VARIANSI: Journal of Statistics and Its Application on Teaching and Research **ISSN 2684-7590 (Online)**
Vol. 7 No. 3 (2025), 241-255

DOI: 10.35580/variansiunm475

Implementation of K-Affinity Propagation in Provincial Grouping in Indonesia Based on Case of Environmental Pollution

Derliani Natalia D.[1], Suwardi Annas, Zulkifli Rais[3]

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, Indonesia

Keywords: K-Affinity Propagation, Environmental Pollution, Cluster Analysis.

Abstract:

Indonesia has varying levels of environmental pollution across its provinces. This study aims to provide an overview and clustering results of Indonesian provinces based on environmental pollution indicators, including water, soil, and air pollution caused by household and industrial waste. The method applied is K-Affinity Propagation (K-AP) with validation using the Davies-Bouldin Index. The analysis results indicate that the optimal number of clusters is two: Cluster 1 consists of three provinces with the highest levels of environmental pollution, while Cluster 2 comprises the remaining 35 provinces with lower pollution levels. Therefore, the government needs to pay special attention to provinces in Cluster 1 through industrial monitoring, waste management, and increasing public awareness of the importance of environmental preservation.

1. Introduction

Data mining is a concept often used to recognize hidden patterns and find relationships between parameters in large data sets (Putri & Wijayanto, 2022). One technique in data mining is clustering, which is the process of grouping data into several groups based on the similarity of their characteristics. This technique aims to solve problems in data grouping or, more precisely, to separate the dataset into subsets (Yunita, 2018). Clustering methods are divided into two methods, namely hierarchical methods and non-hierarchical methods (Annas et al., 2022).

The hierarchical method is an approach to clustering that builds a hierarchical structure from a set of data based on the similarity of object characteristics. In the hierarchical method, there are several types of clusters that are often used, namely: single linkage, complete linkage, average linkage, centroid method, ward, and median cluster (Govender & Sivakumar, 2020). Meanwhile, non-hierarchical methods are used for grouping objects, where the number of clusters to be formed can be determined in advance (Widyadhana et al., 2021). One of the developing non-hierarchical methods is K-Affinity Propagation. In this study, the researchers used K-Affinity Propagation analysis.

Previous studies that have applied K-Affinity Propagation analysis include: Research conducted by (Primantoro et al., 2024) entitled Clustering of Earth Hotspots in Forest Fire Potential Using K-Affinity Propagation, Research results by (A'yuni et al., 2023) entitled MSME Sales Clustering Based on Business Aid Distribution Priority Using K-Affinity Propagation, In addition, research by (Asriny et al., 2021) on K-Affinity Propagation Clustering Algorithm for the Classification of Part-Time Workers Using the Internet.

.

* Corresponding author.
E-mail address: derlianinatalia172@gmail.com

2. Literarure Review

2.1 Data Mining

Data mining is a data analysis technique used to recognize hidden patterns and find relationships between parameters in large data sets (Putri & Wijayanto, 2022). Data mining is a data analysis technique used to recognize hidden patterns and find relationships between parameters in large data sets (Putri & Wijayanto, 2022). Data mining is a combination of several disciplines that integrates techniques from machine learning, pattern recognition, statistics, databases, and visualization for handling problems of extracting information from large databases (Hand, 2008).

Data mining has three objectives: explanatory, which is the process of explaining the observation process; confirmatory, which is the process of confirming an existing hypothesis; and exploratory, which is the process of analyzing new data from unusual relationships (Hoffer et al., 2011). In data mining, there are several methods that are often mentioned, including clustering, classification, association rules, neural networks, genetic algorithms, and others.

2.2 Cluster Analysis

Cluster analysis is a process of grouping data into a number of clusters, where objects within a cluster have a high degree of similarity, while objects between clusters show striking differences. These similarities and differences are generally based on the attribute values possessed by each object, or can also be determined by calculating the distance between objects (Han et al., 2012). Distance measurement is a common approach used to assess the degree of similarity between objects, where two objects that are closer in distance are considered more similar than objects that are farther apart (Supranto, 2004). Euclidean distance is the square root of the sum of the differences for each variable value (Supranto, 2010). Euclidean distance is formulated as follows (Charrad et al., 2014):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Description:

$d(x,y)$: distance of data x to cluster center y
x_i : data x in observation i
y_i : center point y in observation i
n : number of observations

2.3 K-Affinity Propagation

K-Affinity Propagation (K-AP) is an extension of the Affinity Propagation (AP) method designed to produce an ideal number of exemplars. A new clustering technique called K-AP can find exemplars among all available data points, then use the data points surrounding those examples to create clusters. The ideal K value is determined by comparing the number of indices and finding the optimal K (Muhajir & Sari, 2018). The sequence for starting the K-Affinity Propagation algorithm is as follows (Frey & Dueck, 2007b; Zhang et al., 2010)

1) Calculating the similarity matrix **s**(i,k)

$$\{\boldsymbol{s}(i,k)\}_{i,j\in(1,\dots,N)i\neq k}, K \quad (2)$$

Description:

s : similarity matrix between data
i,k : data elements in rows i and k
K : number of clusters

2) Initialize the availability matrix and confidence matrix

$$\boldsymbol{a}(i,k) = 0 \quad (3)$$

$$\eta^{\boldsymbol{out}}(i) = min(s) \quad (4)$$

Description:

a(i,k) : availability matrix
$\eta^{\boldsymbol{out}}$: confidence matrix

3) Calculate responsibilities using the following equation:

$$\boldsymbol{r}(i,k) = \boldsymbol{s}(i,k) - max\eta^{out}(i) + a(i,i)\}, \quad max_{k':k' \in i,k} a(i,k') + s(i,k') \tag{5}$$

Update self-responsibility with equations:

$$\boldsymbol{r}(i,i) = \eta^{out}(i) - max_{k':k' \neq i}\boldsymbol{a}(i,k') + \boldsymbol{s}(i,k') \tag{6}$$

Description:

$\boldsymbol{r}(i,k)$: responsibility matrix
$\mathbf{a}(\mathrm{i},\mathrm{k}')$: availabity matrix

Then update the availability matrix*:*

$$r(i,k) \leftarrow \mathrm{mi\,n}\,0, r(k,k) + \textstyle\sum_{i^i;i^i \notin (i,k)} \max\{0, r(i',k)\}\} \tag{7}$$

Update self-availability with the equation:

$$a(k,k) = \textstyle\sum_{i^i;i^i \notin (i,k)} max\{0, r(i',k)\} \tag{8}$$

4) Update confidence with the equation:

$$\eta^{\boldsymbol{in}}(\boldsymbol{i}) = a(i,i) - max_{k':k' \neq i}\{a(i,k') + s(i,k')\} \tag{9}$$

$$\eta^{\boldsymbol{out}}(\boldsymbol{i}) = -R^K(\{\eta^{in}(j), j \neq i\}) \tag{10}$$

Description:

$\eta^{\boldsymbol{in}}(\boldsymbol{i}), \eta^{\boldsymbol{out}}(\boldsymbol{i})$: confidence matrix
$R^K(\{\eta^{in}(j), j \neq i\})$: the largest K value from η^{in}
K : number of clusters until convergence

5) Add up availability and responsibility:

$$\boldsymbol{c}(i,k) = argmax_j \boldsymbol{a}(i,k) + \boldsymbol{r}(i,k) \tag{11}$$

Description:

$\mathbf{c}(\mathrm{i},\mathrm{k})$: criterion matrix
$\mathbf{a}(\mathrm{i},\mathrm{k})$: availability matrix
$\mathbf{r}(\mathrm{i},\mathrm{k})$: responsibility matrix

2.4 Cluster Validation Test

Cluster validation testing is a cluster analysis process to ensure that the grouping results are in accordance with the characteristics of the data and can be used as a basis for decision making (Santosa, 2007). This study will use an internal test approach. This approach has several types of internal quality indices used to determine the optimal number of clusters. The validation method used is the Davies Bouldin Index (DBI). DBI was introduced by David L. Davies and Donald W. Bouldin in 1979. This index is used to measure cluster validation in grouping methods. DBI maximizes the distance between clusters and minimizes the distance between two and the center point of the cluster. A smaller DBI value indicates that the clusters obtained are better. This method can be calculated using the following equation (Charrad et al., 2014):

$$DBI = \frac{1}{k}\sum_{i=1}^{1} max_{i=1}\left(R_{i,j}\right) \tag{12}$$

Description:

K : number of clusters
$R_{i,j}$: ratio between i and j

The value of $R_{i,j}$ can be obtained from the following equation:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{13}$$

SSW (*Sum of Square Within Cluster*) is a cohesion matrix found in cluster i, with the following equation

$$SSW_i = \frac{1}{m_1}\sum_{j=1}^{mi} d\,(x_i, c_i) \tag{14}$$

Description :

mi : amount of data in cluster i
(xj,ci) : distance of data to centroid in cluster i

SSB (Sum of Square Between Cluster) is a separation matrix that measures the distance between clusters by measuring the distance between the centroid in one cluster and the centroid in another cluster. The following equation is used to measure the distance between cluster i and cluster j:

$$SSB_{i,j} = d(c_i, c_j) \quad (15)$$

Description:
$d(c_i, c_j)$: distance from centroid c_i to centroid c_j

2.5 Data Standardization

Data standardization is the process of transforming the scale of data on each variable so that it has a distribution form that is appropriate for the purpose of analysis. In this process, the same mathematical operation is performed on each piece of data in its original form, with the aim of keeping the differences between data relatively the same. In other words, standardization aims to equalize the scale between variables so that each variable has a balanced influence in statistical analysis, especially in cluster analysis methods. The standardization process can be done using Z-Score, by determining the mean and variance of each variable in the following equation (Han et al., 2012):

$$Z = \frac{x-\mu}{\sigma} \quad (16)$$

Description:
Z : Z-Score value, which is the standardization of the original data value
x : original data value
μ : average of all data (mean)
σ : standard deviation of all data

2.6 Environmental Pollution

The environment is usually defined as something that surrounds life or organisms. Law of the Republic of Indonesia Number 32 of 2009 concerning Environmental Protection and Management states that environmental pollution is defined as the entry or introduction of living things, substances, energy, and/or other components into the environment by human activities that exceed environmental quality standards. Environmental pollution is the introduction or inclusion of living organisms, substances, energy, and/or other components into the environment by human activities, thereby exceeding the established environmental quality standards. Environmental quality standards are the limits or levels of living organisms, substances, energy, or components that exist or should exist, and/or pollutants that are tolerated in a particular resource as part of the environment. Environmental pollution is classified into water pollution, soil pollution, and air pollution (BPS., 2024).

3. Research Metodology

3.1 Types of Research

This type of research is quantitative research that aims to analyze numerical data related to environmental pollution in Indonesia and is processed using a statistical method approach.

3.2 Data Source

The data used in this study is secondary data, which is data that is already available and obtained from official sources. The data used in this study is sourced from the 2024 Central Statistics Agency publication.

3.3 Definition of Operational Variable

The definitions of the variables used in this study are as follows:

1. The number of villages/subdistricts experiencing water pollution from household waste (X_1)
2. The number of villages/subdistricts experiencing water pollution from factory waste (X_2)
3. The number of villages/subdistricts experiencing soil pollution from household waste (X_3)
4. The number of villages/subdistricts experiencing soil pollution from factory waste (X_4)
5. The number of villages/subdistricts experiencing air pollution from household waste(X_5)
6. The Number of villages/subdistricts experiencing air pollution from factory waste (X_6)

3.4 *Data Analysis Methods*

The data analysis techniques used in this study are as follows:

1. Collecting data from the 2024 Central Statistics Agency website related to environmental pollution data based on the number of villages/subdistricts experiencing water, soil, and air pollution from household and industrial waste.
2. Performing descriptive analysis of each pollution indicator to see the characteristics of the data distribution used in the study.
3. Performing data pre-processing.
4. Performing calculations using the K-Affinity Propagation algorithm.
5. Determine the best cluster using the Davies Bouldin Index (DBI).
6. Interpret the analysis results.
7. Draw conclusions.

4. Results and Discussion

4.1. Descriptive Analysis

4.1.1 The number of villages/subdistricts experiencing water pollution from household waste (X_1)

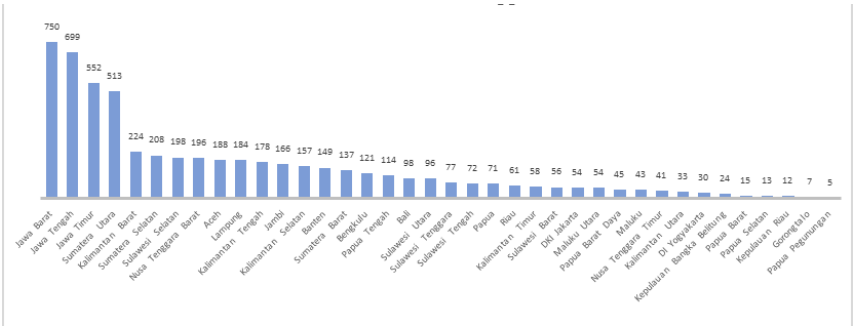

Figure 1 Bar Chart for Variables X_1

Based on Figure 1, the three provinces in Indonesia with the highest number of villages/subdistricts experiencing water pollution due to household waste are West Java with 750 villages/subdistricts, East Java with 699 villages/subdistricts, and Central Java with 552 villages/subdistricts. Meanwhile, the three provinces with the lowest number are Papua Pegunungan with 5 villages/subdistricts, West Papua with 7 villages/subdistricts, and Gorontalo with 12 villages/subdistricts.

4.1.2 The number of villages/subdistricts experiencing water pollution from factory waste (X_2)

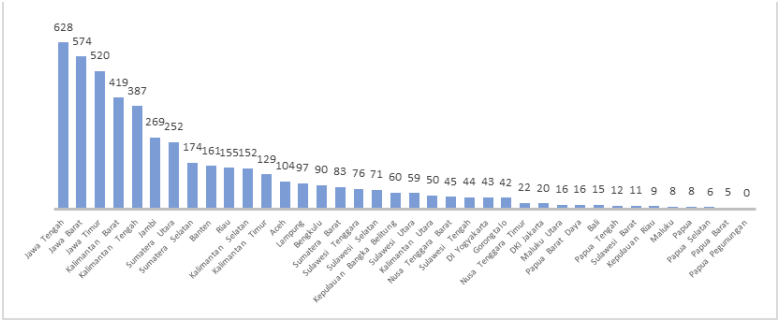

Figure 2 Bar Chart for Variables X_2

Based on Figure 2, the three provinces with the highest number of villages/subdistricts experiencing water pollution due to factory waste in Indonesia are Central Java with 628 villages/subdistricts, West Java with 574

villages/subdistricts, and East Java with 520 villages/subdistricts. Meanwhile, the three provinces with the lowest number of villages/subdistricts experiencing water pollution due to factory waste are West Papua with 6 villages/subdistricts, South Papua with 5 villages/subdistricts, and Papua Pegunungan with 0 villages/subdistricts recorded as experiencing pollution due to factory waste.

4.1.3 The number of villages/subdistricts experiencing soil contamination from household waste (X_3)

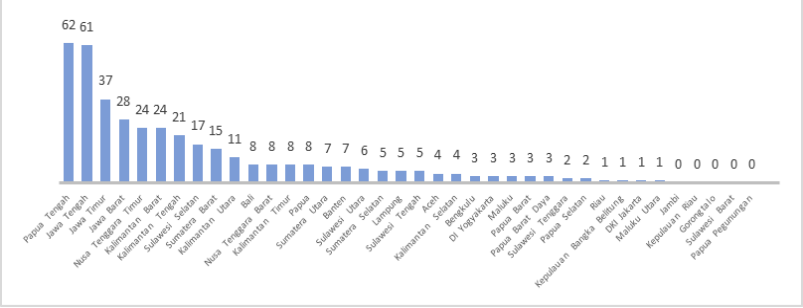

Figure 3 Bar Chart for Variables X_3

Based on Figure 3, the three provinces with the highest number of villages/subdistricts experiencing soil contamination due to household waste in Indonesia are Central Papua with 62 villages/subdistricts, Central Java with 61 villages/subdistricts, and East Java with 37 villages/subdistricts. Meanwhile, several provinces such as North Maluku, Jambi, Riau Islands, Gorontalo, West Sulawesi, and Papua Pengunungan have no villages or subdistricts recorded as experiencing soil contamination due to household waste.

4.1.4 The number of villages/subdistricts experiencing soil contamination from factory waste (X_4)

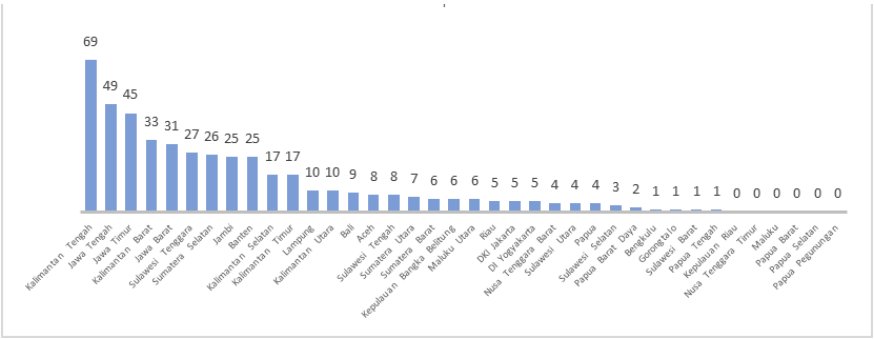

Figure 4 Bar Chart for Variables X_4

Based on Figure 4, the three provinces with the highest number of villages/subdistricts experiencing soil contamination due to factory waste in Indonesia are Central Kalimantan with 69 villages/subdistricts, Central Java with 49 villages/subdistricts, and East Java with 45 villages/subdistricts. Meanwhile, the provinces with the lowest number of villages/subdistricts experiencing soil contamination due to factory waste are Bengkulu with 1 village/subdistrict, Gorontalo with 1 village/subdistrict, West Sulawesi with 1 village/subdistrict, and Central Papua with 1 village/subdistrict. Provinces such as Riau Islands, East Nusa Tenggara, Maluku, West Papua, South Papua, and Papua Mountains are recorded as having no villages/subdistricts experiencing soil contamination due to factory waste.

4.1.5 The number of villages/subdistricts experiencing air pollution from household waste (X_5)

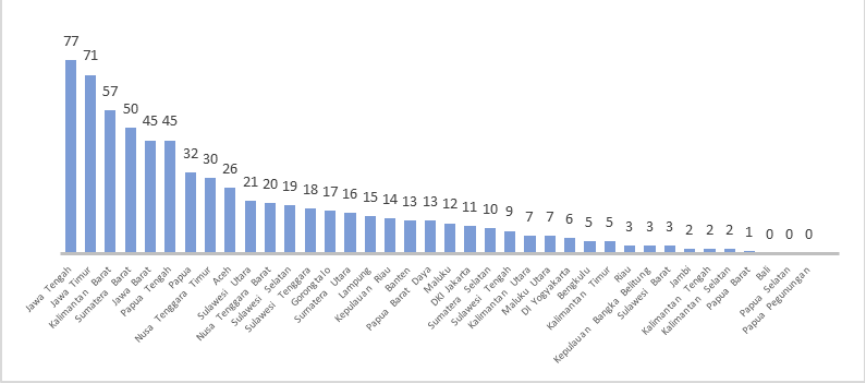

Figure 5 Bar Chart for Variables X_5

Based on Figure 5, the three provinces with the highest number of villages/subdistricts experiencing air pollution due to household waste in Indonesia are Central Java with 77 villages/subdistricts, East Java with 71 villages/subdistricts, and West Kalimantan with 57 villages/subdistricts. Meanwhile, the provinces with the lowest number of villages/subdistricts experiencing air pollution due to household waste are West Papua with 1 village/subdistrict, as well as Bali, South Papua, and Papua Pegunungan, each of which has no affected villages/subdistricts.

4.1.6 The number of villages/subdistricts experiencing air pollution from factory waste (X_6)

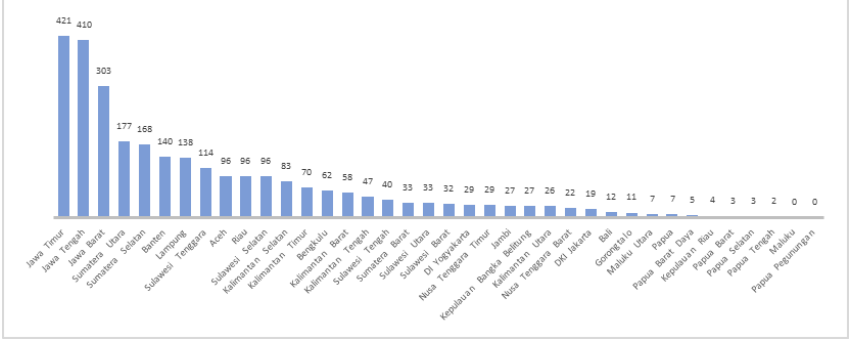

Figure 6 Bar Chart for Variables X_6

Based on Figure 6, the three provinces with the highest number of villages/subdistricts experiencing air pollution due to factory waste in Indonesia are East Java with 421 villages/subdistricts, Central Java with 410 villages/subdistricts, and West Java with 303 villages/subdistricts. Meanwhile, the provinces with the lowest number of villages/subdistricts experiencing air pollution due to household waste are Central Papua with 2 villages/subdistricts, South Papua with 3 villages/subdistricts, and West Papua with 3 villages/subdistricts. In fact, Maluku and Papua Pegunungan are recorded as having no villages/subdistricts experiencing air pollution due to factory waste.

4.2 Clustering K-Affinity Propagation

The K-Affinity Propagation method in this study begins with calculating the similarity level between observations that have undergone standardization. These similarity values are then used to calculate responsibility values as a measure of the suitability of a data point for a candidate exemplar. The next step is to calculate availability values, which indicate the availability of a data point to be selected as an exemplar. The process of updating responsibility and availability values is carried out repeatedly until a stable or convergent condition is reached, indicated by no change in availability values in subsequent iterations. Clustering is determined based on the sum of the responsibility and

availability values. The results of clustering using the K-Affinity Propagation method in the case of environmental pollution were obtained by forming 2 to 6 clusters. The number of clusters was determined through a calculation process using R software. Based on the results of data processing, the following cluster grouping results were obtained:

Table 1 K-AP Cluster Results Using 2 Clusters

Cluster	Eksemplar	Members	Total
1	East Java	West Java, Central Java, East Java.	3
2	Central Sulawesi	Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, East Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Central Papua, Papua Pengunungan.	35

Based on the results of the K-Affinity Propagation analysis using 2 clusters shown in Table 1, there are 3 members in cluster 1, namely West Java, Central Java, and East Java. Meanwhile, there are 35 members in cluster 2, consisting of Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, East Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Central Papua, and Papua Pegunungan.

Table 2 K-AP Cluster Results Using 3 Clusters

Cluster	Eksemplar	Members	Total
1	East Java	West Java, Central Java, East Java	3
2	Banten	North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, Southeast Sulawesi	9
3	Central Sulawesi	Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Yogyakarta Special Region, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Central Papua, Papua Pegunungan	26

Table 2 shows the results of the K-Affinity Propagation analysis using 3 clusters. Cluster 1 has 3 members, namely West Java, Central Java, and East Java. Meanwhile, cluster 2 has 9 members, namely North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, and Southeast Sulawesi. Cluster 3 has 26 members, namely Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Central Papua, and Papua Pegunungan.

Table 3 K-AP Cluster Results Using 4 Clusters

Cluster	Eksemplar	Anggota	Jumlah
1	East Java	West Java, Central Java, East Java	3
2	Banten	North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, Southeast Sulawesi	9
3	Central Sulawesi	Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Yogyakarta Special Region, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Papua Pegunungan	25
4	Central Papua	Central Papua	1

Table 3 shows the results of the K-Affinity Propagation analysis using 4 clusters. Cluster 1 has 3 members, including West Java, Central Java, and East Java. Cluster 2 has 9 members, including North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, and Southeast Sulawesi. Cluster 3 has 25 members, including Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, and Papua Pegunungan. Cluster 4 has 1 member, namely Central Papua.

Table 4 K-AP Cluster Results Using 5 Clusters

Cluster	Eksemplar	Members	Total
1	East Java	West Java, Central Java, East Java	3
2	Banten	North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, Southeast Sulawesi	8
3	Central Sulawesi	Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Yogyakarta Special Region, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Papua Pegunungan	25
4	Central Kalimantan	Central Kalimantan	1
5	Central Papua	Central Papua	1

Table 4 shows the results of the K-Affinity Propagation analysis using 5 clusters. Cluster 1 has 3 members, including West Java, Central Java, and East Java. Cluster 2 has 8 members, including North Sumatra, Jambi, South Sumatra, Lampung, Banten, West Kalimantan, South Kalimantan, and Southeast Sulawesi. Cluster 3 has 25 members, including Aceh, West Sumatra, Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, and Papua Pegunungan. Cluster 4 has 1 member, namely Central Kalimantan. Cluster 5 has 1 member, namely Central Papua.

Table 5 K-AP Cluster Results Using 6 Clusters

Cluster	Eksemplar	Members	Total
1	Lampung	Aceh, North Sumatra, West Sumatra, Jambi, South Sumatra, Lampung, Banten, South Kalimantan, South Sulawesi, Southeast Sulawesi	10
2	DI Yogyakarta	Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, Yogyakarta Special Region, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, South Papua, Papua Pegunungan	22
3	East Java	West Java, Central Java, East Java	3
4	Kalimantan Barat	Kalimantan Barat	1
5	Central Kalimantan	Central Kalimantan	1
6	Central Papua	Central Papua	1

Table 5 shows the results of the K-Affinity Propagation analysis using 5 clusters. Cluster 1 has 10 members, including Aceh, North Sumatra, West Sumatra, Jambi, South Sumatra, Lampung, Banten, South Kalimantan, South Sulawesi, and Southeast Sulawesi. Cluster 2 has 22 members, including Riau, Bengkulu, Bangka Belitung Islands, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, NTB, NTT, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua, and South Papua. Cluster 3 has 3 members, including West Java, Central Java, and East Java. Cluster 4 has 1 member, namely West Kalimantan. Cluster 5 has 1 member, namely Central Kalimantan. Cluster 6 has 1 member, namely Central Papua.

4.3 Evaluation of The Best Cluster

The Davies Bouldin Index validation test can be used to determine the best cluster. The following table shows the DBI results:

Tabel 6 Hasil Uji Validasi *Cluster*

Number of Clusters	Davies *Bouldin Index*
2	**0,4226676**
3	1,021048
4	0,9812245
5	0,8318202
6	0,6729084

Table 6 shows that the cluster validation test results using DBI with a number of clusters ranging from 2 to 6 clusters obtained the smallest DBI value of 0.4226676. Therefore, the best cluster is 2 clusters.

4.4 Cluster Interpretation

From the validation results, the best cluster obtained was 2 clusters. The description of each variable based on the cluster can be seen in the following figure:

4.4.1 The number of villages/subdistricts experiencing water pollution from household waste (X_1)

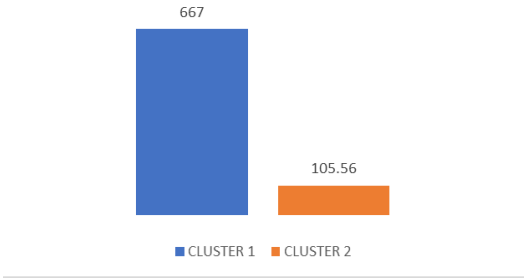

Figure 7 Average of variable X_1

Figure 7 shows that the highest average X_1 is located in Cluster 1 at 667, while in Cluster 2 it is 105.56. The high average value in Cluster 1 is due to the high population, which results in a larger volume of household waste. Conversely, the low average in Cluster 2 is influenced by the low population density and the presence of areas with natural environments.

4.4.2 The number of villages/subdistricts experiencing water pollution from factory waste (X_2)

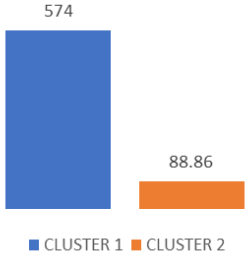

Figure 8 Average of variable X_2

Figure 8 shows that the average variable X_2 in Cluster 1 is 574, which is higher than Cluster 2 at 88.86. This indicates that provinces in Cluster 1 have greater industrial activity, resulting in factory waste that contributes to water pollution. The low average in Cluster 2 is due to the dominance of agricultural areas with a relatively low number of industrial areas.

4.4.3 The number of villages/subdistricts experiencing soil pollution from household waste (X_3)

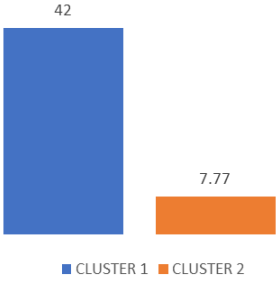

Figure 9 Average of variable X_3

Figure 9 shows that the average of variable X_3 in Cluster 1 is 42, higher than Cluster 2, which is only 7.77. This condition reflects the limited household waste management system in Cluster 1, resulting in frequent soil pollution. Conversely, the pollution level in Cluster 2 is relatively lower due to the smaller volume of household waste.

4.4.4 The number of villages/subdistricts experiencing soil pollution from factory waste (X_4)

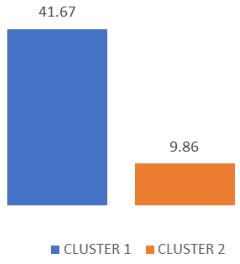

Figure 10 Average of variable X_4

Figure 10 shows that the average of variable X_4 in Cluster 1 is 41.67, while in Cluster 2 it is 9.86. This indicates that provinces in Cluster 1 have more industrial activities that produce solid waste and chemicals that have the potential to contaminate the soil. Meanwhile, the low value in Cluster 2 is influenced by limited industrial activity.

4.4.5 The number of villages/subdistricts experiencing air pollution from household waste (X_5)

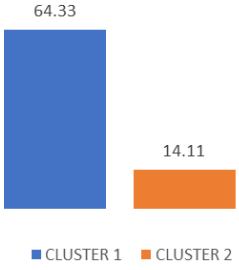

Figure 11 Average of variable X_5

Figure 11 shows that the average of variable X_5 in Cluster 1 is 64.33, while in Cluster 2 it is only 14.11. This indicates that air pollution caused by household activities is more prevalent in provinces in Cluster 1. This condition is related to the high population and the continued use of waste incineration and fuel methods that are not environmentally friendly. Meanwhile, provinces in Cluster 2 have lower levels of air pollution due to relatively low household activity.

4.4.6 The Number of villages/subdistricts experiencing air pollution from factory waste (X_6)

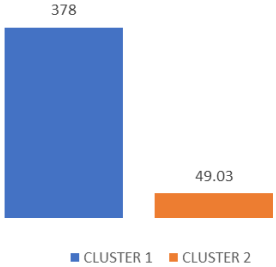

Figure 12 Average of variable X_6

Figure 12 shows that the average of variable X_6 in Cluster 1 reached 378, while Cluster 2 only reached 49.03. This difference indicates that provinces in Cluster 1 have more industrial areas that produce air pollutant emissions. Meanwhile, provinces in Cluster 2 have low levels of air pollution due to minimal industrial activity.

Based on the results obtained, cluster 2, which is at the lowest position, is cluster 2, and cluster 1, which is at the highest position, can be seen in Figure 13:

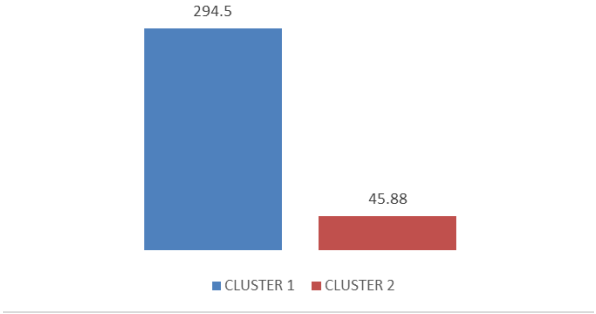

Figure 13 Average for each cluster

` Figure 13 shows that the overall average variable for Cluster 1 is 294.50, which is higher than Cluster 2 at 45.88. This indicates that the areas included in Cluster 1 have a relatively higher level of environmental pollution compared to Cluster 2. The high average in Cluster 1 is due to the dominance of provinces with high industrial activity and population density, which produce various types of waste that impact water, soil, and air pollution. Conversely, the low average in Cluster 2 indicates that the provinces in this cluster generally have lower industrial activity and relatively good environmental conditions, resulting in lower pollution levels.

The map of the grouping results using K-Affinity Propagation analysis is as follows

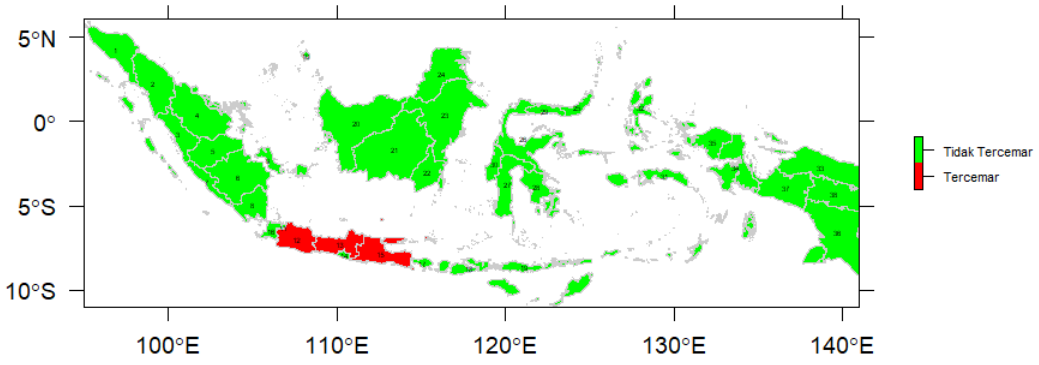

Figure 14 Map of Indonesia Showing the Distribution of Environmental Pollution Cases

Figure 14 shows a map of the results of K-Affinity Propagation cluster analysis using data on environmental pollution cases in Indonesia in 2024 covering 38 provinces. There are two clusters formed and distinguished by different colors. Clusters with relatively high pollution levels (polluted) are shown in red, while clusters with lower pollution

levels (unpolluted) are shown in green. These clustering results indicate differences in the characteristics of environmental pollution levels between provinces in Indonesia based on the indicators used in this study.

5.1 Conclusion

Based on the results of research on environmental pollution in various provinces in Indonesia using the K-Affinity Propagation method, several conclusions can be drawn as follows:

1. The pattern of environmental pollution in Indonesia shows variations in pollution levels between provinces. These differences are influenced by the socioeconomic characteristics and industrial activities in each province. Provinces with high population density and a predominance of industrial activities, such as West Java, Central Java, and East Java, tend to have higher levels of water, soil, and air pollution. Conversely, provinces with lower industrial activity and more natural geographical conditions tend to have lower levels of pollution. This shows that environmental pollution in Indonesia is uneven and related to the level of regional development.
2. The results of grouping provinces in Indonesia based on the level of environmental pollution using the K-Affinity Propagation method produced two clusters. The determination of the best number of clusters was based on the results of a validation test using the Davies-Bouldin Index (DBI), which yielded an optimum value of 0.4226676, so that number was chosen as the most optimal grouping structure. Cluster 1 consists of three provinces, namely West Java, Central Java, and East Java, which have high levels of environmental pollution. Meanwhile, Cluster 2 covers 35 other provinces that are classified as having lower levels of environmental pollution. These results provide important information for the government to prioritize environmental pollution control efforts in provinces included in Cluster 1, particularly through industrial emission monitoring, improving waste management quality, and implementing sustainable environmental policies to prevent further environmental damage in the future.

References

Annas, S., Irwan, I., Safei, R. H., & Rais, Z. (2022). K-Prototypes Algorithm for *Cluster*ing The Tectonic Earthquake in Sulawesi Island. *Jurnal Varian*, *5*(2). https://doi.org/10.30812/varian.v5i2.1908

Asriny, N. I., Muhajir, M., & Andrian, D. (2021). *K-Affinity Propagation cluster*ing algorithm for the classification of part-time workers using the internet. *Indonesian Journal of Electrical Engineering and Computer Science*, *24*(1). https://doi.org/10.11591/ijeecs.v24.i1.pp464-472

A'yuni, T. Q., Febriati, B. N., Effendie, L. I., Muhajir, M., & Yotenka, R. (2023). *MSME sales clustering based on business aid distribution priority using K-affinity propagation*. *Enthusiastic: International Journal of Applied Statistics and Data Science*, *3*(1), 111–124. https://doi.org/10.20885/enthusiastic.vol3.iss1.art10

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of *cluster*s in a data set. *Journal of Statistical Software*, *61*(6). https://doi.org/10.18637/jss.v061.i06

Frey, B. J., & Dueck, D. (2007a). *Cluster*ing by Passing Messages Between Data Points. *Science*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Frey, B. J., & Dueck, D. (2007b). *Cluster*ing by passing messages between data points. *Science*, *315*(5814). https://doi.org/10.1126/science.1136800

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

Hand, D. J. (2008). Data Mining: Methods and Models by D. T. Larose. *Biometrics*, *64*(1). https://doi.org/10.1111/j.1541-0420.2008.00962_9.x

Hoffer, J. A., Ramesh, V., & Topi, H. (2011). *Modern database management* (10th ed.). Pearson.

Jia Muhaji wei Han, M. K. and J. P. (2012). *Data Mining: Concepts and Techniques, Third Edition - Books24x7*. Morgan Kaufmann Publishers.

Muhajir, M., & Sari, N. N. (2018). *K-Affinity Propagation* (K-AP) and K-Means *Cluster*ing for Classification of Earthquakes in Indonesia. *Proceeding - 2018 International Symposium on Advanced Intelligent Informatics: Revolutionize Intelligent Informatics Spectrum for Humanity, SAIN 2018*. https://doi.org/10.1109/SAIN.2018.8673344

Primantoro, S., Goejantoro, R., & Prangga, D. S. (2024). *Clustering Titik Panas Bumi Pada Potensi Kebakaran Hutan Menggunakan K-Affinity Propagation Geothermal Hotspot Clustering on Forest Fire Potential Using K-Affinity Propagation*. *15*(2). https://doi.org/10.30872/eksponensial.v15i2.1299

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, *11*(1). https://doi.org/10.34010/komputika.v11i1.4350

Tarisya Qurrota A'yuni, Baiq Nina Febriati, Lazuardy Ilham Effendie, Muhammad Muhajir, & Yotenka, R. (2023). MSME Sales *Cluster*ing Based on Business Aid Distribution Priority Using K-Affinity Propagation. *Enthusiastic : International Journal of Applied Statistics and Data Science*. https://doi.org/10.20885/enthusiastic.vol3.iss1.art10

Undang-Undang Republik Indonesia Nomor 32 Tahun 2009 tentang Perlindungan dan Pengelolaan Lingkungan Hidup. (3 Oktober 2009). Jakarta: Sekretariat Negara..

Widyadhana, D., Hastuti, R. B., Kharisudin, I., & Fauzi, F. (2021). Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 584–594. https://journal.unnes.ac.id/sju/index.php/prisma/

Zhang, X., Wang, W., Nørvåg, K., & Sebag, M. (2010). K-AP: Generating specified K *cluster*s by efficient Affinity Propagation. *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2010.107

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (Edisi ke-3). Elsevier.

Santosa, B. (2007). *Data mining: Teknik pemanfaatan data untuk keperluan bisnis*. Graha Ilmu.