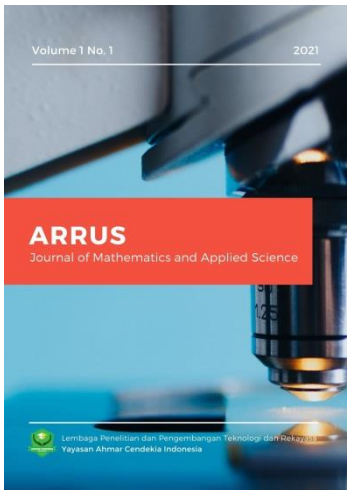
https://doi.org/10.35877/mathscience3396

OPEN ACCESS

ISSN [Print]: 2776-7922



RESEARCH ARTICLE

Geographically Weighted Regression with Bi-Square Kernel Weights for Life Expectancy Data in East Java Province

Sitti Masyitah Meliyana[1*], R. Rusli [2], Abdul Rahman[2]

[1]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, Indonesia.
[2]Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, Indonesia.

*Corresponding author: Sitti Masyitah Meliyana, Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, Indonesia

E-mail: sittimasyitahmr@unm.ac.id

Abstract: This study explores the application of Geographically Weighted Regression (GWR) using a Bi-Square Kernel weighting function to analyze life expectancy data across East Java Province. By incorporating spatial heterogeneity, the GWR model provides more accurate and localized insights compared to traditional global regression models. The results indicate significant spatial variability in the effects of poverty rate, healthcare facilities, sanitation, health complaints, and immunization coverage on life expectancy. Based on the analysis of life expectancy estimates in regencies/cities of East Java Province, the Madura region exhibits lower life expectancy compared to other areas, with Bangkalan Regency having the lowest life expectancy at 61.43 years. Additionally, urban areas generally have higher life expectancy than rural areas, with Surabaya City recording the highest life expectancy in East Java at 72.03 years. This disparity can be attributed to differences in the quality of healthcare services and better access to healthcare in urban areas compared to rural ones.

Keywords: Geographically Weighted Regression, Bi-Square Kernel, Life Expectancy, Spatial Analysis, East Java.

1. INTRODUCTION

Life expectancy is a crucial indicator for evaluating government performance in improving public welfare and health. WHO data shows that life expectancy in Indonesia has been steadily increasing, from 20.7 million people aged over 60 in 2010 to 36 million, with a projected rise to 71 million by 2050. However, this growth remains relatively small compared to the total population, particularly in rural areas where awareness of healthy living practices is limited.

Factors such as poverty levels and the availability of healthcare facilities significantly impact life expectancy. Global analyses, such as ordinary regression, are often used to study these relationships but fail to account for spatial differences. Geographically Weighted Regression (GWR) offers a more suitable approach to addressing spatial heterogeneity. GWR calculates regression parameters locally based on geographic locations, using a weighting matrix dependent on the distance between observations (Charlton & Fotheringham, 2009). The bi-square kernel function is commonly used in GWR to provide optimal weighting based on distance.

https://doi.org/10.35877/mathscience3396

OPEN ACCESS

ISSN [Print]: 2776-7922

This study focuses on East Java, a region with relatively high life expectancy. GWR is applied to build local models and evaluate whether they outperform global models in explaining life expectancy variability across regencies and cities in East Java. The primary objectives are to construct a GWR model using bi-square kernel weighting and compare the performance of local and global models.

2. Literature Review

2.1. *Regression Analysis*

The regression equation, commonly defined using the parameter estimation method of Ordinary Least Squares (OLS), can generally be written as follows:

$$y_i = b_0 + \sum_{k=1}^{p} x_{ik} b_k + e_i \quad , \qquad i = 1, 2, 3, \ldots, n \tag{2.1}$$

Where b_0 is the constant, x_{ik} is the value of the k-th explanatory variable for the i-th observation, b_k is the coefficient of the explanatory variable of $x_{.k}$, p is the number of explanatory variables used in the model, nnn is the number of observations (samples), and e_i is the random error for the i-th observation. The random error vector $e = (e_1, e_2, e_3 \ldots\ldots, e_n)'$ is assumed to be distributed as $N(0, \sigma^2 I)$.

2.2. *Geographically Weighted Regression*

Geographically Weighted Regression (GWR) is a locally linear regression model that produces local parameter estimates for each observation location. Using the Weighted Least Squares (WLS) method, the parameter estimate at the i-th location is formulated as follows:

$$\hat{\mathbf{b}}(i) = (X'\tilde{W}(i)X)^{-1}X'W(i)Y \tag{2.2}$$

With $W(i) = diag[w_1(i), w_2(i), \ldots., w_n(i)]$, and $0 \le w_j(i) \le 1 \, (i, j = 1,2,\ldots., n)$. $W(i)$ is a n×n diagonal matrix, representing the spatial weighting matrix for the i-th location. The values of its diagonal elements are determined by the proximity of the i-th observation (location) to other locations (j-th locations). The closer the locations, the higher the weight assigned to the corresponding element.Association.

According to Fotheringham et al. (2002), several weighting functions in spatial analysis include:

- $w_j(i) = 1$ for all i and j. he GWR model with this weighting will result in an ordinary regression model, where each data point at all locations is assigned the same weight, which is 1, regardless of its position or distance from other locations.
- $w_j(i) = 1$, if $d_{ij} < d$ and $w_j(i) = 0$ for $d_{ij} \ge d$. The value d represents the minimum distance between locations beyond which they are considered to have no influence on each other. If the distance from i-location to j-location is less than d($d_{ij} < d$), all data at that location are used and given the same weight, which is 1.
- $w_j(i) = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]$ Where d_{ij} is the distance from i-location to j-location, and b is the bandwidth, a smoothing parameter that is always positive. This function is commonly referred to as the normal kernel function (Gaussian).

https://doi.org/10.35877/mathscience3396

OPEN ACCESS

ISSN [Print]: 2776-7922

- $w_j(i) = \exp\left[1-\left(\frac{d_{ij}}{b}\right)^2\right]^2$ if $d_{ij} < d$, and $w_j(i) = 0$ for $d_{ij} \geq d$. This function follows the form of the biweight kernel and is commonly referred to as the bi-square kernel weighting function.
- $w_j(i) = \exp\left(-\frac{R_{ij}}{b}\right)$, with Where R_{ij} is the rank of the distance from i-location to j-location (j=1,2,3,…,n). The closest distance results in $w_j(i)$ approaching 1, and the value decreases as the distance between i-location and j-location increases.

In this study, one of the weighting functions involving the distance between villages, the bi-square kernel, is used. A weighting function that incorporates continuous distance is selected, as it is expected to produce a model with better smoothing.
To globally detect the superiority of the GWR model over Ordinary Least Squares (OLS) regression for the case data used, an Analysis of Variance (ANOVA) test, as proposed by Brunsdon et al. (1999), can be applied as follows:

$$F_{hit} = \frac{(JKG_{OLS} - JKG_{RTG})/v_1}{JKG_{RTG}/\delta_1} \tag{2.3}$$

Where JKG_{OLS} is the sum of squared errors from the OLS model, and JKG is the sum of squared errors from the GWR model. The F_{hit} will approximate an F-distribution with degrees of freedom db1 = $\frac{v_1^2}{v2}$ and $db_2 = \frac{\delta_1^2}{\delta_2}$, where $\delta_1 = tr[(1-S)'(1-S)]^i, i = 1,2$. The parameter v_1 is defined as $n - p - 1 - \delta_1$ and $v2$ is $n - p - 1 - 2\delta_1 + \delta_2$, where S is the hat matrix of the GWR model. A small F_{hit} value supports the acceptance of the null hypothesis, indicating that the GWR and OLS models are equally effective in explaining the relationship between variables. At a significance level of α, the null hypothesis will be rejected if:

$$F_{hit} > F_{\alpha}\left(\frac{v_1^2}{v2}, \frac{\delta_1^2}{\delta_2}\right) \tag{2.4}$$

2.3. *The Selection of Bandwidth*

The selection of bandwidth has a significant impact on the results obtained from GWR, acting as a smoothing parameter. A wider bandwidth leads to better smoothing. Over-smoothing in the model results in parameters that are constant across the study area, while under-smoothing produces parameters with excessive local variation, making it difficult to determine their form. The optimal bandwidth strikes a balance between these two extremes. There are three methods for selecting bandwidth, as follows:

- Direct Bandwidth Selection

  Determining the bandwidth size directly is the simplest method but is only applicable when a suitable bandwidth value has already been identified. Alternatively, a strong theoretical basis is needed to determine the bandwidth.

- Selecting Bandwidth that Minimizes the Cross-Validation (CV) Function

https://doi.org/10.35877/mathscience3396

ISSN [Print]: 2776-7922

Cross-validation is one criterion for obtaining the optimal bandwidth. The optimal bandwidth is the one that minimizes the cross-validation coefficient (CV), calculated using the formula:

$$CV = \sum_{i=1}^{n}\left[y_i - \hat{y}_{\neq i}(b)\right]^2 \quad (2.5)$$

Where $\hat{y}_{\neq i}(b)$ is the predicted value y_i (fitting value) for i-observation when the data from i-location is excluded from the prediction process (Fotheringham et al., 2002). The optimal bandwidth is obtained through an iterative process to minimize the CV.

- Selecting Bandwidth that Minimizes the Akaike Information Criterion (AIC)
  Adjusting the bandwidth alters the degrees of freedom in the model. If cross-validation is used as the criterion for bandwidth selection, the score for each bandwidth corresponds to a slightly different model. An alternative criterion for bandwidth selection is the Akaike Information Criterion (AIC). The most appropriate bandwidth is obtained by minimizing the AIC value. AIC accounts for differences in the degrees of freedom between models, making it more accurate for comparison. A model with a lower AIC value is considered better than one with a higher AIC. The AIC is calculated as follows:

$$\mathrm{AIC}_c = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left\{\frac{n + \mathrm{tr}(S)}{n - 2 - \mathrm{tr}(S)}\right\} \quad (2.6)$$

3. Research Method and Materials

The data used in this paper is secondary data regarding the factors that are suspected to influence life expectancy in East Java. There are six variables used, as follows:

Y: Life Expectancy (years)
X1: Number of poor population (people)
X2: Number of healthcare facilities (units)
X3: Percentage of population with access to sanitation facilities (units)
X4: Percentage of health complaints (%)
X5: Percentage of children under five receiving immunizations (%)

This paper will analyze whether these five independent variables influence the level of life expectancy in East Java. The analysis method used is the Geographically Weighted Regression (GWR), so that each district/city in East Java has its own regression model to represent its life expectancy.

4. Results and Discussion

4.1. Check the data distribution

The data distribution needs to be checked first to determine whether the global model used is Classical Linear Regression or Poisson Regression. The results are as follows:

Table 1: Kolmogorov-Smirnov Test

	Y
Kolmogorov-Smirnov Z	1.220
Asymp. Sig. (2-tailed)	0.102

One-sample KS test hypothesis:

H0: The data follows a normal distribution
H1: The data does not follow a normal distribution

https://doi.org/10.35877/mathscience3396

ISSN [Print]: 2776-7922

Conclusion:

Using a 95% confidence level, there is not enough evidence to reject H0 because the p-value = 0.102 > alpha = 0.05. Therefore, the life expectancy data follows a normal distribution. Thus, the appropriate spatial global model for this case is Classical Linear Regression.

4.2. Multicollinearity

After the data is entered, the assumption that there is no multicollinearity between the independent variables is checked. This is to avoid high bias in the model. Independent variables are considered to be correlated with each other if the correlation value is greater than 0.7. Based on the analysis, there are no independent variables that are correlated with each other, thus all independent variables can be used in the regression modeling.

4.3. Spatial Heterogeneity

A regression model should not contain homogeneity of variance between its observations, in this case, spatial homogeneity of variance. This can be tested using the Breusch-Pagan test, with the following results:

Table 2: Breusch-Pagan Test

Breusch-Pagan Test		
BP=13.9884	Df=5	p-value=0.01568

Breusch-Pagan test hypothesis:

H0: The variance between regions is equal

H1: The variance between regions is not equal

Conclusion:
Using a 95% confidence level, there is sufficient evidence to reject H0 because the p-value = 0.01568 < alpha = 0.05. Therefore, we can be confident that there is no spatial homogeneity of variance in the data.

4.4. Global Regression Model

After the assumption testing phase, the next step is to form the global model using geographically weighted regression. The global model here is similar to classical regression, with each city/district treated as an observation. The lm-test package is used in the spatial global regression analysis.

The formed global regression model is as follows:

$$\hat{y} = 50.305359 - 0.010442x_1 + 0.049656x_2 + 0.078070x_3 - 0.037472x_4 + 0.200975x_5$$

If a 90% confidence level is used, only the variable x3 is not significant. Based on the overall F-statistic test, the model is concluded to be significant because the p-value = 2.02E-5 < alpha = 0.05.

Since it is very difficult to obtain a truly perfect model in research, an adjusted R-squared of 51.96% is considered quite good. The AIC of 155.9328 is also regarded as satisfactory. Overall, the spatial global model for life expectancy in East Java cities/districts can be considered good.

4.5. Geographically Weighted Regression with Bisquare Kernel Weight

After obtaining the global model, the next step is to find the local geographically weighted regression model using the Bisquare Kernel model for each city/district. Based on the R analysis results for the geographically weighted regression model on life expectancy data in East Java, the model can be considered quite good with an AIC value of 135.5517.

https://doi.org/10.35877/mathscience3396

OPEN ACCESS

ISSN [Print]: 2776-7922

The formed local GWR with bisquare kernel weight model consists of regression models for each district/city, where each area has a different and unique model.

$$\hat{y} = \widehat{\beta_0}(u_i, v_i) + \widehat{\beta_1}(u_i, v_i)x_1 + \widehat{\beta_2}(u_i, v_i)x_2 + \widehat{\beta_3}(u_i, v_i)x_3 + \widehat{\beta_4}(u_i, v_i)x_4 + \widehat{\beta_5}(u_i, v_i)x_5$$

Where,
i: district/city area in East Java

(u_i, v_i): latitude and longitude coordinates for each district/city area

Example of a local model for Pacitan District:

$$\hat{y} = 49.46359 - 0.00544x_1 + 0.030454x_2 + 0.069332x_3 + 0.027258x_4 + 0.16544x_5$$

Example of a local model for Surabaya City:

$$\hat{y} = 48.88773 - 0.00769x_1 + 0.048365x_2 + 0.083198x_3 + 0.02093x_4 + 0.198936x_5$$

Here is the map showing the visualization of the estimated life expectancy in the districts/cities of East Java province:

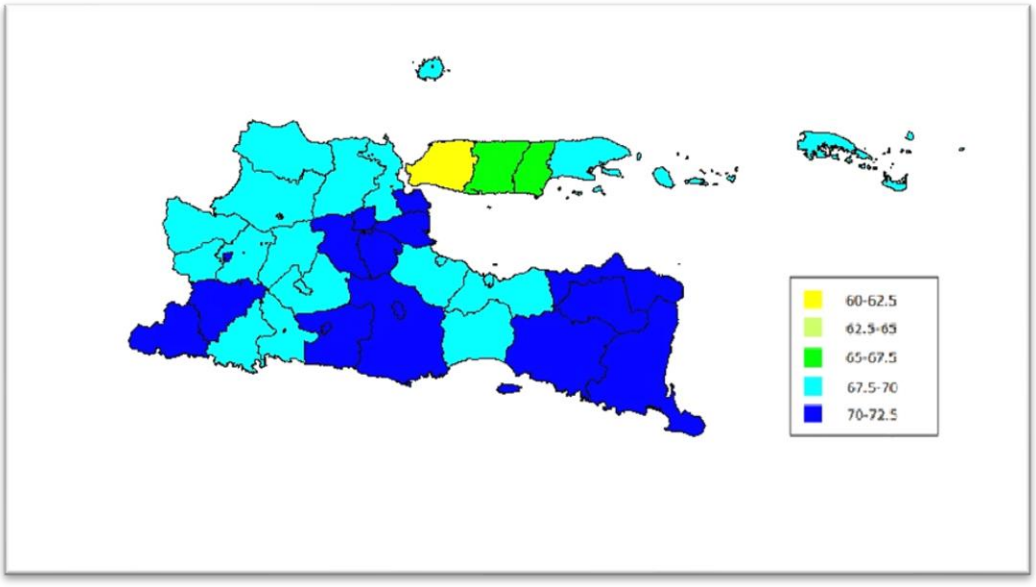


Figure 1. Peta Visualisasi Hasil Estimasi Model GWR with Kernel Bi-Square

The map above shows the visualization of the estimated life expectancy in the districts/cities of East Java province using the local Geographically Weighted Regression (GWR) model. The yellow color indicates areas with life expectancy between 60-62.5 years, light green for areas with life expectancy between 62.5-65, green for areas with life expectancy between 65-67.5, light blue for areas with life expectancy between 67.5-70, and blue for areas with life expectancy between 70-72.5. It can be observed that the Madura region (Bangkalan, Sampangan, and Pamekasan districts) has a lower life expectancy compared to other areas. This can serve as a basis for the local government to improve healthcare services in Madura. Additionally, urban areas generally have higher life expectancy compared to rural areas. This can be understood due to differences in healthcare quality and better access to healthcare in urban areas compared to rural areas. In this case, Surabaya city is the area with the highest life expectancy in East Java, which is 72.03 years, while Bangkalan district has the lowest life expectancy, which is 61.43 years.

https://doi.org/10.35877/mathscience3396

OPEN ACCESS

ISSN [Print]: 2776-7922

5. Conclusion

From the analysis conducted on the Life Expectancy data in the districts/cities of East Java, several conclusions can be drawn is The Local Geographically Weighted Regression (GWR) model has proven to be better than the Global (Classical Regression) model for this data, allowing each district/city in East Java to have its own model in spatially modelling life expectancy data. The map showing the estimation of life expectancy in the districts/cities of East Java reveals that the Madura region has a lower life expectancy compared to other areas. Additionally, urban areas generally have higher life expectancy than rural areas. This can be understood due to differences in healthcare quality and better access to healthcare in urban areas compared to rural areas.

References

Bivand, R. and Yu, D. 2012. Package 'spgwr': Geographically Weighted Regression. CRAN R Project.

Brunsdon C, Fotheringham AS, Chartlon M. 1999. Some notes on parametric significance tests for geographically weighted regression, Journal of Regional Science, Vol. 39, No 3, 497- 524.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. 2002. Geographically Weighted Regression. John Wiley & Sons: West Sussex.

Rahmawati, R. 2010. Regresi Terboboti Geografis Dengan Pembobot Kernel Kuadrat Ganda Untuk Data Kemiskinan Kabupaten Jember. Bogor: Institut Pertanian Bogor.