A Comparison study of DBScan and K-Means Clustering in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007

G C Pamuji 1,*, H Rongtao 2
1,2 School of Computer and Software, Nanjing University of Information Science and Technology, Jiangsu, China.
1 Master of Information System, Faculty of Postgraduate, Universitas Komputer Indonesia, Indonesia.

Email: *gcpamuji@unikom.ac.id

Abstract. The purpose of this study is to compare between two different of cluster analysis algorithm in data mining on the Tropical Rainfall Measuring Mission (TRMM). The TRMM is a joint mission between NASA and the Japan Aerospace Exploration (JAXA) Agency to study rainfall for weather and climate research. The TRMM satellite data-sets used in this research is a 3-hourly rainfall data within 10 years from 1998 to 2007. These data-sets will be analyzed by two different cluster analysis algorithms in data mining which are K-means and DBScan. In this paper, rainfall data in Jakarta based on TRMM was analyzed and compared in the efficiency and the accuracy using each algorithm. The comparison results of the two algorithmic processes can be seen from several parameters, especially from the number of clusters formed and the time needed to process the model.

1. Introduction
A short period of an intense rainfall could cause a flash flood and made an unusual severity [1]. On January 1st 2020, The Indonesian capital of Jakarta was hit by a nearly 400 millimeters of rainwater in overnight [2]. At least 66 people killed and more than 60,000 was evacuated to temporary shelters. The Tropical Rainfall Measuring Mission (TRMM) designed to monitor and study tropical rainfall, it is one of the satellite remote sensing techniques that take an important role in filling data gaps [3].

The Meteorology, Climatology, and Geophysical Agency (BMKG) of Indonesia uses TRMM satellite data to observe the rainfall conditions over Indonesia [4]. It is often used in various research for studying rainfall and climate [5-10]. The previous study, Nuryanto et al. [11] researched the potential flooding rainfall of Greater Jakarta using TRMM satellite data. Prabawadhani et al. [12] used Intensive Observation Period (IOP) to find out the rainfall characteristics in Jakarta. Based on all the previous researches mentioned above, only a few studies mentioning the usage of clustering algorithms, specifically DBScan and K-Means.
The objective of this study is to cluster extreme rainfall in Jakarta region based on rainfall data from TRMM. The cluster analysis is done using the K-means and DBScan algorithms separately. To compare between these algorithms and find out the efficiency and accuracy of each clustering analysis algorithm.

2. Method
Cluster analysis consist in distinguishing, in the set of analyzed data, the groups are called clusters [13]. Both K-means and DBScan are popular algorithms in cluster analysis. K-means clustering aims to partition n observations into k clusters in which observations belong to the cluster with nearest mean. Density-based spatial clustering of applications with noise (DBScan) groups together points that are closely packed together.

The proposed methods have several steps for comparing two different clustering algorithm. The first step is data collection from TRMM satellite in the period 10 years, from 1998-2007. Next, the datasets will be analyzed using K-means and DBScan separately. Lastly, the results of both algorithm will be compared, analyzed, and draw the conclusions. These steps is described in Figure 1.

3. Results and Discussion
3.1. Data Collection
The data-sets used in this study was taken from TRMM satellite data for Jakarta region. It consists 3-hourly of time-stamps and rainfall from nine points of latitude and longitude that defined the area. It is a 10-years period of time from 1998 to 2007. Table 1 shows latitude and longitude of each point. The total data-points which will be processes is 29,213 data. This study did not use more recent data-sets because it will be used in further research.
Table 1. Latitude and longitude.

| Name | Latitude | Longitude |
|------|----------|-----------|
| JKT1 | -5.875   | 106.625   |
| JKT2 | -5.875   | 106.875   |
| JKT3 | -5.875   | 107.125   |
| JKT4 | -6.125   | 106.625   |
| JKT5 | -6.125   | 106.875   |
| JKT6 | -6.125   | 107.125   |
| JKT7 | -6.375   | 106.625   |
| JKT8 | -6.375   | 106.875   |
| JKT9 | -6.375   | 107.125   |

3.2. K-means clustering

K-means algorithm is composed of three steps. The first step is initialization. One of the input of this step is \( k \) value as number of clusters. It will randomly choose data points from the data-sets as initial centroids. The algorithm is iterative nature and random initialization of the centroids at the start of the algorithm. In this study, the value of \( k \) has been determined based on the data-sets. Due to the nature of rainfall, it can be determined as 3 different cluster. It will be representing light, medium, and heavy rainfall.

The second step is cluster assignment. All of the data points that are the closest (similar) to a centroid will create a cluster. Euclidian distance is used as divergence between data points and every centroid. This will draw a line between centroids, then a boundary line divides it into two clusters.

Last step is moving the centroid. After new clusters formed, it will choose a centroids that is the mean of all the data points in a cluster as its center. The process will repeating steps two and three until the centroids stop moving. The final results of centroids formed for each area of Jakarta shown in Table 2.

Table 2. Centroids for each point area of Jakarta.

| Cluster | JKT1 | JKT2 | JKT3 | JKT4 | JKT5 | JKT6 | JKT7 | JKT8 | JKT9 |
|---------|------|------|------|------|------|------|------|------|------|
| Cluster 0 | 0.091 | 0.091 | 0.101 | 0.064 | 0.053 | 0.065 | 0.096 | 0.079 | 0.088 |
| Cluster 1 | 6.919 | 8.125 | 7.246 | 8.712 | 10.956 | 9.549 | 5.037 | 5.625 | 6.026 |
| Cluster 2 | 1.489 | 1.569 | 1.542 | 2.536 | 2.533 | 2.403 | 3.539 | 3.827 | 3.421 |

Based on centroids, JKT5 had the highest centroid compared to other points. Figure 2 illustrates a better view of the centroids.
The gap between clusters are recognizable, which is described in Figure 2. The blue line representing cluster 0, the green line representing cluster 1, and the red line represented cluster 2. Figure 3 is a scatter plot graph focusing in cluster 1 as the heavy rainfall representative, illustrating two area in Jakarta, JKT5 and JKT6.
3.3. **DBScan clustering**

DBSCAN is the most well-known density-based clustering algorithm [15], it groups the points that are close to each other based on a distance measurement, usually Euclidian distance and minimum number of points. First, we set the value of epsilon and minpts (minimal points). Epsilon specify the size of the neighborhood, it has real data-type and the range is from 0 to indefinite adjusted to the data-sets. The minimum number of data-points within the epsilon radius forming a cluster is called as minpts, it is integer and the range is from 1 to indefinite. Due to of a large data-sets, it is hard to determine the value of epsilon and minpts. Based on a trial-error and repetitive experiment, the value of epsilon filled with 3.0 and the minpts is 1,000. Figure 4 is a Scatter plot result from the DBScan algorithm focusing between two points who have different cluster.

![Scatter plot result from the DBScan algorithm focusing between two points who have different cluster.](image)

**Figure 4.** Scatter plot result from the K-means algorithm (Green: Cluster 0, Blue:Cluster 1)

3.4. **Result comparison**

Aarthi [14] had a similar research comparing two algorithms in several parameters, which consists of number of clusters, cluster instance, number of iteration, sum of squared errors, and time taken to build model. This study compared number of clusters, variables, measure types, divergence, cluster instances, and time taken to process algorithm. Table 3 shows the results of the comparison between the two algorithms.
Table 3. Results of the comparison between K-means and DBScan

| Parameters          | K-means               | DBScan               |
|---------------------|-----------------------|----------------------|
| Number of clusters  | 3                     | 2                    |
| Variables           | $k = 3$               | Epsilon =3           |
|                     |                       | Minpts = 1000        |
| Measure types       | Bregman Divergence    | Bregman Divergence   |
| Divergence          | Squared Euclidian distance | Squared Euclidian distance |
| Cluster Instance    | Cluster 0 = 27,671    | Cluster 0 = 2,706    |
|                     | Cluster 1 = 222       | Cluster 1 = 26,507   |
|                     | Cluster 2 = 1,320     |                      |
| Time taken to process algorithm | 2s           | 13m 32s              |

4. Conclusion
Using the data mining techniques K-Means and DBSCAN the rainfall data-sets analyzed. Each algorithm formed different number of clusters. K-Means used a predetermined variable ($k$) with a value of 3, and DBScan has the ability to determine how many cluster, which can be formed based by data-points. The size of the data-sets is affecting the process time. K-means is faster in processing larger data-sets. Each of the algorithm have their own strengths and weakness. Due to the nature of the data-sets, which are from data-sets type and size, K-Means produced a more efficient and accurate results than DBScan.

Acknowledgements
We would like to thank the support from Wido Hanggoro for the datasets he provided, Aprilia Ramdhani Hidajat who provided advice and insights during this research, INCITEST 2020 Committee, and to all parties involved who have assisted in writing this report either directly or indirectly.

References
[1] Archer D R and Fowler H J 2018 Characterising flash flood response to intense rainfall and impacts using historical information and gauged data in Britain *J. Flood Risk Manag.* 11, pp. S121–33
[2] Hays B and Coote D 2020 New Year's Eve flooding kills 16, displaces thousands in Jakarta. UPI Retrieved from https://www.upi.com
[3] Liu Z 2016 Comparison of Integrated Multisatellite Retrievals for GPM (IMERG) and TRMM Multisatellite Precipitation Analysis (TMPA) monthly precipitation products: Initial results *J. Hydrometeorol.* 17, pp. 777–90
[4] Kuswanto H, Setiawan D and Sopaheluwakan A 2019 Clustering of Precipitation Pattern in Indonesia Using TRMM Satellite Data *Eng. Technol. Appl. Sci. Res.* 9, pp. 4484–9
[5] As-Syakur A R, Osawa T, Miura F, Nuarsa I W, Ekayanti N W, Dharma I G B S, Adnyana I W S, Arthana I W and Tanaka T 2016 Maritime Continent rainfall variability during the TRMM era: The role of monsoon, topography and El Niño Modoki *Dyn. Atmos. Ocean.* 75, pp. 58–77
[6] Prakash S, Mitra A K, Pai D S and AghaKouchak A 2016 From TRMM to GPM: How well can heavy rainfall be detected from space? *Adv. Water Resour.* 88, pp. 1–7
[7] Giarno G, Hadi M P, Suprayogi S and Murti S H 2018 Distribution of Accuracy of TRMM Daily Rainfall in Makassar Strait *Forum Geogr.* 32
[8] As-syakur A R 2015 Spatio-Temporal Variations of Rainfall and SST Anomaly over Indonesia during ENSO Modoki Event in 2010 *J. Mar. Aquat. Sci.* 1, pp. 23
[9] Liu J, Duan Z, Jiang J and Zhu A X 2015 Evaluation of three satellite precipitation products TRMM 3B42, CMORPH, and PERSIANN over a subtropical watershed in China Adv. Meteorol. 2015

[10] Maggioni V, Meyers P C and Robinson M D 2016 A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era J. Hydrometeorol. 17, pp.1101–17

[11] Nuryanto D E et al 2017 IOP Conf. Ser.: Earth Environ. Sci. 54 012028

[12] Ratna Prabuwadhani D, Harsoyo B, Handoko Seto T, Bayu Rizky Prayoga M, Besar Teknologi Modifikasi Cuaca – Badan Pengkajian dan Penerapan Teknologi B, Kunci K and Hujan C 2016 Karakteristik temporal dan spasial curah hujan penyebab banjir di wilayah DKI jakarta dan sekitarnya Spatial and Temporal Characteristics of Flood-Induced Rainfall in Jakarta Area and Its Surroundings J. Sains Teknol. Modif. Cuaca 17, pp.21–5

[13] Wierzchon S T and Klopotek M A 2015 Algorithms of Cluster Analysis (Warsaw: Institute of Computer Science Polish Academy of Sciences

[14] Aarthi R 2019 Cluster analysis of extreme rainfall seasons in particular J. of The Gujarat Research Society 21, pp.1–6

[15] Xu D and Tian Y 2015 A Comprehensive Survey of Clustering Algorithms Ann. Data Sci. 2, pp.165–93