GEOGRAPHICALLY WEIGHTED REGRESSION MODELING WITH FIXED GAUSSIAN KERNEL WEIGHTS ON TUBERCULOSIS CASES DATA IN NORTH SUMATRA PROVINCE

(Thesis)

By

EVA SARI B. SILABAN
NPM. 2117031072

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

GEOGRAPHICALLY WEIGHTED REGRESSION MODELING WITH FIXED GAUSSIAN KERNEL WEIGHTS ON TUBERCULOSIS CASES DATA IN NORTH SUMATRA PROVINCE

By

Eva Sari B. Silaban

Multiple linear regression is often used to analyze the relationship between variables, but it is less effective in handling data with spatial heterogeneity. Geographically Weighted Regression (GWR) overcomes this limitation by considering geographical aspects in parameter estimation. This study applies GWR to analyze the factors affecting the number of Tuberculosis (TB) cases in North Sumatra, using variables such as BCG immunization, population density, access to sanitation, education, and health facilities. The GWR model with a fixed Gaussian kernel performed better than multiple linear regression, with an AIC value of 327.431 (smaller than 372.018 in multiple linear regression), confirming that this model is more suitable in capturing the spatial variation of TB distribution. These findings can support location-based health policies for more effective TB control.

Keywords: Geographically Weighted Regression, Tuberculosis, Spatial Heterogeneity, North Sumatra, Spatial Modeling.

GEOGRAPHICALLY WEIGHTED REGRESSION MODELING WITH FIXED GAUSSIAN KERNEL WEIGHTS ON TUBERCULOSIS CASES DATA IN NORTH SUMATRA PROVINCE

EVA SARI B. SILABAN

Thesis

As One of the Requirements to Achieve the Degree
BACHELOR OF MATHEMATICS

In

Department of Mathematics
Faculty of Mathematics and Natural Sciences

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2025

Thesis Title	:	**GEOGRAPHICALLY WEIGHTED REGRESSION MODELING WITH FIXED GAUSSIAN KERNEL WEIGHTS ON TUBERCULOSIS CASES DATA IN NORTH SUMATRA PROVINCE**
Student Name	:	**Eva Sari B. Silaban**
Student Identification Number	:	**2117031072**
Study Program	:	**Mathematics**
Faculty	:	**Mathematics And Natural Sciences**

APPROVE

1. Supervisors

Dr. Khoirin Nisa, S.Si., M.Si.
NIP 197407262000032001

Bernadhita Herindri S.U, S.Si., M.Sc.
NIP 199206302023212034

2. Head of Mathematics Department

Dr.Aang Nuryaman, S.Si.,M.Si.
NIP. 197403162005011001

VALIDATE

1. Examiner Tim

Head : **Dr. Khoirin Nisa, S.Si., M.Si** .....

Secretary : **Bernadhita Herindri S.U, S.Si., M.Sc.** .....

Examiner
Not Supervisor : **Prof. Drs. Mustofa, M.A., Ph.D.** .....

2. Dean of the Faculty of Mathematics and Natural Sciences

Ir. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

The Date of Passing the Examination: **February 20, 2025**

STUDENT THESIS STATEMENT

The Undersigned below:

Name	:	**Eva Sari B. Silaban**
Student Identification Number	:	**2117031072**
Study Program	:	**Mathematics**
Thesis Title	:	**Geographically Weighted Regression Modeling With Fixed Gaussian Kernel Weights On Tuberculosis Cases Data In North Sumatra Province**

Hereby declare that this thesis is the result of my own work. If in the future it is proven that this thesis is the result of a copy or made by someone else, then I am willing to accept sanctions in accordance with applicable academic provisions.

Bandar Lampung, 20 February 2025

Author,

Eva Sari B. Silaban

BIOGRAPHY

The author, Eva Sari B. Silaban, was born in Bandar Lampung on April 8, 2003. She is the third of three siblings, born to K. Silaban and S. Nainggolan.

Eva began her education at TK Bhakti Putra from 2008 to 2009, then continued to SD Negeri 1 Suka Bhakti from 2009 to 2015. She pursued her secondary education at SMP Negeri 1 Gedungaji Baru from 2015 to 2018 and completed her high school studies at SMA Negeri 1 Rawajitu Selatan from 2018 to 2021. She was subsequently admitted to the Mathematics Study Program at the Faculty of Mathematics and Natural Sciences, Universitas Lampung, through the SBMPTN selection pathway and has been undertaking her studies from 2021 to 2025.

During her academic journey, Eva has been actively involved in Christian student organizations. In 2022, she joined the Persekutuan Oikumene Mahasiswa MIPA (POMMIPA), where she served as both secretary and treasurer. Additionally, she participated in an internship program at the Bandar Lampung City Government from January to February 2024. From June to August 2024, she also took part in the Community Service Program (KKN) in Marga Mulya Village, Bumi Agung District, Lampung Timur Regency.

INSPIRATIONAL WORDS

"But seek first the kingdom of God and His righteousness, and all these things will be added to you."

(Matthew 6:33)

"Who you are is God's gift to you, but who you become, is your gift to God".

(Hans Urs von Balthasar)

"Dream big because we have a big God."

(Nick Vujicic)

"People's biggest problem is not a lack of love for God, but not realizing how much God loves them. When we truly understand His love, our lives will be filled with true faith, hope, and joy."

(Sari Silaban)

DEDICATION

This thesis is first dedicated to the Lord Jesus Christ as an expression of deep gratitude for His provision, blessings and grace, which always strengthen me so that this thesis can be completed properly and on time. And with gratitude and happiness, I offer my gratitude to:

My Beloved Mother and Father

There are no words other than a big thank you for my parents. With love and respect, I dedicate this thesis to my beloved parents. Thank you for your endless prayers, love, support, and sacrifice. You are a source of inspiration and strength for me in every step of this journey. I hope this simple work can be a source of pride for you, as you have always been a source of pride for me.

Supervisor and Examiner

Thank you to my supervisors and examiner who have been very helpful, motivating, providing direction and valuable knowledge.

My Beloved Brother and Sister

I also dedicate this thesis to my beloved brother and sister. Thank you for every support, encouragement, and affection you give. You are always a place to share, a source of motivation, and a reminder that I never walk alone. May this success also be a happiness for all of us.

My Friends

To my best friends, thank you for every support, togetherness, and laughter that has colored this journey. You are a family that always encourages in difficult times and shares happiness in every achievement. I love you all, Buddies and Abraham Family.

Beloved Almamater
Universitas Lampung

AKNOWLEDGEMENT

Praise and gratitude to God Almighty for His abundance of blessings and gifts so that the author can complete this thesis entitled “Geographically Weighted Regression Modeling With Fixed Gaussian Kernel Weights On Tuberculosis Cases Data In North Sumatra Province” well and smoothly and right on time.

In the process of preparing this thesis, many parties have helped provide guidance, support, direction, motivation and advice so that this thesis can be completed. Therefore, on this occasion the author would like to thank:

1. Dr. Khoirin Nisa, S.Si., M.Si. as the first supervisor who has taken a lot of time to provide direction, guidance, motivation, advice and support to the author so that he can complete this thesis.
2. Bernadhita Herindri S.U, S.Si., M.Sc. as Supervisor II who has provided direction, guidance and support to the author so that he can complete this thesis.
3. Prof. Drs. Mustofa, M.A., Ph.D as an examiner who has been willing to provide criticism and suggestions and evaluation to the author so that it can be even better.
4. Dr. Aang Nuryaman, S.Si., M.Si. as the Head of Mathematics Department, Faculty of Mathematics and Natural Sciences, University of Lampung.
5. Dr. Agus Sutrisno, S.Si., M.Si. as academic advisor.
6. All lecturers, staff and employees of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung.
7. My parents, sister, brother-in-law, brother, sister-in-law and nephews and nieces. who always provide support, prayers, and motivation during the completion of this thesis.

8. My beloved cousins, Widia and Hani, thank you for all your support, togetherness, and encouragement. Your presence brings happiness and encouragement in every step of this journey. May we always support each other and achieve success together.

9. My best friends Reni, Hotdiana, Widia, Sinta, Sabil, Dera, Meiliana, Nabila, Maya, Tasya, Rahma, Adinda, Dita, Bang Sherina, Vedisya, and Fathan who have helped, encouraged, and motivated me so that I can complete this thesis well.

Hopefully this thesis can be useful for all of us. The author realizes that this thesis is far from perfect, so the author hopes for constructive criticism and suggestions to make this thesis even better.

Bandar Lampung, 20 February 2025

Eva Sari B. Silaban

TABLE OF CONTENTS

TABLE OF CONTENTS . . . . . **xiii**
LIST OF TABLES . . . . . **xiv**
LIST OF FIGURES . . . . . **xv**
I INTRODUCTION **1**
1.1 Background 1
1.2 Problem Formulation 3
1.3 Research Objectives 3
1.4 Research Benefits 3
II LITERATURE REVIEW **5**
2.1 Multiple Linear Regression 5
2.2 Classical Assumptions of Linear Regression 6
2.2.1 Normality Test 6
2.2.2 Autocorrelation Test 7
2.2.3 Multicollinearity Test 7
2.3 Model Parameter Test 8
2.4 Heterogeneity Test 9
2.5 Geographically Weighted Regression (GWR) 9
2.5.1 Spatial Weighting 10
2.5.2 Parameter Estimation of GWR Model 11
2.5.3 Parameter Test of GWR Model 13
2.6 Model Goodness Criteria 14
2.7 TB Disease 15
III RESEARCH METHODOLOGY **17**
3.1 Research Time and Place 17
3.2 Research Data 17
3.3 Research Methods 18
IV RESULTS AND DISCUSSION **21**
4.1 Descriptive Analysis 21
4.1.1 Number of TB Cases 22

4.1.2 Number of BCG Immunized Infants 24
4.1.3 Population Densitys 25
4.1.4 Percentage of Households With Access to Adequate Sanitation 27
4.1.5 Number of Health Workers 28
4.1.6 Number of Health Facilities 29
4.2 Multiple Linear Regression Analysis 31
4.2.1 Classical Assumption Test of Multiple Linear Regression . . 32
4.2.2 Model Parameter Test 34
4.3 Spatial Heterogeneity Test 35
4.4 GWR Modeling 36
4.4.1 Calculation of Euclidian Distance and Bandwidth 36
4.4.2 Parameter Estimation of GWR Model 40
4.4.3 Parameter Testing of the GWR Model 43
4.5 Model Goodness Criteria 45
V CONCLUSION 47
REFERENCES 48
APPENDIX 51

LIST OF TABLES

1. Spatial Data Research Variables 18

2. Descriptive Analysis of TB Data 21
3. Parameter Estimates with Statistical Metrics 31
4. Normality Test 32
5. Autocorrelation Test 33
6. VIF Value of Multicollinearity Test 33
7. Simultaneous Test of Model Parameters 34
8. Partial Test of Model Parameters 34
9. Spatial Heterogeneity Test 35
10. Euclidean distance of each location 37
11. The Weighting of each location 39
12. Parameter Estimation of GWR Model 41
13. Simultaneous Testing of GWR Model 43
14. t_{table} GWR Model 44
15. Significant Independent Variable Groups 45
16. Model AIC Criterion Value 45

LIST OF FIGURES

1. Flow chart of the research steps 20

2. Distribution Map of the Number of TB Cases in North Sumatra . . . 23
3. Distribution Map of the Number of BCG Immunized Infants 24
4. Map of Population Density Distribution 26
5. Distribution Map of Households With Access to Adequate Sanitation 27
6. Distribution Map of number of health workers 28
7. Distribution Map of Number of Health Facilities 30
8. Plot of Y Actual and Y Predicted 42

CHAPTER I

INTRODUCTION

1.1 Background

Regression analysis is one of the most commonly used statistical methods to study the relationship between the dependent variable (response) and one or more independent variables (predictors) (Nurdin et al., 2014). Armstrong in Basri (2019) stated that regression analysis aims to model and analyze the relationship between variables. Regression models can be simple regression (one independent variable) or multiple regression (two or more independent variables).

Multiple linear regression allows us to identify the factors that influence a dependent variable by considering several independent variables simultaneously. However, conventional multiple linear regression has a drawback when applied to geographically dispersed data, this model presumes that the connection between the dependent and independent variables remains consistent across different regions. In reality, the relationship can vary significantly from one location to another, especially if the data exhibits spatial heterogeneity or geographical differences.

The influence of geographical location causes variations in the value of the response variable that is influenced by various factors that differ at each location, known as spatial heterogeneity (Anselin and Getis, 1992). Spatial data itself location information and attribute descriptions indicate a close relationship between the data and the location of the observations, so spatial variations need to be considered in the analysis.

To overcome this problem, the Geographically Weighted Regression (GWR) method was developed, which is a localized form of linear regression that considers spatial aspects (Fotheringham et al., 2002). This model allows local analysis in each

geographical area, thus revealing risk factors that are specific to each region. In GWR, the influence of independent variables on the dependent variable is not assumed to be homogeneous across regions but rather varies locally.

Based on the Ministry of Health report written by Rokom, Dr. Siti Nadia Tarmizi, M.Epid as the Bureau of Communication and Public Services of the Ministry of Health of the Republic of Indonesia, said that in 2022 Indonesia ranked second highest in the world after India regarding the number of TB cases, with 969 thousand cases and 93 thousand deaths per year or equivalent to 11 deaths per hour. North Sumatra is among the top five provinces with the highest number of TB cases in Indonesia. The spread of TB is strongly influenced by various local factors, such as population density, poverty levels, access to health services, and environmental quality.

Using the GWR method, this study aims to identify factors that influence the number of TB cases in various regions in North Sumatra province and analyze the spatial variation of the influence of these factors. The results of this modeling are expected to provide deeper insight into the spatial distribution of TB, so that it can support more targeted health policies, especially in addressing local variations that affect the spread of this disease.

Some previous studies have shown that distance and geographic distribution play a significant role in increasing the number of TB cases. Like the research conducted by Long Viet Bui et al. (2018), the use of GWPR model can identify well the geographical factors that affect the incidence of tuberculosis in Nam Dinh region, Vietnam. Taking another example from the city of Bandung, research by Octavianty et al. (2017), mentioned that modeling the number of TB cases with a semiparametric approach (fixed and flexible) provides a more accurate picture of the distribution of this disease. Thus, the challenge in mapping the number of TB cases lies not only in data collection but also in understanding the complex factors that influence its spread. Later research by Wei et al. (2016) discovered that the GWR model is more effective in geographically differentiating the connection between the average number of BTA-positive TB cases and socio-economic factors, which allows for a better interpretation of the dataset,(adjusted R_2 = 0.912, AICc = 1107.22) than the OLS model (adjusted R_2 = 0.768, AICc = 1196.74).

Researchers are looking to investigate the linear regression method for the number of TB cases in North Sumatra while taking into account spatial factors through the GWR model, this model can help understand the spatial distribution pattern of TB in North Sumatra by considering variables such as BCG immunization, population density, households that have access to proper sanitation, number of health workers and the number of health facilities such as hospitals, Puskesmas, polyclinics, and so on.

1.2 Problem Formulation

Based on the above background, the problem formulations discussed in this study are as follows:

1. How is GWR modeling data on the number of TB cases in North Sumatra province in 2022?
2. What are the factors that influence the number of TB cases in districts/cities in the North Sumatra province in 2022 using the GWR method?

1.3 Research Objectives

The objectives of this research include:

1. Modeling data on the number of TB cases in North Sumatra province in 2022 using GWR.
2. Using the GWR method dentify what factors influence the number of TB cases in districts/cities in North Sumatra province.

1.4 Research Benefits

The benefits of this research are:

1. Providing a deeper understanding of how the number of TB cases in North Sumatra Province is influenced by various factors. By using the GWR method, the analysis can show the variation in the influence of these factors spatially, which helps in understanding the dynamics of disease spread.

2. Assist the government in decision-making so that public health policies can be designed and implemented more effectively to control and prevent the spread of TB.

CHAPTER II

LITERATURE REVIEW

2.1 Multiple Linear Regression

The multiple linear regression model extends simple linear regression to situations involving multiple independent or predictor variables. Thus, the aim of multiple regression is to investigate and measure the association between a numerical dependent variable and one or more qualitative or quantitative predictor variables. The outcome of multiple linear regression is a model that illustrates the connection between two or more independent variables (x) and one dependent variable (y) (Del Águila and Benítez-Parejo, 2011). The general equation as follows:

$$y = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik} + \varepsilon \tag{2.1}$$

where:

y : value of the dependent variable
x_{ik} : the value of the kth independent variable at observation-i
β_0 : constant
β_k : regression parameters of the independent variable k
ε : error, where $\varepsilon \sim IIDN(0, \sigma^2)$
i : 1, 2, 3, ..., m
k : 1, 2, 3, ..., n.

2.2 Classical Assumptions of Linear Regression

According to Sholihah et al. (2023), the classical assumption test is a statistical requirement that must be met in Ordinary Least Squares (OLS) based multiple linear regression analysis. The linear regression model is considered good if it meets these classical assumptions, namely normally distributed residuals, and there is no multicollinearity or heteroscedasticity. The importance of fulfilling these assumptions is so that the resulting regression model is unbiased and the test can be trusted. If any of the classical assumptions are violated, the outcomes of the regression analysis cannot be regarded as BLUE (Best Linear Unbiased Estimator). This study will employ specific tests to evaluate the assumptions.

2.2.1 Normality Test

Normality test is a test conducted to determine whether the regression analysis model, response variables, and predictor variables are both normally distributed or not. So it is necessary to do a normality test. One of the methods that can be used for the normality test is the Shapiro-Wilk test, with the following hypothesis:

H_0 : error is normally distributed
H_1 : errors are not normally distributed

The Shapiro-Wilk test statistic is (Cahyono, 2015):

$$W = \frac{\left(\sum_{i=1}^{n} a_i (x_{(n+1)-i} - x_i)\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad i = 1, 2, 3, \ldots n \tag{2.2}$$

where:

W : calculated Shapiro-Wilk coefficient
$\bar{x}$: average of data
$x_{(n+1)-i}$: data at position $(n+1) - i$
x_i : data at position i.

Accept H_0 if the value of $W_{\text{count}} < W_{\text{table}}$ or if p-value $> \alpha = 0.05$, which means the data is normally distributed or the assumption of normality is met.

2.2.2 Autocorrelation Test

According to Mardiatmoko (2020), autocorrelation is a situation in a regression model where there is a correlation between residuals in period t and residuals in the previous period (t-1). A good regression model is free from autocorrelation. This test can be done with the Durbin-Watson test, as follows:

H_0 : there is no autocorrelation in the residuals

H_1 : there is autocorrelation in the residuals

$$D_W = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}, \quad i = 1, 2, 3, \ldots n \tag{2.3}$$

where:

e_i : residual of the observation at period i
e_{i-1} : residual of the observation at period $i - 1$.

Reject H_0 if p-value $< \alpha = 0.05,$ which means autocorrelation exists in the residuals.

2.2.3 Multicollinearity Test

According to In and Asyik (2019), the multicollinearity test is conducted to check whether there is intercorrelation or collinearity between the independent variables in the regression model. This intercorrelation refers to a strong linear relationship between one independent variable or predictor and another predictor in the model. One method that can be used in detecting multicollinearity is by calculating the VIF (Variance Inflation Factor) value with the formula:

$$VIF = \frac{1}{1 - R_j^2} \tag{2.4}$$

where:

R_j^2 : the coefficient of determination from the results of regressing the independent variable j with other independent variables.

The hypothesis H_0 is that there is no multicollinearity, with the decision criteria:

if $VIF < 10$, then fail to reject H_0, which means that there is no multicollinearity (Montgomery et al., 1992).

2.3 Model Parameter Test

Parameter testing in multiple linear regression models includes simultaneous testing and partial testing. Simultaneous testing aims to determine whether the independent variables collectively have a significant effect on the dependent variable. This test is conducted using the F-test with the following hypotheses (Caraka and Yasin, 2017):

$\text{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : at least one $\beta_j \neq 0,\ j = 1, 2, \ldots, k.$

The F-test statistic is defined as:

$$F_{\text{count}} = \frac{(R^2/k)}{((1 - R^2)/(n - k - 1))} \tag{2.5}$$

where:

R : regression correlation coefficient,
k : number of predictor variables,
n : number of data samples.

Reject H_0 if $\text{F}_{\text{count}} > F_{(\alpha, df_1, df_2)}$ where $\alpha = 0.05$. which means that the independent variables jointly significantly affect the dependent variable.

Furthermore, partial testing is used to determine the variables that significantly affect the response variable. The partial testing hypothesis is as follows:

$\text{H}_0 : \beta_k = 0$

$\text{H}_1 : \beta_k \neq 0 \quad \text{for } k = 1, 2, \ldots, m,\ i = 1, 2, \ldots, n.$

With a significance level of $\alpha = 0.05$, the test statistic is defined as:

$$t_{\text{count}} = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)} \tag{2.6}$$

where $\text{SE}(\hat{\beta}_k)$ represents the standard error of the coefficient $\hat{\beta}_k$. The decision rule states that H_0 is rejected if $|t_{\text{count}}| > t_{\text{table}}(\alpha/2, df)$. which means there is an influence between the dependent variable individually with the independent variable.

2.4 Heterogeneity Test

According to Bakri et al. (2024), the spatial heterogeneity test is a test that aims to determine whether each observation location has its characteristics or uniqueness. This test can be performed using the Breusch Pagan test, with the following hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \cdots = \sigma_n^2$ (no spatial heterogeneity)
H_1 : there is at least one $\sigma_i^2 \neq \sigma_j^2$ where $i \neq j$ (spatial heterogeneity)

With the test statistics used are (Anselin, 1988):

$$BP = \frac{1}{2} f^T Z (Z^T Z)^{-1} Z^T f \tag{2.7}$$

where,

Z : $[Z_1, Z_2, \ldots, Z_p]^T$ is a matrix of size $n \times (p+1)$ which contains predictor variables
f : $[f_1, f_2, \ldots, f_n]^T$ with $f_i = \left(\frac{\epsilon_i^2}{\sigma^2} - 1 \right)$.

With the test criteria, reject H_0 if $BP > \chi^2_{(\alpha,p)}$ or if p-value $< \alpha$, which means that there is spatial heterogeneity or differences in the characteristics of one region with other regions, so it is necessary to do modeling using GWR.

2.5 Geographically Weighted Regression (GWR)

According to Fotheringham et al. (2002), GWR is a statistical method used to analyze spatial heterogeneity, or an extension of the classical linear regression model, which is used to model data that has spatial influence. With this approach, spatial weights represent the magnitude of different spatial influences in each location, based on the idea of GWR model. In the GWR model the dependent variable y is predicted using the independent variables, where the regression coefficient of each variable depends on the location where the data is taken, this location is denoted as (u_i, v_i), which is the two-dimensional coordinate vector (latitude and longitude) for the i-th location. The GWR model can be written according to equation (2.8) (Fotheringham et al., 2002).

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2.8}$$

where:

y_i : observation value of the response variable at location i,
$x_{k,i}$: observation value of the k-th predictor variable at the observation location i,
β_0 : a constant or intercept at the i-th observation,
(u_i, v_i): geographical coordinates (longitude, latitude) of the observation location i,
ε_i : error at observation location i, where $\varepsilon \sim \text{IIDN}(0, \sigma^2)$.

2.5.1 Spatial Weighting

Spatial weighting refers to weights that describe the relative positional relationship between one piece of data and another. This weighting is an important component because it reflects the spatial location of the observed data. Weighting in GWR can be done using various kernel function methods. In this study, the Fixed Gaussian kernel function weighting method will be used, where a bandwidth with the same value is applied to each observation location point. The fixed Gaussian kernel function is as follows (Aliu et al., 2022):

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right) \tag{2.9}$$

where:

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \tag{2.10}$$

with:

d_{ij} : Euclidean distance between location (u_i, v_i) and (u_j, v_j),
b : smoothing parameter (bandwidth),
u_i : longitude of the i-th location,
u_j : longitude of the j-th location,
v_i : latitude of the i-th location,
v_j : latitude of the j-th location.

Euclidean distance is the distance between the regression point i and the location j, where $i \neq j$.

The kernel function functions to provide weights based on the optimum bandwidth. The bandwidth refers to the radius of a circle used to determine the weight of the location center point to the observation points in the regression model. Points that are closer to the i-th observation location will be given greater weight in building a regression model at that location. This weight represents the level of influence of the surrounding points on the parameter values at that observation location.

According to Fotheringham et al. (2002), the optimum bandwidth selection can be done using the Cross Validation (CV) method, where the optimum bandwidth value is indicated by the minimum CV value. The CV method is formulated by the following equation:

$$\text{CV} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(b)\right)^2 \tag{2.11}$$

with:

y_i : observation value of the response variable at location i,
$\hat{y}_{\neq i}(b)$: the estimated value of y_i at location (u_i, v_i), which is omitted from the estimation process,
n : number of samples.

2.5.2 Parameter Estimation of GWR Model

Parameter estimation in the GWR model is done using the Weighted Least Square (WLS) method, which involves giving varying weights to each observation location (Hakim et al., 2015). The first step in applying WLS is to form a diagonal matrix that represents different weights for each location i, as follows:

$$\mathbf{W}(u_i, v_i) = \begin{bmatrix} w_{i,1} & 0 & \cdots & 0 \\ 0 & w_{i,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{i,k} \end{bmatrix} \tag{2.12}$$

Parameter estimation is obtained by minimizing the sum of squares of the weighted

errors, assuming that $\mathbf{W}(u_i, v_i) = \mathbf{W}$. Equation (2.8) can then be written in the following form:

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

or in matrix form:

$$\begin{aligned}
\boldsymbol{\varepsilon}^T \boldsymbol{W} \boldsymbol{\varepsilon} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y}^T - \boldsymbol{\beta}^T \boldsymbol{X}^T) \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} \\
&= \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - \boldsymbol{W} (\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{\beta})^T - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} \\
&= \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} \\
&= \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta}
\end{aligned} \tag{2.13}$$

where:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_i, v_i) \\ \beta_1(u_i, v_i) \\ \vdots \\ \beta_p(u_i, v_i) \end{bmatrix}$$

Equation (2.13) is differentiated with respect to $\beta^T(u_i, v_i)$ and the result is equated to zero. Then, the parameter estimator of the GWR model is obtained as follows:

$$
\begin{aligned}
\frac{\partial(\varepsilon^T \boldsymbol{W} \varepsilon)}{\partial \boldsymbol{\beta}^T} &= 0 \\
\frac{\partial(\boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= 0 \\
-2\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} &= 0 \\
2\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} &= 2\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \\
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}(u_i, v_i) \boldsymbol{y}
\end{aligned}
\tag{2.14}
$$

with:

$\mathbf{X}$	: matrix of predictor variables of order $n \times (k+1)$,
$\mathbf{y}$	: response vector of size $n \times 1$,
$\boldsymbol{W}(u_i, v_i)$	: spatial weighting matrix for the GWR model of size $n \times n$.

2.5.3 Parameter Test of GWR Model

Testing the parameters of the GWR model includes simultaneous testing and partial testing. Simultaneous testing aims to determine whether there is a significant difference between the linear regression model and the GWR model. With

H_0: $\beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \cdots = \beta_k(u_i, v_i) = 0$

H_1: at least one $\beta_j(u_i, v_i) \neq 0,\ j = 1, 2, \ldots, k$

$$
F_{\text{count}} = \frac{\left(\mathrm{RSS}_{\mathrm{OLS}} - \mathrm{RSS}_{\mathrm{GWR}}\right)/v}{\mathrm{RSS}_{\mathrm{OLS}}/\delta_1} \tag{2.15}
$$

with

$\mathrm{RSS}_{\mathrm{OLS}}$	: sum of squared errors of the residuals of the OLS model
$\mathrm{RSS}_{\mathrm{GWR}}$	: sum of squared errors of GWR model residuals
δ_1	: degrees of freedom of the GWR model
v	: degrees of freedom of the OLS model

With the test criteria, reject H_0 if $F_{\text{count}} > F_{(\alpha,df_1,df_2)}$, where $\alpha = 0.05$, which means there is a difference between the linear regression model and the GWR model.

Furthermore, partial testing is used to determine the variables that significantly affect the response variable. The partial testing hypothesis is as follows:

$\text{H}_0 : \beta_k(u_i, v_i) = 0$

$\text{H}_1 : \beta_k(u_i, v_i) \neq 0,\ k = 1, 2, \ldots, m;\ i = 1, 2, \ldots, n.$

with a significance level of $\alpha = 5\%$, the test statistic is

$$t_{\text{count}} = \frac{\hat{\beta}_k(u_i, v_i)}{\text{SE}\hat{\beta}_k(u_i, v_i)} \tag{2.16}$$

where $\text{SE}\hat{\beta}_k(u_i, v_i)$ is the standard error of the coefficient $\hat{\beta}_k(u_i, v_i)$. With the decision to reject H_0 if $|t_{\text{count}}| > t_{\text{table}(\alpha/2,df)}$, which means there is an influence between the dependent variable and the independent variable.

2.6 Model Goodness Criteria

In determining the best model, the criteria used in this study are the Akaike Information Criterion (AIC), which is used to measure the relative quality of a statistical model based on available data. Mathematically, AIC is expressed as in the equation (Fathurahman, 2010):

$$\text{AIC} = e^{\frac{2k}{n}} \left(\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n} \right) \tag{2.17}$$

with

k : number of parameters estimated in the regression model
n : number of observations
e : 2.718
u : residual.

2.7 TB Disease

Tuberculosis is a contagious infectious disease caused by the bacteria *Mycobacterium tuberculosis*, which usually affects the lungs in humans. Transmission occurs through BTA-positive patients, who spread the bacteria through small droplets when coughing or sneezing. These bacteria can survive in the air and be inhaled by healthy people, thus causing infection (Anggraeni and Rahayu, 2018). The source of transmission of these infectious diseases is through the air (airborne disease) (Farrell, 2017).

There are several things that are factors of exposure to TB disease, including: (Apriadisiregar et al. (2018); Nafsi and Rahayu (2020); Suhartono et al. (2021)):

1. **Socioeconomic Factors**
   Low socioeconomic levels are often associated with an increased risk of TB. People living in neighborhoods with high levels of poverty tend to have more limited access to health care, poor nutrition, and crowded environments, all of which increase the risk of TB infection. Research shows that poverty, unemployment, and low education are associated with higher TB prevalence.

2. **Gender and Age**
   Adult men tend to be more at risk of developing TB than women, especially in their productive years (15-54 years). This may be due to differences in exposure to high-risk work environments and behaviors such as smoking or alcohol use that are higher among men. In addition, children and the elderly are also vulnerable due to weaker immune systems.

3. **Population Density**
   Living in areas with high population density increases the risk of spreading TB. Dense environments facilitate airborne transmission of bacteria, especially in poorly ventilated places, such as slum housing or prisons.

4. **Unhygienic Environment**
   A dirty and unhealthy environment increases the risk of TB infection as bacteria spread easily in such conditions. Poor housing, poor ventilation, and poor sanitation are important factors in the transmission of TB.

5. **Lack of Access to Medical Care**
   Limited access to health facilities leads to delays in diagnosis and treatment,

which allows the disease to spread further. In remote or poor areas, limited medical services also lead to inappropriate or incomplete treatment, increasing the risk of drug resistance.

6. **Contact with TB Patients**
   People who are often in close contact with people with active TB (such as family members or coworkers) have a higher risk of being infected. Transmission can occur through inhalation of bacteria released by a person with TB when coughing or sneezing.

7. **Lack of Knowledge about TB Disease**
   Lack of knowledge about how TB is transmitted, prevented, and treated leads to people not being aware of the risks and not seeking early medical care. Lack of education also leads to stigmatization of TB sufferers, which results in delays in diagnosis and treatment.

8. **Incomplete Vaccines, Especially BCG Vaccine**
   Bacillus Calmette-Guerin (BCG) vaccination is an effective vaccine in preventing severe tuberculosis in children. However, incomplete vaccination or low vaccination coverage may increase the risk of infection, especially in TB endemic areas.

CHAPTER III

RESEARCH METHODOLOGY

3.1 Research Time and Place

This research was conducted in the even semester of the 2024/2025 academic year at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Lampung University.

3.2 Research Data

The data used in this study are secondary, namely data on TB cases in North Sumatra province in 2022 obtained from the website of the Central Bureau of Statistics of North Sumatra Province and the Sectoral Statistics Book of the Office of Communication and Information of North Sumatra Province. The variables involved in this study are the number of TB cases per 100000 population as the response variable and predictor variables including the number of children immunized with BCG, population density per square (km^2), the number of health workers and the number of health facilities number of health facilities (public hospitals, puskesmas clinics, etc.).

Table 1. Spatial Data Research Variables

Research Variables	**Indicator**	**Description**
Dependent	y	Number of TB Cases
Independent	x_1	BCG Immunization (number of Children)
	x_2	Population Density (people/km^2)
	x_3	percentage of households with access to adequate sanitation
	x_4	number of health workers (people)
	x_5	number of health facilities (unit)
Spatial	(u_i, v_i)	Coordinate Point (latitude, longitude)

3.3 Research Methods

The steps in this research method are as follows:

1. Describe the research variables, namely the dependent variable (Y) and the independent variable (X) which will be used in model building.
2. Detect multicollinearity between predictor variables using the VIF value.
3. Analyzing the regression model, with the following steps:
   (a) Perform parameter estimation of linear regression models with the OLS method.
   (b) Perform assumption test.
   (c) Testing the significance of linear regression model parameters with simultaneous test in equation (2.5) and partial test according to equation (2.6).
4. Detecting spatial heterogeneity with equation (2.7).
5. Analyzing the GWR model with the following steps:
   (a) Calculate the Euclidean distance between observation locations based on latitude and longitude coordinates with equation (2.10).
   (b) Determine the optimum bandwidth for parameter estimation at the i-th observation location, by selecting the minimum CV value based on equation (2.11).
   (c) Calculate the weight matrix with the fixed Gaussian kernel function in equation (2.9).

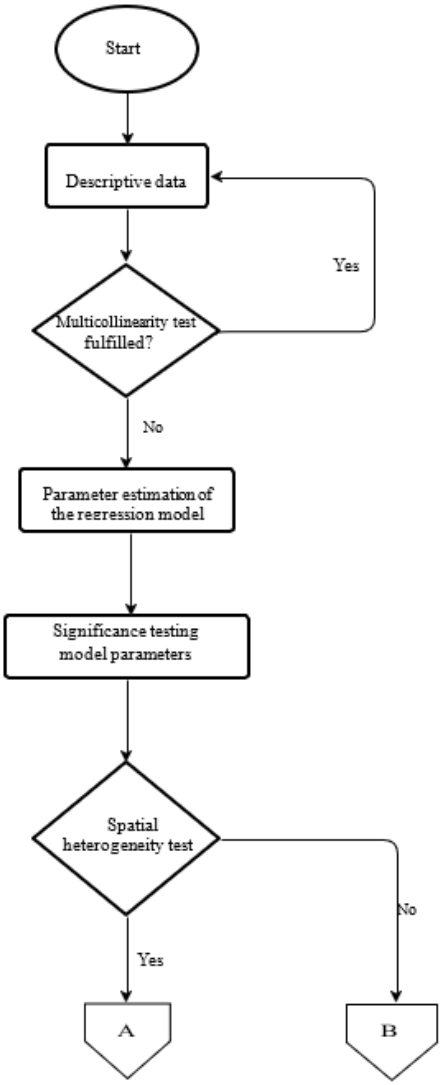
(d) Perform GWR model parameter estimation at the i-th location using the optimum bandwidth obtained in the previous step (b).

(e) Test the significance of GWR model parameters both simultaneously and partially using equations (2.15) and (2.16).

6. Selecting the best model based on the AIC value for each model using equation (2.17).

7. Interpreting results.

The following is a flow chart of the research steps, which is shown in Figure 1.. The analysis process in this study was carried out with the help of the R Program and ArcGIS software.



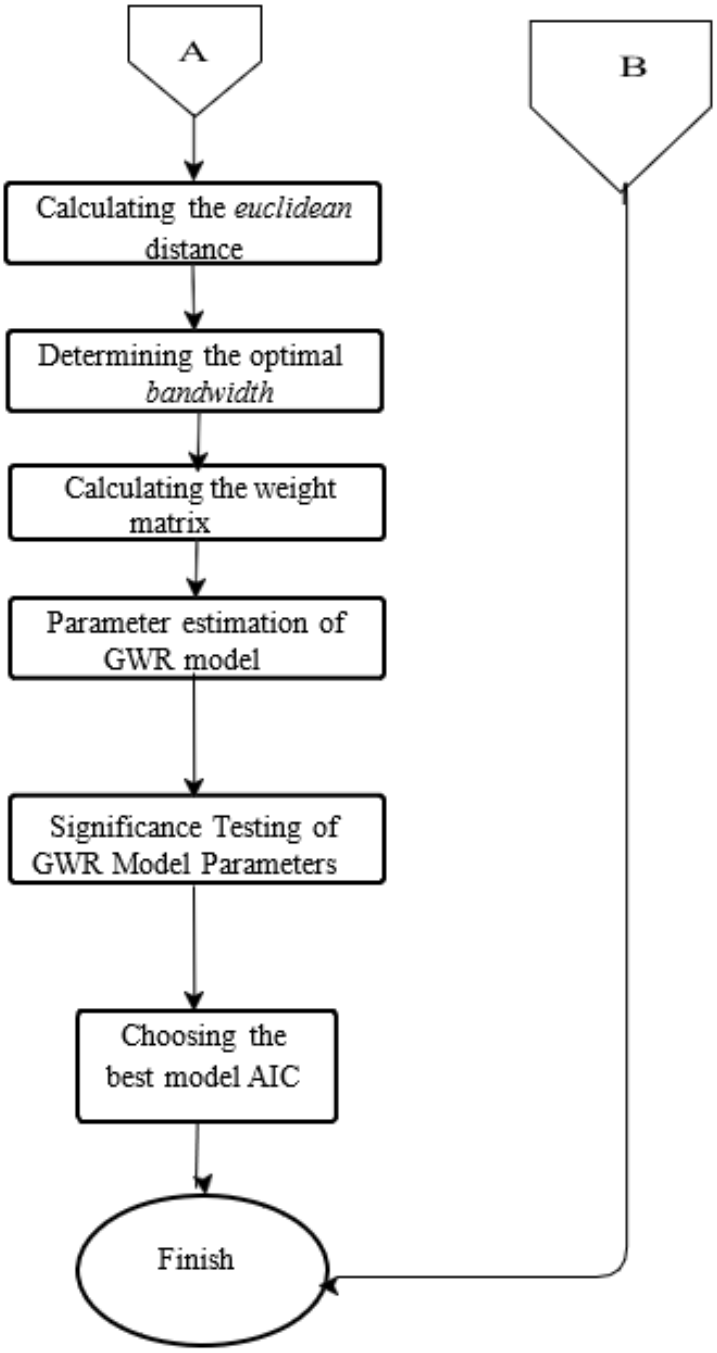


Figure 1. Flow chart of the research steps

CHAPTER V

CONCLUSION

Based on the results of GWR modeling research with fixed gaussian kernel function weights on the number of tuberculosis in the North Sumatera in 2022, it can be concluded that:

1. The results of modeling the number of TB cases in North Sumatra in 2022 using the GWR method with a fixed Gaussian kernel function are:

   $$\hat{Y}_{\text{nias}} = -14.55542 - 0.00027x_1 + 0.06292x_2 + 0.59390x_3 + 0.19263x_4 - 0.38803x_5$$

   $$\hat{Y}_{\text{Asahan}} = -38.48448 - 0.01319x_1 + 0.0377x_2 + 1.43914x_3 + 0.07444x_4 + 0.10814x_5$$

   $$\hat{Y}_{\text{Gunungsitoli}} = -28.748 - 0.00204x_1 + 0.06228x_2 + 0.70635x_3 + 0.20128x_4 - 0.367x_5$$

   Other models can be seen in the Appendix 5.

2. Based on the t-test, it can be concluded that the significant independent variables are divided into eight groups.

REFERENCES

Aliu, M. A., Zubedi, F., Yahya, L., and Oroh, F. A. (2022). The comparison of kernel weighting functions in geographically weighted logistic regression in modeling poverty in indonesia. *Jurnal Matematika, Statistika dan Komputasi*, 18(3):362–384.

Anggraeni, D. E. and Rahayu, S. R. (2018). Gejala klinis tuberkulosis pada keluarga penderita tuberkulosis bta positif. *HIGEIA (Journal of Public Health Research and Development)*, 2(1):92–95.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models (Vol. 4)*. Springer Netherlands.

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, 26(1):19–30.

Apriadisiregar, P., Gurning, F., Eliska, E., and Pratama, M. (2018). Analysis of factors associated with pulmonary tuberculosis incidence of children in sibuhuan general hospital. *Jurnal berkala epidemiologi*, 6(3):268.

Bakri, N. A., Annas, S., and Aidid, M. K. (2024). Pendekatan geographically weighted regression (gwr) untuk menganalisis hubungan pdrb sektor pertanian, kehutanan, dan perikanan dengan faktor pencemaran lingkungan di jawa timur. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, 6(01):11–17.

Basri, H. (2019). Pemodelan regresi berganda untuk data dalam studi kecerdasan emosional. *DIDAKTIKA: Jurnal Kependidikan*, 12(2):103–116.

Cahyono, T. (2015). Statistik uji normalitas. *Yayasan Sanitarian Banyumas, Banyumas, Indonesia*.

Caraka, R. E. and Yasin, H. (2017). Geographically weighted regression (gwr) sebuah pendekatan regresi geografis.

Del Águila, M. R. and Benítez-Parejo, N. (2011). Simple linear and multivariate regression models. *Allergologia et immunopathologia*, 39(3):163.

Farrell, M. (2017). Textbook of medical surgical nursing.

Fathurahman, M. (2010). Pemilihan model regresi terbaik menggunakan akaike's information criterion. *J. EKSPONENSIAL*, 1(2):26–33.

Fotheringham, A., Charlton, M., and Brunsdon, C. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley Sons Ltd, England.

Hakim, A. R., Yasin, H., and Suparti, S. (2015). Pemodelan persentase penduduk miskin di kabupaten dan kota di jawa tengah dengan pendekatan mixed geographically weighted regression. *Jurnal Gaussian*, 3(4):575–584.

In, A. W. K. and Asyik, N. F. (2019). Pengaruh kompetensi dan independensi terhadap kualitas audit dengan etika auditor sebagai variabel pemoderasi. *Jurnal Ilmu dan Riset Akuntansi (JIRA)*, 8(8):6–7.

Long Viet Bui, L. V. B., Mor, Z., Chemtob, D., Son Thai Ha, S. T. H., and Levine, H. (2018). Use of geographically weighted poisson regression to examine the effect of distance on tuberculosis incidence: a case study in nam dinh, vietnam. *Plos One*, 13(11).

Mardiatmoko, G. (2020). Pentingnya uji asumsi klasik pada analisis regresi linier berganda (studi kasus penyusunan persamaan allometrik kenari muda [canarium indicum l.]). *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 14(3):333–342.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (1992). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Nafsi, A. Y. and Rahayu, S. R. (2020). Analisis spasial tuberkulosis paru ditinjau dari faktor demografi dan tingkat kesejahteraan keluarga di wilayah pesisir. *Jurnal Penelitian dan Pengembangan Kesehatan Masyarakat Indonesia*, 1(1):81.

Nurdin, N. N., Raupong, R., and Islamiyati, A. (2014). Penggunaan regresi robust pada data yang mengandung pencilan dengan metode momen. *Jurnal Matematika, Statistika dan Komputasi*, 10(2):115–116.

Octavianty, O., Toharudin, T., and Jaya, I. (2017). Geographically weighted poisson regression semiparametric on modeling of the number of tuberculosis cases

(case study: Bandung city). In *AIP Conference Proceedings*, volume 1827. AIP Publishing.

Rokom (2023). Deteksi tbc capai rekor tertinggi di tahun 2022.

Sholihah, S. M., Aditiya, N. Y., Evani, E. S., and Maghfiroh, S. (2023). Konsep uji asumsi klasik pada regresi linier berganda. *Jurnal Riset Akuntansi Soedirman*, 2(2):103.

Suhartono, S., Raharjo, M., et al. (2021). Faktor-faktor yang mempengaruhi kejadian tuberkulosis: Sebuah review. *Sanitasi: Jurnal Kesehatan Lingkungan*, 13(1):20–25.

Wei, W., Yuan-Yuan, J., Ci, Y., Ahan, A., and Ming-Qin, C. (2016). Local spatial variations analysis of smear-positive tuberculosis in xinjiang using geographically weighted regression model. *BMC Public Health*, 16:1–9.