p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
11(2) 2023: 22-30

Jurnal Statistika
Universitas Muhammadiyah Semarang

GEOGRAPHICALLY WEIGHTED REGRESSION ANALYSIS WITH ADAPTIVE GAUSSIAN IN THE SOCIAL AND ECONOMIC FIELDS FOR TUBERCULOSIS IN SOUTH SUMATRA 2020

Dia Cahya Wati[1], **Ismi Rizqa Lina**[2]
[1,2] Department of Data Science, University of Insan Cita Indonesia, Indonesia

***e-mail**: dia.cahya@uici.ac.id*

Article Info:
Received: July 27, 2023
Accepted: November 7, 2023
Available Online: November 30, 2023

Keywords:
Tuberculosis; Geographically Weighted Regression (GWR); Adaptive Gaussian; South Sumatera.

Abstract: Most of the tuberculosis germs not only attack the lungs, but can also attack other organs. Low income, population density, education level, low public health knowledge, and sanitation in the home environment are sources of transmission for tuberculosis sufferers. In this case, the number of tuberculosis cases varies between districts/cities. This study aims to analyze the factors that influence tuberculosis in South Sumatra at 2020 using the Geographically Weighted Regression (GWR) approach. GWR is a modification of a simple regression model into a weighted regression model that can better explain the relationship between response variables and predictors. Tuberculosis is spread in every sub-district so that it can be identified more deeply in each sub-district using the GWR with an adaptive gaussian weighting function. Based on the results of the study, the distribution of tuberculosis was affected into 5 groups. The dominant group with economic influence is the average expenditure per capita (X3) in the areas of Ogan Komering Ilir Regency, Banyuasin Regency, Ogan Ilir Regency, Palembang City and Prabumulih City and the coefficient of determination is 96.10834%.

1. INTRODUCTION

Mycobacterium Tuberculosis is a bacterium that causes tuberculosis to appear [1]. Most of the tuberculosis bacteria not only attack the lungs, but can also affect other organs of the body. The source of transmission of this disease can be caused by several factors, one of which is the environment. Residential density, walls, temperature, roof conditions, house floors are environmental factors that can influence. In addition, several factors for tuberculosis are also caused by gender, age, income, knowledge and attitudes towards tuberculosis prevention [2].

The number of cases of tuberculosis in South Sumatra Province is quite a lot after diarrheal disease. The results of the study said that the province of South Sumatra had the highest prevalence of tuberculosis. The South Sumatra provincial health office stated that the percentage case detection rate Tuberculosis disease in South Sumatra Province has fluctuated in the last 5 years. In 2018 the percentage of tuberculosis in South Sumatra was 46%, while the strategic plan target was 55%. This means that in 2018 it has not reached the target set with an achievement ratio of 83.64%.

Factors that cause tuberculosis are still high in Indonesia are low income, population density, education level, low health knowledge in the community [3]. Generally, tuberculosis is dominated in rural areas. However, the average people affected by tuberculosis are people

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

who live in urban areas because the air atmosphere and environmental conditions in urban areas are very bad and have a high potential for the emergence and spread of tuberculosis bacteria [4]. Although it does not rule out the possibility that people living in rural areas can develop tuberculosis. This is because there are still many Indonesian people who lack knowledge about the symptoms, risks and transmission of tuberculosis itself. Not a few people who can't tell the difference between a normal cough and a tuberculosis cough. This agrees with [5] which says that the reason for the low coverage of TB cases is the low level of public awareness in the process of healing and treatment. To improve people's welfare, it is hoped that the number of tuberculosis cases can decrease drastically. The difference in the number of tuberculosis sufferers in each region means that the factors that influence it are different in each region.

The assumption of identical residuals is not fulfilled because of the occurrence of cases of heteroscedasticity when the number of tuberculosis sufferers is analyzed using multiple regression analysis. To overcome this, this research will require a statistical method that takes into account geographical location (longitude, latitude) or spatial data. One of the models that affect the spatial effect is Geographically Weighted Regression (GWR). The GWR model is a development of the linear regression model by taking into account the spatial effects of the location of an observation [6].

The GWR model is a regression model in which parameter estimation is carried out at each location and uses spatial weighting, which gives different weights for each observation at each location. Spatial weights represent the effect of location on the parameter estimation. Observational data that is located closer will have a greater influence than observational data that is farther away, so that it will be given greater weight. The distance to the observation location is the weight value in general.

This is related to research conducted [7], by applying GWR to determine the factors that influence cases of malnutrition in children under five in West Java. The results of testing the GWR model with adaptive gaussian kernel weights turned out to be more suitable for modeling cases of malnutrition in children under five in West Java than the Ordinary Linear Regression (OLR) model. This can be seen from the residual squared of the GWR model with an adaptive gaussian kernel weight of 0.2555239 which is the smallest among the residual squares of the OLR model of 1.3451993 and the sum of the squares of the GWR model with a fixed gaussian kernel weight of 0.2951993. It is different from the research conducted [8] that the data used in the research has spatial effects so that the modeling method is continued. Geographically Weighted Regression (GWR) with a gaussian kernel because each observation location has a varying model.

Based on this description, in this study modeling the factors that influence the number of cases of tuberculosis in South Sumatra Province was carried out using the method Geographically Weighted Regression with adaptive gaussian.

2. Geographically Weighted Regression Model (GWR)

Model geographically weighted regression (GWR) is the development of a global linear regression model or Ordinary Linear Regression (OLR) which is used to analyze spatial heterogeneity [9]. To produce estimates of model parameters that are local for each of these data points, each parameter is calculated at each observation location point.
The GWR model can be written as follows:

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

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

with:
y_i = the observed value of the response variable i,
x_{ik} = the observed value of the predictor variable k in the i-th observation,
$\beta_0(u_i, v_i)$ = regression coefficient on the coordinates of the i-th point at a location,
$\beta_k(u_i, v_i)$ = k-th regression coefficient at each location,
ε_i = eror-i.

The error forms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be independent, identical and follow a normal distribution with zero mean and constant variance (ε_i~iid N(0,σ^2)).

2.1 Weighting Function *Adaptive Gaussian*

Adaptive kernel function is a kernel function that has a different bandwidth at each observation location. According to Pierre De Bellefon in the GWR book, the adaptive kernel level is determined by the number of observations from the destination. The lower the density of observations, the smaller the kernel. This is supported in [10] that the adaptive kernel function is better than the fixed kernel function because it assumes that adaptive has different bandwidth at each observation location. Adaptive kernel functions include functions adaptive gaussian and functions adaptive bisquare [9]. Here is the weighting function adaptive gaussian:

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h_i}\right)^2\right) \tag{2}$$

with :
w_{ij} = The weighting function adaptive gaussian
d_{ij} = Distance between the point at location i and location j which is obtained from the distance euclidean
h_i = Non parameter negative is known as the bandwidth or smoothing parameter

2.2 Model Suitability Testing

For testing the GWR model hypothesis, there are two models, namely the suitability test and model parameter testing.

2.2.1 GWR Model Statistical Test

This test was conducted to determine whether the GWR model is significantly better at modeling data than the OLS model or not. The formulation of the hypothesis is:
H_0 : $\beta_k(u_i, v_i) = \beta_k$ for each $k = 0,1,2,\ldots,p$ and $i = 1,2,\ldots,I$
H_1 : there is at least one $\beta_k(u_i, v_i) \neq \beta_k$, $k = 0,1,2,\ldots,p$
The test statistics used can be written as follows [11].

$$F = \frac{RSS(H_o)/df_1}{RSS(H_1)/df_2} \tag{3}$$

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

If F it produces a relatively small value, it can be said that the alternative hypothesis (H_1) is more appropriate to use. If compared to the global regression model, the GWR model has a better goodness of fit. If the level of significance (α) , is given, then the decision is taken by subtracting (H_0) when $F < F_{1-\alpha,df1,df2}$ where $df_1 = \frac{{\delta_1}^2}{\delta_2}$ and $df_2 = (n - p - 1)$ [12].

2.2.2 Model Parameter Testing

This test was conducted to determine which parameters statistically significantly affect the dependent variable. The form of the hypothesis is as follows:

H_0 : $\hat{\beta}_k(u_i, v_i) = 0$ for every k = 0,1,2,...,p

H_1 : at least there is one $\hat{\beta}_k(u_i, v_i) \neq 0$

Statistics the test used is:

$$t = \frac{\hat{\beta}_k(u_i, v_i)}{Se(\hat{\beta}_k(u_i, v_i))} \tag{4}$$

with $Se(\hat{\beta}_k(u_i, v_i))$ is standard error which is obtained from the root of the estimated variance of the GWR parameter. The test criteria used are, if $t > t_{(\alpha/2,df)}$ then the decision H_0 can be rejected or $\hat{\beta}_k(u_i, v_i) \neq 0$.

3. METHODOLOGY

Based on the research objectives, the analytical method used in this study is inferential analysis. Some of the software that will be used to analyze data are Rstudio and ArcGis. The data used is secondary data sourced from the Central Statistics Agency of South Sumatra Province. The variables used in this study are data on average per capita expenditure (X3) in the economic sector, human development index (X1) and life expectancy (X2) in the social sector as independent variables in 17 districts/cities throughout the South Sumatra region in 2020 The variables used in the study are listed in Table 1. The steps in modeling tuberculosis are as follows:

1. Describe each variable
2. Assuming regression data, including:
   a. Linearity Test
   b. Normality test
   c. Multicollinearity Test
   d. Heteroscedasticity Test
3. Analyzing GWR models
4. Mapping of the spread of tuberculosis in South Sumatra
5. Draw conclusions

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

Table 1. Research Variables

Response Variable		Variable Predictor	
Y	Number of TB cases in each district	X1	Total human development index
		X2	Total life expectancy
		X3	The average amount of spending per capita

3.1 Human Development Index (X1)

The Human Development Index (X1) is defined as the process of expanding choices for the population (a process of enlarging the choices of people). Human development achievements based on a number of basic components of quality of life are measured by the Human Development Index (X1). X1 is built through a basic three-dimensional approach as a measure of quality of life. This dimension includes longevity and health, knowledge, and a decent life which has a very broad meaning because it is related to many factors. The index value of each dimension obtained from the calculation of the constituent indicators. According to research [13], it showed that the higher the X1 (Human Development Index) of the country concerned, the lower the number of deaths caused by communicable and non-communicable diseases.

3.2 Life Expectancy (X2)

To evaluate the government's performance in improving the welfare of the population in general requires X2 as a tool to calculate it. In addition, X2 is also used to improve health status in particular. A low X2 in an area must be followed by health development programs and other social programs including environmental health, adequate nutrition and calories including poverty eradication programs [14]–[16]. The X2 indicator is used to measure the dimensions of long and healthy life at birth. X2 is a constituent indicator of the X1 calculation so it is very important to influence each other who are also associated with tuberculosis.

3.3 Average Per-Capita Expenditures (X3)

The costs incurred for the consumption of all household members for a month are the average expenditure per capita, this includes those originating from purchases, gifts, or own production divided by the number of household members in the household. Household consumption is divided into consumption of food and non-food. Expenditures for food consumption were calculated over the past week, while those for non-food were calculated over the past month and 12 months. These results are then converted into a month's average spending. The average per capita consumption/expenditure figures presented in this publication are obtained from the quotient of the total consumption of all households (both consuming food and not) to the total population [14].

4. RESULTS AND DISCUSSION

Regression is a method for measuring the magnitude of the influence of the response variable on predictors. There are several basic assumptions that can produce the best unbiased linear estimator. The regression model is obtained from the least squares method using regression known as the classical assumption. The classical assumptions consist of homoscedasticity, non-autocorrelation, non-multicollinearity and residual normality. Based on the results of the assumption test, the regression conditions are obtained but not homoscedasticity so that it is

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

continued with GWR because the results of breusch pagan test with P-value < 0.05.

The thing that needs to be considered in doing the GWR model is the determination of the kernel function. The kernel function is a weighting function used to estimate parameters in the GWR model. Data that is closer to the regression point *i* will get more weight than data that is farther away. Based on these data in this study used the kernel function adaptive gaussian below with AIC values 319.6779.

GWR model parameter estimation using the method Weighted Least Square (WLS). The weight value represents the observation data with one another. Therefore, weighting in GWR has a very important role. As for the estimation of GWR with a weighting matrix adaptive gaussian can be seen in Table 2.

Table 2. Parameter Estimates in South Sumatra

No	Regency
1	$Y_{\text{ogan komering ulu}}$ = -111.52 X1 + 229.98 X2 + 0.0042 X3
2	$Y_{\text{ogan komering ilir}}$ = -1.59 X1 + 36.74 X2 + 0.0043 X3
3	$Y_{\text{Muara Enim}}$ = -72.36 X1 + 198.65 X2 + 0.0037 X3
4	Y_{Lahat} = -105.90 X1 + 266.22 X2 + 0.0014 X3
5	$Y_{\text{Musi rawas}}$ = -84.63 X1 + 175.96 X2 + 0.00094 X3
6	$Y_{\text{Musi Banyuasin}}$ = -94.44 X1 + 202.85 X2 + 0.0019 X3
7	$Y_{\text{Banyuasin}}$ = -18.64 X1 + 63.53 X2 + 0.0046 X3
8	$Y_{\text{south ogan komering ulu}}$ = -86.68 X1 + 254.32 X2 + 0.0027 X3
9	$Y_{\text{east ogan komering ulu}}$ = -54.68 X1 + 130.64 X2 + 0.0040 X3
10	$Y_{\text{Ogan Ilir}}$ = -39.53 X1 + 57.55 X2 + 0.0051 X3
11	$Y_{\text{Empat Lawang}}$ = -1.45 X1 + 2.02 X2 + 0.4423 X3
12	Y_{Pali} = -0.79 X1 + 2.39 X2 + 4.24 X3
13	$Y_{\text{northern musi rawas}}$ = -2.31 X1 + 2.37 X2 + 0.905 X3
14	$Y_{\text{Palembang}}$ =-0.91 X1 + 0.26 X2 + 5.14 X3
15	$Y_{\text{Prabumulih}}$ = -1.30 X1 + 0.71 X2 + 5.58 X3
16	$Y_{\text{Pagar Alam}}$ = -1.94 X1 + 3.67 X2 + 0.92 X3
17	$Y_{\text{Lubuk Linggau}}$ = -1.84 X1 +2.103 X2 + 0.57 X3

4.1. GWR Model Significance Test

The GWR model significance test was conducted to determine whether the GWR model was significantly better at modeling data than the OLR model or not. The formulation of the hypothesis is:

H_0 : $\beta_k(u_i, v_i) = \beta_k$ for each $k = 0,1,2,\ldots,p$ and $i = 1,2,\ldots,I$

H_1 : there is at least one $\beta_k(u_i, v_i) \neq \beta_k$, $k = 0,1,2,\ldots,p$

Table 3. ANOVA of GWR

Model	df	Sum of Square	Mean Square	F
Regression	4.0000	31645231	4.0148	3.74
Residual	9.1841	2866770	312145	
Total	3.8159	296683	77749	

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

In table 3, with a confidence level of 5%. The results obtained from the R software for the adaptive gaussian weighting function are shown in table 3. The result is (4.0148 > 3.74), which means that there is a significant difference between the regression model and the GWR model.

4.2. Mapping of Tuberculosis Rates

The next step is to map the results of the estimated Tuberculosis disease rate in South Sumatra in 2020. The mapping of the estimation results using the GWR model on the Tuberculosis disease rate for each Regency/City in South Sumatra is as follows:

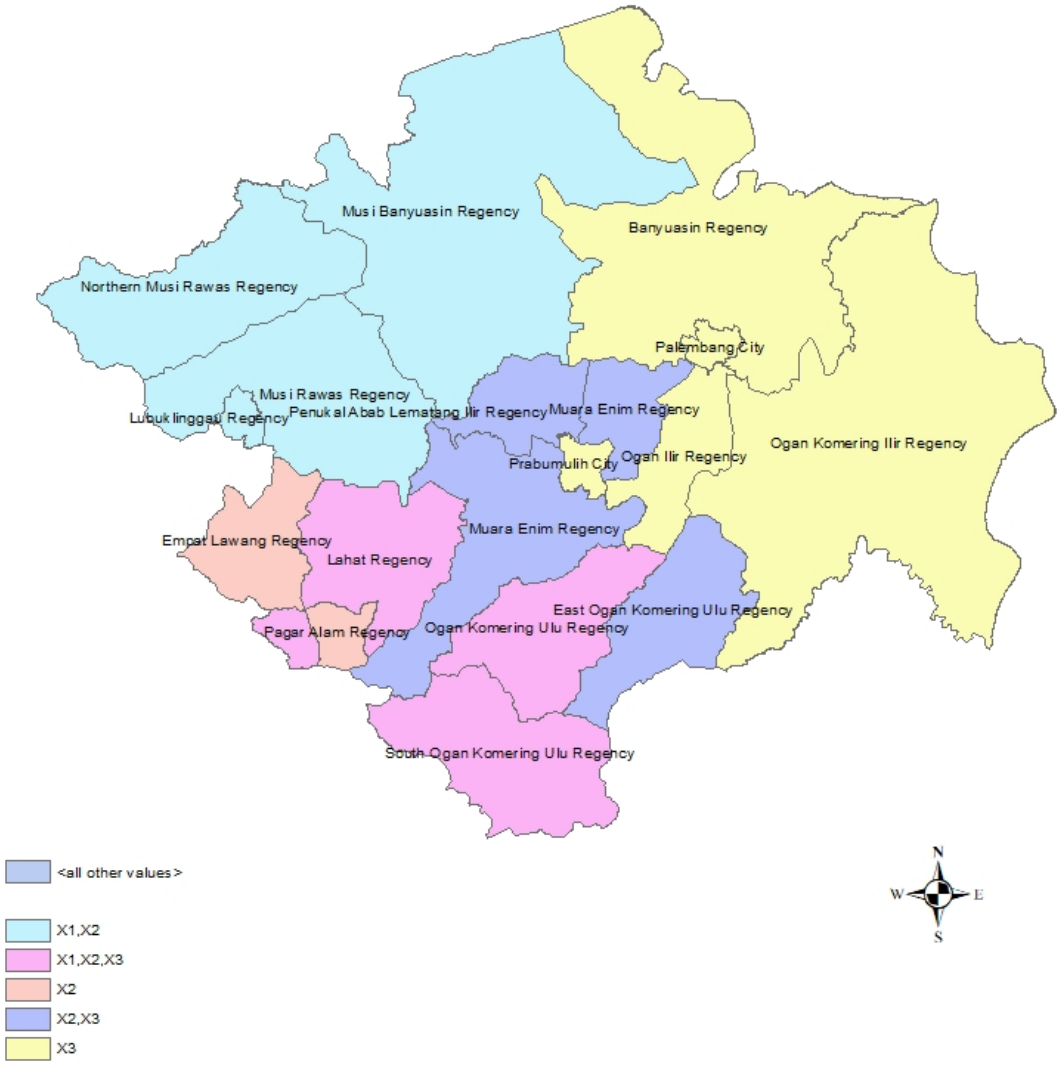

Fig 1. Map of the Distribution of Tuberculosis in Each District/City

Figure 1. explains that there are 5 color classifications on the map. Each color represents the tuberculosis rate in each district/city in South Sumatra. Like the dusty pink color, owned by Pagar Alam Regency and Empat Lawang Regency, it means that tuberculosis in these districts is affected by X2 in the social field. For the blue color it is influenced by X2 in the social sector and X3 in the economic sector with the areas of Muara Enim Regency, East Ogan Komering Ulu Regency, and Pali Regency. For the sky blue color it is influenced by X1 and X2 with the

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

areas of Northern Musi Rawas Regency, Musi Banyuasin Regency, Musi Rawas Regency and Lubuk Linggau Regency. The purple color is influenced by the X1 and X2 in the social sector and the X3 in the economic sector in the areas of Ogan Komering Ulu Regency, Lahat Regency and South Ogan Komering Ulu Regency. As for the yellow color, which is influenced by the average per capita expenditure (X3) in the economic sector, the area is Ogan Komering Ilir Regency, Banyuasin Regency, Ogan Ilir Regency, Palembang City, and Prabumulih City.

5. CONCLUSION

Tuberculosis disease in South Sumatra in 2020 using Geographically weighted regression (GWR) with an adaptive bisquare weighting function is influenced by 5 groups. Where the first group is influenced by life expectancy (X2), the second group is influenced by X2 and average per capita expenditure (X3), the third group is influenced by the human development index (X1) and X2, the fourth group is influenced by X1, X2, and X3 , while the fifth group was influenced by the lesson plan. One that is influenced by the X3, namely Prabumulih City with its GWR model is $Y_{Prabumulih} = -1.30\,X1 + 0.71\,X2 + 5.58\,X3$. GWR with the adaptive bisquare weighting function is very well used to analyze the factors that influence tuberculosis in South Sumatra, as seen by the coefficient of determination that is owned by 96.10834%.

REFERENCES

[1] A. S. N. Zaina, R. S. Pontoh, and B. Tantular, "Pemodelan Dan Pemetaan Penyakit TB Paru di Kota Bandung Menggunakan Geographically Weighted Negative Binomial Regression: Studi Kasus Dinas Kesehatan Kota Bandung," in *Prosiding Seminar Nasional Statistika Aktuaria| Departemen Statistika FMIPA Universitas Padjadjaran*, 2021, pp. 62–71.

[2] S. Ramadhani, "Analisis Spasial Penyebaran Penyakit Tuberkulosis di Sumatera Utara Menggunakan Indeks Moran dan Local Indicator of Spatial Association (LISA)," Universitas Sumatera Utara, 2020.

[3] Z. Azzahra, "Faktor-Faktor yang Mempengaruhi Kejadian Penyakit Tuberkulosis Paru di Wilayah Kerja Puskesmas Muliorejo Kecamatan Sunggal Kabupaten Deli Serdang Tahun 2017," 2017.

[4] S. Suharyo, "Determinasi Penyakit Tuberkulosis di Daerah Pedesaan," *KEMAS: Jurnal Kesehatan Masyarakat*, vol. 9, no. 1, pp. 85–91, 2013.

[5] P. S. BARAT, "PENGETAHUAN, SIKAP DAN PERILAKU MASYARAKAT TENTANG PENYAKIT TUBERKULOSIS (TB) PARU DI KECAMATAN SUNGAI TARAB, KABUPATEN TANAH DATAR".

[6] S. M. Meutuah, H. Yasin, and I. M. Di Asih, "Pemodelan Fixed Effect Geographically Weighted Panel Regression untuk Indeks Pembangunan Manusia di Jawa Tengah," *Jurnal Gaussian*, vol. 6, no. 2, pp. 241–250, 2017.

[7] A. Maulani, N. Herrhyanto, and M. Suherman, "Aplikasi Model Geographically Weighted Regression (GWR) Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Kasus Gizi Buruk Anak Balita Di Jawa Barat," *Jurnal EurekaMatika*, vol. 4, no. 1, pp. 46–63, 2016.

[8] Z. F. Annabilah and H. T. Sutanto, "Pemodelan Indeks Pembangunan Manusia di Jawa Timur Menggunakan Geographically Weighted Regression (GWR)," *MATHunesa: Jurnal Ilmiah Matematika*, vol. 7, no. 1, pp. 14–17, 2019.

[9] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley \& Sons, 2003.

[10] A. K. Lumaela, B. W. Otok, and S. Sutikno, "Pemodelan chemical oxygen demand (cod) sungai di Surabaya dengan metode mixed geographically weighted regression," *Jurnal Sains dan Seni*

[DOI: 10.14710/JSUNIMUS.11.2.22-30]

ITS, vol. 2, no. 1, pp. D100--D105, 2013.

[11] C. Brunsdon, A. S. Fotheringham, and M. Charlton, "Some notes on parametric significance tests for geographically weighted regression," *Journal of regional science*, vol. 39, no. 3, pp. 497–524, 1999.

[12] Y. Leung, C.-L. Mei, and W.-X. Zhang, "Statistical tests for spatial nonstationarity based on the geographically weighted regression model," *Environment and Planning A*, vol. 32, no. 1, pp. 9–32, 2000.

[13] A. Ramani, "Hubungan Indeks Pembangunan Manusia Dengan Indikator Penyakit, Lingkungan, Dan Gizi Masyarakat (Analisis Data Sekunder Negara Anggota UNDP)," *Jurnal Ilmu Kesehatan Masyarakat*, vol. 10, no. 1, 2014.

[14] A. H. Hidup, "MENELAAH RELASI GENDER EQUALITY TERHADAP PERTUMBUHAN EKONOMI DI JAWA TENGAH".

[15] S. Mariani, Wardono, Masrukan, and F. Fauzi, "The ArcView and GeoDa application in optimization of spatial regression estimate," *J Theor Appl Inf Technol*, vol. 95, no. 5, 2017.

[16] M. Y. Darsyah, I. J. Suprayitno, F. Fuzi, B. W. Otok, and B. S. S. Ulama, "Smooth Support Vector Machine (SSVM) for classification of Human Development Index," *J Phys Conf Ser*, vol. 1217, no. 1, p. 012114, 2019, doi: 10.1088/1742-6596/1217/1/012114.