JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 1) JUNE 2024: 41 - 53
https://doi.org/10.21009/JSA.08104

MODELING GROSS ENROLLMENT RATE FOR HIGHER EDUCATION IN CENTRAL JAVA PROVINCE USING PRINCIPAL COMPONENT GEOGRAPHICALLY WEIGHTED REGRESSION APPROACH

Mohd Syafrizal [1*], Widyanti Rahayu[2], Dania Siregar[3]

[1,2,3]Statistics Study Program, Faculty of Mathematics and Natural Sciences, State University of Jakarta, Rawamangun Muka St., East Jakarta, DKI Jakarta, 13220, Indonesia.

*Corresponding author's e-mail: ** ***mohdsyafrizal25@gmail.com***

ABSTRACT

Article History:
Received: 13 April 2024
Revised: 30 April 2024
Accepted: 9 June 2024
Published: 30 June 2024
Available online.

Keywords:
Gross enrollment rate for higher education; Geographically weighted regression; Spatial heterogeneity; Multicollinearity; Principal component analysis.

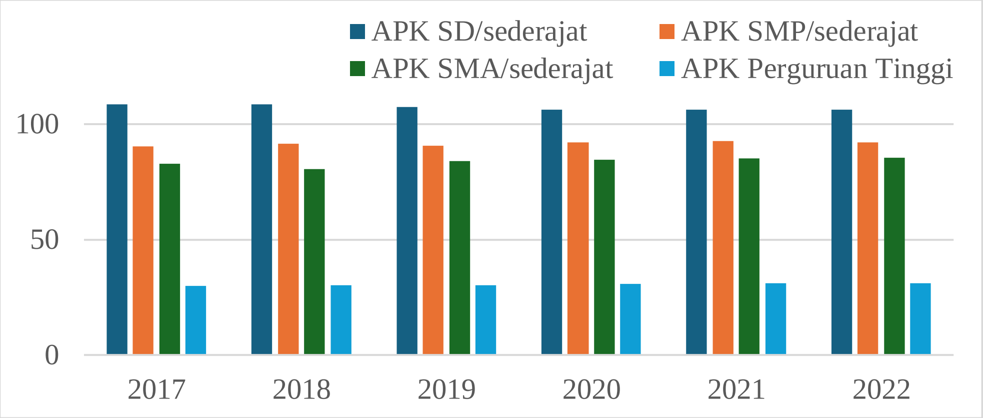
Education in a country is a crucial factor in enhancing human resources. The Gross Enrollment Rate for Higher Education is one of the important indicators used by the government to evaluate the development of the education sector, particularly higher education. Social and cultural diversity, as well as geographical influences, result in varying conditions across different regions, leading to each region having its own unique characteristics, known as spatial heterogeneity. Geographically Weighted Regression (GWR) is a method that can address the problem of spatial heterogeneity. Additionally, a common issue encountered in modeling with many independent variables is multicollinearity, which can lead to high variance in regression parameter estimates and invalid conclusions. Principal Component Analysis (PCA) is a dimensionality reduction method that can address multicollinearity. The aim of this research is to model Gross Enrollment Rate in Higher Education in Central Java using GWR, preceded by handling multicollinearity with PCA. Furthermore, this study aims to determine the factors influencing Gross Enrollment Rate in Higher Education in Central Java. The results, using a fixed Gaussian kernel weighting function, indicate that modeling with PCA and GWR performs better than using the Ordinary Least Square (OLS) method alone, yielding an AIC value of 169.43 and an R^2 of 96.2%.

How to cite this article:
M. Syafrizal, W. Rahayu, D. Siregar, "MODELING GROSS ENROLLMENT RATE FOR HIGHER EDUCATION IN CENTRAL JAVA PROVINCE USING PRINCIPAL COMPONENT GEOGRAPHICALLY WEIGHTED REGRESSION APPROACH", Jurnal Statistika dan Aplikasinya, vol. 8, iss. 1, pp. 41 – 53, June 2024

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

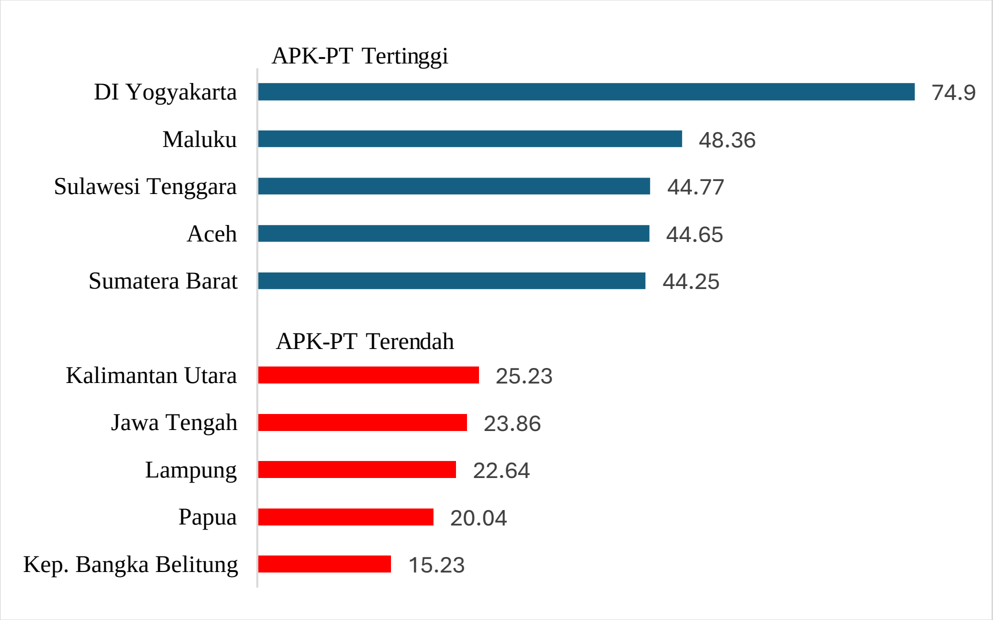
Education is a deliberate and planned effort to enhance skills and potentials in order to benefit future citizens (according to UU No. 20 Tahun 2003). The role of education is crucial in national development because it is a means of improving human resources. The quality of a country's human resources reflects the quality of its education; therefore, addressing educational disparities is essential to realize national progress (Maskar et al., 2022). Gross Enrollment Rate is one of the criteria used by the government to assess the level of success in the field of education.



Source: Badan Pusat Statistik (2022).

Figure 1. Gross Enrollment Rate by Education Level

The higher the level of education of the community, the tendency for the participation rate to decrease. The participation rates for primary to secondary education are above 80%, but the participation rate for higher education is only around 30%. However, higher education holds significant value as it can enhance an individual's qualifications and skills, as well as provide opportunities for better employment and income (Khadijah et al., 2017).



Source: Badan Pusat Statistik (2022).

Figure 2. Provinces with the highest and lowest Gross Enrollment Rate for Higher Education in 2021

The Gross Enrollment Rate for Higher Education in Central Java is among the lowest compared to other provinces. This also indicates that Central Java has the lowest Gross Enrollment Rate for Higher Education among the provinces on the island of Java. The Central Java Provincial Government stated that the lack of progress in education is not solely due to educational infrastructure issues, but also because others have not fully participated (Humas Jateng, 2017). To support planned programs, especially in the education sector, it is important to evaluate and understand the factors influencing the participation rate in pursuing higher education in Central Java.

Regression analysis is a method to determine the factors influencing a dependent variable with one or more independent variables. In the context of Gross Enrollment Rate for Higher Education, regression analysis helps identify the factors that may influence the participation rate of the population in pursuing higher education. One common violation in regression models, especially concerning the geographical and socio-cultural conditions of different regions, is spatial heterogeneity. In such situations, the model may provide less accurate information due to the lack of diversity among regions caused by different observation locations. To address spatial heterogeneity, weighting is needed in regression analysis so that each location can explain the model more accurately. One development in regression analysis that involves weighting is using Geographically Weighted Regression, commonly abbreviated as GWR (Fotheringham et al., 2002).

Another issue often encountered in forming linear regression models involving multiple independent variables is that variables with high correlations can lead to multicollinearity. The presence of multicollinearity can cause the estimation of parameter results to have high variability and lead to invalid conclusions (Montgomery et al., 2012). One way to address multicollinearity is to first transform the independent variables into new variables that are uncorrelated. Principal Component Analysis (PCA) is an analytical method that transforms original variables into smaller dimensions while still explaining the initial variability. The advantage of PCA is that due to the reduction in dimensionality of a set of variables, it can simplify data and interpretation while retaining most of the information.

Jinglei et al. (2013) utilized PCA, which was then applied in GWR modeling to address spatial heterogeneity and multicollinearity in estimating the parameters of the winter wheat water requirement model in Northern China. The study yielded a modeling result using PCA and GWR that was superior to using a global regression model based on AIC and R^2 coefficients. Meanwhile, Sari et al. (2016) conducted research to model Regional Original Income in Central Java, which resulted in a better GWRPCA model compared to PCA regression based on AIC and determination coefficients. Additionally, Zhu et al. (2020) demonstrated that modeling with the Principal Component Geographically Weighted Regression approach would produce clearer and significantly superior performance compared to regression models using OLS.

2. METHODS

Material and Data

The data for this research is secondary data sourced from the Central Java Provincial BPS. The data was extracted from 4 different publications, including "Statistik Pendidikan 2021", "Statistik Kesejahteraan Rakyat Provinsi Jawa Tengah Tahun 2021”, "Statistik Potensi Desa 2021", and " Provinsi Jawa Tengah dalam Angka 2022”. The data the Gross Enrollment Rate for Higher Education for the dependent variable consists of 35 observations covering 29 districts and 6 cities scattered across Central Java Province. Meanwhile, the number of independent variables is 12 suspected to influence the Gross Enrollment Rate for Higher Education, including student-to-university ratio (X_1), student-to-faculty ratio (X_2), percentage of population completing high school or equivalent education (X_3), percentage of households owning a computer/laptop/tablet (X_4), unemployment rate (X_5), percentage of poor population (X_6), gross regional domestic product per capita (X_7), fiscal decentralization ratio (X_8), percentage of households receiving family hope program (X_9), average length of schooling (X_{10}), number of villages/urban neighborhoods with weak or no internet signal (X_{11}), and number of villages/urban neighborhoods with no public transportation available (X_{12}).

Research Method

Multiple linear regression

Multiple linear regression analysis is a method to study the relationship between one response variable and more than one independent variable. The model is written as follows (Montgomery et al., 2012):

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \tag{1}$$

and determining parameter estimators using Ordinary Least Squares (OLS) with:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{2}$$

Multicollinearity

Multicollinearity is a condition of high correlation relationships between two or more independent variables in a regression model. Multicollinearity can be detected by observing the value of VIF (Variance Inflation Factor) calculated by:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3}$$

where R_j^2 is the coefficient of determination of regressing the j-th independent variable (for $j = 1,2,\dots k$) with the other $k-1$ independent variables. Multicollinearity will be detected if $VIF > 10$.

Principal component analysis (PCA)

Let $\boldsymbol{X}' = [X_1, X_2, \dots, X_p\ ]$ have a matrix $\boldsymbol{A}$ as the variance-covariance/correlation matrix which has paired eigenvalues $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_k, e_k)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$, then the formation of its principal components is as follows (Johnson & Wichern, 2007):

$$\begin{aligned} K_1 &= \boldsymbol{e}_1'\boldsymbol{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1k}X_k \\ K_2 &= \boldsymbol{e}_2'\boldsymbol{X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2k}X_k \\ &\vdots \\ K_k &= \boldsymbol{e}_k'\boldsymbol{X} = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kk}X_k \end{aligned} \tag{4}$$

where the eigenvalues of matrix $\boldsymbol{A}$ are computed under the condition:

$$|\boldsymbol{A} - \boldsymbol{\lambda I}| = 0 \tag{5}$$

while the eigenvectors of matrix $\boldsymbol{A}$ are computed by:

$$(\boldsymbol{A} - \boldsymbol{\lambda I})\boldsymbol{e} = 0 \tag{6}$$

The criteria for selecting the number of principal components are determined to represent the original variables with a minimum criterion of explaining at least 70% of the variability and eigenvalues greater than one.

Spatial analysis

Data that includes information about geographical positions or locations obtained through measurements are referred to as spatial data. According to Anselin (1988), data that exhibit spatial influence are referred to as spatial effects. One of the causes of spatial effects resulting from information across regions is the presence of spatial heterogeneity. One way to detect spatial heterogeneity is by conducting the Breusch-Pagan test as follows:

$$BP = \left(\frac{1}{2}\right)\boldsymbol{f}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{f} \tag{7}$$

where $\boldsymbol{f}$ is element vector $(f_i = \frac{\hat{\varepsilon}_i^2}{\sigma^2} - 1)$ and $\boldsymbol{Z}$ is a standardized matrix of size $n \times (k+1)$. If the BP value is less than or equal to $\chi^2_{\alpha;k}$ or p_{value} is greater than α, then the null hypothesis is accepted at the significance level of α. However, if the decision indicates rejection of the null hypothesis, further methods are needed to address spatial heterogeneity issues. One of these methods could be Geographically Weighted Regression (Fotheringham et al., 2002).

Geographically weighted regression (GWR)

Geographically Weighted Regression (GWR) is a method that involves spatial weighting in modeling data. This method extends linear regression models into weighted regressions using Weighted Least Squares (WLS) to estimate regression parameters for each region. The form of the GWR model is as follows(Fotheringham et al., 2002):

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_j(u_i, v_i)x_{ij} + \varepsilon_i \tag{8}$$

with parameter estimators using Weighted Least Squares (WLS):

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) \ = (\boldsymbol{X}'\mathbf{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}(u_i, v_i)\boldsymbol{y} \tag{9}$$

In estimating parameters at a location in the GWR model, spatial weighting is required, as it can represent the proximity of one observation data point to another. The elements of the GWR weighting matrix, W_{iJ}, are determined based on the proximity of the i-th regression point to the J-th. observation point. One known kernel weighting function is the fixed Gaussian kernel weighting function. The fixed Gaussian kernel function is as follows (Fotheringham et al., 2002):

$$w_{iJ} = exp\left[-\left(\frac{d_{iJ}}{b}\right)^2\right] \tag{10}$$

The constant b represents the bandwidth that determines how far the radius from the center location of each observation estimation influences the location i-th. Meanwhile, d_{iJ} indicates the Euclidean distance defined as:

$$d_{iJ} = \sqrt{(u_i - u_J)^2 + (v_i - v_J)^2} \tag{11}$$

The method to identify the optimal bandwidth is by using cross-validation (CV). The optimal bandwidth is the bandwidth that results in the minimum CV. If the predictor valuesy_i are a function of the bandwidth (h) written as $\hat{y}_i(h)$, then the CV method calculates the value of y_i by ignoring the i-th observation. Mathematically, CV is written as:

$$CV(h) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \hat{y}_{\neq i}(h))^2 \tag{12}$$

Principal component geographically weighted regression

After testing for spatial heterogeneity, differences in characteristics between regions are identified, then the GWR method can be applied. If there is multicollinearity issue, then it is necessary to reduce the components of correlated independent variables using the PCA method. Then the modeling process is carried out by regressing the dependent variable with the selected principal components. The resulting model is obtained as follows (Zhu et al., 2020):

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)K_{ik} + \varepsilon_i \tag{13}$$

where p represents the selected principal component from the PCA process.

Hypothesis testing of the model

Goodness-of-fit tests and model adequacy tests are needed to analyze the performance of the obtained model. The model goodness is evaluated using F^*_{test} with degrees of freedom df_1 and df_2 as follows (Caraka & Yasin, 2017):

$$F^*_{test} = \frac{SSE(H_0)/df_1}{SSE(H_1)/df_2} \tag{14}$$

with:

$$SSE(H_0) = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \text{untuk } \boldsymbol{H} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$
$$SSE(H_1) = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{S})'(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{y}$$
$$df_1 = n - p - 1$$
$$df_2 = (n - 2tr(\boldsymbol{S}) + tr(\boldsymbol{S}'\boldsymbol{S}))$$

where the matrix $\boldsymbol{S}$ above is the model projection matrix which represents the value of y to become y hat at the location points (u_i, v_i) and:

$$\boldsymbol{S} = \begin{bmatrix} x_1'(\boldsymbol{X'W}(u_1, v_1)\boldsymbol{X})^{-1}\boldsymbol{X'W}(u_1, v_1) \\ x_2'(\boldsymbol{X'W}(u_2, v_2)\boldsymbol{X})^{-1}\boldsymbol{X'W}(u_2, v_2) \\ \vdots \\ x_n'(\boldsymbol{X'W}(u_n, v_n)\boldsymbol{X})^{-1}\boldsymbol{X'W}(u_n, v_n) \end{bmatrix} \tag{15}$$

As for testing the model parameters, it is done with the criterion $|T_{hitung}| > T_{(\alpha/_2;df_2)}$ which indicates that there are significant parameters $\beta_k(u_i, v_i)$ towards the model. Where:

$$T_{test} = \frac{\hat{\boldsymbol{\beta}}_k(u_i, v_i)}{\hat{\sigma}_w\sqrt{g_{kk}}} \tag{16}$$

with:

$$\hat{\sigma}_w = \sqrt{\frac{SSE(H_0)}{\delta_1}}, \text{ for } \delta_1 = tr(\boldsymbol{S}'\boldsymbol{S})$$

where $\hat{\boldsymbol{\beta}}_k(u_i, v_i)$ is the estimated parameter value at location (u_i, v_i) and g_{kk} is the k-th diagonal element based on the matrix $\boldsymbol{GG}'$ with:

$$\boldsymbol{G} = (\boldsymbol{X}'\boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(u_i, v_i) \tag{17}$$

Selection of the best model

One of the objectives of analysis in modeling is to obtain the best model that explains the relationship between the dependent and independent variables. The best model is the one that performs optimally according to certain criteria. The coefficient of determination (R^2) is one measure used to assess how well the model fits. The coefficient of determination is defined as follows (Montgomery et al., 2012):

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^2}{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2} \tag{18}$$

The coefficient of determination ranges from zero to one, with the criterion that a model is considered better if the coefficient of determination is higher.

Model selection can also be done using AIC by finding the minimum AIC value. The estimation of AIC is determined by (Fotheringham et al., 2002):

$$AIC = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n + tr(\boldsymbol{S}) \tag{19}$$

where $\hat{\sigma} = \sqrt{\frac{SSE}{n}}$ with $\hat{\sigma}$ is the estimated value of residual standard deviation and $\boldsymbol{S}$ is an $n \times n$ projection matrix depicting the value of y to $\hat{y}$ following equation (15).

3. RESULTS

Descriptive statistics

The descriptive statistics of each variable in the study are as follows:

Table 1. Descriptive statistics

Variables	*Min*	*Median*	*Max*	*Mean*	Standard Deviation
Y	7.71	16.03	51.44	19.87	10.59
X_1	41	614.5	6022.8	1551.8	1776.64
X_2	2.929	18.871	46.975	21.695	10.88
X_3	14.21	22.85	44.28	25.12	8.299
X_4	8.29	12.44	47.71	16.56	9.47
X_5	2.430	5.480	9.970	5.874	2.086
X_6	4.560	10.680	17.830	11.393	3.556
X_7	12.76	21.50	87.35	28.81	18.91
X_8	5.41	16.30	46.75	17.60	7.466
X_9	4.86	16.94	26.74	16.49	5.42
X_{10}	6.165	7.665	10.980	8.022	1.278
X_{11}	0	17	82	20.46	17.889

X_{12}	0	36	208	59.23	62.43

Based on Table 1. it provides information that among the 35 Districts/Cities in Central Java Province. the area with the lowest level of higher education participation rate is in Banjarnegara District at 7.71%. and for the highest participation rate in higher education is in Salatiga City at 51.44%. The average Gross Enrollment Rate for Higher Education is 19.87%. indicating that most districts/cities have a Gross Enrollment Rate for Higher Education that is still below the national average.

Multiple linear regression analysis

Here are the results of the regression model using the Ordinary Least Squares (OLS) method:

$$y = -5.507 - 0.0006X_1 + 0.082X_2 + 0.332X_3 + 0.786X_4 + 0.859X_5 - 0.0009X_6 - 0.051X_7 + 0.039X_8 - 0.009X_9 - 0.081X_{10} + 0.014X_{11} - 0.0096X_{12}$$

The F_{test} value obtained using the OLS method yields a model of 15.51 with p_{value} is $4.825e^{-08}$ resulting in a significant model at the 5% level. However. out of the 12 independent variables used in the analysis. only variable X_4 (percentage of households owning a computer/laptop/tablet) has a significant influence on the Gross Enrollment Rate for Higher Education.

Table 2. Partial test

Variables	T_{test}	p_{value}	Variables	T_{test}	p_{value}
X_1	–0.554	0.585	X_7	–0.660	0.516
X_2	0.538	0.596	X_8	0.182	0.857
X_3	0.784	0.462	X_9	–0.021	0.984
X_4	2.731	**0.012**	X_{10}	–0.021	0.984
X_5	1.471	0.155	X_{11}	0.200	0.843
X_6	–0.002	0.999	X_{12}	–0.546	0.591

Multicollinearity

The multicollinearity test was conducted by examining the criteria of the Variance Inflation Factor (VIF) values. VIF values greater than 10 indicate multicollinearity.

Table 3. Multicollinearity

Variables	VIF	Explanation	Variables	VIF	Explanation
X_1	7.353	There is no multicollinearity	X_7	3.952	There is no multicollinearity
X_2	5.162	There is no multicollinearity	X_8	4.750	There is no multicollinearity
X_3	25.45	There is multicollinearity	X_9	10.059	There is multicollinearity
X_4	13.985	There is multicollinearity	X_{10}	46.595	There is multicollinearity
X_5	2.744	There is no multicollinearity	X_{11}	2.861	There is no multicollinearity
X_6	4.624	There is no multicollinearity	X_{12}	2.255	There is no multicollinearity

Based on Table 3. the variables percentage of population completing secondary education (X_3). percentage of households owning a computer/laptop/tablet (X_4). percentage of households receiving the family hope program (X_9). and average years of schooling (X_{10}) indicate multicollinearity. To address this issue. this study applies Principal Component Analysis (PCA) method where it transforms the original variables into new variables while retaining most of the information.

Principal component analysis (PCA)

To determine the number of principal components formed is by looking at eigenvalues greater than one. Since the variables involved in this study use different units. it is recommended to standardize first so that the correlation matrix is used to form eigenvalues. The eigenvalues are obtained as follows:

Table 4. Eigen value

Components	Total Variance	Proportion of Variance	Cumulative Proportion
K_1	2.623	0.591	0.591
K_2	1.153	0.111	0.702
K_3	1.03	0.088	0.79
K_4	0.869	0.063	0.853
K_5	0.722	0.043	0.896
K_6	0.685	0.039	0.935
K_7	0.561	0.026	0.962
K_8	0.415	0.014	0.976
K_9	0.385	0.012	0.988
K_{10}	0.273	0.006	0.994
K_{11}	0.230	0.004	0.998
K_{12}	0.116	0.001	1

When determining the number of principal components formed. it can be examined from eigenvalues greater than 1 and cumulative proportions ranging from 70% to 80%. In this study. 3 variable combinations. namely K_1. K_2. and K_3 meet the requirements for the number of principal components formed. Next is to find the scores of the formed components to create new variables obtained from the following equation:

$$K_1 = 0.291Z_1 + 0.252Z_2 + 0.341Z_3 + 0.35Z_4 + 0.138Z_5 - 0.293Z_6 + 0.313Z_7 + 0.318Z_8 - 0.323Z_9 + 0.348Z_{10} - 0.233Z_{11} - 0.175Z_{12}$$

$$K_2 = 0.127Z_1 + 0.281Z_2 + 0.082Z_3 + 0.046Z_4 - 0.727Z_5 - 0.13Z_6 - 0.126Z_7 - 0.265Z_8 - 0.242Z_9 + 0.096Z_{10} - 0.002Z_{11} + 0.444Z_{12}$$

$$K_3 = -0.521Z_1 - 0.576Z_2 + 0.17Z_3 + 0.001Z_4 + 0.025Z_5 - 0.216Z_6 + 0.089Z_7 + 0.05Z_8 - 0.247Z_9 + 0.094Z_{10} - 0.335Z_{11} + 0.357Z_{12}$$

with $Z_1 . Z_2 . \dots . Z_{12}$ is the standardized original variable.

Spatial analysis

The testing for indications of spatial effects is conducted using the Breusch-Pagan (BP) test with the following results:

Table 5. BP test

BP test	*p-value*
16.497	0.000897

With a significant level of 5%. the decision is to reject H_0. Therefore. it can be concluded that there is a spatial effect in the form of spatial heterogeneity in the observed data. which should be considered to expand the multiple linear regression model to obtain better model performance in predicting the Gross Enrollment Rate for Higher Education in Central Java. Hence. the results obtained from the Breusch-Pagan test serve as the basis for modeling spatial influence data. namely Geographically Weighted Regression (GWR).

Principal component geographically weighted regression

Modeling the Gross Enrollment Rate for Higher Education in Central Java using Geographically Weighted Regression (GWR) with the results from Principal Component Analysis (PCA) components will yield 35 different local models for each district/city. This is due to the differences in environmental and geographical characteristics among the districts/cities in Central Java. Parameter estimation is conducted using the Weighted Least Square (WLS) method. The weighting matrix obtained at each location is substituted to obtain parameter estimates so that each observation location has different parameter estimate values. A summary of the estimation results is presented as follows:

Table 6. Model results

Variables	*Minimum*	*Maximum*	*Mean*
Intercept	13.286	22.221	18.606
K_1	1.424	5.896	3.474
K_2	-2.605	4.763	0.699
K_3	-4.585	5.230	0.499

The next step is to conduct a test of goodness of fit. This aims to determine whether the analyzed data is suitable for modeling. The test results obtained a value of $F^*_{test} = 0.187$ with a $\rho_{value} = 0.0055$. It can be concluded that the data used is suitable for modeling using GWR. Meanwhile. the test of model parameters aims to determine the explanatory variables that have a significant impact on each district/city. If $|T_{test}| > T_{0.05/_{2};31} = 2.04$ or $\rho_{value} < 0.05$. then the null hypothesis is rejected. indicating that the component affects the dependent variable. In summary. partial testing for each district/city yields the following:

Table 7. Partial test of the model predictor

Districts/Cities	Explanation
Banjarnegara. Banyumas. Batang. Blora. Boyolali. Brebes. Cilacap. Demak. Grobogan. Karanganyar. Kebumen. Kendal. Klaten. Kota Magelang. Kota Pekalongan. Kota Semarang. Kota Tegal. Kudus. Kab. Magelang. Pati. Kab. Pekalongan. Pemalang. Purbalingga. Purworejo. Rembang. Kota Salatiga. Kab. Semarang. Sragen. Sukoharjo. Kota Surakarta. Kab. Tegal. Temanggung. Wonogiri. and Wonosobo	Significant against component K_1
Jepara dan Rembang	There is no significant

Based on the results in Table 7. only component K_1 is significant in forming the Gross Enrollment Rate for Higher Education model. Meanwhile. neither component K_2 nor component K_3 provides significance to the Gross Enrollment Rate for Higher Education in any district/city in Central Java. The following shows the results of the formation of local models obtained from the GWR results for each district/city:

Table 8. Local Models

No.	Districts/Cities	Local Models
1	Banjarnegara District	$\hat{y} = 15.724 + 2.648K_1 - 0.283K_2 - 0.279K_3$
2	Banyumas District	$\hat{y} = 19.792 + 3.397K_1 + 0.605K_2 + 1.526K_3$
3	Batang District	$\hat{y} = 17.957 + 2.436K_1 + 1.276K_2 - 0.262K_3$
4	Blora District	$\hat{y} = 17.926 + 2.576K_1 + 0.975K_2 - 1.612K_3$
5	Boyolali District	$\hat{y} = 21.69 + 4.477K_1 - 0.203K_2 - 0.041K_3$
6	Brebes District	$\hat{y} = 17.448 + 3.708K_1 + 2.957K_2 + 0.984K_3$
7	Cilacap District	$\hat{y} = 10.943 + 3.278K_1 + 2.572K_2 + 5.230K_3$
8	Demak District	$\hat{y} = 21.486 + 3.274K_1 + 1.546K_2 - 1.847K_3$
9	Grobogan District	$\hat{y} = 15.724 + 2.648K_1 + 1.376K_2 - 2.797K_3$
10	Jepara District	$\hat{y} = 22.392 + 1.424K_1 + 4.763K_2 - 0.65K_3$
11	Karanganyar District	$\hat{y} = 19.811 + 5.896K_1 - 0.301K_2 - 4.449K_3$
12	Kebumen District	$\hat{y} = 21.419 + 3.077K_1 - 0.893K_2 - 0.308K_3$
13	Kendal District	$\hat{y} = 17.021 + 3.636K_1 - 0.597K_2 + 0.037K_3$
14	Klaten District	$\hat{y} = 23.233 + 4.381K_1 + 1.732K_2 + 0.36K_3$
15	Magelang City	$\hat{y} = 15.812 + 3.995K_1K_1 - 2.485K_2 + 0.786K_3$
16	Pekalongan City	$\hat{y} = 14.827 + 1.885K_1 + 1.753K_2 - 0.556K_3$
17	Semarang City	$\hat{y} = 18.592 + 3.945K_1 - 1.794K_2 - 0.134K_3$
18	Tegal City	$\hat{y} = 19.166 + 3.031K_1 + 2.371K_2 - 2.120K_3$
19	Kudus District	$\hat{y} = 17.22 + 1.972K_1 + 4.334K_2 - 1.372K_3$
20	Magelang District	$\hat{y} = 21.857 + 4.027K_1 - 2.605K_2 + 0.826K_3$
21	Pati District	$\hat{y} = 17.918 + 1.724K_1 + 3.411K_2 - 1.208K_3$
22	Pekalongan District	$\hat{y} = 18.554 + 2.057K_1 + 1.484K_2 - 0.342K_3$
23	Pemalang District	$\hat{y} = 17.041 + 2.706K_1 + 2.679K_2 - 0.873K_3$
24	Purbalingga District	$\hat{y} = 17.472 + 3.122K_1 + 1.191K_2 + 0.145K_3$
25	Purworejo District	$\hat{y} = 19.294 + 3.188K_1 - 1.005K_2 - 0.016K_3$
26	Rembang District	$\hat{y} = 13.455 + 1.521K_1 - 0.732K_2 - 0.372K_3$
27	Salatiga City	$\hat{y} = 23.029 + 4.271K_1 - 2.271K_2 + 0.617K_3$
28	Semarang District	$\hat{y} = 14.727 + 4.250K_1 - 2.536K_2 + 0.651K_3$
29	Sragen District	$\hat{y} = 12.007 + 5.775K_1 - 0.877K_2 - 4.585K_3$
30	Sukoharjo District	$\hat{y} = 20.593 + 5.112K_1 + 2.169K_2 - 1.01K_3$
31	Surakarta City	$\hat{y} = 22.276 + 5.138K_1 + 1.354K_2 - 1.461K_3$
32	Tegal District	$\hat{y} = 19.312 + 3.252K_1 + 2.658K_2 - 0.446K_3$
33	Temanggung District	$\hat{y} = 17.141 + 3.772K_1 - 1.246K_2 + 0.430K_3$
34	Wonogiri District	$\hat{y} = 20.523 + 5.385K_1 + 1.802K_2 - 2.221K_3$
35	Wonosobo District	$\hat{y} = 18.929 + 3.211K_1 - 0.711K_2 - 0.110K_3$

The final step in the modeling process is to obtain the Gross Enrollment Rate for Higher Education model implicitly. so, the model results formed in Table 8 need to be transformed back to the original variables. For example. to obtain the Gross Enrollment Rate for Higher Education model in Banjarnegara District where:

$$\hat{y}_{Banjarnegara} = 15.724 + 2.648K_1 - 0.283K_2 - 0.279K_3$$

since components K_1. K_2. dan K_3 consist of 12 standardized variables. then:

$$\hat{y}_{Banjarnegara} = 15.724 + 2.648(0.291Z_1 + 0.252Z_2 + \dots - 0.175Z_{12}) - 0.283(0.127Z_1 + 0.281Z_2 + \dots + 0.444Z_{12}) - 0.279(-0.521Z_1 + 0.576Z_2 + \dots + 0.357Z_{12})$$

where $Z_1.Z_2.\dots.Z_{12}$ represent the standardization process of the original variables. Finally. to obtain the complete implicit model in Banjarnegara District as follows:

$$\begin{aligned}\hat{y}_{Banjarnegara} = 15.724 + 2.648\left(0.291\left(\frac{x_1 - 1551.8}{1776.64}\right) + 0.252\left(\frac{x_2 - 21.695}{10.88}\right)\dots -0.175\left(\frac{x_{12} - 59.23}{62.43}\right)\right) \\ - 0.283\left(0.127\left(\frac{x_1 - 1551.8}{1776.64}\right) + 0.281\left(\frac{x_2 - 21.695}{10.88}\right)\dots +0.444\left(\frac{x_{12} - 59.23}{62.43}\right)\right) \\ - 0.279\left(0.291\left(\frac{x_1 - 1551.8}{1776.64}\right) + 0.252\left(\frac{x_2 - 21.695}{10.88}\right)\dots +0.357\left(\frac{x_{12} - 59.23}{62.43}\right)\right)\end{aligned}$$

thus, obtaining the Gross Enrollment Rate for Higher Education model from the original variables as follows:

$$\begin{aligned}\hat{y}_{Banjarnegara} = 4.5123 + 0.0005x_1 + 0.0688x_2 + 0.1002x_3 + 0.0965x_4 + 0.2705x_5 - 0.1909x_6 \\ + 0.0444x_7 + 0.1210x_8 - 0.1325x_9 + 0.6793x_{10} - 0.0292x_{11} - 0.0110x_{12}\end{aligned}$$

To obtain the Gross Enrollment Rate model for Higher Education in other districts/cities. the same process as above is carried out.

Selection of best model

The selection of the best model is an evaluation process to determine how well each formed model fits to model the Gross Enrollment Rate for Higher Education data in Central Java Province. The best model is chosen by comparing the criteria of the highest coefficient of determination (R^2) value and the minimum Akaike's Information Criterion (AIC).

Table 9. Selection of best model

Gross Enrollment Rate for Higher Education Models	R^2	AIC
Multiple Linear Regression (12 variables)	0.894	212.397
GWR Model without PCA (12 variables)	0.898	196.58
Model with PC-GWR approach (3 components)	0.962	169.43

Based on Table 9. modeling the Gross Enrollment Rate for Higher Education in Central Java using the PC-GWR (Principal Component Geographically Weighted Regression) approach improves the performance of multiple linear regression models. as indicated by the coefficient of determination (R^2) of 0.962 and the increased accuracy of prediction results. shown by the decrease in AIC value by 169.43. This model explains 96.2% of the variability in the Gross Enrollment Rate for Higher Education in Central Java. while the remaining 3.8% is explained by other variables outside the model. Modeling results using PCA with GWR to address multicollinearity and spatial heterogeneity issues represent the best model for modeling the Gross Enrollment Rate for Higher Education in Central Java.

4. DISCUSSIONS

This study successfully addresses the stated research problem by developing the best model to predict the Gross Enrollment Rate for Higher Education in Central Java Province using the Principal Component Geographically Weighted Regression (PC-GWR) approach. The resulting PC-GWR model shows a significant performance improvement compared to traditional multiple linear regression models and GWR models without PCA. With a coefficient of determination (R^2) of 0.962. this model explains 96.2% of the variability in the Gross Enrollment Rate for Higher Education. while the remaining 3.8% is explained by other variables outside the model.

The analysis results indicate that the first principal component (K_1) significantly influences the Gross Enrollment Rate for Higher Education in most regencies/cities. whereas the other components do not show consistent influence. Therefore. this study concludes that the PCA method applied to the GWR model is an effective solution to address multicollinearity and spatial heterogeneity issues in the data.

The implications of these findings are that local governments and education stakeholders in Central Java can use the PC-GWR model to more accurately and specifically identify factors influencing Gross Enrollment Rate for Higher Education in each region. Additionally. this method can be applied to research in other fields involving spatial data to enhance prediction accuracy and understanding of influential variables.

5. CONCLUSION

The conclusion drawn from the research results is that modeling Gross Enrollment Rate for Higher Education in Central Java using Geographically Weighted Regression (GWR) after conducting Principal Component Analysis (PCA) resulted in 35 local models. with Jepara Regency and Rembang Regency not producing any significant variables at all. Component K_1 significantly influences Gross Enrollment Rate for Higher Education. while other components in regencies/cities do not show influence from the variables under study. This research utilized Principal Component Analysis (PCA) to address multicollinearity issues in Gross Enrollment Rate for Higher Education data. There are alternative methods such as Geographically Weighted LASSO (GWL) or Geographically Weighted Ridge Regression (GWRR) which can identify factors influencing Gross Enrollment Rate for Higher Education directly without reducing the diversity of data information.

6. REFERENCES

[1] H. Al Azies. *Analisis Pengaruh Pengendalian Pencemaran dan Kerusakan Lingkungan Terhadap Pertumbuhan Ekonomi di Indonesia Menggunakan Pendekatan Geographically Weighted Regression Principal Component Analysis (GWRPCA)*. Jember: Prosiding Seminar Nasional Energi 8. 2019.

[2] L. Anselin. *Spatial Econometrics: Methods and Models*. Dordrecht:Kluwer Academic Publishers. 1988.

[3] R. E. Caraka. and H. Yasin. *Geographically Weighted Regression; Sebuah Pendekatan Regresi Geografis*. Yogyakarta: Mobius. 2017.

[4] A. Fotheringham. C. Brunsdon. and M. Charlton. *Geographically Weighted Regression; the analysis of spatially varying relationships*. UK: John Wiley & Sons. 2002.

[5] L. Guo. et. al. *Spatial Modelling of Soil Organic Carbon Stocks with Combined Principal Component Analysis and Geographically Weighted Regression*. The Journal of Agricultural Science. 2018.

[6] D. N. Gujarati. *Basic Econometrics (4th edition)*. USA: McGraw-Hill. 2004.

[7] S. Habibah. Y. P. Putra. and Y. M. Putra. *Faktor-Faktor yang Mempengaruhi Angka Partisipasi Perguruan Tinggi pada 32 Provinsi di Indonesia Tahun 2013-2016*. Jakarta: Jurnal Anggaran dan Keuangan Negara Indonesia. 2019.

[8] P. Harris. C. Brunsdon. and M. Charlton. *Geographically Weighted Principal Component Analysis*. UK: International Journal of Geographical Information Science. 2011.

[9] Humas Jateng. "Pendidikan Kurang Maju Bukan Karena Guru". *Jateng Prov*. 2017 [Online]. Available: https://jatengprov.go.id/publik/pendidikan-kurang-maju-bukan-karena-guru/ [Accessed: 18 March 2023].

[10] W. Jinglei. et. al. *Estimation of Crop Water Requirement Based on Principal Component Analysis and Geographically Weighted Regression*. China: Chinese Science Bulletin. 2013.

[11] R. A. Johnson. and D. W. Wichern. *Applied Multivariate Statistical Analysis (6th edition)*. USA: Prentice Hall. 2007.

[12] LLDIKTI13 KEMDIKBUD. "20 Universitas Terbaik di Indonesia Versi Webometrics Rank 2021". *LLDIKTI13 KEMDIKBUD*. 2021 [Online]. Available: https://lldikti13.kemdikbud.go.id/2021/07/28/20-universitas-terbaik-di-indonesia-versi-webometrics-rank-2021/ [Accessed: 22 August 2023].

[13] Maskar. et. al. *Peningkatan Pemahaman Pentingnya Lanjut Studi ke Perguruan Tinggi bagi Masyarakat Desa Hanura-Pesawaran. Provinsi Lampung*. Community Development Journal. 2022.

[14] D. C. Montgomery. E. A. Peck. and G. G. Vining. *Introduction to Linear Regression Analysis (5th Edition)*. Canada: John Wiley & Sons. 2012.

[15] A. Mutia, "Belum Capai Target. Angka Partisipasi Pendidikan Tinggi di RI 2021 Masih Rendah". *Databoks Katadata*. 2022 [Online]. Available: *https://databoks.katadata.co.id/datapublish/2022/09/29/belum-capai-target-angka-partisipasi-pendidikan-tinggi-di-ri-2021-masih-rendah* [Accessed: 18 March 2023].

[16] S. A. Prabaswara. *Analisis Clustering Kerentanan Sosial terhadap Bencana Alam Menggunakan Geographically Weighted Principal Components Analysis dan Fuzzy Geographically Weighted Clustering (Studi kasus di Seluruh Kabupaten/Kota di Indonesia tahun 2019)*. Jakarta. 2022.

[17] N. Sari. H. Yasin. and A. Prahutama. *Geographically Weighted Regression Principal Component Analysis (GWRPCA) pada Pemodelan Pendapatan Asli Daerah di Jawa Tengah*. Semarang: Jurnal Gaussian. 2016.

[18] Suyono. *Analisis Regresi untuk Penelitian*. Yogyakarta: DEEPUBLISH. 2015.

[19] R. Zhao. et. al. *A Geographically Weighted Regression Model Augmented by Geodetector Analysis and Principal Component Analysis for the Spatial Distribution of PM (2.5).* Elsevier. 2020.

[20] Z. Zhu. et. al. *Socio-Economic Impact Mechanism of Ecosystem Service Value*. *A PCA-GWR Approach*. PJOES. 2021.