ICIC Express Letters
Part B: Applications
Volume **13**, Number **11**, November **2022**
 ISSN 2185-2766
pp. **1199–1205**

RAINFALL ANALYSIS IN SEMARANG CITY USING K-MEANS AND AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS

Prabowo Wahyu Sudarno, Ahmad Ashari and Mardhani Riasetiawan

Computer Science and Electronics Department
Universitas Gadjah Mada
Sekip Utara PO BOX BLS 21, Yogyakarta 55281, Indonesia
prabowowahyu45@mail.ugm.ac.id; {ashari; mardhani}@ugm.ac.id

Received March 2022; accepted June 2022

Abstract. *The state of rainfall can be identified by using several other climate data. This paper analyzes the characteristic cause factor of rainfall in Semarang City through several climate data, such as temperature and humidity, using AHC and K-Means. Clustering methods divide the range of rainfall into three clusters. Besides, we also evaluate the influence of different variables that affect rainfall clusters. This paper evaluates the performance of clustering algorithms using Silhouette score, Calinski-Harabasz score, and Davies Bouldin score. The performance validation correlation between rainfall (RR) and average humidity (RHavg) using the Silhouette score has the highest score of 0.42, the highest Calinski-Harabasz score is 451, and the lowest score using the Davies Bouldin score is 0.89. The performance validation correlation between rainfall (RR) and average temperature (Tavg) using the Silhouette score has the highest score of 0.46, the highest Calinski-Harabasz score is 2438, and the lowest value using the Davies Bouldin score is 0.86. All of these results are generated by the K-Means algorithm, showing the superiority of the K-Means algorithm over AHC in clustering rainfall. The conclusion explained that there is a possibility that high-intensity rain will occur in Semarang City when the humidity is above 80%, and the temperature is between 26 to 29.*
Keywords: Clustering, K-Means, AHC, Machine learning, Rainfall

1. **Introduction.** Indonesia is located in the tropics and has a relatively high annual rainfall. Rainfall affects life activities such as community safety, agricultural production, fisheries, and plantations, according to the conditions in their respective regions. Besides being beneficial, rainfall can also cause landslides and floods. Therefore, it is necessary to know how high and low rainfall is in an area to prevent disasters caused by rain. Rainfall values are different in each region, and the causative factors cannot be generalized [1]. Therefore, it is necessary to determine what causes rain to fall in an area precisely. By building a specific model to an area, we can more accurately predict the arrival of rain and minimize the risk of disasters occurring due to rain.

In terms of weather parameters, most studies monitor temperature and humidity. The weather forecast data in several studies were collected by sensors or from meteorological stations [2]. Rainfall has spatial characteristics that are periodically captured by the station, such as the average temperatures (Tavg) and average humidity (RHavg). These characteristics can be used to group the data into several clusters, providing information on the causes of rainfall intensity. Such information will benefit humankind by making it possible to prepare for and mitigate landslides and floods. In the machine learning domain, this problem of making clusters from data can be overcome by using clustering methods.

Two methods that can be applied to solving this case are the K-Means and Agglomerative Hierarchical Clustering (AHC) algorithm. The K-Means and AHC methods work

DOI: 10.24507/icicelb.13.11.1199

by dividing the data into clusters and making the data precisely in one cluster [3]. AHC is a clustering algorithm that builds nested clusters from individual data point clusters to become a single cluster gradually by combining two pairs of closest clusters [4]. In the context of determining rainfall using climate data, [5] found that the best cluster consisted of around 3 clusters, with a Silhouette score in the range of 0.1 to 0.38. Next, in analyzing one area in Indonesia, [6] identified rainfall using fuzzy clustering and K-Means. The study acquired a Silhouette score of 0.24 to 0.46. These results are certainly influenced by the features of the dataset used and the undoubtedly different tuning. However, instead of comparing the metrics, we add some state-of-the-art comparison with current research that talks about rainfall prediction. A study explained how rain prediction in Indonesia is made using a data mining technique [7]. However, the study only analyzed two clustering target variables: rain and not rain. Thus, it does not yet provide detailed characteristics of the rain intensity prediction compared to other climate data. In this study, we use the K-Means and AHC methods to display the mapping of rainfall grouping results in Semarang City and correlate it with other climate data. This aims to find out how high and low rainfall is in each area easier. In this study, we want to determine what factors are causing rainfall intensity by looking at several clusters produced by several clustering algorithms.

The contents of this study are structured as follows: Section 2 discusses the dataset used in this study; Section 3 explains what clustering is, what methods we are using, and how to evaluate clusters; Section 4 contains the experiment result and discussion about the results, and Section 5 delivers the conclusion of this study.

2. **Dataset.** The climate dataset of Semarang City was taken from Meteorology and Geophysics Agency (BMKG), specifically in Semarang geophysical station. The dataset has 1767 instances with four feature attributes. The feature dataset that is used in this research is explained in Table 1.

Table 1. Data dictionary

Data	**Description**
RR	Rainfall (mm)
Tavg	Average of temperature (Celsius)
RHavg	Average humidity (%)

In this study, we utilize several data preprocessing techniques, such as replacing missing values, removing duplicate data, checking for inconsistent data, and correcting errors in data. Empty data is usually caused by new data for which there is no information [8]. This research will fill null data with predictive values using the KNN imputer method. KNNimputer is a scikit-learn class that helps to predict missing values in a dataset. The missing values for each sample were calculated using the mean of the nearest neighbours found in the training dataset. Two samples are said to be close if none of the features is close.

3. **Clustering.** This section discusses what clustering methods we use and how they work to solve the problem. In addition, this section also discusses how to validate the clustering results. Data mining is research that focuses on statistics, machine learning, and databases. One of the data mining algorithms, the clustering algorithm, is often used to make multiple analyses and predictions. The clustering algorithms work by dividing a dataset into a collection of similar data called clusters. By grouping, we know the patterns in the data. The results were obtained in the form of similarities between objects in the cluster [9]. In this section, Section 3.1 discusses the clustering algorithm used in this study, and Section 3.2 discusses how we evaluate the clustering models' performance.

3.1. **Clustering and model.** Clustering algorithms of machine learning are successfully used to divide the data into several clusters based on similar characteristics. This study utilizes two clustering algorithms, namely the K-Means and the Agglomerative Hierarchical Clustering (AHC).

3.1.1. *K-Means clustering.* K-Means clustering is an unsupervised algorithm that is used to group different objects into clusters. A cluster is a collection of homogeneous data objects in one cluster and heterogeneous with entities in another cluster. A collection of data objects can be grouped as a unit to be considered a type of data compression [10]. K-Means create k clusters from n number of data, where the user assigns the k value. The algorithm is explained as follows [11].

Step 1: Choose k cluster centroid points u at random.

Step 2: Repeat the procedure until convergence is reached.

a) For each data point i, calculate the class it should belong to using the following formula:

$$C^{(i)} := \arg\min \left\| X_i^{(i)} - u_j \right\|^2 \tag{1}$$

where X_i is the features of data point and u_j is the centroid points of class j.

b) Recalculate the class centroid u for each class j using the following formula:

$$u_j := \frac{\sum_{i=1}^{n} 1\left\{C^{(i)} = j\right\} x^i}{\sum_{i=1}^{n} 1\left\{C^{(i)} = j\right\}} \tag{2}$$

3.1.2. *Agglomerative Hierarchical Clustering (AHC).* Agglomerative Hierarchical Clustering (AHC) is a bottom-up approach that starts with individual objects in a cluster, and builds clusters from unrefined clusters until the whole dataset belongs to one cluster [12]. Then the looping process is carried out, wherein in each loop, the algorithm combines the closest pair of clusters that meet the inequality criteria. The loop process ends when all data is included in a cluster. AHC produces a tree-like structure called a dendrogram that visualizes multiple high-level partitions of the data set. AHC helps in generating small clusters, which offer informative data displays. The pseudocode of AHC is explained as follows [13].

Step 1: At first, each data point forms a cluster on its own, i.e., $G_i = \{x_i\}$ $(i = 1, 2, \ldots, n)$.

Step 2: Calculate the distances between each pair of cluster objects and create a distance matrix $D = (D_{ij})c$, where c is the number of clusters.

Step 3: Merge the clusters that are closest to each other (for example, G_p and G_q) into a new cluster G_r, so that $G_r = G_p \cup G_q$, and let $c = c - 1$.

Step 4: Examine the number of clusters. If the class number c exceeds the desired number of clusters, go to Step 2; otherwise, proceed to Step 5.

Step 5: Provide clustering results.

3.2. **Performance validation.** Validation is a performance evaluation step of the clustering algorithm used. In this study, we use several validation techniques, namely Silhouette score [14], the Callinski-Harabasz score [15], and Davies Bouldin score [16].

3.2.1. *Silhouette score.* The Silhouette score S is calculated for each data point i by taking the mean intra-cluster distance a and the mean nearest-cluster distance b into account. An A score close to $+1$ indicates that the data point is in the correct cluster. An A score of 0 indicates that the data point may belong in another cluster. Result -1 indicates that the data point is in an incorrect cluster [17].

$$S = \frac{1}{N} \sum_{i=0}^{N} \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3}$$

where a_i is the average dissimilarity with other data in cluster and b_i is the lowest dissimilarity to any non-member cluster for each data point i.

3.2.2. *The Callinski-Harabasz.* The Callinski-Harabasz is the ratio between-clusters dispersion mean and the within-cluster distribution [18]. Let D' be the clustered data, where $D' = \{x_1, x_2, \ldots, x_n\}$ and $D' \subset D$. The formula for Callinski-Harabasz s with k number of clusters is as follows:

$$S = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1} \tag{4}$$

where $tr(B_k)$ is the between group dispersion matrix and $tr(W_k)$ is the within-cluster dispersion matrix defined by

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{5}$$

$$B_k = \sum_{q=1}^{k} m (c_q - c_{D'})(c_q - c_{D'})^T \tag{6}$$

where C_q is the set point of the q-th cluster; c_q is the center point of cluster q; $c_{D'}$ is the center point of clustered data D'; and m is the number of data that exist in cluster q.

3.2.3. *The Davies Bouldin score.* The Davies Bouldin criterion (DB) is calculated based on a distance ratio between clusters. The index represents the average similarity measure of each cluster to the cluster that is most like it. A higher score indicates that the clusters are closer together and less dispersed [19]. The equation for Davies Bouldin score is as follows:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left\{ \frac{I(c_i) + I(c_j)}{I(c_i, c_j)} \right\} \tag{7}$$

$I(c_i)$ represents the mean of the distances between the objects belonging to cluster C_i and its center. $I(c_i, c_j)$ represents the distance between the centers of the two clusters C_i and C_j. For each cluster i in the partition, we look for cluster j which maximizes the index described as follows:

$$R_{ij} = \frac{I(c_i) + I(c_j)}{I(c_i, c_j)} \tag{8}$$

Therefore, the best clustering is that which minimizes the average of the value calculated for each cluster. In other words, the best clustering is the one that minimizes the similarity between the clusters.

4. **Result and Discussion.** This section discusses the performance and the results using the proposed method in this paper. First, we train the dataset with the K-Means and AHC algorithm. Several optimized parameters' values used to produce the best result are explained in Table 2 and Table 3.

First to be tested is the rainfall (RR) with the average humidity (RHavg). It can be seen in Figure 1 that the results of clustering using K-Means and AHC processing are divided into three clusters. We can also see that high rainfall (RR) intensity mostly occurs when humidity is above 80%. The result is explained in the highlighted dashed rectangle.

Next to be tested is the rainfall (RR) with the average temperature (Tavg). It can be seen in Figure 2 that the results of clustering using K-Means and AHC processing are divided into three clusters. High rainfall (RR) intensity mostly occurs when the temperature is within 26 degrees to 29 degrees. The highest rainfall intensity is explained in the highlighted dashed circle.

The result shows that by applying the clustering using K-Means and AHC for the Semarang City climate, the rainfall level can be divided into several clusters and analyzed.

Table 2. K-Means parameter tunning

K-Means					
Parameter	**n_cluster**	**n_init**	**Max_iter**	**Tol**	**Precompute distance**
Description	The number of clusters that will be generated.	The number of times the K-Means algorithm will be executed with various centroid seeds.	The number of iterations of the K-Means algorithm that can be performed in a single run.	Convergence value to a tolerance of the difference in the cluster centers of two consecutive iterations using the Frobenius norm.	Precompute distances
Value	3	31	412	0.0001	auto

Table 3. AHC parameter tunning

AHC			
Parameter	**n_clusters**	**Affinity**	**Linkage**
Description	The number of clusters that must be discovered	The metric used to calculate the linkage	Which linkage criterion should be used
Value	3	euclidean	ward

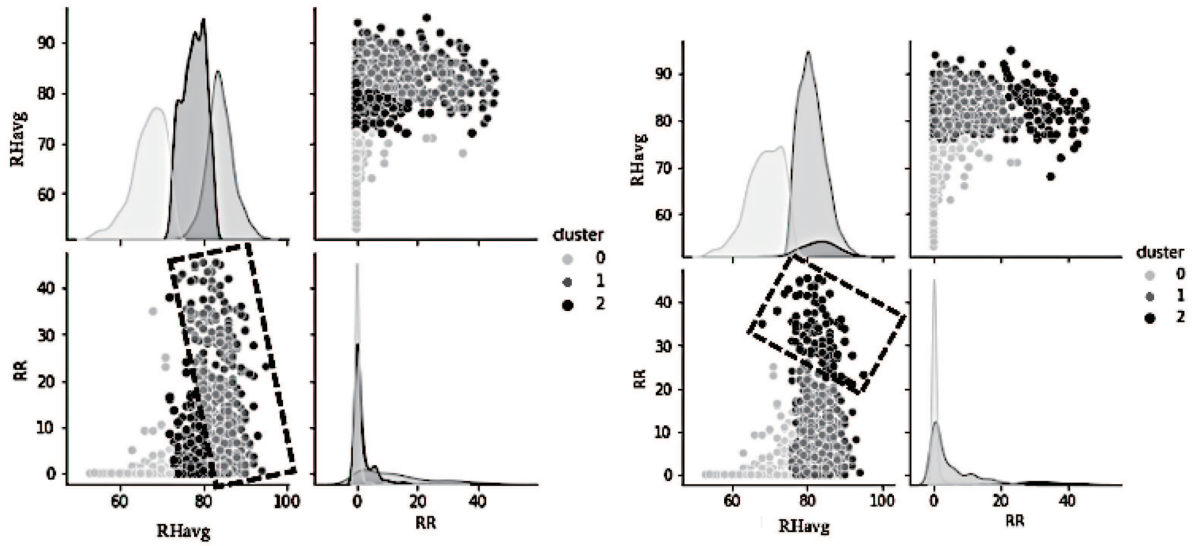

Figure 1. K-Means and AHC clustering RR-RHavg

Table 4 shows that the performance validation correlation between rainfall (RR) and average humidity (RHavg) using the Silhouette score has the highest score of 0.42. The highest Callinski-Harabasz score is 451, and the lowest value using the Davies Bouldin score is 0.89. Table 4 also explained the performance validation correlation between rainfall (RR) and average temperature (Tavg) using the Silhouette score with the highest score of 0.46, the highest Callinski-Harabasz score of 2438, and the lowest value using the Davies Bouldin score is 0.86. All of these results are generated by the K-Means algorithm, showing the superiority of the K-Means algorithm over AHC in clustering rainfall.

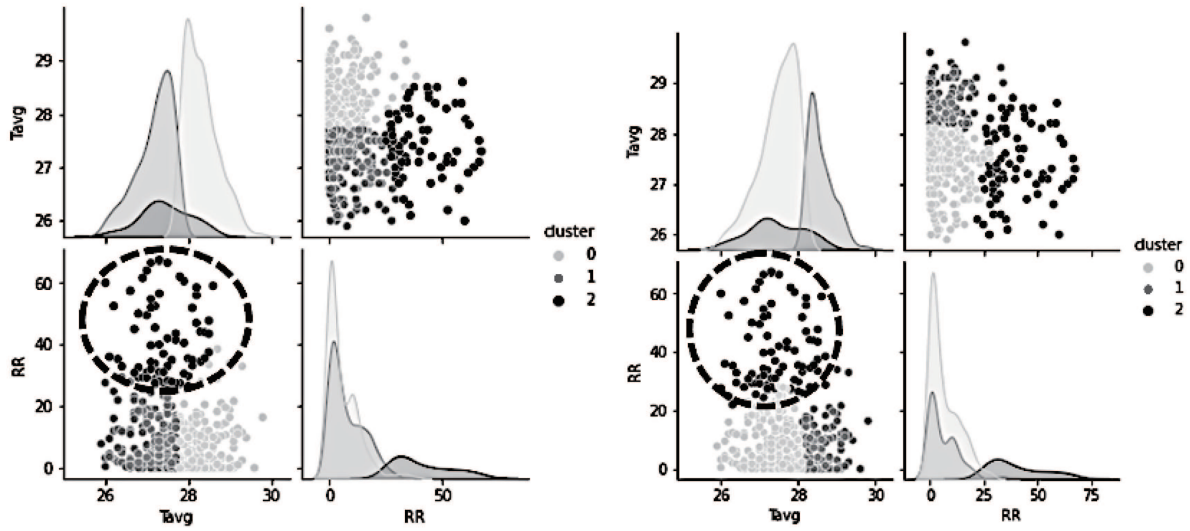

FIGURE 2. K-Means and AHC clustering RR-Tavg

TABLE 4. Result summary

	Silhoutte score		Callinski-Harabasz score		Davies Bouldin score	
Feature	**K-Means**	**AHC**	**K-Means**	**AHC**	**K-Means**	**AHC**
RR-RHavg	**0.42**	0.35	**451**	408	**0.89**	0.94
RR-Tavg	**0.46**	0.42	**2438**	2174	**0.86**	1.12

5. **Conclusion.** Based on the research that has been done, the method using K-Means and AHC provides some information about the clusters. We concluded that high rainfall occurs in Semarang City when the humidity is also high. When the humidity is low, the rain is also low. We analyzed more deeply the causes of high rainfall. As we explained before, high rainfall occurs when the humidity is above 80% (high), so we take sample data with these criteria. The analysis results show that rainfall will be high in Semarang City when the average temperature is 26 to 29 degrees. The conclusion is that the humidity is above 80%, and the temperature is between 26 to 29; there is a possibility that high-intensity rain will occur in Semarang City. It should be reiterated that we cannot generalize the causes of rain from different areas. By knowing the specific causes of precipitation in the Semarang City, later these results can be used to prepare for disaster mitigation. It is hoped that further research can use other clustering methods such as DBSCAN, mean-shift, or Gaussian mixture to analyze the problem from another point of view.

Acknowledgement. In the PMDSU program, researchers acknowledge the Directorate of Research and Community Service, Deputy for Strengthening Research and Development, Ministry of Research, Technology/National Research and Innovation Agency of the Republic of Indonesia.

REFERENCES

[1] S. D. S. Syarifuddin, A. Khurniawan, S. Aulia, D. N. Ramadan and S. Hadiyoso, Rainfall information system using geometry algorithm on IoT platform, *2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pp.195-199, DOI: 10.1109/APWiMob51111.2021.9435219, 2021.

[2] V. W. Samawi, SMCSIS: An IoT based secure multi-crop irrigation system for smart farming, *International Journal of Innovative Computing, Information and Control*, vol.17, no.4, pp.1225-1241, DOI: 10.24507/ijicic.17.04.1225, 2021.

[3] Y. Fang, X. Min, L. Zheng and D. Zhang, *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pp.628-633, DOI: 10.1109/ICSESS47205.2019.9040838, 2019.

[4] K. K. Raihana, S. M. K. Rishad, T. Sadia, S. Ahmed, M. S. Alam and R. M. Rahman, Identifying flood prone regions in Bangladesh by clustering, *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp.556-561, DOI: 10.1109/ICIS.2018.8466533, 2018.

[5] E. Biabiany, D. C. Bernard, V. Page and H. Paugam-Moisy, Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles, *Computers and Geosciences*, vol.145, DOI: 10.1016/j.cageo.2020.104612, 2020.

[6] J. Suryanto, Comparative analysis of the 15 days rainfall grouping of the Yogyakarta Special Region using fuzzy clustering and K-Means clustering, *Journal of AGRIFOR*, vol.16, 2017.

[7] M. T. Anwar, W. Hadikurniawati, E. Winarno and W. Widiyatmoko, Performance comparison of data mining techniques for rain prediction models in Indonesia, *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp.83-88, DOI: 10.1109/ISR ITI51436.2020.9315460, 2020.

[8] O. N. Pratiwi, Predicting student placement class using data mining, *Proc. of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pp.618-621, DOI: 10.1109/TALE.2013.6654511, 2013.

[9] Prihandoko, Bertalya and M. I. Ramadhan, An analysis of natural disaster data by using K-Means and K-medoids algorithm of data mining techniques, *2017 15th International Conference on Quality in Research (QiR): International Symposium on Electrical and Computer Engineering*, pp.221-225, DOI: 10.1109/QIR.2017.8168485, 2017.

[10] S. Kapil, M. Chawla and M. D. Ansari, On K-Means data clustering algorithm with genetic algorithm, *2016 4th International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp.202-206, DOI: 10.1109/PDGC.2016.7913145, 2016.

[11] X. Chu, J. Lei, X. Liu and Z. Wang, KMEANS algorithm clustering for massive AIS data based on the spark platform, *2020 5th International Conference on Control, Robotics and Cybernetics (CRC)*, pp.36-39, DOI: 10.1109/CRC51253.2020.9253451, 2020.

[12] Smarika, N. Mattas, P. Kalra and D. Mehrotra, Agglomerative hierarchical clustering technique for partitioning patent dataset, *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp.1-4, DOI: 10.1109/ICRITO. 2015.7359281, 2015.

[13] S. Zhou, Z. Xu and F. Liu, Method for determining the optimal number of clusters based on agglomerative hierarchical clustering, *IEEE Trans. Neural Networks Learn. Syst.*, vol.28, no.12, pp.3007-3017, DOI: 10.1109/TNNLS.2016.2608001, 2017.

[14] R. D. Jujjuri and M. V. Rao, Evaluation of enhanced subspace clustering validity using Silhouette coefficient internal measure, *J. Adv. Res. Dyn. Control Syst.*, 2019.

[15] A. Chaudhuri, D. Samanta and M. Sarma, Two-stage approach to feature set optimization for unsupervised dataset with heterogeneous attributes, *Expert Syst. Appl.*, DOI: 10.1016/j.eswa.2021.114563, 2021.

[16] A. K. Singh, S. Mittal, P. Malhotra and Y. V. Srivastava, Clustering evaluation by Davies-Bouldin Index (DBI) in cereal data using K-Means, *2020 4th International Conference on Computing Methodologies and Communication (ICCMC)*, pp.306-310, DOI: 10.1109/ICCMC48092.2020.ICCMC-00057, 2020.

[17] K. R. Shahapure and C. Nicholas, Cluster quality analysis using Silhouette score, *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp.747-748, DOI: 10.1109/ DSAA49011.2020.00096, 2020.

[18] T. Caliñski and J. Harabasz, A dendrite method foe cluster analysis, *Commun. Stat.*, DOI: 10.1080/ 03610927408827101, 1974.

[19] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, DOI: 10.1109/TPAMI.1979.4766909, 1979.