Indonesian Forest Fire Data Clustering using Spatiotemporal Data Using Grid Density-Based Clustering Algorithm

Devi Fitrianah [1], Hisyam Fahmi [2] , Ade Putera Kemala[3], Muhammad Edo Syahputra[4]

[1,3,4] Bina Nusantara University, Jakarta 14480, Indonesia
[2] UIN Maulana Malik Ibrahim University, Malang 65144, Indonesia

Abstract. Forest fires are major environmental issues, especially in Indonesia which has the large area of forests. It becomes national problems that must be integrally and systematically resolved. Forest fires prediction and mapping are one of the approaches providing information about potential forest fire areas. The meteorological conditions (e.g., temperature, wind speed, humidity) are known features influencing forest fires to spread. In this research, we combine the information from meteorological data and the forest fires incident in Indonesia for a specific location and time to map and predict forest fire areas. Forest fires data obtained from BNPB website from 2011 until 2023 and then combined with meteorological data at the corresponding time. Grouping closest points into one cluster is the first step to map the data using IMSTAGRID algorithm. This algorithm is the adaptation of the grid density clustering method implemented for spatiotemporal data which provides a good clustering result with Silhouette values up to 0.8175.

Keywords: Forest Fires, Spatiotemporal, Data Mining, Grid Density Based Clustering, HDBSCAN

1 Introduction

Indonesia is a country with a large area of forest and a tropical climate, which often causes forest fires every year. Forest fires have numerous consequences, many of which are detrimental to humans and other ecosystems. Forest fires can affect ecological, climate change, economic and health effects [1]. Fires can destroy the habitat of animals, especially endangered wildlife in Indonesia. The ozone problem in the troposphere is influenced by smoke from fires as well, which can affect climate change. Fire smog causes considerable economic losses, such as tourism, agriculture, forestry, health, and the transport sector, because of poor visibility and heavy breathing [1]. According to the Indonesian National Board for Disaster Management, as of February of 2023 alone there has been 30 occurrences of forest fire in Indonesia. While the most severe year of forest fires in Indonesia happened in 2019 with 757 occurrences.

Forest fires can be caused by natural causes or intentionally by humans. Human-caused forest fires should be prevented and punished by the legal system, but environmental causes can be predicted and avoided. Climate is one of many natural factors that can cause forest fires because it influences the level of surface fuel dryness, the amount of oxygen, and the rate at which the fire spreads.[2]. As a result, an effort should be made to overcome this disaster; one approach is to develop an early warning system using spatiotemporal projection and predict the characteristics of potential forest fire areas.

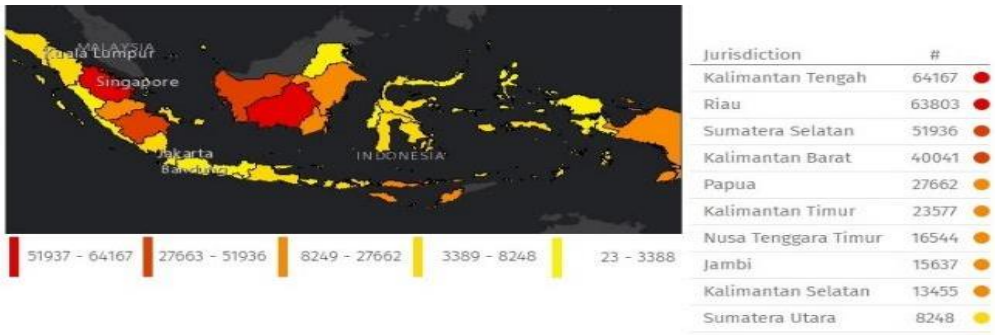


Fig 1. Greatest Number of Fire Alerts by Province (Jan 2013-July 2018) (Global Forest Watch 2014)

The spatiotemporal projection is generated by mapping the values of a 3-dimensional and time-evolving physical quantity into a 2-dimensional space with spatial and temporal axes [3]. Spatiotemporal data mining is broadly used to reveal patterns or phenomena of natural events such as the identification of earthquake disaster areas, forest fires identification, and mapping of potential fishing zones [4]. The clustered results of spatiotemporal mapping will be combined with meteorological features for the corresponding location to predict potential forest fire areas. The meteorological features that are used, for example, temperature, wind, humidity.

Therefore, in this research, we begin by analyzing the causes of forest fires and then develop a system to perform clustering on the dataset using the Imstagrid algorithm [5]. The system utilizes spatiotemporal clustering to cluster the forest fire locations. Meteorological features are combined with the clustering results to create a model that can cluster data and, hopefully, predict the potential area of forest fire in the future.

2 Previous Study

The study of forest fires has been widely implemented by many researchers around the world, which mostly utilize sensory technology. Umamaheshwaran proposed an image mining method using images from the Meteosat-SEVIRI sensor that produced a linear model of forest fires with vegetation and wind direction [6].

A study of the parameters influencing the forest fire process is also being conducted in order to formulate the characteristics of forest fire types. Time of fire, land cover type, altitude of forest land, slope of land, and forest fire statistics are all factors that can influence forest fires. This model has been validated using data from NOAA-AVHRR and Terra MODIS satellites [7]. Another point that needs to be considered in determining potential forest fire is the size of the grid from the image used against existing weather observations [8]. Another study of forest fire-related variables is how to determine the linkages between land cover types and vegetation with fires [9]. The study succeeded in describing the potential vegetation types of forest fires by using imagery from satellite sensing.

Huang et al. conducted research in high-risk areas to understand the spatial distribution of high fire risk using the HDBSCAN algorithm and early warning weather [10]. The data came from satellite data of Yunnan Province during 2015-2019

3 Spatiotemporal Clustering

Spatiotemporal clustering is one method for analyzing spatiotemporal data. Clustering spatiotemporal data can be done directly in 3-dimensional space (time and 2-dimensional space) or alternatively, i.e., in new temporal-spatial data or otherwise. [11].

The most widely used spatiotemporal clustering method for analyzing large spatiotemporal datasets is DBSCAN. It has the ability to find clusters of varying shapes such as linear, oval, concave and other shapes. DBSCAN, unlike other clustering algorithms, does not require the number of clusters to be determined. Birant and Kut proposed S-DBSCAN that improves the performance of DBSCAN by adding 3 marginal extras in DBSCAN to identify core objects, noise objects and adjacent clusters [12].

The spatiotemporal clustering method used for the data of natural phenomena and adjusting to the nature of the data is one of the methods developed based on density and grid. This method proved very robust to handle data of different data types with the result of accuracy reach 82.68% [4].

4 Data, Method, and Implementation of Imstagrid

4.1 Data

The forest fire data utilized in this study is spatiotemporal data which is combined with meteorological data at Indonesian area taken from August 2011 to 2023. The forest fire data obtained from BNPB website [18] includes longitude, latitude, date, month, and year.

Using the obtained latitude and longitude data, meteorological features of the location can be obtained from World Weather Online websites. The site provides a REST API for accessing weather data since July 1, 2008. The API is accessed under the "GET" method and will return weather components such as air temperature (in Celsius and Fahrenheit), weather description, weather icon, and wind speed in JSON (JavaScript Object Notation) format, XML, or CSV. Intake of weather data can be taken by city name, IP address, latitude, and longitude coordinates (in decimal), for UK, US, and Canadian countries can use zip code (World Weather Online 2008).

List parameters of the final dataset comprise of 14 features which are Year, Month, Day, Latitude, Longitude, tempC (temperature), precipMM (precipitation), humidity, cloudcover, HeatIndexC, DewPointC, windspeedKmph, WindGustKmph and aggregate Time with a total of 2846 rows of data. Despite the extensive range of features encompassed by this dataset, it is not without shortcomings. One notable limitation pertains to the unknown extent of severity in the forest fire data, which creates uncertainty regarding the potential impact of the fires on weather conditions. The influence of these fires on weather patterns remains unknown. It is conceivable to treat this type of data as an anomaly or outlier; however, given the dataset size, it is unlikely to exert a substantial influence on the overall results. Figure 2 shows the plot of the forest fire dataset used in this research.

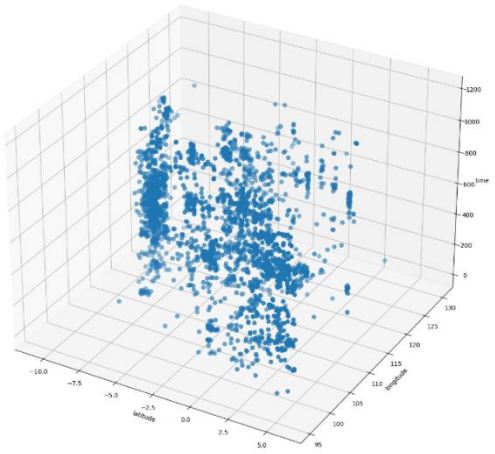


Fig 2. Dataset Visualization

4.2 Method

In this study we take several steps to get to the results. Our study consists of 5 phases as illustrated in Figure 3. For the data acquisition step we collect the data from BNPB website which is a government instantion for disaster management. Then, the dataset is cleaned and the resulting data is used to train IMSTAGRID. The created cluster is

analysed for the best parameters combination and validated using silhouette index method.

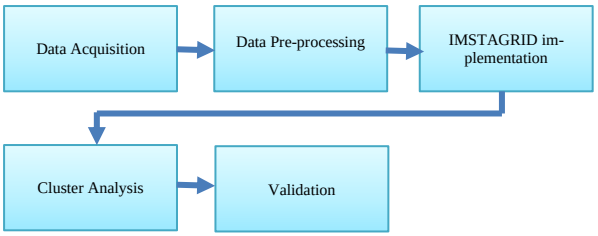


Fig 3. Methodology

4.3 HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [13] is a developed version of DBSCAN [14] algorithm which is used for clustering purposes. HDBSCAN improves density-based clustering method by establishing a hierarchical representation of the clusters [15], the algorithm's execution produces a hierarchy that can be effectively utilized for cluster extraction and outlier detection. HDBSCAN addresses the limitations of DBSCAN by enabling the identification of clusters with varying densities. For this research, we will use HDBSCAN algorithm as the baseline to compare how well the Imstagrid algorithm works.

4.4 Implementation IMSTAGRID Clustering Algorithm

In this study, we adopted the Imstagrid clustering algorithm for processing the spatio-temporal forest fire data. The adaptations for spatiotemporal clustering from the Imstagrid algorithm are in the partitioning phase and computing the distance threshold (r). In terms of partitioning mechanism, Imstagrid outperforms previous clustering methods such as ST-AGRID [16]. Using ST-AGRID would result in unequal data spread and interval, resulting in cells in a different shape of cube and gaps. Imstagrid overcame this problem by recommending a uniform interval (L) value for the spatial and temporal dimensions, resulting in a cube-shaped cell. During the partitioning phase, the data space/object is divided into cubes. It is necessary to perform a cubes interval calculation when determining the number of cubes. The cubes interval (L) value is obtained by dividing each dimension range (upper bound - lower bound) by the number of m cells. The number of intervals (L) is only for spatial dimensions (longitude and latitude), whereas for temporal dimensions, the cell interval for the dimension is using aggregate temporal due to its data structure. Unlike the AGRID+ [17] algorithm approach, which uses only one distance threshold value, the Imstagrid algorithm suggests a unique distance threshold for each dimension (spatial and temporal). Finally, Imstagrid improves the density compensation calculation, which determines the density threshold to determine whether or not a group is a cluster.

As each cube (spatial and temporal) is formed, each data object is stored in each cube that is relevant to its spatiotemporal coordinates. The phase is then adjusted in order to compute the distance threshold (r). Computing the distance threshold much depends on the partitioning phase in finding the value of L.

Where λ is interval weight parameter used, L is interval, and ϵ is a small integer number, so that the value of $r < L/2$. After the adaptation step has been done, the rest of the steps will be implemented similarly to those in the AGRID+.

In this step of our study, our clustering algorithm, IMSTAGRID is an improvement of the ST-AGRID applied to the forest fire dataset.
(1) Partitioning. The entire data space of forest fire data, which includes the location and time of fire, is partitioned into cells based on *m* which is the number of cells and *p* . The coordinates of each object are then assigned to a cube, and non-empty cubes are inserted into a hash table. The cube is a data structure with three dimensions (spatial and temporal), the first two of which are longitude and latitude, and the third of which is time.
(2) Computing distance threshold. Determine the neighbourhood radius (*r*) for each data point to other data point in a neighbour. After that we have and based on the appropriate time unit.
(3) Calculating densities. For each object of data, count the number of objects both in its neighboring cells and in its neighbourhood as its density.
(4) Compensating Densities. For each object of data compute the ratio of the volume of all neighbours and that of neighbours considered and use the product of the ratio and the density of the cell as the new density.
(5) Calculating density threshold (*DT*). The average of all compensated densities is calculated and then the *DT* is computed by finding the average of the density compensation divided by theta parameters which are coefficients that can be tuned in to get a different cluster level.
(6) Clustering automatically. First, each object with a density greater than DT is considered a cluster. Then, for each object, examine each object in the neighboring cells to see if its density exceeds the density threshold and its distance from the object exceeds the distance threshold. If yes, then merge the two clusters to which the two objects belong. Continue the merging procedure described above until all eligible object pairs have been checked.

4.5 Hyperparameter Search

Both the HDBSCAN and Imstagrid algorithms use different hyperparameters that can be tuned to improve clustering results. In this step, we use the Gridsearch method to try and find the best parameters for each algorithm. Tables 1 and 2 show the hyperparameter search space for HDBSCAN and Imstagrid, respectively.

Table 1. HDBSCAN hyperparameters search

Hyperparameter	Search Space
Min_samples	[2, 3, 4, 5, 6, 7, 8, 9]
Min_cluster_size	[2, 3, 4, 5, 6, 7, 8, 9, 10, 12]
Cluster_selection_epsilon	[0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 0.9]
Cluster_selection_method	['eom' , 'leaf']

Table 2. Imstagrid hyperparameter search

Hyperparameter	Search Space
L_spasial	[6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17,18,19,20,21,22,23,24,25]
Theta	[100, 90, 80, 60, 50, 40, 35, 30, 15, 5]
Lambda	[0.2, 0.5, 0.7, 0.8, 0.9]

4.6 Evaluation

The method that will be used to evaluate the clustering model is called silhouette analysis. This method calculates the silhouette index by using the density level within a cluster (intra-cluster distance) and the distance between each cluster (inter-cluster distance).

The best achievable value for the silhouette index is 1, while the lowest value is -1. If the silhouette score approaches 0, it indicates a significant overlap between clusters. Formula 1 shows the way to calculate the silhouette index.

$$Si \frac{bi-ai}{\max(bi,ai)} \quad (1)$$

Si = The silhouette score for data point i
bi = The inter-cluster distance for data point i
ai = The intra-cluster distance for data point i

5 Results

For the HDBSCAN baseline method, a grid search was performed to determine the optimal hyperparameters. The results yielded the following values: min_samples: 9, min_cluster_size: 3, cluster_selection_epsilon: 0.1, and cluster_selection_method: ‘leaf’. The silhouette score obtained for this configuration is 0.5965, it can be understood that HDBSCAN struggled to perform clustering task on this dataset. Figure 4 illustrates the visualization of the clustering outcome.

Table 3 shows the best hyperparameters obtained using the Gridsearch method for the Imstagrid algorithm. 100 different combinations of hyperparameters were investigated and the six best scenarios were displayed on table 3. This method yielded the highest silhouette score of 0.8175, which outperformed our baseline model. Furthermore, significant observations can be drawn from the data presented in Table 3. The clustering results improve as the values of L, lambda, and theta increase. It is worth noting, however, that the theta value has a limit of 30 above which the silhouette score does not improve further. Figure 5 depicts the clustering results obtained using the Imstagrid algorithm.

Table 3. Imstagrid Hyperparameter Search Result

Scenario	L	Agg	Lambda	Theta	Silhouette Score
1	24	1	0.9	30	0.8124
2	24	1	0.9	60	0.8124
3	24	1	0.9	100	0.8124
4	25	1	0.9	30	0.8175
5	25	1	0.9	60	0.8175
6	25	1	0.9	100	0.8175

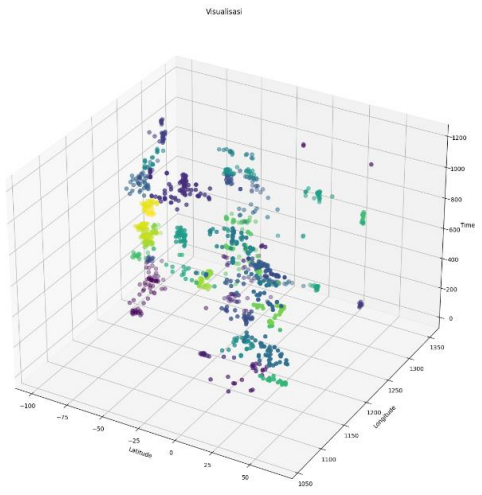
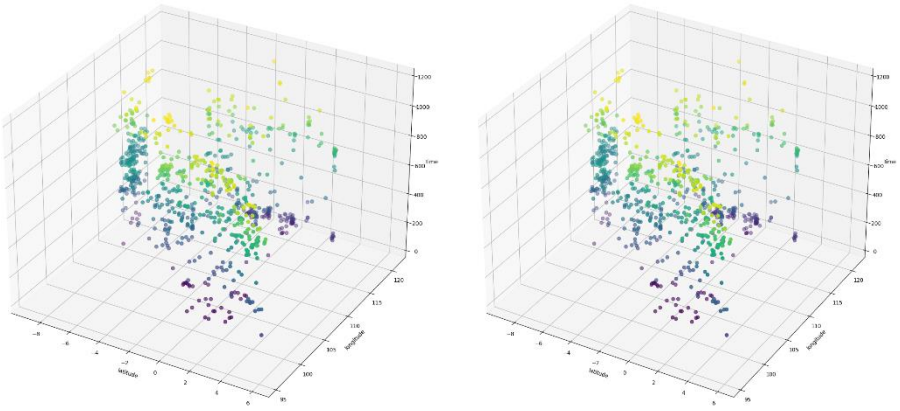


Fig 4. HDBSCAN Clustering result



(a) (b)

Fig 5. a) Scenario 1 Visualization b) Scenario 3 Visualization

6 Conclusion

Real-world data gathering in research demonstrates that conventional algorithms, such as HDBSCAN, may not necessarily perform well in analyzing datasets. This research has successfully obtained information indicating that the Imstagrid method exhibits adequate performance and capability in effectively analyzing and clustering datasets with silhouette score of 0.8175.

The process of hyperparameter searching also significantly influences the improvement of the final model's performance. It is observed that, specifically for this dataset, higher values of L are positively correlated with the model's performance. Furthermore, it is known that the upper limit for the theta value is 30, beyond which the performance does not further improve.

For future research in this field, it is recommended to further develop the analysis of data and mapping of areas prone to forest fires in the Indonesian region.

References

[1] P. Hirschberger, “Forests ablaze: causes and effects of global forest fires,” *WWF: Berlin, Germany*, 2016.

[2] L. Syaufina and D. A. F. Hafni, “VARIABILITAS IKLIM DAN KEJADIAN KEBAKARAN HUTAN DAN LAHAN GAMBUT DI KABUPATEN BENGKALIS, PROVINSI RIAU Variability of Climate and Forest and Peat

Fires Occurrences in Bengkalis Regency, Riau," *Journal of Tropical Silviculture*, vol. 9, no. 1, pp. 60–68, 2018.

[3] H. Nakamura Miyamura, S. Hayashi, Y. Suzuki, and H. Takemiya, "Spatio-Temporal Mapping-A Technique for Overview Visualization of Time-Series Datasets," *Progress in NUCLEAR SCIENCE and TECHNOLOGY*, vol. 2, pp. 603–608, 2011.

[4] D. Fitrianah, A. N. Hidayanto, H. Fahmi, J. Lumban Gaol, and A. M. Arymurthy, "ST-AGRID: A spatio temporal grid density based clustering and its application for determining the potential fishing zones," *International Journal of Software Engineering and its Applications*, vol. 9, no. 1, pp. 13–26, 2015, doi: 10.14257/ijseia.2015.9.1.02.

[5] D. Fitrianah, H. Fahmi, A. N. Hidayanto, and A. M. Arymurthy, "Improved partitioning technique for density cube-based spatio-temporal clustering method," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8234–8244, 2022, doi: 10.1016/j.jksuci.2022.08.006.

[6] R. Umamaheshwaran, W. Bijker, and A. Stein, "Image mining for modeling of forest fires from Meteosat images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 246–253, 2007, doi: 10.1109/TGRS.2006.883460.

[7] P. A. Hernández-Leal, A. González-Calvo, M. Arbelo, A. Barreto, and A. Alonso-Benito, "Synergy of GIS and remote sensing data in forest fire danger modeling," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 1, no. 4, pp. 240–247, 2008, doi: 10.1109/JSTARS.2008.2009043.

[8] N. Khabarov, E. Moltchanova, and M. Obersteiner, "Valuing weather observation systems for forest fire management," *IEEE Syst J*, vol. 2, no. 3, pp. 349–357, 2008, doi: 10.1109/JSYST.2008.925979.

[9] M. A. Tanase and I. Z. Gitas, "An examination of the effects of spatial resolution and image analysis technique on indirect fuel mapping," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 1, no. 4, pp. 220–229, 2008, doi: 10.1109/JSTARS.2009.2012475.

[10] J. Huang *et al.*, "Fire Risk Assessment and Warning Based on Hierarchical Density-Based Spatial Clustering Algorithm and Grey Relational Analysis," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/7339312.

[11] T. Abraham and J. Roddick, "Opportunities for Knowledge Discovery in Spatio-Temporal Information Systems," *Australasian Journal of Information Systems*, vol. 5, no. 2, 1998, doi: 10.3127/ajis.v5i2.338.

[12] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data Knowl Eng*, vol. 60, no. 1, pp. 208–221, 2007, doi: 10.1016/j.datak.2006.01.013.

[13] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Trans Knowl Discov Data*, vol. 10, no. 1, pp. 1–51, 2015, doi: 10.1145/2733381.

[14] M. Daszykowski and B. Walczak, "A Density-Based Algorithm for Discovering Cluster in Large Spatial Databases with Noise," *Comprehensive*

Chemometrics: Chemical and Biochemical Data Analysis, Second Edition: Four Volume Set, vol. 2, pp. 565–580, 1996, doi: 10.1016/B978-0-444-64165-6.03005-6.

[15] G. Stewart and M. Al-Khassaweneh, "An Implementation of the HDBSCAN* Clustering Algorithm," *Applied Sciences (Switzerland)*, vol. 12, no. 5, pp. 1–21, 2022, doi: 10.3390/app12052405.

[16] Fitrianah, D., Hidayanto, A. N., Fahmi, H., Gaol, J. L., & Arymurthy, A. M. (2015). ST-AGRID: A spatio temporal grid density based clustering and its application for determining the potential fishing zones. *International Journal of Software Engineering and Its Applications*, *9*(1), 13-26.

[17] Fitrianah, D., Fahmi, H., Hidayanto, A. N., & Arymurthy, A. M. (2022). Improved partitioning technique for density cube-based spatio-temporal clustering method. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 8234-8244.

[18] BNPB. Data bencana Indonesia [Internet]. [cited 2023 May 15]. Available from: https://dibi.bnpb.go.id/xdibi