K-Means cluster analysis in earthquake epicenter clustering

Pepi Novianti [a,1,*], Dyah Setyorini [a,2], Ulfasari Rafflesia [b,3]
[a] *Department of Statistics, Faculty of Mathematics and Natural Science, University of Bengkulu, Indonesia*
[b] *Department of Mathematics, Faculty of Mathematics and Natural Science, University of Bengkulu, Indonesia*
[1] *pie_novianti@unib.ac.id;* [2] *dyah.setyorini@unib.ac.id;* [3] *ulfasari@unib.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received August 10, 2017
Revised August 15, 2017
Accepted August 18, 2017

Keywords:
Cluster Analysis
K-Means
KL Index
Seismic Activity
Earthquake
Bengkulu Province

ABSTRACT

Bengkulu Province, Indonesia, which lies in two active faults, Semangko fault and Mentawai fault, is an area that has high seismic activity. As earthquake-prone area, the characteristic of each earthquake in Bengkulu Province needs to be studied. This paper presents the earthquake epicenter clustering in Bengkulu Province. Tectonic earthquake data at Bengkulu Province and surrounding areas from January 1970 to December 2015 are used. The data is taken from single-station Agency Meteorology, Climatology and Geophysics (BMKG) Kepahiang Bengkulu. K-Means clustering using Euclidean distance method is used in this analysis. The variables are latitude, longitude and magnitude. The optimum number of cluster is determined by using Krzanowski and Lai (KL) index which is 7. The analysis for each clustering experiment with variation number of cluster is presented.

Copyright © 2017 International Journal of Advances in Intelligent Informatics.
All rights reserved.

I. Introduction

Earthquake is a vibration or a shock that occurs on the Earth's surface due to the sudden release of energy in the earth's crust that creates the seismic waves. The earthquakes are mostly caused by the release of energy produced by pressure exerted by moving plates. The pressure keeps getting larger and eventually reaches the point where it cannot be resisted by the outskirts of the slab so that an earthquake will occur. Based on the cause, the earthquake is divided into tectonic earthquakes, volcanic earthquakes, earthquake debris and artificial earthquake. Tectonic earthquakes are the most common and damaging earthquakes. This earthquake occurs as a result of the movement of the earth's tectonic plates that occurs suddenly, causing vibrations to the surface of the earth. In a tectonic earthquake, not all parts of the earth's plate will have an earthquake. Earthquakes only occur in the areas where two or more plates meet, either on sea or land.

Sumatera is located along the meeting line between The Indo-Australian plate and Eurasian plate. This line meets on the ocean floor at the plate boundary. This plate boundary is called the Sumatran Sub-duction Trench, and it is where the oceanic Indo-Australian plate is slowly descending beneath subducting under the Eurasian continental plate at a rate of about 50-70 mm/year. Moreover, Sumatra passed by 1900 km long fault zone of Sumatra, within or near an active volcanic arc. This active fault divides the Sumatera from the Semangko Gulf to Banda Aceh. These conditions make Sumatra as the earthquake-prone areas which pose major hazards, especially in areas that are high population and surrounding the active fault trace [1-3]. Bengkulu is a province on the Sumatra which lies two active faults: Semangko fault and Mentawai fault. Therefore Bengkulu is also an area that has high seismic activity.

Based on its strength, the largest earthquake in Bengkulu occurred on September 12, 2007 with a strength of 7.9 Ms which is epicenter at 4.67^0S and 101.13^0E. As a result of this disaster, 25 people died, 41 people were seriously injured, and dozens more suffered minor injuries. This earthquake also destroyed thousands of houses, buildings government, houses of worship, educational facilities, health facilities, roads/bridges and irrigation. Other largest earthquake is of 7.3 Ms that occurred on June 4, 2000 and the location of epicenter is 4.77^0S and 102.05^0E within 33 KM under the sea. June 4, 2000 earthquake is one of the destructive earthquakes that occurred in the area of Bengkulu. This earthquake

is approximately 100 km south of Bengkulu, more than 90 people died, hundreds were injured, and thousands of houses and buildings in Bengkulu Province damaged.

Earthquake can cause severe hazard and damage, especially in areas that are close to the epicenter. This should be an important concern by the government and the surrounding community to minimize the impact. Mitigation and preparedness in anticipating the earthquake are still very low and do not have a roadmap in a planned and systematic. In anticipation of the earthquake and its aftermath, it is necessary to study the pattern of earthquake event spatially. Earthquakes involve large amounts of data, particularly time series data. Processing large amounts of data, also known as data mining are growing along with advances in computer technology. It requires the proper way to produce a better conclusion. In this study, the data was analyzed based on the seismic classification by location and magnitude of the earthquakes. Statistically, proximity and characteristic point seismic events can be grouped using cluster analysis [3-4].

Cluster analysis is a multivariate method that searches for patterns in a data set by grouping the observations into clusters. The goal of this method is to find an optimal grouping for which the observations or objects within each cluster are similar (homogeneous). However the clusters are dissimilar to each other (heterogeneous). The distance between the data determines the level of similarity of data. The small distance between the data indicates the high similarity level of the data and on the contrary, the great distance between the data represented the low similarity level of the data [5]. Conventional seismic zoning method to see the characteristics of earthquake in Iran was done with subjective analysis by tectonic structure, crustal characteristic, structural style, age and intensity of deformation, metamorphic activities and other characteristics. But this analysis often occur errors of judgment by eye. Zamani and Hashemi propose an alternative method of earthquake zoning by using hierarchical cluster analysis. This approach is the starting point in earthquake zoning and has a probability to improve and redefine the new data set. This method can also be used in studying neo-tectonics, seism tectonic, seismic zoning, and hazard estimation of the seismogenic areas [6]. The seismic data is one of the most important sources of information to the identification of seismotectonic regions. Pattern recognition of historical and instrumental seismic data provides a more robust and more suitable tool for extracting useful knowledge from large amounts of data. Ansari proposes clustering method based on the fuzzy modification of the maximum likelihood estimation. The comparison between the results of cluster analyses and the seismotectonic models of Iran reveal that it is possible to partition the spatially distributed epicenters of earthquake events into distinct and the comparison shows that the best results will be achieved by the clustering of major events [7].

K-means clustering has been proposed to make the partition of earthquake source zones [8-9]. Weatherill and Burton used K-means clustering method on delineation of shallow seismic source zone in Aegean (incorporating Greece, Albania, Forme Yugoslav Republic of Macedonia (F.Y.R.O.M.), Southern Bulgaria, and Western Turkey). This study developed K-means seismicity partition and estimated seismotectonic in Aegean and also implemented a weighting on K-means analysis. The preferred weighting metric is fault length from earthquake data. The result was that model contained between 20 and 30 clusters which appropriate with seismic source zone in Aegean. Rehman, Burton and Weatherill applied K-means cluster analysis and partitioned earthquake in Pakistan. K-means cluster analysis has used without weighting. This research provided 19 optimal earthquake clusters in Pakistan.

As an earthquake-prone area, the characteristic of each earthquake event in Bengkulu Province needs to be studied and analyzed. Therefore, in this research, authors perform k-means cluster analysis in earthquake epicenter clustering.

II. K-Means Cluster Analysis

K-means cluster analysis seeks to partition the *n* individuals in a set of multivariate data into *K* clusters, where each individual in the dataset is allocated entirely to a particular cluster. As a hard partitioning algorithm, *K*-means cluster analysis is an iterative process. First, data are initially partitioned. Each group is calculated its means and then the data partitioned again by allocating each datum to its nearest means cluster position [8,10]. In its simplest form, this process consists of three stages [11]:

a. Partition objects into *K* initial cluster.

b. Begin by nothing the objects, determine an object into a cluster that which has the closest centroid (mean). (The distance is usually calculated using the Euclidean distance with either standardized or unstandardized observation). Recalculate the cluster centroid to get a new object and for the cluster that lost object. Centroid of the group is calculated by calculating the average value which is performed in (1).

$$C_{kj} = \frac{x_{ikj} + x_{2kj} + \cdots + x_{akj}}{a}, j = 1, 2, \cdots, p \tag{1}$$

where C_{kj} is centroid of group-*k*, variable-*j*, and a for the number of members in the group *k*.

c. Step b is repeated until no more transfer of the object.

Euclidean distance is the distance the most common type selected. Its simplicity is the geometric distance in multiple dimension of space. Euclidean distance is usually calculated from the raw data, and not of the standard data. This method has several advantages, including the distance of any two objects are not affected by the addition of new objects to be analyzed, which may be an outlier. However, the distance can become very large, caused only because of the difference in scale. Euclidean distance is calculated by (2):

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \tag{2}$$

The aim of clustering is partitioning a given data set into disjoint subsets (clusters) so that specific clustering criteria are optimized. The most widely used criterion is the clustering error criterion which for each point computes its squared distance from the corresponding cluster center and then takes the sum of these distances for all points in the data set. This criterion is called clustering error and depends on the cluster centers. [12]. The clustering error is defined in (3).

$$E(m_1, m_2, ..., m_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} I(x_i \in C_k) \|x_i - m_k\|^2 \tag{3}$$

where m_k is the mean of cluster C_k and *I(X)* is 1 if statement *X* is true, 0 otherwise.

Weatherill & Burton state assessment of the results of a clustering is an important consideration in the cluster analysis. There are several index used to determine the optimal number of groups in the cluster analysis, among which is the index Xie & Beni, silhouette index, Calinski & Harabasz index, and the index Krzanowski & Lai [9]. Krzanowski and Lai recommend index measurement to determine the optimal number of clusters of a set of data [13]. Optimal *K* is an index that maximizes Krzanowski & Lai (*KL*). Estimates of *KL* index is expressed by (4) and (5).

$$KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right| \tag{4}$$

$$DIFF(K) = \left[(K-1)^{2/d} WK_{k-1} \right] - \left[K^{2/d} WK_k \right] \tag{5}$$

where d is numbers of dimensions of the data, WK_k is the pooled within-cluster sum of squares of the K partitions, and WK_{k-1} for the pooled within-cluster sum of squares of the K-1 partitions.

III. Methods

In this research, K-Means cluster analysis is performed in earthquake epicenter clustering. The general method is described in the following steps: 1) determine the data to be clustered, 2) Apply K-means cluster analysis to the earthquake data, 3) compute the Krzanowski and Lai criteria for optimum K number of clusters, and 4) discuss each cluster resulted in the analysis.

The object of this study is Bengkulu Province and surrounding areas. Data is taken from single-station BMKG Kepahiang Bengkulu, from January 1970 to December 2015. Data consist of 3325 earthquakes with Ms ≥ 3 and region is 0.54 ºS-8.49ºS and 97.84ºE-105.89ºE. Only earthquakes with magnitude Ms ≥ 5 are considered here. The variables are the latitude, longitude and magnitude of the earthquakes. From the range of longitude and latitude, the earthquakes are not only located on the geographical boundaries of Bengkulu Province. 30.37% of earthquakes spread around the area of West Sumatra, Lampung and the Indian Ocean. In summary, the data used in this study contains 968 events. The spread of epicenter is presented in Fig. 1.

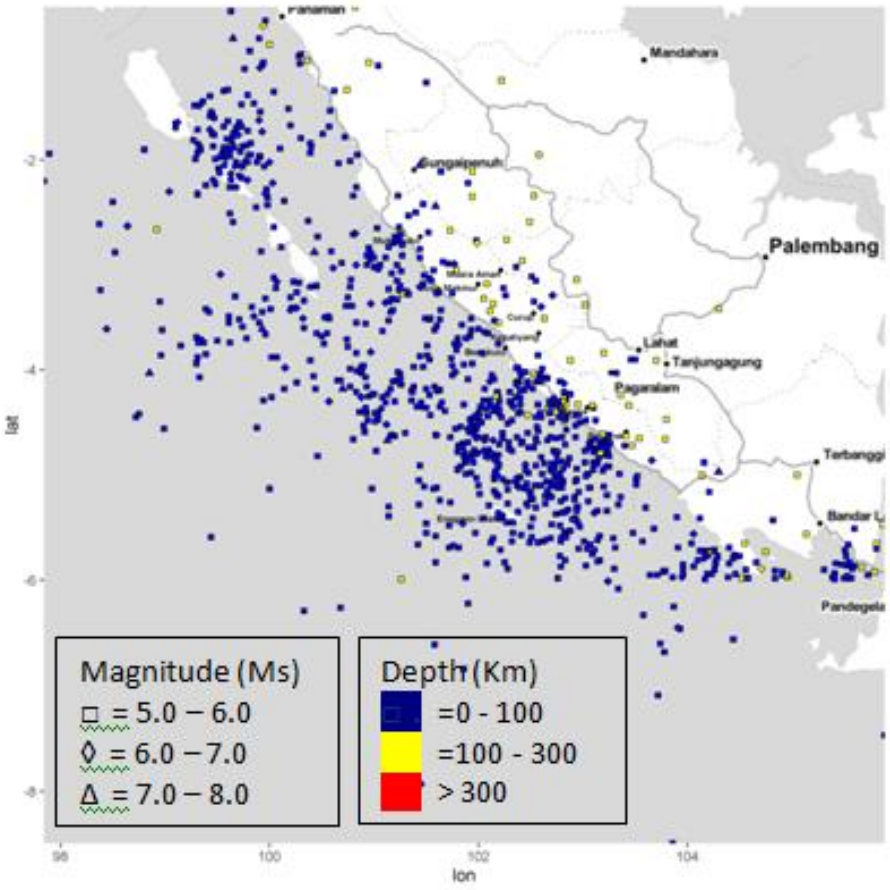

Fig. 1. Earthquakes in Bengkulu Province and its surrounding areas with magnitude ms ≥ 5.0 for period 1970-2015

K-Means cluster analysis is performed using K-Means applied by Weatherill and Burton [8]. K-Means algorithm based on Weatherill and Burton is presented by the flowchart in Fig 2.

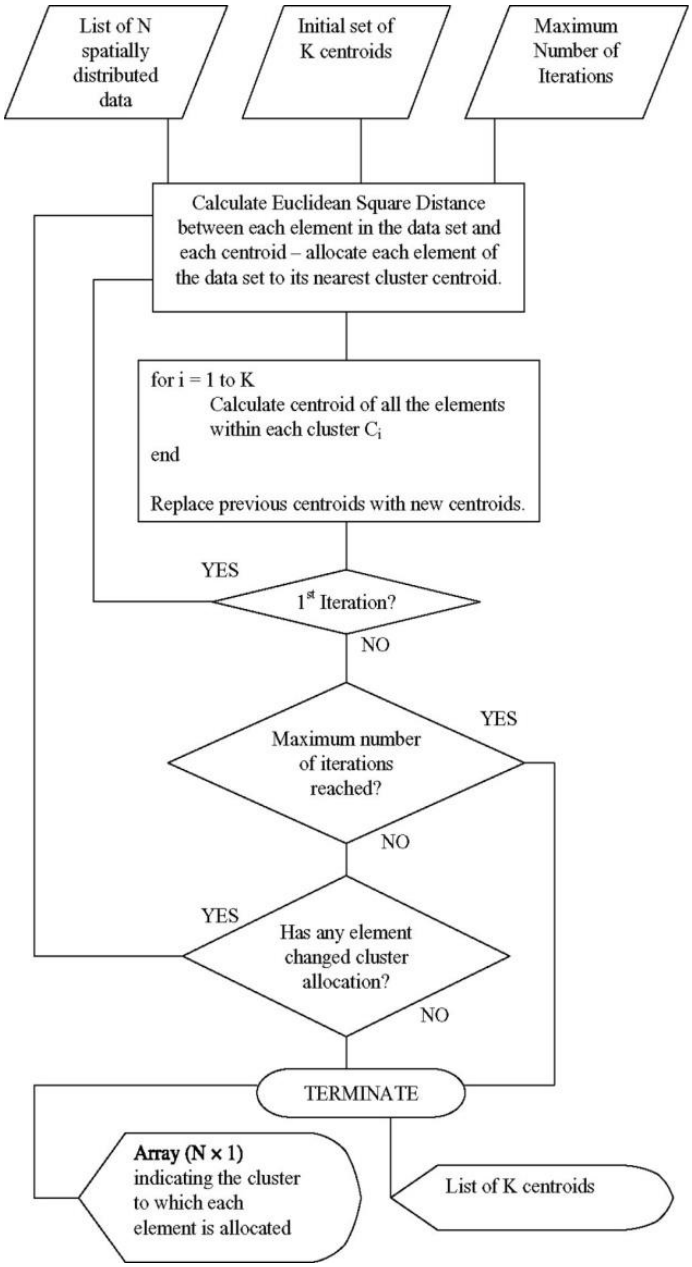

Fig. 2. The flowchart of *K*-Means cluster algorithm

IV. Result and Discussion

A. K-*Means cluster analysis on earthquake data in Bengkulu Province*

The fundamental step in cluster analysis is to determine the optimum K number of cluster. In this study, we use KL index to determine the optimum K. KL index is used to determine the optimum K number of cluster. Based on the KL index of the formed cluster, the largest KL index indicated that the amount of the cluster is the optimal number of cluster.

R Software is used to cluster the earthquakes. Clustering is conducted based on longitude, latitude and magnitude variables. Fig. 3 is the result of KL index value of each cluster formed by the earthquake in Bengkulu Province with longitude, latitude and magnitude variables.

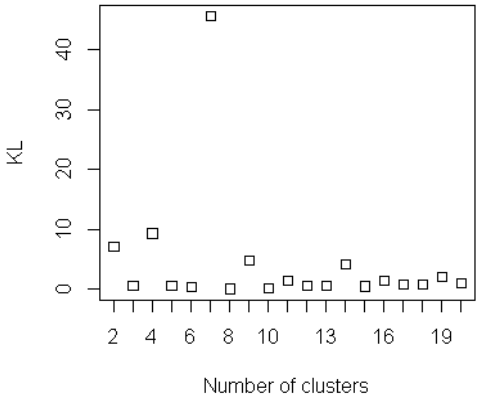

Fig. 3. *KL* Index with *K* numbers of clusters for earthquake during 1970-2015

The largest KL index values are obtained from clustering with K = 7 (KL index = 45.52). Then, the second largest in the cluster with K = 4 followed by 2 clusters 9 clusters and 14 clusters. Meanwhile the other KL index value is less than 2. The next step is clustering earthquake data with the value K = 7, 4, 2, 9 and 14. The result of clustering based on earthquake data is presented in Fig. 4.

As shown in Fig 4. the resulting cluster tends to gather around the epicenter. Clustering results are represented by a different color. This figure is the output of the R software with axis-x as a position longitude and axis-y as latitude position. Figure ellipse in each cluster is a 95% confidence interval based on distribution-t. Ellipse area state a border area of 95% members of each cluster and it does not indicate the epicenter of the earthquake zone.

Fig. 4. (a) shows the results of tectonic earthquake cluster that occurred in Bengkulu Province and surrounding areas to K = 2 cluster. It yields the third largest KL index value after K = 7 and K = 4. This cluster has an epicenter in 102.79ºE and 4.96ºS to the first cluster and 100.38ºE and 2,84ºS to the second cluster. The first cluster with 5.33 Ms occurred around the island of Pagai, Muko-Muko Sea and the Sea of North Bengkulu. The number of seismic activities in this cluster is as many as 591 activities. The smaller the magnitude of the earthquake, the more the incidence of earthquakes. It would be appropriate if we compare it to cluster 2 which has a membership of 377 and an average of large earthquakes is 5.45 Ms.

Fig. 4 (b) is the result of clustering with K = 4. Four cluster division yields an index KL value's second largest, which is 9.49. Fig. 4 (b) shows earthquake occurred around the cluster 1 in Seluma Sea, Manna and Bintuhan with the epicenter position at 102.53ºE and 4.82ºS with a magnitude of an average of 5.32 Ms. The number of members in cluster 1 is as many as 432 activities. The cluster 2 is centered in west coast Lampung and the east end of Sumatra with epicenter 104.54ºE and 5.88 ºS. The average strength of the quake in the cluster 2 is 5.3 Ms and the number of seismic activities is 101 activities. When compared with the results of K = 2 cluster, cluster 1 and cluster 2 at K = 4 are the split of the cluster 1. While cluster 2 is divided into cluster 3 and cluster 4. Cluster 3 at K = 4 occurred in Mentawai Islands of West Sumatra, especially on the island of Sipora. The epicenter of cluster 3 is at 99.74ºE and 1.87ºS with a magnitude of an average of 5.46 Ms. Cluster 4 relatively spreads from epicenter 100.99ºE and 3.61ºS with a membership of 277 seismic activities.

Fig. 4 (c) present the results of clustering for K = 7. Based on KL index value, this cluster is the optimum cluster. At K = 7, cluster 5 is similar to cluster 1 and K = 4, cluster 4 is similar to cluster 2. While cluster 1 and cluster 4 are divided into 5 clusters.

Fig. 4 (d) and (e) is the result of clustering at K = 9 and K = 14. Based on the KL index value, K = 9 and K = 14 is on the order of 4 and 5 with a KL index value of KL 4.71 and 3.74, respectively. In Fig. 4(d), the cluster of earthquakes around the Sipora islands is divided into two, cluster 1 and 7 where is an overlap between two clusters. While the results of clustering K = 14 showed more cluster of overlapping, cluster 5 and cluster 8 located around the sea west Lampung province. Overlap also happened to cluster 14 intersecting with 2, 3, 4, 6, 7, 9, 11, 12 and 13. Custer of K < 7 provided a great diversity in cluster and it caused overlaps in the clustering. Based on the index KL value and cluster results using K-Means analysis, The best cluster is at K = 7. This cluster is expected to be a potential that characterize the seismic activity in Bengkulu.

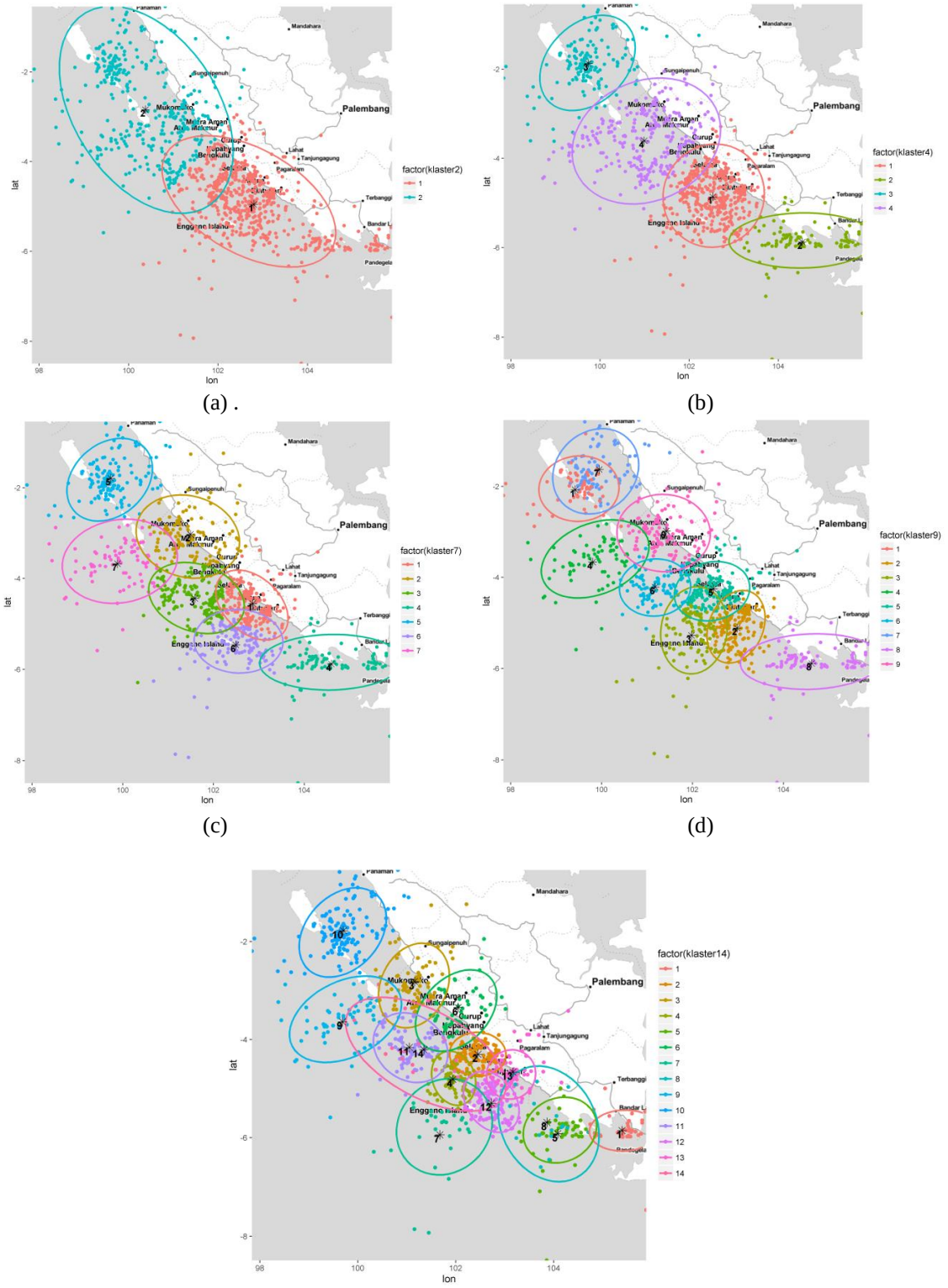

Fig. 4. *K*-Means clustering for earthquake. a. *K*=2, b. *K*=4, c. *K*=7, d. *K*=9 and e. *K*=14

B. Characteristics of 7 clusters in earthquake zone Bengkulu Province

Seismic activity in Bengkulu Province is mostly located in the ocean, i.e. in The Indian Ocean. This is caused by the position of Bengkulu Province is located in the coastal area traversed by traffic lane Indo-Australian tectonic plates and Eurasia west. Besides tectonic earthquakes also occur on land-sourced from Semangko fault that extends in the middle of the island of Sumatra. This seismic

activity using the K-Means grouped into 7 clusters. Based on the analysis of K-Means cluster, the following characteristics of the resulting Table 1.

Table 1. Characteristics of 7 clusters of earthquake in Bengkulu Province and surrounding areas.

Cluster	Longitude Mean	Latitude Mean	Magnitude Mean	Frequency of events	SSE
1	102.86	4.55	5.27	196	73.14
2	101.49	3.03	5.37	122	83.62
3	101.59	4.44	5.43	179	102.33
4	104.59	5.88	5.29	96	79.07
5	99.73	1.79	5.47	141	98.59
6	102.49	5.46	5.35	150	81.84
7	99.88	3.68	5.51	84	72.65

Cluster 1 in Fig. 4(c) is cluster in which the earthquake occurred around Seluma Regency, South Bengkulu Regency and Kaur Regency. both in offshore and in mainland. Cluster 1 has the highest number of members than the other cluster. Based on its depth. There are 85.72% of earthquakes in cluster 1 which include in shallow earthquake and 14.28% of the earthquake is medium earthquakes that frequently occur on the mainland. Cluster 2 is located around Mukomuko Regency, North Bengkulu Regency and Rejang Lebong Regency. The highest earthquake magnitude that has occurred in cluster 2 is the earthquake of 2001 with magnitude 7.4 Ms.

The third cluster has 179 members of earthquakes in the off coast of the city of Bengkulu. The 88.83% members of cluster 3 are the earthquakes with the strength between 5 to 6 Ms. Almost 100% members of cluster 3 are shallow earthquake. Three large earthquakes had occurred in cluster 3. The earthquake of Magnitude 7 Ms with its epicenter located 4.88^0S and 102.19^0E occurred on October 1st, 1975. The large earthquake on June and September 2000 with magnitude 7.3 Ms and 7.9 Ms respectively have occurred in Cluster 3. In Fig. 4 (c) shows that the cluster 4 spreads at sea west of Lampung Province. Seismic events in cluster 4 mostly take place in the sea. The largest earthquake with magnitude 7 Ms of 1994 occurred in 4.97^0S and 104.3^0E.

Cluster 5 consists of seismic event that occurred around the Mentawai Islands of West Sumatra, especially on the island of Sipora. Cluster 5 is a combination of earthquakes in the sea and on land.

Cluster 6 and 7 are the clusters that have 100% of earthquake occurred at the sea of the Indian Ocean and approaching The Indo-Australian plate and Eurasian plate. In cluster 6, 92.67% is an earthquake 5 Ms to 6 Ms and the remaining 7.3% is 6 Ms to 7 Ms. There are 99.33% of the earthquakes those include shallow earthquake. The highest quake of 6.5 Ms occurred in the central point of 102.47^0E and 5.29^0S. On cluster 7 been a big earthquake measuring 7.7 Ms centered in 100.43^0E and 2.88^0S as well as a magnitude 7.2 Ms centered in 99.93^0E and 3.61^0S. One hundred percent members of cluster 7 are shallow earthquakes.

V. Conclusion

The earthquakes in Bengkulu Province and its surrounding areas mostly occur at the Indian Ocean and only small members occur on the mainland. During the period 1971-2015, Around Bengkulu Province are tectonic earthquakes frequently occurred with a strength of 5.0 Ms up to 6.0 Ms and a depth of less than 100 km under the sea. The optimum number of cluster based on KL index is K = 7. The earthquake clustering with K < 7 bring out a large variance in the cluster and earthquake clustering with K > 7 produce overlapping cluster. With the K-Means analysis cluster are obtained 5 earthquakes cluster which are located in the Sea and the mainland province of Bengkulu. One cluster occurred around the Mentawai Islands of West Sumatra province and another occurred at seas of Lampung Province.

VI. Open Problem

The results of the clustering analysis in this study are obtained from the features of geometry using Euclidean distance so that the characteristics of clustering acquired are limited to geometric properties. To make the results of this analysis into the earthquake zone in Bengkulu Province needs to be

reassessed in seismology and geological characteristics. Additionally, these clusters can also be compared with the models of seismic hazard analysis.

Acknowledgements

We thank The Institute of Research and Corporate to society Bengkulu University which has funded this research through funding Research Development. We are also grateful to the BMKG Kepahyang on the given data.

References

[1] D.H. Natawidjaja. "The Sumatran Fault Zone – From Source to Hazard". *Journal of Earthquake and Tsunami*, Vol. 1, No.1, pp 21-47. 2007.

[2] P. Nuannin, O. Kulhánek, and L. Persson. "Variations of b-values preceding large earthquakes in the Andaman-Sumatra Subduction Zone". *Journal of Asian Earth Sciences*, Vol. 61, pp 237–242. 2012.

[3] M. Affan, M. Syukri, L. Wahyuna, and S. Hizir. "Spatial Statistic Analysis of Earthquakes in Aceh Province Year 1921-2014: Cluster Seismicity". *Aceh International Journal of Science and Technology*, Vol.2, 54-62, 2016.

[4] E. Irwansyah and E. Winarko. "Spatial Data Clustering and Zonation of Earthquake Building Damage Hazard Area". *EPJ Web of Conferences*, Vol. 68, No. 5, pp 1-6, 2014.

[5] I.A. Musdar and Azhari. "RCE-Kmeans method for Data Clustering". *International Journal of Advances in Intelligent Informatics*,Vol 1, No 2, pp. 107-114, 2015.

[6] A. Zamani and N. Hashemi. "Computer-Based Self-Organized Tectonic Zoning: a Tentative Pattern Recognition for Iran". *Computers and Geosciences Journal*, Vol. 30, pp 705-718. 2004.

[7] A. Ansari, A. Noorzad, and H. Zafarani. "Clustering analysis of the seismic catalog of Iran". *Computers & Geosciences Journal*. Vol. 35. pp. 475-486. 2009.

[8] G. Weatherill and P.W. Burton. "Delineation of Shallow Seismic Source Zones Using k-means Cluster Analysis. with Application to the Aegean Region". *Geophysical Journal International*, Vol. 176, pp 565-588. 2009.

[9] K. Rehman, P.W. Burton, and G.A. Weatherill. "K-Means Cluster Analysis and Seismicity Partitioning for Pakistan". *Journal of Seismology*, Vol.18, pp. 401-419, 2014.

[10] B. Everitt and T. Hothorn. *An Introduction to Applied Multivariate Analysis with R*. New York: Springer, 2011, pp 163-200.

[11] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. USA: Pearson Education. Inc, 2007, pp 696-702.

[12] A. Likas, N. Vlassis, and J.J. Verbeck. "The global k-means clustering algorithm". *Pattern Recognition*. Vol. 36, pp 451-461, 2003.

[13] W.J. Krzanowski, and Y.T. Lai. "A criterion for determining the number of groups in a data set using sum of squares clustering". *Biometrics*, Vol. 44, pp 23-34, 1988.