Analysis of factors affecting the area of forest and land fires in Indonesia uses spatial regression Geoda and SaTScan

Tuti Purwaningsih[a,1,*], Alya Cintami[a,2]

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

ABSTRACT

This study discusses the factors that influence the extent of forest and land fires in Indonesia that relate several other factors such as rain, fire events, and wind speed which were the events during 2015. Forest fires are one of the environmental and forest problems that is a local and global concern. Countermeasures have been carried out for a long time but are relatively low. By looking for the best regression model with a significance level of 0.05 or 95% using the Spatial Autoregressive Model (SAR) method, the coefficient of determination of 25.00% is obtained which can be obtained by the research regression model and leaves 75.00% needed by other variables that are variables changed.

Keywords:
Rainfall
Wind velocity
Fire event
Area of Fire

I. Introduction

Indonesia has the largest tropical rainforest region in tropical Asia. At present, Indonesia's forest area is 144 million hectares, 64.4 is still forested and contains 7 main forest types with variations of up to 18 types of forests, including bamu forest, nipah forest, sago forest and savanna forest [1]. Among the triggers of Indonesia's tropical forests are forest fires, the distribution of uneven rainfall patterns in an area, the effects of wind speed.

During the 1997 forest fires, national mass media reported that 176 companies were accused of forest fires in land clearing, 133 of which were plantation companies. Therefore, the construction of oil palm plantations was one of the causes of 10 million hectares of forest fires in 1997/98 with economic losses reaching US $ 9.3 billion [2]. Calculations in estimating or estimating the level of forest and land fires can use spatial regression to analyze the relationship between rainfall, fire events, and wind speed. In this case, the existing data has a uniform distribution pattern and makes a pattern of adjacent neighbors.

Regression method used in estimating rain frequency uses Spatial Autoregressive (SAR), Spatial Error Model (SEM) and Spatial Autoregressive Moving Average (SARMA). The weight manager used is the proximity of the queen. The results of the analysis using the above method can determine whether there is a spatial effect of data on existing variables.

The purpose of this study was to analyze the size of rainforests and land and see the influence of rain, fire events, and wind speed.

II. Literature Review

A. *Spatial Statistics*

Spatial data has a special method to analyze. Spatial statistics is a statistical method used to analyze it. 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

DOI: 10.26555/jifo.v12i2.a12340

satellite imagery. Thus the approach of spatial statistical analysis is usually presented in the form of thematic maps [3].

B. Spatial Regression Model

The spatial regression model that is formed from a general regression model that gets spatial influences (location). In the spatial regression model, the value of the response variable in the model is formed. There are four models that can be formed from the General Spatial Model:

1. If $\lambda = 0$ and $\rho = 0$ then the equation becomes:

$$y = X\beta + \varepsilon_1$$

   This equation is called the classical linear regression model, namely the regression model without spatial influence.

2. If $\rho \neq 0$ and $\lambda \neq 0$ then the equation becomes:

$$y = \rho \mathrm{W}_1 \mathrm{y} + X\beta + \varepsilon_1$$

   This equation is called regression *Spatial Lag Model* (SLM) or also called *Spatial Autoregressive Model* (SAR).

3. If $\lambda \neq 0$ and $\rho = 0$ then the equation becomes:

$$y = X\beta + u, u = \lambda \mathrm{W}_2 u + \varepsilon$$

   This equation is called regression *Spatial Error Model* (SEM).

4. If $\lambda = 0$ and $\rho \neq 0$ then the equation becomes:

$$y = \rho \mathrm{W}_1 \mathrm{y} + X\beta + u, u = \lambda \mathrm{W}_2 u + \varepsilon$$

   This equation is referred to as *General Spatial Model* (GSM) or *Model Spatial Autoregressive Moving Average* (SARMA).

C. Spatial Autocorrelation

Calculating a correlation between location is a need in spatial model, it is called as Spatial Autocorrelation. Spatial autocorrelation is an estimate of the correlation between the value of observations relating to spatial locations at the same variable. When the spatial autocorrelation is positive value, it shows the similarity value from adjacent locations and tend to cluster. When the value is negative, it shows that the adjacent locations have different values and tends to spread [4]. 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.

There are many Measurement of spatial autocorrelation. The measurement usually used are 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) [4][7]. This method can be used to detect the onset of spatial randomness. This spatial randomness may indicate clusterisation or forming a trend towards space.

D. Moran's I

Moran's Index is the oldest measurement of spatial autocorrelation. Moran's I is developed from Pearson correlation in the data univariate series. Pearson correlation (ρ) between the predictor variables and the response variable with a lot of data n can be formulated as follows [4]:

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

where $\bar{x}$ 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 [6],[7], and [8], the coefficient of Moran's I used to test the spatial dependency or autocorrelation between observations or location.

III. Method

The research method is one way that consists of steps or sequence of activities that function as general guidelines used to carry out research so that what is the purpose of the research is realized. In carrying out this research the author uses secondary data then the data is analyzed by multiple regression then solved by the SAR, SLM, and SARMA methods.

The data used in this study are secondary data obtained from the Central Agency on Statistics in 2015. 34 data collected from the provinces that were the most in Indonesia with:

Y = Area of Fire (Ha)

X1 = Rainfall

X2 =Fire event

X3 = Wind velocity

The general regression model used is as follows:

$$y = X\beta + u, u = \lambda W_2 u + \varepsilon$$

$$y = \rho W_1 y + X\beta + u, u = \lambda W_2 u + \varepsilon$$

Where Y is the response variable matrix (nx 1), X for the independent variable matrix (nx (p + 1)), β for the regression parameter vector coefficient (p + 1) x1, the spatial autoregression coefficient is ρ, λ for the lag coefficient of regression in error resolution | λ | <1, μ for the error vector is assumed to contain the hanging autocorrelation nx1, ε for the error of the soil vector nx1, the normal distribution with zero averages and variants σ 2I, W is the spatial weight of the vector with nxn, and the amount collected n.

There are four models that can be formed from the General Spatial Model:

1. If $\lambda = 0$ and $\rho = 0$ then the equation becomes:

$$y = X\beta + \varepsilon_1$$

This equation is called the classical linear regression model, namely the regression model without spatial influence.

2. If $\rho \neq 0$ and $\lambda \neq 0$ then the equation becomes:

$$y = \rho W_1 y + X\beta + \varepsilon_1$$

This equation is called regression *Spatial Lag Model* (SLM) or also called *Spatial Autoregressive Model* (SAR).

3. If $\lambda \neq 0$ and $\rho = 0$ then the equation becomes:

$$y = X\beta + u, u = \lambda W_2 u + \varepsilon$$

This equation is called regression *Spatial Error Model* (SEM).

4. If $\lambda = 0$ dan $\rho \neq 0$ then the equation becomes:

$$y = \rho W_1 y + X\beta + u, u = \lambda W_2 u + \varepsilon$$

This equation is referred to as *General Spatial Model* (GSM) or *Model Spatial Autoregressive Moving Average* (SARMA).

The placement method used to determine the presence or absence of spatial data effects. Then hypothesis testing will be carried out by looking at Lagrange Multiplier (LM) Error and Lag

1. H_0: $\rho = 0$ (spatial independence *lag*)
   H_1: $\rho \neq 0$(spatial dependencies *lag*)
2. H_0: $\lambda = 0$ (there is no dependence on spatial effects)
   H_1: $\lambda \neq 0$ (there is a spatial effect)
3. H_0: $\lambda = 0$, $\rho = 0$ (there are no spatial lag dependencies and errors)
   H_1: $\lambda \neq 0$, $\rho \neq 0$ (there are slowness and dependencies of spatial errors)

Furthermore, it was carried out using the homoskedasticity test, the Breusch-Pagan test and the Koenker-Bassett test. To test the hypothesis it is used:

H_0 : asumming datahemogeneity is fulfilled

H_1 : homogeneous assumptions of residual data are not fulfilled

Weighting in neighboring areas uses an approach to the type of queen on a chessboard where only the area around it is included in the area that is considered to have relevance to the scale of neighbors is (1). The adjudication element uses vector and matrix. The queen weighting matrix defines W ij = 1 for adjoining areas. Meet with the area of concern, while W ij = 0 to another area. The spatial weighting matrix is a symmetrical matrix and the main diagonal is always zero.

A. *Descriptive Analysis*

Descriptive analysis is an analysis that aims to describe the state of the data. Descriptive analysis in the form of central symptom measures in the form of mean, median, and mode. The size of the spread is in the form of a range of data (range), deviation (standard deviation and variance). The slope size is the population model, the slope coefficient (kurtosis), and the slope coefficient. To display a summary of data, use the command: summary ().

B. *Multiple Linear Regression Analysis*

Multiple linear regression analysis is a linear relationship between two or more independent variables (X1, X2, ..., Xn) with the dependent variable (Y). This analysis is to determine the direction of the relationship between the independent variable and the dependent variable whether each independent variable is positively or negatively related and to predict the value of the dependent variable if the value of the independent variable increases or decreases. The data used is usually interval or ratio scale.

The multiple linear regression equation is as follows:

$$Y' = a + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_n X_n$$

Explanation:

Y’	= Dependent variable (predicted value)
X1 and X2	= Independent variable
a	= Constants (Y’ value if X1, X2 ... .. Xn = 0)
b	= Regression coefficient (value of increase or decrease)

C. *Proporsi*

The proportion means the number / frequency of certain properties that are comparable. A special form in calculating the ratio is proportion.

D. *SaTScan*

SaTScan is free software that analyzes spatial, temporal and spacetime data using spatial, temporal or space-time scanning statistics. The data will be analyzed on the Y variable, namely the area of fire and the area of non-fire land. Data is obtained from the link https://www.bps.go.id and http://sipongi.menlhk.go.id then for Coordinate data per province in Indonesia obtained by using Google Maps.

In the SaTScan software there are 3 menus, namely the Input column, the Analysis column, and the Output column. First on the Input menu some information is obtained:

1. File case

   Format : <zip=location ID> <number of case> <date>

   Location ID is ID case

   Number of Cases, namely the number of cases

   Date = date of case made in date format (example : 12/31/2017)

2. File control

   Format : <zip=location ID> <number of control> <date>

   Location ID is ID case

   Number of Controls, namely the number of cases

   Date = date of case made in date format (example : 12/31/2017)

3. File coordinates

   Format : <zip=location ID> <latitude> <longitude>

   Location ID is ID case

   Longitude and Latitude is a geographical coordinate system used to determine the location of a place on the surface of the earth.

IV. Results and Discussion

Pigure 1 shows the area of fire as a variable Y in this case and Figure 2 shows the conditional map of rainfall, fire events, and wind velocity.

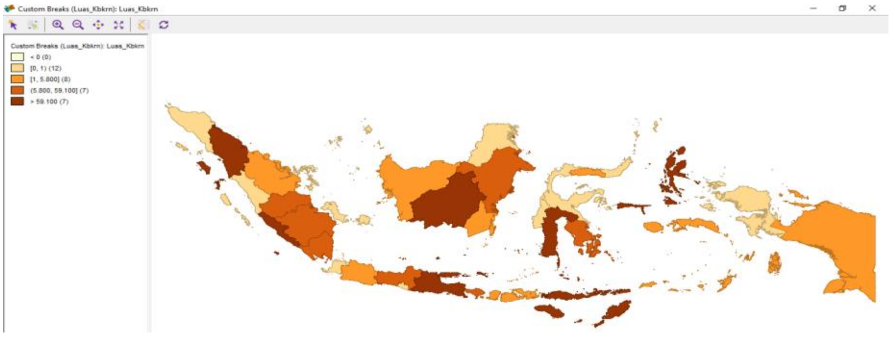

Custom Breaks (Luas_Kbkrn): Luas_Kbkrn
Custom Breaks (Luas_Kbkrn): Luas_Kbkrn
< 0 (0)
[0, 1) (12)
[1, 5.800] (8)
(5.800, 59.100] (7)
> 59.100 (7)

Fig. 1. Area of Fire

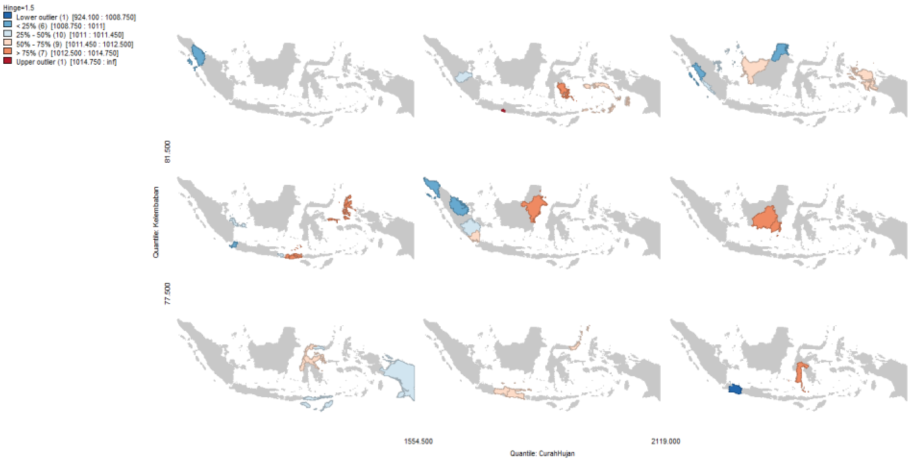


Fig. 2. Conditional Map of Rainfall, Fire Events, and Wind Velocity.

A. *Univariate Moran I Index*

The data in the fire area (Figure 3) shows a spread pattern at one point and has no outlier value. A straight line that has a negative trend because its direction shows downwards means it has positive and not negative values.

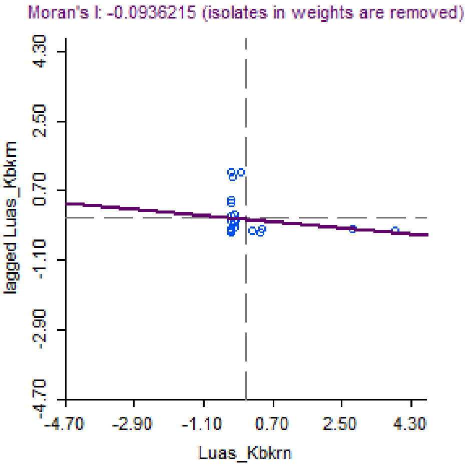


Fig. 3. Moran I Index Area of Fire

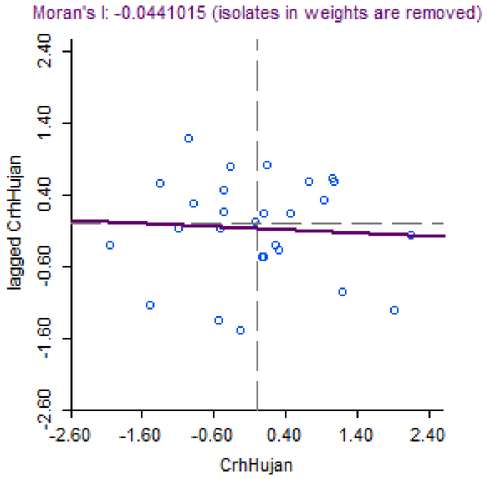


Fig. 4. Moran I Index Rainfall

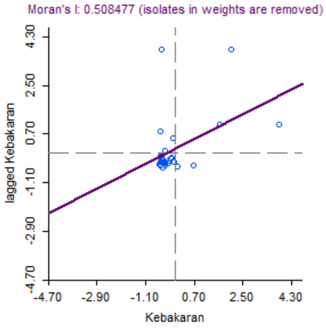


Fig. 5. Moran I Index Fire Event

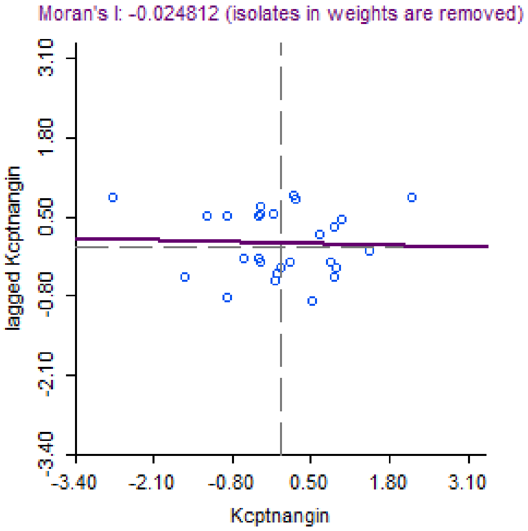


Fig. 6. Moran I Index Wind Velocity

Figure 4 shows the distribution of the Moran I Rainfall Index, Figure 5 shows the distribution of the Moran I Fire Event Index, while Figure 6. shows the distribution of the Moran I Wind Velocity Index.

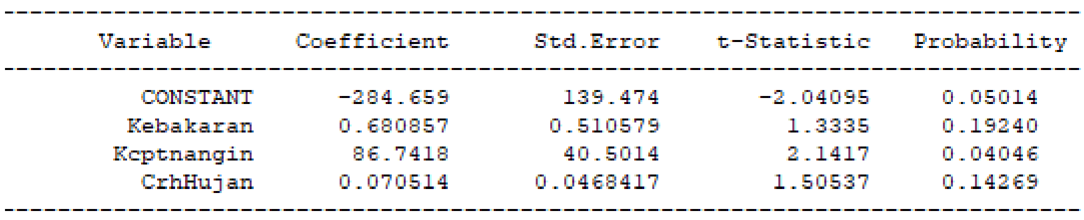
B. *Identification of Variables x and y*

Based on the classical regression analysis the coefficient of determination (R2) is 0.2500 which means the regression model can explain 25% of the total diversity while the remaining 75% is explained by other variables outside the model. The classic regression model that is formed is:

y = 0.680857x1 + 86.7418x2 + 0.070514x3

The interpretation of the above equation is that every 1 increase in X1 will increase Y value by 0.680857, every 1 increase in X2 will reduce Y value by 86.7418, and every 1 increase on X3 will increase the value of Y by 0.070514 (Table 1).

Table 1. Identification of Variables x and y



```
--------------------------------------------------------------------------------
       Variable      Coefficient       Std.Error       t-Statistic   Probability
--------------------------------------------------------------------------------
          CONSTANT      -284.659        139.474        -2.04095       0.05014
         Kebakaran      0.680857       0.510579         1.3335        0.19240
        Kcptnangin       86.7418        40.5014         2.1417        0.04046
          CrhHujan      0.070514      0.0468417        1.50537        0.14269
--------------------------------------------------------------------------------
```

To develop a spatial model, the first step that must be done is to identify using Lagrange Multiplier (LM). This LM test (Table 2) uses weighting on the assignment of queen pattern weights. The value of the test used for analysis includes the value of LM Error and LM Lag.

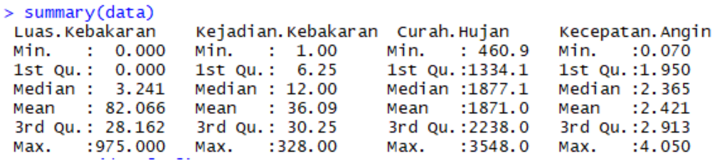
Table 2. Identification Lagrange Multiplier (LM)

Test Dependen Spasial	***Value***	***Probability***
Lagrange Multiplier (Lag)	0.2545	0.61392
Langrange Multiplier (Error)	2.3600	0.12448

Because the probability value of the Lagrange Multiplier Lag is 0.61392 and the Lagrange Multiplier Error is 0.12448 which means more than 0.05, there is no spatial effect on the data.

C. *Descriptive Analysis*

According to the above analysis because there is no spatial effect, it is followed by descriptive analysis. The descriptive analysis output is as in Figure 7.



```
> summary(data)
 Luas.Kebakaran    Kejadian.Kebakaran  Curah.Hujan      Kecepatan.Angin
 Min.   :  0.000   Min.   :  1.00     Min.   : 460.9   Min.   :0.070
 1st Qu.:  0.000   1st Qu.:  6.25     1st Qu.:1334.1   1st Qu.:1.950
 Median :  3.241   Median : 12.00     Median :1877.1   Median :2.365
 Mean   : 82.066   Mean   : 36.09     Mean   :1871.0   Mean   :2.421
 3rd Qu.: 28.162   3rd Qu.: 30.25     3rd Qu.:2238.0   3rd Qu.:2.913
 Max.   :975.000   Max.   :328.00     Max.   :3548.0   Max.   :4.050
```

Fig. 7. Descriptive Analysis

From the picture above, information obtained on the variable area of fire is obtained. Min Value. is the value of respecting the data obtained 0,000. 1 Qu. Is the first quartile of 0,000. Median is 3.241 for the area of fire. Means or an average of 82,066 obtained from the amount of data collected with a lot of data available. 3 Qu. the third quartile is 28,162, while the largest data value is Max. Obtained 975,000.

In the variable Fire event. Min Value. is the value of respecting the data obtained 1.00. 1 Qu. Is the first quartile of 6.25. Median is 12.00 for fire incidents. Means or an average of 36.09 obtained from the amount of data collected with a lot of data available. 3 Qu. the third quartile is 30.25, while the largest data value is Max. obtained 328.00.

In the variable Rainfall. Min Value. is the value of respecting the data obtained 460.9. 1 Qu. Is the first quartile of 1334.1. Median is 1877.1 for rain. Means or an average of 1871.0 obtained from the amount of data collected with a lot of data available. 3 Qu. the third quartile is 2238.0, while the largest data value is Max. Obtained 3548.0.

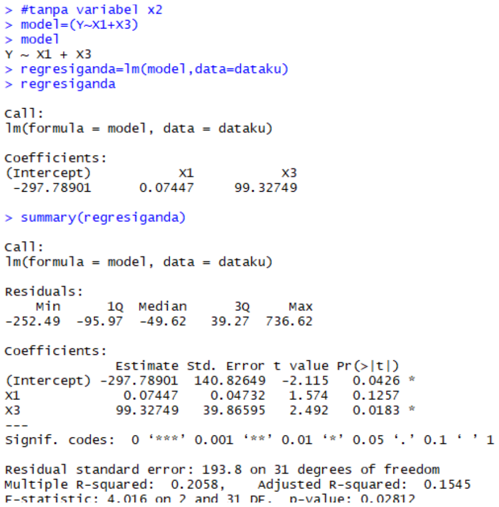
On the variable Wind velocity. Min Value. is the value of appreciating the data obtained by 0.070. 1 Qu. The first quartile is 1,950. Median is 2,365 for wind speed. Means or an average of 2,421 obtained from the amount of data collected with lots of data available. 3 Qu. the third quartile is 2,913, while the largest data value is Max. Obtained 4,050.

D. *Multiple Linear Regression Analysis*

So that the best regression model is obtained based on significant variables, namely a simple linear regression analysis model (Figure 8):

y = b0 + b1X3

y = 0.0426 + 0.0183X3



```
> #tanpa variabel x2
> model=(Y~X1+X3)
> model
Y ~ X1 + X3
> regresiganda=lm(model,data=dataku)
> regresiganda

Call:
lm(formula = model, data = dataku)

Coefficients:
(Intercept)           X1           X3
 -297.78901      0.07447     99.32749

> summary(regresiganda)

Call:
lm(formula = model, data = dataku)

Residuals:
    Min      1Q  Median      3Q     Max
-252.49  -95.97  -49.62   39.27  736.62

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -297.78901  140.82649  -2.115   0.0426 *
X1             0.07447    0.04732   1.574   0.1257
X3            99.32749   39.86595   2.492   0.0183 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 193.8 on 31 degrees of freedom
Multiple R-squared:  0.2058,    Adjusted R-squared:  0.1545
F-statistic: 4.016 on 2 and 31 DF,  p-value: 0.02812
```

Fig. 8. Multiple Linear Regression Analysis

E. *Proportion*

The results of the proportion for calculating the ratio are as follows.

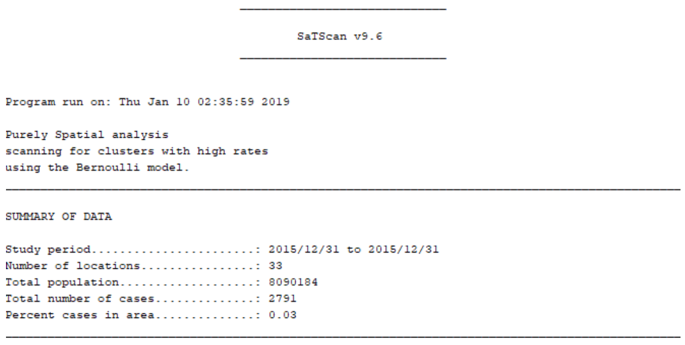
$$proportion = \frac{area\ of\ fire}{number\ of\ provinces}$$

$$= \frac{2790.23}{34}$$

$$= 82.06559$$

F. *SaTScan*

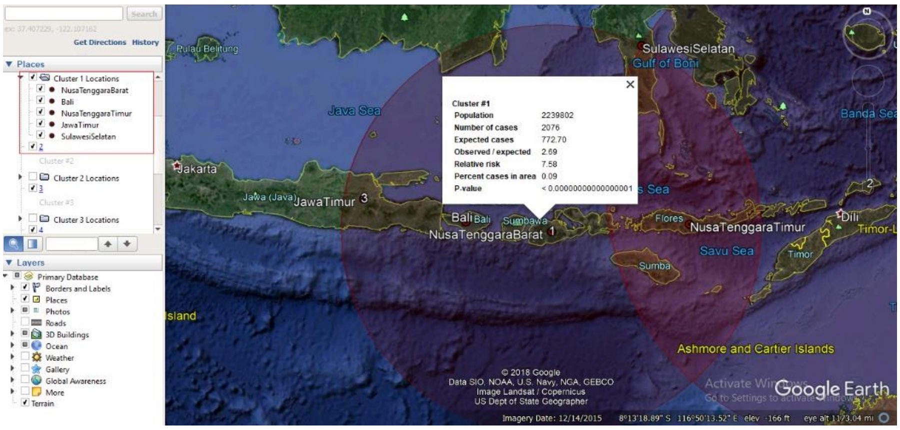
From the fire area data, non-fire land area data, and Coordinate data, a summary of the SaTScan software will be obtained as in Figure 9.



```
                    SaTScan v9.6

Program run on: Thu Jan 10 02:35:59 2019

Purely Spatial analysis
scanning for clusters with high rates
using the Bernoulli model.

SUMMARY OF DATA

Study period.......................: 2015/12/31 to 2015/12/31
Number of locations................: 33
Total population...................: 8090184
Total number of cases..............: 2791
Percent cases in area..............: 0.03
```

Fig. 9. Summary SaTScan

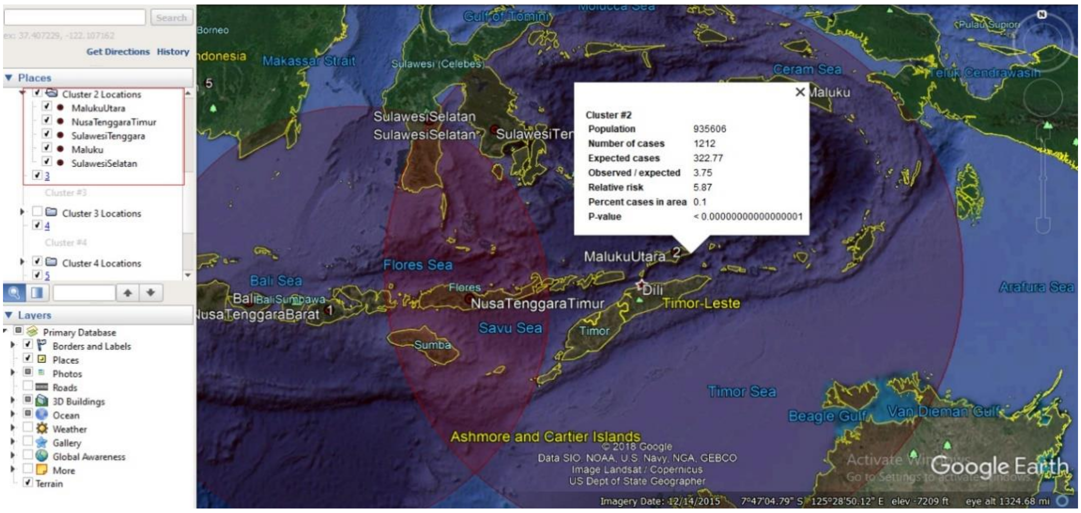
Information was obtained that there were 33 locations inputted in the SaTScan, the total population obtained was 8090184 people. For the total number of cases obtained 2791 and the percentage of cases in area 3%. The next output is the division of clusters. From the Figure 10, it can be seen that there is the first cluster of 2239802 located in the provinces of West Nusa Tenggara, Bali, East Nusa Tenggara, East Java and South South. With the case percentage of the area is 0.09% and p-value <0.00000000000000001. The number of cases is 2076. Conclusions obtained on provinces that have been obtained on the most extensive land



Cluster #1
Population 2239802
Number of cases 2076
Expected cases 772.70
Observed / expected 2.69
Relative risk 7.58
Percent cases in area 0.09
P-value < 0.00000000000000001

Fig. 10. Cluster #1

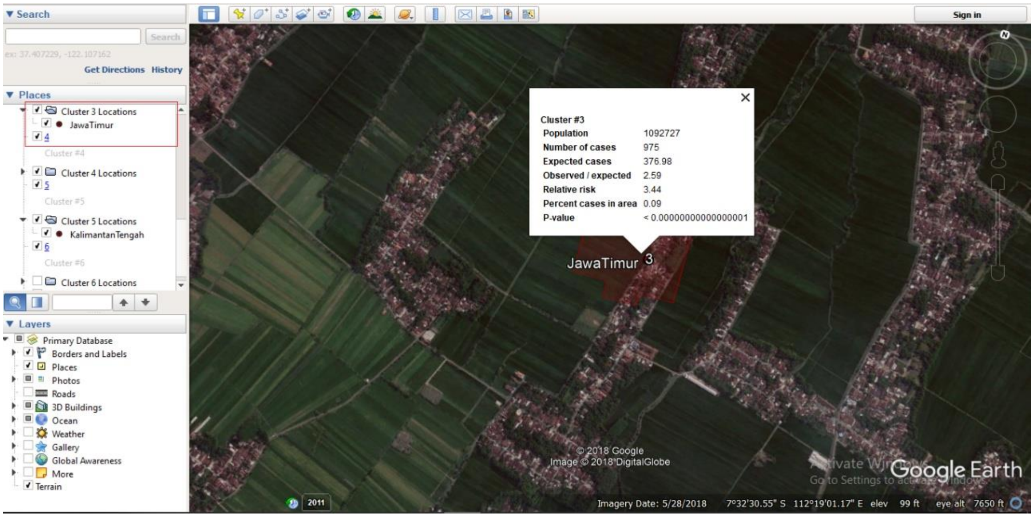
From the Figure 11, it can be seen that the population of the second cluster is 935606 people, where locations are in North Maluku, East Nusa Tenggara, Southeast Sulawesi, Maluku and South Sulawesi provinces. With a percentage of cases in area of 0.1% and having the same p-value with other clusters, which is <0.00000000000000001. Number of cases was 1212 during 2015.



Cluster #2
Population 935606
Number of cases 1212
Expected cases 322.77
Observed / expected 3.75
Relative risk 5.87
Percent cases in area 0.1
P-value < 0.00000000000000001

Fig. 11. Cluster #2

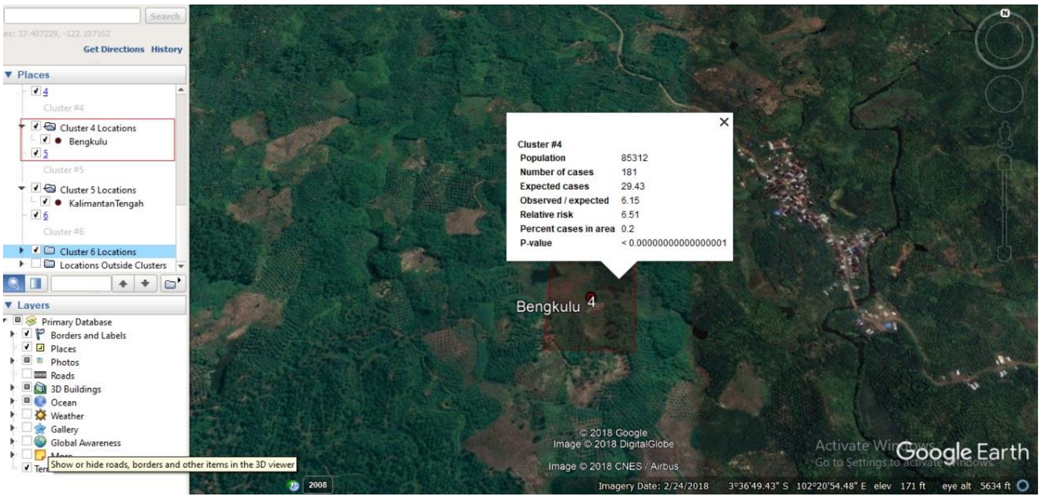
From the Figure 12, it can be seen that the population is 1092727 people, where the location is only in East Java Province. With a percentage of cases in area of 0.09% and have the same p-value with other clusters which is <0.00000000000000001. Number of cases is 975.



Cluster #3	
Population	1092727
Number of cases	975
Expected cases	376.98
Observed / expected	2.59
Relative risk	3.44
Percent cases in area	0.09
P-value	< 0.00000000000000001

Fig. 12. Cluster #3

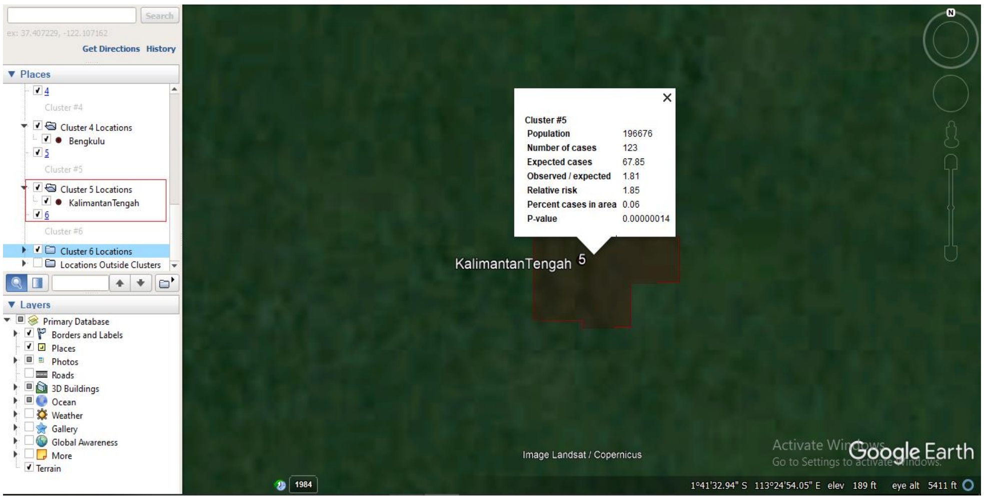
From the Figure 13, it can be seen that the population is 85312 people, where the location is only in Bengkulu province. With a percentage of cases in the area of 0.2% and having the same p-value with other clusters which is <0.00000000000000001. Number of cases is 181.



Cluster #4	
Population	85312
Number of cases	181
Expected cases	29.43
Observed / expected	6.15
Relative risk	6.51
Percent cases in area	0.2
P-value	< 0.00000000000000001

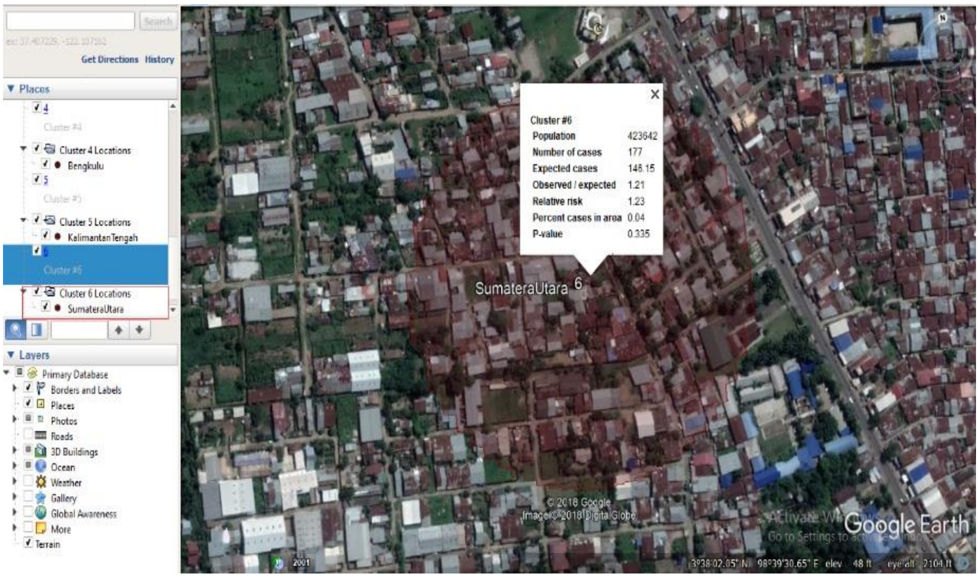
Fig. 13. Cluster #4

From the Figure 14 and Figure 15, it can be seen that the population is 196676 people where the location is only in one province, Central Kalimantan. With a percentage of cases in the area of 0.06% and have the same p-value with other clusters which is equal to 0.00000014. Number of cases as many as 123.



Cluster #5	
Population	196676
Number of cases	123
Expected cases	67.85
Observed / expected	1.81
Relative risk	1.85
Percent cases in area	0.06
P-value	0.00000014

Fig. 14. Cluster #5



Cluster #6	
Population	423642
Number of cases	177
Expected cases	146.15
Observed / expected	1.21
Relative risk	1.23
Percent cases in area	0.04
P-value	0.335

Fig. 15. Cluster #6

V. Conclusion

Based on research using GeoDa and SaTScan software, the results are the most influential factor for the area of fires in 2015 is wind velocity, the determination coefficient is obtained by 25% and only leaves 75% which means that as many as 25% are successfully solved by the regression equation and the rest are explained by other factors.

From the LM Lag and LM Error results, there is no spatial effect on the area of fire data that is affected by rainfall, fire events, and wind velocity. Due to the absence of spatial effects, it was followed by descriptive analysis, multiple regression analysis, and proportions to get the best regression model. Obtained 6 clusters on data on fire land area and area of non-fire land from 34 provinces in Indonesia in 2015, where only 11 provinces have ever experienced land fire. Provinces that have land fires are West Nusa Tenggara, Bali, East Nusa Tenggara, East Java, South Sulawesi, North Maluku, Southeast Sulawesi, Maluku, Bengkulu, Central Kalimantan, and North Sumatra in 2015.

References

[1] https://bps.go.id

[2] https://www.bappenas.go.id/id/

[3] Dubin R. 2009. Spatial Weights. Fotheringham AS, PA Rogerson, editor, Handbook of Spatial Analysis. London: Sage Publications.

[4] Fotheringham, A.S., and Regerson, P.A., (2009), Handbook of Spatial Analysis. Sage Publications, Ltd., London.

[5] Ward MD, Gleditsch KS. 2008. Spatial Regression Models. Los Angeles: Sage Publications, Inc.

[6] Haining Robert. 2004. Spatial Data Analysis Theory and Practice. Cambridge University Press.

[7] Lambert, D.M., Brown, J.P., dan Raymond, J.G.M.F., (2010), A Two-Step Estimator For a Spatial Lag Model of Counts: Theory, Small Sample Performance and Aplication, Journal of Regional Science and Urban Economics, pp. 241-252.

[8] Zhukof, Y., Spatial Autocorrelation, IQQS, Harvard University, Amerika, 2010.