World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Geographically weighted regression with fixed kernel tricube on life expectancy in East Java

Hanin Ruliyani * and Kuntoro

Department of Epidemiology, Biostatistics, Population Study, and Health Promotion, Faculty of Public Health, Airlangga University, Surabaya, East Java, Indonesia.

World Journal of Advanced Research and Reviews, 2023, 18(03), 1560–1566

Publication history: Received on 16 May 2023; revised on 27 June 2023; accepted on 30 June 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.18.3.1290

Abstract

Violation of the homoscedasticity assumption will cause biased and inefficient parameter estimation. Geographically Weighted Regression (GWR) is a spatial regression analysis that will produce local parameter estimates. Data on life expectancy for East Java Province in 2021 shows that there is a violation of the homoscedasticity assumption. The aim of the study is to analyze Geographically Weighted Regression with fixed kernel tricube for life expectancy in East Java Province in 2021. This study use a non-reactive research design using secondary data. The results show that the GWR model is different from the OLS model and produces four regional groups which are divided based on significant predictor variables. Three significant predictor variables are the percentage of households that have access to proper sanitation (X_1), the coverage of exclusive breastfeeding for infants aged less than 6 months (X_2), and the percentage of monthly expenditure per capita for food (X_5). Based on this study, group mapping is obtained showing that areas in the same group tend to be close together and parameter estimates that are local to the observation locations.

Keywords: Fixed Kernel Tricube; Geographically Weighted Regression; Life Expectancy; Spatial

1. Introduction

Regression analysis is an analysis to determine the effect of predictor variables on response variables that have been commonly used in various fields. The most common method of performing regression analysis is Ordinary Least Squares (OLS). One of the assumptions that must be met and emphasized in the OLS regression analysis is the homogeneous error variance across all observations or homoscedasticity. If the homoscedasticity assumption is violated, the resulting estimate will be biased, inconsistent, and inefficient which cannot be categorized as a Best Linear Unbiased Estimator (BLUE). The condition of non-constant error variance or heteroscedasticity will also cause bias in the standard error and statistical test results [1,2].

Geographically Weighted Regression (GWR) is the development of a linear regression method that takes into account spatial effects, namely spatial heterogeneity that occurs in spatial data with heteroscedasticity. The GWR method will produce parameter values that are local so that they only apply to a certain observation location and are different from parameters at other locations [3]. GWR analysis uses the Weighted Least Squares (WLS) method, namely the provision of spatial weights in calculating parameter estimates. The amount of spatial weight can be different at each location depending on the proximity of the distance between the two locations. The closer the distance will cause the higher spatial weight.

Life expectancy is an estimate of the average number of years that a person can live during his life at birth as a reflection of the government's performance in improving the welfare of the population, especially in the aspect of public health status [4]. In 2021, East Java Province is ranked as the 10th province with the highest life expectancy in Indonesia, which

* Corresponding author: Hanin Ruliyani

is 71.38 years. East Java Province has district/cities with different geographical characteristics and success in the development of the health sector. This can be seen from the area of the district which is larger than the area of the city, there are districts that have islands, as well as disparities in supporting of the health system. Preliminary studies conducted by researchers found that data on life expectancy for East Java Province in 2021 violated the assumption of homoscedasticity so that it has a spatial heterogeneity effect. Thus, the research aims to determine Geographically Weighted Regression with fixed kernel tricube on life expectancy in East Java Province in 2021.

2. Material and methods

The research design is a nonreactive research or unobtrusive research using secondary data in examining the relationship between predictor variables and response variables. Secondary data in this research are publication of the 2021 East Java Province Health Profile from the East Java Provincial Health Office and East Java Province in Figures for 2022 from the Central Bureau of Statistics for East Java Province. The unit of observation in this study is district/city in East Java Province. A total of 38 district/cities within the province of East Java is involved in the analysis in this study.

The response variable in the study is life expectancy. Five predictor variables in the study is the percentage of households that had access to proper sanitation (X_1), the coverage of exclusive breastfeeding for infants aged less than 6 months (X_2), the ratio of active posyandu (integrated healthcare center) per 100 children under five (X_3), percentage of infant completeness of immunization (X_4), and the percentage of monthly expenditure per capita for food (X_5). Data is analyzed using the Geographically Weighted Regression (GWR) analysis method with a fixed kernel tricube weighting function. Spatial-based analysis requires coordinate points from observational data. The location coordinates in longitude and latitude in this study are taken from the coordinates of the regent or mayor's office for each district/city in East Java Province obtained from Google Maps. The results of the accuracy of the prediction model from the GWR analysis method will be compared with the prediction. Model of the Ordinary Least Squares (OLS) regression analysis method using the coefficient of determination (R^2) and Akaike Information Criterion (AIC). The model with the largest R^2 and the smallest AIC is the best predictive model. All stages of data analysis is performed using RStudio 4.2.2 software.

3. Results and discussion

Descriptive analysis is carried out to describe the condition of the data in each research variable. The results of the descriptive analysis are shown in table 1 below.

Table 1 Descriptive analysis

Variable	Mean	Minimum	1st Quartile	Median	3rd Quartile	Maximum	SD
Y	71.72	66.89	70.44	72.45	72.84	74.18	1.97
X_1	81.97	39.44	75.81	86.15	90.96	97.31	13.29
X_2	72.81	42.10	64.30	76.35	83.66	92.20	14.32
X_3	1.39	0.54	1.08	1.38	1.66	2.14	0.42
X_4	88.27	40.60	86.42	94.20	97.05	105.50	15.16
X_5	51.27	37.82	46.85	51.94	55.72	64.43	6.19

Descriptive data analysis shows that the distribution of data for each variable varies. The level of data diversity can be seen from the value of the standard deviation (SD). The variable life expectancy (Y), the ratio of active posyandu (integrated healthcare center) per 100 children under five (X_3), and the percentage of monthly expenditure per capita for food (X_5) have a low SD in the research variable data so that it shows lower data variability than the percentage of households with access proper sanitation (X_1), coverage of exclusive breastfeeding in infants aged less than 6 months (X_2), and percentage of infant completeness of immunization (X_4).

3.1. Ordinary Least Squares (OLS) Regression Model

The results of the parameter significance test simultaneously obtained an F value of 9.198 ($> F_{0.05(5;32)} = 2.5123$) dan p-value sebesar 0,0000167 ($< \alpha = 0.05$). Thus, H_0 is rejected so that it can be concluded that the predictor variable simultaneously influences the response variable.

The next test is a partial parameter significance test which aims to determine the predictor variable specifically that affects the response variable shown in table 2. The predictor variable is a significant variable if it has a value |t| greater than $t_{0.025;32}$ (2.036933) and the p-value is smaller than α (0.05). Based on table 2, the variable percentage of households with access to proper sanitation (X_1) and the coverage of exclusive breastfeeding in infants aged less than 6 months (X_2) are significant predictor variables for the variable life expectancy with the linear regression equation as follows.

$Y = 71.36584 + 0.07316X_1 - 0.04696X_2$

Table 2 Results of the partial parameter significance test of the OLS regression model

Estimated Parameter	Coefficient	t	p-value
$\widehat{\beta_0}$	71.36584	15.923	2×10^{-16}
$\widehat{\beta_1}$	0.07316	2.767	0.00932
$\widehat{\beta_2}$	-0.04696	-2.699	0.01103
$\widehat{\beta_3}$	0.60028	1.028	0.31186
$\widehat{\beta_4}$	0.01824	1.094	0.28212
$\widehat{\beta_5}$	-0.09108	-1.737	0.09204

The coefficient of determination (R^2) of the OLS regression model is 58.97%, meaning that the predictor variables in the model can explain 58.97% of the variation in the life expectancy variable in East Java Province, while the remaining 41.03% is explained by other predictor variables not involved in the study.

3.2. Assumption tests

The assumption tests carried out for the OLS linear regression model include error normality tests, multicollinearity, homoscedasticity, and autocorrelation. The results of the Shapiro-Wilk test are at a significance level of 0.05 with a p-value of 0.4493 so that it can be concluded that H_0 is accepted, meaning that the model meets the assumption of normal error. The multicollinearity test on predictor variables in the regression model using the Variance Inflation Factor (VIF) method shows that the VIF values for all predictor variables are less than 10 so that the regression model fulfills the assumption that there is no multicollinearity between predictor variables.

Table 3 VIF values of predictor variables

Variable	VIF Value
X_1	2.468264
X_2	1.242311
X_3	1.210108
X_4	1.279072
X_5	2.108319

The assumption of homoscedasticity is tested using the Breusch-Pagan test with a significance level of 0.05. The p-value in the Breusch-Pagan test is 0.01335 so the test decision, namely H_0, is rejected, meaning that the error variance in each district/city in East Java Province is different so that the data in the modeling has a spatial heterogeneity effect. Regression analysis of the OLS method is not appropriate for analyzing data with spatial heterogeneity effects. The assumption of autocorrelation is the condition of the existence of a correlation to the error in one variable. The results of the Durbin-Watson test showed a p-value of 0.576 (> α = 0.05) so that H_0 is accepted meaning that there is no autocorrelation in the research variables. For spatial data, a spatial autocorrelation test is required using Moran's I to

detect the presence of one of the spatial effects, namely spatial dependencies. Moran's I test results showed a p-value of 0.08283 (> α = 0.05) so it is concluded that there is no spatial autocorrelation.

3.3. Geographically Weighted Regression (GWR) Model with Fixed Kernel Tricube

The tricube fixed kernel weighting function is shown in the following equation.

$$w_{ij}(u_i, v_i) = \begin{Bmatrix} \left[1 - \left(\frac{d_{ij}}{h}\right)^3\right]^3, for\ d_{ij} \leq h \\ 0, for\ d_{ij} > h \end{Bmatrix}$$

d_{ij} is the Euclidean distance which is the distance between the observation areas. The smaller the Euclidean value, the closer the distance between regions is. h is the optimum bandwidth obtained from the Cross Validation (CV) method. The optimum bandwidth for the GWR model with a fixed tricube kernel weight function is 3.013901 with a CV value of 92.39465. The optimum bandwidth value is the same for each observation location because the function used is a fixed kernel. The closer the distance between an area and the point of observation, the greater the spatial weight given by the area to the point of observation. This causes the weight of influence spatially for each location to potentially differ depending on the Euclidean distance. The results of parameter estimation of the GWR model with fixed tricube kernel weights are descriptively explained in table 4.

Table 4 Parameter estimation descriptive analysis

Estimated Parameter	Minimum	1st Quartile	Median	3rd Quartile	Maximum
$\widehat{\beta_0}$	71.722	72.084	72.482	73.688	70.808
$\widehat{\beta_1}$	-0.00009	0.04766	0.06550	0.07629	0.08782
$\widehat{\beta_2}$	-0.06206	-0.05358	-0.04533	-0.03681	-0.02013
$\widehat{\beta_3}$	0.34367	0.49345	0.52525	0.55826	0.57275
$\widehat{\beta_4}$	0.01363	0.01951	0.02075	0.02296	0.02584
$\widehat{\beta_5}$	-0.13676	-0.11476	-0.10481	-0.10176	-0.09810

The variable percentage of households with access to proper sanitation (X_1) has a negative or positive effect on the response variable in several district/cities in East Java Province in the range of -0.00009 to 0.08782. Variable coverage of exclusive breastfeeding in infants aged less than 6 months (X_2) and the percentage of monthly expenditure per capita for food (X_5) in all district/cities have a negative effect on life expectancy. Meanwhile, the variable ratio of active posyandu (integrated healthcare center) per 100 toddlers (X_3) and percentage of infant completeness of immunization (X_4) has a negative effect on life expectancy in all district/cities in East Java Province.

GWR model with fixed tricube kernel weights through goodness of fit test and partial parameter significance test. The results of the F test as a goodness of fit test method show an F value of 2.6026 (> $F_{0.05(14.293;30.716)}$ = 1.893219) and p-value 0.01285 (< α = 0.05) so that it can be concluded that there is a difference between the GWR model using the fixed kernel tricube weights and the OLS method regression model.

The results of establishing the GWR model with a fixed kernel tricube weighting function classify district/cities in East Java Province into four groups based on significant predictor variables. Group four is the group with the most district/cities as well as the group with the most significant variables, namely the percentage of households that have access to proper sanitation (X_1), the coverage of exclusive breastfeeding for infants aged less than 6 months (X_2), and the percentage of monthly expenditure per capita for food (X_5). Group three has the fewest members with a total of four districts.

Table 5 Grouping district/city based on significant variables

Group	Significant Variable	District/City
1	X_5	Pacitan District, Ponorogo District, Trenggalek District, Madiun District, Magetan District, Ngawi District, and Madiun City
2	X_1 dan X_2	Malang District, Banyuwangi District, Pasuruan District, Malang City, Probolinggo City, Pasuruan City, and Batu City
3	X_2 dan X_5	Tulungagung District, Nganjuk District, Bojonegoro District, and Tuban District
4	X_1, X_2, dan X_5	Blitar District, Kediri District, Lumajang District, Jember District, Bondowoso District, Situbondo District, Probolinggo District, Sidoarjo District, Mojokerto District, Jombang District, Jombang District, Lamongan District, Gresik District, Bangkalan District, Sampang District, Pamekasan District, Sumenep District, Kediri City, Blitar City, Mojokerto City, and Surabaya City

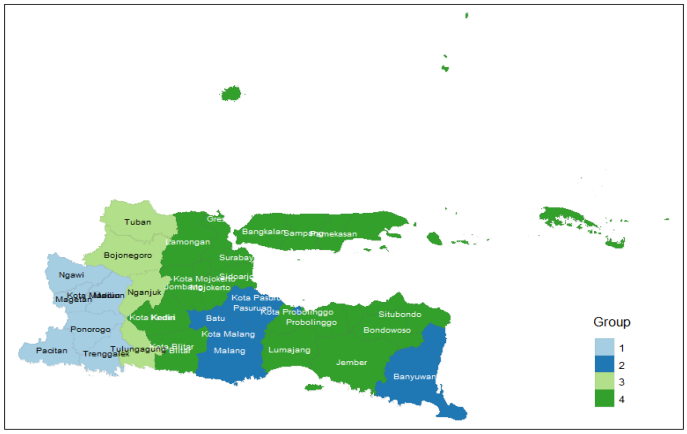

Figure 1 Distribution of district/city groups based on significant variables

The distribution of district/cities based on significant variables is shown in Figure 1. District/cities that are in the same group tend to be in adjacent locations so that it can be seen that areas with certain significant variables can affect the characteristics of adjacent areas so that significant variables in the region are the same even though they have different regression coefficients. One example of GWR modeling results with fixed kernel tricube weights is the following GWR model in Surabaya City.

$$Y = 72.26188 + 0.07024X_1 - 0.04892X_2 - 0.10232X_5$$

Life expectancy in Surabaya City is influenced by three predictor variables. The percentage of households that have access to proper sanitation (X_1) has an effect on increasing life expectancy in the Surabaya City by 0.07024 years for every 1% increase in variables with other variables being constant. The coverage of exclusive breastfeeding in infants aged less than 6 months (X_2) and the percentage of monthly expenditure per capita for food (X_5) have a negative effect on life expectancy in Surabaya City. Every 1% increase in the coverage of exclusive breastfeeding in infants aged less than 6 months can reduce the life expectancy of Surabaya City by 0.04892 years when other variables are held constant. The percentage of monthly expenditure per capita for food has an effect on decreasing life expectancy in Surabaya City by 0.10232 years for every 1% increase with other variables being constant. The local coefficient of determination (R^2) for the Surabaya City is 65.07% meaning that 65.07% of the life expectancy variable for the Surabaya City can be explained by significant predictor variables while the other 34.93% is explained by variables outside the study.

3.4. Best model

Selection of the best model is needed to find out the model with the greatest opportunity to fit the data. The best model selection method is to use AIC and R^2.

Table 6 AIC and R^2 values of OLS and GWR models

Model	AIC	R^2
OLS Regression	138.6841	58.97%
GWR Fixed Kernel Tricube	125.0357	66.02%

The model is decided to be the best model if it has a low AIC value and high R^2. Thus, the GWR model with fixed tricube kernel weights function is a better model than the regression model with the OLS method. This is supported by the GWR method with a fixed tricube kernelwhich can increase R^2 by 8.85% compared to the OLS method and reduce the AIC value by around 13.65.

4. Discussion

Geographically Weighted Regression (GWR) analysis method has a fundamental difference from the Ordinary Least Squares (OLS) regression analysis method. The difference lies in the use of different spatial weights based on the proximity between a location and the center of the regression location in calculating parameter estimates. The goodness of fit test of the GWR model with the fixed kernel tricube weighting function for modeling life expectancy in East Java Province shows that there is a statistical difference between the GWR model with the fixed kernel tricube weighting function and the OLS regression model. The results of this study are not in line with research in Central Java Province which explains that there is no difference between the OLS regression model and the GWR model using fixed kernel tricube weighting function in modeling poverty [5]. The significant difference between the OLS model and the GWR model is caused by the presence of spatial weights with a range from 0 to close to 1 which are different at each location and cause the F statistic to get bigger and the p-value to get smaller. The OLS regression model produces one model that applies globally while the GWR model with the tricube fixed kernel weighting function produces four regional groups based on significant variables with different models in each district/city. The distribution of areas in the resulting groups tends to be close together, meaning that district/cities that are close to each other tend to have the same characteristics, in this case it is a significant variable on life expectancy. This is in accordance with Tobler's Law which states that everything is interconnected but two things that are close to each other have a greater influence than something that is far away. In the GWR model with a fixed kernel tricube weighting function, there are three significant predictor variables, namely the percentage of households that have access to proper sanitation (X_1), the coverage of exclusive breastfeeding for infants aged less than 6 months (X_2), and the percentage of monthly expenditure per capita for food (X_5). The variable percentage of households that have access to proper sanitation (X_1) has a positive effect on life expectancy. The positive effect means that if there is an increase in the percentage of households with access to proper sanitation, life expectancy will also increase. Some categories for proper sanitation are latrines using a latrine water seal system, feces disposal with a septic tank, and sanitation facilities used by a household alone, with other households, or communally [6]. Environmental conditions are important factors in disease development. Good excrement management will create proper sanitation conditions so that it can improve the quality of people's lives by increasing life expectancy [7].

The variable coverage of exclusive breastfeeding for infants aged less than 6 months (X_2) has a negative effect on life expectancy. Any increase in the coverage of exclusive breastfeeding will have an impact on decreasing life expectancy. Exclusive breastfeeding means giving only breast milk to infant from birth until they are 6 months old. The role of exclusive breastfeeding in infant is to increase the baby's body immunity so as to prevent the infant from getting various diseases that can threaten the infant's health [8]. Infant who gets exclusive breastfeeding until the age of 6 months has a survival rate of up to 33 times better than infant who gets exclusive breastfeeding for less than 4 months [5]. Research results that are not in accordance with the theory can be caused by the practice of breastfeeding in infants aged less than 6 months which are actually not exclusive at home. Infants who are given food and drink other than breast milk prematurely, such as formula milk, porridge, water, or bananas can increase the risk of allergies and digestive problems such as gastroenteritis [9]. Culture and beliefs can also influence the behavior of breastfeeding, including baby nurses other than mothers such as grandmothers who will be closely related to cultural beliefs and habits. Thus, socialization and intensive assistance from those closest to the family and health workers in the practice of exclusive breastfeeding so that the infant's nutrition is sufficient so that the infant's body resistance can be maintained.

The variable percentage of monthly expenditure per capita for food (X_5) has a negative effect on life expectancy in all district/cities in East Java Province. This means that for every increase that occurs in the percentage of monthly expenditure per capita for food, life expectancy will also increase. The percentage of monthly expenditure per capita for food describes the percentage of costs incurred for food as consumption by all household members during the month.

A person's quality of life is described by nutritional status which can be influenced by the economic capacity of the household or the amount of income spent in obtaining food [10]. The research results are inverse to the theory. This can be caused by the percentage of commodity groups that fall into the food category, which is dominated by processed food and beverages in the first rank and cigarettes in the second rank [11]. Expenditures that are dominated by these two components have the opportunity to increase the risk of malnutrition which can affect life expectancy. Thus, the impetus for increasing spending per capita for food is in fulfilling balanced nutrition such as whole grains, animal protein, vegetables and fruits.

5. Conclusion

Geographically Weighted Regression (GWR) model with a fixed kernel tricube weighting function is a better model than the OLS regression model in modeling life expectancy in East Java Province in 2021. The predictor variable that influences life expectancy is the percentage of households that have access to sanitation adequate coverage, coverage of exclusive breastfeeding for infants aged less than 6 months, and the percentage of monthly expenditure per capita for food spread across four regional groups based on significant predictor variablesin the GWR model. The mapping results show that the areas in the same group will be close toeach other. Each district/city has different parameter estimates and can be used locally to each region.

Compliance with ethical standards

Acknowledgements

Authors would like to thank reviewer's constructive suggestions to this paper.

Disclosure of Conflict of interest

Authors proclaim no conflict of interest

References

[1] Usman AU, Tukur K, Suleiman A, Abdulkadir A, Ibrahim H. The use of the weighted least squares method when the error variance is heteroscedastic. Benin J Stat. 2019;2:85–93.

[2] Oscar L, Astivia O, Zumbo BD. Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in R and SPSS. Pract Assesment, Res Eval. 2019;24(1):1–16.

[3] Kartika S, Sufri, Kholijah G. Penggunaan metode geographically weighted regression (GWR) untuk mengestimasi faktor dominan yang mempengaruhi penduduk miskin di Provinsi Jambi. J Math Theory Appl. 2020;2(2):37–45.

[4] Maryani H, Kristiana L. Pemodelan angka harapan hidup (AHH) laki-laki dan perempuan di Indonesia tahun 2016. Bul Penelit Sist Kesehat. 2018;21(2):71–81.

[5] Mahara DO, Fauzan A. Impacts of Human Development Index and Percentage of Total Population on Poverty using OLS and GWR models in Central Java, Indonesia. EKSAKTA. 2021;2(2):142–54.

[6] Badan Perencanaan Pembangunan Daerah Provinsi Jawa Timur [Internet]. Rencana penanggulangan kemiskinan daerah Provinsi Jawa Timur tahun 2019-2024; 2021. [cited 2023 June 24]. Available from https://sepakat.bappenas.go.id/assets/media/dokumen/RPKD Jatim 2019-2024.pdf

[7] Ummalla M, Samal A, Zakari A, Lingamurthy S. The effect of sanitation and safe drinking water on child mortality and life expectancy: Evidence from a global sample of 100 countries. Aust Econ Pap. 2022;61(4):778–97.

[8] Alfaridh AY, Azizah AN, Ramadhaningtyas, Anggraini, Maghfiroh DF, Emizia, Amaria H, Mubarokah K, et al. Peningkatan kesadaran dan pengetahuan tentang ASI eksklusif pada remaja dan ibu dengan penyuluhan serta pembentukan kader melalui komunitas "CITALIA." J Pengabdi Kesehat Masy. 2021;1(2):119–27.

[9] Hurek RKK, Esem O. Determinan pemberian makan pada bayi berusia kurang dari enam bulan. ARKESMAS. 2020;5(2):1–8.

[10] Chandra F, Aisah. Hubungan sosial ekonomi terhadap status gizi remaja putri di SMA Negeri 11 Kota Jambi. J Akad Baiturrahim Jambi. 2023;12(1):188–93.

[11] BPS Jawa Timur [Internet]. Provinsi Jawa Timur dalam angka tahun 2022; 2022 [cited 2023 June 24]. Available from https://jatim.bps.go.id/publication/2022/02/25/33699f6fcd84e0e2a0ad96f0/provinsi-jawa-timur-dalam-angka-2022.htm