JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Prasetyadi, A., Nugroho, B., & Tohari, A. (2022). A Hybrid K-Means Hierarchical Algorithm for Natural Disaster Mitigation Clustering. *Journal of Information and Communication Technology, 21*(2), 175-200. https://doi.org/10.32890/jict2022.21.2.2

A Hybrid K-Means Hierarchical Algorithm for Natural Disaster Mitigation Clustering

*[1]Abdurrakhman Prasetyadi, [2]Budi Nugroho & [3]Adrin Tohari
[1&2]Research Center for Informatics,
National Research and Innovation Agency, Indonesia
[3]Research Center for Geotechnology,
National Research and Innovation Agency, Indonesia

*abdurrakhman.prasetyadi@brin.go.id
budi.nugroho@brin.go.id
adrin.tohari@brin.go.id
*Corresponding author

Received: 31/10/2021 Revised: 16/11/2021 Accepted: 18/11/2021 Published: 7/4/2022

ABSTRACT

Cluster methods such as k-means have been widely used to group areas with a relatively equal number of disasters to determine areas prone to natural disasters. Nevertheless, it is difficult to obtain a homogeneous clustering result of the k-means method because this method is sensitive to a random selection of the centers of the cluster. This paper presents the result of a study that aimed to apply a proposed hybrid approach of the combined k-means algorithm and hierarchy to the clustering process of anticipation level datasets of natural disaster mitigation in Indonesia. This study also added keyword and disaster-type fields to provide additional information for a better clustering process. The clustering process produced three clusters for the anticipation level of natural disaster mitigation. Based on the validation from experts, 67 districts/cities (82.7%) fell into Cluster 1 (low anticipation), nine districts/cities (11.1%) were classified into Cluster 2 (medium), and the remaining five districts/cities (6.2%) were categorized in Cluster 3 (high anticipation). From the analysis of the calculation of the silhouette coefficient, the hybrid algorithm provided relatively homogeneous clustering results. Furthermore, applying the hybrid algorithm to the keyword segment and the type of disaster produced a homogeneous clustering as indicated by the calculated purity coefficient and the total purity values. Therefore, the proposed hybrid algorithm can provide relatively homogeneous clustering results in natural disaster mitigation.

Keywords: Clustering, Hybrid, K-means, Mitigation, Natural disaster.

INTRODUCTION

Many countries around the world are prone to natural disasters, including Indonesia. High rainfall, active tectonic and volcanic activities, and natural disasters, including floods, volcanic eruptions, earthquakes, and tsunamis, are very common occurrences in Indonesia. Consequently, disaster mitigation efforts are indispensable to minimize the impact of a disaster in many regions in Indonesia.

A huge variety of research on natural disaster mitigation have been carried out. Natural disaster mitigation is a continuous effort to reduce the impact of disasters against people and property (Sadewo et al., 2018). Prihandoko and Bertalya (2016) studied several factors for natural disasters in Indonesia and found that geographical condition was the main cause for natural disaster occurrence instead of weather condition. Anjayani (2008) suggested that earthquake hypocenters strongly correlated with the locations of many active volcanoes. Moreover, Supriyadi et al. (2018) revealed that floods were the most common natural disaster in Indonesia. In 2007, the Indonesian government passed the Law of the Republic of Indonesia Number 24 of 2007 concerning disaster management as the national reference (Indonesia, 2007). Rachmawati (2018) conducted a study on the community's knowledge in the disaster areas to measure people's general awareness of areas at risk to lessen the consequences of natural disasters.

Numerous previous works have used data and information on natural disaster mitigation compiled by the National Agency for Disaster Countermeasure (BNPB) of Indonesia. Sadewo et al. (2018) conducted a clustering of disaster mitigation anticipation levels at the provincial government using the k-means method. Priatmodjo (2011) stated that disaster mitigation required preparedness, which included analysis of potential disasters and planning for anticipation. He also developed tools for disaster prevention and management. Atasever (2017) revealed a method to determine the level of damage due to a disaster. Han and Kamber (2001) used data mining to process large amounts of disaster data. Meanwhile, Prihandoko et al. (2017) used data mining techniques to analyze and predict disaster mitigation anticipation levels.

Various methods have been used to cluster the anticipation level of natural disaster mitigation. Ediyanto et al. (2013) described hierarchical clustering based on Euclidean distance to calculate the level of similarity. Hierarchical clustering is usually shown in the form of a tree diagram (dendrogram). Whereas for large amounts of data, the k-means method is more often used (Bagirov et al., 2011).

This paper presents the results of the clustering process of datasets of mitigation activities using data mining techniques to determine the anticipation levels for natural disasters. In this study, the k-means method and hierarchy are combined with a hybrid approach in producing a hierarchical k-means hybrid clustering. The clustering process is conducted using datasets originating from various research reports of natural disaster mitigation activities conducted by local and provincial governments available in the National Scientific Repository managed by the National Research and Innovation Agency (BRIN). The keyword fields and disaster types are also added as additional information for a better clustering process. Clustering is a powerful tool for data mining, which applies to virtually every field where large amounts of information are needed for data organization (Abdulsahib & Kamaruddin, 2015).

RELATED WORKS

Natural Disaster Mitigation

A natural disaster is a natural event that has a major impact on the human population. Located on the Pacific Ring of Fire (a region with many tectonic activities), Indonesia must continue and prepare to face the risks of volcanic explosions, earthquakes, floods, and tsunamis. Preparation for natural disasters includes all activities carried out prior to the detection of signs of disaster in order to facilitate the use of available natural resources, request assistance, and plan for rehabilitation in the best possible way and likelihood. Preparedness for natural disasters starts at the local communication level. If local resources are insufficient, the region can request assistance at national and international levels (Sadewo et al., 2018).

Clustering Disaster-Prone Areas and Mitigation

Cluster methods such as k-means have been widely used to group areas with relatively the same number of disaster characteristics to see which areas are prone to natural disasters (Yana et al., 2018). A study by Supriyadi et al. (2018) used k-means to classify disaster-prone areas into three clusters: high, medium, and low. In addition, Yana et al. (2018) found two regional clusters in Indonesia, namely prone to and not prone to natural disasters. Prihandoko and Bertalya (2016) suggested the cluster correlation between natural disasters, the number of victims, and weather conditions using k-means.

Sadewo et al. (2018) classified provinces according to their mitigation efforts using k-means in a disaster mitigation study. Their research results showed three clusters (high, medium, and low mitigation efforts). The regions of West Java, Central Java, and East Java entered a high level of mitigation. In another study, Kandel et al. (2014) discussed a comprehensive assessment of fuzzy techniques for mitigation. They utilized incremental fuzzy clustering to group mitigation data. Nevertheless, the authors did not experiment with other clustering techniques on the same dataset for accuracy measures. Several previous studies above, such as Sadewo et al. (2018), Supriyadi et al. (2018), and Prihandoko and Bertalya (2016), did not validate the results of clustering on the mitigation and disaster grouping by province. In addition, the results have not yet been compared with other clustering algorithms that are more effective and efficient in mitigation/disaster grouping.

Hybrid Clustering

Hybrid k-means and hierarchical clustering have been applied to studying disasters such as air pollution (Govender & Sivakumar, 2020). K-means and hierarchical clustering are two approaches that have different strengths and weaknesses. For instance, hierarchical clustering identifies groups in a tree-like structure yet suffers from computational complexity in large datasets. In contrast, k-means clustering is efficient but designed to identify homogeneous spherically shaped clusters (Peterson et al., 2018). Several studies have combined these two methods, such as Govender and Sivakumar (2020), who applied a combination of k-means and hierarchical clustering techniques to analyze air pollution. Atasever (2017) combined the k-means cluster method and backtracking search optimization algorithm (BSA) clustering to detect damage to natural disaster areas. The data results were grouped with a hybrid approach into two classes: damaged and undamaged areas.

Moreover, some studies compared hybrid k-means cluster methods with other methods. Nugroho (2021) compared the kernel k-means algorithm on bipartite graphs and k-means on the term-document matrix in the COVID-19 research dataset. The result was that the k-means kernel algorithm provided slightly better validation as compared to k-means. Balavand et al. (2018) used a hybrid of the crow search algorithm (CSA) k-means method with data envelopment analysis and compared it with other algorithms.

Nevertheless, other research works used a combination of clustering methods for disaster or other subjects. Wen et al. (2019) developed a combination of geographic information system (GIS) technology and the QUEST cluster algorithm, and the results showed the distribution of drought disaster areas. Ali et al. (2018) discussed disaster management with cluster techniques for emergencies, while Welton-Mitchell et al. (2018) examined clusters of people affected by the disaster. Ng and Khor (2016) evaluated the rapid profiling with clustering algorithms for plantation stocks on Bursa Malaysia. They utilized expectation maximization (EM), k-means, and hierarchical clustering algorithms to cluster the 38 plantation stocks listed on Bursa Malaysia. The results showed that a cluster resulting from EM had a better profile.

This study seeks to address some of the shortcomings of previous research. First, no one has explicitly used hybrid k-means and hierarchical clustering algorithms to suppress the level of disaster mitigation efforts. Second, previous research only surveyed the combination of k-means and hierarchical clustering studies. This study presents the application of the hybrid clustering approach that amalgamates the two methods to identify the general-shaped level of disaster mitigation clusters more efficiently. Specifically, the dataset is first partitioned into groups using the k-means algorithm. The next stage is to combine k-means and hierarchical as a hybrid approach. The hybrid approach is used because the k-means algorithm uses random observational data to determine the initial centroid. The centroid point is initialized randomly so that the resulting data grouping can be different. If the random value for initialization is not good, then the resulting grouping becomes less than optimal. A hybrid k-means and hierarchical algorithm is expected to avoid this problem.

METHODOLOGY

This study clustered the natural disaster (earthquake, tsunami, landslide, volcano eruption) dataset from technical reports on natural disaster research compiled by the National Scientific Repository. This dataset consisted of 237 documents of technical research reports conducted by researchers within and outside BRIN. A total of 81 districts and cities (subsequently named “region”) in Indonesia were included in this dataset. A mitigation category was created for each of the technical reports on natural disaster research. The categories consisted of A, B, C, D, E, and F, and are based on the types of natural disaster mitigation recommended by the National Research and Innovation Agency in each disaster-prone area. Table 1 presents the dataset summary.

Table 1

Summary of Dataset

No.	Region	Keyword	Type	Mitigation	Mitigation Code
1	Kota Padangsidimpuan	Monograph; Volcanic; Basic Data; Andesite;	Earthquake	Assessment of Disaster Risk and Characteristics	C
2	Kab. Simeulue	Active Tectonics; Earthquake; Continuous GPS; Subduction Zone; Megathrust; Sumatra	Earthquake; Tsunami	Preparation and Installation of Early Warning System Instruments	D

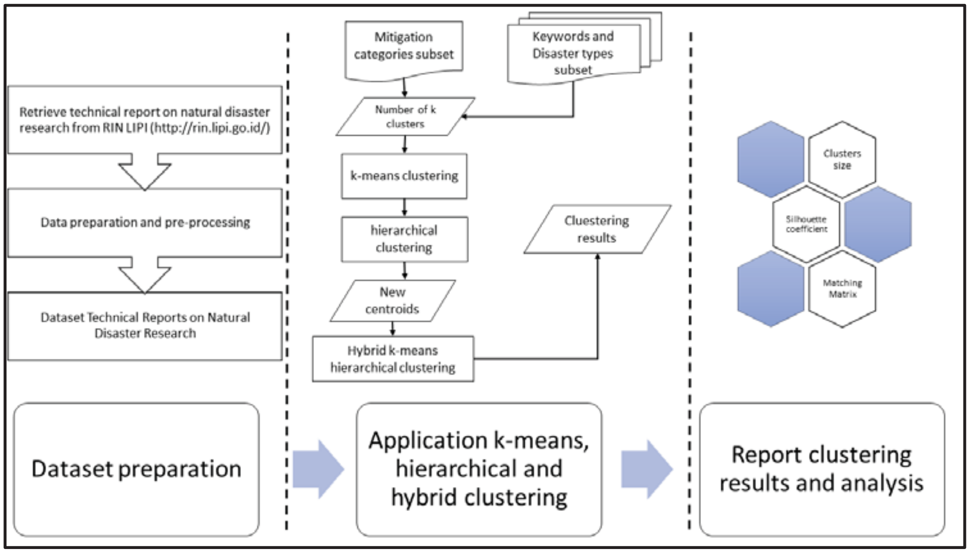
(continued)

No.	Region	Keyword	Type	Mitigation	Mitigation Code
3	Kota Padang	Active Tectonics; Earthquake; Continuous	Earthquake; Tsunami	Assessment of Disaster Risk and Characteristics	C
4	Kab. Sumedang	Rainfall; Slope Instability; Hydrological	Landslide	Construction and Strengthening of Building Structures	A
5	Kab. Kebumen	Weathering; Residual Soil; Physical Properties	Landslide	Assessment of Disaster Risk and Characteristics	C
6	Kab. Tanggamus	Disaster, Tanggamus; Vulnerability; Mitigation; Spatial	Earthquake; Tsunami	Planning and Implementation of Spatial Planning	E
7	Kota Serang	Earthquake; Geotechnical; Liquidation; Decline	Earthquake	Construction and Strengthening of Building Structures	A
8	Kota Bandung	Bandung Basin; Garut and Sumedang; Hazard Zoning; Earthquake; Active Fault	Earthquake	Planning and Implementation of Spatial Planning	E
9	Kab. Kepulauan Talaud	Talaud Regency; Border; Environment	Earthquake	Assessment of Disaster Risk and Characteristics	C
10	Kab. Kepulauan Mentawai	Mentawai Islands; Sumatran GPS Array (Sugar); Coral;	Earthquake	Preparation and Installation of Early Warning System Instruments	D

Three clusters of anticipation levels of disaster mitigation were constructed from each category: high, medium, and low. To ensure validity, the clusters were confirmed with experts in natural disasters from the Research Center for Geotechnology–Indonesian Institute of Sciences. This study used a hybrid approach that combined the k-means and the hierarchical algorithms to categorize the anticipation level. This method adopted Sadewo et al.’s (2018) clustering of anticipated levels of natural disaster mitigation at the provisional level and Atasever's (2017) hybrid approach to detect damage due to natural disasters. The current study also used the R programming language with the factoextra library in computing the application of a hybrid approach (Kassambara & Mundt, 2020). Figure 1 illustrates the exact method.

Figure 1

Detailed Hybrid Approach Flow Chart


Retrieve technical report on natural disaster research from RIN LIPI (http://rin.lipi.go.id/)
Data preparation and pre-processing
Dataset Technical Reports on Natural Disaster Research
Mitigation categories subset
Keywords and Disaster types subset
Number of k clusters
k-means clustering
hierarchical clustering
New centroids
Hybrid k-means hierarchical clustering
Cluestering results
Clusters size
Silhouette coefficient
Matching Matrix
Dataset preparation
Application k-means, hierarchical and hybrid clustering
Report clustering results and analysis

In the first stage, the k-means algorithm was applied with a numerical vector parameter of the mitigation category, with the number of clusters k = 3. The next step employed hierarchical clustering to the dataset of disaster mitigation categories with the parameter number of clusters k = 3and the ward method. Silhouette measures were utilized to determine the validity of this clustering result. The silhouette coefficient measured how well an observation was grouped and estimated the average distance between clusters (i.e., the average silhouette width). The negative silhouettes coefficient indicated that the observations might be grouped in the wrong cluster. To calculate the Silhouette Score, the following formula is used:

$$\text{Silhouette Score} = (b-a)/\max(a,b)$$

where

a= average intra-cluster distance, i.e., the average distance between each point within a cluster.

b= average inter-cluster distance, i.e., the average distance between all clusters.

The next stage was to combine the k-means and hierarchical clustering as a hybrid approach. The hybrid approach calculated the hierarchical clusters and cut the tree into several *k* clusters. It then calculated the centroid of each cluster. Finally, the hybrid approach calculated the k-means using the cluster centroid obtained from the previous calculation as the cluster's initial centroid. Next, hierarchical and k-means clustering results were compared, respectively, with hybrid clustering results using a matching matrix.

The hybrid clustering results from the disaster mitigation category were confirmed with experts to account for their validity. Furthermore, the three clusters were made as a ground truth reference for applying the hybrid k-means hierarchical algorithm to the natural disaster dataset on the subset of keywords and types of disasters.

This study used disaster-type and keyword subsets as the clustering base. The subsets represented the mitigation category relationship. This study carried out the clustering stage from each subset using the k-means algorithm by building a term frequency-inverse document frequency (TF-IDF) matrix to convert the document into a TF-IDF vector. Stop words were eliminated. In the computation, stop words were filtered out before or after processing natural language data (text), such as *the*, *is*, *at*, *which*, and *on*. The stemming process was not applied because the terms used in the sections were specific terms that reflected the contents' technical report documents. The hybrid algorithm of k-means and hierarchical was employed with the number of clusters $k = 3$ according to the anticipated levels of natural disaster mitigation that had been defined.

This study utilized unsupervised learning to divide the input data point with some common properties. In the previous stage, prior knowledge class labels were defined as the ground truth. In order to validate the clustering results, a matching matrix method was intuitively used. As described by Samatova et al. (2013), the matching matrix (Figure 2) is a $V \times W$ matrix, where V is the number of class labels in P and W is the total number of resulting clusters. Each row of the matrix represents one class label, and each column represents a cluster ID. Each m_{ij} entry represents the number of points from Class *i* that are present in cluster g_j. The table was filled based on the prior knowledge P and clusters obtained using U.

In this paper, purity was employed as the validation metric for the hybrid algorithm. Purity (Pu) is a measure to analyze the cluster's homogeneity concerning the class labels. Equation 1 calculates purity as follows:

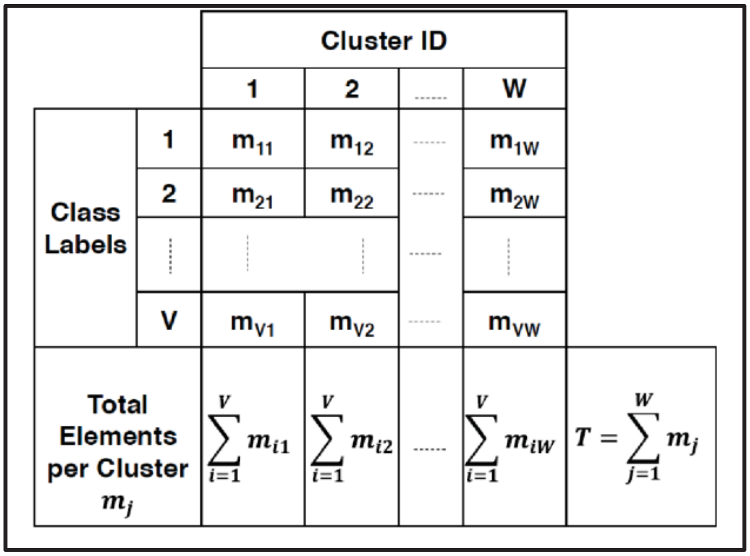
$$Pu_{gj} = \max_{i = 1\ to\ V} P_{ij} \tag{1}$$

This measure takes any value in the range of 1/V to 1. A value of 1 indicates an utterly homogeneous cluster. The total purity (TPu) was calculated for the entire cluster's results. TPu, as denoted by Equation 2 for the whole cluster set, was calculated as the sum of each cluster's purities weighted by the number of elements in each cluster.

$$TPu = \sum_{j=1}^{W} \frac{m_j}{M} P_{ugj} \tag{2}$$

Figure 2

Matching Matrix Template (Samatova et al., 2013)



		Cluster ID				
		1	2	......	W	
Class Labels	1	m_{11}	m_{12}	......	m_{1W}	
	2	m_{21}	m_{22}	......	m_{2W}	
	⋮	⋮	⋮	......	⋮	
	V	m_{V1}	m_{V2}	......	m_{VW}	
Total Elements per Cluster m_j		$\sum_{i=1}^{V} m_{i1}$	$\sum_{i=1}^{V} m_{i2}$	......	$\sum_{i=1}^{V} m_{iW}$	$T = \sum_{j=1}^{W} m_j$

RESULTS

K-means Clustering Overview

In the first stage, the k-means algorithm applied a numerical vector parameter of the mitigation category, with the number of clusters *k* = 3. The results of k-means clustering are presented in Figure 3. According to Figure 3, several regions were classified as Cluster 1, including Kab. Simeulue and Kab. Toba Samosir. Kota Banda Aceh, Kab. Kebumen, and Kab. Cilacap were included in Cluster 2. While Kab. Kep. Mentawai and Kab. Lampung were located in Cluster 3. Each cluster represented a different level of preparedness for natural disaster mitigation. To determine the validity of this clustering result, the silhouette size was used as shown in Figure 4. The silhouette size indicated that the majority of data points were well clustered, as indicated by positive silhouette values, particularly in Clusters 1 and 3. While in Cluster 2, several data points were found with negative values, indicating that they might belong to the incorrect cluster. The silhouette coefficient measured how well an observation was grouped and estimated the average distance between clusters (i.e., the average silhouette width). The negative silhouette coefficient indicated that the observations might be grouped in the wrong cluster. Table 2 presents some data with negative silhouette coefficients, including Kab. Bandung, Kab. Cilacap, and Kab. Tanggamus. These findings revealed that some data points in Cluster 2 were grouped in the wrong cluster.

Figure 3

Cluster Plot of K-means on Mitigation Category

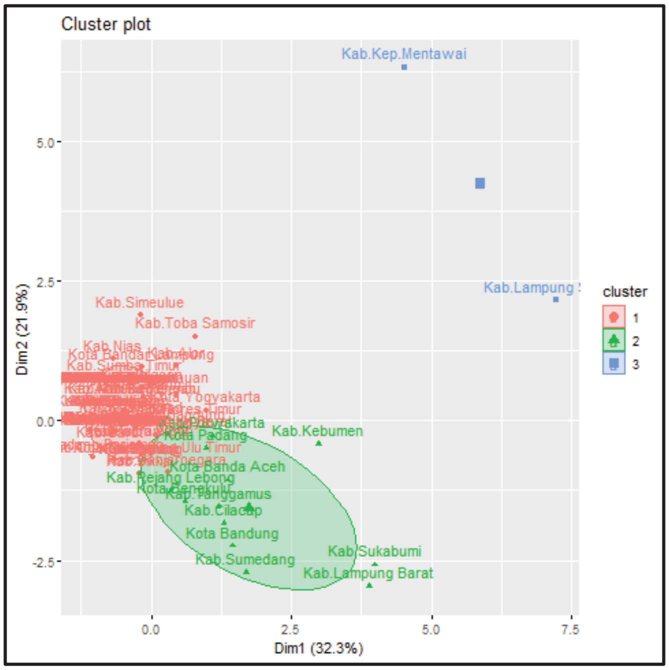


Figure 4

Cluster Silhouette Plot of K-means on Mitigation Category

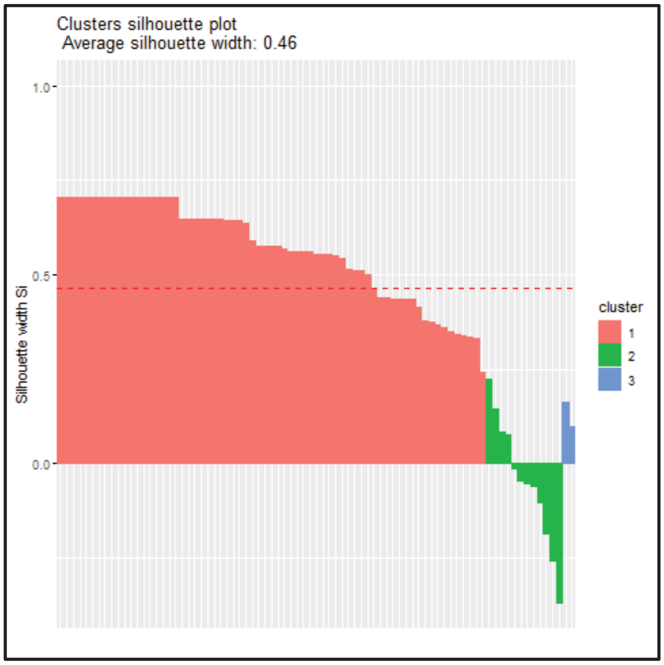


Table 2

Data Points with Negative Silhouette Coefficient in K-means Clustering

Region	Cluster	Neighbor	Sil_width
Kota Bandung	2	1	-0.01
Kab. Cilacap	2	1	-0.04
Kab. Tanggamus	2	1	-0.05
Kota Banda Aceh	2	1	-0.06
Kota Bengkulu	2	1	-0.10
Kab. Purwakarta	2	1	-0.18
Kab. Rejang Lebong	2	1	-0.25
Kota Padang	2	1	-0.36

Hierarchy Clustering Overview

In the next stage, the hierarchical algorithm was applied to the dataset of disaster mitigation categories with the parameter number of k =

3 and the ward method. The hierarchical results are presented in a dendrogram graph, as shown in Figure 5. Based on Table 3, there were only two data points with negative silhouette coefficients. Kab. Banda Aceh and Kab. Kebumen might be incorrectly clustered at this stage of hierarchical clustering, as indicated by a negative silhouette value. Figure 6 shows that the averaged silhouette width of hierarchical clustering was 0.53, which was higher than that of k-means (see Figure 4) for the same clustering category. The clustering results were similar to those of k-means clustering. Most data points were grouped into Cluster 1, followed by Cluster 2, and the remainder into Cluster 3. The difference between hierarchical and k-means clustering was fewer data points with negative silhouette values were found in the former. Therefore, hierarchical might produce a better result than the k-means algorithm in clustering the dataset.

Figure 5

Cluster Dendrogram of Hierarchical Clustering on Mitigation Category

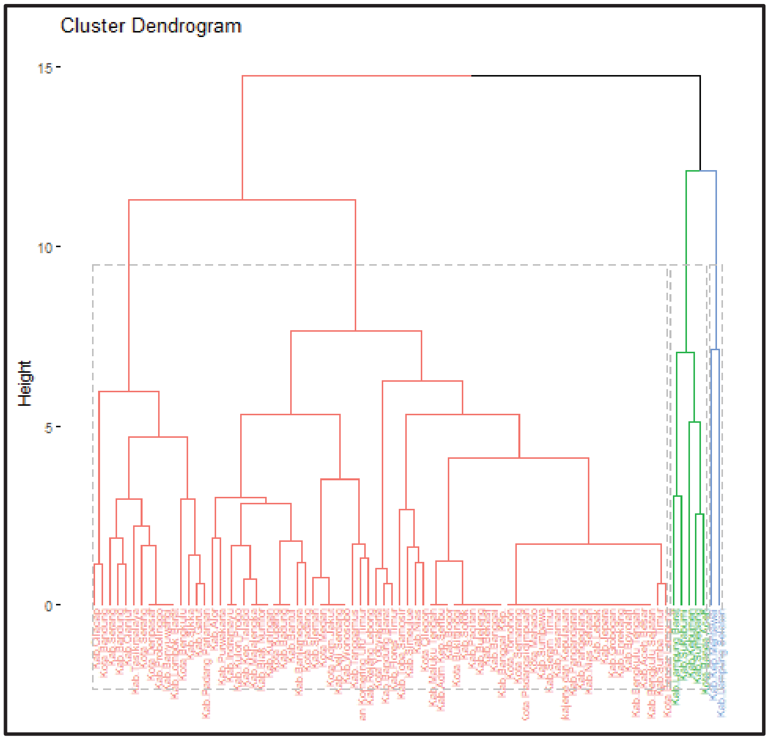


Figure 6

Cluster Silhouette of Hierarchical Clustering on Mitigation Category

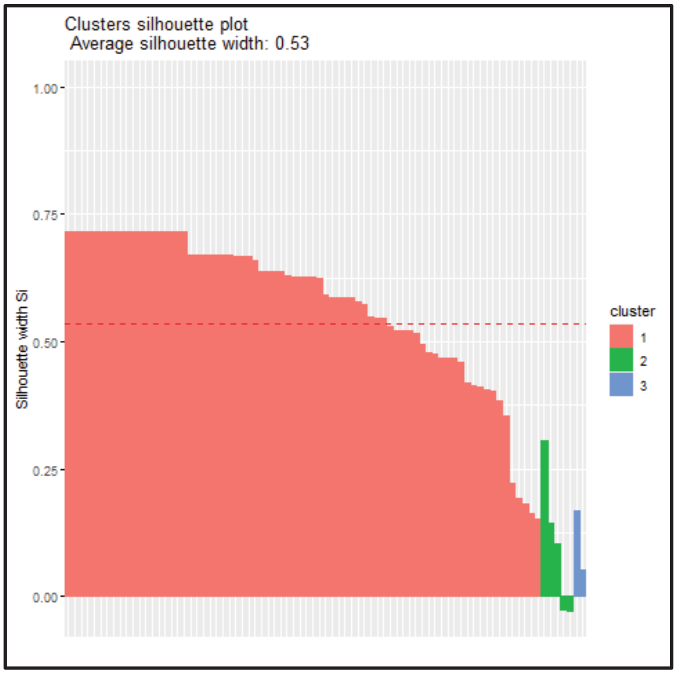


Table 3

Data Points with Negative Silhouette Coefficient in Hierarchical Clustering

Region	Cluster	Neighbor	Sil_width
Kota Banda Aceh	2	1	-0.02
Kab. Kebumen	2	1	-0.02

Hybrid K-means Hierarchical Clustering Overview

The next stage was to combine k-means and hierarchical as a hybrid approach. The hybrid approach was employed because the k-means algorithm uses random observational data to determine the initial centroid. The clustering solution of k-means is very sensitive to a random selection of the centers of the cluster. Therefore, clustering results may vary when recomputing.

The hybrid approach calculated the hierarchical clusters and cut the tree into several *k* clusters. It then calculated the center of each cluster and calculated the k-means using the cluster center obtained from the previous calculation as the cluster's initial center. The new centroids were defined as the mean of the variables in the cluster. Table 4 summarizes the result of the calculation of the new centroid. Next, k-means clustering was applied using the cluster's center above to obtain the cluster results, as presented in Table 4.

Table 4

New Centroids for Hybrid K-means Hierarchical Clustering

Cluster	A	B	C	D	E	F
1	−0,19	−0,14	−0,18	−0,12	−0,12	0
2	2,86	1,37	0,74	−0,02	1,62	0,16
3	0,13	1,86	5,11	4,77	0,66	−0,46

Hierarchical and Hybrid Clustering Results Comparison

Next, the hierarchical and hybrid clustering results were compared using the match matrix. As shown in Table 5, the hybrid algorithm produced better clustering results than the hierarchical algorithm. The data points were clustered homogeneously into each cluster.

Table 5

Matching Matrix of Hierarchical and Hybrid Clustering

		Hybrid Results		
		1	2	3
Hierarchical Results	1	67	0	0
	2	7	5	0
	3	0	0	2

From Figure 7, it can be observed that most data points were clustered homogeneously into a predetermined cluster. Nevertheless, some data did not appear to fit into the cluster. Cluster 2, marked in red, contained seven data points included in Cluster 1. The mis-clustered data points were confirmed by calculating the silhouette value, as illustrated in Figure 7 and Table 5.

Figure 7

Cluster Dendrogram of Hybrid K-means and Hierarchical Clustering on Mitigation Category

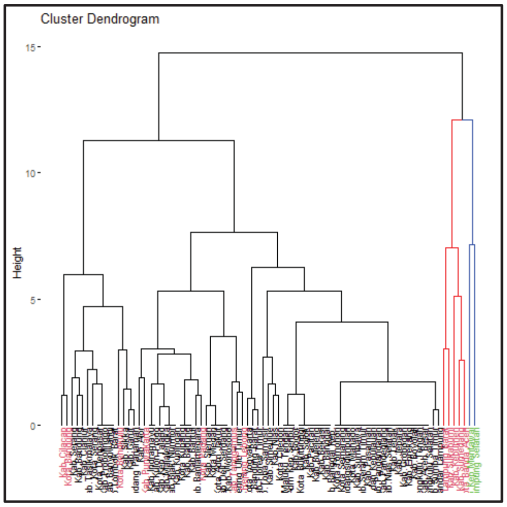


Figure 8 shows that most data points had positive silhouette values, which meant that the data points were clustered into the correct cluster. In contrast, it can also be seen that in Cluster 2, eight data points had negative values. The negative value indicated that there was a possibility that the data points were not clustered correctly in Cluster 2. This finding was confirmed by the silhouette values earlier. The final clustering solution, k-means, regrouped some data.

Figure 8

Cluster Silhouette of Hybrid Clustering on Mitigation Category

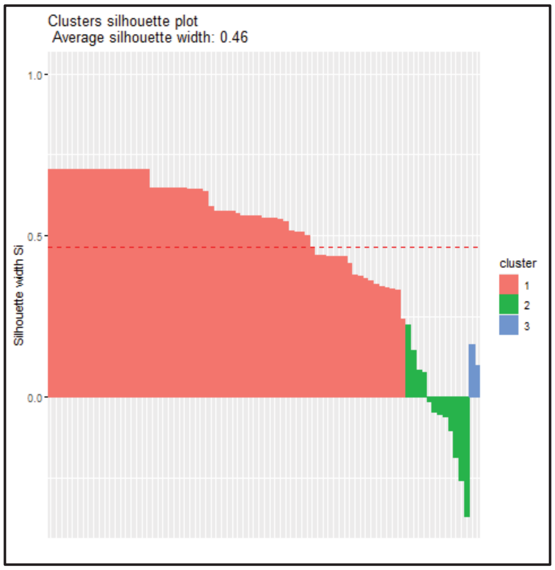


Table 6 presents the negative values of the silhouette. As many as eight data points were indicated in this table as being in the incorrect cluster, including Kota Bandung, Kab. Cilacap, and Kab. Tanggamus.

Table 6

Data Points with Negative Silhouette Coefficient in Hybrid Clustering

Region	Cluster	Neighbor	Sil_width
Kota Bandung	2	1	-0.01
Kab. Cilacap	2	1	-0.04
Kab. Tanggamus	2	1	-0.05
Kota Banda Aceh	2	1	-0.06
Kota Bengkulu	2	1	-0.10
Kab. Purwakarta	2	1	-0.18
Kab. Rejang Lebong	2	1	-0.25
Kota Padang	2	1	-0.36

K-means and Hybrid Clustering Results Comparison

In the same way, using the matching matrix, the clustering of standard k-means was compared with the hybrid approach, as shown in Table 7. Table 7 describes a matching matrix that consolidated the results of the standard and hybrid k-means clustering. Clusters 1, 2, and 3 each showed the clustered data points correctly by the two types of clustering algorithms applied.

Table 7

Matching Matrix of K-means and Hybrid Clustering

		Hybrid Result		
		1	2	3
k-means Result	1	67	0	0
	2	0	12	0
	3	0	0	2

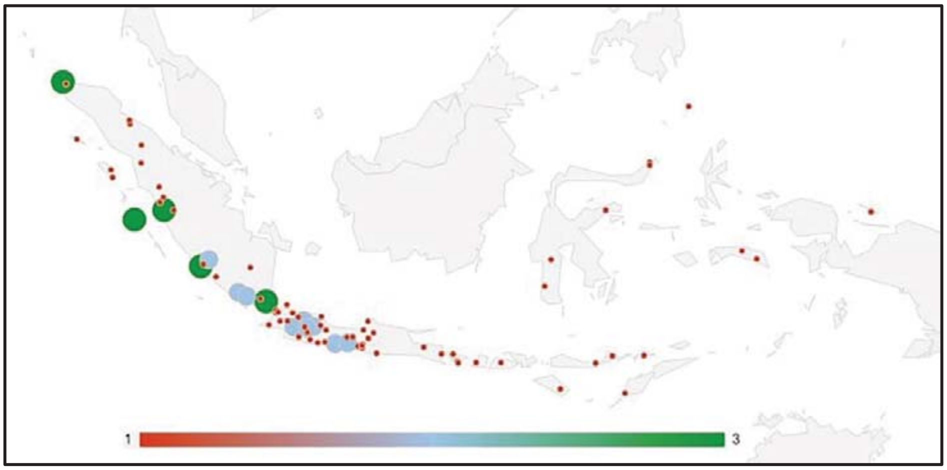
The results of this clustering were consulted with related experts. The clustering results showed that there were two regions, namely Kab.

Mentawai and Kab. South Lampung, in the high anticipation category. Based on the experts' justification, several other cities, i.e., Banda Aceh, Padang City, and Bengkulu City, could be highly anticipated. This difference was due to the lack of research in Category A that discussed construction and strengthening of building structures.

The hybrid clustering results from the disaster mitigation category were also confirmed with experts to account for their validity. The experts were from the Research Center for Geotechnology with expertise in the field of natural disasters. Furthermore, three clusters were defined as a ground truth reference for applying the hybrid k-means hierarchical algorithm. The hybrid algorithm was then used for keyword and disaster-type subsets. Figure 9 presents the ground truth, which consolidated the clustering results of applying the hybrid algorithm and validation from experts. Cluster 1 represented areas with low anticipation levels, Cluster 2 for medium anticipation levels, and Cluster 3 for high anticipation levels.

Figure 9

Consolidated Clustering Result as Ground Truth



Application of Hybrid Algorithm of K-means and Hierarchical Clustering

The hybrid algorithm of k-means and hierarchical was applied with the number of clusters k = 3 according to the anticipated levels of natural disaster mitigation that had been defined. The clustering results, shown in Figures 10 and 11, indicated that most data were grouped in Cluster 1. There were only two data in Cluster 2 and one data in Cluster 3.

Figure 10

Cluster Dendrogram of Hybrid K-means and Hierarchical Clustering on Keywords

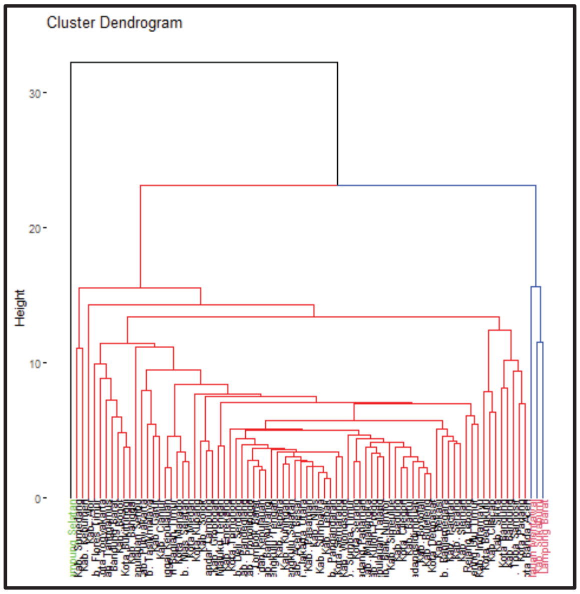


Figure 11

Result of Hybrid K-means and Hierarchical Clustering on Keywords

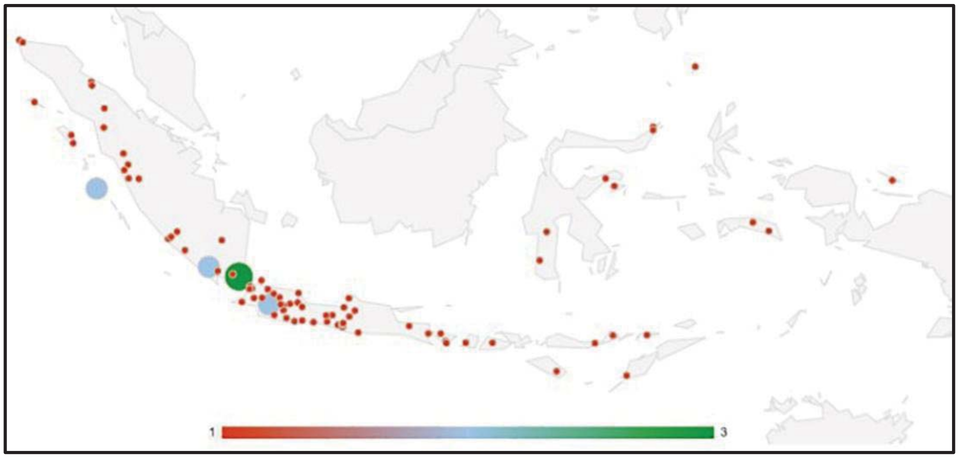
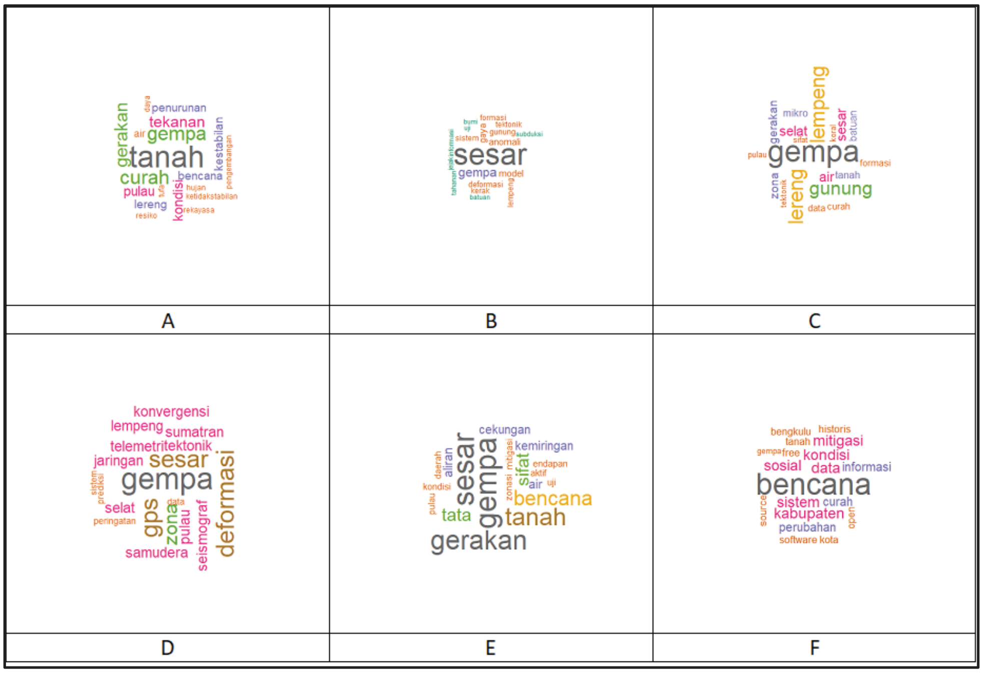


Figure 12 illustrates the keywords that represented each mitigation category. For Category A, which was construction and strengthening of building structures, the keywords that appeared most often included earth movements, pressure, and earthquakes. Category B was for mapping of disaster-prone areas, and the keyword that appeared the most was fault. Then, Category C was for assessment of disaster risks and characteristics, and the keywords that appeared were earthquakes and tectonic plates. While in Category D, which was for preparation and installation of early warning system instrumentation, keywords such as deformation and earthquake fault appeared. Category E was for planning and implementation of spatial planning, and the keyword that most often appeared was earthquakes. Finally, in Category F, which was for outreach and information dissemination, the keyword that often appeared was disaster.

Figure 12

Keyword Word Cloud on Mitigation Category



Figures 13 and 14 show the result of hybrid k-means and hierarchical on disaster types. In contrast with the previous clustering results on the keywords, the clustering on the disaster types resulted in more regions falling into Clusters 2 and 3.

Figure 13

Cluster Dendrogram of Hybrid K-means and Hierarchical Clustering on Disaster Types

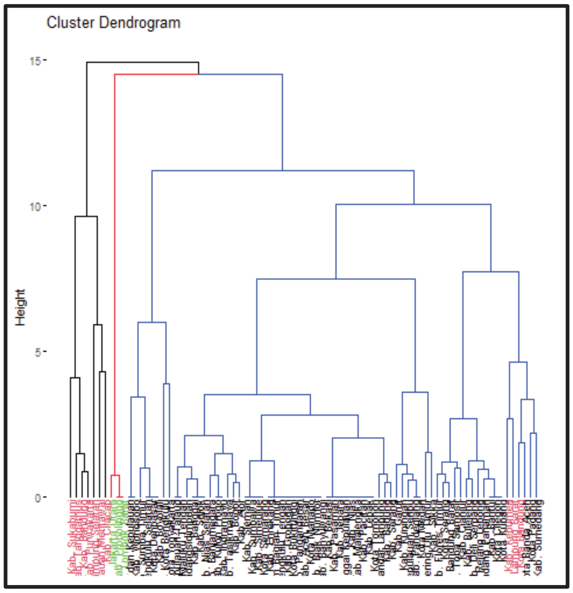


Figure 14

Result of Hybrid K-means and Hierarchical Clustering on Disaster Types

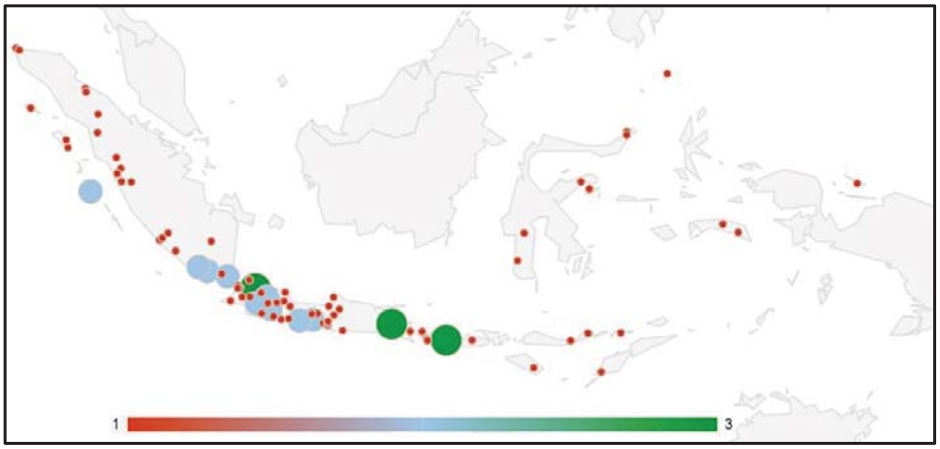
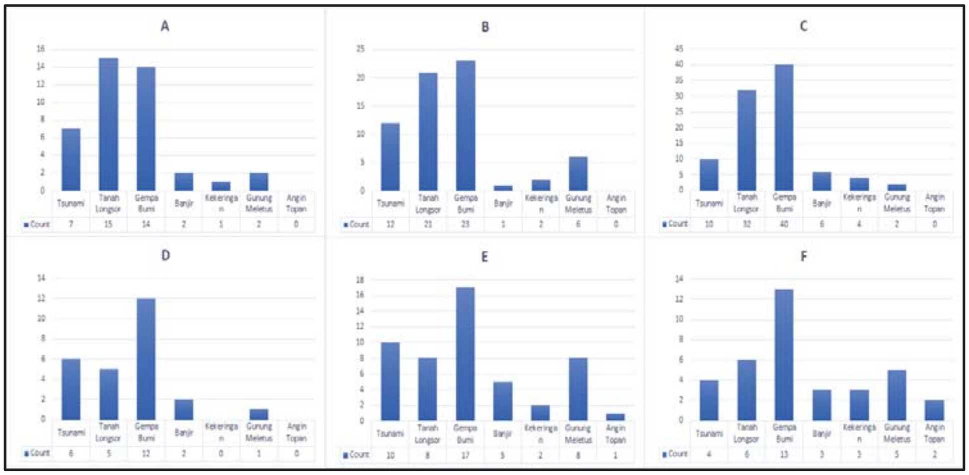


Figure 15 shows the correlation of each category of mitigation anticipation with different types of disasters. For example, in the mitigation category A, which was construction and strengthening of building structures, landslides were the most anticipated. While in Category D, which was for preparation and installation of early warning system instrumentation, earthquake and tsunami disasters were the most anticipated.

Figure 15

Disaster Types Bar Plot on Mitigation Category



To validate the clustering results, a matching matrix was used on keywords, disaster types, and mitigation code to determine purity. As shown in Table 8, the results indicated that the clusters had an average TPu value of 0.88 for the hybrid clustering algorithm, 0.84 for hierarchical, and 0.86 for k-means. The TPu values were close to 1, representing the acceptable results of the hybrid clustering algorithm. From this table, it can be concluded that the hybrid clustering outperformed k-means and hierarchical since its TPu value was the highest.

Table 8

Matching Matrix Validation on Keywords, Disaster Types, and Mitigation Code

	k-means	Hierarchical	Hybrid
TPu on Keywords	0.81	0.81	0.82
TPu on Disaster Type	0.76	0.79	0.82
TPu on Mitigation Code	1	0.91	1
Average TPu	0.85	0.83	0.88

CONCLUSION

This study examined on clustering the natural disaster literature dataset. The clustering process was performed by applying the k-means, hierarchical, and hybrid algorithms. This process produced three clusters for the anticipation levels of natural disaster mitigation: Cluster 1 for low anticipation level, Cluster 2 for medium anticipation level, and Cluster 3 for high anticipation level. In addition, from validation by experts, the clustering results indicated that 67 districts/cities (82.7%) fell into Cluster 1, nine districts/cities (11.1%) were classed into Cluster 2, and the remaining five districts/cities were categorized in Cluster 3 (6.2%). From the analysis of the silhouette coefficient calculation, the hybrid algorithm can provide relatively homogeneous clustering results.

Furthermore, a matching matrix was used on keywords, disaster types, and mitigation code to determine purity to validate the clustering results. The clusters had a TPu close to 1, representing acceptable results of the hybrid clustering algorithm. It was concluded that the hybrid clustering outperformed standard k-means and hierarchical since its TPu value was the highest. The clustering solution of k-means is very sensitive to a random selection of the centers of the cluster. Therefore, clustering results may vary when recomputing. This led the study to use hybrid clustering because the algorithm uses random observational data to determine the initial centroid.

A further study that aims to compare the hybrid clustering algorithm with other algorithms is recommended. The method for determining the disaster mitigation level also needs improvement.

ACKNOWLEDGMENT

The authors gratefully acknowledge the use of service and facilities of the Research Center for Informatics, National Research and Innovation Agency.

REFERENCES

Atasever, U. H. (2017). A new unsupervised change detection approach with hybrid clustering for detecting the areal damage after natural disaster. *Fresenius Environmental Bulletin, 26*(6), 3891–3896.

Bagirov, A. M., Ugon, J., & Webb, D. (2011). Fast modified global k-means algorithm for incremental cluster construction. *Pattern Recognition*, *44*(4), 866–876. https://doi.org/10.1016/j.patcog.2010.10.018

Balavand, A., Kashan, A. H., & Saghaei, A. (2018). Automatic clustering based on crow search algorithm-k-means (CSA-k-means) and data envelopment analysis (DEA). *International Journal of Computational Intelligence Systems*, *11*(1), 1322–1337. https://doi.org/10.2991/ijcis.11.1.98

Ediyanto, M. N. M., & Satyahadewi, N. (2013). Classification of characteristics using the k-means cluster analysis method. *Buletin Ilmiah Matematika Statistik dan Terapannya*, *2*(2), 133–136.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, *11*(1), 40–56. https://doi.org/10.1016/j.apr.2019.09.009

Han J., & Kamber M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

Indonesia, P. R. (2007.). *Undang-undang republik Indonesia nomor 24 tahun 2007 tentang penanggulangan bencana*. DPR RI

Kandel, A., Tamir, D., & Rishe, N. D. (2014). Fuzzy logic and data mining in disaster mitigation. In: Teodorescu HN., Kirschenbaum A., Cojocaru S., Bruderlein C. (eds), *Improving disaster resilience and mitigation - IT means and tools*. NATO Science for Peace and Security Series C: Environmental Security (pp. 167–186). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9136-6_11

Kassambara, A., & Mundt, F. (2020). *Factoextra: Extract and visualize the results of multivariate data analyses*. https://cran.r-project.org/package=factoextra

Abdulsahib, A. K., & Kamaruddin, S. (2015). Graph based text representation for document clustering. *Journal of Theoretical and Applied Information Technology*, *10*(1), 1–13. https://www.researchgate.net/publication/281944315

Ng, K.-H., & Khor, K.-C. (2016). Evaluation on rapid profiling with clustering algorithms for plantation stocks on bursa malaysia. *Journal of Information and Communication Technology*, *15*(2), 63–84. https://doi.org/10.32890/jict2016.15.2.4

Nugroho, B. (2021). Perbandingan aplikasi algoritma kernel k-means pada graf bipartit dan k-means pada matriks dokumen- istilah dalam dataset penelitian covid-19 RISTEKBRIN. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *8*(2), 411–418. http://dx.doi.org/10.25126/jtiik.2021824365

Peterson, A. D., Ghosh, A. P., & Maitra, R. (2018). Merging k-means with hierarchical clustering for identifying general-shaped groups. *Stat*, *7*(1), 1–16. https://doi.org/10.1002/sta4.172

Priatmodjo, D. (2011). Penataan kota bermuatan antisipasi bencana. *Nalars*, *10*(2), 83–104. https://doi.org/10.24853 nalars.10.2.%25p

Prihandoko, P., & Bertalya, B. (2016). A data analysis of the impact of natural disaster using k-means clustering algorithm. *Jurnal Ilmiah Kursor*, *8*(4), 169–174. https://doi.org/10.28961/kursor.v8i4.109.

Prihandoko, P., Bertalya, B., & Ramadhan, M. I. (2017, July). An analysis of natural disaster data by using k-means and k-medoids algorithm of data mining techniques. In *15th International Conference on Quality in Research (QiR): International Symposium on Electrical and Computer Engineering* (pp. 221–225). IEEE. https://doi.org/10.1109/QIR.2017.8168485

Rachmawati, L. (2018). People's knowledge on hazard map and merapi hazard mitigation. *Jurnal Kependudukan Indonesia*, *13*(2), 143–156. https://doi.org/10.14203/jki.v13i2.324

Sadewo, M. G., Perdana Windarto, A., & Wanto, A. (2018). Penerapan algoritma clustering dalam mengelompokkan banyaknya desa/ kelurahan menurut upaya antisipasi/ mitigasi bencana alam menurut provinsi dengan k-means. In *Konferensi Nasional Teknologi Informasi dan Komputer (KOMIK)* (pp. 311–319). STMIK. http://dx.doi.org/10.30865/komik.v2i1.943

Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. (2013). *Practical graph mining with R*. CRC Press.

Supriyadi, B., Windarto, A. P., Soemartono, T., & Mungad. (2018). Classification of natural disaster prone areas in Indonesia using k-means. *International Journal of Grid and Distributed Computing*, *11*(8), 87–98. https://doi.org/10.14257/ijgdc.2018.11.8.08

Welton-Mitchell, C., James, L. E., Khanal, S. N., & James, A. S. (2018). An integrated approach to mental health and disaster preparedness: A cluster comparison with earthquake affected communities in Nepal. *BMC Psychiatry*, *18*(296). https://doi.org/10.1186/s12888-018-1863-z

Wen, L.-H., Shi, Z.-H., & Liu, H.-Y. (2019). Research on risk assessment of natural disaster based on cloud fuzzy clustering algorithm in Taihang mountain. *Journal of Intelligent & Fuzzy Systems*, *37*(4), 4735–4743. https://doi.org/10.3233/JIFS-179308

Yana, M. S., Setiawan, L., Ulfa, E. M., Rusyana, A., Statistika, J., Kuala, S., & Aceh, B. (2018). Penerapan metode k-means dalam pengelompokan wilayah menurut intensitas kejadian bencana alam di Indonesia tahun 2013-2018. *Journal of Data Analysis*, *1*(2), 93–102. https://doi.org/10.24815/jda.v1i2.12584.