COMMUNICATION SCIENCE ET TECHNOLOGIE
Vol 23, N°01, pp01, 12
ENPO – Maurice AUDIN
ASJP/ ESSN 2773-3483

A Clustering-Driven Strategy Utilizing DBSCAN for Detecting Outliers in Water Quality Data

A. Mohamed Imed KHELIL[1*], B. Mohamed LADJAL[1, 2,] C. Mohammed Assam OUALI[1, 2] and D. Hamza BENNACER[1, 3],

[1] LASS, Laboratory of Analysis of Signals and Systems,

[2] Department of Electronics, Faculty of Technology, University Pole of M'sila, 28000, Algeria

[3] ECP3M Laboratory, University- of Mostaganem

ABSTRACT

Monitoring environmental data to ensure the safety and reliability of public resources has become a crucial task in data-driven systems. One key aspect of this monitoring is the detection of anomalies—data points or behaviors that significantly diverge from the norm. This study explores the use of a density-based clustering method, DBSCAN, to identify such anomalies within datasets collected from drinking water treatment facilities. DBSCAN's capability to recognize dense regions and isolate noise makes it well suited for flagging irregularities in complex, real-world data. By applying this method to extensive datasets with diverse attributes, the research aims to enhance the consistency and safety of drinking water production processes, contributing to improved public health outcomes and operational resilience in water management systems.

Keywords: Anomaly detection; DBSCAN; water treatment; clustering algorithms; environmental data analysis.

Author Correspondence, e-mail: mohamed3imed.khelil@gmail.com

1. INTRODUCTION

Anomaly detection refers to the process of identifying data patterns that deviate significantly from established or expected behavior [1]. This process becomes particularly critical when such irregularities provide valuable insights into the underlying system. Anomalies may stem from diverse sources including cyber-attacks, sensor malfunctions, environmental shifts (e.g., climatic variations), or human oversight [1]. Its applicability spans numerous domains, including but not limited to intrusion detection, military reconnaissance, fraudulent transaction identification, healthcare diagnostics, insurance risk analysis, and preemptive fault detection in safety-critical infrastructure [2][3]. A primary advantage of anomaly detection lies in its capacity to transform atypical patterns into actionable intelligence. For instance, unauthorized data exfiltration from a compromised computer could manifest as unusual network traffic, prompting early intervention [2][4]. Similarly, detecting irregularities in MRI scans can assist in diagnosing malignant tumors [2][5], and anomalous telemetry from spacecraft systems may signal component degradation. Likewise, inconsistencies in financial transactions can serve as early indicators of credit card or identity fraud [2][6].

In the context of water treatment and production systems, continuous monitoring of water quality is vital. One of the most technically demanding aspects of this process involves determining the appropriate coagulant dosage, a factor essential for achieving optimal water purification [7][8]. Accurate dosing relies heavily on precise and dependable sensor readings of raw water parameters. Consequently, high-level processes, such as optimizing coagulation tests, must be resilient to sensor anomalies, including transient faults or inaccurate inputs [7][9]. Effective anomaly detection in these sensing systems is thus indispensable for maintaining operational integrity and ensuring high-quality water output. The timely identification of sensor faults, data outliers, and systemic failures has drawn increasing attention due to its implications for minimizing system downtime, enhancing productivity, and upholding safety and reliability standards [7][10][11]. This study aims to detect and validate potential sensor misreadings, data corruption, or anomalous raw water values to enable the reconstruction of trustworthy input for automatic coagulation control systems. By doing so, it ensures the integrity and reliability of data gathered from various water quality sensors [7][9]. However, one of the major challenges in applying supervised machine learning techniques to this problem is the scarcity of labeled anomalous instances [7][9][10]. As a result, unsupervised learning approaches present a more viable alternative in such scenarios [7].

Principal component analysis (PCA) has been widely applied in data mining to study data structure. In PCA, new orthogonal variables (latent variables or principal components) are obtained by maximizing the variance of the data. The number of latent variables (factors) is much smaller than the number of original variables, so the data can be visualized in a low-dimensional PC space. Although PCA significantly reduces the dimensionality of the space, it does not reduce the number of original variables, as it uses all the original variables to generate the new latent variables (principal components). For interpretation or future

investigations, reducing the number of variables would often be very useful. Feature (variable) selection can be achieved either by choosing informative variables or by eliminating redundant variables. [12]

In this research, the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) is employed to perform anomaly detection in the water treatment context. DBSCAN is a well-established density-based clustering technique known for its ability to identify clusters of arbitrary shape while effectively isolating noise [13][14]. The algorithm relies on two principal parameters—Epsilon (Eps) and Minimum Points (MinPts)—to define neighborhood density. Performance metrics include the number of identified clusters, unassigned data points, classification errors, and the time-to-noise ratio [13].

The structure of this paper is as follows: Section 2 outlines the dataset and describes the DBSCAN algorithm alongside the feature selection methodology. Section 3 presents and discusses the experimental results, while Section 4 offers concluding remarks.

2. MATERIALS AND METHODS

2.1. Study Area and dataset

The Cheliff dam is geographically located about 30 km northeast of the city of Mostaganem and 363 km northwest of Algiers (Fig. 1). It is located between the following coordinates: 35° 59' 00" N, 0° 24' 47" E. Mostaganem has a cold semi-arid climate and an average precipitation of about 347 mm/year. The average yearly temperature is 17.9 °C.

In this research, we seek to apply our approach for surface water quality monitoring using several physicochemical parameters. These parameters were collected from the Sidi Lahdjel production station over two years. Our knowledge of the treatment process is limited to data recorded at this station. More quality parameters of the surface water are measured daily by sensors, in addition to laboratory tests, which are carried out every week at all treatment process. The above physicochemical parameters were used to analyze the relationship among these descriptors and to verify the water quality monitoring model. Descriptive statistics of water parameters are given in Table 1 [15].

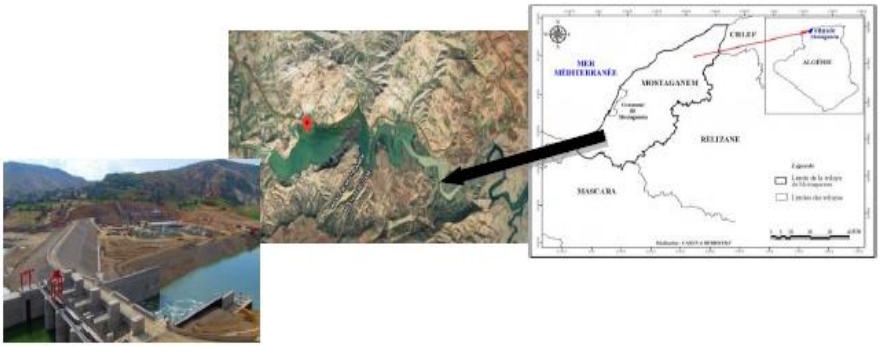


Fig.1. Map showing the region under study: Cheliff dam – Mostaganem – Algeria [Google Maps].

Table 1. Descriptive statistics of water parameters.

Variables	Min	Max	Mean	Standard deviation
Turbidity (NTU)	0.66	21.7	6.5	4.2521
pH	6.25	8.37	7.97	0.2692
Temperature (°C)	11.3	29	19.58	4.9852
Conductivity (µs/cm)	1144	3600	2125.6	408.1714
TDS (mg/L)	689	1728	1208.2	206.1315
OM (mg/L)	2.47	6.7	2.47	0.9347
Chlorine (mg/L)	192	724	425.29	99.3793
Bicarbonate (mg/L)	83	299	160.02	35.6523
Calcium (mg/L)	59	163.5	127.2	22.0475
Magnesium (mg/L)	44	110	74.1	11.1634
Total Hardness (°F)	45	77	62.33	7.4717
Color	11	169	58.1	36.1473
Coagulant (mg/L)	1.2	12	3.81	2.4355

2.2. Principal component analysis (PCA)

The PCA technique (also known as the eigenvector regression filter or the Karhunen-Loeve transform [16][17]) is used for dimensionality reduction, which involves zeroing out one or more of the weakest principal components, resulting in a lower-dimensional projection of the raw feature data that preserves the maximal data variance. The dimensionality reduction process is achieved through an orthogonal, linear projection operation. Without loss of generality, the PCA operation can be defined as

$$Y = XC \qquad (1)$$

With $Y \epsilon R^{SxP}$ is the projected data matrix that contains P principal components of X with $P \leq N$. So the key is to find the projection matrix $C \epsilon R^{NxP}$, which is equivalent to find the eigenvectors of the covariance matrix of X, or alternatively solve a singular value decomposition (SVD) problem for X [17][18]

$$X = U \Sigma VT \qquad (2)$$

Where $U \epsilon R^{SxS}$ and $V \epsilon R^{NxN}$ are the orthogonal matrices for the column and row spaces of X, and Σ is a diagonal matrix containing the singular values, λ_n for n = 0…, N-1, non-increasingly lying along the diagonal. It can be shown [18,19] that the projection matrix C can be obtained from the first P columns of V with

$$V = [V1,….,VN] \qquad (3)$$

And

$$C = [C1,….,CP] \qquad (4)$$

Where $v_n \epsilon R^{Nx1}$ is the nth right singular vector of X, and $c_n = v_n$.

In fact, the singular values contained in Σ in (2) are the standard deviations of X along the principal directions in the space spanned by the columns of C [17][18]. Therefore, λ_n^2

becomes the variance of X projection along the nth principal component direction. It is believed that variance can be explained as a measurement of how much information a component contributes to the data representation. One way to examine this is to look at the cumulative explained variance ratio of the principal components, given as

$$\mathbf{R}_{cev} = \frac{\sum_{i=1}^{p} \lambda_n^2}{\sum_{i=1}^{N} \lambda_n^2} \qquad (5)$$

Moreover, illustrated in Fig. 2. It indicates that keeping only a few principal components could retain over 90% of the full variance or information of X. As a comparative study, a varying number of principal components has been used and examined in the following evaluation section. [17]

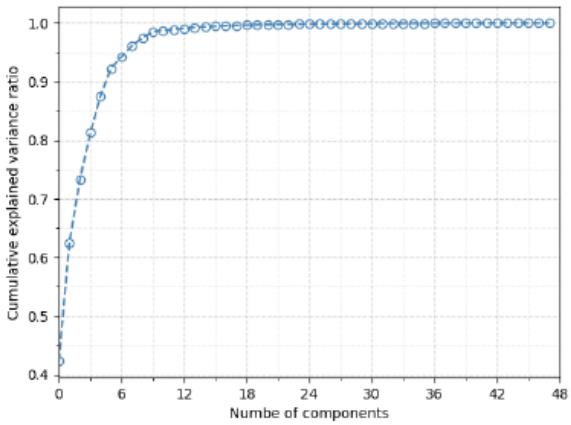


Fig. 2. Cumulative explained variance ratio over components.

2.3. DBSCAN algorithm

Density-based spatial clustering of application with noise, DBSCAN is a data-clustering algorithm that forms clusters with a maximal set of density-connected points. Clusters in the data space are typically high-density regions separated by lower object density regions.

DBSCAN defines the density in terms of the following:

1. ε-Neighborhood: Objects within a radius of ε (eps) from an object and can be represented by the relation,

$$N_\varepsilon(q): \{q | d\ (p, q) \leq \varepsilon \qquad (6)$$

Where p, q are data points in the space and d(p, q) represents the separation between the data points.

2. High density: ε-Neighborhood of an object containing at least minpts of data points. [2]

The algorithm requires two parameters: the neighbourhood distance ε (eps) and the minimum number of the points needed to form a high-density region minpts. The parameters categorize the data points as core points, border points, and outlier points. A core point has more than minpts number of points within the ε (eps) distance and lies at the cluster's interior. A border point is in the neighbourhood of a core point but has fewer than minpts number of points within eps. Outlier points are the anomalous points that are neither a core point nor a border

point and do not fit any cluster.

The DBSCAN algorithm works as follows. An arbitrary point that has not been visited yet is selected, and its ε-neighborhood is retrieved. If the number of neighborhood points is greater than the minpts, a cluster is started; else, the point is marked as noise. If the point being noise is later found to lie in the ε- neighborhood of some other point with apt size, it would be made part of that cluster. If a point lies in a cluster's high-density zone, then its ε-neighbourhood is also a part of that cluster. All points found within the ε- neighbourhood are added to the cluster, as is their own ε-neighborhood if they are dense until it is found that the density-connected cluster is complete. Again, an unvisited point is retrieved and processed as stated above, leading to the determination of a further cluster or noise. [2]

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the aforementioned proposed framework was applied to water quality data from Cheliff dam station in Mostaganem (Algeria). For testing the applicability of the suggested methodology, our monitoring system consists of two steps: features selection and sensors anomalies detection in the different measurements for water quality assessment. The feature selection technique is based on PCA, and sensors anomalies detection technique are based on DBSCAN. All proposed methods were implemented and assessed using MATLAB2019b environment software.

3.1. Features selection using PCA method

The PCA method is used with a variation of 80 to 90% of the eigenvalues, without any transformation of the resulting components that are not correlated. A total of 142 samples of twelve physicochemical parameters of water quality are used in this phase. Parameters such as color, pH, temperature (T°), electrical conductivity (EC) and turbidity (TU) are collected by sensors installed in all treatment processes of the plant. Every week, in the laboratory, some chemical parameters are examined such as: TDS, OM, Chlorine, bicarbonate (B), Calcium, Magnesium (Mg) and Total Hardness (TH). The aforementioned collected data will be applied to verify the water quality assessment model.

First, a PCA analysis is performed to determine the descriptor parameters or input variables most representative of water quality. This involves extracting relevant information such as: correlation matrix, histogram of eigenvalues and correlation circle. It should be noted, however, that all 12 input variables of this database are retained due to the importance of its parameters for water quality and the continuity of their measurements over time. The PCA analysis applied to all the database data provides Table II and the histogram in Figure 3.

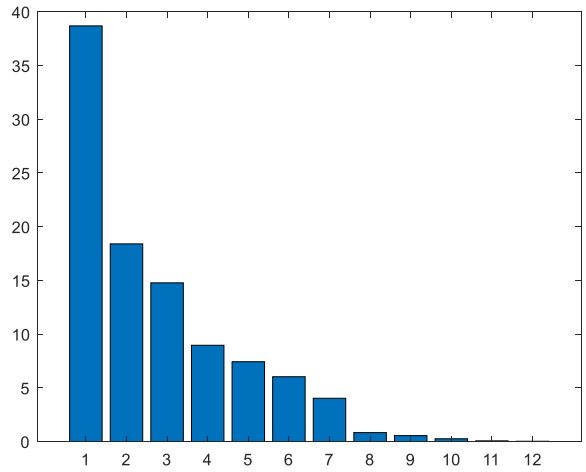


Fig. 3. The PCA analysis

TABLE 2. Variables of eigenvectors obtained by applying PCA.

Variables												
TU	**0.09**	**0.58**	**0.15**	**0.27**	**0.08**	**0.08**	**-0.20**	**0.31**	**-0.62**	**0.06**	**-0.00**	**-0.01**
PH	**0.06**	**0.40**	**0.18**	**-0.42**	**-0.36**	**0.26**	**0.64**	**0.03**	**0.03**	**-0.03**	**0.00**	**0.00**
T°	**-0.07**	**-0.12**	**0.60**	**-0.16**	**0.51**	**-0.00**	**0.08**	**0.27**	**0.16**	**0.16**	**0.07**	**-0.41**
EC	**0.36**	**-0.11**	**0.36**	**-0.10**	**0.32**	**-0.01**	**0.03**	**-0.14**	**-0.10**	**-0.32**	**-0.14**	**0.66**
TDS	**0.45**	**-0.05**	**0.00**	**-0.06**	**-0.01**	**-0.01**	**-0.04**	**-0.39**	**-0.22**	**-0.46**	**0.17**	**-0.57**
OM	**0.01**	**-0.13**	**0.37**	**0.57**	**-0.34**	**-0.50**	**0.36**	**-0.01**	**-0.01**	**-0.01**	**-0.00**	**-0.01**
Ch	**0.45**	**-0.05**	**0.04**	**-0.01**	**-0.01**	**0.03**	**0.00**	**-0.38**	**-0.08**	**0.79**	**-0.02**	**0.01**
B	**-0.02**	**0.22**	**-0.45**	**0.14**	**0.59**	**-0.24**	**0.54**	**-0.11**	**-0.03**	**0.01**	**0.02**	**-0.01**
Cal	**0.36**	**0.06**	**-0.16**	**-0.31**	**-0.08**	**-0.50**	**-0.09**	**0.38**	**0.09**	**0.00**	**-0.53**	**-0.12**
Mg	**0.28**	**-0.24**	**-0.12**	**0.42**	**0.03**	**0.58**	**0.19**	**0.24**	**0.11**	**-0.06**	**-0.42**	**-0.13**
TH	**0.43**	**-0.11**	**-0.17**	**0.02**	**-0.04**	**-0.02**	**0.04**	**0.49**	**0.14**	**0.02**	**0.68**	**0.15**
Co	**0.17**	**0.56**	**0.11**	**0.25**	**0.04**	**0.00**	**-0.23**	**-0.18**	**0.68**	**-0.08**	**0.01**	**-0.00**

PC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**
						Eigenvalues						
	4.64	**2.21**	**1.77**	**1.07**	**0.89**	**0.72**	**0.48**	**0.1**	**0.06**	**0.03**	**0.01**	**0.01**
						Total variance proportion (%)						
	38.68	**18.39**	**14.76**	**8.95**	**7.42**	**6.02**	**4.02**	**0.82**	**0.54**	**0.25**	**0.05**	**0.02**
						Cumulative variance proportion (%)						
	38.68	**57.08**	**71.84**	**80.8**	**88.23**	**94.26**	**98.28**	**99.1**	**99.66**	**99.91**	**99.97**	**100**

A variance-covariance matrix is formed using PCA on the input variables. According to Table II, the PCA results and statistical parameters such as eigenvalues, cumulative variance proportion, and variance proportion are shown. The four PCs represent 80.80% of the total variance proportion of the input samples and eliminate the remaining components, as shown in Table II. These PCs mainly calculate the initial variance of the data. In addition, 1CP

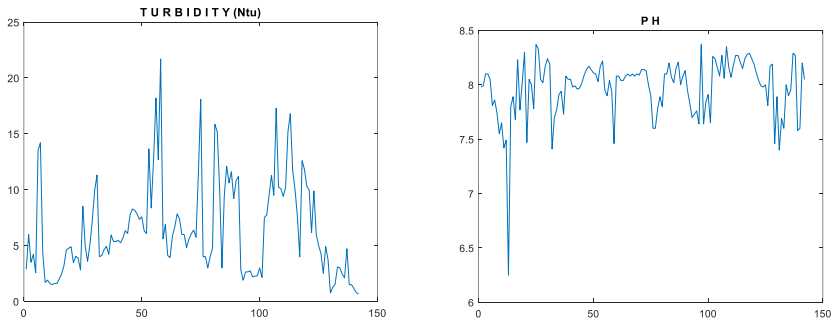
applications are used to obtain eigenvectors to evaluate the coefficients for PC training. The correlations between each variable and the learned principal components are shown in Table II. In this table, the most effective parameters in PC training are shown in red bold. The total variance in the dataset represents 80.80% of the first four principal components combined. The first component (PC1) is 38.68%, 18.39% being the second component (PC2), 14.76% being the third component (PC3) and 8.95% of the total variance being the fourth component (PC4).

In Table II, the rapid decrease of the eigenvalues is apparent. For the evaluation of the predominant physicochemical processes, the eigenvalues of the first four principal components (PC1-PC4) can be used. The EC and B concentrations are very positive (0.59 – 0.66), while the Mg concentration is weakly positive for the first component (0.58). T° and TH have high positive loadings in PC2 (0.60 - 0.68), and the other concentrations show weak positive loadings (0.38-0.45). The TU concentrations in PC3 have high positive loadings (0.58). The pH concentrations for PC4 show high positive loadings (0.64), while Mg displays moderate positive loadings (0.58), and TU and TH show positive loadings (0.58 - 0.68).

According to Table II, the first four PCs are the input characteristics of the evaluated classifiers. The variables selected are: pH, Temperature (T°), Electrical Conductivity (EC) and Turbidity (TU). Consequently, monitoring must take place at the treatment plant and continuously using selected parameters that are the most representative used due to the strong correlations existing between all parameters, as well as the most fundamental and easily measurable by physical sensors in the water quality monitoring system.

3.2. Anomaly detection using DBSCAN method

This segment of the study focuses on identifying anomalies within the sensor-generated data corresponding to selected physicochemical parameters relevant to water quality evaluation. To this end, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is employed as an unsupervised technique for outlier detection. The dataset under analysis comprises four key variables—Temperature (T°), pH, Electrical Conductivity (EC), and Turbidity (TU)—which were previously selected through the Relief-based feature selection process. These refined inputs serve as the basis for the anomaly detection framework illustrated in Figure 4.



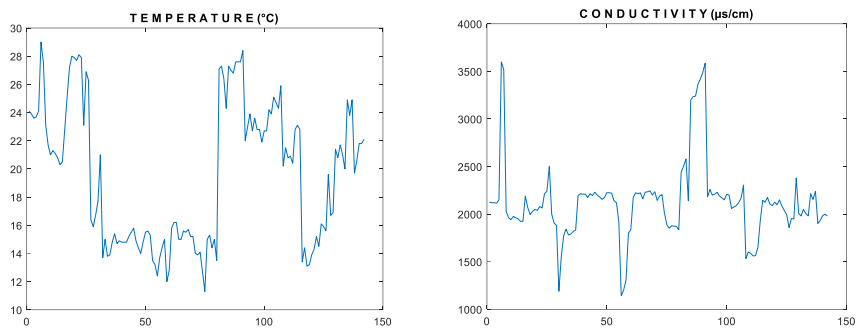


Fig. 4. Evolution of the water quality variables.

In the anomaly detection process, the DBSCAN algorithm parameters were set with minPts = 4 and ε = 2. It is important to note that the number of anomalies detected is sensitive to variations in these parameter values. During the experimental validation phase, intentional faults were introduced into the turbidity and temperature sensors to observe their impact on the data visualization. The algorithm was applied to a dataset comprising 142 samples, with the results illustrated in Figure 5.

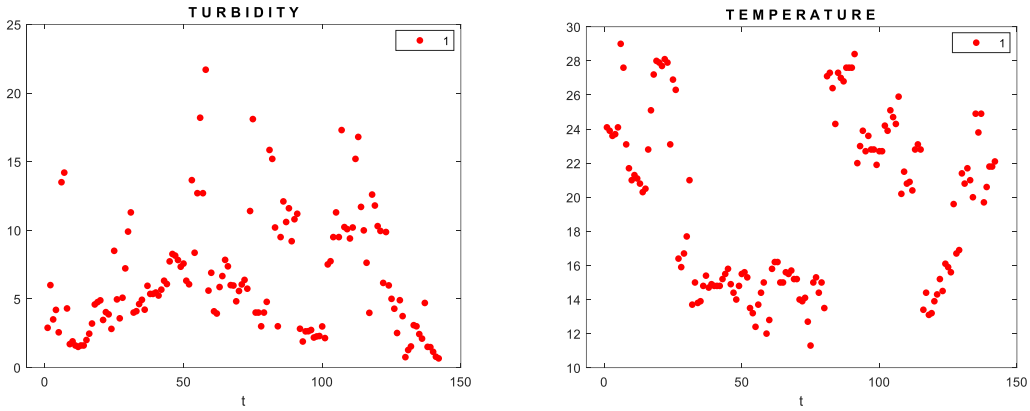


Fig. 5. Results obtained with the DBSCAN algorithm.

To evaluate whether a sensor is producing faulty readings, hypotheses were tested through the generation of test statistics, which are essential tools in process monitoring for anomaly detection. Simulated sensor faults were introduced on day 60, and their effects were analyzed through graphical representations (Figure 6). The presence of outlier points—highlighted in red—confirmed the successful detection of anomalies in both the turbidity and temperature sensors.

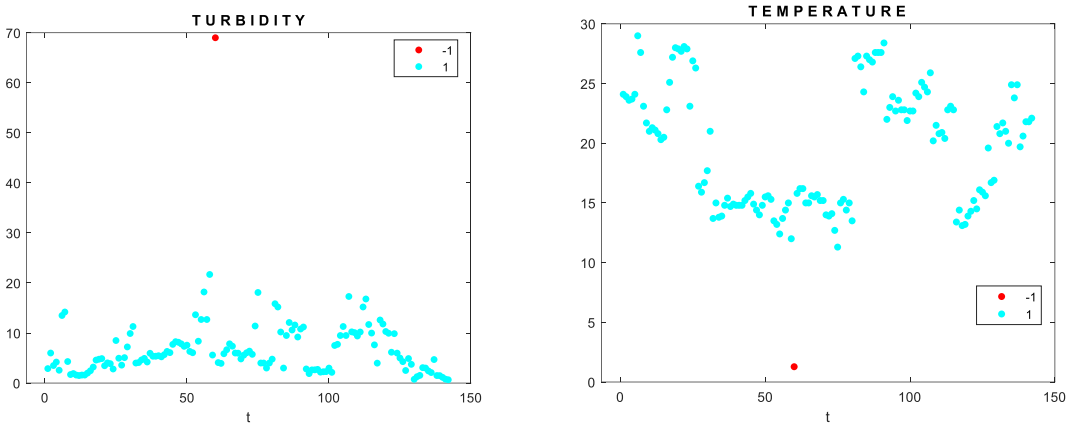


Fig. 6. Fault detection with the DBSCAN algorithm.

Further, a dual-fault scenario was simulated, wherein simultaneous disturbances were injected into the turbidity and temperature sensors on days 100 and 101. As evidenced in Figure 7, the observed deviations in sensor behavior confirmed the malfunction of both units.

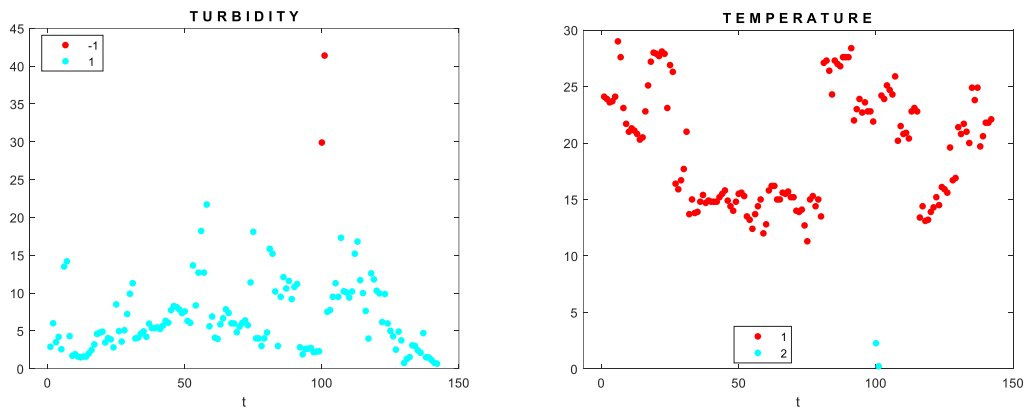


Fig. 7. Simulation of two faults by the DBSCAN algorithm.

4. CONCLUSION

This study presents a comprehensive framework for water quality assessment, integrating two key methodologies: the ReliefF algorithm for feature selection and the DBSCAN clustering technique for sensor anomaly detection. The first contribution lies in dimensionality reduction through Relief, enabling the selection of the most informative physicochemical variables. The second focuses on the application of DBSCAN to identify sensor anomalies in real-time. Accurate anomaly detection in drinking water treatment systems is a critical component of quality assurance. The findings demonstrate that DBSCAN is capable of reliably detecting sensor faults, with performance comparable to other established techniques referenced in [7]. Real-world experimental data from the treatment plant further validate the robustness and efficacy of this approach. Importantly, this methodology also contributes to cost-efficiency in system monitoring by enhancing fault detection capabilities with minimal added complexity.

5. REFERENCES

[1] Mete ÇELİK, Filiz DADAŞER-ÇELİK, Ahmet Şakir DOKUZ. Anomaly Detection in Temperature Data Using DBSCAN Algorithm. International Symposium on INnovations in Intelligent SysTems and Applicat ions. 2011. 1 5 - 1 8 J u n e 2 0 1 1 , I s t a n b ul -Ka d i k ö y , TURKEY.

[2] Praphula Jain, Mani Shankar Bajpai, and Rajendra Pamula, A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality. The International Arab Journal of Information Technology, Vol. 19, No. 1, January 2022.

[3] Chandola V., Banerjee A., and Kumar V.,"Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.

[4] Tan P., Steinbach M., Kumar V., Potter C., Klooster S., and Torregrosa A., "Finding Spatio-Temporal Patterns in Earth Science Data," in KDD 2001 Workshop on Temporal Data Mining, pp. 1-12, 2001.

[5] Quinn J. and Sugiyama M., "A Least-Squares Approach to Anomaly Detection in Static and Sequential Data," Pattern Recognition Letters, vol. 40, pp. 36-40, 2014.

[6] Kalid S., Ng K., Tong G., and Khor K., "A Multiple Classifiers System for Anomaly Detection in Credit Card Data with Unbalanced and Overlapped Classes," IEEE Access, vol. 8, pp. 28210-28221, 2020.

[7] Mohamed Imed KHELIL, Mohamed LADJAL, OUALI Mohammed Assam, BENNACER Hamza. Sensor Anomaly Detection using Hierarchical Clustring and Self Organizing Map for Water Quality Assessment.

[8] Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS).

[9] B. Lamrini • El-K. Lakhal • M-V. Le Lann, L. Wehenkel, "Data validation and missing data reconstruction using self-organizing map for water treatment," Springer (2011),

[10] Y. Zhang, C.M. Bingham and M. Gallimore, "Applied Sensor Fault Detection, Identification and Data Reconstruction, " Advances in Military Technology Vol. 8, No. 2, December 2013.

[11] Monowar H. Bhuyan a , D.K. Bhattacharyya b , J.K. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic," Information Sciences 348 (2016) 243–271,

[12] Q. Guo , W. Wub, D.L. Massart , C. Boucon , S. de Jong "Feature selection in principal component analysis of analytical data, " Chemometrics and Intelligent Laboratory Systems 61 (2002) 123– 132.

[13] Sarika Chaudhary, Pooja Batra Nagpal, Contrivancing DBSCAN Algorithm on Spatial Data Using Matlab, IJITKM.2016. Volume--10, Number-1 Jun-Dec 2016 p. 10-14.

[14] G. Karypis, E. H. Hanand, V. Kumar, "Chameleon: Hierarchical Clustering using Dynamic Modelling," Computer, Aug 1999.

[15] Mohamed Imed KHELIL, OUALI Mohammed Assam, Mohamed LADJAL, BENNACER Hamza, Soft Sensing Modeling Based on Support Vector Machine and Self-Organizing Maps Model Selection for Water Quality Monitoring.

[16] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd Edition. AddisonWesley, 2002.

[17] Jason Zhang, Machine Learning With Feature Selection Using Principal Component Analysis for Malware Detection: A Case Study. (Feb 2019).
[18] Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016.