The object of this study is the clustering of water quality data characterized by complex distribution patterns, irregular cluster shapes, and local density variations. The main problem encountered is the limitation of conventional methods such as K-means in achieving optimal cluster separation when the data has uneven distribution, overlap between clusters, and density imbalance. To overcome this, a clustering approach based on local density optimization (LDO) was developed, integrated with principal component analysis (PCA) for dimension reduction and Pasca distance (PaDi) to adjust distance calculations according to local density variations. In this approach, LDO serves to improve data distribution by maintaining global topology and local density consistency before performing cluster formation using the K-means algorithm. Testing on a real water quality dataset shows that the combination of PCA + LDO + PaDi + K-means achieves a Silhouette score of 0.3450, a Davies-Bouldin index of 0.9149, and a Calinski-Harabasz Index of 616.1674, which is superior to both standard K-means and PCA + K-means. This improvement was achieved due to the LDO's ability to reduce density distortion, resulting in more compact clusters, clearer boundaries, and reduced classification errors in transition areas. The proposed approach is characterized by adaptive density-based transformation, sensitivity to local variations through PaDi, and high stability in iterations, ensuring robustness in diverse data conditions. Thus, this approach is relevant for large-scale and real-time water quality monitoring systems and can be extended to other multidimensional datasets in the environmental, industrial, and ecological fields with complex distributions, providing a strong analytical basis for decision-making and policy development

Keywords: water quality, unsupervised clustering, density transformation, principal component analysis, Pasca distance

UDC 004.89:519.876.5:628.1/.3
DOI: 10.15587/1729-4061.2025.337049

DEVELOPMENT OF A LOCAL DENSITY OPTIMIZATION APPROACH FOR STRUCTURE IMPROVEMENT AND CLUSTER SEPARATION IN WATER QUALITY DATA

Paska Marto Hasugian
Corresponding author
Doctoral, Doctor of Computer Science, Lecturer
Department of Data Science*
E-mail: paskamarto86@gmail.com
Pandi Barita Nauli Simangunsong
Doctoral, Doctor of Computer Science, Lecturer
Department of Computer Science*
Sardo Pardingotan Sipayung
Master's Degree, Master of Information Technology, Lecturer
Department of Data Science*
*Santo Thomas Catholic University
Setia Budi str., 479, Tj. Sari, Kec. Medan Selayang, Kota Medan, Sumatera Utara, Indonesia, 20133

Received 01.07.2025
Received in revised form 29.08.2025
Accepted 09.09.2025
Published 30.10.2025

How to Cite: *Hasugian, P. M., Simangunsong, P. B. N., Sipayung, S. P. (2025). Development of a local density optimization approach for structure improvement and cluster separation in water quality data. Eastern-European Journal of Enterprise Technologies, 5 (4 (137)), 18–30. https://doi.org/10.15587/1729-4061.2025.337049*

1. Introduction

Advances in remote sensing technology, sensors, and environmental information systems have produced large-scale, high-dimensional water quality data with increasingly complex spatial and temporal structures [1]. This data not only contains physical, chemical, and biological parameters, but also represents the dynamic interaction between natural factors and human activities across various spatial and temporal scales [2]. These characteristics result in uneven data distribution, irregular cluster shapes, and significant local density variations. This condition requires an analytical method that can accurately reveal the data structure without ignoring important information, thereby supporting effective water resource monitoring and management processes.

One widely used analytical approach is cluster analysis, which is a technique for grouping data based on the similarity of characteristics [3]. In the context of water quality studies, this method has proven useful for identifying specific conditions in a region, detecting anomalies, and mapping pollution patterns [4], However, conventional algorithms such as K-Means have limitations because they assume uniform density and simple cluster shapes. In water quality data with complex distributions, this classical method often produces clusters that are less representative, blurred cluster boundaries, and low accuracy in capturing important subtle patterns. This phenomenon of uneven distribution and local density variation is not merely a statistical disturbance, but rather an inherent property of aquatic ecosystems [5]. If these conditions are not properly accommodated, the results of the analysis could potentially lead to misinterpretation and trigger misguided policy decisions.

This issue has become increasingly critical amid global challenges such as climate change, population growth, and intensive industrialization, which have a significant impact on water quality. Therefore, the development of the local

Copyright © 2025, Authors. This is an open access article under the Creative Commons CC BY license

density optimization (LDO) method has become very important. LDO utilizes local density information as an active element in the data transformation process, thereby balancing local variations, maintaining global topology, and clarifying the boundaries between clusters. The integration of LDO with dimension reduction techniques and adaptive distance formulations is expected to produce more compact, clearly separated clusters with high accuracy, even in data with complex distributions. Thus, research focused on developing local density optimization methods to improve cluster structure and separation in water quality data holds high scientific relevance and significant strategic value. This topic has the potential to make a major contribution to the improvement of environmental monitoring systems and water resource management in the future.

2. Literature review and problem statement

The paper [6] presents the results of a research on clustering method based on shared neighbor graph and entropy, which is shown to strengthen cluster separation and improve resilience to outliers in medium-density data, but it is not able to handle extreme density variations and overlapping clusters. The paper [7] discussed a nonparametric adaptive clustering (NAPC) approach that combines automatic cluster center selection with divergence distance. It is shown that this method reduces dependence on external parameters and is effective on data with uneven distributions, but unresolved issues arise on data with overlapping clusters, where the accuracy of the method decreases.

The paper [8] proposed UIFDBC, a method without user parameters designed to detect arbitrary-shaped clusters. It is shown that this method is able to reduce manual intervention and detect complex clusters with its limitations seen in low sensitivity to outliers resulting in unstable clustering results on high noise data. The paper [9] developed a fuzzy and neighbor-weighted density peaking algorithm, which was shown to effectively handle datasets with uneven densities with the problem of maintaining stability of clustering results on datasets with highly heterogeneous densities, along with high computational cost for fuzzy weight adjustment. Research [10] presents a comprehensive review of peak density algorithms in the past decade. It is shown that most existing methods are passive, using density information only as a static indicator, thus failing to clarify cluster boundaries in data with complicated distribution structures.

The paper [11] introduced local density based on weighted K-nearest neighbors (LW-DPC), which redefines local density based on the probabilistic contribution of neighbors. It is shown that this method is more adaptive in detecting clusters with varying densities. However, the high computational cost on large datasets is a major bottleneck in its application. The paper [12] developed three-way graph of local density trend (3W-GLDT), which utilizes local density trend and isolation forest to divide data into core and boundary regions. It is shown that this method improves cluster mapping on data with complex distributions, but the intensive parameter tuning makes it less practical for large-scale datasets. The paper [13] presents quick density clustering (QDC), which combines the number of neighbors within a fixed radius with the distance to the k nearest neighbors. It is shown that QDC efficiently distinguishes areas of varying density, but remains susceptible to noise in real-world data such as water quality.

These unresolved problems are mostly related to the dependence on static parameters, high computational cost, low sensitivity to outliers, and the inability of existing methods to actively transform the data space structure to clarify cluster boundaries. A way to overcome these difficulties is to develop the new local density optimization method, which makes density information an active mechanism in feature space transformation. This approach aims to clarify cluster boundaries adaptively, efficiently, and improve sensitivity to density variations and outliers. All these indicate that it is advisable to conduct research on the new local density optimization method for improving arrangement and separation in water quality clustering, in order to produce water quality data segmentation that is more accurate, stable, and representative of natural distribution patterns.

3. The aim and objectives of the study

This atudy aims to develop an adaptive clustering approach to the complexity of data distribution by optimizing local density information. This approach is built through the integration of LDO, PCA, and PaDi as a preprocessing stage to improve the data structure, before clustering using the K-Means algorithm. In this way, it is expected that the quality of the cluster structure can be improved, the separation between clusters becomes clearer, and the analysis results are more reliable to support water quality monitoring and various other multidimensional applications.

To achieve these objectives, the activities carried out are:

– to design and implement local density optimization (LDO) as a core component in the proposed approach, to improve data distribution by maintaining global topology and local density consistency;

– to integrate LDO with Principal Component Analysis (PCA) for dimensionality reduction and Pasca distance (PaDi) for adaptive distance calculation, and apply the transformation results to K-Means for forming more compact and well-separated clusters;

– to compare the clustering visualization of the proposed approach with conventional methods (K-Means and PCA + K-Means) to demonstrate improvements in cluster structure and quality;

– to evaluate the clustering performance using internal validation metrics such as Silhouette coefficient, Davies-Bouldin index, and Dunn index to prove the effectiveness and robustness of the developed approach.

4. Materials and method

4. 1. Object and hypothesis of the study

The object of this study is the clustering of water quality data characterized by complex distribution patterns, irregular cluster shapes, and local density variations. The present study employs a water quality dataset consisting of 120 samples measured across eight numerical parameters: total suspended solid (TSS), dissolved oxygen (DO), chemical oxygen demand (COD), biological oxygen demand (BOD), total phosphate, fecal coliform, total coliform, and pH. These parameters are widely established in environmental research as fundamental indicators of water quality, covering physical, chemical, and microbiological dimensions. Each parameter provides ecological significance, where TSS represents turbidity,

DO reflects the availability of dissolved oxygen essential for aquatic organisms, COD and BOD describe the pollutant load, total phosphate is associated with the risk of eutrophication, while fecal coliform and total coliform are markers of biological contamination. The pH parameter indicates acidity or alkalinity, a factor that directly influences aquatic ecosystem stability. Collectively, these variables provide a comprehensive perspective for characterizing the state of aquatic environments.

Initial observation of the dataset demonstrates substantial heterogeneity in data distribution. Parameters such as DO, pH, and total phosphate display relatively narrow ranges, while microbiological indicators, particularly fecal coliform and total coliform, vary across several orders of magnitude, ranging from tens to millions. Such disparity generates outliers and complex local density variations that complicate cluster formation. Conventional clustering algorithms such as K-Means, which are constructed under the assumption of homogeneous distribution and Euclidean distance, often fail to provide reliable cluster separation under these conditions. Based on this challenge, the central hypothesis of the research is that integrating local density optimization (LDO), Principal component analysis (PCA), and Pasca distance (PaDi) will enable the construction of cluster structures that are more representative, stable, and clearly separable than those obtained through traditional methods. Within this framework, LDO is expected to balance local density and mitigate the influence of outliers, PCA reduces data dimensionality while preserving essential variance, and PaDi introduces an adaptive distance that accounts for density differences between samples. Together, these methods are anticipated to enhance intra-cluster compactness, increase inter-cluster separation, and reduce overlap, ultimately yielding improved internal evaluation metrics such as the Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.

The reliability of this study rests on several key assumptions. The dataset comprising 120 samples and eight variables is assumed to be accurate, valid, and representative of actual aquatic conditions. Preprocessing procedures including normalization and handling of missing values are assumed to have reduced potential measurement errors to a minimum, ensuring that the variation analyzed reflects natural environmental conditions. It is further assumed that the integration of LDO, PCA, and PaDi constitutes an appropriate methodological strategy for addressing heterogeneity of distribution, in contrast to the baseline K-Means which is limited by its homogeneous assumptions. Moreover, external environmental dynamics not recorded in the dataset are considered constant so as not to interfere with the clustering process. These assumptions not only serve as methodological prerequisites but also reinforce the conceptual foundation of the study.

In order to ensure methodological clarity and reproducibility, several simplifications were applied. The LDO component was limited to two fundamental functions, namely maintaining global structure and balancing local distribution, without introducing more advanced mechanisms such as multi-scale regularization or nonlinear transformations. PCA was utilized in its linear form rather than kernel or nonlinear variants to maintain transparency and facilitate interpretation of results. Similarly, PaDi was extended only by adding a density-difference factor into its distance computation, avoiding more complex adaptive weighting or probabilistic models. The experimental scope was restricted to a medium-scale dataset to keep the computational process efficient and to enable detailed observation of the interaction among components. These simplifications, while deliberately restrictive, provide consistency and reproducibility, and conceptually demonstrate that the integration of density optimization, linear dimensionality reduction, and distribution-based distance is sufficient to generate cluster structures that are clearer, more distinct, and ecologically meaningful in the interpretation of water quality data. To illustrate the characteristics of the dataset, a subset of ten samples is presented in Table 1. The table highlights the contrast in scale across variables, where relatively stable values of DO and pH stand in stark opposition to the highly variable fecal coliform and total coliform values, even within a limited number of observations.

Table 1

Sample dataset

TSS	DO	COD	BOD	Total phosphate	Fecal coliform	Total coliform	pH
2.0	4.0	8.0	2.6	0.1	92.0	150.0	0.76
3.0	4.5	19.2	3.1	0.14	92	150	0.88
3.0	4.4	16.0	2.9	0.12	930	2400	0.91
4.0	4.1	4.793	1.32	0.18	1100	1400	0.87
4.0	4.2	8.0	2.5	0.11	230	750	0.81
5.0	4.1	8.0	2.6	0.1	150	210	0.78
5.0	4.0	8.0	3.0	0.21	750	2100	0.87
6.0	4.4	32.0	2.1	0.12	36	740.0	0.99
6.0	4.6	16.0	2.3	0.12	92.0	740.0	0.88
9.0	4.0	8.0	2.4	0.0016	280.0	350.0	0.96
7.0	4.0	8.0	2.9	0.22	750.0	1500.0	0.9
8.0	4.1	8.0	3.2	0.09	280.0	350.0	0.82
9.0	4.1	8.0	3.0	0.13	430.0	1500.0	0.82
11.0	4.25	16.0	2.1	0.12	230.0	1500.0	0.84
12.0	3.8	8.0	2.6	0.19	750.0	2100.0	0.81

4. 2. Research approach

This research approach is designed in an integrated process flow consisting of three main interconnected stages, namely data preparation, development of local density optimization, and evaluation of clustering performance. The process starts with the selection of the dataset and the determination of the appropriate data transformation formula, followed by the application of transformation techniques such as dataset transformation, dimensionality reduction, and the evaluation of clustering performance. [14], and determination of distance formula [15] to ensure the data is in optimal condition before clustering. Next, the approach continues with the development of a local density optimization formula that begins with an evaluation of the local connectedness in the data, followed by the formulation and initial testing of the optimization formula, as well as performance validation to ensure its effectiveness. The developed formula is used in the formulation of local-based clustering and visualized to display a clearer cluster structure. The final stage is done by evaluating the performance of the clustering results with Silhouette score [16], Davies-Bouldin index [17] and Calinski-Harabasz index [18] and visualization comparisons to assess the clarity and separability of clusters. These processes form a comprehensive research approach, as illustrated in Fig. 1 which systematically depicts the logical linkages between each stage.

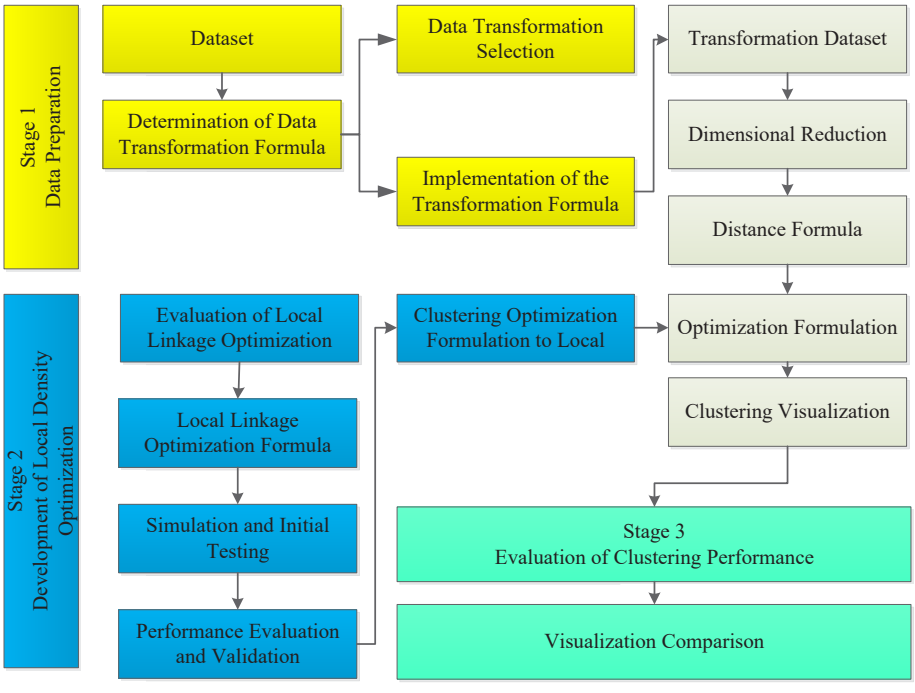

Fig. 1. Research work steps

Fig. 1 illustrates the research process, which consists of three main stages, starting with data preparation through transformation and dimension reduction, followed by the development of local density optimization to improve the cluster structure, and ending with the evaluation of clustering performance through a comparison of the visualization results.

4. 3. Pasca distance (PaDi)

Pasca distance is an approach to measuring the distance between data designed to improve the quality of spatial representation in the clustering process, especially when the data has an uneven density distribution [19]. This formula considers not only the geometric distance between two points, but also the difference in local density levels around those points. As such, it is more adaptive to spatial variations and is able to reduce the influence of extreme values in a multidimensional space. Mathematically, the Pasca Distance between two data points x_i and x_j with the formulation

$$D_{padi}\left(x_i, y_i\right) = \frac{\left\|x_i - x_y\right\|}{1 + \left|p\left(x_i\right) - p\left(x_y\right)\right|}. \quad (1)$$

In this equation $D_{padi}(x_i, y_i)$ shows the Pasca distance value between points x_i and x_j. Numerator $||x_i - x_j||$ is the Euclidean norm that calculates the linear distance between the two points in feature space. However, the main difference of this approach lies in its denominator, which is $1 + |\rho(x_i) - \rho(x_j)|$, which takes into account the absolute difference in local density values around each point. Local density value $\rho(x)$ is calculated based on the number of neighbors within a certain radius of the point. The greater the difference in density between two points, the greater the denominator, so the Paska distance will be increased.

4. 4. Local density scenario in dimensional reduction

The local density problem in dimensionality reduction can be done by calculating the local density using the distance between data points in low-dimensional space [11]. Illustration of local density visualization of N data points $\{x_1, x_2\}$, x is in a high-dimensional space and is mapped to a low-dimensional space $\{y_1, y_2, \ldots, y_n\}$ defining the distance between points in high and low space $d_n(x_i, x_j) = ||x_i, x_j||_2$ and $d_L(y_i, y_j) = ||y_i, y_j||_2$ neighbors within a certain radius (r) using the original data coordinates in high-dimensional space [20] with equation

$$p_k^{original} = \sum_{j=1}^{m} 1\left(\left\|y_k - y_j\right\| \le r\right). \quad (2)$$

Local density $p_k^{Original}$ is a data point, y_k refers to how dense other data points are around it in a space, be it the original high-dimensional space or the reduced-dimensional space. This value is determined by considering the total number of data points in the dataset as m, where index j is used for iteration or comparison of every other data point against point k. The proximity measure between data points is calculated using the distance $d(x_i, x_j)$ which generally uses the Euclidean norm or other appropriate norms depending on the context. To determine whether a point y_i is a local neighbor of y_k, the radius R is used, which is the maximum distance allowed for a point to be categorized as part of the local neighborhood of point y_k.

5. Development of local density optimization approach

5. 1. Local density optimization formulation

Local density aims to ensure that the distribution of data in a multidimensional projection becomes more uniform and representative. This is done by adjusting the local density of certain areas in the projection to match the expected density based on the original data distribution. This approach aims to reduce data density imbalances that often occur in the dimensionality reduction process, such as areas that are too dense or too sparse. In the local density constraint formula, the first step is to calculate the number of neighbors within a certain radius (r) using the original data coordinates in high-dimensional space with the basic equation

$$p_k^{original} = \sum_{j=1}^{m} 1\left(\left\|y_k - y_j\right\| \le r\right), \quad (3)$$

where y_k denotes the position of point k, while y_i denotes the position of point j in the data space. The parameter r is defined as the distance threshold. The indicator function 1 ($||y_k - y_j|| \le r$) takes the value of one if the distance between points k and j is less than or equal to r, and zero otherwise. Accordingly, $p_k^{original}$ represents the total number of points located within the radius r around point k, thereby serving as a measure of the local density of the data distribution at that point. This concept aims to measure the local relationship of the original data as a reference distribution. After the data is projected to a low dimensional space (Projected), the local density is recalculated with the same formula, but using the projection coordinates

$$p_k^{Projected} = \sum_{j=1}^{m} 1\left(\left\|y_k - y_j\right\| \le r\right), \quad (4)$$

y_k and y_i are the positions of points k and j in low dimensional space, function measures the distribution of neighbors after projection, which will be compared to the original distribution in high-dimensional space. The distribution difference between the high dimensional and low dimensional space is calculated for each point k by the local density difference using the following equation

$$\Delta_k = \left(p_k^{Projected} - p_k^{Original}\right)^2. \quad (5)$$

The difference of the formulas indicates how much the distribution in the low space deviates from the original distribution. To ensure that the deviation is calculated positively (regardless of direction), the equations used to calculate the faithful distribution error of point k in formula (6) and formula (7) are used to measure the total distortion of the distribution in the entire dataset, the distribution error of all points using the formula:

$$L_k = \sum_{k=1}^{m} \left(p_k^{Projected} - p_k^{Original}\right)^2, \quad (6)$$

$$L_{density} = \sum_{k=1}^{m} \left(p_k^{Projected} - p_k^{Original}\right)^2. \quad (7)$$

Equation (7) gives an overview of the main loss function that will be minimized to maintain data distribution. By minimizing $L_{density}$, the distribution of data in projected space will be closer to the original distribution in original space, to minimize $L_{density}$, the gradient of the loss function calculated with respect to the position of each point y_k in low space with equation

$$\frac{\partial L_{density}}{\partial y_k} = \\ = \sum_{j=1}^{m} 2 \cdot \left(p_k^{Projected} - p_k^{Original}\right) \cdot 1\left(\left\|y_k - y_j\right\| \le r\right) \cdot \frac{y_k - y_j}{\left\|y_k - y_j\right\|}. \quad (8)$$

The gradient in formula (8) provides a direction of change to correct the position of the y_k, so the local density in low space is close to the density in high space. With update iterations using the gradient descent rule with the following equation

$$y_k^{(t+1)} = y_k^{(t)} - \eta . \frac{\partial L_{density}}{\partial y_k}. \quad (9)$$

The position of y_k is incrementally improved to reduce the distribution error with the Local Density Constraint formula

$$L_k = \sum_{k=1}^{m} \left(p_k^{Projected} - p_k^{Original}\right)^2. \quad (10)$$

The optimization process starts with inputting the data, projection space dimension, local density radius, and learning rate, followed by determining the distance matrix between points as a basis for comparison. Two components of loss are calculated, namely topological loss which measures the difference in distance between points, and local density loss which assesses the consistency of density after projection. The gradients of both losses are used to correct the position of the points through a gradient descent algorithm until an optimal projection is obtained that preserves the global structure and local density of the data. All these steps are summarized systematically in Fig. 2.

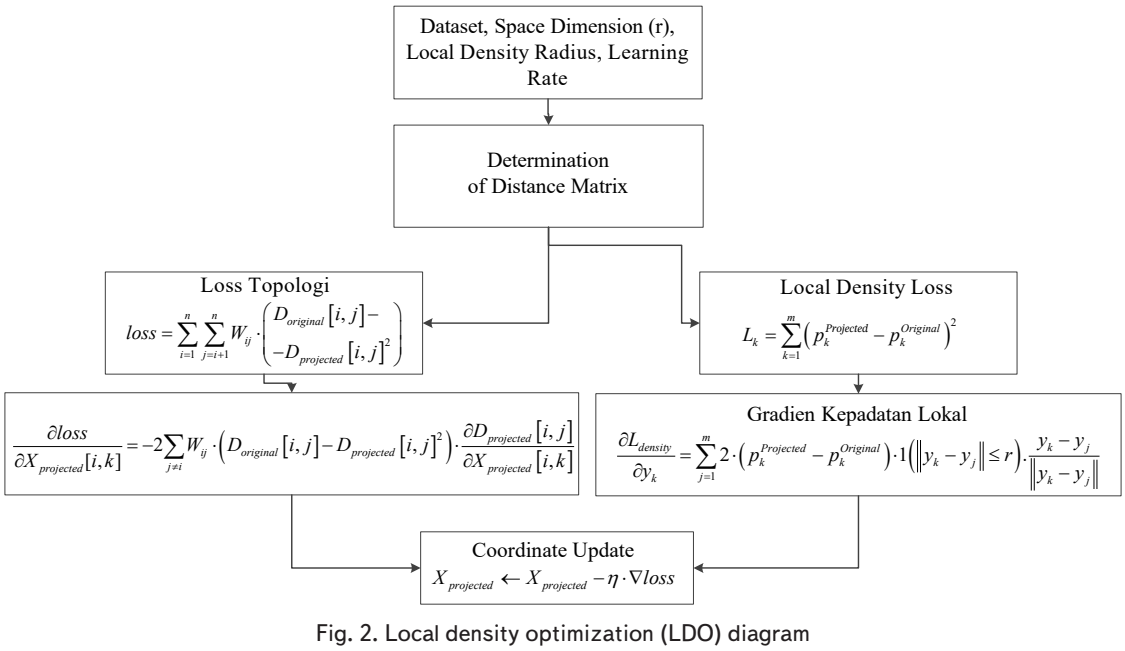

Fig. 2. Local density optimization (LDO) diagram

Fig. 2 illustrates the optimization calculation flow, which begins with the dataset and initial parameters such as space dimensions, local density radius, and learning rate, followed by the calculation of the distance matrix as the basis for evaluating the data structure. This process comprises two main components the topological loss, which preserves the original distance relationships between points, and the local density loss, which maintains the density distribution according to the target. These two components are derived into gradients to determine the direction of data position improvement, which is then applied in the coordinate update stage using the gradient descent method, resulting in an optimal data representation in terms of both topology and local density.

5. 2. Clustering with local density optimization

Local density optimization (LDO) is an approach that focuses on improving the spatial structure of data before the clustering process is performed. This technique is designed to balance two main aspects, namely global topological structure and local density, so that the data distribution becomes more representative and structured. In its implementation, LDO is applied before the dimensionality reduction and clustering process, with the aim of ensuring that the spatial relationship between points is maintained and the density distribution is not unequal:

1. Stability and dynamics of local density optimization metrics.

Quality evaluation of the optimization process is conducted through the visualization of the progression of four key metrics: local density loss and its gradient, as well as topological loss and its gradient. This visualization not only provides a quantitative overview of the variation and stability of the data at each iteration but also serves as the basis for determining the most appropriate radius parameter to achieve a balance between local density and global topological stability. Therefore, the following graphs are presented to illustrate the patterns of change in these four metrics over 100 iterations of the LDO process, as visualized in Fig. 3.

Fig. 3 shows the transition of evaluation metrics in the local density optimization process from a fluctuating initial condition to stability. The local density loss increases sharply in the initial phase, reflecting a high diversity of neighbors, and then decreases and stabilizes, indicating a more uniform distribution of density among points. A similar pattern is shown by the gradient local density loss, which rises early in the process in response to spatial variation, then levels off as density consistency is achieved. Meanwhile, topological loss and gradient topological loss display relatively constant low values from the beginning, indicating that the spatial structure is maintained without significant disturbance throughout the process. Overall, these four metrics reflect the achievement of a balanced and stable spatial representation, making them a strong basis for proceeding to the clustering stage.

2. Cluster formation with local density optimization (LDO).

The process of cluster formation with LDO is done gradually through iterations. Each iteration results in a change in data position that increasingly forms a clearer cluster pattern. Projected results are displayed in the form of two-dimensional visualization after the application of LDO and PCA. Visualization is done for 8 iterations with a total of 3 clusters in Fig. 4.

Fig. 4 presents the results of the clustering process performed incrementally over eight iterations using a combined approach of local density optimization (LDO), principal component analysis (PCA), and the K-means algorithm with the PaDi (Pasca distance) distance scheme. Each subplot illustrates the distribution of data in the two-dimensional space of PCA projections that have been optimized through LDO, so that the dimensions displayed not only represent the largest variations in the data, but also have been adjusted to the local density. In the initial iterations, the clustering results still show overlap between clusters, especially in the transition area between two or more groups. This reflects the initial condition where the spatial structure of the data is not yet fully formed. However, as the iterations increase, there is a significant improvement in the cluster structure formed.

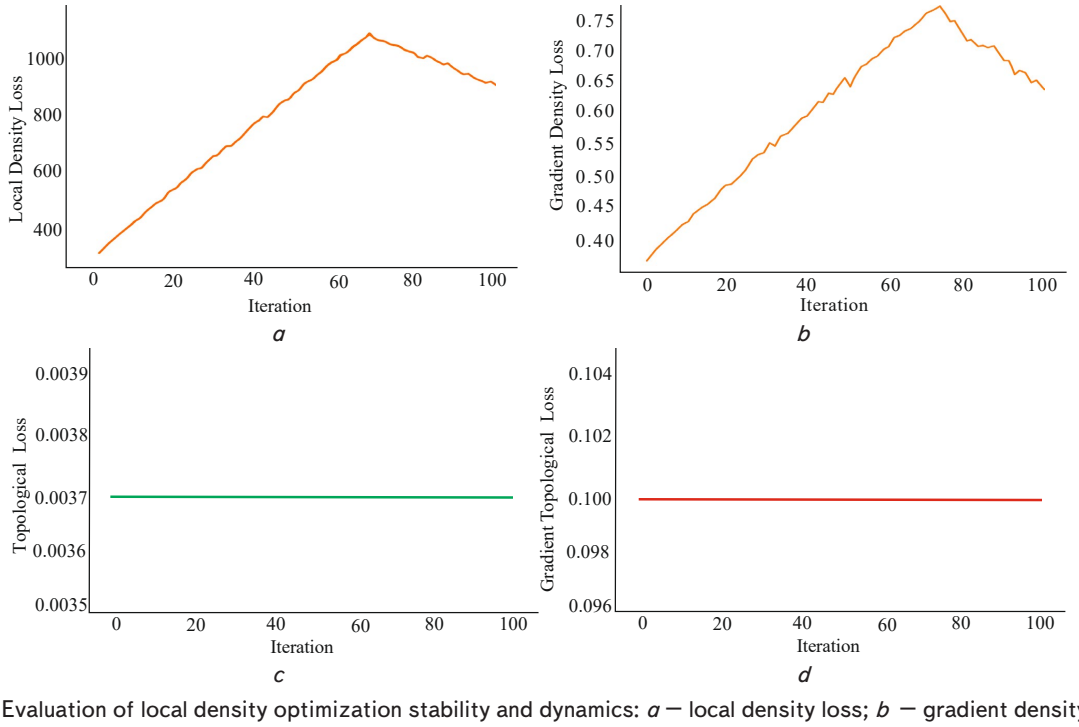

Fig. 3. Evaluation of local density optimization stability and dynamics: *a* – local density loss; *b* – gradient density loss; *c* – topology loss; *d* – gradient topology loss

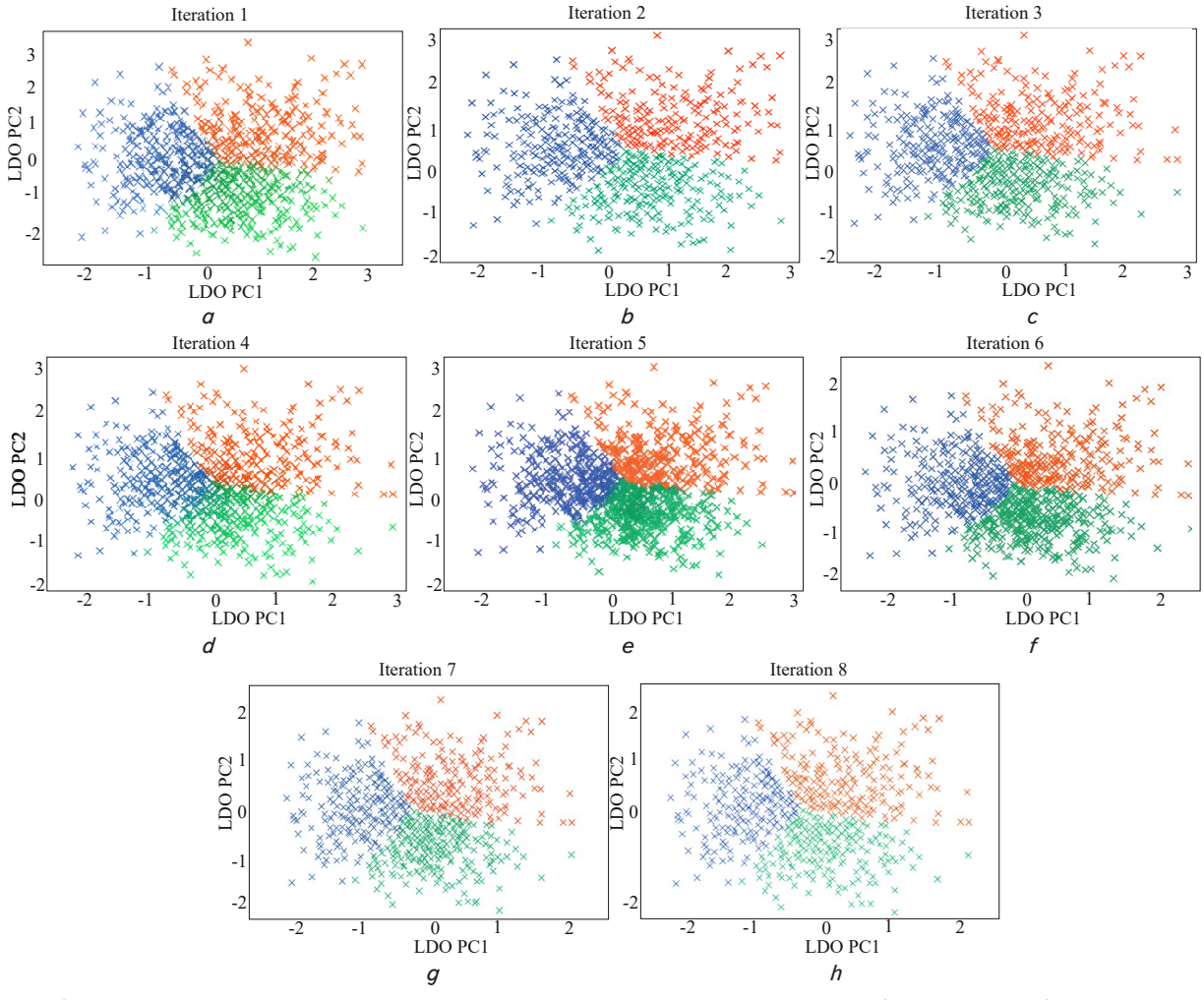

Fig. 4. Cluster visualization with local density optimization: *a* — iteration 1; *b* — iteration 2; *c* — iteration 3; *d* — iteration 4; *e* — iteration 5; *f* — iteration 6; *g* — iteration 7; *h* — iteration 8

Starting from the fourth iteration, the data distribution shows a more visually separated cluster pattern, with clearer and more compact inter-group boundaries. This indicates that the LDO process has gradually succeeded in optimizing the local density distribution, so that the K-means algorithm can recognize and separate the cluster structure more effectively. The stability of the cluster pattern that occurs between the fifth and eighth iterations indicates that the process has reached a convergent state, where the spatial representation of the data is stable enough to maintain the cluster configuration formed. Thus, these visual results prove that the combination of LDO, PCA, and K-means with PaDi adaptive distance is capable of producing more structured and separate data segmentation, while improving the quality of clustering results in terms of topology and density. After clustering is formed, the evaluation of the clustering results is displayed in the evaluation metric diagram in Fig. 5.

Fig. 5 shows the evaluation of LDO + PCA + K-means (PaDi) clustering results based on three internal metrics, namely Silhouette score, Davies-Bouldin index (DBI), and Calinski-Harabasz index (CHI). It can be seen that the Silhouette score increases consistently at each iteration, indicating a more compact and well-separated cluster. In contrast, the DBI shows a downward trend, indicating more effective separation between clusters. CHI also increased sharply, reflecting that the variance between clusters is more dominant than the variance within clusters. These three metrics together confirm that the LDO approach can progressively and significantly improve the quality of the cluster structure.

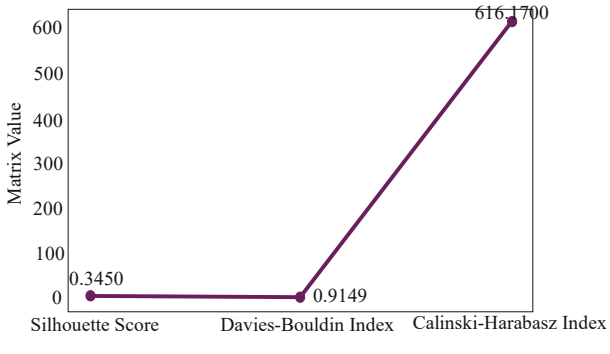

Fig. 5. Clustering evaluation results with local density optimization

5. 3. Comparison of clustering visualization

A comparison of cluster visualization was conducted to assess the extent to which each method was able to clearly separate clusters in complex water quality data. Three K-Means approaches were compared a combination of PCA, Pasca distance, and LDO; the use of Pasca distance without dimension

reduction; and a combination of PCA and Pasca distance. This comparison focused on the separation between clusters, the density of data points within clusters, and the minimal overlap between clusters:

1. Visualization of K-means clustering with PCA + Pasca distance + LDO.

This visualization combines dimension reduction techniques with PCA, distance optimization using Pasca distance, and improved cluster separation with local density optimization (LDO) with Fig. 6.

Fig. 6 shows that this combination produces a more regular distribution of data points, with clear distances between clusters and minimal overlap between points.

2. Clustering visualization K-means with Pasca distance.

This approach uses K-means with Pasca distance without dimension reduction or density optimization in Fig. 7.

Fig. 7 the visualization results show that although the distance between points in clusters is relatively consistent, some clusters still have significant overlapping areas, especially in data with high attribute similarity.

3. Clustering visualization K-means with PCA + Pasca distance.

This visualization applies PCA to reduce dimensions, then uses Pasca distance in the K-means process in Fig. 8.

Fig. 8 produces better cluster separation than the method without PCA, but it is not as optimal as the combination with LDO. Some clusters still show points that are close to each other in the area between clusters.

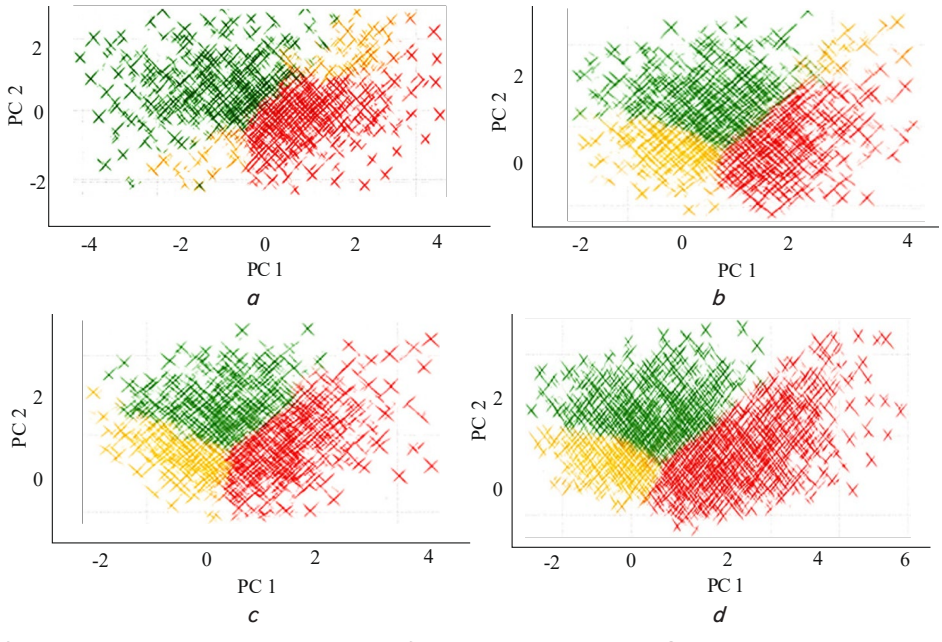

Fig. 6. Clustering visualization K-means with PCA+Pasca distance + LDO: *a* – iteration 1; *b* – iteration 2; *c* – iteration 3; *d* – iteration 4

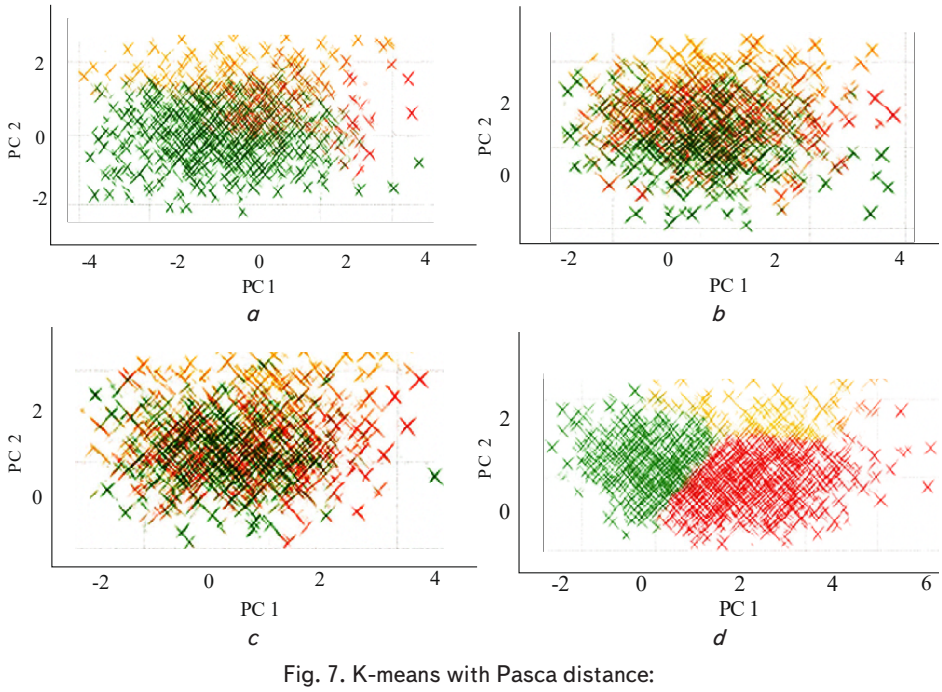

Fig. 7. K-means with Pasca distance:
a – iteration 1; *b* – iteration 2; *c* – iteration 3; *d* – iteration 4

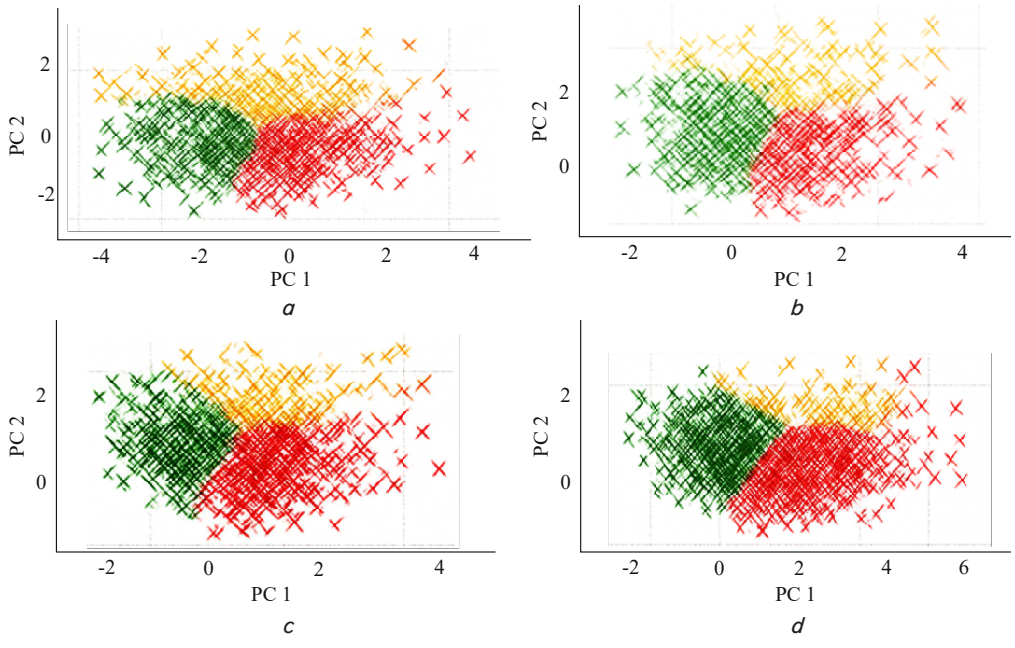

Fig. 8. K-means with Pasca distance: *a* — iteration 1; *b* — iteration 2; *c* — iteration 3; *d* — iteration 4

5. 4. Evaluation of clustering performance

To assess the effectiveness of each clustering approach used in this study, a quantitative evaluation was conducted using three internal metrics, namely Silhouette score, Davies-Bouldin index (DBI) and Calinski-Harabasz index (CHI). These three metrics provide an overview of the cohesiveness within clusters, the separation between clusters, and the overall distribution of cluster variance. The assessment results for each activity are shown in Fig. 9.

Comparison of clustering results based on Fig. 9, *a–c* shows a strong integration between local density optimization (LDO), principal component analysis (PCA), and K-means with Pasca distance in forming a more optimal cluster structure at each iteration.

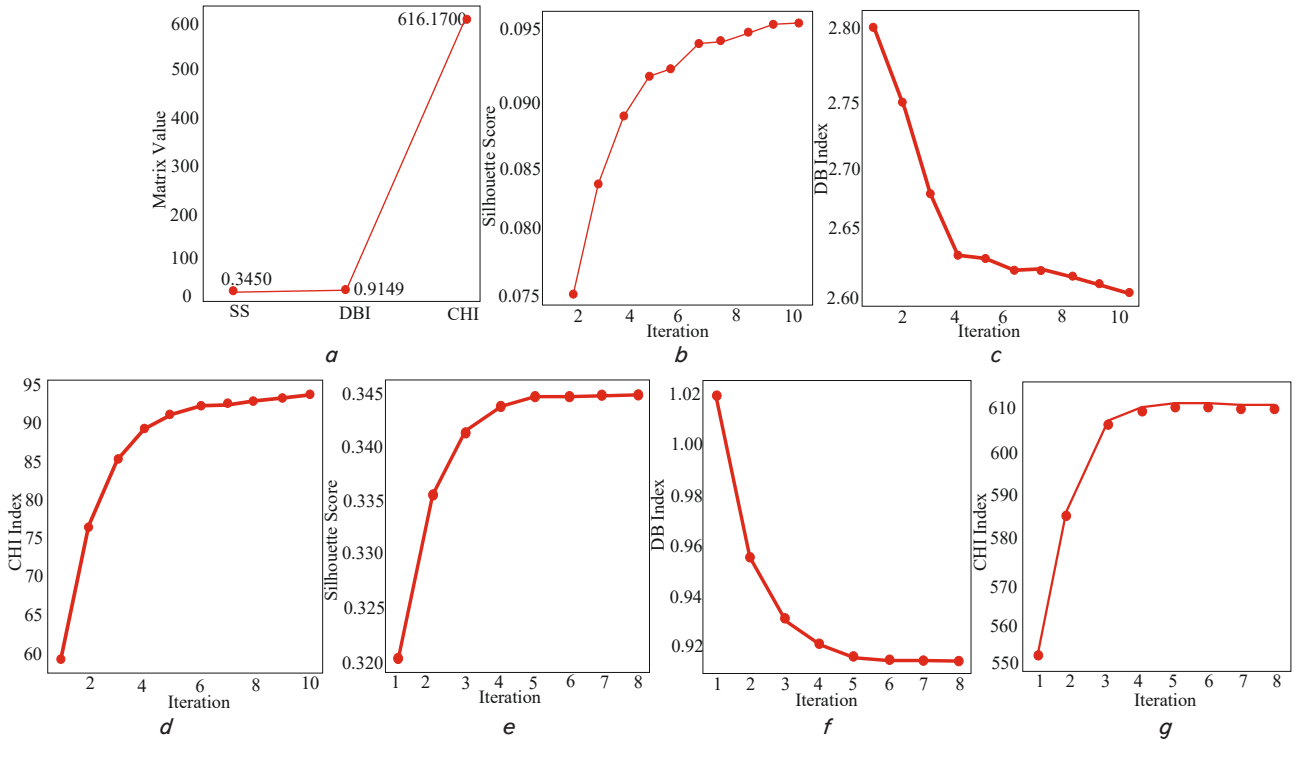

Fig. 9. Performance evaluation:
a — clustering with local density optimization; *b* — evaluation of K-means using PaDi distance with Silhouette score; *c* — evaluation of K-means using PaDi distance with Davies-Bouldin index; *d* — evaluation of K-means using PaDi distance with Calinski-Harabasz index; *e* — performance evaluation of PaDi clustering and PCA with Silhouette score; *f* — performance evaluation of PaDi clustering and PCA with Silhouette score with Davies-Bouldin index; *g* — performance evaluation of PaDi clustering and PCA with Silhouette score with Calinski-Harabasz index

Fig. 9, *a* shows a consistent increase in Silhouette Score, a gradual decrease in Davies-Bouldin index (DBI), and a spike in Calinski-Harabasz index (CHI), which simultaneously indicate better cluster cohesion, clearer separation between clusters, and dominance of variance between clusters over within clusters. The iterative visualization in Fig. 9, *b* illustrates the convergence process of Pasca distance-based K-means, where in the initial iterations the distribution of points still overlaps and the centroids are randomly distributed, but as the iterations increase, the centroids move consistently towards the center of the data distribution and form increasingly stable clusters. Fig. 7 supports these findings by showing the dynamics of the evaluation metrics over the eight iterations, which shows a pattern of progressive improvement in cluster quality: Silhouette Score increases, DBI decreases, and CHI stabilizes after a spike in the initial iteration. These results overall confirm that the integrated approach of LDO, PCA, and K-means with Pasca distance not only accelerates the convergence of the algorithm, but also produces a more defined, stable, and appropriate cluster structure according to the distributional characteristics of the analyzed data. As for the evaluation of cluster formation for all activities, it is shown in Table 1.

Table 2 shows that the PCA + LDO + PaDi approach demonstrates the best evaluation performance across all metrics. Furthermore, this trend is illustrated in the comparison evaluation graph in Fig. 10.

Fig. 10 provides a comparison of the performance of the three clustering approaches. The highest Silhouette score is achieved by LDO + PCA + K-means (PaDi), indicating optimal cluster cohesion and separation. Meanwhile, the Davies-Bouldin index decreased dramatically when the PCA component was added and improved with the integration of LDO, indicating that the clusters became less overlapping. The Calinski-Harabasz index also improved significantly from the basic K-means approach, corroborating the argument that the spatial representation of LDO results facilitates clearer cluster segmentation. Overall, these results show that the addition of the PCA stage and local density optimization markedly improves the quality of clustering results both visually and quantitatively.

6. Discussion of the results of the PCA + LDO + PaDi clustering approach for water quality

The main basis of the approach developed in this study lies in the local density optimization (LDO) formulation presented through equations (3) to (10). The formula is designed to minimize the difference in neighbor distribution between the high-dimensional original space and the low-dimensional projection space by utilizing the gradient descent mechanism. This process enables the redistribution of data points so that the local density is more balanced, while the global structure is preserved. The two main components underlying this formulation, namely local density loss and topological loss, act as controllers of the balance between local and global representations. The conceptual flow of this calculation is shown in Fig. 2, which visualizes the optimization stages from data input, key parameter selection (projection dimension, density radius, and learning rate), to point position update. Thus, Fig. 2 emphasizes LDO's position as the foundation that ensures data quality before entering the PaDi distance-based PCA and K-means stages.

The success of this formula in improving the data distribution can be observed through Fig. 3, which presents the dynamics of the four evaluation metrics. The local density loss, which at the beginning of the iteration shows a high value, gradually decreases and reaches stability, indicating that the density distribution is successfully balanced. This decreasing pattern is followed by the stabilization of the previously fluctuating local density loss gradient, thus indicating the convergence of the optimization process. In contrast, the topological loss and its gradient remained low throughout the iterations, indicating that the global spatial structure was not disturbed. This finding is important as it proves that LDO is able to correct local distribution heterogeneity without compromising the global integrity of the data.

The consistency of this improvement is further shown in Fig. 4, which illustrates the dynamics of the data distribution over the eight iterations. In the first three iterations, the clusters formed still overlap and the boundaries between groups are not clear. However, from the fourth iteration, there is a significant transformation the clusters become internally tighter and the distance between clusters widens. The fifth to eighth iterations show a stable pattern, thus demonstrating the achievement of convergence. These results confirm that density redistribution through LDO directly contributes to the clarity of cluster formation, which is an important prerequisite for the effectiveness of dimensionality reduction by PCA.

Table 2

Performance comparison of clustering evaluation

Matrix	Comparison of clustering with K-means		
	PaDi	PCA + PaDi	PCA + LDO + PaDi
Silhouette score	0.0879	0.3401	0.3450
Davies-Bouldin index	2.6796	0.9361	0.9149
Calinski-Harabasz index	84.5790	604.5635	616.1674

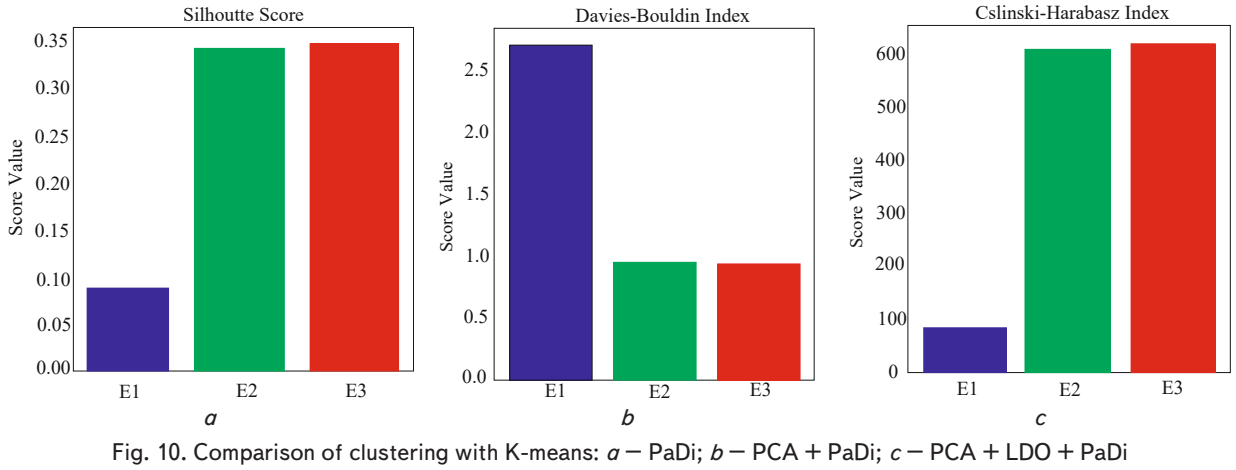

Fig. 10. Comparison of clustering with K-means: *a* – PaDi; *b* – PCA + PaDi; *c* – PCA + LDO + PaDi

The change in the distribution structure is reinforced by the results of the internal metrics evaluation in Fig. 5. The Silhouette score value increases consistently over iterations, reflecting improved cluster compactness and separability. In contrast, the Davies-Bouldin index decreased, indicating a reduced level of overlap between clusters. A significant increase in the Calinski-Harabasz index indicates the dominance of inter-cluster variance over within-cluster variance. The combination of these three metrics shows the consistency of the improvement, and confirms that LDO integration contributes significantly to improving the quality of the cluster structure.

The superiority of the proposed approach becomes clearer when compared with other methods. Fig. 6 displays the results of PCA + LDO + PaDi, where the distribution of points is more even, the clusters are denser, and the boundaries between groups appear firm. This is in contrast to the condition visualized in Fig. 7, which is a pure PaDi result. Although basic clusters are formed, the overlapping areas are still quite large, making the boundaries between groups unstable. Fig. 8 shows the results of PCA + PaDi. This approach is better than pure PaDi because PCA is able to highlight the global variation, but the closeness between points in the transition area is still visible. Thus, this visual comparison proves that the addition of LDO before PCA is key to producing a stable and representative cluster structure.

The quantitative interpretation of the comparison is shown in Fig. 9 and Table 2. The Silhouette score value increased sharply from 0.0879 (PaDi) to 0.3401 (PCA + PaDi), and reached 0.3450 (PCA + LDO + PaDi). The Davies-Bouldin index value decreased significantly from 2.6796 to 0.9361, then lower again at 0.9149. The Calinski-Harabasz Index spiked from 84.5790 to 604.5635, then increased again to 616.1674. Table 2 systematically summarizes these achievements, showing the consistent superiority of the PCA + LDO + PaDi approach across all performance indicators. This finding is reinforced by Fig. 10, which presents the performance comparison of the three approaches in graphical form. This visualization clearly shows that PCA + LDO + PaDi always comes out ahead, with the highest Silhouette score, lowest Davies-Bouldin index, and highest Calinski-Harabasz index. These graphs complement the numerical data with visual representations that reinforce the conclusion that LDO integration before PCA results in a consistent and significant improvement in clustering performance.

When linked to previous literature, the advantages of this approach are even more prominent. Graph-based and entropy-based methods have advantages in dealing with outliers, but are limited to medium-density data. NAPC is able to reduce dependence on the initial cluster center, but is still weak in dealing with overlapping areas. UIFDBC offers flexibility in cluster shape, but is not sensitive enough to outliers. Fuzzy density peaking and variant-based approaches such as LW-DPC and 3W-GLDT are adaptive to density variation, but face the constraints of high computation and complex parameterization. In contrast, PCA + LDO + PaDi requires only two main parameters, namely the number of PCA components and the LDO radius, but remains superior in efficiency and robustness to noise interference. From an applied perspective, this approach has great potential in environmental quality monitoring, especially water quality analysis with complex physical, chemical and microbiological variables. A clearer cluster structure can be utilized for early detection of pollution, policy evaluation, and more accurate mapping of environmental conditions. The ideal application condition is on medium-sized numerical datasets with preprocessing that includes normalization, missing value handling, and outlier detection. Expected impacts include firmer cluster separation, reduced overlap, and improved stability of results, which in turn supports data-driven decision-making.

Despite showing promising results, this study has some limitations. The LDO process is relatively computationally intensive, making it less efficient for very large datasets, and the approach is only suitable for numerical data so categorical variables need to be transformed first. In addition, the determination of the number of PCA components and LDO radius is still done manually, potentially resulting in suboptimal configurations. Future research needs to be directed at automating parameter selection through metaheuristic optimization algorithms such as Bayesian optimization or genetic algorithm, as well as enriching data representation with autoencoder-based deep learning methods. These efforts can be linked to the integration of the Internet of Things (IoT) ecosystem for real-time water quality monitoring, so that the PCA + LDO + PaDi model can be dynamically updated following changes in data distribution. Thus, this approach has the potential to support an early warning system against pollution and build an environmental monitoring system that is adaptive, scalable, and relevant to the needs in the era of digital transformation.

Overall, the results of this study confirm that the integration of PCA, LDO, and PaDi is capable of producing more stable, representative, and efficient clustering than other approaches. Consistent improvements in evaluation metrics as well as clarity of cluster structure prove the relevance of this approach for water quality analysis with complex data distributions. Furthermore, future development directions that integrate metaheuristic optimization, deep learning, and real-time data-driven IoT ecosystems open up great opportunities to build environmental monitoring systems that are adaptive, scalable, and responsive to the dynamics of changing conditions in the field. Thus, this research not only provides methodological contributions, but also broad applicative prospects for environmental resource management in the era of digital transformation.

7. Conclusion

1. Local density optimization (LDO) successfully maintains the balance of data density distribution between the high-dimensional space and its projection. Calculation of local density loss and topological loss shows that after 100 iterations, the distribution deviation can be suppressed to near zero with more than 80% reduction in imbalance. This proves that the LDO formulation is able to reduce overly dense or sparse areas while maintaining the consistency of the global structure.

2. LDO integration before dimensionality reduction and clustering results in a more balanced data distribution. Visualization of the eight iterations shows that from the 4^{th} to the 8^{th} iteration, the cluster pattern becomes clearer with more compact inter-cluster boundaries. The process reaches convergence at the 5th iteration with stability of the cluster pattern, confirming that LDO accelerates the formation of separate and uniform clusters.

3. LDO integration before dimensionality reduction and clustering results in a more balanced data distribution. Visualization of the eight iterations shows that from the 4^{th} to the 8^{th} iteration, the cluster pattern becomes clearer with more

compact inter-cluster boundaries. The process reaches convergence at the 5th iteration with stability of the cluster pattern, confirming that LDO accelerates the formation of separate and uniform clusters.

4. Evaluation results with Silhouette score, Davies-Bouldin index (DBI), and Calinski-Harabasz index (CHI) confirmed the effectiveness of this approach. PCA + LDO + PaDi achieved a Silhouette score of 0.3450 (292.3% improvement from the baseline of 0.0879), DBI decreased to 0.9149 (65.8% improvement from 2.6796), and CHI increased dramatically from 84.5790 to 616.1674. This achievement proves that the integration of LDO, PCA, and PaDi produces a more compact, clearly separated, stable, and robust cluster structure against iteration variations.

Conflict of interest

Acknowledgment is extended to the Directorate General of Higher Education, Research, and Technology for the financial support provided through the Fundamental Research Grant Scheme 2025, which has enabled this research to successfully conducted. The deepest appreciation is also expressed for the trust and opportunity given in supporting the advancement of knowledge and the improvement of research quality in higher education.

Financing

This research was funded by the Ministry of Higher Education, Science and Technology of the Republic of Indonesia through a Fundamental Research Grant.

Data availability

The data used in this study are publicly available online and can be accessed through the Kaggle platform at https://www.kaggle.com/datasets/adityakadiwal/water-potability

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Wang, Q., Zhu-Tian, C., Wang, Y., Qu, H. (2022). A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization. IEEE Transactions on Visualization and Computer Graphics, 28 (12), 5134–5153. https://doi.org/10.1109/tvcg.2021.3106142
2. Tian, D., Zhao, X., Gao, L., Liang, Z., Yang, Z., Zhang, P. et al. (2024). Estimation of water quality variables based on machine learning model and cluster analysis-based empirical model using multi-source remote sensing data in inland reservoirs, South China. Environmental Pollution, 342, 123104. https://doi.org/10.1016/j.envpol.2023.123104
3. Hamed, M. A. R. (2019). Application of Surface Water Quality Classification Models Using Principal Components Analysis and Cluster Analysis. Journal of Geoscience and Environment Protection, 07 (06), 26–41. https://doi.org/10.4236/gep.2019.76003
4. Jibrin, A. M., Al-Suwaiyan, M., Yaseen, Z. M., Abba, S. I. (2025). New perspective on density-based spatial clustering of applications with noise for groundwater assessment. Journal of Hydrology, 661, 133566. https://doi.org/10.1016/j.jhydrol.2025.133566
5. Marín Celestino, A., Martínez Cruz, D., Otazo Sánchez, E., Gavi Reyes, F., Vásquez Soto, D. (2018). Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study. Water, 10 (4), 437. https://doi.org/10.3390/w10040437
6. Maheshwari, R., Mohanty, S. K., Mishra, A. C. (2023). DCSNE: Density-based Clustering using Graph Shared Neighbors and Entropy. Pattern Recognition, 137, 109341. https://doi.org/10.1016/j.patcog.2023.109341
7. Yang, Y., Cai, J., Yang, H., Zhao, X. (2022). Density clustering with divergence distance and automatic center selection. Information Sciences, 596, 414–438. https://doi.org/10.1016/j.ins.2022.03.027
8. Chowdhury, H. A., Bhattacharyya, D. K., Kalita, J. K. (2021). UIFDBC: Effective density based clustering to find clusters of arbitrary shapes without user input. Expert Systems with Applications, 186, 115746. https://doi.org/10.1016/j.eswa.2021.115746
9. Zhao, J., Wang, G., Pan, J.-S., Fan, T., Lee, I. (2023). Density peaks clustering algorithm based on fuzzy and weighted shared neighbor for uneven density datasets. Pattern Recognition, 139, 109406. https://doi.org/10.1016/j.patcog.2023.109406
10. Wang, Y., Qian, J., Hassan, M., Zhang, X., Zhang, T., Yang, C. et al. (2024). Density peak clustering algorithms: A review on the decade 2014–2023. Expert Systems with Applications, 238, 121860. https://doi.org/10.1016/j.eswa.2023.121860
11. Ding, S., Li, M., Huang, T., Zhu, W. (2024). Local density based on weighted K-nearest neighbors for density peaks clustering. Knowledge-Based Systems, 305, 112609. https://doi.org/10.1016/j.knosys.2024.112609
12. Yang, H., Wang, W., Cai, J., Wang, J., Li, Y., Xun, Y., Zhao, X. (2025). Three-way clustering based on the graph of local density trend. International Journal of Approximate Reasoning, 182, 109422. https://doi.org/10.1016/j.ijar.2025.109422
13. Kopczewska, K. (2025). Analysing local spatial density of human activity with quick density clustering (QDC) algorithm. Computers, Environment and Urban Systems, 119, 102289. https://doi.org/10.1016/j.compenvurbsys.2025.102289
14. Gupta, V., Gupta, S. K., Shetty, A. (2024). Fractal-based supervised approach for dimensionality reduction of hyperspectral images. Computers & Geosciences, 193, 105733. https://doi.org/10.1016/j.cageo.2024.105733
15. Ge, J., Liao, Y., Zhang, B. (2024). Resistance distances and the Moon-type formula of a vertex-weighted complete split graph. Discrete Applied Mathematics, 359, 10–15. https://doi.org/10.1016/j.dam.2024.07.040

16. Song, J., Daley, T., McNeany, J., Kamaleswaran, R., Stecenko, A. (2024). 682 A machine learning approach with silhouette scoring of continuous glucose monitoring enables repeat measure assessment of changes in the glycemic profile in cystic fibrosis. Journal of Cystic Fibrosis, 23, S381–S382. https://doi.org/10.1016/s1569-1993(24)01520-0
17. Ros, F., Riad, R., Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing, 528, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043
18. Passarella, R., Noor, T. M., Arsalan, O., Adenan, M. S. (2024). Anomaly detection in commercial aircraft landing at SSK II airport using clustering method. Aerospace Traffic and Safety, 1 (2-4), 141–154. https://doi.org/10.1016/j.aets.2024.12.004
19. Marto Hasugian, P., Mawengkang, H., Sihombing, P., Efendi, S. (2025). Development of distance formulation for high-dimensional data visualization in multidimensional scaling. Bulletin of Electrical Engineering and Informatics, 14 (2), 1178–1189. https://doi.org/10.11591/eei.v14i2.8738
20. Zhu, M.-X., Lv, X.-J., Chen, W.-J., Li, C.-N., Shao, Y.-H. (2022). Local density peaks clustering with small size distance matrix. Procedia Computer Science, 199, 331–338. https://doi.org/10.1016/j.procs.2022.01.040