Village classification index prediction using geographically weighted panel regression

A S Ningrum¹, A Rusgiyono², and A Prahutama³

¹,²,³ Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Corresponding author: alan.prahutama@gmail.com

Abstract. Village classification index is a certain status of the achievements of village development activities. In measuring the achievement of village classification index, it needs to be made in several time periods and must concern to the spatial effects because the geographical conditions of each village are diverse. It is necessary to study the variables that affect the village classification index in several time periods. Statistical methods that used in overcome the spatial effects of panel data type is Geographically Weighted Panel Regression (GWPR), which is a combination of Geographically Weighted Regression (GWR) models and panel data regression. This study focused on the establishment of GWPR model with fixed effects using fixed bisquare kernel on the village classification index in Batang Regency, 2015-2018. The results of this study indicate that the fixed effect model GWPR differ significantly on panel data regression model, and the model generated for each location will be different from one another. In addition, all independent variables namely the community economy, security and order, and community participation in development have a significant effect on the village classification index for all villages with R-square value of 0.3952.

1. Introduction

Village profile is an information system which the village is directly subject to inputting data. Village Profile is a comprehensive picture of the character of villages which includes basic family data, natural resource potential, human resources, institutions, infrastructure and facilities as well as the development of progress and problems faced by villages. The level of Village Development is a certain status of the achievements of the results of development activities that reflect the level of progress and / or success of the community, village government and regional governments in carrying out development in the village [1].

The purpose of the preparation of village profile is to make a database and source of information for development needs. The level of village development which is measured using a village classification index reflects the success of village development every year and every five years is measured from the speed of the community's economic development, security and order, and community’s participation in development.

Batang Regency is a combination of coastal, lowland and mountainous regions so that the villages have a diverse geographical character. The difference in geographical conditions between one location and another location of an area allows for spatial problems that cause a significant influence on the rate of speed of economic development, security and order, and community’s participation in development in each village. Therefore, it is necessary to study variables in several time periods that affect the level of village development. In addition, a statistical modeling method is needed that takes into account the geographical location or factors of the location of observation. The statistical method used to overcome spatial effects, especially the problem of spatial heterogeneity in panel type data is Geographically Weighted Panel Regression [2].
Geographically Weighted Panel Regression (GWPR) is a combination of the Geographically Weighted Regression (GWR) model and the panel data regression model [3]. In this study, the focus is on the establishment of a fixed effect GWPR model using fixed bisquare kernel weighting on the village classification index data in the Village Profiles in Batang Regency in 2015-2018.

2. Literature review

2.1. Data panel regression analysis

Panel data regression is a combination of cross-section data and time series data. The general panel data regression model is [4]:

\[ y_{it} = \alpha + X_{it}\beta + \varepsilon_{it} \] (1)

The one-way residual component model for the panel data regression model is defined in the following equation:

\[ u_{it} = \mu + \varepsilon_{it} \] (2)

Based on Eq. (1) and (2), the general panel data model becomes:

\[ y_{it} = \alpha + \mu + X_{it}\beta + \varepsilon_{it} \] (3)

Fixed effect model is also known as the Least Square Dummy Variable (LSDV), because the observation value on the coefficient \( \mu_i \) is a dummy variable that has different values for each \( i \)-th individual. One of the parameters estimated in the fixed effect model is using within transformation [5]. The fixed effect model as follow as:

\[ y_{it}^* = X_{it}\beta + \varepsilon_{it}^* \] (4)

Random Effect Model (REM) is also called the Error Component Model (ECM). This model has the following equation:

\[ y_{it} = \alpha + \mu + X_{it}\beta + \varepsilon_{it} \] (5)

In this model, \( \mu_i \) is considered as a random variable with an average of \( \mu_0 \). So intercept can be written as \( \mu_i = \mu_0 + \nu_i \) where \( \nu_i \) is a residual with an average value of zero and variance \( \sigma^2 \). Then the equation can be written as:

\[ y_{it} = \alpha + \nu_i + X_{it}\beta + \varepsilon_{it} \] where, \( w_{it} = \varepsilon_{it} + \nu_i \)

Selection of Panel Data Regression Model using Chow test, Hausman test and Lagrange Multiplier test. The Chow test is to testing the model of regression panel data for common effect model or fixed effect model. It is following [5]:

\[ H_0 : \text{common effect model} \]

\[ H_1 : \text{fixed effect model} \]

Statistical test: \( F_0 = \frac{(RRSS-URSS)/(n-1)}{URSS/(nT-n-p)} \) (6)

Where, RRSS = restricted residual sum of squares common effect model ; URSS = unrestricted residual sum of squares fixed effect model. \( H_0 \) is rejected if \( F_0 > F_{label} \) with \( F_{table} = F ((n-1,nT-n-p, \alpha)) \) or p-value < \( \alpha \), which means the model used is the fixed effect model. The Hausman test is testing to model of regression panel data for random effect or fixed effect. It is as follow as [5]:

\[ H_0 : \text{random effect model} \]

\[ H_1 : \text{fixed effect model} \]

Statistical test: \( W = [\hat{\beta}_{REM} - \hat{\beta}_{FEM}] \left[ var [\hat{\beta}_{REM} - \hat{\beta}_{FEM}] \right]^{-1} [\hat{\beta}_{REM} - \hat{\beta}_{FEM}] \) (7)

\( H_0 \) is rejected if \( W > \chi^2_{\alpha,p} \) or p-value < \( \alpha \), which means the model used is the fixed effect model.

The Lagrange multiplier test is to testing model of regression panel data which is common effect model or random effect model. The testing as follow as [5]:

\[ H_0 : \text{common effect model} \]

\[ H_1 : \text{random effect model} \]
H₁ : random effect model
Statistical test: \( LM = \frac{nT}{2(T-1)} \left[ \sum_{i=1}^{n} \left( \sum_{t=1}^{T} \epsilon_{it} \right)^2 / \sum_{i=1}^{n} \epsilon_{it}^2 - 1 \right]^2 \)  
(8)

\( H_0 \) is rejected if \( LM > \chi^2_{1,\alpha} \) or p-value < \( \alpha \), which means the model used is the random effect model.

2.2. The assumptions of data panel regression
There are some assumption of Data Panel regression model such as normality of residual, homoscedasticity of residual; non-autocorrelation of residual; and non-multi-collinearity of predictors. For normality test, we can use Jarque Bera (JB) test, with hypothesis as:

\( H_0: \) Residuals are normally distributed
\( H_1: \) Residuals are not normally distributed

Statistical test: \( JB = n \left[ \frac{S^2}{6} + \frac{(K-3)^2}{24} \right] \)  
(9)

\( H_0 \) is rejected if \( JB > \chi^2_{2,\alpha} \) or p-value < \( \alpha \), which means the residuals are not normally distributed. One test used in analyzing the homoscedasticity assumption is the Glejser test. Glejser suggests regressing the absolute value of residuals with independent variables . Autocorrelation assumption aims to determine the possibility in a model there is a correlation between residuals in period \( t \) with residuals in period \( t-1 \). To detect the existence of autocorrelation can use the Durbin Watson method [6]. Multicollinearity is a linear relationship between independent variables in the model. One way to detect the presence of multicollinearity is by calculating VIF (Variance Inflation Factor) where multicollinearity is detected if the VIF value > 10 [6].

2.3. Fixed effect geographically weighted panel regression model
Geographically Weighted Panel Regression model is a combination of the GWR model and the panel data regression model by involving the time element in the GWR model. The equation of the Fixed Effect Geographically Weighted Panel Regression model is [7]:

\( y_{it} = \beta_k(u_i, v_i)X_{itk} + \epsilon_{it}, \quad i = 1,2,...,n \text{ dan } t = 1,2,...,T \)  
(10)

In spatial modelling, we can test the spatial effect of the data by spatial heterogeneity. Spatial heterogeneity is caused by spatial unit conditions in an inhomogeneous observation area. The Breusch-Pagan test method can be used to test spatial heterogeneity as [1]:

\( H_0: \) there is no spatial heterogeneity
\( H_1: \) there is spatial heterogeneity

Statistical test: \( BP = (1/2)f^T Z^T Z^{-1} Z^T f \sim \chi^2_{(p)} \)  
(11)

\( H_0 \) is rejected if \( BP > \chi^2_{(p)} \) or p-value < \( \alpha \), which means there is spatial heterogeneity. To estimate the parameters of fixed effect geographically weighted panel regression model using weighted least square approach as estimated in GWR.

\( \hat{\beta}(u_i, v_i) = (X^TW(u_i, v_i)X)^{-1}X^TW(u_i, v_i)y \)  
(12)

In the weighting of the GWPR model is the same as the weighting of the GWR model, which depends on the distance between points of observation location. Observations in local sampling locations will be weighted based on kernel functions in GWPR as well as in GWR [7]. One is to determine the weighting matrix by using the kernel function. One type of kernel function used in this study is the fixed kernel function with the bisquare kernel function which is stated by the following formula [7]:

\( w_{ij} = \left( 1 - \left( \frac{d_{ij}}{h} \right)^2 \right)^2, \text{ untuk } d_{ij} < h, \)  
(13)
In the fixed kernel weighting function, there are bandwidth parameters. Bandwidth is analogous to the radius of a circle, so that an observation location that is within the radius of a circle is still considered influential in forming parameters at the i-observation location. There are several methods that can be used to choose the optimum bandwidth and one of them uses Cross Validation (CV). The CV calculation in GWPR is the same as the GWR which is calculated based on the average of the dependent and independent variables for the whole time and is defined as follows [8]:

\[
CV = \sum_{i=1}^{n} \left( \bar{y}_i - \hat{y}_{si}(b) \right)^2
\]

where \(y_i\) is the average over time the dependent variable at the observation location \(i\) and \(\hat{y}_{si}(b)\) is the estimator value of \(y_i\) with bandwidth \(b\) with observations at the location \((u_i, v_i)\) removed from the estimation process. The testing of Fix Geographically Weighted Panel regression are simultaneous and partial test. The simultaneous test as follow as [8]:

\[
H_0 : \beta_k(u_i, v_i) = \beta_k\quad \text{for each } k = 1,2,\ldots,p \text{ and } i = 1,2,\ldots,n
\]

(there is no significant difference between the panel data regression model and GWPR)

\[
H_1 : \text{there is at least one } \beta_k(u_i, v_i) \neq \beta_k\quad \text{for each } k = 1,2,\ldots,p \text{ and } i = 1,2,\ldots,n
\]

(there is significant difference between the panel data regression model and GWPR)

Statistical test: \(F = \frac{RSS(H_0)/df_1}{RSS(H_1)/df_2}\),

\[
H_0 \text{ is rejected if } F > F_{1-\alpha,df_1,df_2} \text{ or } \text{p-value} < \alpha, \text{which means there is significant difference between the panel data regression model and GWPR.}
\]

The partial test as follow as:

\[
H_0 : \beta_k(u_i, v_i) = 0, \quad \text{for each } k = 1,2,\ldots,p \text{ and } i = 1,2,\ldots,n
\]

\[
H_1 : \text{there is at least one } \beta_k(u_i, v_i) \neq 0 \quad \text{for each } k = 1,2,\ldots,p \text{ and } i = 1,2,\ldots,n
\]

Statistical test: \(T_{value} = \frac{\hat{\beta}_k(u_i, v_i)}{\delta / \sqrt{C_{kk}}}\)

Where \(\delta = \frac{RSS(H_1)}{\delta_1}\) and \(C_{kk}\) are k-diagonal element of matrix \(C_1C_1^T\) with \(C = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i)\). \(T_{value}\) will follow the t distribution with free degrees is df = \(\frac{\delta_1^2}{\delta_2}\). \(H_0\) is rejected if \(|T_{value}| > \alpha / 2\).

3. Research methods

The data used in this study are secondary data sourced from the Office of Community and Village Empowerment of Batang Regency. The variables used in this study are the village classification index data as the dependent variable and independent variables, namely the community economy (X1), security and order (X2), and community participation in development (X3) with the observation unit used covering 31 villages in Batang Regency from 2015 to 2018 and the geographical location of each village in Batang Regency. The analytical method used in this study is Fixed Effect Geographically Weighted Panel Regression using R software. The steps taken to analyze the data in this study:

1. Obtain Village Profile data
2. Estimating panel data regression parameters using the common effect, fixed effect, and random effect models.
3. Conduct a Chow Test to select a model between the common effect model and the fixed effect model.
4. Conduct Hausman Test to choose a model between the random effect model and the fixed effect model.
5. Testing the assumptions of panel data regression namely normality test, non-autocorrelation test, heteroscedasticity and non-multicollinearity.
6. Conduct spatial heterogeneity testing.
7. Testing local non-multicollinearity assumptions.
7. Calculating the euclid's distance between the $i$-th location and the $j$-location located at coordinates $(u_i, v_i)$.
8. Calculate the optimum bandwidth with the minimum CV method.
9. Calculate the fixed bisquare weighting matrix using optimum bandwidth.
10. Estimating the parameters of the fixed effect GWPR model using the fixed bisquare weighting matrix.
11. Testing the GWPR fixed effect model.
12. Get the final model and the coefficient of determination.
13. Interpret the model that has been obtained.

4. Results and discussions

4.1. Panel data regression model
There are three model for panel data regression such as common effect, fixed effect, and random effect for modelling village classification index. The estimation results of the common effect as follow as:

$$\hat{y}_{it} = 0.169 + 0.137X_{1it} + 0.388X_{2it} + 0.172X_{3it}$$

The estimation results of the fixed effect as follow as:

$$\hat{y}_{it} = 0.128X_{1it} + 0.263X_{2it}$$

The estimation results of the random effect as follow as:

$$\hat{y}_{it} = 0.302 + 0.134X_{1it} + 0.317X_{2it} + 0.098X_{3it}$$

From the three model, we conclude that the independent variables have positive impact for dependent variable. In fixed effect model, the variable of $X_{3it}$ not significant. The variable of $X_{2it}$ has the biggest impact than others. For selecting the best model, we should get the Chow test and Hausmant test as follow as in table 1.

| Table 1. Selecting panel data regression model |
|-----------------------------------------------|
| Chow Test | Result | Conclusion | Hausman Test | Result | Conclusion |
| F0 = 14.357 and p-value = 0.000 | H0 rejected, which means the model used is the fixed effect model. | W = 10.466 and p-value = 0.000 | H0 rejected, which means the model used is the fixed effect model. |

Based on Table 1, we concluded that the best model for modelling the village classification index is Fixed effect model. After that, we checking the assumptions of Fixed effect model of panel data regression. In testing the panel data regression assumptions, it was concluded that in the residual normality test there was no normal distribution, in the heteroscedasticity test variants of the constant residuals, in the non-autocorrelation test there was no autocorrelation in the residuals, and in the non-multicollinearity test there was no linear relationship between the independent variables.

4.2. Fixed effect geographically weighted panel regression model
Based on the spatial heterogeneity test obtained a BP value of 16.335 and a p-value of 0.0009 which indicates that there is spatial heterogeneity. Meanwhile, the assumption of local non-multicollinearity is fulfilled because the VIF value of each independent variable for each location < 10 indicates that there is
no local multicollinearity. Based on GWPR fixed effect modeling with fixed bisquare kernel weighting for each village location, the estimated parameter values are different because each location has a different weighting value. For estimation model can be seen in table 2. The coefficient of determination (R^2) generated from the fixed effect GWPR model with fixed bisquare kernel weighting formed 39.52%.

Table 2. Estimation parameter of fix geographically weighted panel data regression in village classification index in batang regency, Indonesia

| Area             | Coefficient parameter | T-value | p-value |
|------------------|-----------------------|---------|---------|
|                  | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| Bandung          | 0.213    | 0.696    | -0.057   | 2.4725   | 7.5225   | -0.6582   | 0.015    | 0.000    | 0.512    |
| Bawang           | 0.047    | 0.559    | 0.215    | 0.7759   | 12.8351  | 3.9314    | 0.439    | 0.000    | 0.000    |
| Bawang-Blado     | 0.122    | 0.584    | 0.140    | 2.1895   | 14.1605  | 2.8858    | 0.031    | 0.000    | 0.004    |
| Botolambat       | 0.248    | 0.649    | -0.020   | 3.1519   | 8.5417   | -0.2595   | 0.002    | 0.000    | 0.796    |
| Brayo            | 0.206    | 0.706    | -0.066   | 2.3032   | 7.4155   | -0.7389   | 0.023    | 0.000    | 0.462    |
| Gerlang          | 0.020    | 0.530    | 0.268    | 0.3294   | 14.4132  | 5.5418    | 0.743    | 0.000    | 0.000    |
| Gombong          | 0.190    | 0.718    | -0.067   | 2.1373   | 7.2645   | -0.7409   | 0.035    | 0.000    | 0.460    |
| Gringgingingsari | 0.171    | 0.735    | -0.080   | 2.2710   | 8.8534   | -0.9017   | 0.025    | 0.000    | 0.369    |
| Gumawang         | 0.175    | 0.715    | -0.051   | 2.5174   | 9.5111   | -0.6249   | 0.013    | 0.000    | 0.533    |
| Gumangsari       | 0.019    | 0.544    | 0.250    | 0.3122   | 13.3144  | 4.6139    | 0.756    | 0.000    | 0.000    |
| Harjowinangun Barat | 0.196  | 0.681    | -0.025   | 3.2954   | 11.5113  | -0.3331   | 0.001    | 0.000    | 0.739    |
| Kalimanggis      | 0.230    | 0.677    | -0.045   | 2.7802   | 7.6859   | -0.5321   | 0.006    | 0.000    | 0.596    |
| Kalipucang       | 0.267    | 0.612    | 0.016    | 3.5781   | 9.1557   | 0.2329    | 0.001    | 0.000    | 0.816    |
| Kulon            | 0.022    | 0.535    | 0.261    | 0.3473   | 14.2094  | 5.1888    | 0.729    | 0.000    | 0.000    |
| Kebatuan          | 0.256    | 0.636    | -0.008   | 3.3260   | 8.7759   | -0.1031   | 0.001    | 0.000    | 0.918    |
| Kecepak          | 0.270    | 0.609    | 0.019    | 3.6519   | 9.2743   | 0.2806    | 0.000    | 0.000    | 0.779    |
| Klidang Lor      | 0.200    | 0.697    | -0.046   | 3.1954   | 9.8609   | -0.5658   | 0.002    | 0.000    | 0.573    |
| Kranggan         | 0.068    | 0.567    | 0.193    | 1.1607   | 12.9563  | 3.6075    | 0.248    | 0.000    | 0.000    |
| Pasusukan        | 0.152    | 0.602    | 0.099    | 2.6470   | 14.3043  | 2.0005    | 0.009    | 0.000    | 0.048    |
| Pesantren        | 0.156    | 0.618    | 0.079    | 2.7357   | 13.6764  | 1.4655    | 0.007    | 0.000    | 0.146    |
| Plosovangi       | 0.023    | 0.534    | 0.263    | 0.3758   | 14.4124  | 5.3266    | 0.708    | 0.000    | 0.000    |
| Pranten          | 0.258    | 0.631    | -0.003   | 3.3771   | 8.8500   | -0.0399   | 0.001    | 0.000    | 0.968    |
| Proyonanggian Selatan | 0.186  | 0.660    | 0.008    | 3.1186   | 12.7072  | 0.1257    | 0.002    | 0.000    | 0.900    |
| Rejosari Timur   | 0.118    | 0.589    | 0.139    | 2.0932   | 13.6764  | 2.7187    | 0.039    | 0.000    | 0.007    |
| Sibebek          | 0.163    | 0.740    | -0.074   | 1.9870   | 7.8941   | -0.8194   | 0.049    | 0.000    | 0.414    |
| Sijono           | 0.260    | 0.627    | 0.002    | 3.3079   | 8.8589   | 0.0265    | 0.001    | 0.000    | 0.979    |
| Soka             | 0.068    | 0.566    | 0.194    | 1.1497   | 13.0289  | 3.6657    | 0.253    | 0.000    | 0.003    |
| Tembok           | 0.037    | 0.527    | 0.267    | 0.6540   | 15.5788  | 5.9819    | 0.515    | 0.000    | 0.000    |
| Wanar            | 0.153    | 0.601    | 0.112    | 2.3979   | 13.7569  | 2.1652    | 0.018    | 0.000    | 0.033    |
| Warungasem       | 0.253    | 0.639    | -0.009   | 3.2216   | 8.6333   | -0.1245   | 0.002    | 0.000    | 0.901    |
| Wonokerso        | 0.185    | 0.691    | -0.027   | 2.8853   | 10.9514  | -0.3868   | 0.004    | 0.000    | 0.713    |

The following is one of the fixed effect GWPR models formed in Kranggan Village:

$$\hat{y}_{17t} = 0.200X_{117t} + 0.697X_{217t}$$
With:
\(\hat{y}_{i17t}\) : Estimator of the 17-th Location Classification Index (Kranggan Village) in the t-year
\(X_{117t}\) : Community Economy Location 17th (Kranggan Village) t-year
\(X_{217t}\) : Security and Order 17th location (Kranggan Village) t-year

The equation explains that the village classification index will increase by 0.200 if the community economy increases by 1 unit assuming other variables are constant. The village classification index will increase by 0.697 if security and order increase by 1 unit assuming other variables are constant. While the community participation variable in development did not significantly influence the village classification index in Kranggan Village. The simulataneous testing of fixed effect geographically weighted panel regression model in table 3.

| Table 3. Goodness of fit test |
|-------------------------------|
| Result                        | Conclusion                             |
| F = 14.891                    | \(H_0\) rejected, there is significant difference between the panel data regression model and GWPR |
| and                          |                                        |
| p-value = 0.000               |                                        |

Based on table 3, the goodness of fit test using \(\alpha\) of 5\%, it was concluded that there was a significant difference between the fixed effect panel data regression model and the fixed effect GWPR. Then, the partial test of fixed effect geographically weighted panel regression model in table 4.

| Tabel 4. Parameter significant test |
|-------------------------------------|
| Variable, t-Statistics, p-value, Conclusion |
| X₁  3.1954  0.002  In Kranggan Village, variables X₁ and X₂ namely community economy and security and order affect the village classification index. |
| X₂  9.8609  0.000  Whereas variable X₃, namely community participation in development, does not affect the village classification index. |
| X₃ -0.5658  0.573  |
5. Conclusion
Based on the model selection in the panel regression, the fixed effect model was chosen to be applied to the Geographically Weighted Panel Regression model. In addition, based on testing the suitability of the model shows that the fixed effect GWPR model is significantly different from the panel data regression model, and the resulting model for each location will be different from one another. In the fixed effect GWPR model produces $R^2$ value of 39.52%.

References
[1] Mendagri 2007 *Pedoman Penyusunan Dan Pendayagunaan Data Profil Desa Dan Kelurahan* Jakarta Menteri Dalam Negeri Republik Indonesia
[2] Fotheringham AS Brunsdon C and Charlton M 2002 *Geographically Weighted Regression* England John Wiley & Sons Ltd
[3] Anselin L 1988 *Spatial Econometrics: Methods and Models* Dordrecht Kluwer Academic Publishers
[4] Gujarati DN 2004 *Basic Econometrics* (4 ed.) New York The McGraw-Hill Companies
[5] Baltagi BH 2005 *Econometrics Analysis of Panel Data* 3 ed. England John Wiley & Sons Ltd
[6] Greene WH 2002 *Econometric Analysis* 5 ed. New Jersey Prentice Hall International
[7] Yu D 2010 *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Science* 38
[8] Cai R, Oppenheimer M and Yu D 2014 *Journal of Agricultural and Resource Economics*. 