non-homogeneous variance at each location, or spatial heterogeneity. To address spatial heterogeneity, the Geographically Weighted Regression (GWR) model is used. However, in the GWR model, there is a phenomenon of multicollinearity, which is a strong relationship between independent variables that will reduce the accuracy of parameter estimation. To overcome multicollinearity in the GWR model, the Least Absolute Shrinkage and Selection Operator (LASSO) method is used. The LASSO method estimates the parameters of the GWR model by minimizing the sum of squared errors subject to a constraint function, which is solved using the Least Angle Regression (LARS) algorithm. This results in the Least Absolute Shrinkage and Selection Operator (LASSO) regression model to address the problem of multicollinearity in spatial data. Based on the research results, the LASSO method can overcome multicollinearity by shrinking the coefficients of parameters that contribute less and have a strong correlation with other independent variables in the GWR model, resulting in 33 final models. One of the models is for Nias Regency, where the factors influencing the poverty rate are the open unemployment rate, life expectancy, average length of schooling, gross participation rate, and per capita income. In Nias Regency, the value of s is 0.288 with an R-squared value of 0.9403. In Nias Regency, 94.03% of the variation in the poverty rate is explained by the independent variables in the model, while the remaining 5.97% is attributed to external factors not covered by the model. Coefficient of the Human Development Index variable shrinks to exactly zero, indicating that it has no effect on the poverty rate in Nias Regency.

1. INTRODUCTION

Statistics is the study of data collection, analysis, and interpretation. In statistics, there are several types of data, one of which is spatial data, which is data that contains location effects. The existence of spatial influences on data results in spatial diversity. This spatial diversity is a condition when the independent variable cannot explain its effect due to differences in characteristics between locations.

Geographically Weighted Regression is one of the regression models used to analyze spatial heterogeneity [1]. GWR is a regression analysis performed on each observation location. So that different regression models are obtained at each observation location. This diverse regression model is obtained from the addition of a different weight matrix at each location.

The Geographically Weighted Regression model often suffers from multicollinearity, which is a condition where independent variables are highly correlated. This can significantly affect the accuracy of parameter estimations within the model.

The LASSO (Least Absolute Shrinkage and Selection Operator) technique offers a way to tackle multicollinearity issues in GWR models. By adding a penalty term to the regression model's objective function, LASSO effectively diminishes the impact of less significant or highly correlated parameters. This approach helps identify the subset of independent variables that are most critical in predicting the dependent variable.

This study investigates the factors contributing to the poverty rate in North Sumatra, which remains a significant concern despite a decline in recent years. Between September 2020 and March 2023, the number of individuals living in poverty decreased from 1.3 million to 1.24 million, reflecting

a 0.18% reduction in the poverty rate. However, North Sumatra's poverty rate is still higher than the national average, prompting the need for research to identify the underlying causes.

Recognizing that geographical differences can influence poverty determinants, this study utilizes the Geographically Weighted Regression (GWR) model to analyze these factors in each district/city of North Sumatra. By understanding the unique challenges faced by different regions, this approach enables the design of targeted poverty reduction programs that cater to the specific needs and characteristics of each area.

Several studies have explored methods to address multicollinearity [2]. Conducted research on economic growth modeling in West Kalimantan using the LASSO approach and found it to be the most effective method for handling multicollinearity [3]. Utilized the GWR method to estimate the dominant factors influencing poverty in Jambi Province, revealing that these factors vary across districts/cities within the province [4]. Employed the Ridge method in a GWR model with multicollinearity and found that the GWR model became significant for infant mortality rate data in East Java in 2012, suggesting the use of the LASSO method for further research.

By implementing the LASSO method in GWR models, the accuracy of parameter estimation can be significantly improved by mitigating the effects of multicollinearity. LASSO not only helps to identify the most influential independent variables but also reduces the impact of highly correlated ones. As a result, the parameter estimation in the GWR model becomes more precise and relevant in describing the spatial relationship between the independent and dependent variables.

This research only obtained the final model in addressing the presence of multicollinearity in GWR using the Least Absolute Shrinkage and Selection Operator method.

2. BASIC THEORY

2.1 Multiple linear regression

Multiple linear regression is a model that explains the effect of one dependent variable (Y) with two or more independent variables $(X_1, X_2, \dots, X_p)$. The multiple linear regression equation model can be expressed mathematically as follows [5]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i; i = 1,2,\dots,n \quad (1)$$

where,

Y_i: the dependent variable value at the i^{th} observation,

β_0: intercept,

β_k: the k^{th} regression parameter,

X_{ik}: the i^{th} observable of the k^{th} independent variable, (k=*1,2,3..., p*)

ε_i: error at the i^{th} observation.

2.2 Spatial data

Spatial data refers to data related to location or space. It includes geographic information, such as geographic coordinates, borders, maps, satellite images, topographic data, and other information related to geographic dimensions.

2.3 Spatial heterogeneity test

Spatial heterogeneity is a condition where the global regression model is unable to explain between variables because of the diversity of characteristics between regions [6]. Spatial heterogeneity testing is done through Breusch-Pagan testing [6].

$$BP = \left(\frac{1}{2}\right) f^T Z (Z^T Z)^{-1} Z^T f \sim \chi^2 \quad (2)$$

with:

$$f_i = \frac{\varepsilon_i^2}{\sigma^2} - 1$$

where,

ε_i: error for the i^{th} observation,

σ^2: variance,

Z: matrix of size (n×*1*) which is the standardized value of X for each i^{th} observation.

2.4 Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a development of global regression methods that takes into account location aspects and fulfills the assumption of spatial heterogeneity, resulting in different parameter estimates for each location. The GWR model depends on the weights used at each observation location based on their geographic location. The elements of the weight matrix are determined based on the regression point and the observation point. Large weights are obtained based on points that are close to the observation point. This weighting depends on the selection of the optimal bandwidth [7]. The GWR model is an extension of a regression model in which each parameter is estimated at each point [8]. The GWR model can be expressed as follows [9]:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \quad (3)$$

where,

y_i: the value of the dependent variable at the i^{th} observation location,

X_{ik}: the value of the k^{th} independent variable at the i^{th} observation location,

u_i, v_i: coordinates of point i with u_i is Longitude and v_i is Latitude,

$\beta_k(u_i, v_i)$: k^{th} local parameter at observation location (u_i, v_i),

In estimating GWR parameters, there are several steps, namely:

a. Calculating spatial weights

A spatial weighting matrix is essentially a matrix that shows how different regions relate to each other [10]. Spatial weights are contained in a diagonal matrix that shows the proximity between observation locations whose function is to estimate different parameters at each observation location [11].

$$W(u_i, v_i) = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right] \quad (4)$$

with [11]:

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (5)$$

where, d_{ij} is the euclidean distance measures the straight-line

distance between two locations (u_i, v_i) and (u_j, v_j), while the smoothing parameter (bandwidth) controls the degree of smoothing applied in spatial analysis methods like kernel density estimation.

b. Optimum bandwidth selection

In Geographically Weighted Regression (GWR), bandwidth acts as a weight to achieve a balance between how well the curve fits the data points and the overall smoothness of the curve in the model. The best bandwidth value is the one that minimizes the Cross-Validation (CV) value, which is a measure of the model's predictive accuracy.

$$CV(h) = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(h)\right)^2 \quad (6)$$

c. GWR parameter estimation

Parameter estimation in the GWR model uses the Weighted Least Square (WLS) method, which is by giving different weights to each observation location [12]. Parameter estimates based on the weighting matrix will minimize the sum of squared residuals or Sum Square Error (SSE) so that it is obtained:

$$\sum_{i=1}^{n} W(u_i, v_i)\varepsilon_i^2 = \sum_{i=1}^{n} W(u_i, v_i)[y_i - \beta_0(u_i, v_i)] - \sum_{k=1}^{p} \beta_k(u_i, v_i) X_{ik} \quad (7)$$

so that the parameter estimates of the GWR model for each observation location are obtained as follows:

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i) X]^{-1} X^T W(u_i, v_i) \quad (8)$$

d. Hypothesis testing of GWR model

1) Goodness of Fit (GoF) test on GWR aims to determine the best model between GWR model and multiple linear regression model.
2) t-Test is performed to identify which independent variables have a significant individual impact on each dependent variable.

2.5 Multicollinearity

Multicollinearity was initially identified by Ragnar Frisch, who observed a linear relationship among some or all of the independent variables in a regression model [13]. When multicollinearity symptoms appear in the model, it leads to an increase in the variance of the regression coefficients. This rise in variance results in the following effects [14]:

1) Testing regression parameters using the t test becomes invalid.
2) There is a contradiction between the results of simultaneous parameter hypothesis testing through the F test and the results of partial regression parameter testing through the t test.

According to the study conducted by Darnius and Tambunan [14], multicollinearity can be detected using the Variance Inflation Factor (VIF) value. The VIF value can be found using the formula:

$$VIF_k = \frac{1}{(1 - R_k^2)}, k = 1,2,\ldots,p \quad (9)$$

R_k^2 is the coefficient of determination obtained from the independent variable X_k regressed with other independent variables. If the VIF_k value is greater than 10, there is a multicollinearity problem.

2.6 Least Absolute Shrinkage and Selection Operator

Definition

Consider data $(\boldsymbol{x^i}, \boldsymbol{y_i}), \boldsymbol{i = 1, 2, \ldots, n}$ with $\boldsymbol{x^i} = \left(\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{ip}}\right)^{\boldsymbol{T}}$ representing the independent variables and $\boldsymbol{y_i}$ representing the dependent variables. As in the standard regression framework, we either assume that the observations are independent or that $\boldsymbol{y_i s}$ are conditionally independent given $\boldsymbol{x_{ij} s}$ [15]. We assume the $\boldsymbol{x_{ij}}$ are standardized such that $\frac{\sum_i x_i}{n} = \boldsymbol{0}$ and $\frac{\sum_i x_{ik}^2}{n} = \boldsymbol{1}$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}, \ldots, \widehat{\boldsymbol{\beta}}_{\boldsymbol{p}})$, the LASSO estimate $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}})$ is defined by:

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_k \beta_k x_{ik}\right)^2\right\} \quad (10)$$

Subject to $\sum_j |\beta_j| \leq t$.

Here $\boldsymbol{t \geq 0}$ tuning parameter. For any value of $\boldsymbol{t}$, the solution for $\boldsymbol{\beta_0}$ is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} = \overline{\boldsymbol{y}}$. Without loss of generality, we can assume $\overline{\boldsymbol{y}} = \boldsymbol{0}$ and thus omit $\boldsymbol{\beta_0}$ [16].

The parameter $\boldsymbol{t \geq 0}$ governs the degree of shrinkage applied to the estimates. Let $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{\boldsymbol{0}}$ denote the full least squares estimates, and let $\boldsymbol{t_0} = \sum |\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{\boldsymbol{0}}|$. Values $\boldsymbol{t < t_0}$ will lead to shrinkage of the estimates towards zero, with some coefficients potentially being exactly zero.

LASSO is a technique used to address multicollinearity, a problem where independent variables are highly correlated. Introduced by Tibshirani in 1996 [17], LASSO works by reducing the regression coefficients of independent variables that are strongly correlated with the error term to zero or near zero [18]. LASSO estimation is obtained from the following equation:

$$\boldsymbol{\beta^{lasso}} = \boldsymbol{argmin_{\beta}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2 + \lambda\sum_{k=1}^{p}|\beta_k|\right\} \quad (11)$$

provided that $\sum_{k=1}^{p} |\beta_k| \leq t$. The value of t is a tuning parameter that controls the shrinkage of the LASSO coefficient with $t \geq 0$. According to Djuraidah's study [19], LASSO coefficient estimation uses quadratic programming with inequality constraints.

2.7 Parameter estimation of GWR model with LASSO method

Estimation of GWR model parameters with the LASSO method is only the addition of a weighting matrix so that it is obtained:

$$\hat{\beta}^{lasso} = argmin\{\sum_{i=1}^{n} W(u_i, v_i)\left(y_i - \beta_{0(u_i.v_i)} - \sum_{k=1}^{p}\beta_{k(u_i.v_i)}x_{ik} + \lambda\sum_{k=1}^{p}|\beta_{k(u_i.v_i)}|\right)^2\} \quad (12)$$

Since the equation has an absolute value constraint on the

regression coefficient, it causes a nonlinear pattern so that the solution process uses a quadratic program assisted by the LARS algorithm.

2.8 Least Angle Regression

LASSO has been solved using the LARS algorithm [20]. The LARS algorithm is generally as follows:

1) Assume all regression coefficients β_k are zero so that $\varepsilon = 0$.
2) Select the independent variable that has the highest correlation coefficient with ε.
3) Estimating the coefficient β_k for x_{ik} that has the highest correlation with ε.
4) Calculate the residuals $\varepsilon = y - \hat{y}$ with the independent variable x_k included in the model.
5) Calculate the partial correlation between the remaining independent variables and the new ε.
6) Repeating steps 3 to 6 until all independent variables are included in the model and stopping if the correlation $(y, x_{ik}) = 0$.

3. RESEARCH MODEL

3.1 Literature study

This research uses literature studies, namely reviewing and studying various sources both books, theses and various journals related to the discussion in this study.

3.2 Data

The study utilized secondary data on poverty rates for each regency/city in North Sumatra in 2022, sourced from the Central Bureau of Statistics (BPS). North Sumatra consists of 33 regions, comprising 25 regencies and 8 cities. To identify the presence of multicollinearity in the data, the Variance Inflation Factor (VIF) values were examined.

3.3 Research variables

The variables used in this study are presented in Table 1:

Table 1. Research variables

Variable	Definition
Dependent Variable: Poverty Rate of Districts/Cities in North Sumatra (Y)	The percentage of the population of North Sumatra that is below the poverty line.
Independent Variable:	
a. Human Development Index (X_1)	A measure of human development achievement based on certain components.
b. Open Unemployment Rate (X_2)	Percentage of unemployed to the overall workforce.
c. Life Expectancy (X_3)	The average number of years that a newborn baby will live in a given year.
d. Average Years of Schooling (X_4)	The average number of years spent by the population aged 15 years and over in all types of education.
e. Gross Enrollment Rate in Higher Education (X_5)	The Gross Enrollment Rate (APK) in Higher Education (HE) is the ratio between the number of people studying in Higher Education (HE) (regardless of the age of the population) and the number of people who are officially eligible for school age at the HE level (19-23 years old).
f. Per Capita Income (X_6)	Per Capita Income is a measure of the amount of money earned per person in a country or geographic area (district/city).
Symptoms of Multicollinearity	To determine whether there is a multicollinearity value in the data, the VIF value is used. If the VIF value is>10, then the data contains multicollinearity. The following independent variables have VIF values greater than 10: $X_1 = 163{,}8894;\ X_3 = 10{,}6086;\ X_4 = 42{,}2006;\ X_6 = 32{,}9169$

3.4 Data processing methodology

The steps of this research are as follows:

1) Preparing data.
2) Spatial Heterogeneity Test.
3) Perform Geographically Weighted Regression modeling.
   a. Calculating the Euclidean distance of each district/city.
   b. Identifying the optimal bandwidth value using the Cross-Validation method.
   c. Calculating the weight matrix of each district/city with a fixed Gaussian kernel.
   d. Calculating the parameter estimation value for each district/city based on the bandwidth value and kernel weights that have been determined.
4) Detect local multicollinearity by looking at the VIF (Variance Inflation Factor) value.
5) Perform GWR modeling with LASSO to overcome local multicollinearity using the LARS algorithm.
   a. Calculate the weight matrix (W) for each district/city.
   b. Calculate the square root of the weight matrix $W^{\frac{1}{2}}(i) = sqrt(diag(W(i))$ and $W^{\frac{1}{2}}(i) = 0$ to eliminate the its location.
   c. Calculates $X_W = W^{\frac{1}{2}}(i)X$ and $y_W = W^{\frac{1}{2}}(i)y$ at each i^{th} location.
   d. Call the $lars(X_W, y_W)$ algorithm. Save the LASSO results and check for the best LASSO solution, in this case the one that minimizes the error for y_i.
6) Save the kernel estimation result.
7) Choose the optimal s (shrinkage) value and select the model coefficients.
8) Determine the final LASSO model.

The Geographically Weighted Regression (GWR) method was chosen because it is able to capture spatial variations in data that can influence the relationship between independent and dependent variables. In addition, the LASSO (Least Absolute Shrinkage and Selection Operator) method was chosen because of its ability to overcome multicollinearity by selecting variables and shrinking coefficients, which is very important in spatial data analysis which often experiences multicollinearity problems. The LASSO method uses the Least Angle Regression (LARS) algorithm, which starts with the

assumption that all regression coefficients are zero and then iteratively adds independent variables that have the highest correlation with the residuals until all variables enter the model or until the correlation between the independent variables and the residuals becomes zero.

4. RESULT AND DISCUSSION

4.1 Data exploration

As this study involves spatial data, exploratory data analysis was conducted to gain initial insights. The data utilized for this study is the poverty rate data in North Sumatra for the year 2022. The results of this descriptive analysis are presented in Table 2.

Table 2. Data descriptive statistical analysis

Var	Average	Standard Deviation	Variance	Min	Max
X_1	10.3193	4.4537	19.836	3.62	24.75
X_2	71.7930	4.4197	19.534	62.93	81.76
X_3	4.6512	2.7248	7.425	0.26	9.36
X_4	9.2866	1.3661	1.8664	5.88	11.5
X_5	24.2075	6.5742	43.2206	14.25	40.73
X_6	10716.09	2092.463	4378401.9	6152	15503

4.2 Multiple linear regression with ordinary least square method

Below is a multiple linear regression model utilizing the OLS method to analyze the factors affecting the poverty rate in North Sumatra, as implemented in R Studio:

$$Y = 61.2976 - 2.8220X_1 + 0.0683X_2 + 1.4975X_3 + 2.9227X_4 + 0.0591X_5 + 0.0017X_6 \quad (13)$$

4.3 Spatial heterogeneity test

In Table 3, the p-value of 0.0136 is smaller than the value of α=0.05, so $\boldsymbol{H_o}$ is rejected. This means that there is spatial heterogeneity in the model, namely the difference in characteristics between observation locations. Therefore. it is necessary to do GWR modeling.

Table 3. Results of spatial heterogeneity testing with Breunsch Pagan

BP Value	p-Value
16.026	0.0136

4.4 GWR modeling

Calculating Euclidean distance

The initial step in Geographically Weighted Regression (GWR) modeling involves establishing the geographical coordinates (longitude and latitude) for each regency/city in North Sumatra, representing the observation location of each region. Then calculate the Euclidean distance between (u_i, v_i) and (u_j, v_j) using Eq. (5).

The following is an example of calculating the Euclidean distance of several districts:

- Nias

$$d_{11} = \sqrt{(1.033 - 1.033)^2 + (97.766 - 97.766)^2} = \sqrt{0 + 0} = 0 \quad (14)$$

- Mandailing Natal

$$d_{12} = \sqrt{(0.783 - 1.033)^2 + (99.254 - 97.766)^2} = \sqrt{0.062 + 2.214} = 1.509 \quad (15)$$

The Euclidean distance of each regency/city in North Sumatra is presented in Table 4.

Table 4. Euclidean distance between regencies/cities in North Sumatra

Regencies/Cities	Euclidean Distance
Nias	0
Mandailing Natal	1.509056845
Tapanuli Selatan	0.7329067162
…	…
…	…
…	…
Padangsidimpuan	2.3557689
Gunungsitoli	1.656309923

Determining the bandwidth value

The next step is to determine the bandwidth value (h) obtained from the minimum CV value. The weighting function used is the Gaussian Kernel. Determination of bandwidth with the Cross Validation method using R-Studio software. For the minimum CV value and Bandwidth can be seen in Table 5.

Table 5. CV value and optimum bandwidth value

Weighting Function	CV Minimum	Bandwidth
Kernel Gaussian	204.7099	0.6571

Furthermore, the GWR weighting function is calculated using Eq. (9). With the value of $h = 0.6145937$ then obtained:

$$\mathrm{W}(\mathrm{u_i}, \mathrm{v_i}) = \exp\left(-\frac{1}{2}\left(\frac{\mathrm{d_{i,i+1}}}{0.6571}\right)^2\right) \quad (16)$$

Next, calculate the weight matrix based on the Fixed Gaussian Kernel weight function, namely:

$$\mathrm{W}(\mathrm{u_1}, \mathrm{v_1}) = \exp\left(-\frac{1}{2}\left(\frac{\mathrm{d_{11}}}{\mathrm{h}}\right)^2\right) = \exp\left(-\frac{1}{2}\left(\frac{0}{0.6571}\right)^2\right) = 1 \quad (17)$$

Substitute the weight values from each location into the matrix to obtain the overall *W* matrix as follows.

$$W = \begin{bmatrix} 1 & 0.0715 & 0.0597 & \cdots & 0.9026 \\ 0.0715 & 1 & 0.5368 & \cdots & 0.0329 \\ 0.0597 & 0.5368 & 1 & \cdots & 0.0425 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.9026 & 0.3297 & 0.0425 & \cdots & 1 \end{bmatrix} \quad (18)$$

Table 6. Parameter estimation results of GWR model

	Min.	Quartil 1	Median	Quartil 2	Max
Intercept	-10.0689	15.663	29.798	38.548	53.242
IPM	-3.1888	-1.9809	-0.2984	0.5068	1.9747
TPT	-0.44505	-0.0852	0.0062	0.1961	0.9069
AHH	-1.0240	-0.5470	0.2285	1.1648	2.3009
RLS	-4.0642	-1.2066	-0.0445	2.3259	5.3544
APK	-0.3879	-0.0346	-0.0151	0.0002	0.0290
PPP	-0.0019	-0.0005	-0.0002	0.0008	0.0020

GWR parameter estimation

The results of the parameter estimation are listed in Table 6.

GWR model hypothesis test

GWR model fit test. The $\boldsymbol{F_{count}}$ value of GWR with $\boldsymbol{df_1}$=26 and $\boldsymbol{df_2}$=9.8272 is 4.876 with $\boldsymbol{F_{table} = 2.3552}$, $\boldsymbol{p - value}$=0.0065. Because the $\boldsymbol{p - value < \alpha = 0.05}$ and the value of $\boldsymbol{F_{count} > F_{table}}$, the decision is made to reject $\boldsymbol{H_0}$, this indicates a substantial difference between the global multiple linear regression model and the local Geographically Weighted Regression (GWR) model.

GWR model parameter significance test. Parameter significance test is determined by testing the parameters partially. The $\boldsymbol{t_{table} = 2.0595}$ was obtained. If the $\boldsymbol{|t_{count}|}$ value of each district/city is greater than 2.0595. This indicates that the independent variable affects the dependent variable.

Table 7. Summary of t test on each independent variable

Variable	District
X_1	Gunungsitoli
X_2	Nias Utara
X_3	Nias, Tapanuli Selatan, Nias Selatan, Nias Utara,Nias Barat, Sibolga, Padang Sidempuan, Gunungsitoli
X_4	-
X_5	Nias, Nias Utara, Nias Selatan, Nias Barat, Gunungsitoli
X_6	-

Based on the results of the t-test as shown in Table 7, the independent variables demonstrate varying significance across different regions: Variable X_1 is significant in Gunungsitoli, indicating that the factors represented by X_1 have a noticeable impact on the outcomes in this area. This may be due to unique local conditions or specific policies that amplify the influence of this variable. Variable X_2 shows significance in North Nias, which could be attributed to demographic or economic characteristics in the region that align with the aspects measured by X_2. Variable X_3 is significant in several regions, including Nias, South Tapanuli, South Nias, North Nias, West Nias, Sibolga, Padang Sidempuan, and Gunungsitoli. This widespread influence suggests that X_3 may be related to broader regional factors affecting many areas, such as development policies or widespread environmental issues. Variable X_5 shows significance in Nias, North Nias, South Nias, West Nias, and Gunungsitoli, indicating that X_5 has specific relevance in these areas, possibly linked to social structures or relevant infrastructure impacting this variable. These results highlight the importance of contextual analysis when assessing the impact of independent variables, as their influence can vary significantly based on geographic location and the specific characteristics of each region.

4.5 Detection of local multicollinearity

The next step is to assess the presence of multicollinearity in the GWR model, which can be determined by examining the local VIF (Variance Inflation Factor) values. From the table, it is evident that some VIF values exceed 10. For instance, the HDI (Human Development Index) variable in Nias Regency has a VIF value of 133.31897.

This indicates that the standard error of the estimated coefficient for the HDI variable will be significantly higher $\boldsymbol{11.5463}$ ($\sqrt{\boldsymbol{133.3189}}$) times than it would be if there were no correlation with other independent variables. A summary of VIF values in the research data is presented in Table 8.

Table 8. Summary of local multicollinearity

Variable	X_1	X_2	X_3	X_4	X_5	X_6
Maximum	228.4629	2.9955	33.4230	59.6546	4.1245	34.1849
Mean	82.7906	1.0700	6.3378	17.9088	1.4008	15.3581
Minimum	133.4862	1.8164	15.2454	31.4169	2.3653	23.7751
VIF>10	33	0	24	33	0	33

4.6 GWR modeling with LASSO

Since the GWR model exhibits local multicollinearity, it can be addressed using the LASSO method. The following steps will be presented. GWR Modeling with LASSO in Nias Regency from the LARS algorithm obtained Table 9.

The selection of the best LASSO model is done by the 5-fold Cross Validation method with fraction mode, which is by calculating the cross-validation value for each step with one variable entered into the model.

Based on Figure 1, it is shown that there are 100 steps generated by mode fraction. Therefore, the resulting s value is also 100 different values. Where for the best minimum s value is when the cross-validation value is minimum. In this model, the smallest cross validation value is at step 31, namely with a cross validation value of 0.0322 with an s value of 0.303. Next, the s value obtained from the fraction mode will be compared with the s value of the LARS algorithm.

Table 9. Stages in the Nias Regency LARS algorithm

Step	X_1	X_2	X_3	X_4	X_5	X_6
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	-0.995
3	0.000	0.000	0.000	0.000	-0.629	-0.964
4	0.000	0.000	0.058	0.000	-0.665	-0.954
5	0.000	0.000	0.082	0.066	-0.667	-1.026
6	0.000	0.096	0.110	0.112	-0.701	-1.109
7	-2.414	0.309	0.937	1.368	-0.493	-0.184

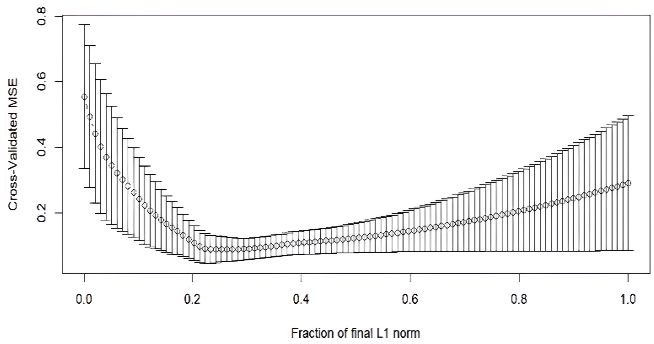

Figure 1. CV value using mode fraction Nias Regency

The coefficient value at each stage will be compared with the sum of the OLS method coefficients and the *s* value obtained in the LARS algorithm will be adjusted to the s value in the fraction method and the s value is the shrinkage value.

Because the value of s in the fraction mode is 0.303 and in the LARS method the value of s that is close to 0.303 is 0.288, namely at the $\boldsymbol{6^{th}}$ stage in Table 10.

The value of s is 0.288 with an R-squared value of 0.9403. This indicates that in Nias Regency, 94.03% of the variation in the poverty rate is explained by the independent variables in

the model, while the remaining 5.97% is influenced by factors not included in the model.

GWR Modeling with LASSO in Medan City. The outcomes of the LARS algorithm are shown in Table 11.

The selection of the best LASSO model is done with the n-fold Cross Validation process with mode fraction, namely by calculating the cross validation value for each step with one variable entered into the model.

Based on Figure 2, it is shown that there are 100 steps generated by mode fraction. Therefore, the resulting s value is also 100 different values. Where for the best minimum s value is when the cross validation value is minimum. In this model, the smallest cross validation value is at step 14, namely with a cross validation value of 0.0215 with an s value of 0.131. Next, the s value obtained from the fraction mode will be compared with the s value of the LARS algorithm.

Table 10. Comparison of LASSO coefficient values with OLS coefficients Nias Regency

Step	$\sum_{k=1}^{p} \lvert\widehat{\beta}_k^{LASSO}\rvert$	$\sum_{k=1}^{p} \lvert\widehat{\beta}_k^{OLS}\rvert$	$S = \frac{\sum_{k=1}^{p} \lvert\widehat{\beta}_J^{LASSO}\rvert}{\sum_{k=1}^{p} \lvert\widehat{\beta}_J^{OLS}\rvert}$
1	0	7.3731	0
2	0.9955	7.3731	0.1350
3	1.5933	7.3731	0.2160
4	1.6778	7.3731	0.2275
5	1.8428	7.3731	0.2499
6	2.1299	7.3731	0.2888
7	5.708	7.3731	0.7741

Note: The final model is obtained:
$Y = 0.0962X_2 + 0.1101X_3 + 0.1125X_4 - 0.7019X_5 - 1.1092X_6$

Table 11. Stages in LARS algorithm for Medan City

Step	X_1	X_2	X_3	X_4	X_5	X_6
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	-0.1607	0.0000	0.0000	0.0000
3	0.0000	-0.0945	-0.2873	0.0000	0.0000	0.0000
4	0.0000	-0.1065	-0.2816	-0.0301	0.0000	0.0000
5	0.0000	-0.1551	-0.2970	-0.1091	0.1095	0.0000
6	0.0000	-0.2255	-0.3152	-0.1936	0.1598	0.1206
7	0.3214	-0.2356	-0.4015	-0.3476	0.1419	0.0000
8	0.3242	-0.2362	-0.4024	-0.3495	0.1419	0.0000
9	0.37777	-0.2373	-0.4166	-0.3745	0.1387	-0.0211

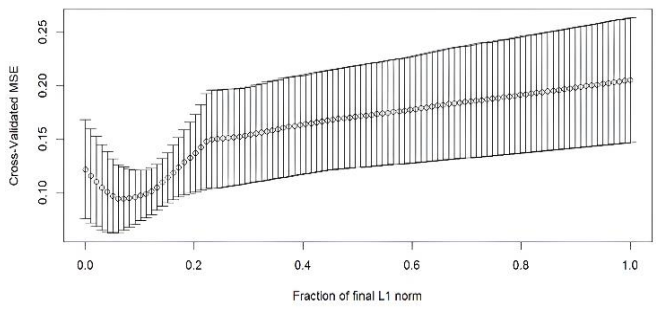

Figure 2. CV value using mode fraction Medan City

The coefficient value at each stage will be compared with the sum of the OLS method coefficients and the s value obtained in the LARS algorithm will be adjusted to the s value in the fraction method and the s value is the shrinkage value.

Because the value of s in the fraction mode is 0.131 and in the LARS method the value of s that is close to 0.131 is 0.1376, namely at the 6^{th} stage presented in Table 12.

Table 12. Comparison of LASSO coefficient value with OLS coefficient of Medan City

Step	$\sum_{k=1}^{p} \lvert\widehat{\beta}_k^{LASSO}\rvert$	$\sum_{k=1}^{p} \lvert\widehat{\beta}_k^{OLS}\rvert$	$S = \frac{\sum_{k=1}^{p} \lvert\widehat{\beta}_J^{LASSO}\rvert}{\sum_{k=1}^{p} \lvert\widehat{\beta}_J^{OLS}\rvert}$
1	0	7.3731	0
2	0.1607	7.3731	0.0217
3	0.3818	7.3731	0.0517
4	0.4182	7.3731	0.0567
5	0.6707	7.3731	0.0909
6	1.0147	7.3731	0.1376
7	1.448	7.3731	0.1963
8	1.4537	7.3731	0.1971
9	1.5659	7.3731	0.2123

Note: The final model is obtained:
$Y = -0.2255X_2 - 0.3152X_3 - 0.1936X_4 + 0.1598X_5 + 0.1206X_6$

The value of s is 0.137 with an R-squared value of 0.5143. This indicates that in Medan, 51.43% of the variation in the poverty rate is explained by the independent variables in the model, while the remaining 48.57% is affected by other factors outside the model.

Spatial patterns in parameter coefficient measures regarding geographic heterogeneity in the determinants of poverty focus on understanding how the variables that influence poverty differ across locations or regions. LASSO helps reduce model complexity and automatically selects the most significant variables. This is particularly relevant in the context of Geographically Weighted Regression (GWR), where local multicollinearity can influence the results. LASSO's advantage in identifying and measuring differences in poverty factors across regions is that it takes into account spatial influences in parameter estimation, thereby providing a better understanding of how poverty determinant variables behave at the local level.

Additionally, LASSO automatically simplifies the model by shrinking the coefficients of less significant variables to zero, increasing interpretability. From a statistical perspective, in Nias Regency, the LASSO model produces an R-squared of 0.9403, which means that 94.03% of the variation in poverty levels can be explained by independent variables. In contrast, in Medan City, the R-squared was recorded at 0.5143, indicating 51.43% of the variation was explained by the model. AIC is also used to evaluate model quality, where lower values indicate a good balance between goodness of fit and complexity. With its computational efficiency, LASSO not only fixes technical problems such as multicollinearity but also provides more precise insights into the determinants of poverty.

4.7 Limitations and future research

In this analysis, there are several data limitations and model assumptions that can affect the accuracy of the results. First, data limitations include temporal coverage limited to one year (2022), which may not reflect temporal variations in poverty trends. Therefore, the use of data from multiple years can provide more in-depth longitudinal analysis. In addition, the variables currently selected only include life expectancy, open unemployment, life expectancy, average years of schooling, and per capita income. Exploring additional socio-economic and environmental variables can provide a more comprehensive understanding. In terms of model assumptions, the spatial independence assumption may not fully hold across regions, so integrating spatial autocorrelation measures can improve model accuracy.

Although LASSO helps overcome multicollinearity, interactions between variables may not have been fully detected, so adding interaction terms or non-linear models can provide deeper insights. From a methodological perspective, experiments with different kernel functions other than Gaussian can still help to better capture spatial dependencies. Validating the model with out-of-sample data is also important to test the model's predictive power on unseen data. Further research should consider geographic differences in factors influencing poverty to guide more targeted policy interventions, as well as use causal inference techniques to understand the relationships between variables and poverty outcomes.

5. CONCLUSION

Based on the results of the data analysis, the following conclusions can be made:

1) The Least Absolute Shrinkage and Selection Operator method can overcome multicollinearity in spatial data, namely with the Geographically Weighted Regression model at 33 observation locations.
2) The LASSO model in overcoming multicollinearity in Nias Regency is as follows:

$$Y = 0.0962X_2 + 0.1101X_3 + 0.1125X_4 - 0.7019X_5 - 1.1092X_6$$

From this model, it is obtained that the factors that influence the poverty rate of Nias Regency using the LASSO method are the open unemployment rate (X_2), life expectancy rate (X_3), average years of schooling (X_4), gross enrollment rate (X_5) and per capita income (X_6). The value of s is 0.288 with an R-squared value of 0.9403. This indicates that in Nias Regency, 94.03% of the variation in the poverty rate can be attributed to the independent variables in the model, while the remaining 5.97% is affected by factors outside the model.

3) The LASSO model in overcoming multicollinearity in Medan City is as follows:

$$Y = -0.2255X_2 - 0.3152X_3 - 0.1936X_4 + 0.1598X_5 + 0.1206X_6$$

From the model, it is obtained that the factors that affect the Poverty Level of Medan City using the LASSO method are the Open Unemployment Rate (X_2), Life Expectancy Rate (X_3), Average Years of Schooling (X_4), Gross Enrollment Rate (X_5) and Per Capita Income (X_6). The value of s is 0.137 with an R-squared value of 0.5143. This indicates that in Medan, 51.43% of the variation in the poverty rate can be attributed to the independent variables in the model, while the remaining 48.57% is affected by factors outside the model.

As a suggestion to readers who want to use the LASSO method on spatial data, particularly with a GWR model that contains multicollinearity, consider using kernel weights other than fixed Gaussian and incorporating additional independent variables into the dataset to achieve better model accuracy.

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to the Rector of the University of Sumatera Utara for providing financial support and resources in the preparation of this article through the TALENTA Project 2024 (Grant No.: 104/UN5.4.10.S/PPM/KP-TALENTA/RB1/2024). Without this support and collaboration, this article would not have been completed. Thank you for the opportunity to develop this research and for the meaningful contribution to the academic field.

REFERENCES

[1] Caraka, E.R., Yasin, H. (2017). Geographically Weighted Regression. MOBIUS, Graha Ilmu Yogyakarta.
[2] Oktavianus Frans, L., Rizki, S.W., Kusnandar, D. Pemodelan pertumbuhan ekonomi kalimantan barat menggunakan pendekatan Least Absolute Shrinkage and Selection Operator (LASSO). Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya, 11(1): 111-120. http://dx.doi.org/10.26418/bbimst.v11i1.51613
[3] Kartika, S., Sufri, Kholijah, G. (2020). THE Using the Geograhically Weighted Regression (GWR) method to estimate the dominant factors affecting the poor in Jambi Province. Journal of Mathematics: Theory and Applications, 2(2): 37-45. https://doi.org/10.31605/jomta.v2i2.998
[4] Nadhori, A.K. (2015). Estimasi parameter model Geographically Weighted Regression (GWR) yang mengandung multikolinearitas dengan metode regresi ridge. Doctoral Dissertation, Universitas Islam Negeri Maulana Malik Ibrahim, 13(3). http://etheses.uin-malang.ac.id/id/eprint/3781.
[5] Fotheringham, A.S., Brunsdon, C., Charlton, M.E. (2009). Geographically Weighted Regression. In the Sage Handbook of Spatial Analysis, Chichester, John Wiley and Sons, pp. 243-254.
[6] Tibshirani, R.J., Efron, B. (1993). An introduction to the bootstrap. In Monographs on Statistics and Applied Probability, London, Chapman & Hall, 57(1): 1-436.
[7] Anselin, L. (1988) Spatial Econometrics: Methods and Models. Kluwer Academic, Dordrecht. https://doi.org/10.1007/978-94-015-7799-1
[8] Hakim, A.R., Yasin, H., Suparti, S. (2014). Pemodelan persentase penduduk miskin di Kabupaten dan Kota di Jawa tengah dengan pendekatan mixed Geographically Weighted Regression. Jurnal Gaussian, 3(4): 575-584. https://doi.org/10.14710/j.gauss.3.4.575-584
[9] Nguyen, N.T., Pham, C.C., Nguyen, T.N.H. (2023). Applying geographically weighted regression to quantify the impact of impervious surface density on land surface temperature in Ho Chi Minh City, Vietnam. IOP Conference Series: Earth and Environmental Science, 1170: 012018. https://doi.org/10.1088/1755-1315/1170/1/012018
[10] Mubarak, R. (2021). Pengantar Ekonometrika. Pamekasan: Duta Media Publishing. http://repository.iainmadura.ac.id/766/.
[11] Montgomery, D.C, Runger, G.C. (2011). Applied Statistics and Probability for Engineers. New York, John Wiley & Sons.
[12] Suritman, Raupong, Kalondeng, A. (2023). Pemodelan mixed Geographically Weighted Regression yang mengandung multikolinearitas dengan regresi ridge. Journal of Statistics and Its Application, 1: 100-113. https://doi.org/10.20956/ejsa.vi.25426
[13] Darnius, O., Manurung, A. (2019). Model selection in regression linear: A simulation based on Akaike's information criterion. In Journal of Physics: Conference

Series, 1321(2): 022085. https://doi.org/10.1088/1742-6596/1321/2/022085
[14] Darnius, O., Tambunan, Y. (2023). Spatial regression model of poverty in the province of North Sumatera on 2017. In Journal of Physics: Conference Series, 2421(1): 012003. https://doi.org/10.1088/1742-6596/2421/1/012003
[15] Bager, A.S.M., Hussein, A. (2021). Comparing between the imported and local bottled drinking water by LASSO regression. IOP Conference Series: Materials Science and Engineering, 1090: 012052. https://doi.org/10.1088/1757-899X/1090/1/012052
[16] Chasco, C., Garcia, I., dan Vicens, J., (2007). Modelling spatial variations in household disposable income with geographically weighted regression. Munich Personal RePEc Arkhive (MPRA), 1682(12). https://ideas.repec.org/p/pra/mprapa/1682.html.
[17] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1): 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[18] Muhlis, Fatmawati, Rahim, I., Syamsia. (2021). Evaluation of the accuracy of spatial data in detecting the rate of land change in Sinjai District. Journal of Physics: Conference Serie, 1899: 012096. https://doi.org/10.1088/1742-6596/1899/1/012096
[19] Djuraidah, A. (2020). Monograph Penerapan dan Pengembangna Regresi Spasial Dengan Studi Kasus Pada Kesehatan, Sosial, dan Ekonomi. PT Penerbit IPB Press.
[20] Gujarati, D.N., Porter, D.C. (2009). Basic Econometrics. McGraw-Hill, New York.