ISSN: 1648-6897 / eISSN: 1822-4199

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of ENVIRONMENTAL ENGINEERING & LANDSCAPE MANAGEMENT

2025
Volume 33
Issue 1
Pages 110–117
https://doi.org/10.3846/jeelm.2025.22953

AUTOMATIC MONITORING OF TREATED WATER RELEASED FROM WASTEWATER TREATMENT PLANTS USING MODEL-BASED CLUSTERING WITH DENSITY ESTIMATION

Sheik Faritha BEGUM[1]✉, K LOKESHWARAN[2]

[1]*Department of CSE, PSNA College of Engineering and Technology, Tamil Nadu, India*
[2]*Department of CSE (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, India*

Highlights:

- in this paper a novel integrated approach is proposed for continuously evaluating the performance of wastewater treatment plants (WWTPs), with a focus on treated wastewater quality assessment and reuse of treated water for beneficial purposes like irrigation, aquarium, groundwater recharge, and in river water discharge based on pollution level in treated water;
- a model-based clustering with density estimation is implemented to generate the non-overlapped clusters to categorize the clusters by assuming that each data object originate from the mixture of underlying probability distributions;
- cluster analysis using the Euclidean distance resulted in three clusters labeled under a specified category of water polluted: non-polluted, lightly polluted, highly polluted or slightly polluted. Water quality parameters like suspended solids (SS) have been considered for the analysis;
- further, it motivates the reuse of treated water for beneficial purposes like irrigation, aquarium, groundwater recharge, and river water discharge based on pollution level in treated water.

Article History:

- received 02 May 2024
- accepted 13 November 2024

Abstract. One of the most promising efforts to fight against the water scarcity threat is to reuse the treated water released from WasteWater Treatment Plants (WWTP). The objective of this paper is to propose an integrated approach for continuously evaluating the performance of wastewater treatment plants (WWTPs), with a focus on treated wastewater quality assessment and reuse of treated water for beneficial purposes like irrigation, aquarium, groundwater recharge, and in river water discharge based on pollution level in treated water. This paper implemented a model-based clustering with density estimation to generate the non-overlapped clusters to categorize the clusters. Cluster analysis using the Euclidean distance resulted in three clusters labeled under a specified category of water polluted: non-polluted, lightly polluted, highly polluted or slightly polluted. Unlike standard clustering algorithms like K-means, hierarchical that produce optimized clusters in statistical terms that deviate from naturally categorized clusters, model-based clustering with density estimation operates on the assumption that each data object originates from the mixture of underlying probability distributions. Water quality parameters like suspended solids (SS) have been considered for the analysis. Our experimental results conclusively show the polluted levels of wastewater from WWTP using a model-based clustering approach. The Dataset used in this work has been derived from the wastewater treatment plant located in Manresa, a town of 100,000 inhabitants near Barcelona (Catalonia). The plant treats a flow of 35,000 m^3/day, mainly domestic wastewater, although wastewater from industries located inside the town is received in the plant too. In this research, the plant's behavior over 527 days are under consideration. Model-based density clustering algorithm discovers 3 clusters, with half lying in size range of 14–89 and a maximum size of 352. With the help of natural clusters generated, our results show that out of 445 days, in 352 days, the treated water is almost non-polluted. By this, we can assess the performance of the wastewater treatment plant.

Keywords: clustering, density estimation, pollution, water quality.

✉Corresponding author. E-mail: ***sfarithaphd@gmail.com***

1. Introduction

Clustering problems arise in numerous disciplines, including medicine, biology, and environmental issues. Several approaches have been proposed for defining quality criteria for solutions (clusters) and the methods to interpret the solution obtained. During cluster analysis, one wishes to partition objects into groups based on the given features of each entity so that the groups are homogenous and well-separated. It is also known as the unsupervised classification of patterns and has many applications in different areas. Several clustering methods can perform partitions but need help during cluster characterization (Jain et al., 1999). The clustering objective is to find the co-clusters to be merged, choosing the sets of objects with the same properties. An appropriate variable required for cluster-

ing has to be selected at the beginning of the chosen clustering process so that it has to differentiate the data segments perfectly, relatively satisfying the objective functions. Generally, one should avoid using many numbers of clustering variables since it stops the iteration process very quickly, as all the data segments are no longer dissimilar. Our problem comes under the above case. Since the objective is to categorize or label the cluster in terms of pollution, the selected variable should significantly impact the pollution. Cluster heads can be used as service repository to send or publish service request (Buvana & Suganthi, 2015).

Finally, the selection of clustering variables depends on data availability or the objective function to satisfy (Mukhopadhyay et al., 2012). Next, by selecting a particular clustering procedure, one can determine how clusters have to be formed. It involves optimizing, such as minimizing the within-cluster variance or maximizing the distance between the objects or clusters. Constrained multiobjective problems can be solved by modifying the dominace in dense population (Deb et al., 2002). But fuzzy algorithms leads to formation of uninterested optimal clusters which is not suited for formation of well seperated clusters (Dunn, 1973). Our paper uses model-based clustering along with density estimation to achieve our objective. In model-based clustering, the data d is viewed as taking from mixture density $p(d) = \sum_{c=1}^{G} bcpc(d)$ where p_c is the probability density function of the observations in group c, and b_c is the probability that observation comes from the c^{th} mixture component ($b_c \in (0, 1)$. The Gaussian distribution consistently models every component. Finite mixture models provide a basic statistical approach for clustering. Each single component probability corresponds to an individual cluster, and comparisons can be made for the models which differ in the component distributions or several components using some statistical criteria. Model estimation for the data is done by the clustering process, which allows non-overlapping clusters, resulting in a probabilistic clustering that measures the uncertainty of observations belonging to the components of the finite mixture. In the final step, the solution has to be interpreted by labeling and defining the generated clusters. This work has been done by assessing the values of water quality parameters. A similarity measure cam also be used to identify and measure the separative property of the formed clusters (Davies & Bouldin, 1979). Model based clustering can also be used to produce overlapped clusters in discriminat analysis and multivariate analysis (Fraley & Raftery, 2007).

This paper's principal objective is to suggest an integrated method for the automated monitoring and evaluation of treated water quality from wastewater treatment plants (WWTP). With the use of clustering techniques, this method seeks to classify the pollutant levels of treated wastewater in order to make it easier to potentially reuse treated water for beneficial uses including irrigation, aquariums, groundwater recharging, and river discharge.

The study focuses on the following in particular: Using suspended solids (SS) as the primary water quality metric, different clustering techniques (K-means, Hierarchical, and Model-based clustering) are implemented to classify treated water into distinct contamination levels.

Assessing how well various clustering techniques provide categories that accurately represent the quality of treated water, such as non-contaminated, lightly polluted, and highly polluted.

Recommending useful uses for reusing cleaned water in accordance with the contamination level found by the cluster analysis.

2. Related work

Anaokar et al. (2018) analyzed the efficiency of nine wastewater characteristics from 6 wastewater treatment plants. The values were compared with limits suggested by Central Pollution Control Board of India. Quantitative and qualitative approaches have been used for ranking the water quality parameters. Ashfaq et al. (2010) systematically evaluated the performance of common effluent treatment plant (CETP). The Collected samples are examined for the following parameters Ammoniacal Nitrogen, COD pH, BOD, TSS, and TDS. Results state that major pollutants can be reduced after effective treatment. Baki et al. (2019) developed a prediction model to estimate the value of Biochemical Oxygen Demand (BOD) by using the values of other water quality parameters like suspended solids, electrical conductivity, total nitrogen, pH, and chemical oxygen demand. Mazhar et al. (2019) treated wastewater by three biological treatments such as sequential, aerobic, and anaerobic, to predict the water quality of industrial wastewater from paper and pulp. Along with this, simulation modeling by Mamdani Fuzzy Logic (MFL) is associated with some selected parameters. The treated water was irrigated to determine its phytotoxic effects. Nadiri et al. (2018) predicted the water quality indices like Biological Oxygen Demand (BOD), total suspended solids, temperature, pH and chemical oxygen demand (COD) of Tabriz wastewater treatment plant (TWWTP), introduced a predictive ensemble model as supervised committee FL (SCFL). In the testing process, the mean absolute percentage error (MAPE) for BOD, COD, and TSS varies from 10% to 13% for each FL model (Nourani et al., 2018). Predicts the performance of Nicosia wastewater treatment using three artificial intelligence models Support Vector Machine (SVM), Feed Forward Neural Network (FFNN), and Adaptive Neuro-Fuzzy Inference System (ANFIS). The performance will be calculated using chemical oxygen demand, biological oxygen demand, and total nitrogen. An Ensemble model has been developed with daily data to improve the prediction ratio (Padalkar & Kumar et al., 2018). In Common effluent treatment plants (CETPs), observations have shown that the removal efficiencies and reliability of each measure (BOD, COD, and TSS) varied. It can be enhanced by optimising treatment procedures, particularly primary clari-

flocculators and aeration tanks, which are crucial components of any CETP, and releasing effluent with hydraulic and organic loading to the CETP in accordance with requirements. Sharghi et al. (2019) explain that for selecting water effluent parameters of WWTP for Artificial Neural Network (ANN) modeling self-organization map as AI-based clustering method was employed. Also, two other models using principle component analysis (PCA) and the variables found using the mutual information (MI) measure that is not linear were developed and compared with the ANN model. In cluster analysis, automatic classification using different cluster algorithms is discussed to provide the guidelines in order to choose the algorithm for application (Kaufman & Rousseeuw, 2009). Based on the distance between centroids, data set and distance measure, a validity function was proposed to identify the compactness of clusters (Xie & Beni, 1991).

In order to address the pressing problem of water shortage, this study suggests an integrated method for monitoring treated water quality in wastewater treatment facilities (WWTPs) in order to regularly analyze their performance. The study classifies treated water into natural groupings according to pollution levels using model-based clustering with density estimation, providing a more accurate assessment than conventional techniques. Results from a WWTP in Manresa, Catalonia, show that the plant is effective because treated water is almost completely free of pollutants for a significant amount of the evaluation period. This research contributes to resilience against water scarcity concerns by supporting sustainable water management practices and increasing the possibilities for recycling treated water for purposes such as agriculture and groundwater recharge.

3. Water quality assessment

Reuse of treated wastewater has been demonstrated as a dependable option water asset, which can constitute a noteworthy segment of coordinated water assets administration and give a powerful answer for adapting to water shortage conditions. Researchers have significantly improved water quality from wastewater treatment plants (WWTP) in recent years by building many models. After implementing a model, the effort has to be made to assess the quality of water yield by that model. This paper aims to assess the water quality parameters and to classify them where the treated water can be effectively used.

River Pollution Index (RPI):

The "River Pollution Index," or RPI for short, is a comprehensive index that serves as the foundation for the Environmental Pollution Act's (E.P.A.) current assessment of river quality (Diller, 2013). RPI is a coordinated marker that determines a waterway's degree of contamination. The concentration of four water quality parameters – dissolved oxygen (DO), suspended solids (SS), biochemical oxygen demand (BOD5), and ammonia nitrogen (NH_3-N) is used to determine the index value. Below are the RPI computation and comparative baselines (Table 1) (Diller, 2013).

We have used the dataset from the plant located in Manresa, a town in Catalonia with 100,000 residents, close to Barcelona. The 35,000 m^3/day flow of wastewater that the facility handles is mostly from homes, though it also receives wastewater from businesses in the area. Eight of the system variables which are quality indicators are measured on a daily basis at various plant locations, including the input (P1), the pretreatment stage (P2), the biological reactor information (P3), and the plant's water output (P4). This results in a set of 38 values each day, nine of which are performance percentages. This study has taken into account the plant's behaviour over a period of 527 days.

This paper assesses water quality parameters like suspended solids (SS), pH, temperature, turbidity, total dissolved solids (TDS), and biochemical oxygen demand (BOD). Water quality studies that include total suspended solids (TSS) are necessary for wastewater treatment operations since TSS is a crucial indication of environmental health. Large amounts of suspended organic and inorganic particles are present in wastewater, which needs to be eliminated by filtration, settling/flotation, or screening. before environmental discharge. High amounts of TSS can reduce the quality of the water in the receiving environment if they are not adequately eliminated by treatment. Because of the suspended solids' absorption of light, the water's temperature rises and its oxygen content falls, making it unsuitable for aquatic life. The objective is to classify the water released from the plant at the end of the treatment process. All the instances in the dataset are collected from the sensors daily. domain has been described as an ill-structured domain (Lichman, 2013).

Number of occurrences: 527

Number of Features: 07

Attribute Information:

Every attribute is continuous and numerical.

1. PH-S (output pH)
2. DBO-S (output biological demand of oxygen)

Table 1. RPI baselines comparison (Diller, 2013)

Water quality parameter	Non polluted	Lightly polluted	Moderately polluted	Highly polluted
Ammonia Nitrogen (NH_3-N) mg/L	NH_3-N ≦ 0.50	0.50 > NH_3-N ≦ 0.99	1.00 ≦ NH_3-N ≦ 3.00	NH_3-N > 3.00
Dissolved Oxygen (DO) mg/L	DO ≧ 6.5	6.5 < DO ≧ 4.6	4.5 ≧ DO ≧ 2.0	DO < 2.0
Suspended Solids (SS) mg/L	SS ≦ 20.0	20.0 > SS ≦ 49.9	50.0 ≦ SS ≦100	SS > 100
Biochemical Oxygen Demand (BOD5) mg/L	BOD5 ≦ 3.0	3.0 < BOD5 ≦ 4.9	5.0 ≦ BOD5 ≦ 15.0	BOD5 > 15.0

3. DQO-S (output chemical demand of oxygen)
4. SS-S (output suspended solids)
5. SSV-S (output volatile suspended solids)
6. SED-S (output sediments)
7. COND-S (output conductivity)

Method of Sampling: Over the course of 527 days, wastewater samples were taken at regular intervals from the Manresa WWTP. These samples show wastewater intakes from both residential and commercial sources.

Investigated Parameters: In addition to other pertinent indicators including pH, temperature, turbidity, total dissolved solids (TDS), and biochemical oxygen demand (BOD), Suspended solids (SS) are the main water quality parameters that were examined in this study.

Period of Investigation: The investigation period extended 527 consecutive days, providing for a full analysis of the plant's performance under varied conditions.

4. Methods

The K-means clustering algorithm and Hierarchical clustering algorithm is explained in this section. And our proposed model-based clustering algorithm with density estimation is explained in detail.

4.1. K-means clustering

K-means is an unsupervised clustering algorithm that finds groups within the data (Burkardt, 2009) K-means clustering seeks to divide a collection of observations (x_1; x_2; ... ; x_n) into a set of k clusters ($\leq n$) such as $S = S_1$; S_2; ... ; S_k in order to minimise the within-cluster sum of squares. Each observation is a d-dimensional vector which is defined as the sum of distance functions between each cluster point and the k centres. K-means' objective functionis to identify

$$\arg\min s \sum k\, i = 1 \sum x \in S_t \left\| x_i - c_i \right\|^2, \quad (1)$$

where c_i is the centroid of points in S_i.

The K-means clustering algorithm's sensitivity to initial centroids means that various initial centroids may produce different clusters, which is one of the method's main limitations. A preprocessor can also be used in Kmeans algorithm to produce well seperated clusters (Baridam, 2012). To optimize the categorization of clusters in numerical dataset, K-means algorithm is well suited (Begum et al., 2016).

4.2. Hierarchical clustering

The following represents a general agglomerative algorithm for hierarchical clustering:

Algorithm 1

Step 1. Determine all inter-object dissimilarities.

Step 2. Generate clusters from the two closest clusters or objects.

Step 3. Redefine differences that exist between the new cluster and other objects or clusters, leaving the other interpoint dissimilarities unaltered.

Step 4. Return to Step 2 until every item is found in a single cluster.

Step 1 often calls for $O(N^2)$ computations, or $N(N - l)/2$ inter-object dissimilarities. Although the number of variables influences the amount of time needed for calculations, they are typically thought of as constants for each given data collection. In Steps 2 and 3, it might be worthwhile to think about maintaining a sorted list of all the dissimilarities that are being considered; this would require an initial sorting time of $O(N^2 \log N)$ and subsequent updating time. If not, any execution of Step 2 will need $O(N^2)$ time. The hierarchy calls for $N - 1$ iterations (Steps 2, 3, and 4) in order to form the $N - 1$ cluster at most. Step 3 includes applying the Lance-Williams combinatorial formula and can be completed in O(N) time. If k is any other object or cluster and the recently merged objects or clusters are indexed by $/$ and y, the formula is:

$$D(i + j, k) = a(i)d(i, k) + a(j)d(j, k) + bd(i, j) + c|d(i, k) - d(j, k)|, \quad (2)$$

where the values of a, b and c depend on the clustering strategy. In addition to uniting what may initially appear to be multiple operations, the aforementioned recurrence formula also helps with the study of subordinate questions, such as the circumstances under which inversions (or reversals) occur. i.e.

$$d(i + j, k) \geq d(i, j) \text{ for some } i, j, k \quad (3)$$

(Murtagh, 1983; Batagelj, 1981).

The main disadvantage of hierarchical clustering is that it often won't provide the best solution. When missing data is found, it will swamp up and work poorly with mixed data types. Hierarchical clustering is not suitable for huge data sets.

4.3. Proposed algorithm

We adopted model-based clustering, which assumes that each data object originates from a combination of the probability distributions at work. Every cluster has a distinct distribution associated with it. The likelihood of the expression data is maximized to estimate the parameters of each distribution, or cluster. Figure 1 shows the flow of our algorithm, and the following process explains how non-overlapping clusters are generated.

4.3.1. Preprocessing and selecting attributes for water quality assessment

Since the experiment's dataset were derived from the urban wastewater treatment plant in real-time, the data has to be preprocessed and filtered by removing missing values and omitting attributes like categorical values, which could not be considered during clustering. Also, the attribute for density estimation is fixed in this stage. In this experiment an attribute named date has been omitted. An attribute named SS-S (Suspended solids) has been selected for density estimation since this attribute is mainly used in water quality assessment.

4.3.2. Model-based clustering with density estimation

Due to recent advances in strategies and programming for model-based clustering and to the interpretability of the outcomes, clustering procedures based on likelihood models are progressively favored over heuristic techniques. The clustering process evaluates a model for the information that allows for overlapping clusters that measure the instability of perceptions belonging to components of the mixture. The subsequent grouping model can likewise be utilized for other vital issues in multivariate analysis, including density estimation and discriminant analysis. While participation in segments is vital in clustering, the mixture likelihood (Banfield, & Raftery, 1993) itself, or its incentive at given focuses, is the focus of concentration in density estimation (Silverman, 1986). The fitted probability can be utilized to uncover or think about information patterns. Model-based clustering computes the Bayesian Information Criterion (BIC) (Schwarz, 1978), considering the number of components in the model, the data dimensions, and the maximized log-likelihood for the model. The maximum log-likelihood with a penalty on the number of model parameters is the BIC. It permits the examination of models with varying parameterizations and contrasting quantities of clusters. When all is said is done, the bigger the estimation lower the BIC, the more clusters and robust the model's evidence. Here, BIC chooses a two-component mixture of Gaussian variables with a similar change. The parameter appraisals can be perused from the outline yield using R gives mixing probabilities, means, and variances.

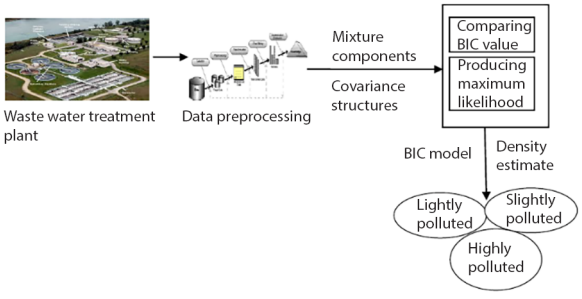

Figure 1. Architecture diagram of wastewater treatment data using density-based clustering

4.3.3. Optimization in terms of yielding natural clusters

As we mentioned earlier, the objective of our experiment refers to generating more natural clusters. Though familiar with clustering algorithms like K-means, hierarchy produces optimized clusters in statistical terms that deviate from naturally categorized clusters. In our experiment, we mean that natural clusters should possess an equal range of water quality parameters and hence can be categorized each cluster under some specific category of water quality assessment in terms of the pollution level. We have proved this assessment in Section 5.

5. Results and discussion

All the clustering algorithms discussed in this paper are implemented with RStudio on a Windows platform. All algorithms have been implemented on the benchmark dataset in which wastewater samples were taken at regular intervals from the Manresa WWTP. In this dataset, each data object refers to the output of the wastewater treatment plant each day. The experimental results for the average sum of intracluster distance are calculated on the dataset discussed in Section 2 using three different clustering methods provided in Tables 2 and 3. The results are collected over a single run generating 3 clusters by fixing K = 3.

After generating the clusters, the value of suspended solids of data objects in each cluster is noted, out of which minimum and maximum values are analyzed. By using K-means clustering, when K = 3, At cluster number 1: Minimum and Maximum value of suspended solids are 98 and 238; at cluster number 2: Minimum and Maximum value of suspended solids are 30 and 84; at cluster number 3: Minimum and Maximum value of suspended solids are 6 and 29. The results given in Table 2 are obtained over 512 data points supervised over a single attribute (Suspended Solids). By Applying Hierarchical Clustering, when 3 clusters are generated, at cluster number 1: Minimum and Maximum value of suspended solids are 10 and 35; at cluster number 2: Minimum and Maximum value of suspended solids are 26 and 54; and at cluster number 3: Minimum and Maximum value of suspended solids are 53 and 238.

As per the comparison, the baseline of RPI discussed in Section 2, the range of suspended solids in non-polluted level water is less than 20; in lightly polluted, it is between 20 and 50; and in highly polluted, it is greater than 50 and less than 100. However, analyzing the results obtained over the K-means and Hierarchical clustering methods in Table 2, each cluster has data objects which do not fit into any above-said ranges. For example, when using K-means Clustering at cluster number 2, the minimum and maximum values of suspended solids are 30 and 84. We cannot conclude whether the data objects (particular days) the treated water is lightly polluted or highly polluted.

Table 2. Range of clusters using K-means and hierarchical

	K-means			Hierarchical		
	1	2	3	1	2	3
Min	98	30	6	10	26	53
Max	238	84	29	35	54	238

Model-based clustering uses the probabilistic approach to find the clusters with the data objects limited to a fixed range, which once more depends on the first decision of parameter density estimation. A model-based clustering algorithm can generate non-overlapped clusters in which the minimum and maximum value of the quality parameter lies within the fixed range. Table 3 summarizes

Table 3. Range of clusters using model-based clustering with density estimation

Cluster no.	n	Mean	SD	Median	IQR	Min	Max	Pollution level
1	352	16.9346	4.2040	17	6	6	25	Slightly
2	89	33.1910	6.6825	32	10	26	53	Lightly
3	14	92.1428	47.2454	76	35	54	238	Highly

the results of Waste Water Treatment Plant data, including the different statistical measures like mean, median, and standard deviation of all data objects with three generated clusters. Model-based density clustering algorithm discovers 3 clusters, with half lying in size range of 14–89 and a maximum size of 352.

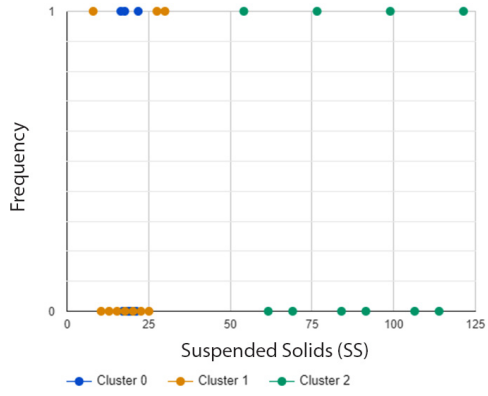

Figure 2. Distribution of suspended solids by clusters (model based clustering)

Figure 2 shows that the Cluster 1 has the largest number of observations (352 days) and is concentrated in the lower range of SS values. The majority of data points in this cluster fall within the range of 6 to 25. Cluster 2 has a smaller number of observations (89 days) compared to Cluster 1, and the SS values are distributed within a narrower range of 26 to 53. Cluster 3 has the smallest number of observations (14 days) and is concentrated in the higher range of SS values. The majority of data points in this cluster fall within the range of 54 to 238. The overall distribution of SS values is skewed to the right, indicating that there are a few observations with very high SS values, while the majority of observations are in the lower range. The three clusters clearly show distinct patterns in the distribution of SS values, suggesting that the clustering algorithm effectively identified groups of observations with similar characteristics.

Cluster 1 likely represents treated wastewater with relatively low levels of suspended solids, indicating effective treatment and potential suitability for reuse in applications like irrigation or groundwater recharge. Cluster 2 represents treated wastewater with moderate levels of suspended solids, requiring further treatment or careful consideration for reuse. Cluster 3 represents treated wastewater with high levels of suspended solids, indicating potential issues with the treatment process or the need for more stringent treatment before reuse.

Cumulative distribution function (CDF) for the suspended solids (SS) is shown in Figure 3 represents values in the clusters generated using model-based clustering. The CDF curve for Cluster 1 shows a rapid increase, indicating that a large proportion of observations in this cluster have relatively low SS values. The curve plateaus at a lower value, suggesting that there are fewer observations with higher SS values. The CDF curve for Cluster 2 starts at a higher value than Cluster 1, indicating that a larger proportion of observations in this cluster have higher SS values. The curve plateaus at a slightly lower value than Cluster 1, suggesting that there are fewer observations with extremely high SS values. The CDF curve for Cluster 3 starts at the highest value, indicating that most observations in this cluster have very high SS values. The curve plateaus at a relatively low value, suggesting that there are fewer observations with lower SS values. The model-based clustering approach successfully identified distinct clusters based on suspended solids (SS) values, demonstrating its ability to group data points with similar characteristics. The clusters likely represent different levels of water quality, ranging from low to high SS concentrations.

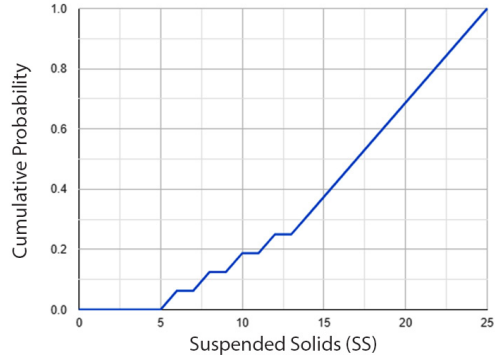

Figure 3. Cumulative distribution function of suspended solids (model based clustering)

6. Conclusions

In assessing the water quality in the wastewater treatment plant (WWTP), an effort has been made to use the dataset derived from daily measures of the urban wastewater treatment plant. An optimized clustering method is identified by employing different clustering methods, which yields the natural clusters with the optimal number of clusters.

The plant under study is situated in Manresa, a 100,000-person town close to Barcelona, in Catalonia. The

35,000 m^3/day flow of wastewater that the facility handles is mostly from homes, though it also receives wastewater from businesses in the area. Eight of the system variable which are quality indicators are measured on a daily basis at various plant locations, including the input (P1), the biological reactor's information (P3), the plant's water output (P4), and after pretreatment (P2). This results in a set of 38 values each day, nine of which are performance percentages. This study has taken into account the plant's behavior over a period of 527 days. In this paper, we presented a novel approach to checking the pollutant level of treated water and assessing the performance level of wastewater treatment plants. Model-based clustering with density estimation is used to calculate the wastewater treatment plant (WWTP) performance. This model will categorize the treated water into three levels: Lightly polluted, Moderately Polluted, and Highly Polluted. Based on the number of data objects clustered in each category, the performance of WWTP is evaluated clearly. In model-based clustering with density estimation, overlapping clusters have been completely avoided. By doing so, each cluster can be categorized in terms of water quality mentioning the level of pollution in the water. The inspiration for this algorithm is from forming various groups of days which are non-overlapped and categorized. While identifying the pollution level in treated water will classify the operational state of WWTP to predict the faults at each stage of the WWTP. It generates the non-overlapped clusters by assuming that each data object originates from the mixture of underlying probability distributions and will consecutively assess the performance of the wastewater treatment plant. With the help of natural clusters generated, our results show that out of 445 days, in 352 days, the treated water is almost non-polluted. By this, we can assess the performance of the wastewater treatment plant. Further, it motivates the reuse of treated water for beneficial purposes like irrigation, aquarium, groundwater harvesting, and river water discharge based on pollution level in treated water.

References

Anaokar, G. S., Khambete, A. K., & Christian, R. A. (2018). Evaluation of a performance index for municipal wastewater treatment plants using MCDM – TOPSIS. *International Journal of Technology*, *9*(4), 715–726. https://doi.org/10.14716/ijtech.v9i4.102

Ashfaq, A., Saadia, A., & Sharma, S. (2010). Performance evaluation of a common effluent treatment plant in Delhi, India. *Journal of Industrial Pollution Control*, *26*(2), 157–160.

Baki, O. T., Aras, E., Akdemir, U. O., & Yilmaz, B. (2019). Biochemical oxygen demand prediction in wastewater treatment plants using different regression analysis models. *Desalination and Water Treatment*, *157*, 79–89. https://doi.org/10.5004/dwt.2019.24158

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821. https://doi.org/10.2307/2532201

Baridam, B. B. (2012). More work on K-means clustering algorithm: The dimensionality problem. *International Journal of Computer Applications*, *44*(2), 23–30. https://doi.org/10.5120/6236-8332

Batagelj, V. (1981). Note on ultrametric hierarchical clustering algorithms. *Psychometrika*, *46*(3), 351–352. https://doi.org/10.1007/BF02293743

Begum, S. F., Rajesh, A., & Kaliyamurthie, K. P. (2016). Multi-objective clustering and optimization. *International Journal of Control Theory and Applications*, *9*(28), 217–223. https://doi.org/10.17485/ijst/2016/v9i12/89282

Burkardt, J. (2009). *K-means clustering*. Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics.

Buvana, M., & Suganthi, M. (2015). An efficient cluster based service discovery model for mobile ad hoc network. *KSII Transactions on Internet and Information Systems*, *9*(2), 680–699. https://doi.org/10.3837/tiis.2015.02.011

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Diller, P. (2013). *Environmental protection administration, executive yuan guidelines concerning the establishment and oversight of non-profit corporations dedicated to environmental protection*.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Fraley, C., & Raftery, A. E. (2007). Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, *18*(6), 1–13. https://doi.org/10.18637/jss.v018.i06

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Lichman, M. (2013). *UCI machine learning repository*. The University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Mazhar, S., Ditta, A., Bulgariu, L., Ahmad, I., Ahmed, M., & Nadiri, A. A. (2019). Sequential treatment of paper and pulp industrial wastewater: Prediction of water quality parameters by Mamdani Fuzzy Logic model and phytotoxicity assessment. *Chemosphere*, *227*, 256–268. https://doi.org/10.1016/j.chemosphere.2019.04.022

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2012). An interactive approach to multiobjective clustering of gene expression patterns. *IEEE Transactions on Biomedical Engineering*, *60*(1), 35–41. https://doi.org/10.1109/TBME.2012.2220765

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. https://doi.org/10.1093/comjnl/26.4.354

Nadiri, A. A., Shokri, S., Tsai, F. T. C., & Asghari Moghaddam, A. (2018). Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model. *Journal of Cleaner Production*, *180*, 539–549. https://doi.org/10.1016/j.jclepro.2018.01.139

Nourani, V., Elkiran, G., & Abba, S. I. (2018). Wastewater treatment plant performance analysis using artificial intelligence–

an ensemble approach. *Water Science and Technology*, *78*(10), 2064–2076. https://doi.org/10.2166/wst.2018.477

Padalkar, A. V., & Kumar, R. (2018). Common effluent treatment plant (CETP): Reliability analysis and performance evaluation. *Water Science and Engineering*, *11*(3), 205–213. https://doi.org/10.1016/j.wse.2018.10.002

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sharghi, E., Nourani, V., Aliashrafi, A., & Gökçekuş, H. (2019). Monitoring effluent quality of wastewater treatment plant by clustering-based artificial neural network method. *Desalination and Water Treatment*, *164*, 86–97. https://doi.org/10.5004/dwt.2019.24385

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC Press.

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(8), 841–847. https://doi.org/10.1109/34.85677