International Journal of Advanced Science and Technology
Vol. 29, No. 8, (2020), pp. 4932-4941

K-Mean Clustering and Linear Regression Model for Grouping and Treatment Prediction on Natural Disasters and Poverty Average Percentage to Decrease the Victims in 2015 – 2019

Bambang Suharjo
Lecturer, Master Program of System Analysis and Operation Research, Sekolah Tinggi TeknologiAngkatanLaut
E-mail: bambang_suharjo@tnial.mil.id; arief. wibowo@budiluhur.ac.id

Abstract

Indonesia as an archipelagic country prone to storms and tsunamis, a country that is a meeting between three active tectonic plates that are prone to earthquakes and volcanoes, as a country in the tropics that is raw against floods and landslides. Various studies have been conducted to examine the natural disasters and poverty and the many victims that arise. Clustering is needed to be done to investigate the number of natural disasters, poverty and disaster victims so that they can be identified in each group and expected to make it easier to map and contribute to efforts to suppress victims and at the same time improve welfare. And then, regression analysis is also needed to make sure the prediction about number of disasters influence poverty and victims. The best cluster was three cluster and the regression analysis give the recommendation that number of disasters influence poverty and victims. Poverty also influence the victims. So, we recommended to develop human resources and their wealth to reduce victims when the disasters happen.

Keywords: *natural disaster, poverty, victims, cluster, regression*

1. Introduction

a. Background

Geographically, Indonesia as an archipelagic country prone to storms and tsunamis, a country that is a meeting between three active tectonic plates that are prone to earthquakes and volcanoes, as a country in the tropics that is raw against floods and landslides. According to Statistics Center Board in the 5 years from 2015 to 2019 there were 10,844 disasters which included volcanic eruption, earthquakes, floods, etc. The number of dead and missing victims reached 6,046. Besides that, as a developing country, the average percentage of poor people is 11.34%.

Various studies have been conducted to examine the natural disasters and poverty and the many victims that arise. Researches[1] and [1a] explains the effect of natural disasters on poverty. Researches [2],[3] explained the effect of poverty on victims. As well as [1] and the National Statistics Board (2020) explained the effects of natural disasters on victims. In addition, various studies have also been carried out to cluster regions based on the level of disaster and the number of victims.

Clustering is needed to be done to investigate the number of natural disasters, poverty and disaster victims so that they can be identified in each group and expected to make it easier to map and contribute to efforts to suppress victims and at the same time improve welfare. And then, regression analysis is also needed to make sure the prediction about number of disasters influence poverty and victims.

ISSN: 2005-4238 IJAST

b. Research Problem

1) How the cluster of regions on disasters, poverty and number of victims
2) How the influence of disaster and poverty on the number of victims

c. Research Benefit

1) Can be used as a model for national disaster management agencies in order to map disasters and victims
2) Can be used as a model to carry out treatment in order to reduce the number of victims of natural disasters

2. Literature Review

a. Disaster, The Victims and Poverty

Natural disasters affect human and animal lives and properties all around the world. In many cases natural disaster caused by differences natural happen. A brief technical description of some of the major natural disasters is as follows:

1) Earthquake: The sudden movement of the earth's crust, causing destruction due to violent activity caused by volcanic action beneath the surface of the earth.
2) Landslides: Sudden collapse of the earth or rock mass from mountains or cliffs due to vibrations on the surface of the earth.
3) Storm: Bad weather in the form of rain or snow caused by strong winds or air currents formed due to unexpected changes in air pressure on the surface of the earth.
4) Flood: Overflow of large water masses exceeds normal limits on dry land. Every year, millions of lives of people, cows and agricultural crops are destroyed due to lack of proper planning and weather forecast.
5) Tsunamis: High ocean waves which are large volumes of displaced water, caused by earthquakes, volcanic eruptions, or other underwater explosions.
6) Volcanic eruption: This is a sudden and violent release of steam, gas, ash, rock, or lava from the surface of the earth that has ejected to a height and spread for several miles.

b. Natural Disaster and The Victims

Natural disaster fatalities include total dead, total affected and total economic losses [3]. Indonesia Statistics Board (2020) defined natural disaster victim are: dead or loss, injured, and suffer. So, from the two sources we can conclude that it is important to decrease fatalities in dead, losses, injured and suffer implied by natural disaster.

c. Natural Disaster and Poverty

The low-income and vulnerable populations who suffer most in natural disasters are women, children, the elderly, the disabled, and ethnic minorities. This study uses correlation and regression analysis to find the relationship between the effects of disasters and different conditions of poverty. He revealed that people living in poverty have a significant positive relationship with deaths from disasters and damage to property, which shows that natural disasters tend to increase poverty. In addition, districts with socially disadvantaged groups are

ISSN: 2005-4238 IJAST

more vulnerable to disasters [1]. Similar research was done by [1a]. He found that disasters implied in socio economy.

d. Poverty and Natural Disaster Victims

Socially and economically marginalized people and environmentally vulnerable areas are disproportionately affected by natural hazards. Identifying populations and places that are vulnerable to disasters is important for disaster management, and it is important to reduce their economic consequences. The influence of vulnerability is one of the factors that influences the vulnerability of impacts due to natural disasters [2]. The other side, [3] said that economic development for poor people implied to decrease the victim disaster. So, it is necessary to look at economic vulnerability when the discussion on vulnerability as a disaster victim will be sought for anticipation.

e. Conceptual Model

From the above literature disaster, poverty, and victims, we could make the conceptual model, as follows.

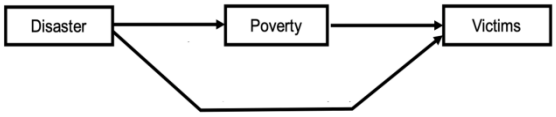

Figure 1. Conceptual model of this research

The conceptual model explained briefly, that disaster influenced the poverty and victims, and poverty influence victims. This model gives me the reason to use the model to make a proposal to decrease poverty and also to decrease the vulnerability of the people from natural disaster explicitly using regression analysis. Before we make regression model, we need clustering method to make sure a biggest population (represented by a biggest cluster) to be first priority for the treatment.

f. K-Mean Cluster

The analysis to be carried out is a very important analysis in data mining, namely cluster analysis. Cluster analysis is one of the most important research directions in the field of data mining. Compared to other data mining methods, clustering can complete data classification without prior knowledge. Grouping algorithm is the process of dividing physical or abstract objects into a collection of similar objects. For grouping, it is done by getting the objects as close as possible to the group [4].

K-Means Clustering Algorithm as one of the clustering methods by partitioning data set into cluster K. This is a distance-based clustering algorithm that divides data into a number of clusters in numeric attributes. The clustering steps are as follows [5]:

1). Determine the number of clusters K and the maximum number of iterations.
2). Perform the K midpoint cluster initialization process, then the centroid count feature equation:

ISSN: 2005-4238 IJAST

$$C_i = \frac{1}{M}\sum_{j=1}^{m} x_j$$

The equation is carried out as many dimensions as p from i = 1 to i = p

3). Connect all observational data to the nearest cluster. Euclidean distance measurements can be found using the equation:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

4). Reallocation of data to each group based on the comparison of distances between data and centroids of each group:

$$a_{ij} = \begin{cases} 1 & d = min\{D(x_i, c_i)\} \\ 0 & others \end{cases}$$

5). Recalculate the position of the cluster midpoint. Then, aij is the membership value of point xi to group centers c1, d is the shortest distance from data xi to group K after comparison, and c1 is the center of group 1. The objective function used by this method is based on the distance and value of data membership in the group. Objective functions can be determined using equations. n is the amount of data, k is the number of groups, ai1 is the membership value from data point xi to group c1 followed by a value of 0 or 1. If the data is a member of a group, the value is ai1 = 1. If not, the value of ai1 = 0.

$$J = \sum_{i=1}^{n}\sum_{l=1}^{k} a_{ic} D(x_i, c_i)^2$$

6). If there is a change in the position of the cluster midpoint or the number of iterations <maximum number of iterations, return to step 3. If not, return the results of the grouping.

g. Linear Regression Analysis

It explained that regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables [6]. In data mining independent variables are attributes already known and response variables are what we want to predict. Types of regression methods: Linear Regression, Multivariate Linear Regression, Nonlinear Regression, and Multivariate Nonlinear Regression.

It explained that multiple regression analysis is a method for estimating and predicting characteristics or the trend of population elicited by analyzing collected data [7],[8]. The main purpose is to estimate a value of the dependent variable when designating the value of the independent variable. Multiple regression analysis shows the straight-line relationship of the first function between more than two independent variables and a dependent variable, as shown in Formula:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + + \beta iXi + \varepsilon$$

ISSN: 2005-4238 IJAST

X herein, refers to the independent variable and Y is the dependent variable. ε is the constant number and β0, β1, β2, ... , βi are regression coefficients. It is important to review the validity of estimated multiple regression analysis and representation and accuracy on the given data. R^2, which is a coefficient of determination and VIF, is a variance inflation factor, are applied as a method to specify the degree of multiple regression analysis among various specification methods.

3. Research Method

a. **Collecting data**. Data collected from sources obtained from the Statistics Central Board data (2015-2019) consists of 2 tables, namely a table on disasters containing the number of disasters, deaths and disappearances, injuries, and suffering for each province during 2015 - 2019. The second table contains about the average percentage of poverty in each province during 2015 - 2019.
b. **Cleaning data**. Done starting from checking whether there is data that is wrong, blank or different types. Because data is taken from tables that have been exported to pdf.
c. **Merging data**. After the data is cleared, merging into one data by paying attention to the order of data according to the province so that data placement errors do not occur.
d. **Cluster data**. Data clustering was performed using the k-mean cluster method so that a grouping was obtained.
e. **Regressing data in the most dominant cluster**. Regression was conducted to determine the effect of the variable number of disasters on poverty, the effect of poverty on the number of victims and the effect of the number of disasters on many victims.

4. Result and Discussion

Data was obtained from two tables from the Central Statistics Board (2015-2019), cleared and merged into table as below.

Table 1. Data Provinces, number of disasters, dead or loss, injured, suffer and average of poverty (%)

Province	Number of Disasters	Dead or loss	Injured	Suffer	Average of Poverty (%)
1. Aceh	464	134	1,357	1,313,683	16.398
2. sumaterautara	315	137	117	552,775	9.83
3. sumaterabarat	329	86	174	98,132	6.868
4. riau	158	10	3	400,689	13.298
5. jambi	116	34	8	262,500	7.73
6. sumateraselatan	276	16	34	333,593	13.298
7. Bengkulu	72	48	12	11,344	16.462
8. lampung	89	146	4,024	276,184	13.618
9. Kepulauan Bangka Belitung	56	6	7	36,381	5.138
10. kepulauanriau	35	4	2	1,653	6.076
11. dki Jakarta	71	15	13	307,307	3.698
12. jawabarat	1,452	270	400	2,946,687	8.31
13. jawatengah	3,066	274	537	2,117,308	12.396

ISSN: 2005-4238 IJAST

14. di Yogyakarta	137	30	47	282,434	13.02
15. jawatimur	1,829	191	341	1,578,500	11.502
16. banten	206	339	10,061	654,050	5.42
17. bali	120	34	43	134,725	4.208
18. Nusa Tenggara Barat	178	591	3,298	3,290,735	15.792
19. nusatenggaratimur	102	24	27	882,131	21.818
20. kalimantanbarat	105	14	6	108,500	7.808
20. kalimantanbarat	105	14	6	108,500	7.808
21. kalimantantengah	160	5	5	360,720	5.424
22. kalimantanselatan	224	3	23	130,434	4.732
23. kalimantantimur	336	20	15	292,801	6.1
24. kalimantanutara	27	2	10	6,235	6.682
25. sulawesiutara	102	26	36	73,484	8.11
26. sulawesitengah	58	3,492	4,486	526,702	14.148
27. sulawesiselatan	342	34	168	317,398	9.184
28. sulawesitenggara	82	9	20	80,937	12.292
29. Gorontalo	79	13	0	792,745	17.204
30. sulawesibarat	44	1	8	56,426	11.542
31. maluku	67	6	27	14,053	18.59
32. malukuutara	68	4	74	39,859	6.586
33. papuabarat	25	11	62	7,835	24.306
34. papua	54	17	1,127	15,382	27.92

From the table, we begin to process data to make clusters using Rapid Miner as follow.

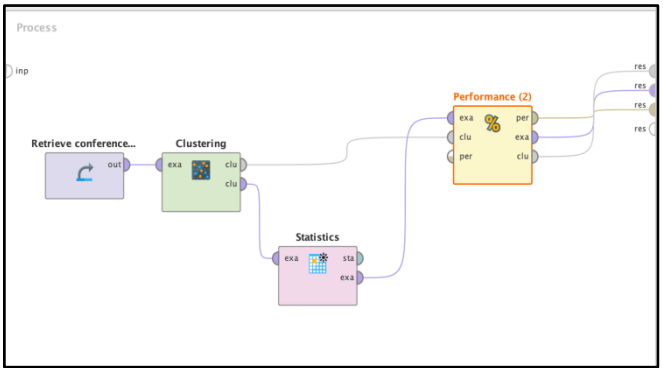

Figure 2. Rapid miner process of clustering

It process was done with many k, such as k=3, k=4 and k=5. The brief report as follows.

Table 2. Brief report in many k

	K=3	K=4	K=5
Davies Bouldin	0.339	0.421	0.330
Cluster	Cluster 0: 2 items Cluster 1: 29 items Cluster 2: 3 items Total number of items: 34	Cluster 0: 24 items Cluster 1: 2 items Cluster 2: 3 items Cluster 3: 5 items Total number of items: 34	Cluster 0: 24 items Cluster 1: 2 items Cluster 2: 2 items Cluster 3: 5 items Cluster 4: 1 item Total number of items: 34

ISSN: 2005-4238 IJAST

Using the table, we recommend to use 3 or 4 cluster. Because the value of Davies Bouldin little near to zero, and every cluster include more than 1 item. From the recommendation, choose k=3 to be analyzed.

Table 3. Attribute and cluster descriptive

Cluster	Provinces	Disaster	Poverty (%)	Dead or loss	Injured	Suffer
Cluster_0	Jawa Barat, Nusa Tenggara Barat	815	12.051	430.500	1849	3,118,711
Cluster_1	All other Provinces (29 Provinces)	132.931	11.073	158.138	711.690	243,358.931
Cluster_2	Aceh, Jawa Tengah, Jawa Timur	1786.333	13.432	199.667	745	169,830.333

From the table 3, we can conclude that:

1. Cluster _0 is characterized by middle number of disasters, middle percentage of poverty and high number of dead or losses, high number of injured and low number of suffer. The provinces are: Jawa Barat, and Nusa Tenggara Barat.
2. Cluster_2 is characterized by high number of disasters, high percentage of poverty, middle number of dead or losses, middle number of injured, and low number of suffer. The provinces are Aceh, Jawa Tengah and Jawa Timur.
3. Cluster_1 is characterized by low number of disasters, low percentage of poverty, low number of dead or losses, middle number of injured and high number of suffer. There are other provinces (29 provinces).

After clustering, we propose the regression process to give a prediction.

Regression process

Regression processes were done using Python, as follows:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
df = pd.read_csv (r'C:\Users\Bambang_\conferenceregres2020.csv')
df.head()
df_x = df.iloc[:, 0:5]
df_x.head()
x_array = np.array(df_x)
print(x_array)
import numpy as np
import statsmodels.api as sm
df_xvar = df.iloc[:, 0:1]
df_xvar.head()
df_yvar = df.iloc[:,1:2]
df_yvar.head()
model = sm.OLS(df_yvar, df_xvar)
results = model.fit()
print(results.summary())
```


ISSN: 2005-4238 IJAST

Result from this python script was:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Poverty   R-squared (uncentered):                   0.374
Model:                            OLS   Adj. R-squared (uncentered):              0.352
Method:                 Least Squares   F-statistic:                              16.76
Date:                Sat, 09 May 2020   Prob (F-statistic):                    0.000326
Time:                        07:47:37   Log-Likelihood:                         -108.00
No. Observations:                  29   AIC:                                      218.0
Df Residuals:                      28   BIC:                                      219.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Disasters      0.0470      0.011      4.094      0.000       0.023       0.071
==============================================================================
Omnibus:                        1.224   Durbin-Watson:                   0.932
Prob(Omnibus):                  0.542   Jarque-Bera (JB):                1.035
Skew:                           0.437   Prob(JB):                        0.596
Kurtosis:                       2.698   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 3. Result of regression process of Disaster to Poverty using python

Continuing with variable Poverty to variable DeadOrLoss, as follows:

```
df_xvar = df.iloc[:, 1:2]
df_xvar.head()
df_yvar = df.iloc[:,2:3]
df_yvar.head()
model = sm.OLS(df_yvar, df_xvar)
results = model.fit()
print(results.summary())
```

The result was:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             DeadOrLoss   R-squared (uncentered):                   0.061
Model:                            OLS   Adj. R-squared (uncentered):              0.027
Method:                 Least Squares   F-statistic:                              1.811
Date:                Sat, 09 May 2020   Prob (F-statistic):                       0.189
Time:                        07:42:09   Log-Likelihood:                         -228.21
No. Observations:                  29   AIC:                                      458.4
Df Residuals:                      28   BIC:                                      459.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Poverty       12.6943      9.434      1.346      0.189      -6.630      32.019
==============================================================================
Omnibus:                       67.608   Durbin-Watson:                   2.024
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              722.520
Skew:                           4.860   Prob(JB):                    1.28e-157
Kurtosis:                      25.438   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 4. Result of regression process of Poverty to DeadOrLoss using python

From the regression model in Figure 3 and Figure 4, we get

Poverty = 0.0470(Disaster).
$R^2 = 0.374$

ISSN: 2005-4238 IJAST

So, we can conclude that is disaster reduce 1 time, the percentage of poverty reduce 4.7% with the determination coefficient = 37%.

DeadOrLoss = 12.6943(Poverty).
$R^2 = 0.061$

So, we can reduce Dead or Losses using reduce of percentage poverty. Every percentage of poverty reduce 1%, Dead or losses reduce 12.69 people with the determination coefficient = 6.1%.
From the two result, we need that disaster should be reduced and also percentage of poverty to be reduced to make sure that dead or loss are reduced.

5. Conclusion and Recommendation

Indonesia disaster, percentage of poverty and victims can be cluster in 3 cluster with its specific characteristic such as: Cluster _0 is characterized by middle number of disasters, middle percentage of poverty and high number of dead or losses, high number of injured and low number of suffer. The provinces are: Jawa Barat, and Nusa Tenggara Barat.Cluster_2 is characterized by high number of disasters, high percentage of poverty, middle number of dead or losses, middle number of injured, and low number of suffer. The provinces are Aceh, Jawa Tengah and Jawa Timur. Cluster_1 is characterized by low number of disasters, low percentage of poverty, low number of dead or losses, middle number of injured and high number of suffer. There are other provinces (29 provinces). From regression analysis, it can be conluded that disaster reduce 1 time, the percentage of poverty reduce 4.7% with the determination coefficient = 37% and Every percentage of poverty reduce 1%, Dead or losses reduce 12.69 people. From the two results, we need to reduce disasters and also percentage of poverty to be reduced to make sure that dead or loss are reduced. In conclusion, the authorproposes that integrating socially disadvantaged groups' vulnerability into disaster mitigation policies can fundamentally reduce the loss of human lives and the economic loss of a community from natural disasters.

6. References

6.1 Journal Article

[1] A.M.Sufiyan, Disaster and Poverty: The Differential Impacts of Disaster on The Poor in The Gulf Coast Region. Dissertation. Department of Urban and Public Affairs University of Louisville Louisville, Kentucky **(2014)**

[2] S.Jeong and D.K. Yoon, Examining Vulnerability Factors to Natural Disasters with a Spatial Autoregressive Model: The Case of South Korea. Sustainability 2018, 10, 651; doi:10.3390/su10051651 **(2018)**

[3] J. Padli, M.S. Habibullah andA.H. Baharom, The impact of human development on natural disaster fatalities and damage: panel data evidence, Economic Research-EkonomskaIstraživanja, 31:1, 1557-1573, DOI: 10.1080/1331677X.2018.1504689**(2018)**

ISSN: 2005-4238 IJAST

[4] C.YuangandH. Yang, Research on K-Value Selection Method of K-Means Clustering Algorithm. Multidiciplinary Scientific Journal. J 2019, 2, 16; doi:10.3390/j2020016.**(2019)**

[5] M.A.Syakur, Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering, 336(1). https://doi.org/ 10.1088/1757- 899X/336/1/012017**(2018)**

[6] FestimHaliliand AvniRustemi,Predictive Modeling: Data Mining Regression Technique Applied in a Prototype. International Journal of Computer Science and Mobile Computing, Vol.5 Issue.8, August- 2016, pg. 207-215**(2019)**

[7] S. Gupta, A Multiple Regression Technique in Data Mining. International Journal of Computer Applications (0975 – 8887) Volume 126 – No.5, September 2015. **(2015)**

[8] Y.S. Song and M.J. Park, Development of Damage Prediction Formula for Natural Disasters Considering Economic Indicators. Sustainability 2019, 11, 868; doi:10.3390/su11030868 **(2019)**

6.2. Book

[1] A.GunadiBrata, The Socio-Economic Impacts of Natural Disasters: Empirical Studies on Indonesia. VU Universiteit Amsterdam: Dissertation **(2017)**

Authors

Author's Name :Bambang Suharjo
NIDN :4706107001
Bird of Date : Sukoharjo, 6 October1970
Address : Grand Victoria – AS5 no 11JatisariPermai - Bekasi
Telp/Email : +6281235258341/ (031)3955208, bambang_suharjo@tnial.mil.id

Education
Bachelorin Mathematics, Faculty of Mathematics and Natural Science, Universitas GadjahMada (1995)
Master in Statistics,Faculty of Mathematics and Natural Science, ITS– Surabaya (2003)
PhD in Mathematics Education, Universitas Negeri Surabaya (2011)

Publication
[1] Using System Dynamics to Analyze the Leadership Style on Motivation and Soldier's Performance.E3S Web of Conferences **125**, 22002 (2019)
[2] The Naval Harbours Priority Development Using Zero-One Matrix Decision Variable (ZOMDV) And Fuzzy MCDM Methods; A Case Study. *International Journal of Civil Engineering and Technology (IJCIET)* Volume 10, Issue 02, February 2019,pp.623-634 (2019)
[3] The Simulation of Navy Fleet Placement Model using Covering Technique and BinaryMatrixDecisionVariable.*JournalofEngineeringandAppliedScience*.Year 2019. Volume 14. Issue 12. Page No. 4139 –4145 (2019)
[4] Failure Mode Effect and Criticality Analysis (FMECA) For Determination Time Interval Replacement of Critical Components In Warships Radar. *Journal of Theoretical and Applied Information Technology 31st May 2019. Vol.97. No10 (2019)*
[5] Failure Risk Analysis Glass Bowl Production Process Using Failure Mode Effect Analysis and Fault Tree Analysis Methods (A Case Study). Vol 9 No 2 (2018): *International Journal ofASRO (2018)*

ISSN: 2005-4238 IJAST