Comparison of Clustering K-Means, Fuzzy C-Means, and Linkage for Nasa Active Fire Dataset

Muchamad Kurniawan[1], Rani Rotul Muhima[2], Siti Agustini[3]*
[1,2,3]Teknik Informatika ITATS Arief Rachman Hakim 100, Surabaya, Indonesia
[1] muchamadkurniawan@itats.ac.id; [2]ranimuhima@itats.ac.id; [3]*sitiagustini@itats.ac.id
*corresponding author

ABSTRACT

One of the causes of forest fires is the lack of speed of handling when a fire occurs. This can be anticipated by determining how many extinguishing units are in the center of the hot spot. To get hotspots, NASA has provided an active fire dataset. The clustering method is used to get the most optimal centroid point. The clustering methods we use are K-Means, Fuzzy C-Means (FCM), and Average Linkage. The reason for using K-means is a simple method and has been applied in various areas. FCM is a partition-based clustering algorithm which is a development of the K-means method. The hierarchical based clustering method is represented by the Average Linkage method. The measurement technique that uses is the sum of the internal distance of each cluster. Elbow evaluation is used to evaluate the optimal cluster. The results obtained after conducting the K-Means trial obtained the best results with a total distance of 145.35 km, and the best clusters from this method were 4 clusters. Meanwhile, the total distance values obtained from the FCM and Linkage methods were 154.13 km and 266.61 km.

Keywords : Active fire dataset, K-Means, FCM, Linkage, Elbow Clustering.

Article History		
Received : September, 7th 2020	Revised : November, 26th 2020	Accepted : November, 28th 2020

I. INTRODUCTION

The active fire dataset of the National Aeronautics and Space Administration (NASA) is data obtained from the Visible Infrared Imaging Radiometer Suite Sensor (VIIRS), And the resulting image is a spectroradiometer image as shown in Figure 1. In this dataset, the data features have eight features: latitude, longitude, brightness, scan, track, acq date, acq time, satellite, confidence, version, bright_t31, frp, daylight. This dataset is from NASA's official website (https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/viirs-i-band-active-fire-data).This dataset was studied before. In this study [1], The purpose active fire data are used to prevent forest fires. The features used have been reduced to 2 features: longitude and latitude. The dataset regions are taken only in South and Southeast Asia. The algorithm used is a combination of Local Outlier Factor (LOF) and K-Means. Implementation of the LOF, the accuracy value of K-Means increases compared to Simple K-Means. Certain studies using this dataset include the [2] [3] [4] [5] [6] [7] for the prevention, clustering, and monitoring of forest cover, flare monitoring [8].

Group analysis is to group data into several groups based on data similarity. If there is new data, the similarities in its features will be seen and will be included in certain groups. In group analysis, there are two types of algorithmic approaches, partition-based approaches and hierarchy-based approaches. Partition-based methods include K-means. K-harmonic means, K-modes, Fuzzy C-means, K-Medoid. Meanwhile, based on hierarchy, there are agglomerative linkage methods (single, complete, average), density-based clustering (DBScan), Spectral, and Graph Clustering [9] [10].

For the optimum number of clusters, a partial clustering algorithm needs to be analyzed. Research [11] has contributed to Davies-Bouldin's technical development aside from the advancement of the K-Means method itself. Around the same period, Davies-Boulding and Silhouette index were used to measure the performance of the clustering method [12] [13]. The Dunn and Silhouette experiments are also used to measure clusters on Clustering Large Application (CLARA) and K-Means. By using a statistical approach, research [14] improves the performance of the Dunn index using the K-Means cluster method. Existing Clustering Quality Matrix (CQMs) has been used for internal cluster validity [15].

Our research contributes to the evaluation of the clustering method that best fits this dataset by comparing several methods with cluster measurement using various techniques. In this study, we will use the active fire dataset from NASA with a comparison of partial clustering and hierarchical clustering: the K-means, Fuzzy C-means (FCM), and Linkage. As for the internal cluster analysis, we will use Elbow. The partition of this document shall be divided into four parts: the first part explains the introduction,

DOI: 10.25139/ijair.v2i2.3030

the second part describes the theoretical basis of the methods/techniques used in the research, the results and discussion are in the third part, what conclusions are in the last section.

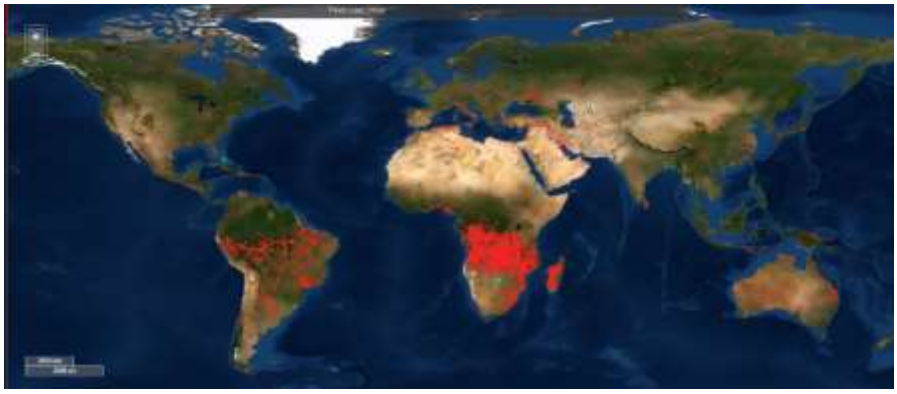


Fig. 1. Digital Image Active Fire

II. METHOD

A. K-Means

The clustering method is used to divide large data into several clusters. There are two types of clustering, namely Hierarchical and Non-Hierarchical. K-Means is one of the non-hierarchical clustering methods. K-Means works as analyzing, modeling, and clustering data by partition system. K-Means is used to cluster data into clusters where the data in a cluster have the same characteristics between each other and have different characteristics with another cluster. In other words, the aim of K-Means Clustering is minimizing the objective function. Minimizing objective function can be obtained by minimizing data variant with another cluster. K-Means algorithm is an iterative algorithm which attempts to partition the dataset into cluster K. The algorithm is continuing as follows:

1. Select initial cluster centers k (centroid)
2. Calculate point-to-cluster centroid distances of each centroid from all observations.
3. Assign each observation of the nearest centroid to the cluster.
4. Assign each of the nearest centroid observations to the cluster.
5. Assign observations to another centroid on a stand-alone basis if the reassignment reduces the sum of the in-cluster, point-to-cluster-centroid distances.
6. Compute centroid each K
7. Repeat steps 2 – 6 until the value of each centroid does not change.

B. Fuzzy C-Means

Fuzzy C-Means (FCM) is a clustering approach that allows numerous clusters with different degrees of membership to belong to each data point. FCM has known as an improved partition clustering system. FCM is based on the following objective function being minimized. The concept of FCM is based on determining the center of the cluster that will mark the average location for each cluster. Each data has a degree of membership for each formed cluster. In the beginning, the cluster center is still inaccurate and repeatedly repairs itself till it is located at the right point. This loop is based on minimizing the objective function, which describes the distance from a given data point to the center of the cluster weighted by the degree of membership of that data point. From the loop, it can be seen that the longer the center of the cluster will move towards that location is right. FCM is satisfied as equation (1).

$$J_m = \sum_{i=1}^{D} \sum_{j=1}^{N} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad (1)$$

where :

- *D* variable is the number of data points.
- *N* is the number of clusters.
- *m* variable is a fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with $m > 1$. Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is, the number of data points that have significant membership in more than one cluster.
- x_i variable is the i^{th} data point.

DOI: 10.25139/ijair.v2i2.3030

- c_j is the center of the j^{th} cluster.
- μ_{ij} is the degree of membership of x_i in the j^{th} cluster. For a given data point, x_i, the sum of the membership values for all clusters is one.

During clustering, FCM performs the following steps:

1. Randomly initialize the cluster membership values, μ_{ij}.
2. Calculate the cluster centers :

$$c_j = \frac{\sum_{i=1}^{D} \mu_{ij}^m x_i}{\sum_{i=1}^{D} \mu_{ij}^m} \quad (2)$$

3. Update μ_{ij} according to the following:

$$u_{ij} = \frac{1}{\sum_{k=1}^{N} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

4. Calculate the objective function, J_m.
5. Repeat steps 2–4 until J_m improves by less than a specified minimum threshold or until after a specified maximum number of iterations.

C. Linkage

There are three Hierarchical clusters such as single linkage, complete linkage, and average linkage. This research uses an average linkage. Average linkage gives results if some clusters are gathered according to the average distance between the pair of cluster membership. A linkage is a gap between two clusters. The following notation describes the linkages used by the various methods:

- Cluster r is formed from clusters p and q.
- n_r is the number of objects in cluster r.
- x_{ri} is the i^{th} object in cluster r.

Average linkage uses the average distance between all pairs of objects in any two clusters as shown in equation (4).

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \quad (4)$$

D. Elbow Evaluation

Elbow is a heuristic method for analyzing and determining the optimal cluster of datasets. First, this method begins with plotting the values as a result of the function of the number of clusters and mark them to the elbow of the curve. This curve gives information about the number of clusters to use. The algorithm of this method is following this step:

1. A initial number of maximum clusters.
2. Repeat until the maximum cluster
   - for i=1 to max cluster
     Calculate the sum of distance each data for its cluster $sumD_i$
   - end
3. Calculate the optimal class by measuring the widest distance to *sumD*

E. Data

From the data taken on 18/08/2020, it is focused on the South Asia region. The results of plotting the data that have been taken are shown in Figure 2. Based on Figure 2.a, there is a red dot that indicates a hot spot on the island of Borneo. Figure 2.b is an enlargement of the hotspots on the island of Borneo only. To obtain data on the island of Borneo, we must first limit the values of latitude and longitude as in Figure 3, with coordinates of A (6,491º, 108,790 º), B (6,491º, 118,524 º), C (-4,039 º, 108,790 º) and D (-4,039 º, 118,524 º). The number of hotspots in the Southeast Asia region is 2582 data, and the amount of data in the Borneo island region is 393 data. This Data will be used as research data.

DOI: 10.25139/ijair.v2i2.3030

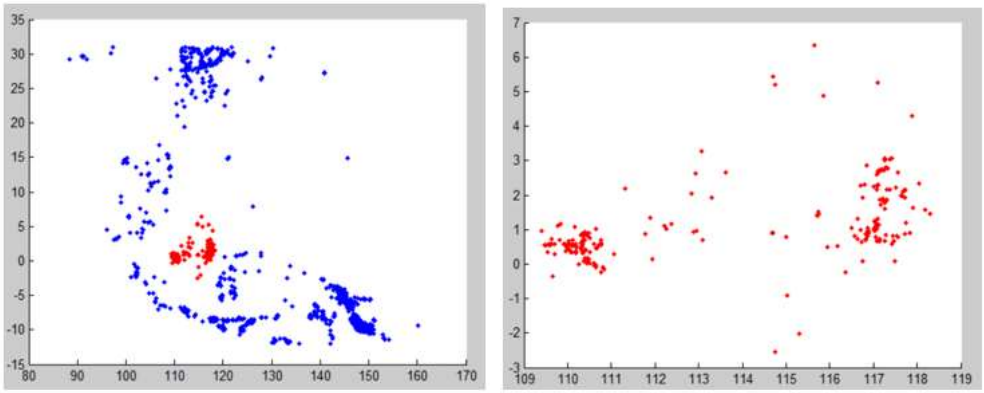


Fig. 2. Plotting data active fire

Fig. 3. Border coordinate Borneo island

F. Methodology

The methodology that we propose can be seen in Figure 4. The first step is to prepare a dataset. Clustering algorithms have been taken to get clusters and members of clusters. The third step is to calculate the sum of the distance from each cluster from its centroid. Tests conducted are until the maximum cluster is achieved, the maximum cluster we use is 20 clusters. After getting the total value of all distances, each cluster that is set will then be analyzed with Elbow Graphic to get the optimal number of clusters.

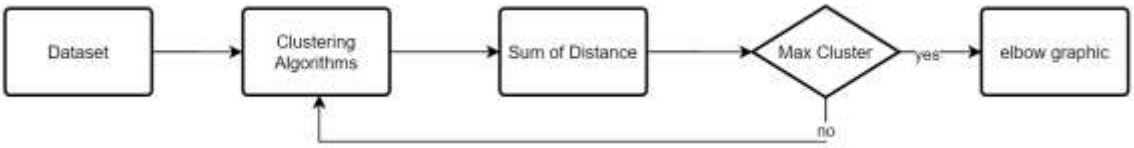


Fig. 4. Research methodology

III. RESULT AND DISCUSSION

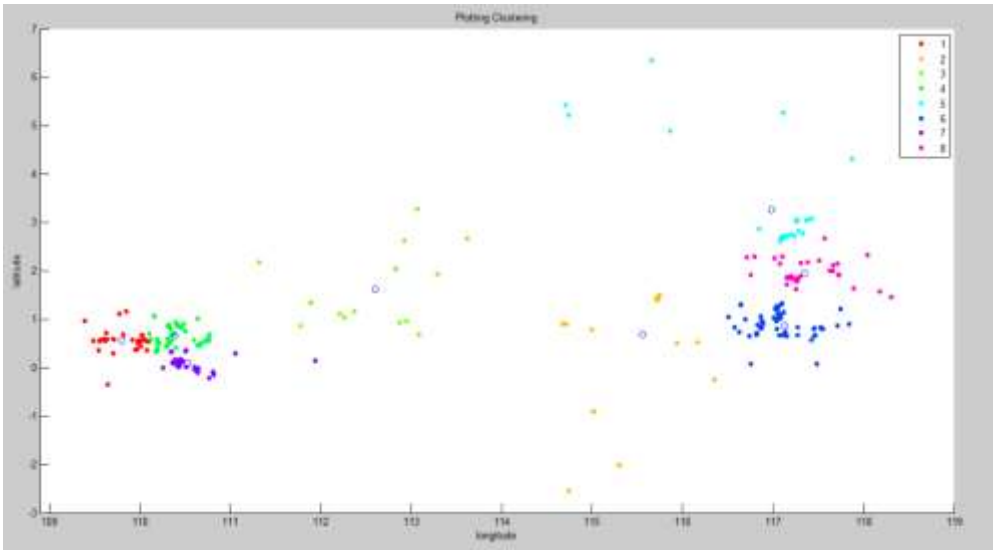
The first trial used was a predetermined number of clusters. The number of clusters determined in eight clusters. The result obtained is the sum of the distance from centroid with members of each cluster, as in Table 1. The results obtained in this table are the results of measuring the internal distance of each cluster. From the results of the sum of distance, the best values are the K-means, FCM, and linkage methods with the results of 145.35, 154.13, and 266.61. All methods have different patterns of cluster member retrieval. This can be concluded from looking at the internal distance analysis for each cluster. The average data deviation value from each cluster in all methods was 22.55. This average deviation value strengthens the previous analysis that each method has a different approach to classifying data. The plotting results of this experiment illustrate the correction from the analysis of the results of table 1 that each clustering method has a different approach. The plotting results of these results can be seen in Figure 5. Figure 5.a is the result of the K-Means method. In this figure, the left and right data is divided into three clusters; the rest is in the middle of the data. Figure 5.b is an approach to the FCM method, the left and right sides of the image are divided into two clusters,

and the rest of the clusters occur in the middle part of the data. In Figure 5.c, the linkage approach has the most different approach from the two previous methods. The result is that the data on the right and left are only one cluster each, and the rest of the clusters are in the middle data section.

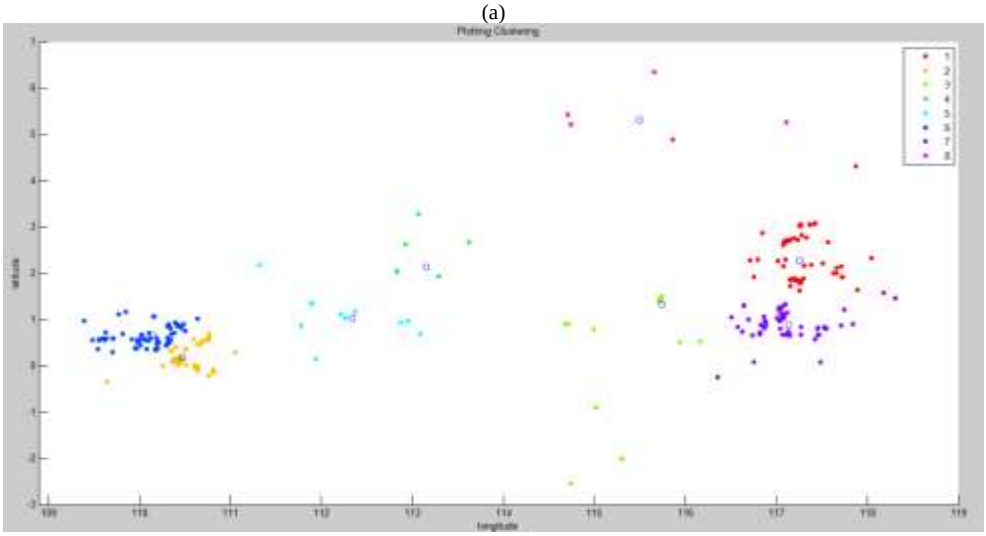
TABLE I
RESULT DISTANCE OF 8 CLUSTERS

sum distance each cluster	n cluster								Sum
	1	2	3	4	5	6	7	8	
K-Means	3.78	45.33	15.77	6.69	47.46	12.64	4.82	8.86	145.35
FCM	31.91	21.28	14.97	36.72	15.92	16.34	10.36	6.63	154.13
Linkage	3.61	6.51	5.58	129.77	6.44	17.13	71.88	25.69	266.61

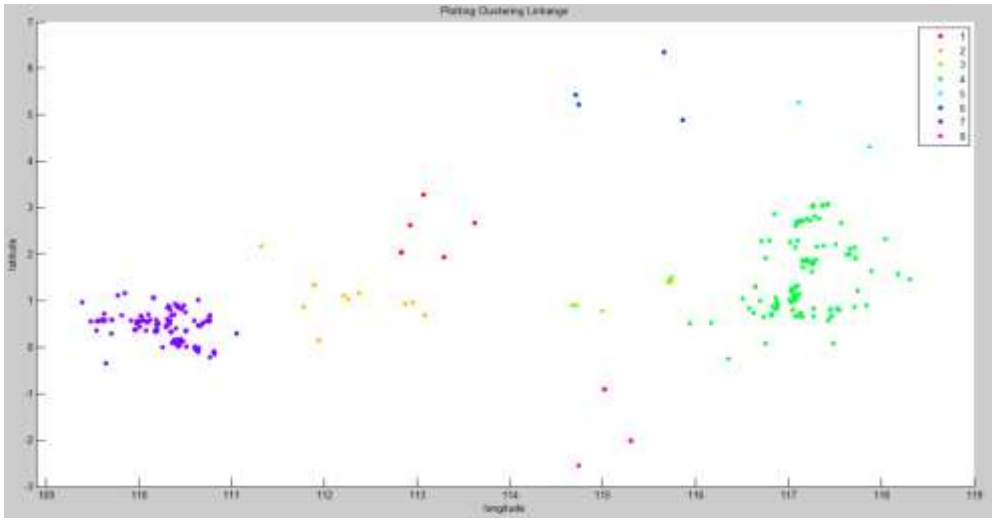
From the results obtained, it can be analyzed that the partial clustering method gets better results when viewed from the total distance obtained. Meanwhile, the hierarchical clustering algorithm brought a difference of 75% greater than the entire distance of the partial algorithm. But if seen from the plotting results, the Linkage algorithm maps each cluster according to the proximity of its neighbors. It is suitable if applied to an island country like Indonesia. Geographically, the distance between the islands is quite far apart from the sea. If we use the Linkage algorithm, the centroid obtained can be right in the middle of the island. Inversely proportional to K-Means or FCM, if applied very likely that the centroid point is received in the middle of the sea.



(a)



(b)



(c)

Fig. 5. Plotting result eight cluster, (a) K-Means; (b) FCM; (c) Linkage.

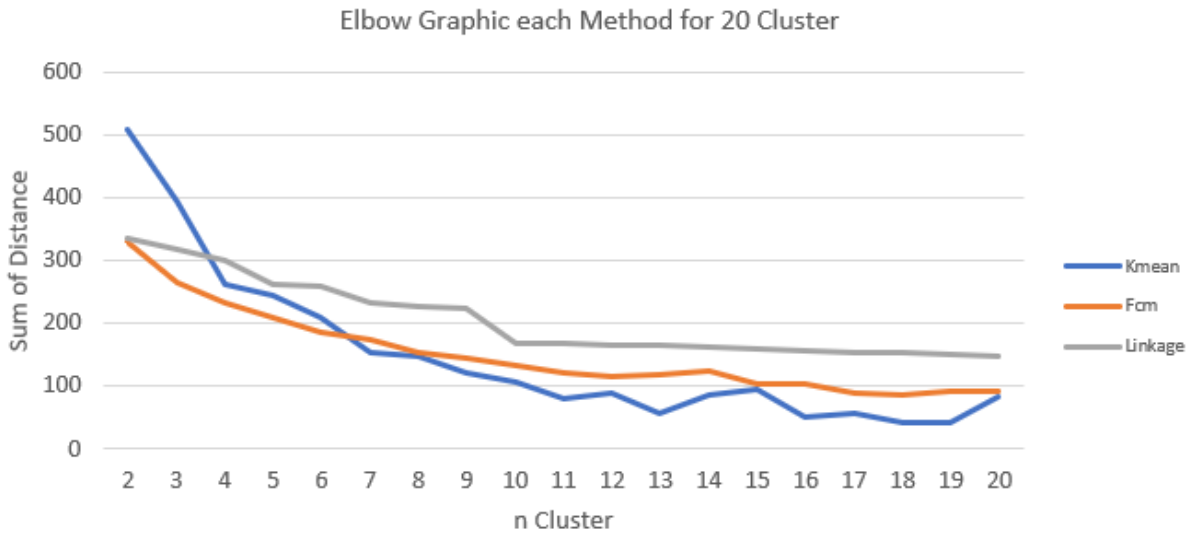


Fig. 6. Elbow Evaluation Clustering

After experimenting with two to twenty n-clusters, the results obtained from the elbow diagram are obtained from Figure 6. The pattern obtained from the graph is almost the same. The higher number cluster gets better internal distance. Detail number result from elbow graphic that presents in Table II. The highest and lowest values were obtained from the K-Means method, namely with a value of 508.7 in n clusters 2 and 40.0 in n clusters 19. Although it did not get the maximum results in the FCM method grouping, which got the most stable value, this can be seen from the standard difference between each. The *n* variable the smallest cluster. Comparable to the first experiment with a value of 8 clusters, the linkage method obtained less competitive results than other methods in this study.

TABLE II
SUM OF DISTANCE EACH CLUSTER

n Cluster	2	3	4	5	6	7	8	9	10
K-Mean	508.7	393.4	260.7	243.5	208.9	152.6	147.9	119.7	105.8
Fcm	329.9	264.5	230.6	207.3	185.7	174.6	154.1	143.0	132.1
Linkage	335.0	316.2	300.5	261.6	258.4	231.7	224.8	223.5	167.8

n Cluster	11	12	13	14	15	16	17	18	19	20
Kmean	79.8	87.8	56.0	83.8	94.6	49.0	55.6	40.0	40.0	81.6
Fcm	119.2	113.1	118.1	124.7	102.4	103.7	86.7	85.5	90.2	90.7
Linkage	166.5	164.3	163.0	161.6	157.8	155.2	154.1	153.1	148.3	146.9

DOI: 10.25139/ijair.v2i2.3030

To get the best n-cluster from the elbow graphic, what needs to be done is to calculate the difference between the n-cluster value and the previous cluster. The optimal n-cluster is the n-cluster with the largest distance value. As in Figure 7, the three methods used have different results: 4 clusters are the best results from the K-Means method, 3 clusters for the FCM method, and Linkage gets the best 10 clusters according to the internal elbow analysis. The highest gap of the K-Means method is located between 2 clusters and 4 clusters, and the value approaches 140. Meanwhile, FCM has the highest gap approaches 70 between 2 clusters and 3 clusters. These results showed that FCM and K-Means were obtaining the optimal n cluster at the beginning of the c cluster. Linkage obtained an optimal n cluster at a median of 10 clusters with a gap value is 55.

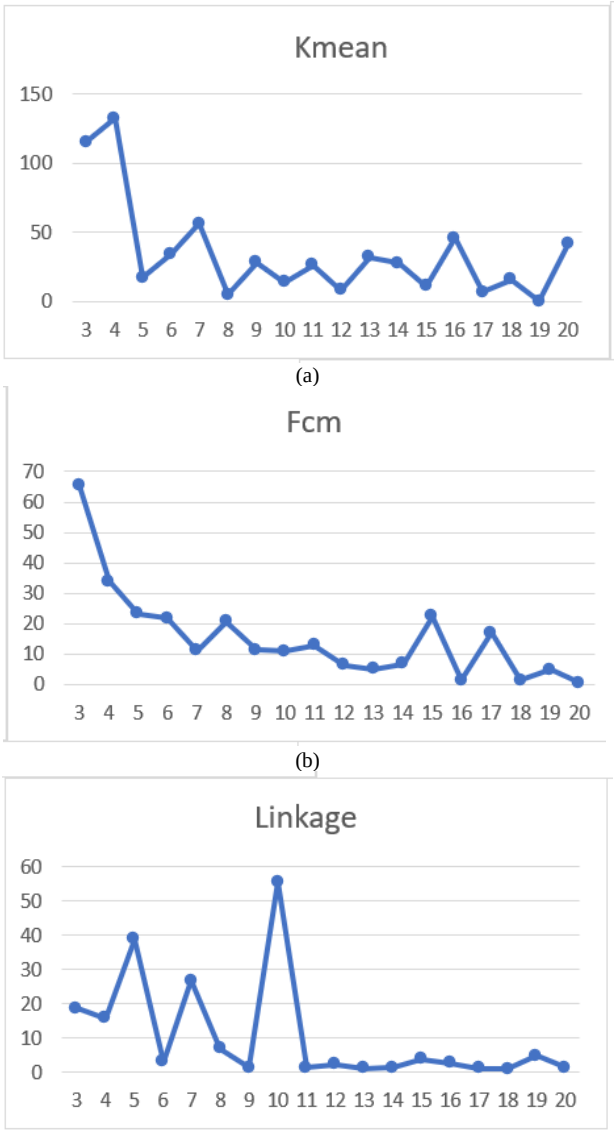


Fig. 7. Gap value distance n-cluster. (a) K-Means (b) FCM (c) Linkage

IV. CONCLUSION

After experiments from the active fire dataset in the Borneo island region with the partial clustering and Linkage hierarchical clustering method, the conclusion that can be obtained is the clustering method gets a smaller total distance value compared to hierarchical clustering. Each technique that has been tested turns out to have an optimal n-cluster amount that varies according to the elbow graph measurement. In general, the most competitive method of internal clustering evaluation is K-Means. From a computational point of view, K-Means is the method that requires the least computation. The limitation of K-Means lies in determining the initial centroid. If the initial centroid point is less precise, the results are also less than optimal. The FCM method gets the closest result to K-means because of the same approach. The disadvantages of the FCM method are the same as the K-means method, namely the initial centroid determinant, and it is more wasteful than K-Means due to the addition of the Fuzzy

DOI: 10.25139/ijair.v2i2.3030

membership function process. Meanwhile, the linkage method from the internal distance results is not good. This result is obtained because the dataset used has a high spread. And in terms of computation, this method is the most expensive because it continuously evaluates intra and extra clusters for each data. For the future research are to focus on the partial clustering method to complete the active fire dataset from NASA, more specifically the K-Means method. K-Means is a simple method but requires a lot of effort to optimize it.

REFERENCES

[1] R. R. Muhima, M. Kurniawan, and O. T. Pambudi, "A LOF K - Means Clustering on Hotspot Data," *Int. J. Artif. Intell. Robot.*, vol. 2, no. 1, pp. 29–33, 2020, doi: 10.25139/ijair.v2i1.2634.

[2] E. Çolak and F. Sunar, "The importance of ground-truth and crowdsourcing data for the statistical and spatial analyses of the NASA FIRMS active fires in the Mediterranean Turkish forests," *Remote Sens. Appl. Soc. Environ.*, vol. 19, no. April, p. 100327, 2020, doi: 10.1016/j.rsase.2020.100327.

[3] T. V Loboda, L. Giglio, L. Boschetti, and C. O. Justice, "Regional fire monitoring and characterization using global NASA MODIS fire products in dry lands of Central Asia," *Front. Earth Sci.*, vol. 6, no. 2, pp. 196–205, 2012, doi: 10.1007/s11707-012-0313-3.

[4] R. V. Virgil Petrescu, R. Aversa, T. M. Abu-Lebdeh, A. Apicella, and F. I. T. Petrescu, "NASA Satellites Help us to Quickly Detect Forest Fires," *Am. J. Eng. Appl. Sci.*, vol. 11, no. 1, pp. 288–296, 2018, doi: 10.3844/ajeassp.2018.288.296.

[5] P. Li, C. Xiao, Z. Feng, W. Li, and X. Zhang, "Occurrence frequencies and regional variations in Visible Infrared Imaging Radiometer Suite (VIIRS) global active fires," *Glob. Chang. Biol.*, vol. 26, no. 5, pp. 2970–2987, 2020, doi: 10.1111/gcb.15034.

[6] X. Wei, G. Wang, T. Chen, D. F. T. Hagan, and W. Ullah, "A spatio-temporal analysis of active fires over China during 2003-2016," *Remote Sens.*, vol. 12, no. 11, 2020, doi: 10.3390/rs12111787.

[7] A. A. Pereira *et al.*, "Burned area mapping in the Brazilian Savanna using a one-class support vector machine trained by active fires," *Remote Sens.*, vol. 9, no. 11, 2017, doi: 10.3390/rs9111161.

[8] O. C. D. Anejionu, G. A. Blackburn, and J. D. Whyatt, "Detecting gas flares and estimating flaring volumes at individual flow stations using MODIS data," *Remote Sens. Environ.*, vol. 158, pp. 81–94, 2015, doi: 10.1016/j.rse.2014.11.018.

[9] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, vol. 2001. 2001.

[10] W. M. J. Mohammed J. Zaki, *Data Mining and Analysis: Fundamental Concepts and Algorithms*, vol. 27. Cambridge University, 2013.

[11] A. Martino, A. Rizzi, and F. M. F. Mascioli, "Distance Matrix Pre-Caching and Distributed Computation of Internal Validation Indices in k-medoids Clustering," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, pp. 1–8, 2018, doi: 10.1109/IJCNN.2018.8489101.

[12] P. A. Murena, J. Sublime, B. Matei, and A. Cornuéjols, "An information theory based approach to multisource clustering," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, no. July, pp. 2581–2587, 2018, doi: 10.24963/ijcai.2018/358.

[13] V. S. Akondi, V. Menon, J. Baudry, and J. Whittle, "Novel K-Means Clustering-based Undersampling and Feature Selection for Drug Discovery Applications," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 2771–2778, doi: 10.1109/BIBM47256.2019.8983213.

[14] M. Misuraca, M. Spano, and S. Balbi, "BMS: An improved Dunn index for Document Clustering validation," *Commun. Stat. - Theory Methods*, vol. 48, no. 20, pp. 5036–5049, 2019, doi: 10.1080/03610926.2018.1504968.

[15] J. Lipor and L. Balzano, "Clustering quality metrics for subspace clustering," *Pattern Recognit.*, vol. 104, p. 107328, 2020, doi: 10.1016/j.patcog.2020.107328.