Water Science & Technology

Water Science & Technology Vol 89 No 7, 1757 doi: 10.2166/wst.2024.087

Using unsupervised learning to classify inlet water for more stable design of water reuse in industrial parks

Kan Chen[a,b], Xiaofei Shi[a], Zhihao Zhang[a], Shijun Chen[b], Ji Ma[b], Tong Zheng [a,*] and Leonardo Alfonso [c]

[a] School of Environment, Harbin Institute of Technology, 73 Huanghe Road, Harbin, Heilongjiang, China
[b] Suzhou Sujing Environmental Engineering Co., Ltd, 2 Weixin Road, Suzhou, Jiangsu, China
[c] IHE Delft Institute of Water Education, Westvest 7, 2611AX Delft, The Netherlands
*Corresponding author. E-mail: zhengtong@hit.edu.cn

TZ, 0000-0002-7054-0980

ABSTRACT

The water reuse facilities of industrial parks face the challenge of managing a growing variety of wastewater sources as their inlet water. Typically, this clustering outcome is designed by engineers with extensive expertise. This paper presents an innovative application of unsupervised learning methods to classify inlet water in Chinese water reuse stations, aiming to reduce reliance on engineer experience. The concept of 'water quality distance' was incorporated into three unsupervised learning clustering algorithms (K-means, DBSCAN, and AGNES), which were validated through six case studies. Of the six cases, three were employed to illustrate the feasibility of the unsupervised learning clustering algorithm. The results indicated that the clustering algorithm exhibited greater stability and excellence compared to both artificial clustering and ChatGPT-based clustering. The remaining three cases were utilized to showcase the reliability of the three clustering algorithms. The findings revealed that the AGNES algorithm demonstrated superior potential application ability. The average purity in six cases of K-means, DBSCAN, and AGNES were 0.947, 0.852, and 0.955, respectively.

Key words: AGNES, DBSCAN, inlet water classification, K-means, unsupervised learning, water reuse

HIGHLIGHTS

- Validation of three unsupervised learning clustering algorithms in six real cases.
- L2 distance improved based on water quality in the clustering algorithms.
- Practical engineering applications consideration.
- Machine clustering compared with human and ChatGPT's clustering.

1. INTRODUCTION

Water scarcity has emerged as an increasingly pressing issue due to the continuous development of the global economy and the impact of climate change (Lee & Jepson 2020; Verhuelsdonk *et al.* 2021). The Asia-Pacific region, in particular, faces the challenge of limited water resources (UNESCO 2009). To ensure their sustainable development, industrial parks, as key contributors to economic growth, must address water scarcity (Bauer *et al.* 2019). Wastewater reuse and management play vital roles in alleviating the demand for water resources and addressing water quality degradation (Dairi *et al.* 2023). By efficiently treating and reusing wastewater, industrial parks can mitigate water shortages and meet various needs, including irrigation and industrial activities (Lahlou *et al.* 2021). Moreover, this approach can lead to reduced energy consumption and greenhouse gas emissions, contributing to environmental sustainability (Chang *et al.* 2017).

A water reuse station, also known as a wastewater reclamation plant, refers to a facility that treats wastewater for reuse, typically for the regeneration of water resources. These stations employ various technologies and processes to clean wastewater to specific water quality standards, which are then used for purposes, such as flushing, irrigation, industrial production, or other applications. Modern industrial parks increasingly prioritize clustering enterprises within the same industrial chain and centralizing wastewater treatment and reuse to achieve economies of scale. However, this approach

also introduces complexities in managing multiple streams of wastewater received by water reuse stations, as these wastewater streams serve as the inlet water for reuse systems.

Figure 1 illustrates a typical water system for an industrial park's water reuse facility. The inlet water of the water reuse station consists of the effluent from each factory's self-built sewage station, the wastewater from the domestic sewage plant within the factory, as well as low-pollution effluents discharged by public infrastructures in the park (such as cooling water and desalination stations). It may also include rainwater collected from drainage ponds.

There have been some studies on water source classification based on water quality. Sa'ad *et al.*'s (2022) research has demonstrated that classifying different qualities of wastewater into multiple grades in industrial areas can lead to higher economic benefits for water reuse stations, while also reducing freshwater consumption. In Elsayed *et al.*'s (2022) research, employing various machine learning methods to classify influent wastewater in sewage treatment plants facilitates swift adjustment of subsequent operational parameters, thereby promoting optimized operation and effective management of wastewater treatment facilities. Therefore, it is crucial to scientifically classify inlet water in reuse stations. Nonetheless, the existence of numerous sources of wastewater poses a significant challenge in designing an efficient and cost-effective treatment system.

However, numerous studies have been conducted on water quality classification in the field of environmental engineering, with a predominant focus on the categorization of water quality levels within river basins. Most of these studies employ supervised learning using labeled data, such as the use of multi-layer perceptron-K nearest neighbors (MLP-kNN) for water quality classification in Thailand's Wang River (Northep *et al.* 2020) and fusion algorithms for water quality classification in the Yellow River Delta wetlands (Zhao *et al.* 2023). Although the rise of artificial intelligence (AI) technology has led to its application in simulating process conditions in certain water reuse units (Amitesh *et al.* 2023), there is a lack of research on the classification of inlet water in reuse stations. The classification of inlet water represents an unsupervised learning process for computers, given that each water reuse station possesses unique characteristics and lacks numerical boundaries for inlet water clusters. Clustering algorithms serve as a type of unsupervised learning classification process and have been implemented across various domains. Not only are they used extensively for processing a large number of samples and data, such as image recognition analysis (Zhou *et al.* 2023), electronic text analysis (Pauletic *et al.* 2019), and network

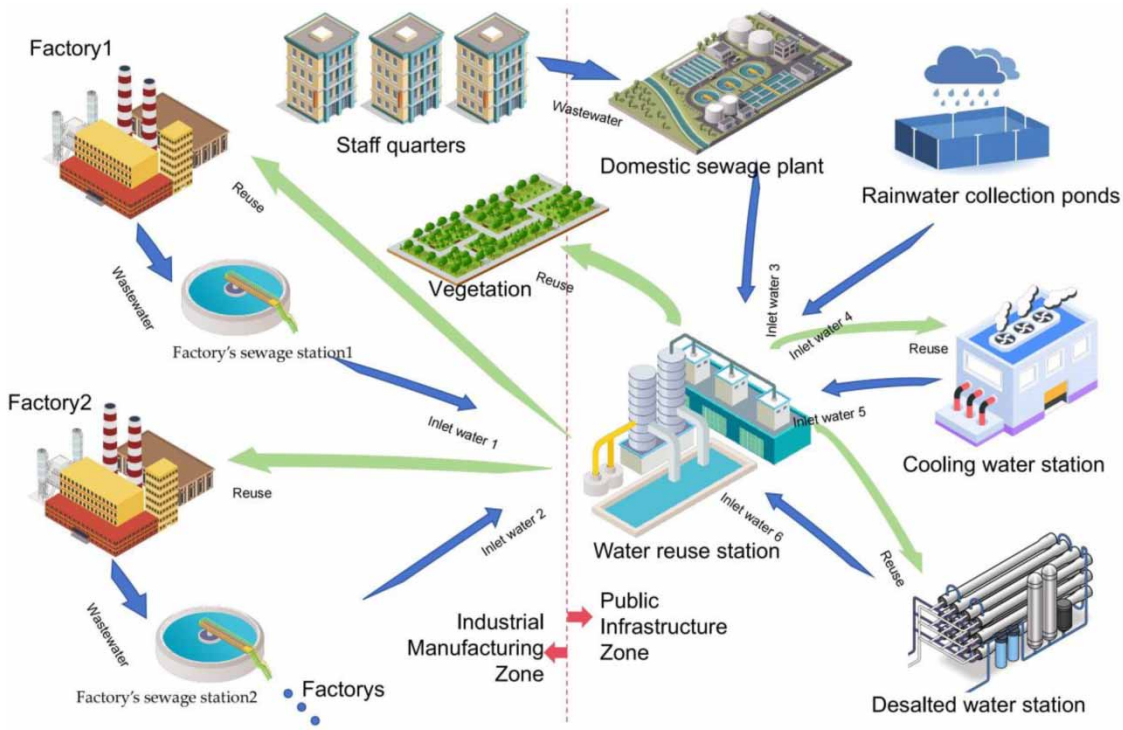


Figure 1 | A typical water system of the reuse water stations' inlet water.

data analysis (Qiao *et al.* 2023), but they also find applications in specific environmental protection domains, including climate change analysis (Biabiany *et al.* 2020), identification of air pollution sources (Zhang & Yang 2022), water resources management decisions (Sahraei & Asadzadeh 2021; Kumar *et al.* 2022), and environmental protection equipment design (Lee *et al.* 2020).

This paper explores the application of unsupervised learning clustering algorithms to classify treated wastewater in industrial parks, aiming to reduce construction and operation costs while unveiling objective mathematical principles. By employing unsupervised learning methods to analyze extensive datasets, engineers can discern patterns and make more precise and rapid classification, regarding the behavior of diverse water sources within a treatment system.

2. MATERIALS AND METHODS

2.1. Datasets

When engineers design water reuse stations, the classification of inlet water quality is determined based on the analysis of water quality data. Therefore, when applying unsupervised learning algorithms for clustering, it is essential to carefully design the clustering feature set and select pertinent and concise indexes of water quality. Water reuse stations primarily treat treated wastewater, and the discharge indexes for first- and second-class pollutants outlined in the 'Integrated Wastewater Discharge Standard (GB8978-1996)' commonly serve as the discharge requirements of the upstream sewage station for the reuse water system. Additionally, other crucial indexes, such as conductivity, hardness, and alkalinity, play a vital role in membrane system design. These two components collectively constitute 37 indexes in the clustering feature set, as listed in 'Appendix A. Water Quality Distance Indexes and Index Benchmarks'. The data collected for a specific case only include a subset of the 37 indicators, excluding other pollutants that are either absent or do not require removal at the reuse station, or have no impact on the reuse system.

This study analyzed the inlet water of six reuse stations. The inlet water data from the water reuse stations of three industrial parks were utilized as the preliminary research cases, which were from three industrial parks: JM, JDD, and FRX. Through the analysis of these three cases, we established a definitive methodology for determining the optimal cluster number of various unsupervised clustering algorithms. Furthermore, the machine-generated clustering results were assessed in comparison to those produced by human experts and a large-scale language model (ChatGPT), thereby validating the feasibility of unsupervised learning clustering methodologies. Subsequently, three additional testing cases (ZL, HL, and FM) were employed to further validate the effectiveness of the unsupervised clustering method. And the inlet water quality for each park was listed in 'Appendix B, C, D, E, F, G' (in supplementary materials, indexes not listed in the tables were considered to be 0). A concise description of each industrial park is provided in Table S1 (in supplementary materials).

2.2. Clustering algorithms

To cluster the inlet water data from the three industrial parks, three unsupervised learning clustering algorithms were used: K-means for prototype clustering (Jain & Dubes 1988), DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for density clustering (Ester *et al.* 1996), and AGNES (AGglomerative NESting) for condensed hierarchical clustering (Kaufman & Rousseeuw 2009). The algorithms were implemented using Python programming language, utilizing the sklearn module in Python for calculation processes.

2.3. Clustering performance evaluation

Appropriate indexes should be utilized to evaluate clustering results, which can be classified into internal and external indexes. External indexes rely on known labels to assess the quality of clustering, while internal indexes solely utilize information derived from the clustering process itself to evaluate its quality. The number of clusters in unsupervised learning, particularly the K-means algorithm, was determined using external indexes in this study.

2.3.1. Internal indexes

In this study, several internal indexes (Liu *et al.* 2010) were employed, including the sum of squares due to error (SSE) (Thorndike 1953), Silhouette coefficient (SC) (Rousseeuw 1987), and Calinski–Harabasz index (CH) (Caliński & Harabasz 1974). The SSE holds particular significance in the K-means algorithm, as it aids in identifying the optimal model by obtaining the minimum SSE for a given clustering number. A smaller SSE value indicates a higher degree of internal aggregation within each cluster in the results. However, an excessively high level of aggregation will result in an excessive number of clusters.

Therefore, the 'elbow method' based on SSE was employed to determine the optimal number of clusters in the K-means algorithm by identifying the inflection point where SSE exhibits a significant change with the number of clusters.

The SC is an evaluation method that combines cohesion and separation, producing values within the range of [−1, 1]. A value closer to 1 indicates a better degree of cohesion and separation. On the other hand, the CH coefficient serves as another index to assess the clustering effect. A larger CH value indicates a better clustering effect. In this study, the CH coefficient will be employed to evaluate the clustering effect in cases where the SSE and SC coefficients fail to provide precise judgments.

2.3.2. External indexes

External indexes, which relied on the true labels, encompass purity, NMI (normalized mutual information), and ARI (adjusted rand index). The true labels for this study were the clustering results determined by engineers in real project cases for each water plant, which were utilized to validate the model. Purity assesses the proportion of correctly clustered samples, with a value range of [0, 1], where a higher value indicates greater accuracy. NMI quantifies the degree of correlation between clustering results and real labels, with larger values indicating a higher correlation. ARI adjusts the Rand Index to mitigate the issue of random clustering results yielding values close to 0, with a range of [−1, 1], where higher values indicate better performance.

The purity was calculated as follows:

$$\text{Purity}\ (C, W) = \frac{1}{N}\sum_{j}\max_{k}|c_j \cap w_k|$$

where N was the number of samples, $C = \{c_1, c_2, \ldots, c_k\}$ represented the clustering results, and $W = \{w_1, w_2, \ldots, w_k\}$ represented the true labels (clustering results of engineering applications). Purity $\in [0,1]$, and a value closer to 1 denoted better clustering result accuracy.

The calculation methods for NMI and ARI were obtained from the references (de Souto *et al.* 2012).

2.3.3. Dendrogram

The dendrogram in AGNES serves as a valuable tool for visualizing the clustering process and summarizing the hierarchy of clusters through a tree diagram. The x-axis represents samples, with similar ones connected by straight lines and vertical lengths indicating the distance between them. Greater differences in height signify greater dissimilarity between samples. By identifying natural splitting points where the dendrogram branches, this tool can assist in determining an appropriate number of clusters, avoiding arbitrary selection.

2.3.4. Economic indexes

This study examined six real-world engineering cases and compared the results of machine clustering and human clustering with the final engineering application outcomes. The sum of construction and operation costs per year was computed by a budget engineer and served as the baseline economic index (E_0). The same budget engineer also calculated the sum of construction and operation costs of both machine- and human-generated clusters, which were used as the economic index for each clustering process (E_i). The ratio of these two economic indexes (E_0/E_i) was employed as an economic factor (EF) to evaluate cluster effectiveness. An EF value of 1 was assigned to represent optimal performance in terms of engineering application outcomes, while higher values indicated more cost-effective clustering results.

2.4. Design of the water quality based distance

First, it was crucial to standardize each index in order to mitigate the clustering distance bias toward indexes with larger values resulting from variations in their respective ranges. In this paper, we introduced the concept of 'water quality distance' by employing water quality benchmarks. The discharge index value of the first-level Integrated Wastewater Discharge Standard of China (GB8978-1996) was used as the benchmark for most indexes, as it typically represented the discharge requirement for the upstream sewage station of the water reuse system. However, for hardness, alkalinity, and conductivity, which were not covered by the discharge standard, we used the empirical limited index value of the inlet water of the reverse osmosis membrane in the water reuse systems as the index benchmarks. The final index benchmarks were listed in 'Appendix A: Water Quality Distance Indexes and Index Benchmarks'.

To calculate the water quality distance, we obtained a 'base value' for each index by dividing the actual index value by the corresponding index benchmark (excluding pH). The water quality distance was then calculated based on this 'base value' using the following method:

$$L_{WQ} = \sqrt{\sum_{i=1}^{n} (c_1(i)/d(i) - c_2(i)/d(i))^2}$$

where $c_1(i)$ represents the index value of the first inlet water for index i, $c_2(i)$ represents the index value of the second inlet water for index i, and $d(i)$ is the corresponding index benchmark. For pH (Index 14), the calculated values of c_1 and c_2 are as follows:

$$c_{1\text{ or }2}(14) = |\text{pH} - 7|$$

The water quality distance can be defined as follows: first, the original water quality data are standardized by dividing it with the corresponding water quality benchmarks to obtain the 'base value'. Subsequently, this 'base value' is utilized in the L_2 distance calculation method, and the resulting L_2 distance is normalized by the water quality benchmarks to determine the water quality distance.

The clustering algorithms used in this study employed the 'water quality distance' as the basis for calculating the distances between data points.

3. RESULTS AND DISCUSSION

3.1. Results of K-means

3.1.1. Internal indexes analysis of K-means

In the context of the K-means clustering algorithm, the initial step for three engineering cases involved computing SSE, SC, and CH. By employing the 'elbow method' based on the SSE curve in conjunction with SC and CH values, a comprehensive assessment was conducted to determine the optimal number of clusters. Figure 2 illustrates the evolution of SSE, SC, and CH as the number of clusters increases.

Based on Figure 2, in the JM case, SSE decreased significantly as the number of clusters increased from 1 to 2, and then remained stable. SC reached its maximum value at 2, while CH did not change significantly between 2 and 3. The optimal number of clusters was determined to be 2, using the 'elbow method'.

In the JDD case, SSE exhibited a significant decrease as the number of clusters increased from 1 to 3, with no further reduction observed thereafter. SC attained its maximum value at two clusters, while CH was higher for three clusters than for 2. Thus, the elbow method may be employed subjectively to determine three clusters.

In the FRX case, the SSE exhibited a sharp decrease as the number of clusters increased from 1 to 3 and subsequently stabilized beyond that point. The SC and CH both reached their highest values at three clusters, thus leading us to select this as the optimal number of clusters for our algorithm.

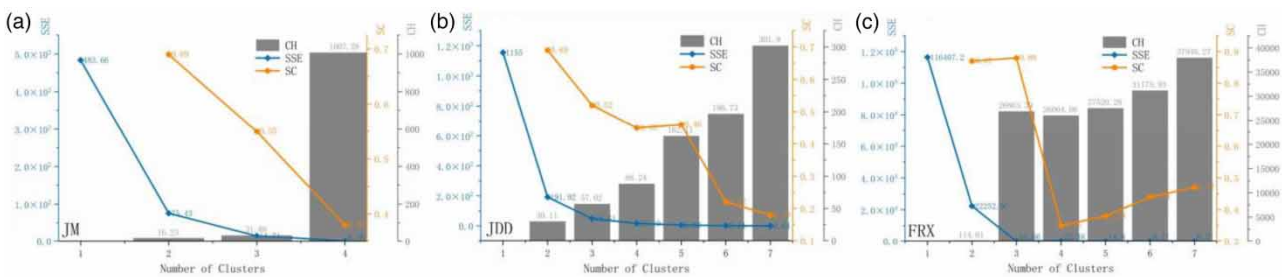


Figure 2 | Evolution of SSE, SC, and CH of K-means in three cases with an increasing number of clusters. (a) JM, (b) JDD, and (c) FRX.

3.1.2. Clustering results of K-means

The final clustering results were obtained using two clusters for JM, and three clusters for both JDD and FRX. The detailed clustering results are shown in Tables S2–S4 of supplementary materials, where real labels stand for clustering results of engineering applications. Different colors and numbers represent different clusters. And the SN refers to the wastewater source code provided in the appendix, specifically indicating the code of the inlet water.

3.1.3. External indexes analysis of K-means

The K-means clustering results were compared with the engineering application results of the three cases, and the clustering performance was evaluated using three external indexes (purity, NMI, and ARI) (as shown in Figure 3). Based on this evolutionary graph, we can analyze whether using the previously determined number of clusters yields the most accurate results.

As depicted in Figure 3, the JM case exhibited maximization 1 of all three external indexes at a cluster number of 2, showing perfect cluster prediction performance.

In the JDD case, the clustering result could be acceptable, with 0.88 of purity, 0.81 of NMI, 0.71 of ARI, which were all highest at a cluster number of 3. Therefore, the choice of three clusters for the JDD case in the previous section was deemed reasonable.

In the FRX project, the K-means algorithm demonstrated excellent clustering performance with a data volume of 19 water source groups, achieving the highest levels of purity (0.95), NMI (0.902), and ARI (0.982).

3.1.4. Conclusion of K-means

The aforementioned analysis demonstrated that K-means was a classical clustering algorithm that yields favorable clustering outcomes. However, it also presented certain limitations. The most challenging aspect of K-means lies in the subjective process of determining the number of clusters through the elbow method, which could be considered a weakness of this algorithm. Nevertheless, with meticulous selection of the number of clusters, K-means had the potential to generate highly precise clustering results.

3.2. Results of DBSCAN

3.2.1. Parameters selection

In the context of the DBSCAN clustering algorithm, the determination of two parameters, namely min_samples and epsilon distance (eps), is required.

The min_samples parameter represents the minimum number of sample points required to classify a data point as a core point within a cluster. It is important to note that this value should not exceed half of the total samples, as higher values tend to result in fewer clusters. The epsilon distance was utilized to determine the density of data points by creating data distances around each point. As previously mentioned in the calculation process for water quality distance, these distances were calculated based on multiples of index benchmarks and could be interpreted as representing the multiple relationships between pollutants in two sets of samples.

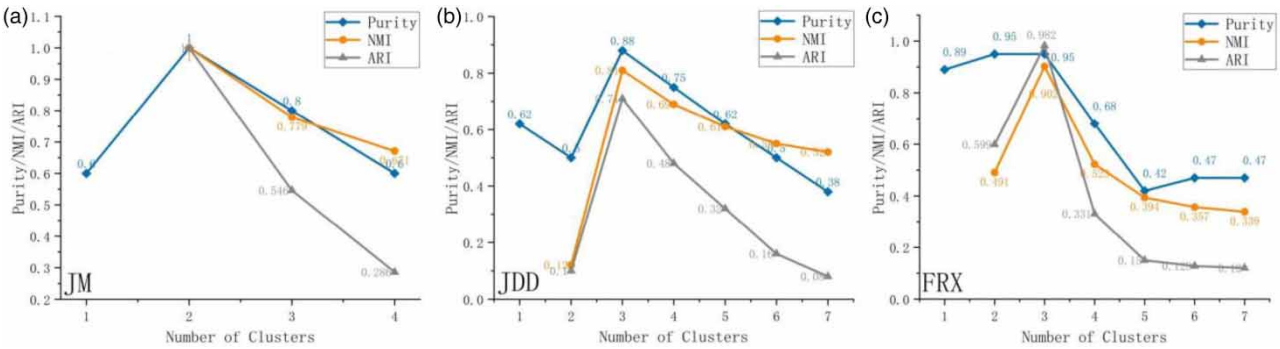


Figure 3 | Evolution of purity, NMI, and ARI of K-means in three cases with an increasing number of clusters. (a) JM, (b) JDD, and (c) FRX.

In order to achieve the best economy of scale, it is advisable to keep the number of clusters as low as possible. Therefore, we selected min_samples as an integer slightly less than half the total number of samples, specifically JM = 2, JDD = 3, and FRX = 9 for our three projects, respectively.

For epsilon distance, we believe that there is no meaningful merging between the two sets of water quality data if the distance exceeds three times the standard deviation. Therefore, we uniformly choose epsilon distance as 3.

3.2.2. Clustering results of DBSCAN

According to the aforementioned criteria for parameter selection, Tables S2–S4 present the corresponding clustering outcomes.

The best results have been achieved in all cases, thus resulting in all external indexes (purity, NMI, ARI) being equal to 1.

3.2.3. Conclusion of DBSCAN

The optimal results can be achieved by setting the min_samples value to its maximum within 50% of the sample size (2, 3, and 9 for each case, respectively), and setting eps to 3.

When designing for water reuse, it is important to consider both project cost and the scale economy effect of equipment. As a result, the actual design process tends to have fewer clusters. Therefore, the parameter selection scheme proposed in this study not only provides excellent results but also has practical significance.

3.3. Results of AGNES

In this section, we will evaluate the effectiveness of AGNES in classifying inlet water from three water reuse stations and generate dendrograms to assist designers in determining the final clustering outcomes.

3.3.1. Dengrograms of AGNES

For the JM case, and the distance between clusters 0, 1, and clusters 3, 2, 4 was the largest in the dendrogram. The optimal number of clusters was, therefore, determined to be 2 (Figure 4).

Similarly, in the case of JDD, the dendrogram indicates that clustering into two or three clusters is feasible. However, due to the exceptionally high hardness levels of water sources 3, 4, and 5, they need to be treated separately for hard removal. Therefore, based on subjective judgment from the auxiliary manual, we have selected a total of three clusters here (Figure 4).

In the FRX case, the dendrogram indicated that the 19 inlet water samples could be classified into either 2 or 3 clusters. However, due to a greater vertical distance, it was deemed more appropriate to group them into two distinct clusters.

3.3.2. Clustering results of AGNES

When the number of clusters is 2 in JM, 3 in JDD, and 2 in FRX, the detailed clustering results are shown in Tables S2–S4 of supplementary materials.

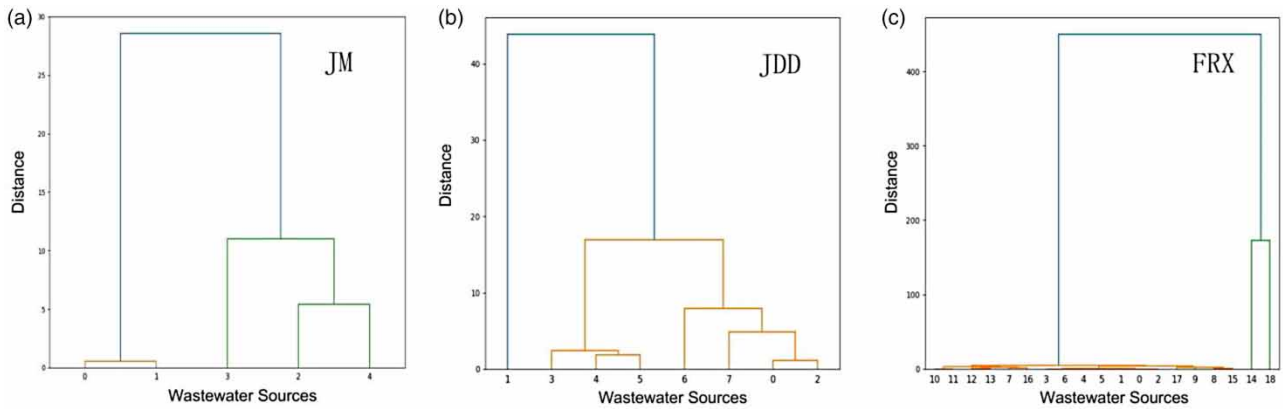


Figure 4 | Diagram of the evolution of SC of AGNES with the increase of the number of clusters and the dendrogram of AGNES in three cases. (a) JM, (b) JDD, and (c) FRX.

3.3.3. External indexes analysis of AGNES

Based on the tree diagram in the previous section, we compared the AGNES clustering results with the engineering application results for different cluster numbers in the three cases and analyzed the evolution of the clustering results under different cluster numbers using three external indexes: purity, NMI, and ARI, and drew an evolution diagram. Based on this evolution diagram, we can analyze whether the most accurate results were obtained by using the previously determined cluster number.

In Figure 5, for the JM case, all three external indexes were maximized when the number of clusters was 2, indicating that AGNES clustering with two clusters was reasonable.

For the JDD case, clustering into three clusters using AGNES resulted in high accuracy with 0.88 purity, 0.81 NMI, and 0.71 ARI. However, this approach has inherent limitations for achieving good clustering results for data with arbitrary shapes. The algorithm mistakenly identified water source JDD01 as a sample with a prominent shape due to its high ammonia nitrogen and chemical oxygen demand (COD) levels.

For the FRX case, all three external indexes were maximized when the number of clusters was 2, indicating that AGNES clustering with two clusters was reasonable.

3.3.4. Conclusion of AGNES

Based on the aforementioned analysis, AGNES is capable of providing design engineers with clustering suggestions through the output of a dendrogram. However, significant deviations may occur in certain cases due to arbitrary shapes or data noise in real-world data. Nevertheless, when the selection of clustering numbers is reasonable, AGNES' clustering results are deemed acceptable with a purity rate exceeding 0.88.

3.4. Comparison of human–machine results

In this section, we present the clustering results of three reuse water stations generated independently by four human experts with varying levels of experience in water treatment design (15, 12, 10, and 4 years of experience, denoted as M15, M12, M10, and M4, respectively) and ChatGPT, a state-of-the-art language model that has gained recent popularity. These results were compared against those obtained from three clustering algorithms. The problem descriptions provided to both the ChatGPT and human experts are identical in Chinese, with the Q&A results included in the supplementary materials. Due to the instability of ChatGPT's output, we performed 10 iterations and selected the clustering results based on their highest frequency.

After analyzing external indexes, we compared the clustering results of each algorithm with those obtained by four human experts. The comparison is presented in Figure 6.

The comparison of human and machine results in JM is presented in Figure 6(a). In the JM case, where there was less variation in the type of inlet water, both the computer algorithms and most human engineers demonstrated good results. This was also reflected in the economic index of the construction and operation costs, which followed the same pattern. The ChatGPT, along with the less experienced engineer M4, yielded suboptimal results in terms of purity, NMI, ARI, and EF scores of 0.8, 0.78, 0.55, and 0.8, respectively.

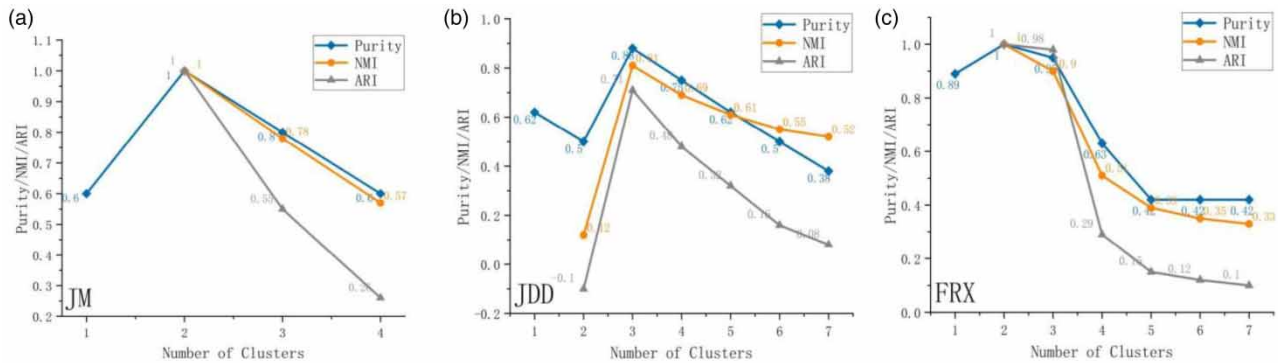
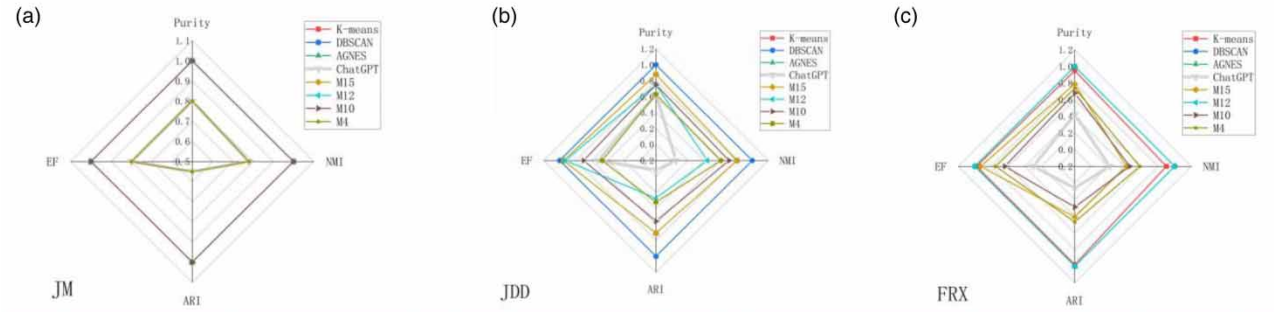


Figure 5 | Evolution of purity, NMI, and ARI of AGNES in three cases with an increase in the number of clusters. (a) JM, (b) JDD, and (c) FRX.



Indexes	K-means	DBSCAN	AGNES	ChatGPT	M15	M12	M10	M4
Purity	1	1	1	0.8	1	1	1	0.8
NMI	1	1	1	0.78	1	1	1	0.78
ARI	1	1	1	0.55	1	1	1	0.55
EF	1	1	1	0.8	1	1	1	0.8

Indexes	K-means	DBSCAN	AGNES	ChatGPT	M15	M12	M10	M4
Purity	0.88	1	0.88	0.63	0.88	0.75	0.75	0.63
NMI	0.81	1	0.81	0.05	0.81	0.44	0.71	0.61
ARI	0.71	1	0.71	-0.08	0.71	0.27	0.56	0.32
EF	0.95	1	0.95	0.52	0.95	0.93	0.7	0.47

Indexes	K-means	DBSCAN	AGNES	ChatGPT	M15	M12	M10	M4
Purity	0.95	1	1	0.42	0.79	1	0.68	0.74
NMI	0.9	1	1	0.22	0.42	1	0.46	0.58
ARI	0.98	1	1	0.06	0.41	1	0.29	0.47
EF	0.98	1	1	0.3	0.93	1	0.64	0.75

Figure 6 | Comparison of human–machine results in three cases. (a) JM, (b) JDD, and (c) FRX.

The comparison of human and machine results in the JDD case is presented in Figure 6(b). In this case, there were eight categories of inlet water, and the accuracy of human clustering was lower than that of machine clustering, with only M15 achieving relatively ideal results. The purity's extreme difference of artificial clustering increased to 0.25, indicating the instability of artificial clustering performance. Simultaneously, the optimal purity of artificial clustering achieved a mere 0.88, with similar trends observed in NMI, ARI, and EF metrics. Specifically, these optimal values decreased by 0.19, 0.29, and 0.05, respectively, compared to the JM case study results. Consequently, as data volume increases, the efficacy of clustering performed by an individual engineer within a limited timeframe diminishes significantly. However, the accuracy of machine results was consistent. DBSCAN achieved 100% purity, with perfect NMI, ARI and EF, demonstrating the advantages of this algorithm. And the optimal outcomes of manual classification can also be achieved by K-means and AGNES algorithms. In addition, it is noteworthy that the performance of ChatGPT exhibited a significant deterioration in comparison to the results obtained in the JM case, ranking consistently at the lowest positions for both NMI and ARI.

Notably, K-means, AGNES, and all four human engineers believed that the second inlet water (JDD01) should be treated separately (shown in Table S3). However, in reality, despite having relatively high levels of ammonia nitrogen and COD (shown in Appendix C), it was feasible to be incorporated into other wastewater for treatment, and it did not affect the subsequent treatment processes. As we analyzed earlier, the difference in clustering results was caused by the arbitrary shape of water quality data in high-dimensional space or the existence of 'data noise' (Duan *et al.* 2007). Therefore, the ability of DBSCAN to detect noise led to the best clustering effect.

The comparison of human and machine results in JDD is presented in Figure 6(c). In this case, a substantial amount of data was analyzed. Although the M12 engineer achieved the best results with all evaluation indicators being 1, the extreme difference of purity in artificial clustering results further increased to 0.32 compared to the previous two cases, further highlighting the substantial impact of increasing data volume on the stability of artificial clustering performance. However, computer clustering results remained stable and even surpassed human clustering results with DBSCAN and AGNES achieving all the highest evaluation indicators. It can be concluded that when dealing with a large amount of data, machine clustering results are more stable and accurate than human clustering results. In this case, the results obtained from ChatGPT were inferior compared to previous instances, with lower external index values. Although ChatGPT served as a useful guide, caution should be exercised in relying solely on it for our design.

Based on the analysis of three cases, the K-means clustering algorithm achieved an average purity of 0.943, while DBSCAN achieved a perfect purity of 1.00. AGNES had an average purity of 0.96, and ChatGPT had an average purity of 0.617. In contrast, the average purity of manual classification by four human engineers ranged from 0.723 to 0.916, with an overall average of 0.832. Overall, the machine clustering algorithms outperformed both the human engineers and ChatGPT in terms of purity.

The visual comparison results of each individual external index and the EF for human–machine interaction in three different cases are presented in Figure 7.

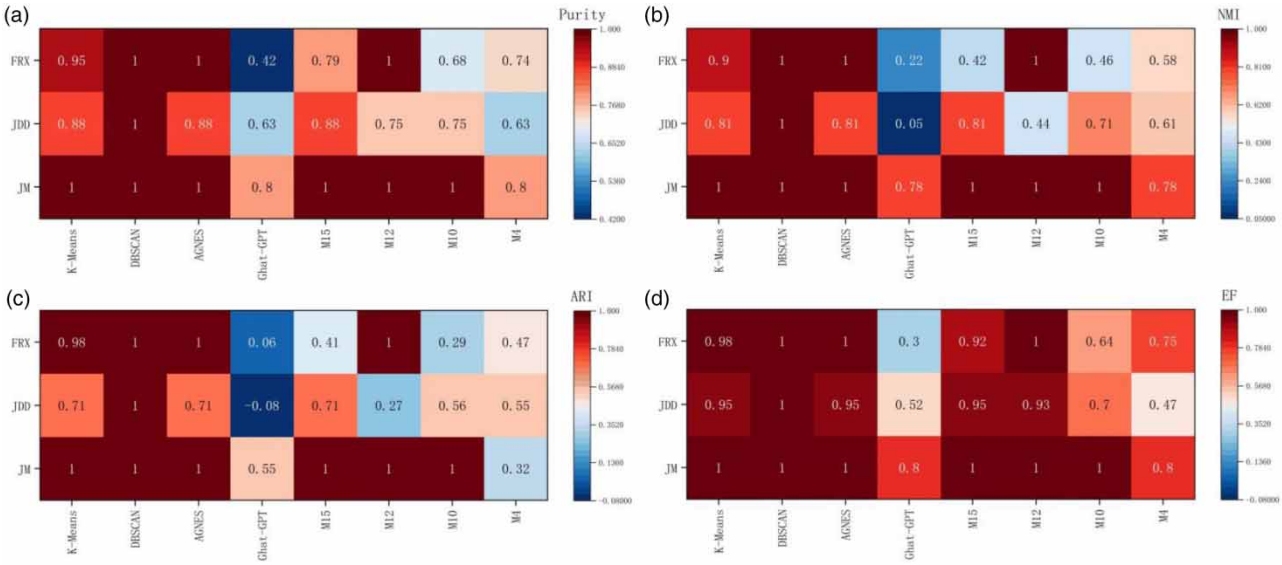


Figure 7 | Human–machine comparison of external indexes. (a) Purity, (b) NMI, (c) ARI, and (d) EF.

The figure above illustrates the outcomes of machine clustering algorithms in the left three columns, where the extreme differences in purity, NMI, ARI, and EF values across these three projects were 0.12, 0.19, 0.29, and 0.05, respectively. In contrast, four engineers working on these projects achieved extreme differences of 0.37 in purity, 0.58 in NMI, 0.73 in ARI, and 0.53 in EF, respectively. As a reference, ChatGPT achieved extreme differences of 0.38, 0.73, 0.63, and 0.5 in these indexes, respectively.

Based on these findings, it can be concluded that machine clustering results were significantly more stable than those obtained by humans or ChatGPT. The analysis conducted on the aforementioned three projects in this article suggests that machine clustering accuracy was at least equivalent to human classification. The utilization of three unsupervised learning algorithms for inlet water classification at reuse stations was deemed feasible.

In addition, through the above three cases, we also found that the 'elbow method' with SC and CH coefficients could effectively obtain the number of clusters in the K-means algorithm. In DBSCAN algorithm, by selecting min_samples as the maximum value within 50% of the sample size and eps as 3, better clustering results could be obtained. AGNES algorithm can yield robust clustering results through the dendrograms.

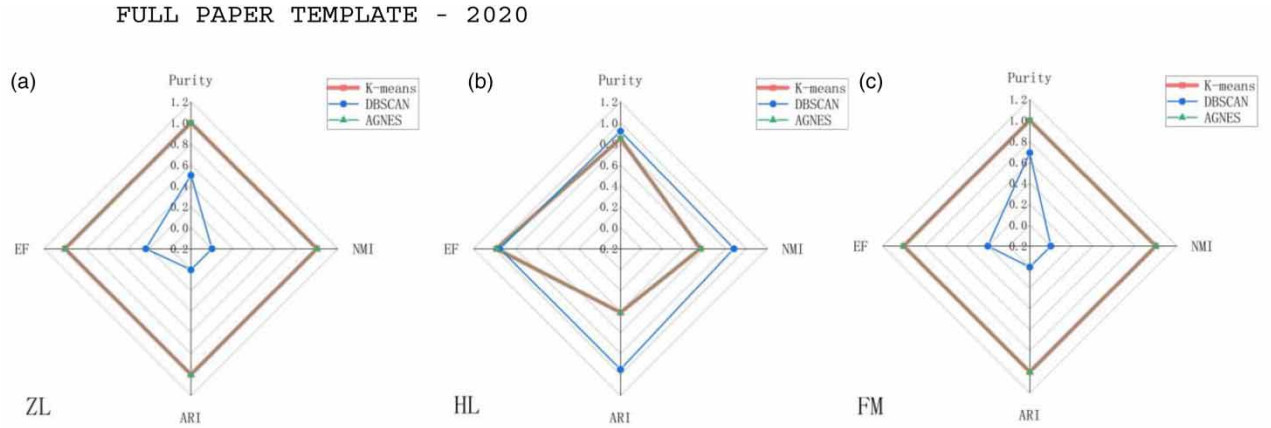
3.5. Three testing cases

The unsupervised learning clustering analysis method mentioned earlier was applied to the inlet water classification design of water reuse stations in three additional industrial parks. This application aimed to further validate the reliability of the unsupervised learning clustering algorithm. For brief introductions to these three cases, please refer to Table S1, and detailed data on water quality and process design are available in Appendix E–G.

The external indexes of the clustering results for these three testing cases are illustrated in Figure 8.

From Figure 8, it was observed that in these three cases, DBSCAN exhibited unstable performance (no better results were obtained after returning the hyperparameters). Specifically, in the ZL and FM cases, the purity of DBSCAN was only 0.5 and 0.69, respectively. However, in the HL case, DBSCAN performed well with a purity of 0.92. On the other hand, both k-means and AGNES showed similar and stable performance across the three cases. In particular, the purity of ZL and FM reached 1, while the purity of HL was 0.85 for both k-means and AGNES.

On the other hand, through the AGNES' dendrogram (Figure 9), it was observed that in the HL case, despite the considerable distance between water source 11 and the other water sources, when it was classified separately, the overall classification could still be divided into three clusters (Cluster 1: 11, Cluster 2: 6 and 12, Cluster: others). In this condition, the results from AGNES were completely consistent with practical engineering applications (as shown in Figure S3b, purity, NMI, ARI all reaching 1). This result could also be achieved using K-means when selecting a cluster number of 3 (as shown in



Indexes	K-means	DBSCAN	AGNES
Purity	1	0.5	1
NMI	1	0	1
ARI	1	0	1
EF	1	0.23	1

Indexes	K-means	DBSCAN	AGNES
Purity	0.85	0.92	0.85
NMI	0.56	0.88	0.56
ARI	0.41	0.95	0.41
EF	0.98	0.95	0.98

Indexes	K-means	DBSCAN	AGNES
Purity	1	0.69	1
NMI	1	0	1
ARI	1	0	1
EF	1	0.2	1

Figure 8 | Comparison of human–machine results in three testing cases. (a) ZL, (b) HL, and (c) FM.

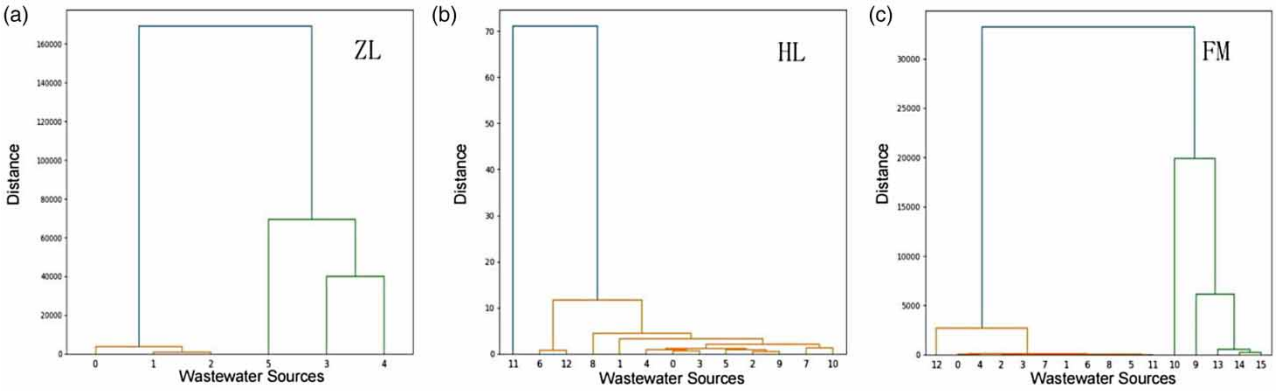


Figure 9 | Diagram of the evolution of SC of AGNES with the increase of the number of clusters and the dendrogram of AGNES in three testing cases. (a) ZL, (b) HL, and (c) FM.

Figure S2). However, the 'elbow method' from K-means' SSE curve and a higher SC index guided us to choose a cluster number of 2 (as shown in Figure S1), leading to the inability to conclude a cluster number of 3. This highlighted a significant advantage of AGNES in guiding clustering decisions: it could visually provide hidden clustering possibilities through the hierarchical clustering dendrogram.

Finally, we compared the purity of the three unsupervised learning clustering algorithms across six cases and compared them with the clustering results of similar studies, as shown in Table 1.

From the data presented in Table 1, it was evident that both K-means and AGNES demonstrated superior stability, achieving average purity indicators of 0.947 and 0.955, respectively. These values had reached the level of similar research. While DBSCAN exhibited commendable performance in scenarios featuring noisy data, such as the JDD project, its efficacy appeared limited in the ZL and FM cases, consequently impacting its overall performance.

4. CONCLUSION AND PERSPECTIVES

4.1. Conclusion

This paper innovatively established the distance calculation method of clustering algorithm based on water quality. The machine clustering results using water quality distance were satisfactory and could serve as a basis for inlet water classification for reuse purposes.

Table 1 | Comparison with similar research results

Research contents	Algorithms	Purity			References
		Minimum	Maximum	Average	
This study	K-means	0.850	1.000	0.947	
	DBSCAN	0.500	1.000	0.852	
	AGNES	0.850	1.000	0.955	
Document analysis	K-means	0.927 (maximum)			Abuaiadah (2016)
Population census	DBSCAN	0.970 (maximum)			Kumar *et al.* (2018)
Data streams analysis	AGNES	1.000 (maximum)			Sangma *et al.* (2022)

This paper also innovatively applied three unsupervised learning clustering algorithms to address the inlet water classification problem in a water reuse station. The results revealed that these three clustering algorithms exhibited exceptional performance when contrasted with outcomes derived from artificial clustering and ChatGPT-based clustering methodologies. K-means demonstrated favorable clustering results with an average purity of 0.947. However, it required to determine the number of clusters relied on internal indexes. DBSCAN performed the best on clustering tasks with noisy data and could automatically cluster using selected parameters. It achieved an average purity of 0.852 by setting the minimum number of samples to the maximum value within 50% of the sample size and an epsilon distance of 3. AGNES possesses an highest average purity of 0.955. In addition, AGNES, visually depicted through dendrograms, offered a significant advantage by providing the designer with a more extensive range of potential clustering possibilities. This advantage enhances the potential applicability of AGNES for future research endeavors.

4.2. Limitations and perspectives

This study investigates the application of three unsupervised learning algorithms for inlet water classification in water reuse systems, yielding promising results. However, certain limitations exist, including a limited quantity of research data and incomplete index benchmarks that may not encompass all scenarios encountered in water reuse engineering. These limitations also point to directions for future research: first, further validation of algorithm effectiveness across a broader range of engineering cases is necessary; second, there is a need to establish a more comprehensive set of index benchmarks that can adapt to a broader array of scenarios.

In addition, the classification of the inlet only represents the initial stage of the design of the water reuse treatment system. The subsequent design process also includes process selection, equipment selection, energy consumption calculation, investment prediction, etc. In the water reuse system, the research on clustering algorithms for inlet classification is the cornerstone for realizing the fully AI-driven design of future water treatment projects.

ACKNOWLEDGEMENTS

The authors wish to thank all engineers of Sujing Engineering Department II for the cooperation in this project and Qingyun Xu, Yigang Chen, Qiankui Fang, and Xiang Qiu for their dedication to this project.

FUNDING

This research was supported by National Key Research and Development Program of China (No.2018YFC0408001).

AUTHOR CONTRIBUTIONS

K. C. performed the investigation, data analysis, and writing; X. S. performed the methodology; Z. Z. performed the resources and formal analysis; S. C. performed the validation; J. M. performed the supervision; T. Z. performed the conceptualization and funding acquisition.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abuaiadah, D. 2016 Using bisect k-means clustering technique in the analysis of Arabic documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **15** (3), 1–13.

Amitesh, D. D., Jyoti, G., Rekhete, C., Dubey, S. & Prajapati, A. K. 2023 Optimization of electrocoagulation process for the removal of chromium from simulated water using the response surface methodology. *Journal of Water Chemistry and Technology* **45** (5), 429–439.

Bauer, S., Behnisch, J., Dell, A., Gahr, A. & Wagner, M. 2019 Water reuse fit for purpose by a sustainable industrial wastewater management concept. *Chemie Ingenieur Technik* **91** (2), 1429–1479.

Biabiany, E., Bernard, D. C., Pagé, V. & Paugam-Moisy, H. 2020 Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles. *Computers & Geosciences* **145**, 104612.

Caliński, T. & Harabasz, J. 1974 A dendrite method for cluster analysis. *Communications in Statistics* **3** (1), 1–27.

Chang, J., Lee, W. & Yoon, S. 2017 Energy consumptions and associated greenhouse gas emissions in operation phases of urban water reuse systems in Korea. *Journal of Cleaner Production* **141**, 728–736.

Dairi, S., Mrad, D., Bouamrane, A., Djebbar, Y. & Abida, H. 2023 Wastewater reclamation and reuse trends in Algeria: Opportunities and challenges. *Doklady Earth Sciences* **511** (2), 753–760.

de Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V. & Costa, I. G. 2012 A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *2012 Brazilian Symposium on Neural Networks*, pp. 49–54.

Duan, L., Xu, L., Guo, F., Lee, J. & Yan, B. 2007 A local-density based spatial clustering algorithm with noise. *Information Systems* **32** (7), 978–986.

Elsayed, A., Siam, A. & El-Dakhakhni, W. 2022 Machine learning classification algorithms for inadequate wastewater treatment risk mitigation. *Transactions of the Institution of Chemical Engineers. Process Safety and Environmental Protection Part B* **159**, 1224–1235.

Ester, M., Kriegel, H. P., Sander, J. & Xu, X. 1996 A density-based algorithm for discovering clusters in large spatial databases with noise. *In kdd* **96** (34), 226–231.

Jain, A. & Dubes, R. 1988 *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, USA.

Kaufman, L. & Rousseeuw, P. J. 2009 *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken, NJ, USA.

Kumar, R. S., Rao, S. S. & Srinivasrao, P. 2018 An efficient clustering approach using DBSCAN. *Helix* **8** (3), 3399–3405.

Kumar, M., Tiwari, N. K. & Ranjan, S. 2022 Application of machine learning methods in estimating the oxygenation performance of various configurations of plunging hollow jet aerators. *Journal of Environmental Engineering* **148** (11), 04022070.

Lahlou, F. Z., Mackey, H. R. & Al-Ansari, T. 2021 Wastewater reuse for livestock feed irrigation as a sustainable practice: A socio-environmental-economic review. *Journal of Cleaner Production* **294**, 126331.

Lee, K. & Jepson, W. 2020 Drivers and barriers to urban water reuse: A systematic review. *Water Security* **11**, 100073.

Lee, S., Kim, J., Hwang, J., Lee, E. J. & Heo, T. Y. 2020 Clustering of time series water quality data using dynamic time warping: A case study from the Bukhan river water quality monitoring network. *Water* **12** (9), 2411.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. 2010 Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pp. 911–916.

Northep, K., Srijiranon, K. & Eiamkanitchat, N. 2020 Water quality classification using data mining techniques: A case study on Wang River in Thailand. In *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, Kitakyushu, Japan, pp. 1–8.

Pauletic, I., Prskalo, L. N. & Bakaric, M. B. 2019 An overview of clustering models with an application to document clustering. In *2019, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1659–1664.

Qiao, C., Brown, K. N., Zhang, F. & Tian, Z. 2023 Adaptive asynchronous clustering algorithms for wireless mesh networks. *IEEE Transactions on Knowledge and Data Engineering* **35** (3), 2610–2627.

Rousseeuw, P. J. 1987 Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65.

Sa'ad, S. F., Alwi, S. R. W., Lim, J. S. & Abd Manan, Z. 2022 The economic study of centralised water reuse exchange system in the industrial park considering wastewater segregation. *Computers & Chemical Engineering* **164**, 107863.

Sahraei, S. & Asadzadeh, M. 2021 Cluster-based multi-objective optimization for identifying diverse design options: Application to water resources problems. *Environmental Modelling & Software* **135**, 104902.

Sangma, J. W., Sarkar, M., Pal, V., Agrawal, A. & Yogita, 2022 Hierarchical clustering for multiple nominal data streams with evolving behaviour. *Complex & Intelligent Systems* **8** (2), 1737–1761.

Thorndike, R. L. 1953 Who belongs in the family? *Psychometrika* **18** (4), 267–276.

UNESCO 2009 The United Nations World Water Development Report 2009, United Nations, New York, NY, USA.

Verhuelsdonk, M., Glas, K. & Parlar, H. 2021 Economic evaluation of the reuse of brewery wastewater. *Journal of Environmental Management* **281**, 111804.

Zhang, L. & Yang, G. 2022 Cluster analysis of PM2.5 pollution in China using the frequent itemset clustering approach. *Environmental Research* **204**, 12209.

Zhao, X., Wu, Y., Lan, Y., Guo, Y., Shao, Y. & Hu, Z. 2023 Application of adaptive weight fusion algorithm to wetland classification in Yellows River Delta. In: *International Conference on Geographic Information and Remote Sensing Technology (GIRST 2022)*, Vol. 12552, pp. 740–746.

Zhou, W., Wang, L., Han, X., Wang, Y., Zhang, Y. & Jia, Z. 2023 Adaptive density spatial clustering method fusing chameleon swarm algorithm. *Entropy* **25** (5), 782.

First received 27 November 2023; accepted in revised form 6 March 2024. Available online 19 March 2024