EKSAKTA
Journal of Sciences and Data Analysis
E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol7.iss1.art7

Clustering of Oil and Gas Pipeline Sectors with Block-Based K-Medoids Method

Dimas Zahran Wicaksana[1], Kariyam[2*], Tri Suryanto[3]

[1,2] Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Jl. Kaliurang KM 14.5, Sleman, Special Region of Yogyakarta, 55584, Indonesia
[3] Analyst, PT Pertamina Hulu Rokan Duri, Pertamina Complex, Duri, Bengkalis Regency, Riau 28784, Indonesia

[*] *Corresponding author: kariyam@uii.ac.id*

Abstract: Effective oil and gas pipeline management requires a data-driven approach to identify segments with varying risk characteristics. This study aims to classify pipeline segments based on protective infrastructure conditions using the Block-Based K-Medoids clustering method. The analysis considers six variables: Pipeline Burial, Pipe Along Road, Pipe Guards, Berm/Rail/Guard Condition, Public Road, and ROW HCA along a 59-kilometer pipeline corridor. Data were normalized, and the optimal number of clusters was determined using the Deviation Ratio Index based on Medoid (DRIM), which indicated three clusters as the most representative structure. The results demonstrate clear differentiation among segments in terms of exposure level, protective condition, and HCA involvement, enabling classification into low-, moderate-, and high-risk groups. Spatial visualization further confirms systematic risk distribution along the route. These findings provide a structured basis for prioritizing inspection, maintenance, and mitigation strategies in pipeline infrastructure management.

Keywords: Pipeline, Pipeline protection, Clustering, Block-based K-Medoids, Deviation Ratio Index

Introduction

Pipelines play a crucial role in transporting energy resources, such as oil and gas, from production facilities to processing plants or export terminals [1], This infrastructure enables the efficient and safe distribution of energy from the source to end users, thereby supporting economic stability and regional energy resilience [2]. Despite being designed for long-term durability, pipelines remain vulnerable to various internal and external factors that may cause damage, disrupt operations, and pose serious risks to the surrounding environment and communities [3].

Pipeline routes often extend beyond industrial or company-controlled areas and pass through residential zones and public environments. This condition increases the likelihood of interaction between pipeline infrastructure and community activities, which can trigger accidents with severe consequences [4], Pipeline failures may result in leaks, explosions, or environmental contamination, causing fire hazards, public health disturbances, and substantial economic losses for pipeline operators [5]. Therefore, pipeline risk is not limited to operational concerns but also represents a significant threat to public safety and environmental sustainability. To reduce the probability and impact of such incidents, systematic monitoring and maintenance strategies are essential components of pipeline

infrastructure management [6]. Various analytical approaches have been developed to support risk mitigation, including failure analysis and risk-based maintenance models [7]. However, most existing studies focus primarily on identifying causes of failure rather than classifying pipeline segments based on infrastructure characteristics that influence exposure and vulnerability. This indicates a need for data-driven segmentation methods capable of identifying groups of pipeline sectors with similar protection and environmental conditions.

In this study, the *Block-based K-Medoids* method is used to group pipeline segments based on the condition of the protection infrastructure, considering variables such as pipe burial depth (*Pipeline Burial*), pipe along the road (*Pipe Along Road*), the presence of road guards (*Pipe Guards*), protective conditions (*Berm/Rail/Guard Condition*), exposure to public roads (Public Road), and potential impacts (*ROW HCA*). Pipelines are categorized into groups of 1 km to obtain a more structured mapping of their characteristics. The Block-Based K-Medoids method was selected because it provides a more stable and structured initialization process compared to classical K-Medoids and is more robust to outliers than partitioning methods such as K-Means. In addition, the medoid-based approach ensures that cluster centers correspond to actual observations, improving interpretability for infrastructure decision-making. Several previous studies have discussed pipeline risk analysis using different approaches. For example, a study by Zemanova et al. (2023) [8] on risk management in the water industri and Stania et al. (2024) [9] on SPAM pipeline risk management employed the Fault Tree Analysis method to identify the causes of leaks in drinking water supply pipes. Additionally, a study by Dewi Fatmawati (2020) [10] examined pipeline leaks in Balikpapan Bay and Noussia (2011) [11] Discuss the responsibility for environmental pollution due to oil spills and other legal consequences. These studies show the importance of comprehensive risk analysis in pipeline management.

Through this study, the researcher aims not only to group pipeline segments based on physical and environmental characteristics using the Block-Based K-Medoids method [12], but also to determine the optimal number of clusters using validation indices and to interpret cluster characteristics in relation to infrastructure exposure and risk conditions. The study seeks to provide a structured, data-driven framework that supports risk-based inspection prioritization and maintenance planning. By identifying homogeneous groups of pipeline sectors, the results contribute to more effective and efficient pipeline management and offer practical decision-support insights for infrastructure risk mitigation in indonesia.

Materials and Methods

Data and data source

The data used in this study consist of secondary data obtained from field measurements and the recording system of the PT.X oil and gas pipeline from Area A to Area B. The original observations were recorded at 10-meter intervals along a 59-kilometer pipeline and then aggregated into 1-kilometer segments to provide a more structured spatial representation of infrastructure conditions. As a result, the final dataset consists of 59 pipeline segments (n = 59), each summarizing exposure and protection characteristics related to Pipe Hit Incident (PHI) risk.

The study employs six technical and environmental variables, as presented in Table 1, including Pipeline Burial, Pipe Along Road, Pipe Guard, Berm/Rail/Guard Condition, Public Road, and ROW HCA. All variables are expressed as numerical percentage-based indicators reflecting infrastructure protection level and environmental exposure. The analytical procedures include data preprocessing, descriptive analysis, outlier detection, multicollinearity testing, cluster determination, clustering using the Block-Based K-Medoids method, and cluster profiling. The analysis was conducted using R statistical software to ensure methodological reproducibility.

Table 1. Definition of Research Variables

Variables	Information	Definition Operational Variables
X_1	Pipeline Burial	The depth of the pipe in the ground.
X_2	Pipe Along Road	Whether the pipe is along the road or not.
X_3	Pipe Guard	Level of protection against
X_4	Berm/Rail/Guard Condition	The physical condition of existing road barriers or barriers.
X_5	Public Road	pipeline under/adjacent to a public road
X_6	ROW HCA	potential impact If happen incident.

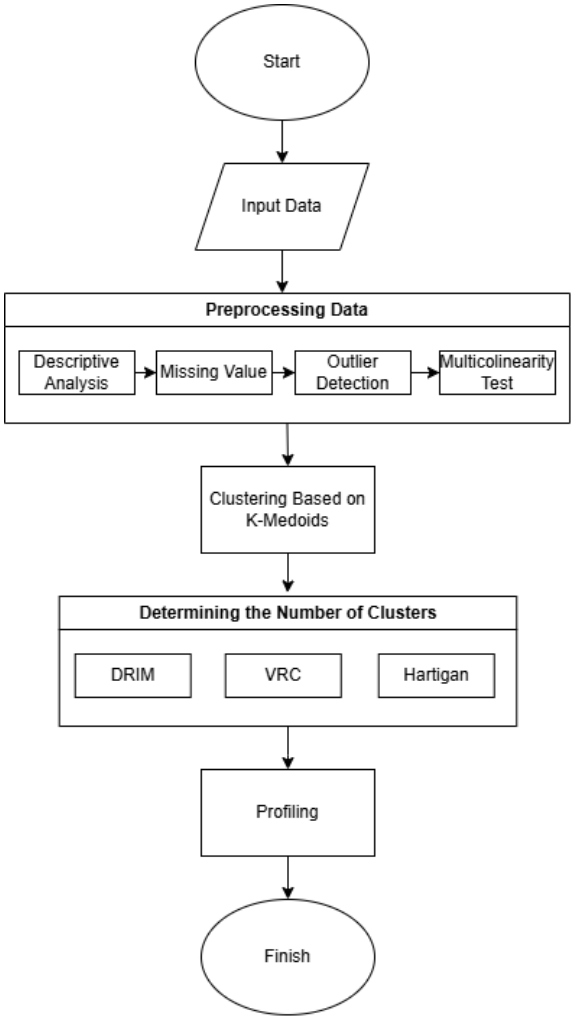

Figure 1. Data Analysis Flowchart

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol7.iss1.art7

The analysis data flow is as shown in Figure 1. with an outline of the precess stages as follows:

(i) Inputting data obtained from PT.X into the Rstudio software.
(ii) Performing data preprocessing, which includes descriptive analysis, handling missing values, outlier detection using univariate analysis, and conducting a multicollinearity test to determine the linear relationships among independent variablels.
(iii) Performing pipeline *clustering*
(iv) Determine the number of clusters using the Deviation Ratio Index based on Medoids (DRIM), Variance Ratio Criterion (VRC), and Hartigan index.
(v) Profiling of cluster results using Quantum Geographic Information System (QGIS) software.

Data Transformation

Data preprocessing is a critical phase in the analytical process to ensure that all variables contribute proportionally to the clustering results. Although the variables in this study are expressed in percentage-based values, differences in their ranges may still influence distance-based clustering calculations [12]. Therefore, data rescaling was applied prior to clustering to prevent variables with relatively larger ranges from dominating the dissimilarity measurement.

The Min–Max normalization method was employed to transform each variable into a standardized interval between 0 and 1 without altering the original distribution characteristics [13]. This transformation ensures comparability across variables and improves the stability of the clustering process. The Min–Max normalization is defined as follows [12].

$$x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \quad (1)$$

where x_{ij} denotes the original value of variable j for object i, $\min(x_j)$ and $\max(x_j)$ represent the minimum and maximum values of variable j across all objects, respectively, and x'_{ij} is the normalized value of variable j for object i after Min–Max transformation.

Multicolinearity Test

Multicollinearity refers to a situation in multiple linear regression analysis where two or more independent variables exhibit a strong correlation with each other. This condition may lead to instability in the estimation of regression coefficients and reduce the accuracy of model interpretation [14]. One of the most common approaches to identify multicollinearity is by computing the Variance Inflation Factor (VIF), as shown in Equation (2). A high VIF value suggests the presence of substantial multicollinearity, whereas a low VIF value indicates that multicollinearity is not a significant concern within the regression model.

$$VIF_j = \frac{1}{1 - R_j^2} \quad (2)$$

Where VIF_j is the Variance Inflation Factor of the j-th independent variable, and R_j^2 denotes the coefficient of determination obtained from regressing the j-th independent variable on all other independent variables in the model.

Block-Based K-Medoids Partitioning Method

The Block-KM algorithm was proposed to address the shortcomings of the Simple and Fast K-Medoids (SFKM) and Simple K-Medoids (SKM) methods, especially those related to the random selection of initial medoids and the possibility of forming empty clusters. Block-KM incorporates a more structured approach for initializing medoids by utilizing the statistical characteristics of the data. This enhancement contributes to greater stability and efficiency in the overall clustering process [12].

1. Determination of initial medoid
   a. For each object $i(i = 1, 2, \ldots, n)$, two parameters are calculated based on the standard deviation as defined in Equation (3), and the sum value as defined in Equation (4).

$$u_i = \sqrt{\frac{\sum_{l=1}^{p}(x_{il} - \bar{x}_l)^2}{p - 1}} \quad (3)$$

   Where u_i denotes the standard deviation of object i, x_{il} represents the value of the l-th variable for object i, $\bar{x}_i$ is the mean value of all variables for object i, and p denotes the total number of variables used in the clustering process.

   The mean value $\bar{x}_i$ is defined as:

$$\bar{x}_i = \frac{w_i}{p} \quad (4)$$

   where w_i is the total sum of all variables for object i, computed as:

$$w_i = \sum_{l=1}^{p} x_{il} \quad (5)$$

   With $i = 1, 2, \ldots, n$ and $l = 1, 2, .., p$. These parameters are used as references to select the initial medoid.

   b. Arrange all objects in ascending order according to Equation (3). For blocks that have the same standard deviation (if any), reorder the objects within those blocks in ascending order based on Equation (4).
   c. For the first k blocks of the combination u_i and w_i (or just u_i), select the first object from each block as the initial medoid.
   d. Assign objects to their respective initial groups according to the shortest distance between each object and the nearest medoid.

2. Get Data Partition
   a. Update the medoids in each group formed based on the objects that minimize the average distance to other group members. The average distance calculation is calculated using Equation (5) $\left(j = r = 1, 2, \ldots, n_g \ \ and \ g = 1, 2, \ldots, k\right)$:

$$\bar{D}_j = \frac{1}{n_g} \sum_{r=1}^{n_g} d(x_i, x_r) \tag{6}$$

   where $\bar{D}_j$ denotes the average distance between object i and all other objects within group j, n_g is the number of objects in group j, $d(x_i, x_r)$ represents the distance between object i and object r, and x_r refers to the r-th object belonging to group j.
   b. Determine clusters by assigning each object to the nearest *medoid* and calculate the sum of distance within group, $SDW(k)$), as in Equation (6), where m_i is the group medoid containing object x_i.

$$SDW(k) = \sum_{i=1}^{n} d(x_i, m_i) \tag{7}$$

   where $SDW(k)$ denotes the total within-cluster distance for k clusters, n is the total number of objects, $d(x_i, m_i)$ represents the distance between object i and the medoid of the cluster to which it belongs, and m_i denotes the medoid associated with object i.
   c. Repeat steps a and b until the value $SDW(k)$ is the same as one of the previous steps, or a predetermined number of iterations is reached, or the number of *medoids* does not change.

Deviation Ratio Index Based on Medoid

The Deviation Ratio Index based on Medoids (DRIM) is a cluster evaluation technique used to determine the optimal number of clusters within a dataset. This method analyzes the deviation ratio across different cluster numbers (k) to identify the most suitable and representative clustering structure [15]. Based on the final medoids obtained from the clustering process for a given number of clusters k, the Deviation Ratio (DR) is defined as follows:

$$DR(k) = \frac{SDW(k)/(n-k)}{SDB(k)/(k-1)} \tag{8}$$

where $SDB(k)$ represents the total distance from each object to the medoids of the other clusters. The $SDB(k)$ can be expressed as:

$$SDB(k) = \sum_{i=1}^{n} \sum_{k=1}^{g-1} d(x_i, m_k) \tag{9}$$

Where m_k denotes the medoid of the k cluster, and $d(x_i, m_k)$ refers to the distance between the i object and the medoid of the k cluster. The deviation ratio index is then calculated to compare the cluster structure between the n- $k$and n-th number of groups $k+1$, which is formulated as follows:

$$DRI(k) = \frac{DR(k)}{DR(k+1)} \tag{10}$$

Where $DRI(k)$denotes the Deviation Ratio Index for $k$clusters, and $DR(k)$and $DR(k+1)$represent the deviation ratios for $k$and $k+1$clusters, respectively. The selection of the optimal number of clusters is determined based on the first smallest k when the value $DRI(k) < 1$, which indicates that the cluster structure of k is better than $k+1$. This process is carried out by trying the value of k from a minimum of two clusters (k = 2), and continues to increase until the value $DR(k)$obtained is greater than $DR(k+1)$or the value $DR(k) < DR(k+1)$.

Hartigan

The Hartigan Index is one of the internal evaluation methods used in cluster analysis to determine the optimal number of clusters within a dataset [16]. The index evaluates the relative improvement in within-cluster dispersion when increasing the number of clusters from $k$to $k+1$. The Hartigan index is formulated as shown in Equation (10) [15].

$$H(k) = \left(\frac{WSS(k)}{WSS(k+1)} - 1\right) \times (n-k-1) \tag{11}$$

In this study, $WSS(k)$is replaced with $SDW(k)$, which represents the total within-cluster distance as defined in Equation (7). The optimal number of clusters is generally indicated when the value of $H(k)$ falls below a certain threshold or shows a substantial decrease, indicating diminishing improvement when increasing the number of clusters.

Variance Ratio Citeration

The Variance Ratio Criterion (VRC), also known as the Calinski–Harabasz Index, is an internal cluster validation method used to assess clustering quality by comparing between-cluster dispersion and within-cluster dispersion [16]. The index is defined as shown in Equation (12):

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{12}$$

where $CH(k)$denotes the Calinski–Harabasz index for $k$clusters, $B(k)$represents the between-cluster variance, $W(k)$is the within-cluster variance, $n$is the total number of objects, and k denotes the number of clusters.

In this study, $W(k)$is replaced with $SDW(k)$, representing the total within-cluster distance, and $B(k)$is replaced with $SDB(k)$, representing the total between-cluster distance, to maintain consistency with the distance-based formulation used in the Block-Based K-Medoids method. A higher value of $CH(k)$indicates better cluster separation and compactness. Therefore, the optimal number of clusters is determined as the value of $k$that maximizes $CH(k)$.

Result and Discussion

Descriptive Analysis

The dataset consists of 59 observations representing the condition of oil and gas pipeline infrastructure based on six indicators. Descriptive statistics, including minimum, maximum, and mean values, were computed using R software, and the results are presented in Table 2. All variables are expressed in percentage form, ranging approximately from 0 to 100, indicating proportional exposure or condition levels for each pipeline segment.

Table 2. Descriptive Analysis

Variables		N	Min	Mean	Max
X_1	Pipeline Burial	59	0.00	69.14	100.00
X_2	Pipe Along Road	59	0.00	94.10	100.00
X_3	Pipe Guard	59	0.00	12.71	81.00
X_4	Berm/Rail/Guard Condition	59	0.00	57.76	100.00
X_5	Public Road	59	0.00	60.08	101.00
X_6	ROW HCA	59	0.00	19.78	100.00

Table 2 indicates substantial variation across variables, reflecting meaningful differences in pipeline risk conditions. The mean value of Pipeline Burial (69.14%) shows that a large proportion of segments are located above ground, increasing exposure to external damage. Pipe Along Road has the highest mean (94.10%), indicating that most pipelines are situated along public roads, which heightens third-party interference risk. Meanwhile, Berm/Rail/Guard Condition (57.76%) suggests that more than half of protective structures are in poor condition. Public Road (60.08%) further reflects considerable public exposure, while ROW HCA (19.78%) identifies segments located in high-consequence areas. Overall, these variations confirm that the dataset contains sufficient heterogeneity to justify clustering using the Block-Based K-Medoids method.

Outlier Detection

Outlier detection is performed to identify extreme values that can significantly affect the analysis results. In this study, outlier detection was performed univariately on each variable using boxplot visualization.

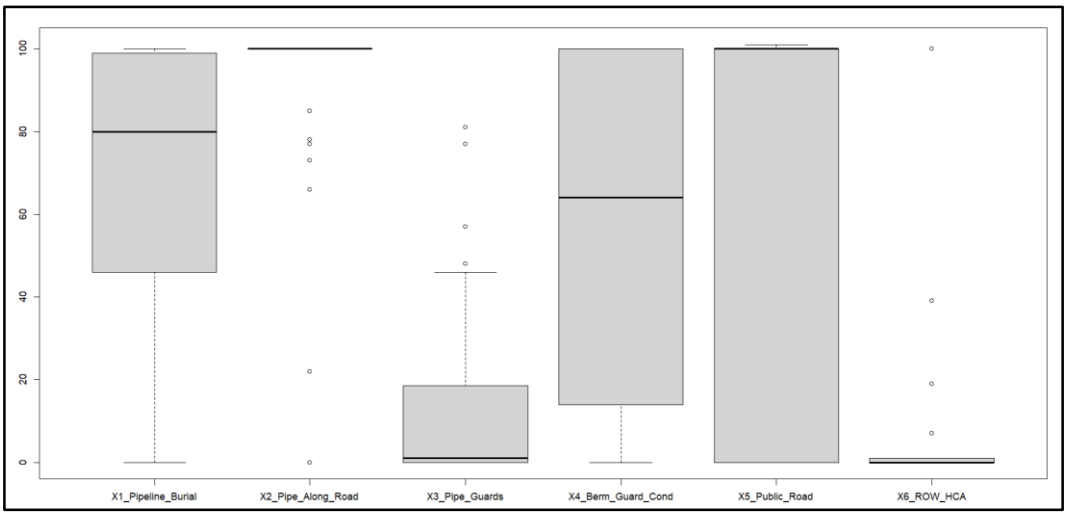

Figure 2. Boxplot Results

From Figure 2. it can be seen that variables X2 (Pipe Along Road), X5 (Public Road), and X6 (ROW HCA) have outlier points that fall outside the upper limit of the whisker. These outliers indicate extreme values that differ significantly from the majority of the other data.

Multicolinearity Test

A multicollinearity test is used to identify whether there is a strong linear correlation among the independent variables. This test is performed using the Variance Inflation Factor (VIF). If the VIF value exceeds 10, it suggests the presence of high multicollinearity.

Table 3. Multicollinearity Test

	Variables	VIF	Information
X_1	Pipeline Burial	2.6462	There is no multicollinearity
X_2	Pipe Along Road	1.2858	There is no multicollinearity
X_3	Pipe Guard	1.6609	There is no multicollinearity
X_4	Berm/Rail/Guard Condition	3.0802	There is no multicollinearity
X_5	Public Road	2.1338	There is no multicollinearity
X_6	ROW HCA	1.6079	There is no multicollinearity

Based on Table 3, all variables have VIF values below 5, which means there is no multicollinearity among the variables used. Thus, all variables are suitable for use in cluster analysis without the need for variable reduction.

Estimation of the Number of Cluster

Following the preprocessing and Min–Max normalization procedures described in the Materials and Methods section, the normalized dataset was used for cluster validation. The optimal number of clusters was first determined using the Deviation Ratio Index based on Medoid (DRIM), which evaluates the ratio between within-cluster and between-cluster deviations for values of k ranging from 2 to 7.

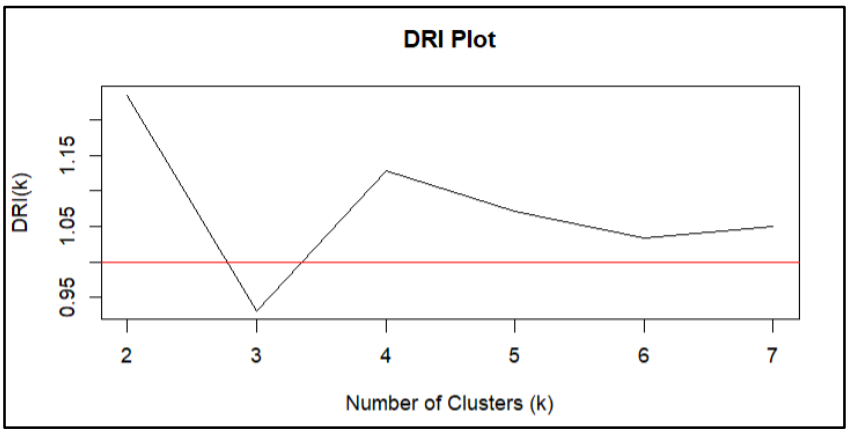

Figure 3. Plot of DRI Values

Figure 3 presents the DRI values obtained for each k. The optimal number of clusters is determined as the smallest value of $k$for which $DRI(k) < 1$. Based on the figure, this condition is first satisfied at $k = 3$, indicating that three clusters provide a more stable and efficient structure compared to $k + 1$. The comparison between $DR(k)$and $DR(k + 1)$also shows a notable decrease at $k = 3$, suggesting improved cluster separation and compactness. Since DRIM is specifically designed for medoid-based clustering, this result serves as the primary reference in determining the optimal number of clusters.

Apart from using the DRIM approach, validation of the cluster amount can also be reinforced with other methods, namely Variance Ratio Criterion (VRC), which can be visualized in the picture below.

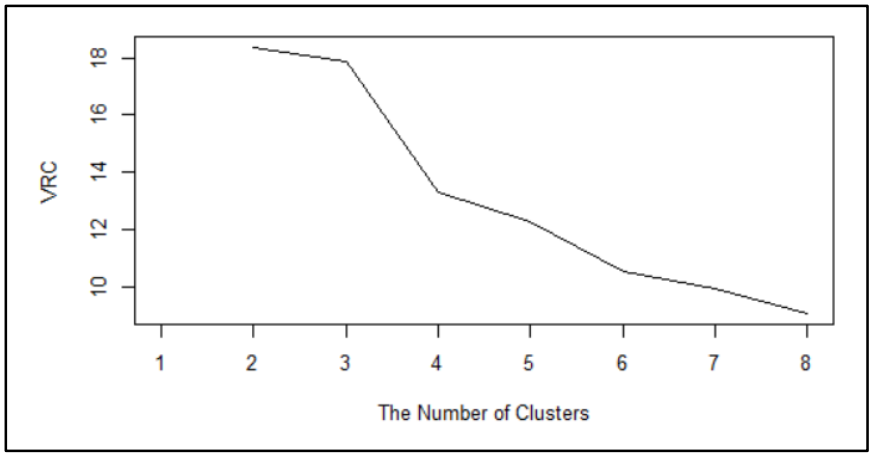

Figure 4. Plot of VRC Values

Based on Figure 4, the VRC reaches its maximum value at $k = 2$, followed by a decline at $k = 3$ and subsequent values. This pattern indicates that two clusters provide the strongest variance-based separation. However, the decrease at $k = 3$is not drastic, suggesting that three clusters still maintain reasonable separation before further reductions occur. Therefore, from the VRC perspective, $k = 2$ emerges as a strong candidate, while $k = 3$ remains a plausible alternative.

To obtain maximum results, other methods can be used as a comparison in determining the number of clusters. The method that can be used is the Hartigan method. Diana can be visualized in the following image.

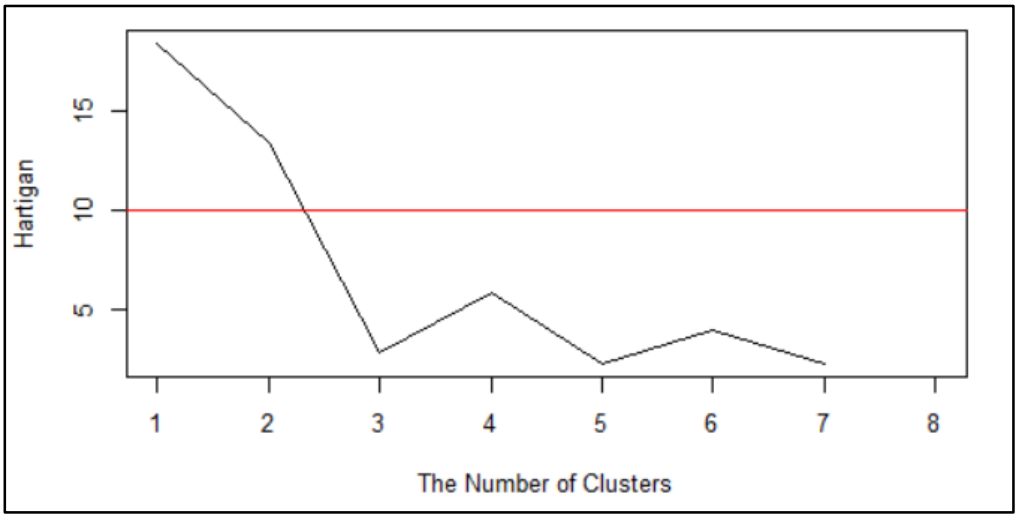

Figure 5. Hartigan Index Plot

Based on Figure 5, it can be seen that the Hartigan index *drastically* decreases from k=2 to k=3, then remains below the threshold of 10 at k>3. According to the general rule, if the Hartigan index < 10, then the addition of clusters no longer provides significant separation. Therefore, the number of clusters of three is considered the optimal point, as there is no significant increase in information beyond this point, and the Hartigan value indicates stability.

Based on the results obtained from DRIM, VRC, and the Hartigan method, the optimal number of clusters is determined to be three. Although VRC suggests that two clusters yield the highest variance ratio, both DRIM and Hartigan indicate that three clusters provide a better balance between compactness and separation. Considering that DRIM is methodologically consistent with the Block-Based K-Medoids approach, three clusters are selected as the most representative solution for the analyzed dataset.

Table 4. PKM Data Grouping

Cluster	Cluster Members	Total
1	KM 24, KM 25, KM 26, KM 27, KM 28, KM 29, KM 30, KM 31, KM 32, KM 33, KM 34, KM 35, KM 36, KM 37, KM 38, KM 39, KM 40, KM 41, KM 50, KM 51, KM 52	21
2	KM 2, KM 3, KM 4, KM 5, KM 6, KM 7, KM 8, KM 9, KM 10, KM 11, KM 12, KM 13, KM 14, KM 15, KM 16, KM 17, KM 18, KM 19, KM 20, KM 21, KM 22, KM 23	22
3	KM 1, KM 42, KM 43, KM 44, KM 45, KM 46, KM 47, KM 48, KM 49, KM 53, KM 54, KM 55, KM 56, KM 57, KM 58, KM 59	16

Table 4 presents the grouping results obtained from the Block-Based K-Medoids method. Cluster 1 consists of 21 members (KM 24–KM 41 and KM 50–KM 52), Cluster 2 includes 22 members (KM 2–KM 23), and Cluster 3 contains 16 members (KM 1, KM 42–KM 49, and KM 53–KM 59). The relatively balanced distribution of members across clusters indicates that the selected number of clusters (k = 3) provides a reasonable partition of the dataset.

After clustering using the Deviation Ratio Index based on Medoid (DRIM), each cluster is represented by a medoid that reflects the central characteristics of the group. The medoids correspond to the 38th, 3rd, and 48th observations for Clusters 1, 2, and 3, respectively. These medoids serve as representative profiles, obtained through the minimization of internal deviations within each cluster.

Table 5. Profiling

Cluster	X1(Pipeline Burial)	X2(Pipe Along Road)	X3(Pipe Guard)	X4 (Berm/Rail/Guard Condition)	X5 (Public Road)	X6 (ROW HCA)
1	74	100	16	81	100	0
2	100	100	0	9	0	0
3	36.5	100	7.5	100	100	100

Based on Table 5, median analysis per variable shows that every cluster has its own specific characteristics related to the condition of the infrastructure of the pipe **Cluster 1** shows that a considerable proportion of pipeline segments are located above ground (*Pipeline Burial* = 74), indicating substantial exposure. All segments pass along public roads (*Pipe Along Road* = 100; *Public Road* = 100), although they are not situated within High Consequence Areas (*ROW HCA* = 0). The protective infrastructure also shows vulnerability, as 16% of segments do not have pipe guards (*Pipe Guard* = 16), and the condition of berm/rail/guard structures is largely poor (*Berm/Rail/Guard Condition* = 81). These characteristics indicate notable exposure and moderate structural weakness, yet without HCA involvement. Based on this median profile, Cluster 1 can be categorized as having a moderate risk level.

Cluster 2 indicates that all pipeline segments are located above ground (*Pipeline Burial* = 100). However, all segments are equipped with pipe guards (*Pipe Guard* = 0), and the berm/rail/guard condition is relatively good (*Berm/Rail/Guard Condition* = 9). In addition, these segments are neither located along public roads nor within High

Consequence Areas (*Public Road* = 0; *ROW HCA* = 0). Although the burial condition increases physical exposure, the presence of protective structures and absence of public interaction significantly reduce the potential impact. Therefore, based on its median characteristics, Cluster 2 can be classified as a low-risk cluster.

Cluster 3 shows that all pipeline segments pass through public roads and High Consequence Areas (*Pipe Along Road* = 100; *Public Road* = 100; *ROW HCA* = 100), indicating the highest level of public exposure. Although a smaller proportion of segments are located above ground compared to Cluster 1 (*Pipeline Burial* = 36.5), protective infrastructure conditions are the poorest among all clusters (*Berm/Rail/Guard Condition* = 100), and 7.5% of segments lack pipe guards (*Pipe Guard* = 7.5). The combination of public interaction, HCA involvement, and degraded protective conditions indicates that Cluster 3 represents the highest risk level among the three clusters.

To clarify results profiling that has been done previously, results *clustering* can be performed using a form map with QGIS software. This visualization aims to present a clearer and more detailed graphical representation of the grouping results. Here is a display of the visualization.

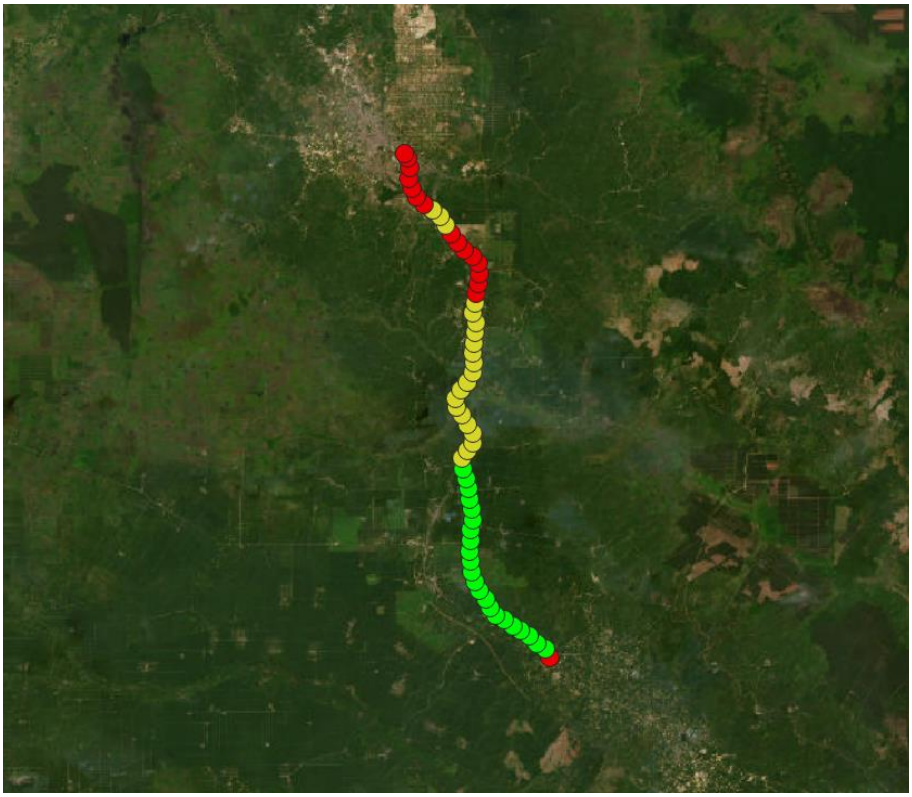

Figure 6. Cluster Visualization

Cluster 3 is concentrated toward the final section of the route and at KM 1, indicating areas with greater public and high-consequence exposure. Cluster 2 dominates the middle section (KM 2–KM 23), reflecting relatively controlled infrastructure conditions. Meanwhile, Cluster 1 occupies KM 24–KM 41 and partially KM 50–KM 52, representing segments with public exposure but without HCA involvement. This spatial

differentiation strengthens the justification for classifying the clusters into low, moderate, and highrisk levels based on their median infrastructure characteristics.

Conclusion

This study applied the Block-Based K-Medoids clustering method with the Deviation Ratio Index based on Medoid (DRIM) to determine the optimal segmentation of pipeline infrastructure from KM 1 to KM 59. The validation results consistently indicated three clusters as the most representative structure. The clustering analysis reveals clear differences in exposure level, protective infrastructure condition, and spatial distribution along the route, enabling classification into low-, moderate-, and high-risk segments. From a practical perspective, the results provide a structured basis for prioritizing inspection, maintenance, and mitigation strategies. Segments located along public roads and within High Consequence Areas require priority intervention, while segments with adequate protection but above-ground exposure require continuous monitoring. Despite its contribution to risk-based pipeline segmentation, this study is limited to static infrastructure indicators. Future research may integrate temporal data or hybrid modeling approaches to improve predictive capability.

References

[1] P. Gautam, R. K. Khutey, S. K. Rath, A. Srivastava, and V. K. Singh, "Risk Analysis of Oil and Natural Gas Pipelines Due to Hazards," *Int. J. Res. Eng. Sci.*, vol. 10, no. 8, pp. 165–176, 2022.

[2] J. A. Ali *et al.*, "Investigating the Influence of Environmental Factors on Corrosion in Pipelines Using Geospatial Modeling," *UHD J. Sci. Technol.*, vol. 8, no. 1, pp. 1–12, 2024, doi: 10.21928/uhdjst.v8n1y2024.pp1-12.

[3] T. Hu and J. Guo, "Development and Application of New Technologies and Equipment for In-line Pipeline Inspection," *Nat. Gas Ind. B*, vol. 6, no. 4, pp. 404–411, 2019, doi: 10.1016/j.ngib.2019.01.017.

[4] Q. Ma *et al.*, "Pipeline in-line inspection method, instrumentation and data management," *Sensors*, vol. 21, no. 11, p. 3862, 2021, doi: 10.3390/s21113862.

[5] A. Nurissa'adah, E. Ismiyah, and A. W. Rizqi, "Analysis of Occupational Health, and Safety (K3) in the Workshop Area Using the HIRA and 5S Methods at PT. Ravana Jaya," *Motiv. J. Mech. Electr. Ind. Eng.*, vol. 4, no. 2, pp. 161–174, 2022, doi: 10.46574/motivection.v4i2.122.

[6] Y. Jianxing, W. Shibo, C. Haicheng, Y. Yang, F. Haizhao, and L. Jiahao, "Risk assessment of submarine pipelines using modified FMEA approach based on cloud model and extended VIKOR method," *Process Saf. Environ. Prot.*, vol. 155, pp. 555–574, 2021, doi: 10.1016/j.psep.2021.09.047.

[7] P. K. Dey, S. O. Ogunlana, and S. Naksuksakul, "Risk-based maintenance model for offshore oil and gas pipelines: A case study," *J. Qual. Maint. Eng.*, vol. 10, no. 3,

pp. 169–183, 2004, doi: 10.1108/13552510410553226.

[8] Z. Zemanova, S. Krocova, and P. Sirotiak, "Risk Management in the Water Industry," *Eng. Proc.*, vol. 57, p. 20, 2023, doi: 10.3390/engproc2023057020.

[9] S. E. Bitty, L. A. Hendratta, A. H. Thambas, and G. Malingkas, "Manajemen risiko pada sistem penyediaan air minum (SPAM) perpipaan dengan metode failure mode and effect analysis dan fault tree analysis di Kabupaten Minahasa Utara," *Padur. J. Tek. Sipil Univ. Warmadewa*, vol. 13, no. 2, pp. 138–147, 2024.

[10] D. Fatmawaty, "Analisis Pertanggungjawaban Pencemaran Lingkungan Akibat Tumpahan Minyak (Studi Kasus: Kebocoran Pipa Minyak di Teluk Balikpapan)," *Bumi Lestari J. Environ.*, vol. 20, no. 1, p. 14, 2020, doi: 10.24843/blje.2020.v20.i01.p03.

[11] K. Noussia, "The BP Oil Spill Environmental Pollution Liability and Other Legal Ramifications," *Eur. Energy Environ. Law Rev.*, vol. 20, no. 3, pp. 98–107, 2011, doi: 10.54648/EELR2011009.

[12] Kariyam, Abdurakhman, Subanar, H. Utami, and A. R. Effendie, "Block-Based K-Medoids Partitioning Method with Standardized Data to Improve Clustering Accuracy," *Math. Model. Eng. Probl.*, vol. 9, no. 6, pp. 1613–1621, 2022, doi: 10.18280/MMEP.090622.

[13] Kariyam, Abdurakhman, Subanar, and H. Utami, "The Initialization of Flexible K-Medoids Partitioning Method Using a Combination of Deviation and Sum of Variable Values," *Math. Stat.*, vol. 10, no. 5, pp. 895–908, 2022, doi: 10.13189/ms.2022.100501.

[14] J. J. Soria, O. Poma, D. A. Sumire, J. H. F. Rojas, and S. M. R. Chipa, "Multiple Linear Regression Model of Environmental Variables, Predictors of Global Solar Radiation in the Area of East Lima, Peru," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1006, no. 1, p. 012009, 2022, doi: 10.1088/1755-1315/1006/1/012009.

[15] Kariyam, Abdurakhman, and A. R. Effendie, "A medoid-based deviation ratio index to determine the number of clusters in a dataset," *MethodsX*, vol. 10, p. 102084, 2023, doi: 10.1016/j.mex.2023.102084.

[16] R. Dangl and F. Leisch, "Effects of Resampling in Determining the Number of Clusters in a Data Set," *J. Classif.*, vol. 37, no. 3, pp. 558–583, 2020, doi: 10.1007/s00357-019-09328-2.