International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

To cite this article: Netti Herawati., Nufus Aulia*, Dorrah Aziz. and Khoirin Nisa (2024). COMPARATIVE STUDY IN ADDRESSING MULTICOLLINEARITY USING LOCALLY COMPENSATED RIDGE-GEOGRAPHICALLY WEIGHTED REGRESSION (LCR-GWR) AND GEOGRAPHICALLY WEIGHTED LASSO (GWL), International Journal of Applied Science and Engineering Review (IJASER) 5 (2): 36-50 Article No. 183 Sub Id 284

# COMPARATIVE STUDY IN ADDRESSING MULTICOLLINEARITY USING LOCALLY COMPENSATED RIDGE-GEOGRAPHICALLY WEIGHTED REGRESSION (LCR-GWR) AND GEOGRAPHICALLY WEIGHTED LASSO (GWL)

**Netti Herawati., Nufus Aulia*, Dorrah Aziz. and Khoirin Nisa**

[1]University of Lampung, Department of Mathematics,
Bandar Lampung, Indonesia

DOI: https://doi.org/10.52267/IJASER.2024.5203

## ABSTRACT

In spatial data, multicollinearity and spatial heterogeneity are often encountered simultaneously. To overcome the problem of heterogeneity in spatial data, GWR method can be used but this method can only overcome heterogeneity but not multicollinearity. Therefore, another method is needed to overcome multicollinearity in spatial data. The purpose of this study is to look at the ability of LCR-GWR and GWL methods to overcome multicollinearity problems simultaneously. The best method is determined by the results of the study which has smaller AIC and RMSE values. The results showed that the GWL method has lower AIC and RMSE values compared to the LCR-GWR model. Therefore, it can be said that GWL is better able to overcome multicollinearity and spatial heterogeneity in Income data compared to LCR-GWR.

**KEYWORDS:** Multicollinearity, Spatial, LCR-GWR Regression, GWL Regression.

## 1. INTRODUCTION

Spatial data is defined as geographically oriented data and has a specific coordinate system and is characterized by spatial dependence and heterogeneity (spatial structure) which is then referred to as spatial effects [1]. To analyze spatial data on point approach, the method can be used is the Geographically Weighted Regression [2]. *Geographically Weighted Regression* (GWR) is a regression analysis method with parameter estimation using the *Weighted Least Square* (WLS) procedure to handle spatial heterogeneity problems by forming a local model that can model spatially varying relationships between response variables and predictor variables for each location by giving greater weight to adjacent data points compared to data points that are far apart [3].

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

In spatial data that contains multicollinearity, the GWR method can no longer be used so a method that can handle multicollinearity in spatial data is needed. Some methods to handle multicollinearity problems in regression analysis include the Ridge Regression and Lasso Regression methods. Methods to overcome local multicollinearity and spatial heterogeneity in GWR models are the *Locally Compensated Ridge-Geographically Weighted Regression* (LCR-GWR) method first introduced by [4] and the *Geographically Weighted Lasso* (GWL) method introduced by [5]. The GWL method develops the concept of *Least Absolute Shrinkage and Selection Operator* (LASSO) with the solution used is the Least Angle Regression (LARS) algorithm so that the insignificant parameter coefficient will shrink to zero [6]. Meanwhile, the LCR-GWR method develops the Ridge concept in the parameter estimation process by using different bias coefficients at each observation location [7]. With this method, each region will have a different regression model according to its own characteristics.

Previous research on the GWL method such as research by [8] who conducted GWL modeling on poverty data in Indonesia concluded that the GWL method can take a better solution than the GWR method if multicollinearity in spatial data can be overcome. Furthermore, research on the LCR-GWR method in the case of stunting in East Nusa Tenggara concluded that the LCR-GWR method was able to produce a better model in handling local multicollinearity compared to the GWR method [7]. Based on the description above, the author is interested in conducting research using the LCR-GWR and GWL methods to handle multicollinearity in Income data in the United States in 2022.

## 2. GEOGRAPHICAL WEIGHTED METHOD FOR SPATIAL DATA

Spatial data is characterized by spatial dependence and heterogeneity (spatial structure) which is then referred to as spatial effects. Based on Tobler's law I (1979), which says that everything that is interconnected will have a greater influence if it is close to each other. Spatial dependence, also known as spatial autocorrelation, is a condition in which observations in one location affect the location of other observations that are located nearby [1]. To test spatial dependence, the Moran Index is used [9]. The statistical test for Moran's index is formulated as follows:

$$Z_{count} = \frac{I - E(I)}{\sqrt{Var(I)}}$$

Where $I$ = Moran's I value, $E(I)$ = Means of I, $Var(I)$ = Variance of I. The value of I is in the interval -1 and 1. Data contains positive spatial dependence if I > E(I) [10]

The next characteristic of spatial data is the presence of spatial heterogeneity. Spatial heterogeneity occurs because it is caused by the diversity of characteristics of a location and its geographical location [11]. Spatial heterogeneity can be known by the diversity test, namely the Breusch-Pagan Test with the following formula:

$$BP = \left(\frac{1}{2}\right)\boldsymbol{f}^{\boldsymbol{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})^{-\mathbf{1}}\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{f} \sim \chi_p^2$$

with $\boldsymbol{f} = (f_1, f_2, \dots, f_n\,)^T$ :

$$\boldsymbol{f}_{\boldsymbol{i}} = \frac{e_i^2}{\sigma^2} - 1$$

Where $e_i^2$ = Error for the $i^{\text{th}}$ observation with a matrix of size (n×1), $\boldsymbol{f}$ =Vector of size $(n \times 1), n$ =Number of observation areas, $\sigma^2$ = Variance of $e_i^2$, $\boldsymbol{Z}$ = Matrix of size n×(p+1) containing the vector of **X**.

### *2.1* Geographically Weighted Regression (GWR)

*Geographically Weighted Regression* (GWR) is a multiple regression to handle the problem of spatial heterogeneity that models spatially varying relationships between response variables and predictor variables for each location by giving greater weight to data points that are close together compared to data points that are far apart. [3]. The GWR model is written in the following form:

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

With $y_i$ = Observation value of response variable for $i^{\text{th}}$ location, $x_{ik}$= Observation value of predictor variable $k$ at $i^{\text{th}}$ location, $\beta_0(u_i, v_i)$ = Intercept at the $i^{\text{th}}$ observation location, $\beta_k(u_i, v_i)$= Coefficient of $k^{\text{th}}$ local regression at $i^{\text{th}}$ observation location, $\varepsilon_{ik}$ = Error at $i^{\text{th}}$ observation location, $(u_i, v_i)$ = Geographic coordinates of the $i^{\text{th}}$ observation location

In the GWR model, where parameters are estimated by the *Weighted Least Square* (WLS) procedure, making the weighting system dependent on location in geographic space. Weights are assigned according to their proximity to location *i*. Data from observation locations close to *i* are given greater weight than data from observation locations further away. Suppose the weight for each $i^{\text{th}}$ location is $w_{ij}$ where, then

International Journal of Applied Science and Engineering Review
IJASER
ISSN: 2582-6271
Vol. 5, Issue.2, March-April 2024, page no. 36-50

$j = 1, 2, 3, \dots, n$ the parameters for the $i^{th}$ location are estimated by minimizing the sum of the following error squares:

$$\sum_{j=1}^{n} w_j\,(u_i, v_i\,)\varepsilon_j^2 = \sum_{j=1}^{n} w_j\,(u_i, v_i\,)\left[y_j - \beta_0(u_i, v_i) - \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{jk}\right]^2$$

So, the parameter estimation of the GWR model for each observation location is as follows:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) \;= (\boldsymbol{X^T}\boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{W}(u_i, v_i)\boldsymbol{Y}$$

The $\boldsymbol{W}(u_i, v_i)$ matrix is written in the following form:

$$\begin{bmatrix} w_1(u_i, v_i) & \cdots & 0 & 0 \\ 0 & w_2(u_i, v_i) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n(u_i, v_i) \end{bmatrix}$$

To form a weighting matrix, a weighting function is needed which is influenced by the size of the neighborhood called the *bandwidth* and then adjusted to the proximity of the $i^{th}$ location point [6]. *Bandwidth* is the radius of a circle where points within the radius of the circle are still considered influential in shaping the model parameters of observation location *i* [12]. The weighting function is calculated from a kernel function that makes the observation location closer to the $i^{th}$ location point have a greater weight. The kernel function used is the exponential kernel function. The function has the same *bandwidth* for each observation with the following:

$$w_j(u_i, v_i) = exp\left[\frac{-d_{ij}}{h}\right]$$

With $d_{ij}$ is the distance between the $i^{th}$ observation location point and the $j^{th}$ location which is calculated by Euclidean distance based on the coordinates of the spatial data to produce weights between observations. The Euclidean distance in the GWR model is written in the following form:

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

Meanwhile, $h$ is the optimum *bandwidth* obtained using the *Cross Validation* (CV) method. CV is an iterative process that aims to find the kernel *bandwidth* that minimizes the prediction error of all observed outcome variables [3]. CV is defined as follows:

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(h)]^2$$

Where $\hat{y}_{\neq i}(h)$ is the estimated value for $y_i$ by omitting the $i^{\text{th}}$ location point observation in the prediction process and the optimum *bandwidth (h)*. The *bandwidth* that minimizes the CV value is the most suitable *bandwidth* to maximize the predictive power of the model.

In the GWR model, there are two hypothesis tests, namely the model fit test and the model parameter significance test. The GWR model fit test serves to explain whether there is a significant difference between the GWR model and the global linear regression model. [12]. The GWR model fit test is calculated with the following formula:

$$F_{count} = \frac{\dfrac{(SSE(H_0) - SSE(H_1)}{v}}{\dfrac{SSE(H_1)}{\delta_1}}$$

With the decision criteria reject $H_0$ if $F_{count} > F_{\alpha, df_1, df_2}$ or if *p-value* < 0.05. The degrees of freedom used are $df_1 = \frac{v^2}{v^*}$ and $df_2 = \frac{\delta_1^2}{\delta_2}$ with $v^* = tr[(R_0 - R_1)^2]$ and $\delta_2 = tr[(R_1)^2]$.[13]
The significance test of the GWR model aims to see which parameters have a significant effect on the response variable [14]. The parameter significance test or partial test of the GWR model parameters is written in the following formula:

$$t_{count} = \frac{\hat{\beta}_k(u_i, v_i)}{SE\left[\hat{\beta}_k(u_i, v_i)\right]}$$

With SE is the Standard Error of $\hat{\beta}_k(u_i, v_i)$ and the decision criterion is to reject $H_0$ if $|t_{count}| > t_{\alpha/2(n-p-1)}$ or if *p-value* < 0,05.

### *2.2* Locally Compensated Ridge-Geographically Weighted Regression (LCR-GWR)

In the case of spatial data with multicollinearity problems, there is a development of the GWR method using the Ridge Regression and *Least Absolute Shrinkage and Selection Operator* (LASSO) concept. In Ridge regression, a constraint (bias term) is added to the least squares, so that the coefficient decreases and approaches zero [15]. The coefficient estimates in Ridge Regression are obtained by minimizing the following equation:

$$\hat{\beta}^{R} = \underset{\beta}{\arg min} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{k=1}^{p} X_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} \beta_k^2 \right\}$$

With the constraint $\sum_{k=1}^{p} \beta_k^2 \geq \lambda$, where $\lambda$ is the amount that controls the amount of shrinkage and the value of $\lambda \geq 0$. When $\lambda=0$, the Ridge Regression will produce the same estimate as OLS. When $\lambda \to \infty$, the coefficient estimates approach zero [6]. One of the applications of the Ridge Regression concept in the GWR method is the *Locally Compensated Ridge Geographically Weighted Regression* (LCR-GWR) method. The LCR-GWR method is a development of Ridge regression to overcome multicollinearity in spatial data analysis. This method uses one bias coefficient for a particular observation location, which means it will produce Ridge bias coefficients locally. Ridge parameters are allowed to vary in each region to adjust to the effects of multicollinearity between predictor variables in each observation area so that the parameter coefficients in the model are expected to be more accurate [7]. The estimation of the LCR-GWR model is written in the following form:

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{LCR}}((u_i, v_i) = \left( \boldsymbol{X}^{*\boldsymbol{T}} \boldsymbol{W}(u_i, v_i) \boldsymbol{X}^* + \boldsymbol{\lambda I}(u_i, v_i) \right)^{-1} \boldsymbol{X}^{*\boldsymbol{T}} \boldsymbol{W}(u_i, v_i) \boldsymbol{Y}^*$$

$\lambda \boldsymbol{I}(u_i, v_i)$ which is the *Locally Compensated* (LC) value of $\lambda$ in the observation $(u_i, v_i)$ region Ridge regression parameter values are obtained by correlating the eigenvalues and conditional number (c) obtained from multiplying the $\boldsymbol{X^T W}(u_i, v_i)\boldsymbol{X}$ matrix. The conditional number (c) is obtained from dividing the largest eigenvalue by the smallest eigenvalue. According to [16], multicollinearity can be detected if the value of the conditional number (c) is greater than 30. When the value of the conditional number (c) obtained is greater than 30, it forces the *Locally Compensated* (LC) condition value not to exceed the same threshold, so follow the convention and determine the threshold of 30 [4].

In the LCR-GWR method, there is hypothesis testing which aims to see which parameters have a significant influence on the response variable. The parameter significance test in LCR-GWR is written in the following formula:

$$t_{count} = \frac{\hat{\beta}_k(u_i, v_i, \lambda_i)}{SE[\hat{\beta}_k(u_i, v_i, \lambda_i)]}$$

With SE is the Standard Error of $\hat{\beta}_k(u_i, v_i, \lambda_i)$ and the decision criterion is to reject $H_0$ if $|t_{count}| > t_{\alpha/2(n-p-1)}$ or if the *p-value* < 0.05.

### *2.3* Geographically Weighted Lasso (GWL)

The second concept in dealing with multicollinearity is the LASSO method. LASSO is a method to shrink highly correlated and insignificant coefficients to zero first proposed by [17]. Like Ridge regression which causes the coefficients to shrink towards zero. However, in LASSO some coefficients shrink to zero [18]. The estimation in the LASSO method is formulated as follows:

$$\hat{\beta}^L = \underset{\beta}{\arg min} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{k=1}^{p} X_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right\}$$

The application of the LASSO method to overcome local multicollinearity and spatial heterogeneity in GWR models is the *Geographically Weighted Lasso* (GWL) method. The estimation of the GWL model is formulated as follows:

$$\hat{\beta}^L = \underset{\beta}{\arg min} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0(u_i, v_i) - \sum_{k=1}^{p} X_{ik}\beta_k (u_i, v_i) \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k (u_i, v_i)| \right\}$$

With the constraint $\sum_{k=1}^{p} |\beta_k (u_i, v_i)| \le t$ which must be weighted at each location so that the shrinkage parameter will be different for each location [6]. The steps in estimating the GWL model are as follows:

1. Calculating the optimum *bandwidth* value with the CV method
2. By using the optimum *bandwidth* obtained, then calculate the weight matrix $\boldsymbol{W}$ which is n×n with an exponential function.
3. At each location $i = 1, 2, 3, \dots, n$

a) Calculate $\boldsymbol{W}^{\frac{1}{2}}(i) = sqrt\left(diag(\boldsymbol{W}(i))\right)$

b) Calculate $\boldsymbol{X}_W = \boldsymbol{W}^{\frac{1}{2}}(i)\boldsymbol{X}$ and $\boldsymbol{y}_W = \boldsymbol{W}^{\frac{1}{2}}(i)\boldsymbol{y}$ using the root of the weight $\boldsymbol{W}(i)$ at each $i^{th}$ location.

c) Determine the solution of LASSO corresponding to the CV based on the value of t at each $i^{th}$ location with the LARS algorithm.

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

The value of parameter $t$ at each $i^{th}$ location $(t_i)$ is calculated with the following formula:

$$t_i = \frac{\sum_{k=1}^{p} |\hat{\beta}_k(u_i, v_i)|}{\sum_{k=1}^{p} |\hat{\beta}_k^{OLS}|}$$

## 3. METHODOLOGY

Data used in this study is the United States Income data in 2022 obtained from the Bureau of Economic Analysis U.S. Department of Analysis (https://www.bea.gov/). The sample amounted to 51 observations with 9 predictor variables, namely Gross Domestic Product $(x_1)$, Personal Consumption Expenditures $(x_2)$, Population Total $(x_3)$, Labor Force Participation Not Seasonally $(x_4)$, Population Density $(x_5)$, Labor Force Participation Seasonally $(x_6)$, High School Graduate Rate $(x_7)$, Bachelor Degree Rate$(x_8)$, and Associate Degree Rate $(x_9)$. First descriptive statistical analysis was performed on the data. and multicollinearity test by calculating the *Variance Inflation Factor* (VIF). Furthermore, Moran's Index is calculated to conduct spatial dependency test with Z test while spatial heterogeneity test with Breusch-Pagan test. After it is found that the data has spatial heterogeneity, the data is analyzed using GWR, LCR-GWR, and GWL methods. To measure the best model, it is seen from the smallest *Root Mean Squared Error* (RMSE) and *Akaike Information Criterion Score* (AIC) values. The AIC and RMSE values are calculated with the following formula:

$$AIC = 2k + n \ln\left(\frac{SSE}{n}\right)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

## 4. RESULT AND DISCUSSION

Descriptive statistical analysis of the United States income data is presented in Table 1 below.

**Table 1 Descriptive Statistic**

| Variable | Minimum | Maximum | Means | Standard Deviation |
|---|---|---|---|---|
| $Y$ | 46370 | 95970 | 63886 | 9701,527 |
| $X_1$ | 40,62 | 3598,04 | 496,18 | 643,488 |
| $X_2$ | 39678 | 85723 | 517288 | 7547,836 |
| $X_3$ | 0,580 | 39,030 | 6,533 | 7,42433 |
| $X_4$ | 54,80 | 71,10 | 63,06 | 4,0187 |
| $X_5$ | 1,30 | 10332,90 | 408,51 | 1444,171 |
| $X_6$ | 54,70 | 71,10 | 62,72 | 3,9484 |
| $X_7$ | 15,50 | 40,10 | 27,53 | 4,2286 |
| $X_8$ | 13,0 | 26,7 | 20,3 | 3,0068 |
| $X_9$ | 3,0 | 14,10 | 9,057 | 1,7286 |

Table 1 above show that Income in the United States has a minimum value of 46370 and a maximum of 95970 with an average of 63886. The Associate Degree Rate $(x_8)$ has the smallest standard deviation compared to other variables. This shows that the population has a bachelor's degree that is fairly evenly distributed between regions in the United States.

Furthermore, to check for multicollinearity in the United States Income data in 2022 is done using the VIF value. A VIF value greater than 10 indicates multicollinearity. The VIF value on the Income data can be seen in table 2 below

**Table 2 VIF Value**

| Variable | VIF |
|---|---|
| $X_1$ | 36,5988 |
| $X_2$ | 6,72749 |
| $X_3$ | 34,7723 |
| $X_4$ | 228,5206 |
| $X_5$ | 4,4890 |
| $X_6$ | 220,9434 |
| $X_7$ | 3,0304 |
| $X_8$ | 5,7613 |

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

| | |
|---|---|
| $X_9$ | 1,9127 |

Table 2 above shows that some independent variables contain multicollinearity, which has a VIF value> 10 is the variable $X_1$, $X_3$, $X_4$, and $X_6$. While the other independent variables have a VIF value < 10. To ensure this, the correlation calculation between the independent variables is also carried out and the results are as follows:

**Table 3 Correlation between Predictor Variables**

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 1 | 0,1847 | 0,9810 | -0,0714 | -0,0368 | -0,0531 | -0,2955 | -0,1898 | 0,1534 |
| $X_2$ | 0,1847 | 1 | 0,0999 | 0,5135 | 0,7072 | 0,5017 | -0,6070 | -0,3717 | 0,7023 |
| $X_3$ | 0,9810 | 0,0999 | 1 | -0,1337 | -0,0787 | -0,1131 | -0,2269 | -0,1762 | 0,0852 |
| $X_4$ | -0,0714 | 0,5135 | -0,1337 | 1 | 0,2880 | 0,9962 | -0,5523 | 0,1429 | 0,7540 |
| $X_5$ | -0,0368 | 0,7072 | -0,0787 | 0,2880 | 1 | 0,3051 | -0,4229 | -0,5625 | 0,3026 |
| $X_6$ | -0,0531 | 0,5017 | -0,1131 | 0,9962 | 0,3051 | 1 | -0,5669 | 0,1216 | 0,7473 |
| $X_7$ | -0,2955 | -0,6070 | -0,2269 | -0,5523 | -0,4229 | -0,5669 | 1 | 0,1422 | -0,7428 |
| $X_8$ | -0,1898 | -0,3717 | -0,1762 | 0,1429 | -0,5625 | 0,1216 | 0,1422 | 1 | -0,0273 |
| $X_9$ | 0,1534 | 0,7023 | 0,0852 | 0,7540 | 0,3026 | 0,7473 | -0,7428 | -0,0273 | 1 |

From Table 3 it can be seen that there is a positive correlation between variables $X_1$ and $X_3$ with a correlation value of 0,9810, variables $X_2$ and $X_5$ with a correlation value of 0,7072, variables $X_2$ and $X_9$ with a correlation value of 0,7023, variables $X_4$ and $X_6$ with a correlation value of 0,9962, variable $X_4$ and $X_9$ with a correlation value of 0,7540, variables $X_6$ and $X_9$ with a correlation value of 0,7473, and negative correlation between variables $X_7$ and $X_9$ with a correlation value of -0,7428.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

Next is to test spatial dependence and spatial heterogeneity in the Income data. Spatial dependence is tested with the Z test which is based on the value of the Moran Index ($H_0$: There is no spatial dependence vs $H_1$: There is spatial dependence) with a significance level of 5% so that $Z_{table}$ = 1.96 is obtained and the results can be seen in table 4.

**Table 4 Spatial Dependency Test**

| Moran's I | $E(I)$ | $Var(I)$ | $Z_{count}$ |
|---|---|---|---|
| 0,4328 | -0,0200 | 0,0286 | 2,6242 |

The $Z_{count}$ is = 2,6242 > 1,96, therefore $H_0$ is rejected. It can be concluded that there is no spatial dependence, which means that the income value in one observation area not depends or is influenced by other observation areas.

After the spatial dependency test is carried out, spatial heterogeneity was tested using the Breusch-Pagan test statistic ($H_0$: There is no spatial heterogeneity vs $H_1$: There is spatial heterogeneity). The results of the analysis are presented in the table below:

**Table 5 Spatial Heterogeneity Test**

| Breusch-Pagan | DF | *p-value* | $\chi^2_{(0,05;9)}$ |
|---|---|---|---|
| 16,954 | 9 | 0,04944 | 16,919 |

The Breusch-Pagan value in tables is 16.954 which greater than $\chi^2_{(0,05;9)}$=16,919 with a *p-value* of 0.04944, hence $H_0$ is rejected and it can be concluded that there is an effect of spatial heterogeneity in each observation location at a real level of 0.05. If the analysis using OLS is still applied to the data, the estimation results obtained will have a large variety of parameter estimates.

Breusch-Pagan test also given an information that there is spatial diversity in Income data, namely the variety that is not homogeneous between observation locations. Due to spatial diversity, a model that can overcome spatial diversity is needed, namely by forming a regression model at each observation location. The *bandwidth* used in this GWR model is 2,558. Next is to form the weight matrix obtained for all observation locations. The weight matrix for all Income data locations is a follow:

$$\begin{pmatrix} 1 & 1{,}81e-07 & \cdots & 0{,}0003 & 0{,}0663 \\ 1{,}81e-07 & 1 & \cdots & 0{,}0001 & 1{,}65e-06 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0{,}0003 & 0{,}0001 & \cdots & 1 & 0{,}0042 \\ 0{,}0663 & 1{,}65e-06 & \cdots & 0{,}0042 & 1 \end{pmatrix}$$

Fitted model test in GWR is presented in the following table:

**Table 6 Fitted Model Test**

| Model | DF | F | *p-value* |
|---|---|---|---|
| GWR | 27 | 1,8290 | 0,03249 |

The value of $F_{count}$ = 1.8290 > $F_{table}$ = 1.8292 and *p-value* = 0.03249 < 0.05 so that $H_0$is rejected. So, it can be concluded that there is a significant difference between the global regression model and the GWR model.

The presence of multicollinearity can cause the parameter estimation results to have a large variance, which can lead to errors in model interpretation. In addition, unaddressed multicollinearity will result in unstable model estimates. Therefore, to overcome the multicollinearity, the LCR-GWR and GWL methods are used. In the LCR-GWR model, the value of the ridge coefficient is determined from the conditional number (c) value obtained. In the Income data used, after calculating the multiplication of the $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}(u_{\mathrm{i}}, v_{\mathrm{i}})\boldsymbol{X}$ matrix, the c value is more than 30, which means that the data is detected to contain multicollinearity. In GWL modelling, based on the concept of LASSO, some coefficients in the GWL model shrink to zero in some observation locations, which makes the variable insignificant to the model. Modelling with GWL also occurs variable selection process. Modelling results using GWL with all predictor variables have an influence that varies between negative and positive in some observation areas.

The following is an example of the results of parameter estimation and parameter significance testing in District of Columbia:

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

**Table 7 LCR-GWR and GWL Model Parameter Significance Test**

| Parameters | LCR-GWR | | GWL | |
|---|---|---|---|---|
| | Estimation | Decision | Estimation | Decision |
| $\hat{\beta}_1$ | 0,285448 | $H_0$ rejected | 0,068131 | $H_0$ rejected |
| $\hat{\beta}_2$ | 0,902948 | $H_0$ rejected | 0,324291 | $H_0$ rejected |
| $\hat{\beta}_3$ | -0,120542 | $H_0$ rejected | 0 | $H_0$ accepted |
| $\hat{\beta}_4$ | 0,040305 | $H_0$ accepted | 0,245935 | $H_0$ rejected |
| $\hat{\beta}_5$ | -0,389905 | $H_0$ rejected | 0 | $H_0$ accepted |
| $\hat{\beta}_6$ | 0,271407 | $H_0$ rejected | 0 | $H_0$ accepted |
| $\hat{\beta}_7$ | -0,180173 | $H_0$ rejected | 0 | $H_0$ accepted |
| $\hat{\beta}_8$ | -0,109396 | $H_0$ rejected | 0,328402 | $H_0$ rejected |
| $\hat{\beta}_9$ | -0,283906 | $H_0$ rejected | -0,234351 | $H_0$ rejected |

Table 7 shows that in the analysis of Income data using LCR-GWR, the independent variables that affect the response variable in the District of Columbia are Gross Domestic Product ($X_1$), Personal Consumption Expenditures ($X_2$), Population Total ($X_3$), Population Density ($X_5$), Labor Force Participation Seasonally ($X_6$), High School Graduate Rate ($X_7$), Bachelor Degree Rate($X_8$), and Assosiate Degree Rate ($X_9$). because they have a *p-value* <0.05.

Whereas in the Income data analysis using GWL, a value of 0 indicates that the predictor variable is not significant at that observation location. Therefore, District of Columbia has factors that affect income including Gross Domestic Product ($X_1$), Personal Consumption Expenditures ($X_2$), Labor Force Participation Not Seasonally ($X_4$), Bachelor Degree Rate ($X_8$), Associate Degree Rate and ($X_9$)

After analyzing the data using LCR-GWR and GWL, the next step is to evaluate the model with AIC and RMSE values to see which model is able to provide a better solution if multicollinearity can be overcome. The RMSE and AIC values for LCR-GWR and GWL models are displayed in the Table 8.

**Table 8 AIC and RMSE Values**

| Model | LCR-GWR | GWL |
|---|---|---|
| AIC | -206.492 | -308,141 |
| RMSE | 0,1107 | 0,0401 |

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

It is clear that AIC and RMSE values for LCR-GWR and GWL are different. AIC and RMSE value of GWL is lower than LCR-GWR. It indicates that GWL model provides better solution for controlling multicollinearity in spatial data.

## 5. CONCLUSION

From the result of controlling spatial heterogeneity using GWR and controlling multicollinearity using LCR-GWR and GWL in Unites States Income data in 2022, we can conclude that GWR can overcome spatial heterogeneity while LCR-GWR and GWL can overcome multicollinearity in spatial data. When comparing LCR-GWR dan GWL using AIC and RMSE value it is found that GWL is better than LCR-GWR in overcoming the multicollinearity in spatial. It is indicated by the lower of AIC and RMSE value in GWL.

## REFERENCES

[1] Anselin, L. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Netherlands,1998.

[2] Lutfiani, N., Sugiman, & Mariani, S. “Pemodelan Geographically Weighted Regression (GWR) dengan Fungsi Pembobot Kernel Gaussian dan Bi-Square,” UNNES Journal of Mathematics,VIII (1), pp. 82-91, 2017.

[3] Klar, R. G. Geographically Weighted Regression based Investigation of Transport Policies for Increased Public Transport Ridership A Case Study of Stockholm. Stocholm, Sweden, 2021.

[4] Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. “GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models,” Journal of Statistical Software, LXII (17), pp. 1-50, 2014.

[5] Wheeler, D.C. “Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: The Geographically Weighted Lasso,” Journal of Environment and Planning A, XLI (3), pp. 722-742, 2009.

[6] Yulita, T. “Pemodelan Geographically Weighted Ridge Regression dan Geographically Weighted Lasso pada Data Spasial dengan Multikolinieritas,” Thesis, Sekolah Pascasarjana, Institut Pertanian Bogor, Bogor, 2016.

[7] Fadliana, A., Pramoedyo, H., & Fitriani, R. “Implementation of Locally Compensated Ridge-Geographically Weighted Regression Model in Spatial Data with Multicollinearity Problems (Case Study: Stunting among Children Aged under Five Years in East Nusa Tenggara Province),” Media Statistika, XIII (2), pp. 125-135, 2020.

[8] Setiyorini, A., Suprijadi, J., & Handoko, B. “Implementations of Geographically Weighted Lasso in Spatial Data with Multicollinearity (Case Study: Poverty modeling of Java Island),” AIP Conference Proceedings, pp. 1-14, 2017.

ISSN: 2582-6271

Vol. 5, Issue.2, March-April 2024, page no. 36-50

[9] Amin, C., Sari, D. N., Priyono, K. D., & Hidayah, B. "The Spatial Pattern of COVID-19 Incidence in Relation to Poverty Across Central Java Province," Proceedings of the International Conference of Geography and Disaster Management (ICGDM 2022), pp. 450-463, 2023

[10] Nadya, M. "Analisis Geographically Weighted Regression (GWR) pada Kasus Pneumonia Balita di Provinsi Jawa Barat," Skripsi, Universitas Negeri Jakarta, Jakarta, 2017.

[11] Fotheringham, A. S., Brunsdon, C., & Charlton, M. "Geographically Weighted Regression: The Analysis of Spatially Varying Relationship," John Wiley & Sons, Hoboken, 2002.

[12] Fadhilah, N. "Geographically Weighted Regression dan Spatial Pattern Analysis untuk Pemodelan Kejadian Penyakit Malaria dan Faktor yang Mempengaruhi di Provinsi Papua," Skripsi, Institut Teknologi Sepuluh Nopember, Surabaya, 2015.

[13] Leung, Y., Mei, C. L., & Zhang, W. X. "Statistical Tests for Spatial Nonstationarity Based on The Geographically Weighted Regression Model," Journal of Environment and Planning A, XXXII, pp. 9-32, 2000.

[14] Putra, R., Tyas, S. W., & Fadhlurrahman, M. G. "Geographically Weighted Regression with The Best Kernel Function on Open Unemployment Rate Data in East Java Province," Enthusiastic International Journal of Statistics and Data Science, II (1), pp. 26-36, 2022.

[15] Ali, R.G. & Nugraha, J. "Penerapan Metode Regresi Ridge dalam Mengatasi Masalah Multikolinearitas pada Kasus Indeks Pembangunan Manusia di Indonesia Tahun 2017," Prosiding Sendika. pp. 239-248, 2019.

[16] Hocking, R. R. "Methods and Applications of Linear Models," John Wiley & Sons, New York, 2003.

[17] Tibshirani, R. "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), LVIII (1), pp. 267-288, 1996.

[18] Herawati, N., Nisa, K., Setiawan, E., Nusyirwan, & Tiryono. "Regularized Multiple Regression Methods to Deal with Severe Multicollinearity," International Journal of Statistics and Applications, VIII (4), pp. 167-172, 2018.