Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Implementation Of Clara Clustering Algorithm On Modis Data For Detection Of Forest Fire Potential In Indonesia

Holbed Joshua Petty[1], Sri Yulianto Joko Prasetyo[2]
[1,2]Informatics Engineering, Faculty of Information Technology, Satya Wacana Christian University, Indonesia

Article Info
Keywords:
CLARA,
Clustering,
Forest and land fires,
MODIS,
Silhouette Coefficient.

ABSTRACT
Forest fires are a recurring issue every year in various countries, especially in those with extensive forests like Indonesia. An initial step in fire prevention is detecting the potential occurrence of fires, which can be achieved by utilizing satellite data, such as MODIS data. In this study, clustering or grouping of MODIS data in Indonesia for the years 2021 and 2022 was conducted using the CLARA algorithm due to its robustness against outliers and efficiency in handling large datasets. The application of clustering with the CLARA algorithm on both datasets resulted in two clusters, and the evaluation using the Silhouette Coefficient yielded values of 0.89 and 0.88 for both years. The analysis revealed that both clusters in both datasets exhibited similar characteristics. In the data for the years 2021 and 2022, the first cluster displayed a moderate to high potential for fire, while the second cluster indicated a low potential for fire. The results of this study can be used as a reference for authorities to identify the level of forest/land fire potential from observed hotspots in Indonesia, thus enabling early prevention measures such as early extinguishment to prevent further spread of the fire.

Corresponding Author:
Holbed Joshua Petty
Informatics Engineering, Faculty of Information Technology, Satya Wacana Christian University, Indonesia
holbed17@gmail.com

INTRODUCTION

Forests are ecosystems where various types of living creatures live with a high level of biodiversity, and play a very important role in maintaining the harmony of natural resources. Forests also provide various materials needed to meet human needs (Mala et al., 2017). Fires generally occur in forests and on land, which can occur due to two main factors: natural factors and human intervention. Natural factors include plants drying out due to the long dry season, while human factors include open burning activities to expand land (Humam et al., 2020).

Hot spots are significant sources of smoke and serve as a source of pollution containing harmful elements that can pose health risks when inhaled in high concentrations (Rozi, 2020). Minister of Forestry Regulation No. P.12/Menhut-II/2009 explains that a hot spot or hotspot is a sign or sign of a forest fire by identifying locations that have a relatively higher temperature compared to the temperature in the surrounding area. Hot spots can be

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

monitored using Terra/Aqua satellite technology and using the MODIS sensor owned by NASA.

The hot spot parameters consist of several factors, namely (Kresna, 2018):

a. Brightness is the level of brightness of a hot pixel measured in Kelvin temperature.
b. Fire radiative power (FRP) describes the intensity of heat radiation in Megawatt (MW) units detected by satellites.
c. Hotspot Confidence Level (confidence) which ranges from 0 to 100 percent, indicates the level of confidence in the existence of hotspots.

Hot spots will be detected by satellites if the temperature at the pixel exceeds a certain limit. The temperature limit used to identify a hot spot is when the temperature exceeds 300°K (Pramesti, 2017).

Research on hotspot clustering using CLARA was previously carried out by Lidrawati et al. (2022) with the title "Implementation of the CLARA Clustering Method for Grouping Data on Potential Forest/Land Fires Based on the Distribution of Hot Spots," this research was conducted with the aim of finding out the ability of the Clara algorithm to identify potential fires by grouping hot spot data with 7532 rows of data, and Two clusters were produced where the first cluster was identified as a hot spot with high fire potential and the second cluster was identified as a hot spot with low fire potential.

As a result of the fire that occurred, various parties, including the government and society, have been urged to take early preventive action. One of the first steps that can be taken is to group data based on hot spots to evaluate the potential for fires in forest and land areas. This is useful for estimating hot spots that may experience fire (Lindrawati, 2022).

METHODS

Theoretical Foundations

MODIS

MODIS is a sensor tool found on the Terra satellite (EOS AM-1) and the Aqua satellite (EOS PM1) owned by NASA. This sensor plays a role in observing various conditions on the earth's surface, such as land and sea temperatures, vegetation index, land cover, forest/land fires, volcanic activity, cloud formation, aerosol concentration, and recording temperature profiles and water vapor levels (ESA, 2015). MODIS provides data regarding hotspot locations that are indications of fire occurrence. This identification process utilizes an algorithm that takes into account the context by utilizing infrared radiation originating from fires. MODIS uses information from band channels 21/22 and 31 to determine where the hot spots are located and their distribution (Pramesti et al., 2017). MODIS products are generated in square tile units with a Sinusoidal grid (SIN) projection, each tile covering approximately 1200 km at the equator (Dou et al, 2020).

Hotspot

Hot spots are significant sources of smoke and serve as a source of pollution containing harmful elements that can pose health risks when inhaled in high concentrations (Rozi et al., 2020). Minister of Forestry Regulation No. P.12/Menhut-II/2009 explains that a hot spot or

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

hotspot is a sign or sign of a forest fire by identifying locations that have a relatively higher temperature compared to the temperature in the surrounding area. Hot spots can be monitored using Terra/Aqua satellite technology and using the MODIS sensor owned by NASA.

The hot spot parameters consist of several factors, namely (Kresna Amijaya et al., 2018):

a. Brightness is the level of brightness of a hot pixel measured in Kelvin temperature.
b. Fire radiative power (FRP) describes the intensity of heat radiation in Megawatt (MW) units detected by satellites.
c. Hotspot Confidence Level (confidence) which ranges from 0 to 100 percent, indicates the level of confidence in the existence of hotspots.

Hot spots will be detected by satellites if the temperature at the pixel exceeds a certain limit. The temperature limit used to identify a hot spot is when the temperature exceeds 300°K (Pramesti et al., 2017).

As a result of the fire that occurred, various parties, including the government and society, have been urged to take early preventive action. One of the first steps that can be taken is to group data based on hot spots to evaluate the potential for fires in forest and land areas. This is useful for estimating hot spots that may experience fire (Lidrawati et al., 2022).

Clustering

Data mining is the rigorous process of extracting and analyzing extensive and complex data to discover valuable patterns, information, or knowledge that may be hidden. The main goal of data mining is to reveal insights that can be used in decision making, predictions, and supporting strategic planning (Zulfa & Hadiana, 2021). Cluster-based analysis is an approach that organizes data objects into groups or clusters based on the information available in the data, reflecting the properties and relationships between objects in the data (AS et al., 2019). The main goal is to make objects in one group have similarities or relationships with each other, while being different from objects in other groups. The higher the level of similarity within a group and the greater the differences between groups, the more optimal the grouping results will be. In the data clustering process, it is not necessary to have a class label associated with each data being processed. New class labels can be given after a group or cluster is formed. Therefore, in this context, clustering is often referred to as unsupervised learning (Hidayatur Rifa et al., 2019). Several clustering methods commonly used in data analysis are K-Means, Fuzzy C-Means, Improved K-Means, K-Medoids (PAM - Partitioning Around Medoids), DBSCAN, CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based upon RANdomized Search) and Fuzzy Subtractive Clustering (Pramesti et al., 2017).

Clara

K-Medoids clustering, which is often referred to as Partitioning Around Medoids (PAM), is a variation of the K-Means method. In contrast to K-Means which uses the average value to represent the core of the cluster. K-Medoids uses medoids as cluster centers. Medoids are

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

data points that are the center of a cluster and have values that are closest to other data in the cluster (Retsya Lapiza et al., 2023).

CLARA is a development of K-Medoids which is focused on adaptability to large dataset sizes. After seeing the complexity of the Partition Around Medoids (PAM) algorithm working at high speed, Kauffman and Rousseeuw (1990) developed a sample-based version to handle large datasets. They measure clustering accuracy based on the average difference between all items in the original dataset, not just the selected samples (Pakhira, 2008). The CLARA algorithm is a sampling-based algorithm, where samples are taken randomly. CLARA takes samples that are representative of the dataset as a whole. Medoids were selected from these samples by a method similar to K-medoids. If the sampling process is carried out correctly, the medoids selected from the sample will be similar to those that would be selected from the dataset as a whole. Therefore, the performance of CLARA is highly dependent on the sample size used (Chinchmalatpure & Dhore, 2020). Different from several other medoid methods, CLARA is resistant to outlier data and can be used on large datasets. This method shows better efficiency in terms of computing time and big data management compared to other approaches. (Susilowati & Sihombing, 2020). The CLARA algorithm is as follows (Nugraha et al., 2020):

"CLARA Algorithm
Input: A collection of d-dimensional patterns in set S, the desired number of clusters K, and the number of samples T.
Output: K medoids.
begin
for i = 1 to T, do the following
begin
Randomly select a sample of size 40 + 2K from the entire dataset;
Apply the PAM algorithm to this sampled subset to determine the medoids;
Assign each object in the original dataset to its closest medoid;
Compute the average distance between data items and their respective medoids.
Store the best result achieved thus far;

end"
In the CLARA algorithm to calculate the distance between objects in the dataset, the Manhattan distance equation is used, which is also known as L1 distance or City Block distance. The Manhattan distance is used to measure the distance between two objects by calculating the total absolute difference between the values of the objects. The Manhattan distance equation can be written as follows [18]:

$$d_{man}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)| \qquad (1)$$

Where:
d_man (x,y) : distance of object x to object y

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

x_i : object x attribute i
y_i : object y attribute i

Silhouette Coefficient

The Silhouette Coefficient method is a grouping validation method that combines the concept of coherence to assess the level of closeness between objects in a cluster and separation to measure the extent to which a cluster is scattered from other clusters (Paembonan & Abduh, 2021). The Silhouette Coefficient value is as follows in table 1 (Athifaturrofifah et al., 2019).

Table 1. Silhouette Coefficient

Scale	Explaination
0,7 < scale ≤ 1	Strong
0,5 < scale ≤ 0.7	Good
0,25 < scale ≤ 0,5	Bad
scale ≤ 0,25	Poor

Research Stages

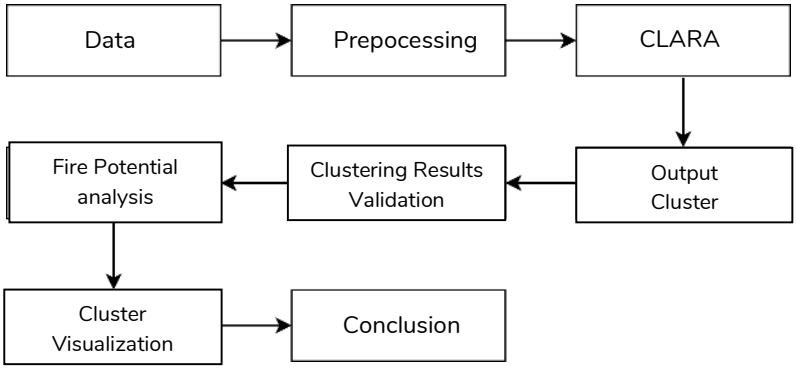

Figure 1. Research stages

As can be seen in Figure 1, this research began by obtaining MODIS data in Indonesia for 2021 and 2022. Next, preprocessing was carried out on the data, namely data selection to select the attributes that would be used in the research. The next stage is the clustering process using CLARA. After obtaining the results from the clustering process, the cluster results are validated and fire potential analysis is carried out, then visualization of the cluster results is carried out and the final stage draws conclusions. In conducting cluster analysis, the R programming language version 4.3.1 and Rstudio version 2023.06.1+524 were used.

Data Collection

The MODIS data used in this research is secondary data containing hotspot data in Indonesia in 2021 and 2022, totaling 14,210 and 11,279 rows respectively which were downloaded from the NASA website on January 1 2023 via the link https://firms.modaps .eosdis.nasa.gov/.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

The MODIS data used in this research is secondary data containing hotspot data in Indonesia in 2021 and 2022, which was downloaded from the NASA website on January 1 2023 via the link https://firms.modaps.eosdis.nasa.gov/. The dataset obtained is in CSV format. This dataset contains information on hot spots in Indonesia from January 1 2021 to December 31 2021 and January 1 2022 to December 31 2022, totaling 14,210 and 11,279 rows respectively and has 15 attributes.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 2. MODIS Data for 2021

latitude	Longi-tude	Bright-ness	scan	track	acq_ date	acq_ time	Satel-lite	Instru-ment	Confi-dence	Ver-sion	bright_ t31	frp	Day night	type
-3.8988	122.4209	312.2	1.8	1.3	1/1/2021	239	Terra	MODIS	39	6.03	285.9	16	D	2
-5.2174	119.4392	317.1	1.2	1.1	1/1/2021	239	Terra	MODIS	33	6.03	294.3	7.7	D	0
-8.9146	116.7551	313.1	1	1	1/1/2021	240	Terra	MODIS	0	6.03	293.8	7	D	0
-1.3673	121.2896	328.5	1	1	1/1/2021	534	Aqua	MODIS	85	6.03	296.6	21.7	D	0
-1.366	121.2988	334.5	1	1	1/1/2021	534	Aqua	MODIS	88	6.03	296.6	29.9	D	0
-3.9928	122.0997	317.4	1	1	1/1/2021	534	Aqua	MODIS	66	6.03	294.9	8.7	D	0
2.1875	117.4451	308.7	1.1	1.1	1/1/2021	535	Aqua	MODIS	16	6.03	288	3.6	D	0
-4.5623	136.8921	313.7	2.1	1.4	1/2/2021	143	Terra	MODIS	34	6.03	291.9	14.8	D	2
-1.4514	127.4426	323.1	1	1	1/2/2021	1701	Aqua	MODIS	93	6.03	277.3	0	N	2
-3.9076	122.4142	302.5	1.1	1.1	1/3/2021	227	Terra	MODIS	35	6.03	272.5	3.5	D	2
-10.192	120.6432	325.3	1.1	1	1/3/2021	228	Terra	MODIS	66	6.03	296.8	15.6	D	0
-8.7457	116.8524	315	1.1	1.1	1/3/2021	228	Terra	MODIS	59	6.03	295.6	7.1	D	2
-2.6124	140.6634	311.8	2.9	1.6	1/3/2021	343	Aqua	MODIS	0	6.03	292.9	27.5	D	0
-10.189	120.6416	328.4	1.6	1.3	1/3/2021	520	Aqua	MODIS	81	6.03	291.9	34.7	D	0
-4.5623	121.7299	310.3	1.2	1.1	1/3/2021	521	Aqua	MODIS	23	6.03	287.2	7.2	D	0
-4.0312	121.7244	311.3	1.1	1.1	1/3/2021	521	Aqua	MODIS	47	6.03	290.6	8.2	D	0
-1.7928	121.6766	318.1	1.1	1	1/3/2021	522	Aqua	MODIS	69	6.03	284.1	14	D	0
-1.206	120.5553	309	1.2	1.1	1/3/2021	522	Aqua	MODIS	52	6.03	289	5.7	D	0
1.9868	117.3604	315.3	1.8	1.3	1/3/2021	523	Aqua	MODIS	33	6.03	294.2	17.7	D	0
-3.7283	115.2961	311.7	2	1.4	1/4/2021	310	Terra	MODIS	50	6.03	291.7	15.8	D	0

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 3. Data Attribute Description

No.	Attribute	Information
1.	*latitude*	Fire pixel centers per 1 km are based on Earth's latitude, but are not always the actual location of a fire because one or more fires can be detected within a single 1 km pixel.
2.	*longitude*	Fire pixel centers per 1 km are based on Earth's longitude, but are not always the actual location of a fire because one or more fires can be detected within a single 1 km pixel.
3.	*brightness*	The brightness temperature of channel 21/22 of the fire pixel is measured in degrees Kelvin.
4.	*scan*	The scan value represents the spatial resolution in the East-West direction of the scan.
5.	*track*	The track value represents the spatial resolution from North to South of the scan.
6.	*acq_date*	(Acquisition Date) Date of data acquisition by MODIS.
7.	*acq_time*	(Acquisition Time) Time of data acquisition/MODIS satellite pass (in UTC).
8.	*satellite*	Satellites that take data.
9.	*instrument*	Instruments used to collect data.
10.	*confidence*	The trust value of a hot spot. This value is based on a set of intermediate algorithm values used in the detection process. Confidence estimates range from 0 to 100% and are assigned one of three fire classes (low confidence fires, nominal confidence fires, or high confidence fires).
11.	*version*	The version identifies the data processing collection type (for example, MODIS Collection 6.3 Processing Standard).
12.	*bright_t31*	The brightness temperature of channel 31 of the fire pixel is measured in degrees Kelvin.
13.	*frp*	(Fire Radiative Power) Describes the pixel's integrated fire radiation power in MW (megawatts).
14.	*daynight*	Data collection time during the day or night.
15.	*type*	Hot spot type. 0 = vegetation fires, 1 = active volcano, 2 = other static land sources, 3 = offshore.

Data Cleaning

Data cleaning is carried out to remove noise in the form of punctuation marks in the data. In this research, the data cleaning process was carried out manually using the data formatting feature in Microsoft Excel.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Data Selection

At this stage, the attributes selected will be used in the research. After going through the cleaning process, the data has 15 attributes. Of the 15 attributes, there are several attributes that need to be removed as follows:

a. There are two attributes that have identical purposes, namely the brightness and bright_t31 attributes. Because these two attributes have the same purpose, namely to show the level of brightness but use different bands, only one of them will be used, namely the brightness attribute.
b. There are other attributes that are not used in this research because they do not have relevant values for the clustering process, namely the scan, track, acq_date, acq_time, satellite, instrument, version, daynight and type attributes.

Of the 15 attributes after data selection, five attributes remained. The five attributes used are latitude, longitude, brightness, confidence and frp.

RESULTS AND DISCUSSION

Data

The following is data that has been processed through the cleaning and data selection stages so that data is produced as in the table. There are five data attributes used to carry out the clustering process.

Table 4. Data From Preprocessing Results

latitude	longitude	brightness	confidence	frp
-38988	1224209	312	39	16
-52174	1194392	317	33	8
-89146	1167551	313	10	7
-13673	1212896	329	85	22
-1366	1212988	335	88	30
-39928	1220997	317	66	9
21875	1174451	309	16	4
-45623	1368921	314	34	15
-14514	1274426	323	93	1
-39176	1224142	303	35	4

Testing the Number of Clusters

In the clustering process, the first step is a cluster test to determine the optimal number of clusters from the data. In this research, various numbers of clusters (k) will be tested, ranging from 2 to 10. To find optimal clusters, the Silhouette method is used. The following are the results of testing the number of clusters using the Silhouette method.

Table 5. Silhouette Coefficient Values for 2021 MODIS data

K value	*Silhouette Coefficient value*
2	0.89
3	0.57
4	0.51

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

K value	*Silhouette Coefficient value*
5	0.49
6	0.55
7	0.58
8	0.59
9	0.60
10	0.58

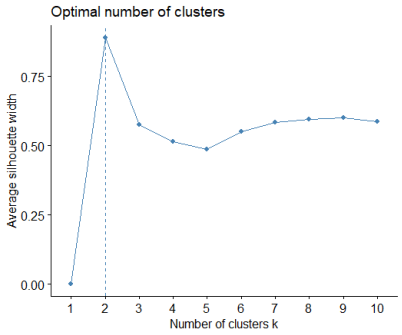

Figure 2. Results of The Test For The Number Of Clusters In 2021 MODIS Data

Table 6. Silhouette Coefficient Values for MODIS data for 2022

K value	*Silhouette Coefficient Value*
2	0.88
3	0.56
4	0.51
5	0.54
6	0.55
7	0.59
8	0.59
9	0.56
10	0.56

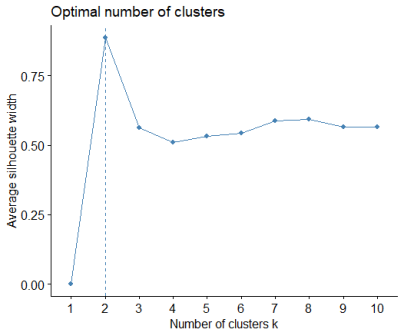

Figure 3. Test Results For The Number of Clusters in 2022 MODIS Data

The graphs in Figures 2 and 3 show that the optimal cluster for 2021 and 2022 data is obtained at the value k=2, where the resulting Silhouette Coefficient value can be seen in the

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

table at 0.89 and 0.88. It can also be seen that when k > 2, the value of the Silhouette Coefficient obtained has a lower value. This is caused by an increase in the number of neighbors involved in the clustering process. In this context, calculating neighbor distances becomes a critical factor that is taken into account in the calculation process. Thus, the distance between objects in one cluster increases, while the distance between one cluster and another cluster decreases.

Clustering

After testing the number of clusters that produced the highest Silhouette Coefficient value at k=2, the clustering process was carried out with k=2 using CLARA.

Table 7. MODIS Data Clustering in 2021

	Call: clara (x = Data, k = 2, pamLike = T)				
	Medoids:				
	Latitude	**Longitude**	**Brightness**	**Confidence**	**frp**
[1,]	-25386	1111405	327	83	19
[2,]	-20958	121081	312	66	11
Objective function:			85933		
Clustering vector:			int [1:14210] 1 1 1 1 1 1 1		
Cluster sizes:			12788 1422		

Table 8. MODIS Data Clustering in 2022

	Call: clara (x = Data, k = 2, pamLike = T)				
	Medoids:				
	Latitude	**Longitude**	**Brightness**	**Confidence**	**frp**
[1,]	-11871	1124201	313	60	8
[2,]	-21537	111832	311	43	12
Objective function:			87706.08		
Clustering vector:			int [1:11279] 1 1 1 1 1 1 1		
Cluster sizes:			10173 1106		

Table and 7 and 8 show the results of clustering using CLARA on MODIS data for 2021 and 2022 where k=2 produces the following data.

Table 8. 2021 MODIS Data Clustering Results

Cluster	Amount Member	Medoids brightness	Medoids confidence	Medoids frp
1	12788	327	83	19
2	1422	312	66	11

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 9. MODIS Data Clustering Results in 2022

Cluster	Amount member	Medoids brightness	Medoids confidence	Medoids frp
1	10173	313	60	8
2	1106	311	43	12

Based on tables 8 and 9, it can be seen that in the 2021 data, after clustering with k=2 using CLARA, it was concluded that the first cluster had more members and produced medoids with higher values. Meanwhile, in the 2022 data, almost the same conclusions were obtained with differences in FRP medoids values where the second cluster had a slightly higher value than the first cluster. The following is a data table that has been clustered using CLARA.

Table 10. MODIS Data Clusters in 2021

latitude	longitude	brightness	confidence	frp	cluster
15415	1248981	314	56	6	1
-14593	127441	312	10	11	2
-25808	121374	318	10	22	2
17221	98805	312	40	6	2
9765	1019345	306	10	6	1
-20635	1013958	308	58	9	1
4141	1018701	308	10	6	1
1488	1276326	311	35	9	1
-29119	1015229	318	59	12	1
-21844	1009431	320	78	13	1

Table 11. MODIS Data Clusters in 2022

latitude	longitude	brightness	confidence	frp	cluster
-24464	1331389	317	64	9	1
-69783	106831	313	27	5	2
-64485	1074511	309	38	4	1
-38681	1224053	303	49	4	1
-81091	112929	306	35	10	2
-81104	1129198	318	97	21	1
-81085	1129242	315	90	17	1
-81098	1129331	311	80	13	1
-8120	1129408	300	18	7	1
37837	1172595	302	33	28	1

Validation

Validation of cluster results was carried out using the Silhouette Coefficient method. The validation results are as follows.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table. 12 Silhouette Coefficient cluster data for 2021

cluster	size	ave.sil.width
1	12788	0.88
2	1422	0.94

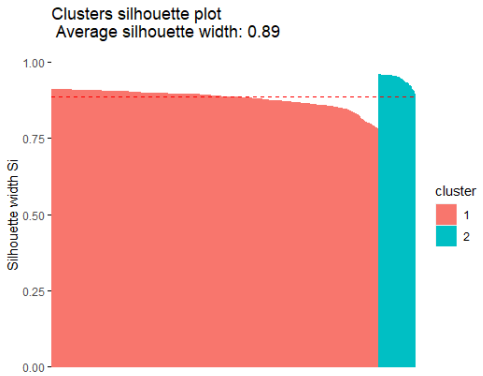

Figure 7. Average Silhouette Coefficient Data for 2021

Table 13. Silhouette Coefficient cluster data for 2022

cluster	size	ave.sil.width
1	10173	0.88
2	1106	0.94

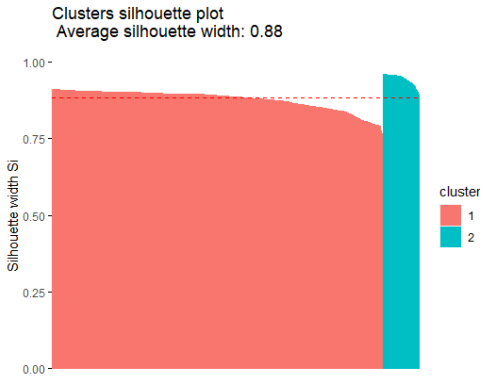

Figure 9. Average Silhouette Coefficient Data for 2022

Based on table 12 and figure 7, it shows that the Silhouette Coefficient value produced in clustering data for 2021 is 0.88 in the first cluster and 0.94 in the second cluster and produces an average silhouette coefficient value of 0.89. Meanwhile, table 13 and figure 9 show that the Silhouette Coefficient value produced in clustering data for 2022 is 0.88 in the first cluster and 0.94 in the second cluster and produces an average silhouette coefficient value of 0.88. Based on the validation test results, it can be concluded that the resulting cluster has a strong structure because both data produce an average silhouette coefficient value of > 0.7 so that the cluster results can be considered valid.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Fire Potential Analysis

Based on the results of the clustering process carried out on MODIS data for 2021 and 2022, it was found that members of the first cluster are hot spots that have medium to high potential, while the second cluster has low potential. This is supported by the medoids value in the first cluster which has a higher value than the second cluster. These results can be used as a reference in determining potential fire locations from hot spots observed directly through the hot spot satellite image information system in Indonesia.

Cluster Visualization

The following is a visualization of the clustering results using CLARA.

Figure 10. Plot of 2021 Data Clustering Results

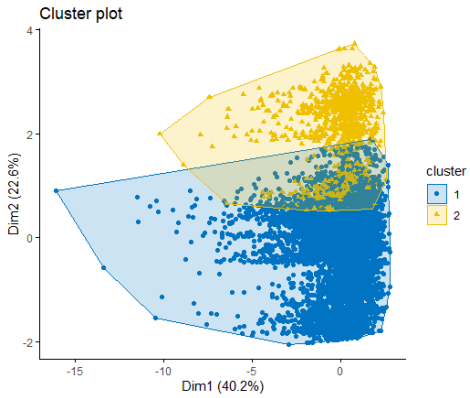

Figure 11. Plot of Data Clustering Results for 2022

Figures 10 and 11 show the plots produced by clustering using CLARA. There are several outlier data in each cluster which can also be clustered well using CLARA.

Discussion

From the results of the research that has been carried out, it can be concluded that clustering using CLARA can produce good clusters in MODIS Indonesia data for 2021 and 2022 and can be a reference for detecting potential fires in Indonesia. Previous research on clustering using CLARA was also carried out by (Lidrawati et al., 2022) with the title "Implementation

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 02, 2024, DOI 10.58471/infosains.v14i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

of the CLARA Clustering Method for Grouping Data on Potential Forest/Land Fires Based on the Distribution of Hotspots" where this research can produce two optimal clusters. Other related research was conducted by (Fitrayana & Saputro, 2022) entitled "Large Application Clustering Algorithm (CLARA) for Handling Outlier Data", (Zulyanti & Noeryanti, 2022) entitled "Comparison of Grouping of Micro, Small and Medium Enterprises in Klaten Regency in 2022 2019 Using the K-Means Method and Clustering Large Application", (Hidayatur Rifa et al., 2019) entitled "Implementation of the Clara Algorithm for Earthquake Data in Indonesia", and (Retsya Lapiza et al., 2023) entitled "Grouping The Districts in Sumatra Region Based on Economic Development Indicators Using K-Medoids and CLARA Methods" concluded that the CLARA algorithm can produce good accuracy in large-scale data clustering.

CONCLUSION

Based on the results of research that has been carried out, the clustering method using CLARA on MODIS data shows its ability to produce significant clusters, with a total of two clusters and a Silhouette Coefficient reaching 0.89 and 0.88. Analysis of the fire potential in the clusters shows that the first cluster has medium to high potential, while the second cluster shows a low level of potential. The results of this research can be used as a reference for authorities such as the Karlahuta task force and volunteers from the community to identify the level of potential forest/land fires from hot spots observed in Indonesia, so that direct monitoring can be carried out at hot spot locations to carry out early prevention of fires such as Early extinguishing to prevent wider spread of fire.

REFERENCE

AS, W., Aidid, M. K., & Nusrang, M. (2019). Pengelompokan Kabupaten/Kota Provinsi Sulawesi Selatan dan Barat Berdasarkan Angka Partisipasi Pendidikan SMA/SMK/MA Menggunakan K-Medoid dan CLARA. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *1*(3), 48. https://doi.org/10.35580/variansiunm12899

Athifaturrofifah, Goejantoro, R., & Yuniarti, D. (2019). Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Studi Kasus : Data Titik Panas Di Indonesia Pada 28 April 2018) Comparison of K-Means And K-Medoids Grouping Data on Potential Forest / Land Fires Based on Hotspots Distribution (Case study : Hotspots Data in Indonesia on April 28, 2018). *Jurnal EKSPONENSIAL*, *10*(2).

Chinchmalatpure, M., & Dhore, M. P. (2020). Quality Healthcare Prediction using K Means And Clara Partition Based Clustering Algorithm For Big Data Analytics. *International Journal of Engineering and Advanced Technology*, *9*(3), 1140–1144. https://doi.org/10.35940/ijeat.C4828.029320

Dou, Y., Huang, R., Mansaray, L. R., & Huang, J. (2020). Mapping high temperature damaged area of paddy rice along the Yangtze River using Moderate Resolution Imaging

Spectroradiometer data. *International Journal of Remote Sensing*, *41*(2), 471–486. https://doi.org/10.1080/01431161.2019.1643936

Fitrayana, P. R., & Saputro, D. R. S. (2022). Algoritme Clustering Large Application (CLARA) untuk Menangani Data Outlier. *PRISMA, Prosiding Seminar Nasional Matematika*, *5*, 721–725. https://journal.unnes.ac.id/sju/index.php/prisma/

Hidayatur Rifa, I., Pratiwi, H., & Respatiwulan. (2019). Implementasi Algoritma Clara Untuk Data Gempa Bumi Di Indonesia. *SEMINAR NASIONAL PENELITIAN PENDIDIKAN MATEMATIKA*, 161–166.

Humam, A., Hidayat, M., Nurrochman, A., Anestatia, A. I., Yuliantina, A., & Aji, S. P. (2020). Identifikasi Daerah Kerawanan Kebakaran Hutan dan Lahan Menggunakan Sistem Informasi Geografis dan Penginderaan Jauh di Kawasan Tanjung Jabung Barat Provinsi Jambi. *Jurnal Geosains Dan Remote Sensing*, *1*(1), 32–42. https://doi.org/10.23960/jgrs.2020.v1i1.14

Kresna Amijaya, B., Furqon, M. T., & Dewi, C. (2018). Clustering Titik Panas Bumi Menggunakan Algoritme Affinity Propagation. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3835–3842. http://j-ptiik.ub.ac.id

Lidrawati, E., Bahri, S., Zubaedi, U. F., Carolina, V. P., Kusrini, K., & Maulina, D. (2022). Implementasi Metode CLARA Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 507–511. https://doi.org/10.47065/josyc.v3i4.2006

Mala, V., Kusuma, A., Tanzil Furqon, M., & Muflikhah, L. (2017). Implementasi Metode Fuzzy Subtractive Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(9), 876–884. http://j-ptiik.ub.ac.id

Nugraha, W., Maulana, M. S., & Sasongko, A. (2020). Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012014

Paembonan, S., & Abduh, H. (2021). Penerapan Metode Silhouette Coeficient Untuk Evaluasi Clutering Obat. *Jurnal Ilmiah Ilmu - Ilmu Teknik*, *6*(2), 48–54. https://ojs.unanda.ac.id/index.php/jiit/index

Pakhira, M. K. (2008). Fast image segmentation using modified CLARA algorithm. *2008 International Conference on Information Technology*, 14–18.

Pramesti, D. F., Tanzil Furqon, M., & Dewi, C. (2017). Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(9), 723–732. http://j-ptiik.ub.ac.id

Retsya Lapiza, Syafriandi, Amalita, N., & Fitria, D. (2023). Grouping The Districts in Sumatera Region Based on Economic Development Indicators Using K-Medoids and CLARA Methods. *UNP Journal of Statistics and Data Science*, *1*(1), 16–22. https://doi.org/10.24036/ujsds/vol1-iss1/13

Rozi, F., Akbar, A. A., & Kadaria, U. (2020). Hubungan sebaran titik panas (hotspot) terhadap kesehatan masyarakat kota Pontianak. *Jurnal TEKNIK-SIPIL*, *20*(2).

Susilowati, B. E., & Sihombing, R. (2020). Metode ROBPCA (Robust Principal Component Analysis) dan Clara (Clustering Large Area) pada Data dengan Outlier (Studi Kasus Data Laporan Indeks Kebahagiaan Dunia Tahun 2018). *Jurnal Ilmu Komputer*, *8*(2), 88–98.

Zulfa, N. S. L., & Hadiana, A. (2021). Kajian Data Mining Menggunakan Algoritma K-Means Dan K-Medoids Dalam Strategi Promosi (Studi Kasus: Universitas Islam Al-Ihya Kuningan). *Jurnal Fakultas Teknik Kuningan*, *2*(2), 57–62.

Zulyanti, T., & Noeryanti. (2022). Perbandingan Pengelompokan Usaha Mikro Kecil Dan Menengah Di Kabupaten Klaten Tahun 2019 Dengan Metode K-Means Dan Clustering Large Application. *Jurnal Statistika Industri Dan Komputasi*, *7*(1), 46–59.