Building Spatial Regression Model to Know The influence of Road’s Access & Education to Population Growth of Germany in 2018

Tuti Purwaningsih [1,a], Aliffian Wahyu Raharjo[a]

[1]tuti.purwaningsih@uii.ac.id
[a] Statistics Department, Universitas Islam Indonesia, Indonesia

ABSTRACT

Article history:
Received: December 30, 2019
Revised: January 13, 2020
Accepted: February 20, 2020

Keywords:
GeoDa
Germany Population
Infrastructure
SAR
Spatial Regression

Population growth are one of the core information of a nation to make a decision of economics strategy in the future. This study aims to see what infrastructure or education factors affect population growth in Germany using Spatial Regression method through SAR model. Creation of spatial regression model using GeoDa software. The results of this study indicate that education and road infrastructure has a significant effect on population growth in Germany in 2018, the result of modelling proccess can be explained as the following equation: Population = 482443 + (-0.10695 ((11,293) (Education) + (113,8) (Road))) with R^2=99,6%.

I. Introduction

The Federal Republic of Germany is one of the developed countries in the Central European region. Having an average per capita income of $ 41,955, Germany became a country famous for its technological advancements. The country is directly bordered by 9 countries, including the Netherlands, Belgium, Luxembourg, France, Switzerland, Austria, Czech, Poland, and Denmark.

It is strategically located in Central Europe and surrounded by many countries, making the development of road infrastructure in Germany has never stopped. Recorded at the end of 2018, the length of national roads that exist throughout Germany reached 229,970 KM. In addition to the highway infrastructure, Germany also has many of the largest Hospitals and Institutions of Higher Education in their respective regions.

The population is a group of individuals with similar characteristics (species) that live in the same place and have the ability to reproduce among themselves. While the population is a person who occupies a certain area and is bound by the laws applicable. Thus, the population is a group of people who occupy a particular area and are bound by the laws applicable in the area. Highway infrastructure is the minimum length of a motorway that can be passed by car in a country measured by kilometers. The calculation of this highway covers the country's main roads, roads that are bypassed by public transport and ordinary public roads that are minimally passable by a car. The hospital is a professional health care institution whose services are provided by doctors, nurses, and other health professionals. The hospital data used is the number of rooms available for the patient. Higher education institutions are post-secondary institutions that include diploma programs, undergraduate programs, master programs, doctoral programs, and professional programs, as well as specialist programs, run by both government and private sector.

The relationship between these three aspects of infrastructure and population numbers can be modeled by using spatial regression analysis. Spatial regression is a method to model a data that has spatial elements.

One of the objectives of spatial regression in the field of infrastructure is for the preparation of government development programs in infrastructure. Therefore, this study has problem formulation: (1) Is spatial regression suitable for the number of infrastructure development in Germany in 2018

DOI:

(2) How is the regression model suitable for the large number of infrastructure developments taking place in Germany in 2018 (3) What are the factors that affect the population development.

II. Method

Researchers use secondary data obtained in the form of ready-made or already collected from other sources and obtained from other parties such as literature books, journals, internet, and records or sources related to the problem under study are not data obtained through research directly to the field.

The analysis used by the researcher is spatial regression using SAR. Then do some analysis test to determine the best model. The steps taken in solving this problem are:

A. *Determine the Problems*

The problem faced is whether the development of infrastructure, education, hospitals, and highways has a significant impact on population growth in a region. In this case regions is provinces in Germany

B. *Literature review*

1. Spatial Statistics

Spatial statistics is a statistical method used to analyze spatial data. Spatial data is data that contains information "location", so not only "what" measurable but indicates the location where the data is located. Spatial data may include information regarding the geographic location such as the location of the latitude and longitude of each border region and between regions. Simply put spatial data is expressed as the address information. In another form, spatial data is expressed in the form of grid coordinates as in the grain map or in the form of pixels as in the form of satellite imagery. Thus the approach of spatial statistical analysis is usually presented in the form of thematic map (Fotheringham, 2009).

2. Spatial Autocorrelation

Spatial autocorrelation is an estimate of the correlation between the value of observations relating to spatial locations at the same variable. Positive spatial autocorrelation shows the similarity value from adjacent locations and tend to cluster. While negative spatial autocorrelation shows that the adjacent locations have different values and tends to spread (Fotheringham, 2009). Characteristics of spatial autocorrelation expressed by Kosfeld, namely:

1. If there is a systematic pattern in the spatial distribution of observed variables, then there is spatial autocorrelation.
2. If the proximity or adjacency between regions closer, it can be said there is positive spatial autocorrelation.
3. negative spatial autocorrelation illustrates a pattern adjacency unsystematic.
4. The random pattern of spatial data showed no spatial autocorrelation.

Measurement of spatial autocorrelation for spatial data can be calculated using the Moran's Index (Moran), Geary's C, and Tango's excess. In this study, the analysis method is limited only to the method of Moran's Index (Moran) (Gleditsch, 2008). This method can be used to detect the onset of spatial randomness. This spatial randomness may indicate clusterisation or forming a trend towards space.

3. Moran's Index

Moran's I is a development of the Pearson correlation in the data univariate series. Pearson correlation (ρ) between the predictor variables and the response variable with a lot of data n using formula (1).

$$\rho = \frac{\sum_{i,=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i,=1}^{n}(x_i - \bar{x})^2 \sum_{i,=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

$\bar{x}$ and $\bar{y}$ the Pearson correlation equation is an average sample of predictor variables and the response. P value is used to measure whether the predictor variables and the response correlated.

According to [5]–[7], the coefficient of Moran's I used to test the spatial dependency or autocorrelation between observations or location. The test statistic used formula (2)-(9). The hypothesis is:

H0: I = 0 (no autocorrelation between locations)

H1: I 0 (no autocorrelation between locations)

$$Z_{hitung} = \frac{I - I_o}{\sqrt{\text{var}(I)}} \sim N(0,1) \tag{2}$$

where :

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

$$E(I) = I_o = -\frac{1}{n-1} \tag{4}$$

$$\text{var}(I) = \frac{n^2 S_1 - nS_2 + 3S_o^2}{(n^2 - 1)S_o^2} \tag{5}$$

$$S_1 = \frac{1}{2}\sum_{i \neq j}^{n}(w_{ji} + w_{ij})^2 \tag{6}$$

$$S_o = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \tag{7}$$

$$w_{io} = \sum_{j=1}^{n} w_{ij} \tag{8}$$

$$w_{oi} = \sum_{j=1}^{n} w_{ji} \tag{9}$$

Where xi are variable data to location-i (i = 1, 2, ..., n), xj as the variable data to location-j(j = 1, 2, ..., n), $\bar{x}$ for Data Average, var (I) ar varians Moran’s I, and E(I) is an expected value Moran’s I.

Decision-making reject Ho if $\left|Z_{hitung}\right| > Z_{\alpha/2}$. The value of the index I is between -1 and 1. If I> Io, the data has a positive autocorrelation, if I < Io, the data has negative autocorrelation, and Moran index value is zero indicating no groups. Moran index value does not guarantee the accuracy of measurement if the weighting matrix used are not standardized weighting.

4. Data collection

As explained earlier that the data used is secondary data obtained from the website of the German statistics agency (*destatis.de*). Here is the data obtained:

Table 1. Data

Province	Room of Hospital	Higher Education Institution	Highway	Population
Baden-Wurttemberg	55940	361855	27420	9355239
Bayern	76128	389080	41893	11379653
Berlin	20127	187107	246	2918072
Brandenburg	15291	49269	12190	2413079
Bremen	5184	37149	114	578877
Hamburg	12547	107455	190	1492489
Hessen	36170	260184	16106	5307140
Mecklenburg-Vorpommern	10369	39137	10005	1583154
Niedersachsen	41942	209770	28035	7352720
Nordhein-Westfalen	119645	768840	29536	15932038
Rheinland-Pfalz	25248	123211	18370	3717802
Saarland	6490	31517	2048	933397
Sachsen	25902	111550	13436	3979538
Sachsen-Anhalt	15894	54212	10945	2247873
Schleswig-Holstein	16053	62057	9874	2683060
Thuringen	15866	49832	9562	2155853

III. Results and Discussion

Researchers want to know the spatial regression of the population number as the variable y (dependent) and the road infrastructure, the infrastructure of educational institutions, and the number of rooms in the hospital as variable x. in carrying out this research, researchers use the help of GeoDa software.

A. *Scatter Plot Matrix*

Scatter plot is used to determine the relationship between variables in the form of linear or not. It is said to be linear when data tends to follow a straight line pattern or can be assumed to follow a line pattern. If the line starts from the left the more right upwards then it is called positive linear, whereas if the opposite is called negative linear.

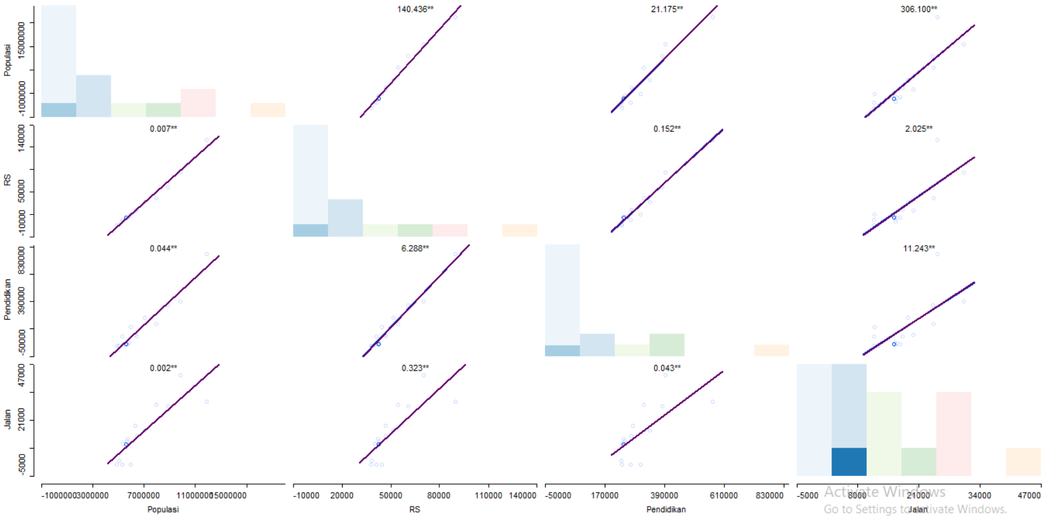


Fig. 1. Scatter Plot Matrix

The four variables are linear positive, it is possible that the variable is included in the best model to be searched. From the figure above it is known that the population tends to form a linear line. Other variables can also be assumed to be linear.

B. *Parallel Coordinate Plot*

Parallel coordinate plot is used to see the relation of the number of values of each variable.

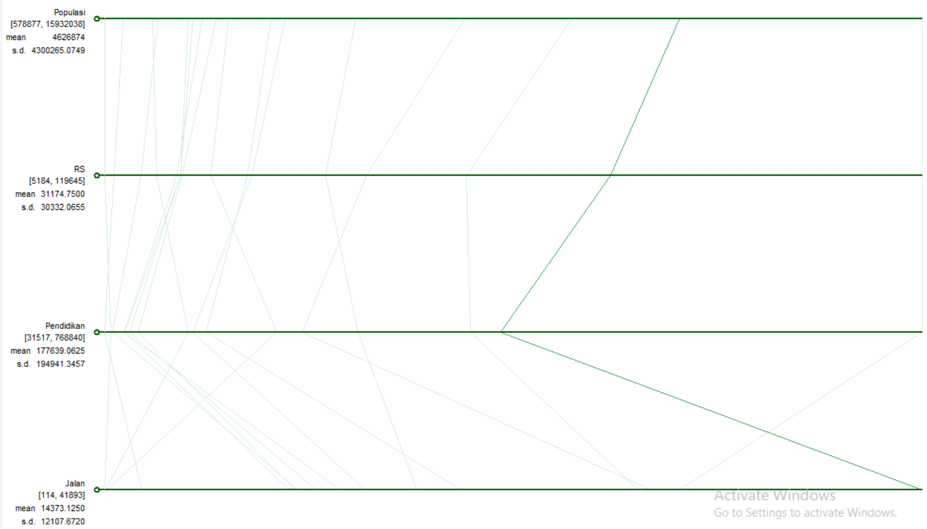


Fig. 2. Parallel Coordinate Plot

On the left side, there is a description of the variable name, the lowest data value, the highest data value, the average, and the standard deviation. Suppose that on the colored green line, the line indicates that the number of population in a particular province is still more than the number of rooms available in the hospital, and the number of rooms in the hospital is still more than the number of educational institutions, but the number of roads is still more than the three previous variables. The parallel coordinate plot is more suitable to compare the number of each variable in different provinces than to use a histogram.

C. *Moran's Index Test*

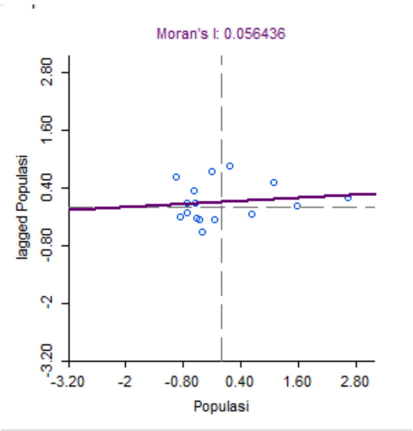
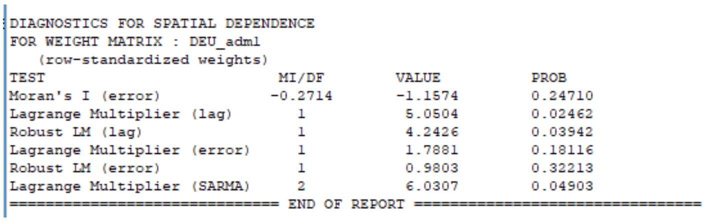


Fig. 3. Moran's Index Test

In the picture above is the result of the Moran index of the dependent variable (Population) in Germany in 2018. It can be seen that the value of Moran's I of 0.056436, this means Moran's I is located at the interval 0 <I <1 which means there is a positive autocorrelation. The pattern of population data in the above picture clustered in quadrant 1.

D. *Lagrange Multiplier Test*

This LM test is the test used for the initial identification of which model is suitable for this case.



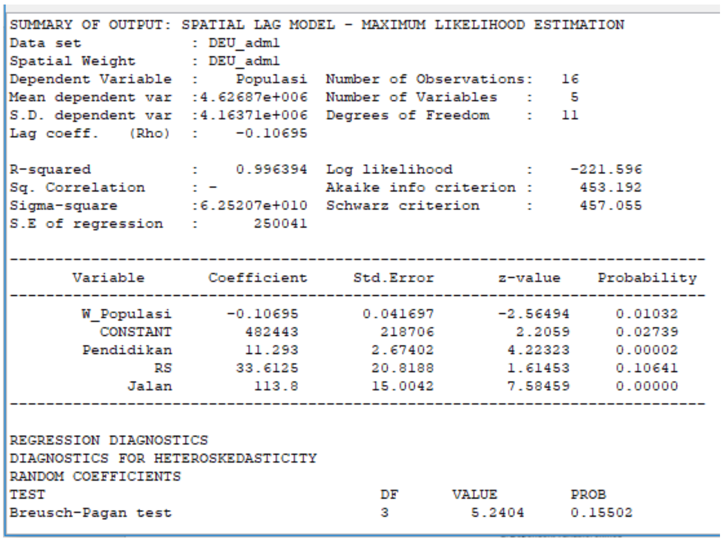
```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : DEU_adm1
   (row-standardized weights)
TEST                           MI/DF        VALUE          PROB
Moran's I (error)             -0.2714      -1.1574         0.24710
Lagrange Multiplier (lag)        1           5.0504         0.02462
Robust LM (lag)                  1           4.2426         0.03942
Lagrange Multiplier (error)      1           1.7881         0.18116
Robust LM (error)                1           0.9803         0.32213
Lagrange Multiplier (SARMA)      2           6.0307         0.04903
============================== END OF REPORT ================================
```

Fig. 4. Lagrange Multiplier Test

If the probability value of lagrange multiplier (lag) <α, then the model made is SAR. If the probability value of lagrange multiplier (error) <α, then model made SEM. If the lagrange multiplier value (SARMA) <α, then the model created is SARMA or a combination of SAR and SEM. α in this case is 0.05. The LM test (lag) aims to identify inter-district linkages. In this case the LM (lag) and LM (SARMA) values are less than 0.05, it is necessary to proceed to SAR or SARMA. But here researchers prefer to use SAR model.

E. *SAR Model*

The suitable model is SAR, the following figure is a summary of SAR regression model analysis. R2 value of 0.996394 which means there is a correlation of 99.64% between the dependent variable with the independent variable.



```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : DEU_adm1
Spatial Weight      : DEU_adm1
Dependent Variable  :    Populasi  Number of Observations:   16
Mean dependent var  :4.62687e+006  Number of Variables   :    5
S.D. dependent var  :4.16371e+006  Degrees of Freedom    :   11
Lag coeff.   (Rho)  :    -0.10695

R-squared           :    0.996394  Log likelihood        :    -221.596
Sq. Correlation     : -            Akaike info criterion :     453.192
Sigma-square        :6.25207e+010  Schwarz criterion     :     457.055
S.E of regression   :      250041

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
        W_Populasi      -0.10695       0.041697       -2.56494     0.01032
          CONSTANT        482443         218706         2.2059     0.02739
        Pendidikan        11.293        2.67402        4.22323     0.00002
                RS       33.6125        20.8188        1.61453     0.10641
             Jalan         113.8        15.0042        7.58459     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF       VALUE        PROB
Breusch-Pagan test                      3         5.2404      0.15502
```

Fig. 5. SAR Model

Based on the hypothesis test, the second variable (hospital) is not significant, so it does not need to be included into the model. To determine the model, we use the coefficient value of the significant variables.

Population = 482443 + (-0,10695 ((11,293) (Education) + (0) (RS) + (113,8) (Road)))

This means that Education has negative impacts to population growth, so if education level increased can be predict that population growth will decrease. It is different with the impact of Road’s Access like Highway, if the number of Highway Increase, can be predicted that the population growth also increased.

IV. Conclusion

Spatial regression is particularly suitable in the case of the large number of populations in Germany in 2018 which is influenced by the construction of educational infrastructure, hospitals, and roads. The model used is: Population = 482443 + (-0,10695 ((11,293) (Education) + (113,8) (Road))). Factors that affect the population significantly are the education and road infrastructure.

References

[1] A. Fotheringham and P. Rogerson, *The SAGE Handbook of Spatial Analysis*. 1 Oliver’s Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications, Ltd, 2009.

[2] M. D. Ward and K. S. Gleditsch, *An Introduction to Spatial Regression Models in the Social Sciences*. Los Angeles: SAGE Publications, Ltd, 2008.

[3] Purwaningsih et.al. 2017.”Spatial data modeling in disposable income per capita in china using nationwide spatial autoregressive (SAR)”International Journal of Advances in Intelligent Informatics: Vol 3, No 2.

[4] PSP, Luh, dkk. 2013. *“Analisis Kemiskinan dengan Pendekatan Model Regresi Spasial Durbin (Studi Kasus Gianyar)”.*

[5] Rahmawati, Rita, dkk. 2015. *“Analisis Spasial Pengaruh Tingkat Pengangguran terhadap Kemiskinan di Indonesia (Studi Kasus Provinsi Jawa Tengah)”.*

[6] Rati, Musfika. 2013. Model Regresi Spasial Untuk Anak Tidak Bersekolah Usia Kurang 15 Tahun di ota Medan.