Vol. 09, No. 01: 11 — 16
April 2020
DOI: 10.22487/25411969.2019.v9.i1.15040

Natural Science
Journal of Science and Technology

Original article

Spatial logistic regression modeling with inverse weighting distance for open unemployment in districts/cities on the island of Sulawesi

Broklyn Pippo Marchegiani Baebae*, Nur'eni, Iman Setiawan*

[1]Statistics Studies Program, Tadulako University, Jl Soekarno-Hatta, Tondo Palu 94118

Keywords: Unemployment, Open unemployment rate, spatial logistic regression, Moran's I Test, Weighting Matrix

Article history:
Received 17 March 2020
Accepted 26 April 2020
Published 29 April 2020

* Coresponding Author :
pippotongo@gmail.com
i.setiawan@untad.ac.id

Abstract

Unemployment is a condition where a person does not have a job, but is looking for a job. To see the unemployment situation in an area, logistic regression analysis can be used. Logistic regression is an analysis used to see the relationship between the response variable (Y) which is binary and the explanatory variable (X) which is categorical or continuous. The application of logistic regression often has a spatial influence on the model. In this study, to model the open unemployment rate, a spatial logistic regression method is used to determine whether the state of the open unemployment rate on Sulawesi Island is influenced by spatial factors. Spatial logistic regression is logistic regression analysis by incorporating spatial influences into the model. Spatial dependency testing is used by Moran's I Test. The weighting matrix used is the distance inverse weighting matrix. The results obtained, the value of Moran's I Test with a p-value of 2.14×10^{-12} <α (0.05), meaning that there is a spatial influence on the level of open unemployment on the island of Sulawesi. With a value of $Z_{(0,05;2)}$ = 1.960 and a significant value <α = 0.05, it was found that the Spatial Factor Variable (Z) affected the level of open unemployment and other variables that affected the level of open unemployment on Sulawesi Island were population density variable (x3), average length of school variables (x7).

INTRODUCTION

Unemployment is a condition which a person belongs to the *labor force* category who do not have a job but are actively looking for work (Nanga, 2001). Indonesia is a developing country that have a problem about Open Unemployment Rate. The Open Unemployment Rate (OUR) is an indicator that can be used to detect the level of labor supply that is not used, or is not absorbed by the labor market. In August 2017 the Open Unemployment Rate in Indonesia was 5.50%. In 2017 the Central Statistics Agency reported that were128.06 million Indonesians were a workforce, 121.02 million people were worked and 7.04 million were unemployed (BPS, 2017).

Sulawesi is the fourth largest island in Indonesia with an area of 174,600 km^2, and the third largest population in Indonesia about 19,93 million or 7.33% of population. The number of unemployed people in Sulawesi in 2017 were 431,372, which means 6.13% of the unemployed in Indonesia and the Open Unemployment Rate were 4.94%. Based on the unemployment rate, the condition of a country whether the economy was developing, slow or experiencing setbacks. Based on these data, it is necessary to conduct the study on factors that affect the level of open unemployment in Sulawesi (BPS, 2017).

In general, the analysis to determine the relationship between the response variable with the

predictor variable is a regression method. Regression is a study about dependent variable response (*dependent variable*), with one or more explanatory variables (*independent variable*). The value of the response variable depends on the explanatory variable (Gujarati & Porter, 1999). However, when a variable has a statistical model with the response variable is binary of "success" and "failure", then the form that used is logistic responses (Agresti, 2003). In 2017 the number of unemployed people on Sulawesi Island was 431,372 people, which means 6.13% of the unemployed in Indonesia and the Open Unemployment Rate of 4.94%. So this study used logistic regression analysis to see the factors that influence the level of open unemployment on the island of Sulawesi.

Logistic regression is a statistical analysis method that describes the relationship between response variables with two categories (*dichotomus*) and explanatory variables that are categorized or continuous (Hosmer & Lameshow, 2000). The application of logistic regression is very broad, but sometimes in its application are found that there are spatial influences that affect the model. Law I Geography stated that '*everything is related to everything else, but near things are more related than distant things'*, which meaning that everything is related to each other, but something close to each other has a closer relationship than those who are further away (Lee & Wong, 2001). Logistic regression analysis by including spatial factors into the model is called spatial logistic regression analysis (Ward & Gleditsch, 2018). Based on Law 1 of Geography above, so an analysis is needed to see whether the open unemployment rate on Sulawesi Island is influenced by the spatial dependence, and the analysis that can be used is spatial logistic regression. Spatial logistic regression is logistic regression by including spatial factors in modeling.

The recent study has been carried out by Tiara (2016) about the spatial logistic regression model with a case study of pulmonary tuberculosis spread in Samarinda. It was concluded that spatial factors influence the model. In this study used a spatial logistic regression model to explain the effect of spatial factors on the level of open unemployment in Sulawesi.

MATERIALS AND METHODS

The data that used in this study was the secondary data, which obtained in Sulawesi in 2017 with 81 districts. The variables used in this study were :

a. Response variable is open unemployment rate (Y), with categories:
   0 : Low OUR is ≤ 4%
   1 : High OUR is > 4%
b. The explanatory variables are the Number of population (x_1), Number of population > 15 years (x_2), population density (x_3), Poor population (x_4), HDI (x_5), Life expectancy (x_6), Average length of school (x_7), expectation of long school (x_8), economic growth (x_9), and APK for high school equivalents (x_{10})

The method that used in this study was spatial logistic regression analysis, with the following stages of analysis:

1. Retrieve data.
2. Conduct descriptive analysis.
3. Make a distance inverse weighting matrix, with *latitude* and *longitude* for each region.
4. Test the spatial effect of the response variable *Y* with *Moran's I test.*
5. Creating a new explanatory variable, the spatial influence variable by multiplying the weighting matrix with the Y variable (WY).
6. Creating a spatial logistic regression analysis model.
7. Perform simultaneous test of model parameters (*Likelihood ratio* test) and partial test of model parameters (*Wald* test).
8. Test the suitability of the model.
9. Interpretation.

Table 1. Descriptive Statistics

Var	Min	Max	Mean	Std. Dev
Y	0,47	10,96	4,4543	2,254
X_1	33.212	1.489.011	246.070,98	207.963,94
X_2	1,22	33,39	7,40	6,17
X_3	11,93	8.471,00	395,22	1.055,085
X_4	2,20	21,85	11,69	4,90
X_5	62,35	81,83	68,07	4,44
X_6	60,79	73,02	68,29	2,54
X_7	5,98	11,68	8,08	1,20
X_8	11,16	16,06	12,77	0,99
X_9	3,07	14, 42	6,70	1,41
X_{10}	50,66	116,94	91,96	15,19

RESULTS AND DISCUSSION

Descriptive Statistics

Descriptive analysis was aimed to describe the data. Descriptive statistics of each variable can be seen in Table 1.

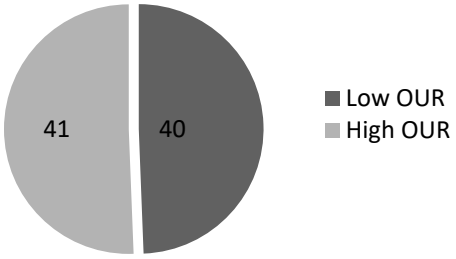

Figure 1. Number of districts by open unemployment category

Spatial Weighting Matrix

Spatial weighting matrix is a matrix that states the relationship of the observation area *n x n* and symbolized by **W**. The weighting matrix W with order 81x81 calculated by using R software are in the following:

$$\begin{matrix} & 1 & 2 & 3 & \cdots & 81 \\ \begin{matrix}1\\2\\3\\\vdots\\81\end{matrix} & \begin{bmatrix} 0.0000000 & 1.6658975 & 0.5569179 & \cdots & 0.4359717 \\ 1.6658975 & 0.0000000 & 0.5835072 & \cdots & 0.4322094 \\ 0.5569179 & 0.5835072 & 0.0000000 & \cdots & 0.2495034 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0.4359717 & 0.4322094 & 0.2495034 & \cdots & 0.0000000 \end{bmatrix} & \end{matrix}$$

Moran's I Test

Moran's I Test was used to detect the dependency of spatial autocorrelation between observations or locations. The results of *Moran's I Test* is:

Table 2. *Moran's I Test*

Statistics	***p-value***
0,198372689	$2{,}14\ x10^{-12}$

Moran's I was obtained with a *p-value* of $2{,}14\ x10^{-12}$ less than α (0.05) H_0. was declined. This showed that there was a spatial autocorrelation of the open unemployment rate.

Spatial Influence Variables

The spatial effect as a new explanatory variable (Z), was obtained by multiplying the matrix **W** to the percentage of the open unemployment rate (Y) .

$$\boldsymbol{Z} = \mathbf{W} \times \mathbf{Y}$$

$$\mathrm{Z} = \begin{bmatrix} 0.0000000 & 1.6658975 & 0.5569179 & \cdots & 0.4359717 \\ 1.6658975 & 0.0000000 & 0.5835072 & \cdots & 0.4322094 \\ 0.5569179 & 0.5835072 & 0.0000000 & \cdots & 0.2495034 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0.4359717 & 0.4322094 & 0.2495034 & \cdots & 0.0000000 \end{bmatrix} \begin{bmatrix} 3.24 \\ 2.94 \\ 2.72 \\ \vdots \\ 5.71 \end{bmatrix}$$

The Z variable was obtained :

Table 3. Spatial variables

Districts	Z
Banggai Kepulauan	121.58708
Banggai	122.59701
Morowali	130.13037
⋮	⋮
Kota Kotamobago	193.35245

Spatial Logistic Regression Model

The estimated spatial logistic regression parameters were :

Table 4. Estimates of spatial logistic regression parameters

Parameter	Estimasi
β_0	-4,848
β_1	0,000002885
β_2	-0,0473
β_3	0,006669
β_4	-0,04263
β_5	-0,269
β_6	0,1642
β_7	1,531
β_8	-0,1581
β_9	-0,2208
β_{10}	-0,009732
ρ	0,01871

Spatial logistic regression equations were obtained with 10 explanatory variables (X) and spatial influence variables (Z) on the open unemployment rate variable (Y) in order to obtain the following equation:

$$\pi(x) = \frac{\exp\begin{bmatrix} -4{,}848+0{,}000002885(x_1)-0{,}0473(x_2)+0{,}006669(x_3) \\ -0{,}04263(x_4)-0{,}269(x_5)+0{,}1642(x_6)+1{,}531(x_7)-0{,}1581(x_8) \\ -0{,}2208(x_9)-0{,}009732(x_{10})+0{,}01871(z) \end{bmatrix}}{1+\exp\begin{bmatrix} -4{,}848+0{,}000002885(x_1)-0{,}0473(x_2)+0{,}006669(x_3) \\ -0{,}04263(x_4)-0{,}269(x_5)+0{,}1642(x_6)+1{,}531(x_7)-0{,}1581(x_8) \\ -0{,}2208(x_9)-0{,}009732(x_{10})+0{,}01871(z) \end{bmatrix}}$$

The equation are:

$$Logit\,\pi(x) = g(x) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$$

$$\begin{aligned} g(x) = & -4{,}848+0{,}000002885(\mathrm{x}_1) -0{,}0473(\mathrm{x}_2) \\ & 0{,}006669(\mathrm{x}_3)-0{,}04263(\mathrm{x}_4)-0{,}269(\mathrm{x}_5) +0{,}1642(\mathrm{x}_6) \\ & +1{,}531(\mathrm{x}_7) -0{,}1581(\mathrm{x}_8) -0{,}2208(\mathrm{x}_9) \\ & -0{,}009732(\mathrm{x}_{10}) +0{,}01871(\mathrm{z}) \end{aligned}$$

Simultaneous Testing of Parameters

This test was aimed to explain whether explanatory variables (X) were simultaneously affect to the response variable (Y). The results of the simultaneous test of the model parameters :

Table 5. Simultaneous test results

Chi-square	***p-value***
41,264	2,171 x 10^{-05}

Based on the data in table 5 showed that the value of the *likelihood ratio* test were 41.254, with a value $\chi^2_{(0,05;11)}$ = 19.6751, then 41.254 > 19.6751, or with a significance value of 2.171 x 10^{-05}<0.05. Then the H_0 was declined, it meaning that simultaneously explanatory variables (X) were affected to the variable level of open unemployment (Y).

Partial Parameters Testing

This test was aimed to determine how the effect of each explanatory variable (X) on the response variable (Y). the partial test results of the model :

Table 6. Partial test results

Parameter	*Wald*	*p-value*
Intercept	0,424	0,6715
β_1	1,303	0,1924
β_2	-0,683	0,4947
β_3	**2,011**	**0,0433**
β_4	-0,557	0,5773
β_5	-1,178	0,2388
β_6	0,946	0,3443
β_7	**2,266**	**0,0234**
β_8	-0,263	0,7928
β_9	-0,838	0,4023
β_{10}	-0,437	0,6624
ρ	**2,202**	**0,0277**

Based on the data of table 6 showed that the *Wald* test values from each explanatory variable. With a value $Z_{(0,05;2)}$ = 1.960 and a significant value α = 0.05. The variables that affected the level of open unemployment rate were the population density variable (x_3), the average length of schooling variable (x_7), and the spatial influence variable (z).

Model Conformity Testing

This test was aimed to determine whether the spatial logistic regression model obtained was feasible to be used. the results of testing the suitability of the model :

Tabel 7. *Hosmer and Lameshow test*

$\hat{C}$	*Df*	*p-value*
5,3513	8	0,7194

Based on the data of table 7 showed that the value $\hat{C}$ were 5.3513, and the significance value were 0.7194, where the value $\chi^2_{(0,05;6)}$ were 12.5916, then 5.3513 < 12.5916 or with significant value 0.7194> 0.05, then H_0 was accepted, so it can be concluded that there were no difference between the observations with the predicted results.

Odds Ratio

Odds ratio is a measure that reflects the size of certain tendencies appear among the results of a group that has a certain character than the comparison group . Here were the *Odds Ratio* values:

Table 8. The Odds ratio

Variable	*Odds Ratio*
X_1	1,000002885
X_2	0,953868673
X_3	1,006690858
X_4	0,958265876
X_5	0,764166906
X_6	1,178468293
X_7	4,622033272
X_8	0,853743863
X_9	0,801877875
X_{10}	0,990315525
Z	1,018889919

Based on the data of table 8 can be interpreted: for variable X_1 (population) *odds ratio* = 1.000002885, meaning that when there were two districts with a population difference of 1 person, the tendency for districts to have a higher HDI at high open unemployment rates (y = 1) increases equal to 1.000002885 times compared to low open unemployment (y = 0).

For variable X_2(population> 15 years) *odds ratio* = 0.953868673, meaning that if there were two districts with a difference in population> 15 years by 1%, the tendency for districts to have a population of> 15 years is higher at the high open unemployment rate (y= 1), decreased by 0.953868673 times compared to the low open unemployment rate (y = 0).

For variable X_3(population density) *odds ratio* = 1.006690858, meaning that if there were two districts with a population density difference of 1 person / km^2, then the tendency for districts / cities to have districts / cities with higher population density at the unemployment rate open high (y = 1) increased by 1.006690858 times compared to the open unemployment rate (y = 0).

For variable X_4(poor population) *odds ratio* = 0.958265876, meaning that if there are two districts / cities where the difference in poverty is 1%, then the tendency for districts / cities to have poorer populations is higher at high open unemployment rates (y = 1) decreased by 0.958265876 times compared to the low open unemployment rate (y = 0) .

For variable X_5(HDI) *odds ratio* = 0.764166906, meaning that if there are two regencies / cities with a HDI difference of 1%, the tendency for districts / cities to have a higher HDI at high open unemployment rates (y = 1) decreases by 0.764166906 times compared to the low open unemployment rate (y = 0).

For variable X_6(life expectancy) *odds ratio* = 1.178468298, meaning that if there are two districts / cities that have a difference in life expectancy of 1%, then the tendency for districts / cities to have a higher life expectancy at high open unemployment rates (y = 1) increased by 1.178468298 times compared to the low open unemployment rate (y = 0).

For variable X_7(average length of school) *odds ratio* = 4.622033272, meaning that if there are two districts / cities that have an average difference in length of schooling of 1 year, then the tendency for districts / cities to have a higher average length of schooling the high open unemployment rate (y = 1) increased by 4.622033272 times compared to the low open unemployment rate (y = 0).

For variable X_8(Expectations of old school days) *odds ratio* = 0.853743863, meaning that if there are two districts / cities that have a difference in expectations of one year of schooling, then the tendency for districts / cities to have higher schooling expectations is at high open unemployment rates (y = 1) decreased by 0.853743863 times compared to the low open unemployment rate (y = 0).

For variable X_9(economic growth) *odds ratio* = 0.801877875, meaning that if there are two districts / cities with a difference in economic growth of 1%, then the tendency for districts / cities to have economic growth at high open unemployment rates (y = 1) decreases amounted to 0.801877875 times compared to the low open unemployment rate (y = 0).

For variable X_{10}(APK for equivalent SMA) *odds ratio* = 0.990315525, meaning that if there are two regencies / cities that have a difference in APK for SMA equal to 1%, then the tendency for districts / cities to have a SMA APK equivalent in high open unemployment rates (y= 1) decreased by 0.990315525 times compared to the low unemployment rate (y = 0).

For the variable Z (Spatial effect) *odds ratio* = 1.018889919, meaning that if there are two districts / cities that have a spatial effect difference of 1 unit, the tendency for districts / cities to have a higher spatial influence on high open unemployment rates (y = 1) increased by 1.018889919 times compared to the low unemployment rate (y = 0).

Nagelkerke R^2

Nagelkerke R^2 were used to determine the explanatory variables that explain the variable response. Based on the value *Nagelkerke* R^2 using the *software* R was 0.532. This showed that the variable population (X_1), population <15 years (X_2), population density (X_3), poor population (X_4), HDI (X_5), life expectancy (X_6), average length of school (X_7), long expectation of school (X_8), economic growth (X_9), APK high school equivalent (X_{10}), spatial influence variable (Z) were able to explain the variable of the open unemployment rate (Y) about 53,2%, and the remaining 46.8% was explained by variables of the variables that used in the model.

Accuracy of Classification

The accuracy of the model in classifying was aimed to evaluate the model. The results of the accuracy classification :

Table 9. accuracy of classification

Actual	Prediction		Total
	Low OUR	High OUR	
TPT Low	35	10	45
TPT High	5	31	36

Based on the table 9, the percentage of classification accuracy can be obtained :

$$1 - APER = \left(\frac{35+31}{45+36}\right) x100\% = 81{,}48\%$$

Based on the calculation, showed that the spatial logistic regression model had the ability to classify observations correctly to 81.48%.

The *sensitivity* and *specificity* values are :

$$Sensitivity = \left(\frac{35}{35+5}\right) x100\% = 87{,}5\%$$

$$Specificity = \left(\frac{31}{31+10}\right) x100\% = 75{,}6\%$$

The *sensitivity* calculation results showed that the spatial logistic regression model had the ability to classify correctly for the category of open low unemployment rate as a category of low open

unemployment rate about 87.5%, and the results of the calculation of *specitivity* showed that the spatial logistic regression model has the ability to classify correctly for high open unemployment rate category as high open unemployment rate category of 75.6% .

Based on the percentage of calculation accuracy classification results, the sensitivity and specificity values can be concluded that the spatial logistic regression model obtained is good, and can be used in modeling the open unemployment rate on Sulawesi Island. Spatial logistic regression model showed that there is a spatial influence on the level of open unemployment on the island of Sulawesi. It was found that the Spatial Factor Variable (Z) affected the level of open unemployment and other variables that affected the level of open unemployment on Sulawesi Island were population density variable (x3), average length of school variables (x7). For further research can measure the extent to which the spatial dependence can affect level of open unemployment by adding several new variables and using different spatial weighting matrix.

Acknowledgement

The author would like to thank to the Central Statistics Agency for allowing the authors to retrieve data to complete this study.

References

Agresti, A. 2003. *Categorical data analysis*, 2nd edn. John Wiley & Sons. New York.

[BPS] Badan Pusat Statistik.(2017). *Keadaan Ketenagakerjaan Indonesia*. Indonesia: BPS Indonesia

Gujarati, DN., Porter, DC. 1999. *Essentials of econometrics*. 4th ed. McGraw-Hill/Irwin. Singapore.

Hosmer, D. W., Lemeshow, S. (2000). Applied Logistic Regression, 2nd ed. John Wiley & Sons. New York.

Lee, J., Wong, DW. 2001. *Statistical analysis with ArcView GIS*. John Wiley & Sons. New York.

Nanga, M. 2001. Makro Ekonomi Teori, Masalah dan Kebijakan Edisi Pertama: PT Raja Grafindo Persada. Jakarta, Indonesia

Tiara, N.M. 2016. *Model Regresi Logistik Spasial Pada Penyebaran Penyakit Tuberkulosis Paru di Setiap Kelurahan di Kota Samarinda Pada Tahun 2013.* [Thesis]. Universitas Mulawarman Samarinda, [Indonesia]

Ward, M. D., & Gleditsch, K. S. (2018). *Spatial regression models* 2nd ed. Sage Publications. United States.