Jurnal Ilmiah
KURSOR
Menuju Solusi Teknologi Informasi

Vol. 8, No. 4, Desember 2016

ISSN 0216 – 0544
e-ISSN 2301– 6914

A DATA ANALYSIS OF THE IMPACT OF NATURAL DISASTER USING K-MEANS CLUSTERING ALGORITHM

[a]Prihandoko, [b]Bertalya
[a,b]The Faculty of Computer Science and Information Technology
Gunadarma University, Jalan Margonda Raya 100, Depok, Jawa Barat, Indonesia
E-mail: pri@staff.gunadarma.ac.id

Abstract

Indonesia is one of the country with a lot of natural disasters occurred every year. The victims of natural disasters, are quite high in terms of the number of deaths, missing people, injuries, sufferings and the number of refugees. Unfortunately, the number of victims is growing from year to year in the last ten years. Thus, based on this condition, this research is carried out in order to analyze the data of the natural disasters and their victims for the last five years. The analysis is intended to know what is the main cause of natural disaster. The series of data about the natural disaster and the weather condition is collected from the government office website. The analysis was carried out by implementing clustering technique to the data, by using k-means algorithm, after data preprocessing completed. The result of the research shows that the weather condition is not the main cause of the occurrence of natural disaster, but the geographical condition is the main trigger of the problem. In addition, this research also found that the data published by the government need to be updated regularly.

Keywords: Natural Disaster, Data Mining, K-Means Algorithm, Clustering

INTRODUCTION

In some developing countries, natural disasters happened because of a high frequency of events and a lack of measures to prevent disaster harm [1]. As a result, damages and deaths resulting from disasters are much higher in developing countries, such as Indonesia, compared to what happened in developed countries [1][2].

Floods, earthquakes, volcanic eruptions, tsunamis, and forest fires are all common in Indonesia, and the scale of these events can sometimes be massive. The geographically dispersed population and uneven development leaves large sections of the population exposed to some level of disaster. However, Indonesia's total risk exposure is considerably lower than many other Asian nations' exposure.

The United Nation's (UN) 2014 World Risk Report named Indonesia the 38th most 'at risk' country for disaster. This is far behind neighboring Philippines 2nd most 'at risk' status. Other Asian nations with higher overall risk levels include Bangladesh, Cambodia, Papua New Guinea, Timor-Leste, Japan, and Vietnam [3].

Natural hazards are the most prevalent threat in Indonesia, driven by its geographic position on the 'Ring of Fire' and location at the boundaries of three tectonic plates. These geographic features essentially cause very high seismicity and a proliferation of active volcanoes. On average, every year nearly half of Indonesia's districts experience at least one natural disaster and many experience multiple disasters. At least, there are three basic measures of disaster impact i.e.: deaths, injuries, and damage to housing resulting from disasters.

In order to analyze the behavior of the climate changes, and to see their impacts to natural disaster and the victims of the disaster, an official data from BNPB (*Badan Nasional Penanggulangan Bencana* or The Agency of Disaster Management) was taken from its website www.bnpb.go.id. In addition, the official data of climate changes was gathered from BMKG (*Badan Meteorologi, Klimatologi dan Geofisika*, or The Agency of Climatology and Geophysics) at www.bmkg.go.id.

Then, the data obtainedwas examined using the data mining technique. Data mining is the process to discover interesting knowledge from large amounts of data [4]. The techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining, and some new techniques such as social network analysis and sentiment analysis.

Clustering is a division of data into groups of similar objects. In this research, we use one of the clustering techniques which is called k-means algorithm. K-means [5] is one of the unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The problem of clustering, according to its attributes, has been widely studied due to its application in areas such as machine learning [6], data mining and knowledge discovery [7, 9], pattern recognition and pattern classification [8]. The aim of clustering is to partition a set of objects that has associated multi-dimensional attribute vectors into homogeneous groups such that the patterns within each group are similar. Several of unsupervised learning algorithms have been proposed which partition the set of objects into a given number of groups according to an optimization criterion. One of the most popular and widely studied clustering methods is k-means [9]. It is algorithmically simple, relatively robust and gives "good enough" answers over a wide variety of data sets.

RESEARCH METHOD

For this research, we took only the data from the province of West Java, as the population is the highest in the country and the number of disasters is quite high. According to the data obtained, we found that the natural disaster mostly happened in West Java is floods. For the need of analysis, we took the data for the last five years, between January 2011 and December 2015.

The number of data available in the website of BNPB which consists of the victims of disasters is 2909 records, from all provinces in Indonesia. For the province of West Java, there are 333 records. Among them, we found some data are missing, and filled with numbers 8888 and 9999. In order to make the analysis more accurate, we eliminated these kinds of data.

In this research, we analyze the relation between the numbers of victims of natural disaster to the weather condition. Usually, when the weather is above normal, for instance when the rainfall is very high, floods are happened in some areas. Thus, the condition of weather is very important to be considered and measured in handling natural disaster.

In order to analyze the cause of floods due to the weather condition, we took data about the climate and weather condition in the province of West Java from BMKG site. The total data captured is 6515 records, but some of them are missing, with numbers 8888 and 9999. After omitting the missing values, the total data captured is 5411 records.

As shown in Figure 1, the research starts with gathering data from the official website of BNPB and BMKG. The data is processed in pre-processing stage, where the data is cleaned. The two sources of data are then joined together in order to make the analysis easier.

The result of data pre-processing is a structured and integrated data from those two sources. Then, the clustering data is carried out in order to analyze the causes and impacts of the weather to the floods.

The field structure of the climate data from BMKG consists of*: climate station, WMO (code of climate station), date, minimum temperature, maximum temperature, average temperature, humidity, rainfall.*

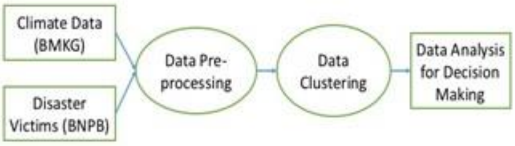

Figure 1. The methodology of research

The field structure of the data from BNPB consists of: *type of disaster, province, city, the date of event, number of death, number of missing, number of injuries, number of suffering, number of refugees.* Type of disasters contains flood, earthquake, landslide, volcanic eruption.

In combining the data, we group the data by monthly and yearly. Thus, we take the average temperature of the weather within a month from the weather table, and add up the total of the victims within the same month from the second table.

After combining the two sources of data, the new table consists of fields: *year, month, average minimum temperature, average maximum temperature, average temperature, average humidity, average rainfall, number of deaths, number of missing, number of injuries, number of suffering, and number of refugees*.

After finishing the data pre-processing, the next process is data clustering. The method of clustering that is used is k-means algorithm. The k-means algorithm is one of the most used clustering algorithms that was first described by Macqueen [5]. It was designed to cluster numerical data in which each cluster has a center called the mean.

The k-means algorithm is classified as a partitioned or non hierarchical clustering method [10]. In this algorithm, the number of clusters k is assumed to be fixed. There is an error function in this algorithm. It proceeds, for a given initial k clusters, by allocating the remaining data to the nearest clusters and then repeatedly changing the membership of the clusters according to the error function until the error function does not change significantly or the membership of the clusters no longer changes. The conventional k-means algorithm [11][12] is briefly described in the following functions.

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the n data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

RESULT AND DISCUSSION

The first process of the research is taking data from the official institutions, which are BNPB and BMKG. In order to have the information about the impact of weather condition to the disaster, we combine the data of the weather with the data of natural victims by a database manipulation process by using SQL statement as follows:

```
select a.Year, a.Month, b.Year, b.Month,
avg(b.MinTemp), avg(b.MaxTemp),
avg(b.AvgTemp),avg(b.AvgHum),
avg(b.Rainfall), sum(a.NumDeaths), sum
(a.NumMissing), sum(a.NumInjured),
sum(a.NumSuffering), sum(a.NumRefugees)
from distwjava a, climwjava b
where a.Year = b.Year and a.Month =b.Month
group bya.Year, a.Month.
```

The variable *distwjava* is the table of disaster victims caused by floods, which consists of the number of deaths, missing people, injuries, suffering, and refugees. It contains 333 records. Another table is called *climwjava*, which is the table of the weather condition, consists of the average minimum temperature, the average maximum temperature, the average temperature, the average humidity, and the average rainfall. It contains 5411 records.

The two tables are joined and grouped by the same year and the month. The result is a table with 48 records.

After completing the cleaning process and integrating the two data, the process proceeds to data clustering, in order to get the pattern of data distribution. The algorithm that is used to do data clustering is k-means algorithm with the value of k = 3. This algorithm gives results in clustering vectors as series of numbers as follows:

2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 1 1 2 2 2 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Number 1, 2 and 3 represent the cluster number. These numbers means that the record from no 1 to no 13 goes into cluster 2, record no 14 goes to cluster 1, records no 15 to 18 goes to cluster 2, record no 19 goes to cluster 1, and so on. Thus, cluster 1 consists of 6 records, cluster 2 consists of 41 records, and cluster 3 consists of 1 record. In cluster 1 there is 6 records, i.e., record no 14, 19, 23, 25, 26, 33.

In cluster 3 there is only record no 30. The record no 30 is really interesting because the number of suffering achieved 21.878.338, while the number of refugees reach 16.310.292.

Table 1 shows that the number of sufferings more than 2 million people in a particular month. This number is also still very high, but in some extreme situations, it could be happened.

Table 1. Data in Cluster 1

Records No	#Deaths	#Missing	#Injuries	#Sufferings	#Refugees
14	0	0	0	2.762.934	191.296
19	123	0	0	4.983.837	235.545
23	102	0	0	4.556.034	154.428
25	0	0	0	2.553.000	15.600
26	0	0	0	2.607.176	46.696
33	0	0	0	2.587.032	6.300

If we compare the number of sufferings and the number of refugees for each record, we find that they are not consistent. For instance, in record no 14, the number of sufferings reach 2.762.934 and the number of refugees is 191.296.

However, in record 33, where the number of sufferings achieves 2.587.032, but the number of refugees is only 6.300. The difference between the two records is so high.

In Table 1 we also find that the number of sufferings are so high (more than 2 million people), and the number of refugees are more than 100.000 people, but the number of missing and injuries are 0. Normally, when the natural disaster happened, and the number of suffering peoples reaches millions of people, there should be missing and injured people among them.

Table 2 shows the relation between the year and the clusters. The record that we are interesting to discuss is record no 33, which is in cluster 3 and in year 2014.

Table 2. Cluster by Year

Year/Cluster	1	2	3
2011	0	10	0
2012	1	7	0
2013	4	7	0
2014	1	8	1
2015	0	9	0

Table 3. Cluster by Month

Month/Cluster	1	2	3
1	1	3	1
2	2	3	0
3	0	5	0
4	1	4	0
5	1	4	0
6	0	2	0
7	0	3	0
8	0	2	0
9	0	2	0
10	0	4	0
11	0	5	0
12	1	4	0

Table 3 shows the relationship between the cluster and the month where the disaster happened. It shows that the disaster with the strange figure is in January. Usually, January is indeed the month with sometimes extreme weather condition.

Figure 2 shows the number of deaths, missing and injuries in correlation with the rainfall. From Figure 2 we could see that, when the rainfall is high, the victim of disaster is also high. The plots in the figure 2 are clustered within rainfall between 5 to 20 mm. However, in the same rainfall value, there is a condition where we got a very high number of deaths or injuries, as in the case of rainfall between 5 and 10 causing 1600 deaths and 1500 injuries. However, in the rainfall between 20 to 25, we also find the number of deaths reach 1500 people.

In the two cases, we believe that there should be other factors that cause the number of victims are so high. As we know, the number of victims in a natural disaster are not only influenced by the weather, but also the ground conditions, the density of the people living around the area, and the conditions of the houses or buildings in the area. In the next research, we will compare the data with the data gathered from some online media.

Figure 3 shows the average of temperature, the average of humidity and the average of rainfall within a month. In terms of temperature, we could see that the temperature is relatively constant within five years, ranging from 20 to 30 degree Celsius. In terms of rainfall, it is high in month January, November and December. The rainfall is low during July, August, September and October. In terms of humidity, it goes low in September and October. It is high in January and February.

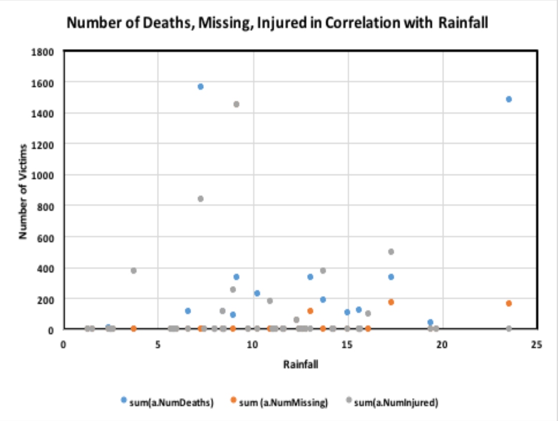

Figure 2. Number of deaths, missing, injuries in correlation with rainfall

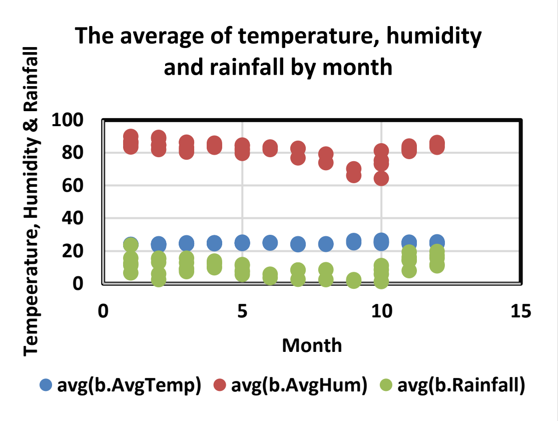

Figure 3. The average of temperature, humidity and rainfall by month

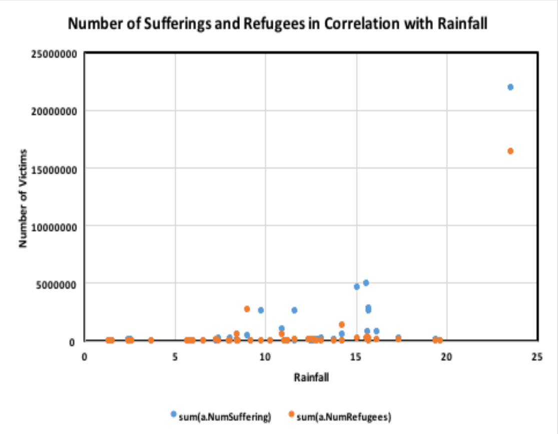

Figure 4. The number of sufferings and refugees in correlation with rainfall

Figure 4 shows the number of people sufferings and the number of refugees as an impact of rainfall. The figure presents the number of refugees is low for the rainfall between 2 to 20 mm. In terms of people suffering because of the rainfall, we found that when the rainfall reached 15 to 16, the number of suffering people reach the number of 5000.000. However, there is an interesting figures when the rainfall around 23 mm, the number of suffering people raised very high to 22 million people, and the number of refugees reach 16 million people.

CONCLUSION

The natural disaster in Indonesia frequently happened, due to the geographical position of the country. Thus, natural disasters mostly occurred as an impact of the natural condition. However, the weather and climate condition has also influenced and triggered the disasters. This research has completed an analysis to the data published by BNPB and BMKG to find out the correlation between the natural disasters happened, the number of victims and the weather condition.

The analysis is completed by cleaning the data, combining them, and clustering them. The clustering process is carried out by using k-means algorithm, a well-known algorithm to do clustering process. The result of clustering shows that the figures do not show a consistent data. For instance, there is a condition where the number of death, number of missing and number of injuries are all 0, where the number of sufferings for that records are more than 2 million people. This condition is almost impossible to happen.

Therefore, this research recommends the government to make correction to the data published officially to public. At the same time, the effort of disaster mitigation should be taken more seriously and consistently in order to minimize the number of victims, which are still quite high, especially in terms of number of sufferings and refugees.

In the future, the research will be continued to obtain the data from all over the country, not only west java province, and with the use of more complete analysis, so that the government or related institution could make a better anticipation work as a mitigation effort.

REFERENCES

[1] P. K. Freeman, M. Keen, and M. Mani, “Being prepared,” *Finance Dev.*, vol. 40, no. 3, pp. 42–5, 2003.

[2] L. J. Henderson, “Emergency and Disaster: Pervasive Risk and Public Bureaucracy in Developing Nations,” *Public Organ. Rev.*, vol. 4, no. 2, pp. 103–119, Jun. 2004.

[3] The World Bank and GFDRR, “Advancing a National Disaster Risk Financing Strategy – Options for Consideration,” Oct. 2011.

[4] W. I. D. Mining, “Data Mining: Concepts and Techniques,” *Morgan Kaufinann*, 2006.

[5] L. M. L. Cam and J. Neyman, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Statistics*. University of California Press, 1967.

[6] Moh’d Belal Al- Zoubi, Amjad Hudaib, Ammar Huneiti, and Bassam Hammo, “New Efficient Strategy to Accelerate k-Means Clustering Algorithm,” *Dep. Comput. Inf. Syst.*, vol. 5, no. 9, 2008.

[7] G. H. Ball and D. J. Hall, “A clustering technique for summarizing multivariate data,” *Behav. Sci.*, vol. 12, no. 2, pp. 153–155, Mar. 1967.

[8] B. Firouzi, T. Niknam, and M. Nayeripour, “A new evolutionary algorithm for cluster analysis,” *World Acad. Sci. Eng. Technol.*, vol. 36, pp. 605–609, 2008.

[9] D. H. Fisher, “Knowledge acquisition via incremental conceptual clustering,” *Mach. Learn.*, vol. 2, no. 2, pp. 139–172, 1987.

[10] Jain, A. and Dubes, R., *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[11] J. A. Hartigan, *Clustering algorithms*. Wiley, 1975.

[12] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979.