Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 24 NO 03 2023, pp 436-452
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol24-iss03/378

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Water Quality Model in Jambi Province using Geographically Weighted Regression

Article Info

Article history :

Received January 14, 2023
Revised August 03, 2023
Accepted August 15,2023
Published September 30, 2023

Keywords :
PDAM, water quality, geographically weighted regression (GWR), BOD, TSD

Sherli Yuinanda[1], Cut Multahadah[1*], Hasnaul Marisa[2], Mohammad Abdullah[3]

[1]Department of Mathematics, Faculty of Science and Technology (FST), Universitas Jambi, Jambi, Indonesia
[2]Department of Biology, Faculty of Science and Technology (FST), Universitas Jambi, Jambi, Indonesia
[3]Chemical Engineering Studies, College of Engineering, Universiti Teknologi MARA Johor Branch, Pasir Gudang, Johor Bahru, Johor, Malaysia

Abstract. PDAM (Regional Water Supply Company) functions to serve the needs of many people's lives by providing quality water for the community. Based on Permenkes no. 492/menkes/per/iv/2010 clean water quality parameters, namely the feasibility of water used in daily life related to physical, chemical and microbiological parameters including *Biological Oxygen Demand* (BOD), *Total Dissolved Solid* (TDS), *Cloride* (Cl), and Nitrates (NO_3). The aim of this research is to test the water quality of PDAMs in Jambi Province based on minimum quality standards for clean water Using Geographically Weighted Regression (GWR) with the assumption. The method can be used to model the relationship between the dependent variable and the independent variable has regional influence. The results showed that the BOD values of all regions in Jambi province met except for Merangin, which was 2.1 mg/l with a threshold value of 2.0 mg/l. For other parameters, namely TDS, Cl and NO_3, they meet the threshold values. Based on the results of the GWR model, the coefficient of R^2 is 0.669 , this means that there is a relationship between TDS, Cl and NO_3 to BOD and is positive.

Corresponding Author :
Sherli Yuinanda
Department of Mathematics, Faculty of Science and Technology (FST),
Universitas Jambi, Jambi, Indonesia
Email : sherliyurinanda@unja.ac.id

1. Introduction

Water is very important for life. Water used in daily life plays a role in determining the degree of public health [1]. This was confirmed by [2] who stated that 40% of the degree of public health is influenced by the physical environment, namely the availability of clean water. Besides being useful for meeting the needs of many people, water can also act as a medium for transmitting diseases such as cholera, dysentery and typhus. To increase the availability of clean water, the government through BUMD (Regional Owned Enterprises) has the mandate to provide supplies of goods to meet the needs of the people through PDAM (Regional Water Supply Companies). PDAM according to [3] functions to serve the needs of the livelihood of many people by providing quality water for the community [4].

Efforts to provide quality water require an increase in clean water services and environmental sanitation. According to [5] increasing population results in an increase in the number of activities and facilities thereby increasing the need for clean water as a support. Clean water quality is the quality of water that can be used by the community in carrying out their daily activities. Based on Permenkes no. 492/menkes/per/iv/2010 clean water quality parameters related to physical, chemical and microbiological parameters [6], including Biological Oxygen Demand (BOD) Parameters, Total Dissolved Solid (TDS), Chloride (Cl), and Nitrates (NO_3). Through this water quality feasibility test, it is possible to detect the dominant content of the water supplied by the PDAM to homes that are used to meet people's daily needs.

From a mathematical perspective, PDAM water quality in Jambi Province is the dependent variable and the physical, chemical and microbiological parameters are called variable X. [7] states that a method that can model the relationship between 2 dependent variables and independent variables is the regression model. If the conditions in the field are spatial data then multiple regression alone is not appropriate.

Observation of water in the PDAM area really needs to be done because it can detect if there are dominant parameters that interfere with the quality of the water that flows to homes which are the people's daily consumption can be overcome [8]. And the observation of water in a PDAM area affects other areas in Jambi Province, so the data has regional influence. The statistical method that can be used to handle spatial data is Geographically Weighted Regression (GWR) [9][10]. The GWR model uses estimates by providing different weights for each location where the data is collected. Weight is very important in this model because the value of the weight represents the location of the observation data [11]. According to stated that to see PDAM water quality the GWR model is better to use when compared to multiple linear regression analysis so that in this study GWR analysis was used. Stated that to see PDAM water quality the GWR model is better to use when compared to multiple linear regression analysis so that in this study GWR analysis was used [12]. Based on the background, this study analyzes the PDAM Water Quality Model in Jambi Province Using Geographically Weighted Regression (GWR).

2. Literature Reviews

2.1 Water

Water has an important role in fulfilling human needs. Water besides being useful for humans is also capable of being a means of disease transmission [13][14]. For this reason, it is necessary to distribute quality water in the community. According to [6] quality water is water that meets the feasibility of use based on physical, chemical and microbiological parameters.

To fulfill this, the government established a clean water distribution platform, namely PDAM (Regional Water Supply Company). In general, the water flowing in PDAM pipes in Jambi Province is sourced from Batanghari River water. The quality of the Batanghari river water has decreased. To see the quality of water, the government issued government regulations containing thresholds for each physical, chemical and microbiological parameter. Based on the government regulation of the Republic of Indonesia No. 20 of 1990 concerning the control of water pollution The threshold value

(maximum or minimum amount) that may be in the waters can be seen in the eligibility criteria for water use divided into parameters *Biological Oxygen Demand* (BOD), *Total Dissolved Solid* (TDS), *Chloride* (Cl), and Nitrates (NO_3) [15].

2.2 Regression Analysis

According to [16] linear regression is the relationship between two variables where one variable is considered to affect the other variable. Linear regression can also be used to see the effect of the dependent variable on the independent variable. If there is more than one independent variable, then the prediction of the relationship between the variables uses multiple linear regression. The general form of the equation is as follows [17]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_n + \varepsilon$$

Description:

y : dependent variable

$x_1, x_2, , x_n$ Independent variable

β_0 : Constant

$\beta_1, \beta_2, \beta_n$: Regression Coefficient

ε : Residual Value

To model a problem using regression, according to [18] the assumptions that must be met include:

1. The heteroscedasticity test was carried out to test whether in the regression model there is an inequality of variance from the residual one observation to another [19] .
2. Non-autocorrelation means that there is no influence of the variables in the model through time intervals or there is no correlation between the random errors. The test that is often used is the run test with the following conditions:

   a. If the sig value is greater than 0.05, it can be concluded that the residual is random.

   b. If the sig value is less than 0.05, it can be concluded that the residuals are not random.

3. Non-Multicollinearity means that between one independent variable and another independent variable in the regression model, there is no near-perfect relationship or perfect relationship. Multicollinearity can be seen from the value of the variance inflation factor (VIF) with testing criteria if the VIF value < 10 then there is no Multicollinearity between the independent variables on the other hand if the VIF value is > 10 then there is multicollinearity between the independent variables.
4. The distribution of errors (error) is normal. To see this, Ujikolmogorov Smirnov is used with the following provisions[20]:
   a. If the sig value is greater than 5%, it can be concluded that the residuals are normally distributed.
   b. If the sig value is less than 5%, it can be concluded that the residuals are not normally distributed.
5. The average value of the population error in the stochastic model is zero.
6. The independent variable has a constant value for each experiment that is repeated (nonstochastic variable).

   Furthermore, the significance test of the regression model was carried out

1. The coefficient of determination R^2

   The formula used to find the coefficient of determination is:

$$R^2 = \frac{ESS}{TSS}$$

Information :
R^2 = coefficient of determination
ESS= explained sum of square (explained variation Y)
TSS= sum of square (total Y variation)
Values $R^2$are in the range 0-1 where the closer to 1, the better the regression, namely the model is able to explain the actual data [21].

2. Test F
The F test aims to determine the effect of the dependent and independent variables simultaneously [18][22]. TestF formula as follows:

$$F = \frac{R^2(n-k-1)}{k(1-R^2)}$$

Information :
R^2 : coefficient of determination
n : number of subjects
k : number of independent variables
the statistical test process :
a. statistical formulation
H_0: There is no relationship between X1, X2… Xn with Y
H_1: There is a relationship between X1, X2… Xn with Y
b. real level (α)and Ftable
The real rates used are usually 5 % or1 %
Table (db)values have degrees of freedom F,
$V1 = m - 1;\ V2 = n - m\ F\alpha; (V1)(V2)$
c. testing criteria
H_0: accepted (H_1rejected) if$F0 \leq F\alpha; (V1)(V2)$
H_0: rejected (H_1accepted) if$F0 > F\alpha; (V1)(V2)$
3. Test T
The test T is used to determine the effect of each independent variable partially with the following procedure [23] :
a. hypothesis formulation
H_0: $B_i = B_0$ (no positive relationship between X_i and Y)
H_1 : $B_i \neq B_0$ (there is a positive relationship between X_i and Y)
H_0: $B_i < B_0$ (no positive relationship between X_i and Y)
H_1: $B_i > B_0$ (there is a positive relationship between X_i and Y)
Determine the real level (α) and ttable
The level used is usually 5% or 1%
values thave degrees of freedom(db) = n – 2
$t_\alpha; n-2 or t_{\alpha/2}; n-2.$
b. testing criteria
H_0: there is no positive relationship between X_i and Y
H_1: there is a positive relationship between X_i and Y
H_0accepted (H_1: rejected) if$t_0 \leq t_\alpha$
H_0rejected (H_1: accepted) if$t_0 > t_\alpha$
H_0: there is no negative relationship between X_i and Y
H_1: there is a negative relationship between X_i and Y

H_0accepted (H_1: rejected) when$t_0 \leq -t_\alpha$
H_0rejected (H_1: accepted) when$t_0 > -t_\alpha$
H_0: there is no negative relationship between X_i and Y
H_1: there is a negative relationship between X_i and Y
H_0accepted (H_1: rejected) when$-t_{\alpha/2} \leq t_0 \leq t_{\alpha/2}$
H_0rejected (H_1: accepted) when $t_0 > t_\alpha$or$t_0 \leq -t_{\alpha/2}$

c. value Test statistic (value t_0)

$$t_0 = \frac{b_1 - B_1}{S_{b1}}, i = 1,2,3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4)$$

d. Draw conclusions [24]

2.3 Geographically Weighted Regression (GWR)

GWR is a regression model developed where each parameter is calculated at each observation location, so that the regression parameters will be different at each observation location [25]. In running this model, the assumptions that must be met are other than normally distributed data, zero mean and variance σ^2 [26].

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

With

y_i : Observation value of the dependent variable at the i-th observation location
x_{ik} : The value of the k-th independent variable observation at the i-th observation location
$\beta_0(u_i, v_i)$: Constants/ *intercept* on the i-th observation
(u_i, v_i) : State the coordinates of the geographic location of the i-th observation location
$\beta_k(u_i, v_i)$: Regression coefficient of the k-th independent variable at the i-th observation location
ε_i : Error of the i-observation.

2.4 GWR Model Parameter Estimation

Parameter estimation in the GWR Model is the WLS (*Weighted Least Square) method* , namely by giving different weights to each location where the data is taken [25][27]. Suppose the weight for each location $(u_i, v_i)$is $w_i(u_i, v_i)$i = 1,2,...,n, then the parameter of the observation location is (u_i, v_i)estimated by minimizing the *sum square residual* from equation (2.8)

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

Or in the SSR matrix is

$$\begin{aligned}
\varepsilon^T W_I \varepsilon &= (y - X\beta_I)^T W_I (y - X\beta_I) \\
&= (y^T - \beta_I^T X^T) W_I (y - X\beta_I) \\
&= y^T W_I y - W_I y^T X\beta_I - \beta_I^T X^T W_I y + \beta_I^T X^T W_I X\beta_I \\
&= y^T W_I y - W_I (y^T X\beta_I)^T - \beta_I^T X^T W_I y + \beta_I^T X^T W_I X\beta_I \\
&= y^T W_I y - \beta_I^T X^T W_I y - \beta_I^T X^T W_I y + \beta_I^T X^T W_I X\beta_I \\
&= y^T W_I y - 2\beta_I^T X^T W_I y + \beta_I^T X^T W_I X\beta_I
\end{aligned}$$

With

$$\beta_I = \begin{pmatrix} \beta_0(U_i, v_i) \\ \beta_1(U_i, v_i) \\ . \\ . \\ \beta_p(U_i, v_i) \end{pmatrix} \text{ and } \quad W_I = diag(w_1(u_i, v_i), w_2(u_i, v_i) \ldots, w_n(u_i, v_i)$$

in obtaining parameter estimators $\beta(u_i, v_i)$by reducing equation (2.10) to $\beta^T(u_i, v_i)$the following:

$$\frac{\partial \varepsilon^T W_I \varepsilon}{\partial \beta^T} = \frac{\partial (y^T W_I y - 2\beta_I^T X^T W_I y + \beta_I^T X^T W_I X \beta_I)}{\partial \beta^T}$$

$$\mathbf{0} = 0 - 2X^T W_I y + X^T W_I X \beta_I + W_I (X^T \beta^T X)^T$$

$$0 = -2X^T W_I y + X^T W_I X \beta_I + X^T W_I X \beta_I$$

$$0 = -2X^T W_I y + 2X^T W_I X \beta_I$$

$$2X^T W_I y = 2X^T W_I X \beta_I$$

$$X^T W_I y = X^T W_I X \beta_I$$

$$\beta_I = (X^T W_I X)^{-1} X^T W_I y$$

GWR model parameter *estimator is* obtained

$$:\beta(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i)\, y$$

2.5 Bandwidth Model GWR

According to [28][29] The role of weights in GWR represents the location of the observation data from one another. The method that can be used in weighting is the Kernel Function. In forming a weighting matrix, the diagonal elements of the matrix are filled with calculated values from the Kernel *Fixed Bandwidth* or Kernel *Adaptive Bandwidth* for each location and the other elements are filled with zero values. As for using the Kernel Function, it can be done with 2 types of calculations, namely

1. Kernel *Fixed Exponential*

Fixed Kernel for each observation location has the same *bandwidth value* [30][31]. The *Fixed Kernel Function* formula at all locations is obtained from the *Exponential* Kernel Function weights as follows:

$$W_f(u_i, v_i) = exp(\frac{-d_{ij}}{h})$$

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

$u_{i=}$ *lattitude* coordinates (latitude) at the i-th location

v_i= *longitude coordinates* (longitude) at the i-th location

h= bandwidth at all locations

2. Kernel *Adaptive* Bandwidth

Adaptive Kernel for each observation point has a different *bandwidth value* [30]. This is because the *Adaptive Kernel function* can be adjusted to the conditions of the observation point [32]. The Exponential Kernel Function with the following formula:

$$W_a(u_i, v_i) = exp(\frac{-d_{ij}}{h_i})$$

With h $_i$ = *bandwidth* at the i-th location.

2.6 Cross Validation (CV)

According to [25] One method that can be used to select the *optimal bandwidth* is to use *Cross Validation* (CV) which can be written systematically as follows:

$$\text{CV} = \sum_{I=1}^{n} (y_i - y_{\neq i}(h))^2$$

With $y_{\neq i}(h)$the estimator value $y_i$at the observation location (u_i, v_i)removed from the estimation process to get the optimum bandwidth value, it is obtained from h which produces the minimum CV.

2.7 Model Fitment Test (*Goodness of fit)*

In this case the parameter test performed is the similarity test between the Multiple Linear Regression Model and the GWR Model [21][33]. According to, this test was carried out using the following hypothesis:

H_0 : $\beta_k(u_i, v_i) = \beta_k$for each$k = 1,2, \dots, p \; dan \; i = 1,2, \dots p$

(no significant difference between Multiple Regression Model and GWR)

H_1: there is at least one $\beta_k(u_i, v_i) \neq \beta_kfork = 1,2, \dots, p \; dan \; i = 1,2, \dots p$
(there is a significant difference between the Multiple Regression Model and GWR)

Test Statistics :

$$F_1 = \frac{SSE(H_1)/df_1}{SSE(H_0)/df_2}$$

With :

$SSE\ (H_{0)} = y^T(I - H)y$ where$H = X(X^TX)^{-1}X^T$

$$SSE\ (H_{1)} = y^T(I - L)^T(I - L)y$$

$df_1 = \frac{\partial_1^2}{\partial_2}$, where$\partial_i = tr\ ([(I - L)^T(I - L)]^i), i = 1,2$

$$df_2 = n - p - 1$$

Information

I : sized identity matrixnxn

L : is the projection matrix of the GWR Model.

2.8 AIC (Akaike's Information Criterion)

According to [25][34] one that can determine the best model is the smallest *AIC* (*Akaike's Information Criterion) value* with the following formula:

$$AIC = 2n \log(\sigma) + nlog(2\pi) + n + tr(L)$$

Description:

σ : Estimator standard deviation value of the maximum estimation error likelihood.

L : The projection matrix where$\hat{y} = Ly$.

3. Method

3.1 Data Source

The data used in this study is primary data obtained by researchers by taking water samples from each PDAM in Jambi Province. Jambi Province has regencies and cities including Batanghari, Bungo, Kerinci, Merangin, Muaro Jambi, Sarolangun, West Tanjung Jabung, East Tanjung Jabung, Tebo, Jambi City and Sungai Full City.

3.2 Research Variable

Variables in this study must have value and can be measured. The variables in this study are parameters that indicate the feasibility of PDAM water that can be used to meet daily needs, including:

Table 1 . Research variables

Variable	Variable name
y	PDAM Water Quality (Biological Oxygen Demand (BOD))
x_1	Total Disolved Solids (TDS)
x_2	Cl (Clear)
x_3	NO_3 (Nitrate)
u_i	Latitude coordinates
v_i	Longitude coordinates

3.3 Research Materials

This research material is in the form of tools and materials as follows:

1. Materials needed for this research activity include
   a. Vision and Mission, as well as the profile of the Faculty of Science and Technology, University of Jambi.
   b. Parameters that show the quality of water that is suitable for use by the community to carry out their daily activities. Furthermore, these parameters are arranged into *dependent* and *independent variables*. To test the water quality, water samples were taken from each PDAM in Jambi Province.
2. Research tool is a tool used in conducting research. The devices used in this study are:
   a. Computer equipment to build, collect and process the data obtained.
   b. Water measuring device namely Digital WA-2017SD (Lutron)
   c. Flashdisk
   d. Printers
   e. *SPSS software*
   f. *Software R Studio*

4. Results and Discussion

Data collection was obtained directly by taking water samples from each PDAM in Jambi Province, namely the districts of Batanghari, Bungo, Kerinci, Merangin, Muaro Jambi, Sarolangun, Tanjung Jabung Barat, Tanjung Jabung Timur, Tebo, Jambi City and Sungai Full City. The data obtained was then tested for parameters at the UPTD Environmental Laboratory of the Provincial Government of Jambi. The parameters tested in this study were BOD, TDS, Cl and NO_3.The parameter test results that affect water quality are as follows

Table 2. Parameter test results

Regency/City	BOD5 (mg/L)	TDS (mg/L)	Cl (mg/L)	NO_3 (mg/L)
Threshold	$\leq$ 2mg/l	$\leq 500mg/l$	$\leq 250mg/l$	$\leq 50mg/l$
Winking	2.1	23	2.46	0.11
Muaro Jambi	1.3	49	0.27	0.24
Jambi City	1.5	48	0.27	0.18
Sarolangun	1.7	42	0.27	0.13
Full River	1.3	53	0.27	0.11
Tebo	1.7	58	0.27	0.08
Batanghari	1.7	73	0.27	0.18
Kerinci	1.3	38	0.14	0.15
Bungo	1.7	53	0.27	0.12
East Cape	1.7	84	0.27	0.12
West Cape	1.5	87	0.27	0.17

Table 2 can explain that the parameter values of BOD5, TDS, CL, NO_3 at each observation location produced different parameter test results. The BOD values for all regions in Jambi province met except Merangin, which was 2.1 mg/l. This analysis was carried out at 11 observation points, namely in districts/cities in Jambi Province. *Geographically Weighted Regression (GWR)* analysis is an influence analysis method that is determined by geographical location or observation location so that in this study using latitude and longitude as weights in the GWR analysis at 11 observation points. Latitude and longitude at each observation location researchers get on google maps. The following is the latitude and longitude in each Regency/City in Jambi Province

Table 3. Latitude and longitude of districts/cities in Jambi Province

Regency/City	Latitude	Longitude
Winking	-2.54998	102.79182
Muaro Jambi	-1.55214	103.82163
Jambi City	-1.60997	103.60725
Sarolangun	-2.32304	102.71351
Full River	-2.06894	101.41688
Tebo	-1.25930	102.34639
Batanghari	-1.70839	103.08179
Kerinci	-1.87205	101.43392
Bungo	-1.64013	101.88917
East Cape	-1.10244	103.82163
West Cape	-1.10585	103.08179

Table 3 can explain that in each district/city in Jambi Province the location of latitude and longitude also varies.

4.1 Classic Assumption Test

The classic assumption test required in regression is the Normality Test, Multicollinearity Test and Autocorrelation Test.

4.2 Normality Test

The Normality Test is useful to find out whether the data used is normally distributed and the Normality Test used is the *Kolmogrov Smirnov Test* with the hypothesis:

Ho: Data X is normally distributed.
Ha: Data X is not normally distributed.
Decision-making:
If $Sig.(p) > 0{,}05$then Ho is accepted
If $Sig.(p) < 0{,}05$then Ho is rejected.

Table 4 . *One-Sample Kolmogorov-Smirnov Test*

		Unstandardized Residuals
N		11
Normal Parameters [a,b]	Means	.0000000
	std. Deviation	.13963231
Most Extreme Differences	absolute	.1 15
	Positive	.089
	Negative	-.1 15
Kolmogorov-Smirnov Z		.382
asymp. Sig. (2-tailed)		.999

Based on table 4 above, the resulting sigma value is 0.999 , this means that the significant value is greater than 0.5 so it $Ho$is accepted, which means the data is normally distributed. Then do the autocorrelation test.

4.3 Autocorrelation Test

The autocorrelation test aims to determine the correlation that occurs between residuals. The autocorrelation test used is a run test with the following hypothesis:
Ho: random residual (no autocorrelation)
Ha: residual is not random (autocorrelation occurs)
Decision-making:
If $Sig.(p) > 0{,}05$then Ho is accepted
If $Sig.(p) < 0{,}05$then Ho is rejected.
Test *run* can be seen in the following table

Table 5. Run tests

	Unstandardized Residuals
Test Value [a]	.01444
Cases < Test Value	5
Cases >= Test Value	6
Total Cases	11
Number of Runs	7
Z	.029
asymp. Sig. (2-tailed)	.977

Based on table 5 above, the results of the *run* test are obtained with a *sig* 0.977 value where the *sig value* >0,05 which means that the Horeceived residual is random and there is no autocorrelation. Next, the Multicollinearity Test was carried out.

4.4 Multicollinearity Test

Multicollinearity test aims to determine the correlation between the independent variables in the model. Multicollinearity detection can be seen from the VIF value with the testing criteria if VIF <10 then there is no Multicollinearity between the independent variables otherwise if the VIF value is >

10 then between the independent variables there is Multicollinearity. Multicollinearity test can be seen in the following table:

Table 6 . VIF value

Model	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
TDS	.682	1.465
CL	.935	1.070 _
NO_3	.719	1.390

Based on Table 6 above, the value $VIF$of each independent variable is less than 10, so it can be concluded that there is no multicollinearity between the independent variables.

4.5 Heteroscedasticity Test

The heteroscedasticity test was carried out to test whether in the regression model there is an inequality of variance from the residual one observation to another. The results of the heteroscedasticity test can be seen in Table 7 below

Table 7. Heteroscedasticity test

Statistic test	Value
χ^2_{count}	155.8922

Bartlett test results show the nilai χ^2_{count} =155.8922 > χ^2_{table} = 18,307 which means accept H_0which means heteroscedasticity occurs. The data in this study is spatial data, so it can be assumed that the hetero that occurs is spatial heterogeneity. The existence of spatial heterogeneity causes multiple linear regression to be inappropriate to use, so this problem can be solved using a point approach wit.

4.6 Model Form

Based on the output generated by SPSS in the *coefficient table*, the multiple regression model that can be formed is

$$y = 1.498 + 0.289\,X_1 - 1.697\,X_2 + 0.004\,X_3$$

The above equation shows a 1.498 positive coefficient value, this means that when the variable x_1, x_2, x_3value is zero, the variable $y$will increase. As for the regression coefficients that have positive values such as the coefficients on the variable x_1 dan $x_3$this means the moment variablex_1 dan $x_3$increases, the value of the variable y also increases with the coefficient. However, if the value of the regression coefficient is negative, this indicates a negative relationship with the y variable where when the variable value $x$increases, the value of the y variable will decrease by the regression coefficient and vice versa when the value of the variable $x$decreases, the value of the y variable will decrease by the regression coefficient.

The model obtained by BOD will increase when TDS and NO_3 increase because the TDS coefficient is positive by 0.289 and the NO_3 coefficient is positive by0,004 which means when TDS increases, BOD will increase by 0.289 and when NO_3 increases, BOD will increase by 0,004. Cl has a negative coefficient value of 1.697 which means that when Cl decreases, BOD will increase by 1.697.

4.7 Model Feasibility Test (*Goodness of Fit*)

The model feasibility test is needed to find out whether the regression model is suitable for use in this model, so several tests are needed to determine the feasibility of the model, namely the coefficient of determination test, F test and T test.

The coefficient of determination test is used to determine the best level of accuracy in the regression analysis which is expressed by the coefficient of determination R^2, $R^2 = 1$ means the independent variable has a perfect effect on the dependent variable. Conversely, if $R^2 = 0$it means the independent variable has no effect on the dependent variable. Based on the results of the SPSS output, the coefficient of determination is R^2 0.669 , this means that there is a relationship between TDS, Cl and NO_3 to BOD and is positive because the R value obtained is close to 1.

The effect of the independent variable on the dependent variable is determined from the R square value. If the value R^2 is equal to 0, then there is not the slightest percentage of the influence contribution given by the independent variable to the dependent variable. On the other hand, the closer to 1 the percentage of influence contribution given by the independent variable to the dependent variable is closer to perfection. The R square value is 0.669 . This indicates that the variable $x$explaining BOD is equal to 66.9 % explained by other factors. Furthermore, to determine the significance of the model as a whole, the F test is used. The F test in multiple regression analysis aims to determine the effect of the independent variables simultaneously. With the following hypothesis:

H_0 : There is no relationship between $X1, X2 \ldots Xn$ with Y
H_1: There is a relationship between $X1, X2 \ldots Xn$ with Y

Table 8. ANOVA

	Model	Sum of Squares	Df	MeanSquare	F count	Sig.
	Regression	0.394	3	0.131	4.717	.04 2 [b]
1	Residual	0.195	7	0.028		
	Total	0.589	1 0			

Based on Table 8 above, the calculated F value is 4.717 and the F table value with the number of independent variables is 3 and the number of data 11 is 4.46 . It can be seen that the calculated F value is greater than F table, so the independent variables have a simultaneous effect on the dependent variable. this can also be proven by looking at the significance value obtained in the table, namely 0.04 2 .The significance value is less than 0.05, so this indicates that there is an influence between the independent variables on the dependent variable simultaneously. Then a T-test is carried out to determine the effect of the variables partially with the following hypotheses:

H $_0$: t count <t table = no significant effect

H $_1$: t count > t table = there is a significant effect

Table 9. T Value count

Variable	T value count
x_1	1.163
x_2	3.007
x_3	-1400 _

Based on Table 9 above, the TDS variable with a t count value of 1.163 . The t table value with a significance level of 0.05 is 2.365 . This shows that the calculated t value is smaller than the t table value so that TDS does not partially affect BOD. Variable Cl with a t-count value of 3.007 has a t-count value that is greater than t-table so that Cl has a partial effect on BOD5 . Variable NO_3 with a t value of -1.400 shows a smaller value than t table so it can be concluded that NO_3 has no partial effect on NO_3.

4.8 Geographically Weighted Regression Model

The GWR model is built using two Kernel Functions, namely the Adaptive Bandwidth Function and the Fixed Bandwidth Function. The step to build these two models is to choose the optimum bandwidth to get the elements of the weighting matrix which will then be used to estimate the model parameters. The bandwidth value is obtained from processing in the GWR 4 software.

Determination *of the optimum bandwidth* using the *Cross Validation* (CV) method which is then used in determining the *Adaptive Kernel Gaussian weighting function* with the aim of providing one region and another region that still provide a relationship (neighborhood). To get the optimum bandwidth, you can use the *R Studio software*, namely the value obtained is the bandwidth value 0.9999408 with a CV score of 11.2646.

Table 10. Model feasibility test

Source	Sum of Squares	Df	MeanSquare	F_{count}
OLS *Residuals*	0.194972	4		
GWR *Improvement*	0.016355	0.62603	0.0026125	
GWR *Residuals*	0.17617	6.37397	0.028023	0.9323

Based on Table 10 above GWR Model with *Fixed* Kernel *Function bandwidth* , the calculated F value is 0.9323 and the F table value is obtained with the number of independent variables 3 and the number of data 11 is 4.46, it can be seen that the calculated F value is smaller than the F table, which means that the GWR model does not have a significant difference with regression models.

4.9 Model GWR Partial Parameter Significance Test

Parameter testing was carried out with the aim of knowing the variables that significantly influence rice production so that it can increase in the following year. Parameter test is done by testing each independent variable on the dependent variable partially. Thus each Regency/City has a model with different parameter characteristics from other regions. The following is a hypothesis from the partial test:

H_0 : t count <t table = no significant effect
H_1 : t count > t table = there is a significant effect

Table 11. Values t_{count} in the GWR Model per Regency/City

Regency/City	TDS	Cl	NO_3
Winking	1.09971	2.939379	-1.438031
Muaro Jambi	1.02217	2.871618	-1.566017
Jambi City	1.016051	2.862819	-1.567754
Sarolangun	1.085656	2.937957	-1.439205
Full River	1.244993	3.139579	-1.318869

Tebo	1.112457	2.997172	-1.445448
Batanghari	1.004429	2.848460	-1.556520
Kerinci	1.245207	3.140395	-1.322697
Bungo	1.215964	3.107930	-1.350694
East Cape	1.029896	2.886230	-1.565302
West Cape	1.021043	2.882752	-1.560899

Based on Table 11 it can be decided that $H_0$the variable Cl is rejected. This is because these variables in each district/city have a significant value by comparing the values of $t_{count\ Cl} > t_{table} = 2.365$. While it can be said that there are other factors outside the model that affect BOD in districts/cities that are not significantly affected by the variables mentioned above.

4.10 Formation of the GWR Model

The GWR modeling for each Regency/City will be different. As previously discussed regarding significant variables, that BOD5 in each district/city in Jambi Province is significantly affected by 5%, namely Cl. The complete GWR model formed for each Regency/City based on the variables that significantly influence it is shown in Table 12.

Table 12. The GWR model formed for each Regency/City

Regency/City	**GWR Model Formed**
Winking	$y = 0.2789779x_2$
Muaro Jambi	$y = 0.2718648x_2$
Jambi City	$y = 0.2711296x_2$
Sarolangun	$y = 0.2778047x_2$
Full River	$y = 0.2962636x_2$
Tebo	$y = 0.2827482x_2$
Batanghari	$y = 0.2698843x_2$
Kerinci	$y = 0.2931179x_2$
Bungo	$y = 0.2931179x_2$
East Cape	$y = 0.2730832x_2$
West Cape	$y = 0.2727917x_2$

4.11 Best Model Selection

Selection of the best model is done by comparing the GWR Model and the GWR Model by looking at the Coefficient of Determination R^2 and *Akaike Information Criterion* (*AIC*) values in each model.

Table 13. Selection of the best model

Model	R^2	*AIC*
Multiple Regression	0.669029	-3.144102
Adaptive GWR	0.6967926	-9.774071

Based on Table 13, it can be concluded that the GWR Model with Kernel *Adaptive Bandwidth Function* is the best model because it has the R^2 largest value of 0.6967926 and has a minimum *AIC value of* -9.774071. Next, we look for the best model of the *Adaptive Bandwidth* Kernel *Function* by eliminating one of the variables. The following is a comparison of the *AIC values* and R^2 the *Adaptive Bandwidth* Kernel *Function* .

5. Conclusion

Based on the results of the study, it was found that the BOD values for all regions in Jambi province fulfilled except Merangin, which was 2.1 mg/l with a threshold value of 2.0 mg/l. For other parameters, namely TDS, Cl and NO_3, they meet the threshold values. Geographically Weighted Regression (GWR) with a value of 69.7% and in multiple regression with a R^2 value of 66.9% which means that modeling using Geographically Weighted Regression (GWR) is better than multiple regression modeling.

6. Acknowledgement

PDAM pays more attention to parameters that are below the standard threshold. In addition, the PDAM can determine alternative tools or methods that can be used so that the water can become clearer.

References

[1] Diana, J. S., Szyper, J. P., Batterson, T. R., Boyd, C. E., & Piedrahita, R. H. (2017). Water quality in ponds. *Dynamics of pond aquaculture*, 53-71.

[2] Blum, H. L., & Knollmueller, R. N. (1975). *Planning for health; development and application of social change theory* (Vol. 75, No. 8, p. 1388). LWW.

[3] BPPKPD. (2017). *Peranan PDAM dalam Meningkatkan PAD*.

[4] Ali, S. F., Hassan, F. M., & Abdul-Jabar, R. A. (2017). Water quality assessment by diatoms in Tigris River, Iraq. *International Journal of Environment & Water*, *6*(2), 53-64.

[5] Nelwan, F., Wuisan, E. M., & Tanudjaja, L. (2013). Perencanaan Jaringan Air Bersih Desa Kima Bajo Kecamatan Wori. *Jurnal Sipil Statik*, *1*(10).

[6] Adianto, J., Gabe, R. T., & Djaja, K. (2021). The Challenge of Reclaiming the Commons of the Ciliwung River in Depok City. *International Journal of Design Management & Professional Practice*, *15*(1).

[7] Taghipour Javi, S., Malekmohammadi, B., & Mokhtari, H. (2014). Application of geographically weighted regression model to analysis of spatiotemporal varying relationships between groundwater quantity and land use changes (case study: Khanmirza Plain, Iran). *Environmental monitoring and assessment*, *186*, 3123-3138.

[8] Hasan, M. K., Khan, M. R. I., Nesha, M. K., & Happy, M. A. (2014). Analysis of water quality using chemical parameters and metal status of Balu River at Dhaka, Bangladesh. *Open J. Water Pollut. Treat*, *1*(2), 58-74.

[9] Yu, H., Fotheringham, A. S., Li, Z., Oshan, T., Kang, W., & Wolf, L. J. (2020). Inference in multiscale geographically weighted regression. *Geographical Analysis*, *52*(1), 87-106.

[10] Li, Z., & Fotheringham, A. S. (2020). Computational improvements to multi-scale geographically weighted regression. *International Journal of Geographical Information Science*, *34*(7), 1378-1397.

[11] O'Sullivan, D. (2003). Geographically weighted regression: the analysis of spatially varying relationships. *Geographical analysis*, *35*(3), 272-275.

[12] Fitriyani, F., Yurinanda, S., & Multahadah, C. (2023). Penerapan Metode Geographically Weighted Regression Pada Tingkat Pencemaran Air Berdasarkan Total Coliform Di Provinsi

Jambi. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, *4*(1), 603-613.
[13] Wikurendra, E. A., Syafiuddin, A., Nurika, G., & Elisanti, A. D. (2022). Water quality analysis of pucang river, sidoarjo regency to control water pollution. *Environmental Quality Management*, *32*(1), 133-144.
[14] Yaroshenko, I., Kirsanov, D., Marjanovic, M., Lieberzeit, P. A., Korostynska, O., Mason, A., ... & Legin, A. (2020). Real-time water quality monitoring with chemical sensors. *Sensors*, *20*(12), 3432.
[15] Jaybhaye, R., Nandusekar, P., Awale, M., Paul, D., Kulkarni, U., Jadhav, J., ... & Kamble, P. (2022). Analysis of seasonal variation in surface water quality and water quality index (WQI) of Amba River from Dolvi Region, Maharashtra, India. *Arabian Journal of Geosciences*, *15*(14), 1261.
[16] Suyono, M. S. (2015). *Analisis Regresi untuk Penelitian*. Deepublish.
[17] Alita, D., Putra, A. D., & Darwis, D. (2021). Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(3), 295-306.
[18] Hasan, M. I. (2013). Analisis Data Statistik Penelitian dengan Statistik.
[19] Ratmono, D. (2017). Analisis Multivariat Dan Ekonometrika Teori, Konsep, Dan Aplikasi Dengan Eviews 10.
[20] Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of clinical epidemiology*, *98*, 146-151.
[21] Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), 329-342.
[22] Hawley, S., Ali, M. S., Berencsi, K., Judge, A., & Prieto-Alhambra, D. (2019). Sample size and power considerations for ordinary least squares interrupted time series analysis: a simulation study. *Clinical epidemiology*, 197-205.
[23] Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, 387-401.
[24] Basuki, A. T., & Prawoto, N. (2016). Analisis regresi dalam penelitian ekonomi dan bisnis.
[25] Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.
[26] Hasibuan, D. O., Bekti, R. D., Sutanta, E., & Pradnyana, I. W. J. (2022). Application of the Geographically Weighted Regression Method to the Human Development Index and Visualization on the Tableau Dashboard. *IC-ITECHS*, *3*(1), 39-51.
[27] Cohen, A., & Migliorati, G. (2017). Optimal weighted least-squares methods. *The SMAI journal of computational mathematics*, *3*, 181-203.
[28] LeSage, J. P. (2004). A family of geographically weighted regression models. In *Advances in spatial econometrics: methodology, tools and applications* (pp. 241-264). Berlin, Heidelberg: Springer Berlin Heidelberg.
[29] Comber, A., Wang, Y., Lü, Y., Zhang, X., & Harris, P. (2018). Hyper-local geographically weighted regression: extending GWR through local model selection and local bandwidth optimization. *Journal of Spatial Information Science*, (17), 63-84.
[30] Sasongko, T. B., Arifin, O., & Al Fatta, H. (2019, July). Optimization of hyper parameter bandwidth on naïve Bayes kernel density estimation for the breast cancer classification. In *2019 International Conference on Information and Communications Technology (ICOIACT)* (pp. 226-231). IEEE.
[31] Liu, X., Tang, B. H., Li, Z. L., & Shang, G. (2021). Development of Kernel-Driven Models With Fixed Hotspot Width Under a General Modeling Framework in the Thermal Infrared

Domain. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 9187-9195.

[32] Roy, S. K., Manna, S., Song, T., & Bruzzone, L. (2020). Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7831-7843.

[33] Goual, H., Yousof, H. M., & Ali, M. M. (2020). Lomax inverse Weibull model: properties, applications, and a modified Chi-squared goodness-of-fit test for validation. *Journal of Nonlinear Sciences & Applications (JNSA)*, *13*(6).

[34] Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3), e1460.