Available online at http://scik.org
J. Math. Comput. Sci. 2022, 12:122
https://doi.org/10.28919/jmcs/7264
ISSN: 1927-5307

AUTOCORRELATION TESTING ON RESIDUAL SPATIAL LOGISTIC REGRESSION MODEL WITH EUCLIDEAN DISTANCE MATRIX APPROACH

DEVI YANTI[1], TONI TOHARUDIN[2,*], YUSEP SUPARMAN[2]

[1]Post-Graduate Program in Applied Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

[2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

Abstract: The existence of spatial effects should not be ignored because it will reduce the goodness of the model. One type of regression analysis that is quite widely used is logistic regression analysis. Spatial logistic regression modelling incorporates spatial effects into the logistic regression model with the expectation that the residuals generated from the model are independent or there is no autocorrelation. The purpose of this study was to obtain the results of spatial autocorrelation testing using a spatial logistic regression model with a Euclidean matrix approach. The results of the study were applied to natural disaster mitigation data in Kupang Regency, Nusa Tenggara Timur Province in 2020, where the distribution of areas in Kupang Regency by village/urban village has spatial autocorrelation. Spatial autocorrelation testing was carried out with Moran's I test to determine the presence of spatial autocorrelation. In this study, a standardized Euclidean distance matrix approach was used to accommodate this spatial effect. The results of the autocorrelation test of the binary spatial logistic model with the Euclidean distance matrix approach were able to accommodate the spatial effect.

---

[*]Corresponding author
E-mail address: toni.toharudin@unpad.ac.id
Received February 13, 2022

Keywords: spatial; R software; Moran's I; logistic regression.

2010 AMS Subject Classification: 91B72, 91D25.

1. INTRODUCTION

In classical modelling it is assumed that between spaces are independent or there is no autocorrelation. When applied to data that is influenced by location, the use of regression analysis is not appropriate because it ignores the influence of that location. Spatial influence should not be ignored because it will reduce the goodness of the model [1].

Classical modelling which assumes that between independent space becomes less relevant is included in logistic regression analysis. Spatial logistic regression modelling incorporates spatial effects into the regression model. The spatial logistic model with the contiguity matrix approach has been able to accommodate this spatial effect. This is indicated by the absence of autocorrelation in the remainder of the spatial logistic regression model. Based on this, in this study, spatial autocorrelation will be tested against the remainder of the binary logistic spatial regression model using another approach, namely the Euclidean distance matrix [2].

According to BNPB, Kupang Regency is the district with the highest disaster risk index in Nusa Tenggara Timur Province, with an index value of 168.48. So that this research is applied to natural disaster mitigation data in Kupang Regency, Nusa Tenggara Timur Province in 2020. The distribution of the area has spatial autocorrelation. So, in the modelling it is necessary to pay attention to the spatial effect. The existence of spatial influence in the spatial regression model consists of spatial autocorrelation and spatial dependence. In this study, only the spatial dependence test will be carried out using the Moran's I Test [3].

2. PRELIMINARIES

2.1 Spatial Autocorrelation

An important aspect in determining spatial autocorrelation is determining the relationship between the closest region, the area around the observed area is thought to have an influence on the observed area. According to Tobler's first law, everything is related, but something closer is more related [4].

According to Lembo (2006) Spatial autocorrelation is the correlation between variables and themselves based on space or can also be interpreted as a measure of the similarity of objects in a space (distance, time and area) [5].

Spatial data is geographically oriented data, has a certain coordinate system as a reference base and has two important parts that make it different from other data, namely location (spatial) and descriptive (attribute) information. Location information (spatial) is related to a geographic coordinate, namely latitude and longitude [6]. Descriptive information (attribute) or non-spatial information, a location that has some information such as population or vegetation type [7].

Analysis of spatial data to determine the existence of spatial autocorrelation requires the main component, namely the weighting matrix. The purpose of the weighting matrix is to determine the weights between the observed regions based on the proximity of the distance of an area [8].

Spatial autocorrelation is the correlation between the values of a variable and other values on the same variable. If there is a systematic pattern in the distribution of a variable, then there is a spatial autocorrelation. The existence of spatial autocorrelation indicates that the attribute value in a certain area is related to the attribute value in other areas that are adjacent or neighboring [9].

2.2 Spatial Weighting Matrix

Spatial weighting matrix (W) is an important element in describing the proximity of one location to another and is determined based on information or proximity between one location and another [10]. One of the spatial autocorrelation tests used is the Euclidean distance matrix. The formation of a distance weighting matrix is obtained from the calculation of the Euclidean distances between research locations based on the degrees of Latitude and Longitude with the Euclidean distance formula between the I and j locations as follows [11]:

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

where:

$u_i, u_j =$ two latitude distance vectors of the location to be calculated.

$v_i, v_j =$ two longitude distance vectors of the location to be calculated.

Based on the distance obtained, a weighting matrix $W_{(denominator)}$ will be made which is formed based on the boundaries of the area or area (r) determined by the researcher. The formation of the $W_{ij}(d)$ matrix based on the distance matrix obtained, then a binary weighting matrix is made with the provisions of the value 1 if $d_{ij} < r$, and 0 otherwise. The general form of the distance weighting matrix is as follows [12]:

$$W_{ij} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ w_{31} & w_{32} & w_{33} & \cdots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nn} \end{bmatrix}$$

2.3 Spatial Logistics Regression

Logistics Regression is used to find the relationship between the dependent variable (Y) which is dichotomous (nominal or ordinal with 2 categories) or polychotomous (nominal or ordinal scale > 2 categories) with one or more independent variables (X) which are continuous or categorical [13].

One of the very simple logistic regressions is binary logistic regression. Binary logistic regression method is a method used to describe the relationship of one or more independent variables to the dependent variable. The dependent variable used is discrete category with two possibilities, namely success and failure. Success events are usually denoted by Y=1, while failure events are denoted by Y=0 [14].

If some of the independent variables have nominal or ordinal scales, then these variables will not be appropriate if they are included in the logit model because the numbers used to express these levels are only for identification and have no numerical value in such a situation, a dummy variable is needed. For independent variables with ordinal and nominal scales with k categories, k-1 dummy variables will be needed [15].

Spatial logistic regression is a regression model that incorporates spatial effects into the model to overcome the existence of spatial relationships. The spatial effect in question is to form a new variable called the Spatial variable [16].

2.4 Moran's Index (Moran's I)

Moran's I is a statistical test used to identify a location from a spatial grouping or spatial autocorrelation [17]. Global testing through Moran's I statistics is a test of the existence of autocorrelation with the assumption that the location is the same but the variables are different and based on covariance [18]. A Moran's I value of zero indicates not in groups, a positive Moran's I value indicates a positive spatial autocorrelation which means that adjacent locations have similar values and tend to be clustered (Clusters), a positive Moran's I value indicates a negative spatial autocorrelation which means that adjacent locations have different values [19]. Calculation of spatial autocorrelation using Moran's Index can be done with the following hypothesis [20]:

H_0: $I = 0$ (no spatial dependency/spatial autocorrelation)

H_1: $I \neq 0$ (there are spatial dependencies/spatial autocorrelation)

Test statistics with the following formula:

$$Z_i = \frac{I - E(I)}{\sqrt{\widehat{var}(I)}}$$

where:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}) \sum_{i=1}^{n} (y_i - \bar{y})^2} \; , \qquad \widehat{var}(I) = \frac{n^2 S_1 - nS_2 + 3(w)^2}{(w)^2 (n^2 - 1)}$$

$$E(I) = \frac{-1}{(n-1)} \qquad w = \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}$$

n = number of observations

y_i = observation value at location i

$\bar{y}$ = mean of from n locations

y_j = observation value at location j

$$S_1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (c_{ij} + c_{ji})^2}{2}$$

$$S_2 = \sum_{i=1}^{n} (c_{i.} + c_{.i})^2$$

I = Moran Index

Z(I) = Moran index test statistic value

E(I) = Expected value of Moran index

W_{ij} = weighting matrix element

cij = Elements of the Euclidean distance matrix

ci. = Number of rows i Euclidean distance

c.i = Number of column i Euclidean distance

The decision-making criteria is if the p value < (0.95) then reject H_0 (there is an autocorrelation). Conversely, if the p value > (0.95) then accept H_0 (no autocorrelation).

2.5 Materials and Methods

The data used in this study is secondary data derived from the 2020 Village Potential Survey (PODES) raw data conducted by BPS. The unit of analysis used in this research is 177 villages/urban village which are included in areas prone to natural disasters in Kupang Regency, Nusa Tenggara Timur Province [21].

The variable used in this research is the dependent variable (Y) natural disaster mitigation/preparedness which consists of two categories, namely ready and not ready. The independent variable (X) consists of regional classification, waste management, access to information, community participation, experience of natural disasters, and education level. Spatial variables are formed from the spatial weighting matrix multiplied by the response variable vector (Y) [22]. The research variables are presented in Table 1.

Table 1. Research Variables

Variables	Category
Natural Disaster Mitigation (Y)	1= ready, 0= not ready
Region Classification (X1)	1= urban, 0= rural
Waste Management (X2)	1= good, 0= not good
Access to Information (X3)	1= good, 0= not good
Community Participation (X4)	2= high, 1= moderate, 0= low
Experience of Natural Disasters (X5)	1= yes, 0= none
Education Level (X6)	1= high, 0= low

The steps taken are to look for the residuals from the results of the formed spatial logistic regression and look for the statistical value I (Moran index) and test the hypothesis. Data processing was carried out using software R version 4.1.2 and Microsoft Excel 2013.

3. MAIN RESULTS

Based on the results of processing the spatial logistic regression model with the Euclidian distance matrix approach, it has slightly better results in predicting natural disaster mitigation in Kupang Regency compared to non-spatial logistic regression. This can be seen from the higher Correct Classification Rate (CCR) and the smaller Akaike Information Criterion (AIC) for spatial logistic regression compared to non-spatial logistic regression (Table 2). The logistic regression model that has the smallest AIC value is the better model [23] and the CCR is a value that shows how accurately the model predicts the actual data [24].

Table 2. Comparison of the Best Value of Spatial and Non-Spatial Models

	Spatial Logistics Regression	Non-Spatial Logistics Regression
AIC	32.074	83.615
CCR	98.87%	94.35%

This research is in line when viewed from the side of spatial autocorrelation where the level of natural disaster mitigation between villages/urban village has a positive spatial correlation based on the results of Moran's I test (Table 3). This means that there is a grouping based on the level of natural disaster mitigation in the village. Based on this information, an estimation of natural disaster mitigation is carried out using spatial logistic regression. Villages that are included in the category of being unprepared for natural disasters because of the distance from the surrounding villages are also categorized as unprepared. Likewise, villages that are included in the category of being ready to face natural disasters because of the distance from the surrounding villages are also categorized as unprepared (Figure 1). The ready and not ready categories are obtained from a combination of four indicators, namely natural disaster early warning systems, safety equipment, signs & evacuation routes, and construction, maintenance, or normalization: rivers, canals, embankments, drainage ditches, reservoirs, beaches, etc. [25].

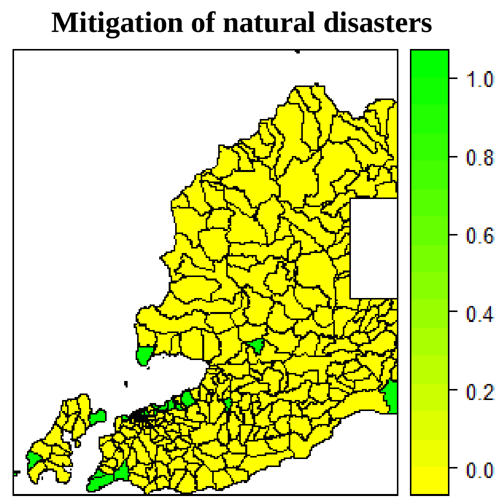

Figure 1. Map of the distribution of natural disaster mitigation

Based on the research of Wulandari et al. (2017) natural disaster mitigation is influenced by education level of natural disaster experience, participation in community activities, and access to information, while the regional classification variable has no effect [26]. In Wijayanti's research (2013) entitled "Opportunities for Waste Management as a Mitigation Strategy in Realizing Climate Resilience in Semarang City", urban waste management is one of the mitigation strategies in creating the resilience of the City of Semarang to climate change. Waste management strategies can form climate resilience. Waste management strategies by third parties at the TPA have a greater chance of contributing to the development of climate resilience than waste management strategies at the Integrated Waste Processing Site (TPST) [27]. In this study, the factors of regional classification, waste management, access to information, community participation, experience of natural disasters, and level of education have no significant effect on natural disaster mitigation.

Furthermore, in this study, spatial autocorrelation with Moran's index will be tested on the residual results of the spatial logistic regression model. Spatial autocorrelation test on the residuals shows that the residuals are independent (Table 3).

Table 3. Moran Test Results for Natural Disaster Mitigation Levels and Spatial Logistics Regression Model Residual

Test	Statistic	P-Value
Level of natural disaster mitigation	2.7817	1.421e-05
Residual spatial logistic regression model	4.4891	0.0519

Based on the results of the Moran test at the level of natural disaster mitigation, the test statistic value is 2.7817 with a p-value of 1.421e-05, which is smaller than α (0,05) then rejects H_0 which means that there is a spatial autocorrelation. Based on this, the research is continued by incorporating spatial effects into the model. Furthermore, the spatial autocorrelation test was carried out on the remainder of the spatial logistic regression model. The results of the Moran test on the remainder of the spatial logistic regression model obtained a p-value of 0.0519, which is greater than α (0,05) so it failed to reject H_0 so it can be concluded that there is no spatial autocorrelation or the residuals are independent.

4. CONCLUSION

Based on the results and discussion, it can be concluded that the test spatial autocorrelation to the remainder of the spatial logistic model using distance matrix approach Euclidian has been able to accommodate the spatial autocorrelation between villages. This matter It is shown from the residual results of the spatial logistic regression model approach that they are independent of each other (no spatial autocorrelation). This study only uses Moran's test and the approach Euclidian distance matrix in spatial autocorrelation analysis, where the distance matrix approach Euclidian has one drawback that is less able to classify spatial data irregular so it is recommended for other research to use the test and another better approach.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] Hasriana, Pemodelan Kemiskinan Menggunakan Geographicallu Weighted Logistic Regression dengan Fungsi Pembobot Fixed Kernel, 2013. Accessed: Dec. 14, 2021. https://core.ac.uk/download/pdf/77625358.pdf

[2] U. Dyah Syafitri and B. Sartono, Pengujian Autokorelasi terhadap Sisaan Model Spatial Logistik, 2008. Accessed: Dec. 15, 2021. https://core.ac.uk/download/pdf/11064481.pdf

[3] BNPB, Indeks Risiko Bencana Indonesia. 2020.

[4] M.M. Fischer, J. Wang, Spatial Data Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[5] A.J. Lembo, Spatial Autocorrelation, 2006. Accessed: Dec. 18, 2021. https://www.researchgate.net/publication/319871707_Autocorrelation_Spatial.

[6] A. Puntodewo, S. Dewi, and J. Tarigan, Sistem Informasi Geografis Untuk pengelolaan sumberdaya alam, 2003. http://www.cifor.cgiar.org

[7] A. Maulani, Aplikasi Model Geographically Weighted Regression (GWR) Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Kasus Gizi Buruk Anak Balita Di Jawa Barat, 2013. Accessed: Dec. 19, 2021. http://repository.upi.edu/523/

[8] H.A. Khoirunissa, Pemetaan Risiko Penyakit Tuberkulosis (TBC) di Kota Surakarta dengan Spatial Empirical Bayes, Indones. J. Appl. Stat. 3 (2020), 78-84.

[9] A. J. Curtis and W.-A. A. Lee, Spatial patterns of diabetes related health problems for vulnerable populations in Los Angeles, 2010. http://www.ij-healthgeographics.com/content/9/1/43

[10] J. P. Lesage, The Theory and Practice of Spatial Econometrics, 1999. http://www.econ.utoledo.edu

[11] Y. Leung, C. L. Mei, and W. X. Zhang, Statistical tests for spatial nonstationarity based on the geographically weighted regression model, Environ. Plan. A, 32 (2000), 9–32.

[12] D.G. Eka, Multidirectional Optimum Ecotope Based on Algorithm (AMOEBA) and Near Neighbourhood Kernel Weighted Matrix in Getis Ord-G Statistic (Case Study: Grouping of District /Cities in Central Java Based on Characteristic of Economic Inequality), Yogyakarta, 2020.

[13] A. Agresti, An introduction to categorical data analysis, Second Edition. Wiley, New York, 2007.

[14] D.W. Hosmer and S. Lemeshow, Applied logistic regression, Second Edition. John Wiley and Sons, New York, 2000.

[15] G. Gendalasari, Pengaruh Motivasi Belajar Dan Status Ekonomi Orang Tua Terhadap Minat Studi Lanjut Siswa/Siswi Sma Dan Smk Di Bogor Pada Bidang Ekonomi Dan Manajemen, 2016. Accessed: Dec. 18, 2021. https://core.ac.uk/download/pdf/198236403.pdf

[16] H.J. Andriana, Analisis Regresi Logistik Spasial Untuk Menduga Status Kemiskinan Desa Di Kabupaten Majalengka, 2013. Accessed: Dec. 17, 2021. https://text-id.123dok.com/document/rz3w647q-analisis-regresi-

logistik-spasial-untuk-menduga-status-kemiskinan-desa-di-kabupaten-majalengka.html

[17] E. Sulistianto, W. Fahrizal, and F. Maryanto, Spatial Autocorelation of Catch Fisheries Production in East Kalimantan, Jurnal Harpodon Borneo, 11 (2018), 23-31.

[18] N.E. Khoirun, Identifikasi Spatial Pattern dan Spatial Autocorrelation Pada Indeks Pembangunan Manusia Provinsi Papua Barat Tahun 2012, 2012, Accessed: Dec. 19, 2021. https://journal.walisongo.ac.id/index.php/attaqaddum/article/view/1914

[19] D. Pfeiffer, ed., Spatial analysis in epidemiology, Oxford University Press, Oxford ; New York, 2008.

[20] J. Lee Jay and D.W.S.Wong David, Statistical Analysis with Arcview GIS. John Wiley & Sons, New York, 2001.

[21] BPS, Kabupaten Kupang Dalam Angka 2019. Badan Pusat Statistik, 2019. https://kupangkab.bps.go.id/publication/2019/08/16/b086aa5e1c1af027afee5eee/kabupaten-kupang-dalam-angka-2019.html

[22] T. Purwaningsih, "Penerapan regresi logistik ordinal spasial untuk menduga status kemiskinan kabupaten di pulau Jawa," 2011.

[23] W. Agus, Ekonometrika: Teori dan Aplikasi Untuk Ekonomi dan Bisnis. Ekonisia Fakultas Ekonomi Universitas Islam Indonesia, Yogyakarta, 2007.

[24] Nurlatifah, Determinan Ketahanan Pangan Regional dan Rumah Tangga di Provinsi Jawa Timur, 2011. Accessed: Dec. 22, 2021. https://repository.ipb.ac.id/handle/123456789/50915

[25] BPS, Penjelasan Teknis Data Updating Podes2020, 2020. Accessed: Dec. 20, 2022. https://silastik.bps.go.id/v3/index.php/mikrodata/detail/Y2dkSDJZVzFOenJJM3BmM0tIYVNUQT09

[26] A. Wulandari, F. M. Faruk, and F. Sisilia Doven, Penerapan Metode Regresi Logistik Biner untuk Mengetahui Determinan Kesiapsiagaan Rumah Tangga dalam Menghadapi Bencana (Application of Binary Logistic Regression Method to Know the Determinant of Household Preparedness in Facing Natural Disasters), 2017. Accessed: Dec. 18, 2021. https://prosiding.stis.ac.id/index.php/semnasoffstat/article/download/146/43/

[27] W. P. Wijayanti, Peluang Pengelolaan Sampah Sebagai Strategi Mitigasi dalam Mewujudkan Ketahanan Iklim Kota Semarang, Pembangunan Wilayah dan Kota, 9 (2013), 152–162. Accessed: Dec. 17, 2021. https://core.ac.uk/display/186474767