Modelling of Income Inequality in East Java Using Geographically Weighted Panel Regression

Modelling of Income Inequality in East Java Using Geographically Weighted Panel Regression

Chusnul Chotimah[1*], Sutikno[1], Setiawan[1]

[1]Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

*Corresponding author: chus_chot@yahoo.co.id

Abstract. Regression analysis is one of the statistical methods that study the relationship between response variables and predictor variables. Parameter estimates in classical linear regression produce regression coefficients that are thought to apply globally to the entire observation unit. But in fact, the existence of factors from the spatial aspect causes conditions between one location and another to be different. This spatial aspect allows the emergence of spatial heterogeneity. Geographically Weighted Regression (GWR) is a local development regression technique from ordinary regression using spatial data. In addition, in a study data is needed in a certain period of time involving cross-section data and time series or referred to as panel data. Geographically Weighted Panel Regression (GWPR) is a combination of GWR and panel data regression. The purpose of this study is to model Geographically Weighted Panel Regression using Fixed Effect Model (FEM) within estimators with adaptive bisquare kernel weight for data on income inequality (Gini ratio) in East Java Province from 2010 to 2014. In addition, to obtain factors that influence significant income inequality in each district/city of East Java Province. The results of this study indicate that the GWPR fixed effect model differs significantly in the panel data regression model, and the models produced for each location will be different from each other. Districts/cities in East Java Province have twenty-eight groups based on significant variables. The variables that significantly influences income inequality are the percentage of the poor, percentage of GDP regional in the category of fisheries forestry agriculture, percentage of GDP regional in the processing industry category, percentage of GDP regional gross fixed capital formation, per-capita GDP regional, and dependency ratio. In the GWPR model, the R^2 value is 99.953%, with Root Mean Square (RMSE) is 0.0061035. While the FEM model within estimator produces an R^2 value of 22.844% with RMSE is 0.1035616.

Keyword: GWR, GWPR, Income Inequality, Panel Data Regression

1. Introduction

Geographically Weighted Regression (GWR) is a local development regression technique from ordinary regression using spatial data. The GWR model is one model with a point approach that is based on the position of latitude and longitude. This method was developed to overcome the parameter estimation of classical linear regression which results in a regression coefficient that is assumed to apply globally to the entire observation unit. The global equation model will provide accurate information for the local area if there is none or there is little diversity between the local regions [1]. But in reality, sometimes the conditions between one location and another are affected by spatial aspects that allow for spatial heterogeneity.

In addition, in a study, it is not enough to just observe the observation unit at a given time, but it is also necessary to observe the unit at various time periods. For this reason, data is needed which is a composite data between cross-section data and time series which is called panel data. The advantage of using panel data is that data is more informative, varied and efficient, avoids multicollinearity, increases freedom of degrees and is more efficient, can measure unobservable effects on pure cross-section data and pure time series, and by making data available in more numbers a lot of panel data can minimize the bias that can occur when aggregating individuals into broader aggregates [2]. Regression using panel data is called the panel data regression model.

The Geographically Weighted Panel Regression (GWPR) method or GWR-Panel is a method that combines GWR and panel data regression [3]. This method was first carried out by Yu, who applied the technique of locally weighted panel data from data based on spatial dimensions. The idea of GWPR is simple but powerful because it allows obtaining local panel data estimates [4]. The results of several studies that studied GWPR [3-6], showed that GWPR did produce better and clearer results than GWR cross-sectional and panel data models.

In this study, it was focused on establishing a fixed effect GWPR model using adaptive bisquare kernel weight on data on income inequality of districts/cities in East Java in 2010-2014 to identify variables that affect income inequality of districts/cities in East Java.

2. Theoretical Review

2.1. Panel Regression

Panel data is data which is a combination of cross section data and time series data. In other words, panel data is a group of individuals (cross section data) observed over time (time series data). So in the panel data there will be a number of N individuals ($i = 1,2,3, ..., N$) in the period of time T ($t = 1,2,3, ..., T$) then with panel data we will has a total observation of as many as NT. The panel data regression model in general can be stated in the following equation [7]:

$$y_{it} = \alpha_{it} + \boldsymbol{\beta}^T\mathbf{x}_{\mathrm{it}} + \varepsilon_{it} \quad , \quad i = 1,2,\dots,N \;, \quad t = 1,2,\dots,T \tag{1}$$

where:

y_{it} = Observation for the first unit cross section in the t-th period
α_{it} = Intercept is the group/ individual effect of the first unit cross section in the t-th period
$\mathbf{x}_{it}^T$ = $(x_{1it}, x_{2it}, \dots, x_{pit})$ shows an observation vector on a *1xp* predictor variable
$\boldsymbol{\beta}^T$ = $(\beta_1, \beta_2, \dots, \beta_p)$ is a *1xp* size vector (coefficient slope)
ε_{it} = regression error from the *i-th* individual for the t-th period
it is assumed that the ε_{it} does not correlate with the predictor variable and $IIDN\ (0, \sigma^2)$ distribution.

2.1.1. Approach and Estimation Method in Panel Data Regression Model

a. Common Effect Model (CEM) [8], the slope and intercept are the same

$$y_{it} = \alpha + \boldsymbol{\beta}^T\mathbf{x}_{\mathrm{it}} + \varepsilon_{it} \quad , \quad i = 1,2,\dots,N \;, \quad t = 1,2,\dots,T \tag{2}$$

b. Fixed Effect Model (FEM) [8], fixed slope and different intercept

$$y_{it} = \alpha_{it} + \boldsymbol{\beta}^T\mathbf{x}_{\mathrm{it}} + \varepsilon_{it} \quad , \quad i = 1,2,\dots,N \;, \quad t = 1,2,\dots,T \tag{3}$$

c. Random Effect Model (REM) [8], fixed slope and random intercept

$$y_{it} = (\alpha + v_i) + \boldsymbol{\beta}^T\mathbf{x}_{\mathrm{it}} + \varepsilon_{it} \quad , \quad i = 1,2,\dots,N \;, \quad t = 1,2,\dots,T \tag{4}$$

2.1.2. Selection of Estimation Method for Panel Data Regression Model

a. Chow Test

The Chow Test is used to select both models between the Common Effect Model and the Fixed Effect Model. The hypothesis is as follows:
H_0: $\alpha_1 = \alpha_2 = \dots = \alpha_N = \alpha$ (Common Effect Model)

H_1: at least one intercept (α_i) is not the same (Fixed Effect Model)

Test statistics [2]:

$$F_0 = \frac{(SSE_{CEM} - SSE_{FEM})/(N-1)}{(SSE_{FEM})/(NT-N-k)}$$

F_0 follow distribution $F_{\alpha,N-1,N(T-1)-k}$. If the value of F_0 is greater than F table then H_0 is rejected, which means the right model is FEM.

b. Hausman Test
The Hausman test is used to compare the Fixed Effect Model with Random Effect. The hypothesis is as follows:
H_0: $corr\ (\mathbf{x}_{it}, \boldsymbol{\varepsilon}_i) = 0$ (Random Effect Model)
H_1: $corr\ (\mathbf{x}_{it}, \boldsymbol{\varepsilon}_i) \neq 0$ (Fixed Effect Model)
Test statistics:

$$W = (\widehat{\boldsymbol{\beta}}_{FEM} - \widehat{\boldsymbol{\beta}}_{REM})^T \left[var(\widehat{\boldsymbol{\beta}}_{FEM}) - var(\widehat{\boldsymbol{\beta}}_{REM})\right]^{-1} (\widehat{\boldsymbol{\beta}}_{FEM} - \widehat{\boldsymbol{\beta}}_{REM})$$

with $\widehat{\boldsymbol{\beta}}_{FEM}$ is a vector of estimation of FEM and parameters $\widehat{\boldsymbol{\beta}}_{REM}$ is a vector of estimated REM parameters. The Hausman test statistic follows the Chi-Square distribution with free degrees of *k-1*, where *k* is the number of variables of the predictor. If the Hausman statistical value is greater than the critical value, H_0 is rejected, which means the right model is FEM.

c. Lagrange Multiplier Test
To find out whether the Random Effect model is better than the Common Effect model. The hypothesis used is:
H_0: $\sigma_u^2 = 0$ (CEM model)
H_1: $\sigma_u^2 \neq 0$ (REM model)
Test statistics:

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} [\sum_{t=1}^{T} \varepsilon_{it}]^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \varepsilon_{it}^2} - 1 \right]^2$$

LM test statistics follow Chi-Square distribution with free degree 1. If the LM test value is greater than $\chi^2_{(1)}$ then $H_0$is rejected, which means the REM model is more appropriate.

2.2. Geographically Weighted Panel Regression
The GWR model is a method used to explore spatial non-stationers, which are defined as the properties and significant relationships between different variables at one location to another [1]. The parameters for the regression model at each location will produce different values. The GWR model is written as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i\,, \qquad i = 1,2,\ldots,n \tag{5}$$

with y_i is the observation value of the *i*-th response variable, x_{ik} is the observation value of the predictor variable k in the k-th observation, β is the regression coefficient, and (u_i, v_i) is the location coordinates i, and ε_i is the *i*-th error. The error forms *error* $\varepsilon_i, \varepsilon_2, \ldots, \varepsilon_n$are assumed to be independent, identical and follow a normal distribution with zero mean and constant variant ($\varepsilon_i \sim iid\ N(0, \sigma^2)$.
Estimating parameters in the GWR model is carried out by the Weighted Least Square (WLS) method, which is by giving different weightings to each location where the data is observed. Estimating the parameters of the GWR model requires spatial weighting to represent the location of the observation data with each other. In this study used adaptive bisquare kernel weights which can be formulated as follows:

$$w_{ij} = \begin{cases} \left(1-\left(\frac{d_{ij}}{h_i}\right)^2\right)^2 & \text{, for } d_{ij} \le h_i \\ 0 & \text{, others} \end{cases}$$

where $(d_{ij})^2 = (u_i - u_j)^2 + (v_i - v_j)^2$ in equation (6) is the Euclidean distance between the point at location i and location j and h are non-negative parameters known as bandwidth or smoothing parameters. To get optimum bandwidth, it can be done by calculating cross-validation (CV). If the CV value gets smaller, the optimum bandwidth is obtained [1] by using the following formula:

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

With $\hat{\boldsymbol{y}}_{\neq \boldsymbol{i}}(\boldsymbol{h})$ is the estimated value for $\boldsymbol{y_i}$ by removing observations at point i from the parameter testing process.

Testing of spatial influences using the spatial heterogeneity test using the Breusch-Pagan test (BP) can be formulated as follows:

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

where $\mathbf{f} = (f_1, f_2, \dots, f_n)^T$ with $\mathbf{f} = \left(\frac{e_i^2}{\sigma^2} - \mathbf{1}\right)$, $\varepsilon_i = y_i - \hat{y}_i$ is the least squares residual for the i-th observation. Z is a matrix sized n × (p + 1) that contains a vector that is normally standardized for each observation. Reject H_0 if BP> $\chi^2_{(p)}$or if $p - value < \alpha$ with p is the number of predictors, which means there is spatial heterogeneity.

Local Non-Multicollinearity assumptions as one of the conditions that must be met in a regression with several predictors is that there is no correlation between one predictor and another predictor by looking at the VIF (Inflation Factor Variance) value of each j- th predictor at the smaller i- th location of 10. VIF values can be formulated as follows:

$$VIF_j(u_i, v_j) = \frac{1}{1 - R_j^2(u_i, v_j)}$$

The main idea of the GWR-Panel is the same as the GWR cross-sectional analysis. In GWR-panel it is assumed that the time series of observations in a geographical location is a realization of a smooth spatiotemporal process. This process follows a distribution where the closest observation (one geographical location or at a time) is more related than distant observation. The GWPR method is a local regression with repetition of data at the location point for each spatial observation. In other words, GWPR is more focused on repetitive spatial observations for each location [4]. GWR-Panel Model The effect of fixed time trends can be written in the equation as follows

$$y_{it} = \sum_{k=1}^{p} \beta_k(u_i, v_i)X_{itk} + \varepsilon_{it}, \qquad i = 1,2,\dots,n \text{ dan } t = 1,2,\dots T \tag{6}$$

Equation (6) is obtained from the results of transformation namely within transformation [4]. This transformation consists of reducing the equation of the fixed influence model with the equation of the average model. The fixed effect model is expressed in the following equation:

$$y_{it} = \beta_0(u_i, v_i) + \mu_i + \sum_{k=1}^{p} \beta_k(u_i, v_i)X_{itk} + \varepsilon_{it} \tag{7}$$

Meanwhile, the average equation is expressed in the following equation:

$$\bar{y}_i = \beta_0(u_i, v_i) + \mu_i + \sum_{k=1}^{p} \beta_k(u_i, v_i)\bar{X}_{ik} + \bar{\varepsilon}_i \tag{8}$$

The result of the transformation from the reduction of equation (7) to equation (8) is stated as follows:

$$y_{it} - \bar{y}_i = \beta_0(u_i, v_i) - \beta_0(u_i, v_i) + \mu_i - \mu_i + \sum_{k=1}^{p} \beta_k(u_i, v_i)(X_{itk} - \bar{X}_{ik}) + \varepsilon_{it} - \bar{\varepsilon}_i \quad (9)$$

Equation (9) can be simplified into the following equation:

$$y_{it}^* = \sum_{k=1}^{p} \beta_k(u_i, v_i) X_{itk}^* + \varepsilon_{it}^* \quad (10)$$

with:

$y_{it}^* = y_{it} - \bar{y}_i$; $X_{itk}^* = X_{itk} - \bar{X}_{ik}$; $\varepsilon_{it}^* = \varepsilon_{it} - \bar{\varepsilon}_i$

Estimating parameters of the GWR-Panel model can use the Weighted Least approach Square (WLS) such as estimation on the GWR model which can be formulated as follows:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X^T W}(u_i, v_i)\mathbf{X}\right)^{-\mathbf{1}} \mathbf{X^T W}(u_i, v_i)\mathbf{y} \quad (11)$$

where $\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\hat{\beta}_{i0}, \hat{\beta}_{i1}, \hat{\beta}_{i2}, \ldots, \hat{\beta}_{ip}\right)^T$ is a vector of local regression coefficients and $\mathbf{W}(u_i, v_i)$ is a diagonal matrix with elements in the diagonals which are geographical weights for each data for the location of the *i*-th observation, and the other elements are zeros.

Compatibility testing of the GWR (Goodness of Fit) model is carried out by testing the suitability of the parameters simultaneously. The hypothesis of testing the suitability of the GWR model is as follows.

H_0: $\beta_k(u_i, v_i) = \beta_k$ for each k = 1,2, ..., p

(there was no significant difference between the FEM and GWR regression models)

H_1 : There is at least one $\beta_k(u_i, v_i) \neq \beta_k$, k = 1,2, ..., p

(there is a significant difference between the FEM and GWR regression models)

Test Statistics:

$$F^* = \frac{SSE(H_1)/\frac{\delta_1^2}{\delta_2}}{SSE(H_0)/(n-p-1)}$$

If F^* produces a relatively small value, it can be said that the alternative hypothesis (H_1) is more suitable to use. In other words, the GWR model has better goodness of fit than the global regression model. If a significance level (α) is given, then the decision is taken by rejecting the null hypothesis (H_0) if $F^* < F_{1-\alpha, df1, df2}$ where $df1 = \frac{\delta_1^2}{\delta_2}$ and $df2 = (n - p - 1)$ (Leung dan Zhang, 2000).

Testing the parameters of the GWR model is done when the GWR model is appropriate to describe the data. Testing the parameters of the GWR model is done to determine the independent variable x_k that has an effect on the *i*- th location. The hypotheses used in testing the parameters of the GWR model are as follows:

H_0: $\beta_k(u_i, v_i) = 0$

H_1: $\beta_k(u_i, v_i) \neq 0$ with k = 1,2, ..., p

The parameter estimator $\hat{\boldsymbol{\beta}}(i)$ will follow a normal distribution with the average $\boldsymbol{\beta}(i)$ and the covariant variant matrix $\mathbf{C}_i\mathbf{C}_i^T\sigma^2$, where $\mathrm{C}_i = \left(\mathbf{X^T W}(i)\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X^T W}(i)$, so that it is obtained:

$$\frac{\hat{\beta}_k(i) - \beta_k(i)}{\sigma\sqrt{c_{kk}}} \sim N(0,1)$$

with c_{kk} is the k th diagonal element of the matrix $\mathbf{C}_i\mathbf{C}_i^T$. Distribution $\frac{SSE(H_1)}{c\sigma^2} = \frac{\delta_1^2\hat{\sigma}^2}{\delta_2\sigma^2}$ can be approached with distribution χ^2 with free degrees $\frac{\delta_1^2}{\delta_2}$. So the test statistics used are [1]:

$$F^* = T_{count} = \frac{\hat{\beta}_k(u_i, v_i)}{\hat{\sigma}\sqrt{c_{kk}}} \tag{12}$$

Under H_0 , T_{count} will follow the distribution of t with a free degree $\left(\frac{\delta_1^2}{\delta_2}\right)$. $\hat{\sigma}$ obtained by rooting $\hat{\sigma}^2 = \frac{SSE\ (H_1)}{\delta_1}$. If a significance level (α) is given, then the decision is taken by rejecting the hypothesis nol (H_0) if $|T_{count}| > t_{\alpha/2,df}$, where $df = \left(\frac{\delta_1^2}{\delta_2}\right)$.

2.3. The Concept of Income Inequality

One measure to calculate income inequality is to use the Gini coefficient/Gini ratio. The Gini coefficient is a measure of imbalance or inequality which ranges from zero (perfect equalization) to one (perfect inequality). The formula for calculating the Gini ratio:

$$G = 1 - \sum_{i=1}^{k} P_i\,(Q_i + Q_{i-1})$$

where:
Pi: percentage of households or residents in class *i*-th
Qi: cumulative percentage of total income or expenditure up to class *i*-th
The Gini ratio value ranges from 0 and 1, if:

- G < 0,3 = low inequality
- 0,3 ≤ G ≤ 0,5 = moderate inequality
- G > 0,5 = high inequality

3. Methodology

3.1. Data Sources and Research Variables

This study uses secondary data derived from publications published by BPS East Java Province. This study uses panel data, consisting of time series data from 2010-2014 and cross-section data covering 38 districts/cities in East Java Province. The response variable in this study is the Gini ratio and the predictor variables are as follows:

x_1 : Human Development Index (HDI)
x_2 : Workforce Participation Rate
x_3 : Percentage of Poor Population
x_4 : Percentage of GDP Regional in the category of Fisheries Forestry Agriculture
x_5 : Percentage of GDP Regional in the Processing Industry category
x_6 : Percentage of GDP Regional in the Construction category
x_7 : Percentage of GDP Regional Large and Retail Trade, Repair and Maintenance of Cars and Motorbikes
x_8 : Economic Growth
x_9 : Percentage of GDP Regional Gross Fixed Capital Formation
x_{10} : Percentage of GDP Regional Information and Communication
x_{11} : Per-capita GDP Regional
x_{12} : Dependency Ratio

3.2. Stage of Research

The analytical method used in this study is Fixed Effect Geographically Weighted Panel Regression using software R. Following are the steps taken to analyze the data in this study:

1. Obtain data Gini ratio along with the variables that influence it.
2. Estimating the fixed effect model.
3. Conduct the Chow Test and Hausman Test to select models among common effect models with fixed effect models and fixed effect models with random effect models.
4. Test panel regression assumptions, namely normality test, non-autocorrelation test, homoskedasticity, and non-multicollinearity.
5. Conduct testing of spatial heterogeneity and assumptions of local non-multicollinearity.
6. Calculates the Euclidean distance between the *i-th* location which is located in the coordinates (u_i,v_i) of the *j-th* location which is located in the coordinates (u_i,v_i).
7. Calculate the optimum bandwidth with local sample data (average data for all time) using the CV method and adaptive bisquare kernel weighting matrix.
8. Perform parameter estimation Fixed Effect model Geographically Weighted Panel Regression uses the deviation between the data with the average data for each unit location to the time unit using the adaptive bisquare kernel weighting matrix.
9. Testing the Fixed Effect Geographically Weighted Panel Regression model.
10. Get the final model and coefficient of determination and interpret the model.

4. Results and Discussion

4.1. Panel Data Regression Model

Common Effect Model (CEM)

The estimation results of the common effect model using software R, the regression model is obtained as follows:

$$\begin{aligned}\hat{y}_{it} = {} & -0{,}20005 + 0{,}004898\,X_{1it} + 0{,}001522\,X_{2it} - 0{,}0001955\,X_{3it} + 0{,}0001438\,X_{4it} \\ & - 0{,}0002591\,X_{5it} + 0{,}005801\,X_{6it} + 0{,}001717\,X_{7it} + 0{,}001089\,X_{8it} \\ & - 0{,}002975\,X_{9it} - 0{,}003122\,X_{10it} + 0{,}0002006\,X_{11it} + 0{,}001839\,X_{12it}\end{aligned}$$

Fixed Effect Model (FEM)

The results of the fixed effect estimation model using R software, the regression model is obtained as follows:

$$\begin{aligned}\hat{y}_{it} = {} & 0{,}020310\,X_{1it} - 0{,}0012882\,X_{2it} - 0{,}0041776\,X_{3it} + 0{,}0023257\,X_{4it} + 0{,}0017433\,X_{5it} \\ & - 0{,}0096911\,X_{6it} + 0{,}012795\,X_{7it} + 0{,}0021746\,X_{8it} + 0{,}0029441\,X_{9it} \\ & - 0{,}0028707\,X_{10it} - 0{,}000084453\,X_{11it} - 0{,}026688\,X_{12it}\end{aligned}$$

Random Effect Model (REM)

The estimation results of the random effect model using R software, the regression model is obtained as follows:

$$\begin{aligned}\hat{y}_{it} = {} & -0{,}18514833 + 0{,}00501941 + 0{,}00123246X_{2it} - 0{,}00048579\,X_{3it} \\ & - 0{,}0001031902\,X_{4it} - 0{,}00031902\,X_{5it} + 0{,}00530969\,X_{6it} \\ & + 0{,}00158295\,X_{7it} + 0{,}00127924\,X_{8it} - 0{,}00273082\,X_{9it} \\ & - 0{,}00346951\,X_{10it} - 0{,}00018494\,X_{11it} - 0{,}00189143\,X_{12it}\end{aligned}$$

4.2. Chow Test, Hausman Test, and Lagrange Multiplier

Table 1. Test Selection of the Panel Regression Model

Description	Result	Conclusion
Chow Test	F= 1.9093 p-value = 0.00387 (Significant) α=0,15	H_0 rejected, the model used the fixedeffect model (FEM)
Hausman Test	W= 18.14 p-value = 0.01115 (Significant)	H_0 rejected, the model used the fixedeffect model (FEM)

	α=0,15	
Lagrange Multiplier Test	LM= 0.82029 p-value = 0,3651 (Not Significant) α=0,15	Failed to reject H_0, the model used the common effect model (CEM)

Based on table 1 above, it can be concluded that the panel regression model selected is a fixed effect model (FEM), which will then be tested for panel assumption regression.

4.3. Testing Panel Regression Assumptions

In testing panel data regression assumptions, it was concluded that the residuals were normally distributed, in the test the variance of homoscedasticity from residuals was not constant, in the non-multicollinearity test there was no linear relationship between the independent variables and in the non-autocorrelation test there was no residual autocorrelation.

4.4. Spatial Heterogeneity Testing

In testing panel data regression assumptions, there are unmet assumptions, namely the presence of heterogeneity. Based on the unit cross-section which is the region, the possibility of heterogeneity that occurs is due to the condition of the non-homogeneous observation area resulting in spatial heterogeneity. Based on spatial heterogeneity testing with α = 0.15, the BP value is 18.798 and p-value is 0.09352 or p-value < 0.15, this means that the p-value is significant so that the conclusion H_0 is rejected, this indicates that there is spatial heterogeneity.

4.5. Fixed Effect Geographically Weighted Panel Regression Model

The first GWPR Fixed Effect modeling procedure is to determine the location of each sample to be used, namely geographical location. Then, calculating the average dependent and independent variables for the entire time at each location to get the optimum bandwidth value using the cross-validation (CV) criteria and weighting values. The weighting matrix is a diagonal weighting value that has been obtained and repeats as much as the unit of time to obtain a parameter estimate. The bandwidth used in this study is bisquare adaptive. The model formed will be different at each location.

Table 2. Optimum Bandwidth and value of *Cross Validation*

Kernel Function	Bandwith	Cross Validation (CV)
Gaussian Fixed	0.7680358	0.1463827
Gaussian Adaptive	0.03846647	0.09789518
Bisquare Fixed	2.077035	0.1456459
Bisquare Adaptive	0.08596422*	5.724455e-25*

*optimum value

Based on table 2, the optimum bandwidth used in this study is bisquare adaptive, with the smallest CV value. The model formed will be different at each location. Next is one of the GWPR fixed effect models that was formed at the Surabaya City observation location:

$$\begin{aligned}\hat{y}_{37} = &-(4.3247\mathrm{E}-17) + 0{,}049297\, X_{1(37)} + \ 0{,}000339\, X_{2(37)} - 0{,}10368\, X_{3(37)} \\ &- \ 0{,}068874\, X_{4(37)} - 0{,}01283\, X_{5(37)} + \ 0{,}1682\, X_{6(37)} - \ 0{,}0306\, X_{7(37)} \\ &+ 0{,}004436\ \ X_{8((37)} + \ 0{,}014305\, X_{9(37)} - \ 0{,}16783\, X_{10(37)} + \ 0{,}00146\, X_{11(37)} \\ &+ \ 0{,}0157239\, X_{12(37)}\end{aligned}$$

Based on the above equation, one of the interpretations is that if $X_{8(37)}$ (Economic Growth) rises by one unit, the income inequality will increase by 0,004436 units assumsing the value of other predictor variables remains. Thus, If the Surabaya city government wants to reduce income inequality by controlling economic growth and carrying out development in all fields that can be enjoyed by all levels of society.

4.6. Testing Fixed Effect of Geographically Weighted Panel Regression Model

Based on table 3 the results of the model match test it can be shown that the value of F_{count}> F_{table} and p-value < 0,05, it can be concluded that there is a significant difference between the fixed effect panel data regression model and the GWPR fixed effect.

Table 3. Model Compatibility Test

Result	Conclusion
F_{count} = 19.662 F_{table} (0.05, 164.12) = 2.3291 *p*-value = 0.00006259 (Significant) α=0.05	H_0 rejected, showing that there is a significant difference between the fixed effect panel data regression model and GWPR fixed effect.

Based on table 4, the selection of the best model is compared by comparing the value of R^2 and RMSE fixed effect panel data regression model with GWPR fixed effect. From the table 4, it can be seen that the R^2 of the GWPR fixed effect model is greater than the panel fixed effect (FEM) data regression and the GWPR fixed effect RMSE value is smaller than the panel fixed effect (FEM) data regression. So that the best model chosen is the GWPR fixed effect.

Table 4. Selection of the Best Models

Model	R^2	RMSE
GWPR	99.953	0.0061035
FEM within Estimator	22.844	0.1035616

Based on figure 1. The testing significance of parameters using α =0,05 districts/cities in East Java can be grouped based on variables that significantly affect inequality resulting in 28 groups of variables as shown in Figure 1. Each location requires different significant variables. For examples Sumenep and Pamekasan district therae are two significant variables, namely variables X_{11} (Per-capita GDP Regional) and X_{12} Dependency Ratio.

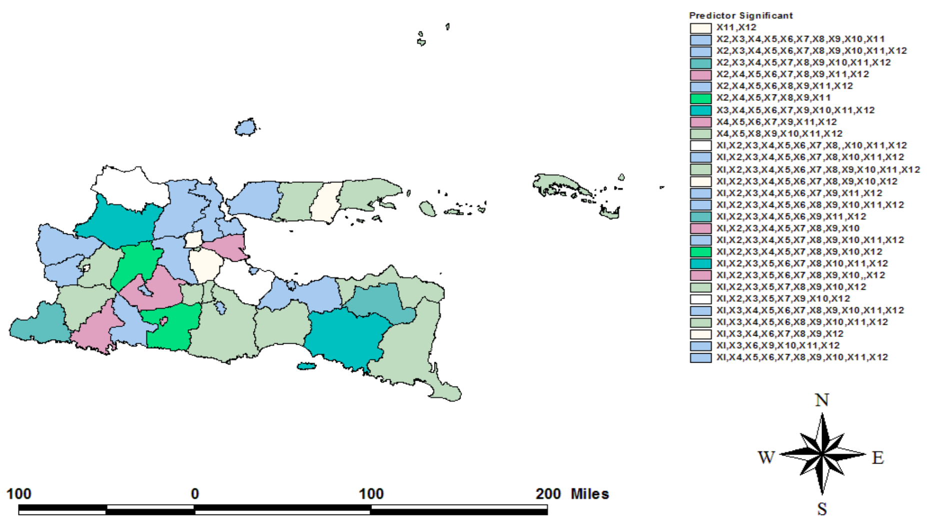

Figure 1. Revenue Gap Map in East Java based on Significant Variables

The following is a partial test of the district/city parameters, partial test parameters for the city of Surabaya are as follows:

Table 5. Partial Test of Surabaya City Parameters

Variables	t-Statistics	*p*-value	Variable	t-Statistic	p-value
X_1	5.280957	0.000573429	X_7	3.253143	0.006998
X_2	0.234175	0.41000*	X_8	3.040013	0.010000
X_3	16.600073	3.52E-07	X_9	3.261313	0.006920
X_4	5.490371	0000000	X_{10}	5.50259	0.000000
X_5	1.951643	0.045972	X_{11}	2.371407	0.024752
X_6	7.233371	0.000000	X_{12}	4.615153	0.00000

*α=0.05 not significant

Based on table 5 above, it can be seen that all predictor variables have a significant effect on income inequality in the city of Surabaya, except variable X_2 (Workforce Participation Rate) does not significantly influence income inequality in the city of Surabaya.

5. Summary

The estimated parameters in GWPR are obtained by WLS. Goodness test follows distribution F and partial test follows distribution t. The goodness of fit of district/city income inequality in East Java in 2010-2014 uses GWPR from fixed effect panel regression. Partial tests provide different models in each location. The variables that significantly influences income inequality are the percentage of the poor, percentage of GDP regional in the category of fisheries forestry agriculture, percentage of GDP regional in the processing industry category, percentage of GDP regional gross fixed capital formation, per-capita GDP regional, and dependency ratio. R^2 of the GWPR model is 99.953% with RMSE of 0.0061035. While R^2 from the fixed effect model is 22.854% with RMSE of 0.1035616. So it can be concluded that GWPR model is better than FEM.

References

[1] Fotheringham, A S Brunsdon, C and Charlton, M. 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (West Sussex: John Wiley & Sons Ltd)
[2] Baltagi, B H 2005 *Econometric Analysis of Panel Regression* (New York : John & Wiley Ltd)
[3] Yu, Danlin 2010 Exploring Spatiotemporally Varying Regressed Relationships The Geographically Weighted Panel Regression Analysis *The International Archives of the Photogrammet Remote Sensing and Spatial Information Sciences* page 134-139
[4] Bruna, Fernando and Yu, Danlin 2013 Geographically Weighted Panel Regression *A Coruna: s.n., XI Congreso Galego de Estatica e Investigacion de Operations*
[5] Cai, Ruohong, Yu, Danlin and Oppenheimer, Michael 2014 Estimating the Spatial Varying Responses of Corn Yields to Weather Variations using Geographically Weigted Panel Regression *Journal of Agricultural and Resource Economics* page 230-252
[6] Bruna, Fernando and Yu, Danlin 2016 Geographically Weighted Panel Regression and Development Accounting for European Region *International Conference on Regional Science.*
[7] Hsiao, C 2003 *Analysis of Panel Data* (New York : Cambridge University Press)
[8] Greene, W H 2000 *Econometric Analysis* (New Jersey : Prentice-Hall Inc.)