Bulletin of the*Transilvania* University of Braşov
Series IX: Sciences of Human Kinetics • Vol. 17(66) No. 1 – 2024
https://doi.org/10.31926/but.shk.2024.17.66.1.9

COMPARISON OF UNSUPERVISED LEARNING ALGORITHMS FOR CLUSTERING CUBAN CITIZENS USING A LIFESTYLE QUESTIONNAIRE

S. TORRES[1] D. ALONSO[1] N. MARTÍNEZ[1] S. MERCED[2]

Abstract: *This study uses information technologies to analyze lifestyles of Cubans. Cluster analysis is used to identify similarities in habits and lifestyles. Clustering results are compared using K-Means, DBSCAN and HDBSCAN algorithms. Principal Component Analysis is applied to visualize the dataset. Internal validation metrics are defined to evaluate the performance of the algorithms. The results indicate that K-Means provides better clustering for this dataset.*

Key words: *algorithms, clustering, density, habits, life.*

1. Introduction

At present, Information and Communication Technologies (ICT) have emerged as one of the main engines that propel knowledge and research, forcing human beings to advance in technological terms. For this reason, the relevance of their use is highlighted, as well as the pros and cons involved in promoting a healthy lifestyle. ICTs have revolutionized our way of life, including how we access knowledge and carry out our daily activities. They have opened up a range of opportunities for education and learning, research and the development of new technologies [12].

In relation to the promotion of a healthy lifestyle, Information and Communication Technologies (ICT) can be a valuable resource to inform and educate about the relevance of physical activity and balanced nutrition [2]. There are multiple applications and digital tools that can assist people in monitoring their physical activity, such as step count, distance traveled, calories burned, heart rate, diet, among other health-related aspects, providing real-time feedback. In this context, there are a wide variety of applications and electronic devices available that can facilitate people to record their physical activity and share recommendations in this regard.

In order for the recommendations given to people to be in accordance with their lifestyles, they should be classified into groups that reflect their lifestyles. In this direction, this work intends to develop population groups of Cubans in terms of

[1] University of Informatics Sciences, Cuba.
[2] University of the Sciences of the Physical Culture and the Sport "Manuel Fajardo", Cuba.

their lifestyles through the analysis of data collected in a diagnosis. For this purpose, a study is presented that uses a methodology based on cluster analysis to identify similarities among Cuban citizens in terms of their habits and lifestyles. The study considered data collected by a diagnosis carried out on 528 Cuban citizens from all over the country and from all age groups over 15 years old, surveyed in the year 2023.

The specific objective of the research is to compare the clustering results produced by the application of the artificial intelligence algorithms K-Means, DBSCAN and HDBSCAN to a dataset that measures lifestyles of Cuban citizens.

2. Algoritmos

2.1. Algoritmo K-Means

The K-Means algorithm is a partitioning algorithm; it divides objects into a pre-specified number of clusters, without regard to a hierarchical structure [20]. It can be applied for "similarity clustering" problems and can help the researcher to gain a qualitative and quantitative understanding of large amounts of N-dimensional data [10-20].

The K-Means algorithm starts with a preliminary solution, which is obtained randomly from the data set. The objective is to iteratively improve this solution until a local minimum is reached. First, the elements of the set are divided into a group of M clusters, associating each object with the centroid of the closest cluster from the previous iteration [19]. Then, the centroids of each cluster are recalculated, considering the new partition. The quality of the new solution must be equal or superior to the previous one. The algorithm continues as long as there is improvement [10].

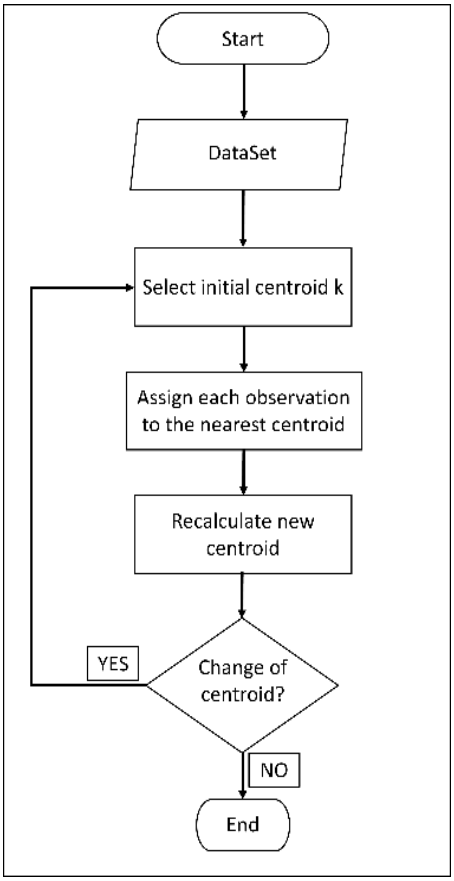

Fig. 1. *Scheme of execution of the K-Means algorithm*

2.2. DBSCAN Algorithm

The DBSCAN algorithm is the first density-based algorithm, the concepts of center point (points having in their neighborhood a number of points greater than or equal to a specified threshold), edge and noise are defined [7].

The algorithm starts by selecting an arbitrary point p. If p is a central point, a group is constructed and all dense-reachable objects are placed in its group from p. If p is not a central point, another object from the dataset is visited. The

process continues until all objects have been processed. Points outside the formed groups are called noise points; points that are neither noise nor central are called edge points [7].

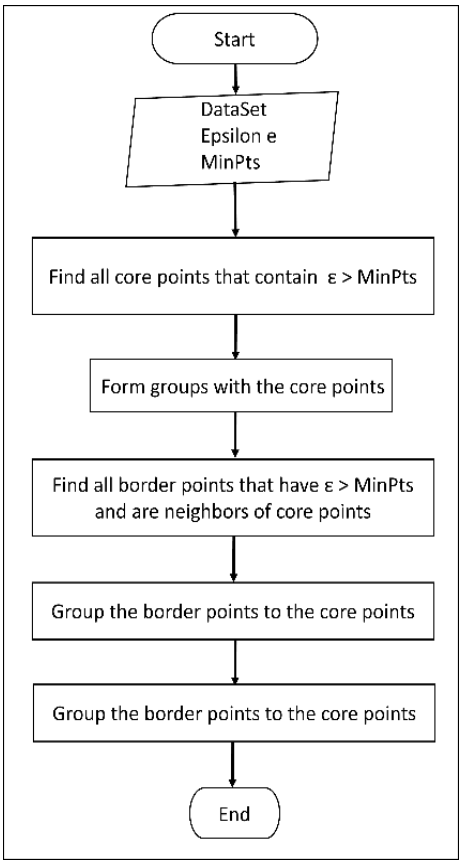

Fig. 2. *Execution scheme of the DBSCAN algorithm*

2.3. HDBSCAN Algorithm

The HDBSCAN algorithm is an optimized version of DBSCAN, developed by the same authors, which inherits the benefits of the hierarchical and density algorithms. It extends DBSCAN by converting it into a hierarchical algorithm and then extracting a flat cluster structure [15].

The algorithm begins by calculating the density of each data point based on the distance to its nearest neighbors. This density is used to construct a hierarchical tree of clusters, the Stability Tree. A cutting algorithm is then applied to this tree to obtain a flat clustering, preserving the most stable clusters as clusters and considering the less stable clusters as noise [15].

```
  Input: Location data: LD, Parameter: Eps and Minpts,
S-Tree: Height
  Output: LD with cluster lable and Spatial_Tree was
built
1. DBSCAN_ OBJECT Root=Joint(LD,Eps,Minpts); // root
   node of Tree
2. ENQUEUE(Q, Root) ;        // push DBSCAN object into
   Queue
3. front:=0, last:=0, level=0;
4. while(Queue<>empty and front<=last) DO
5.   DBSCAN_ OBJECT node= DEQueue(Q); // Pull data from
   Queue
6.    front++;             //
7.    Data_OBJECT Childern    =DBSCAN.getCluster(node);
   //Call DBSCAN
8.    if(level > Height)
9.       break;
10.
11.    For i FROM 1 TO Childern.size  DO
12.      Data child=Childern.get(i);
13.      DBSCAN_ OBJECT Root=Joint(child,Eps,Minpts);
14.      ENQUEUE(Q,DBSCAN_ OBJECT) ;
15.    end For
16.
17.    if(front>last) // members in one level have been
   searched
18.       last= Q.size()+front-1;
19.       level ++;
20.    end if
21. end while
```

Fig. 3. *Pseudo code of HDBSCAN*

3. Dimensionality Reduction

Dimensionality reduction is a technique that transforms high dimensional data to a lower dimensionality space, thus reducing the number of variables or features in a data set, but retaining the essential information. This technique is mainly used to compress data, reduce noise and as a preliminary step to classification. In addition, it allows the visualization of high dimensionality data sets that, due to their large number of attributes, would be impossible to represent graphically without this technique, as it is in this case

[3-19].

One of the most widely used algorithms in dimensionality reduction and that we will use in this research is the Principal Component Analysis (PCA) method [21]. This is a linear and unsupervised dimensionality reduction technique. Its objective is to identify the main directions of variability in the data and represent them in a lower dimensionality space. In this way, it manages to condense almost all the information into a few components. It transforms a set of correlated variables into a set of orthogonal variables, known as principal components [5]. The first principal component is the one that explains the greatest variability in the data set, and so on, until the desired number of components is reached [3-19].

4. Internal Validation Metrics

Internal Validation Metrics are indicators that evaluate the quality of a clustering with data information only. They do not need additional information to the result of the clustering algorithm [4].

In this article we will use three of the most commonly used metrics in clustering:

Silhouette Index: This coefficient evaluates the cohesion of group a(x) by calculating the average distance from the centroid (x) to all other points in the same cluster. It also measures the separation of groups b(x) by calculating the average distance from the centroid (x) to all the points in the nearest cluster [4]. This coefficient is bounded between the values -1 and 1, with -1 being a bad clustering, 0 an indifferent clustering and 1 a good clustering [9].

$$s(x) = \frac{b(x)-a(x)}{\max\{a(x),b(x)\}} \quad (1)$$

$$SC = \frac{1}{N}\sum_{i=1}^{k} s(x) \quad (2)$$

Formula 1 refers to the calculation of the value of the silhouette coefficient for point x, while formula 2 shows the value of the coefficient in question for the whole grouping, where N is the number of groups formed [9].

Davies-Bouldin Index: is a metric proposed by David L. Davies and Donald W. Bouldin in 1979 to evaluate clustering algorithms. It treats each cluster individually, measuring its similarity to the nearest cluster. A smaller value indicates more compact and separate clusters [4].

$$DB = \frac{1}{k}\sum_{i=1,i\neq j}^{k} \max\left(\frac{\sigma_j+\sigma_i}{d(c_i,c_j)}\right) \quad (3)$$

Formula 3 is used to calculate Davies-Bouldin Index. Where k is the number of clusters, σi is the average distance between each point in cluster i and the cluster centroid, σj is the average distance between each point in cluster j and the cluster centroid, and (c_i, c_j) is the distance between the centroids [16].

Calinski-Harabasz Index: the variance ratio criterion, introduced by T. Calinski and J. Harabasz in 1974, evaluates clustering algorithms. It measures the similarity of an object to its cluster (cohesion) and the separation between clusters. Cohesion is based on the distances to the cluster centroid, and separation on the distance of the centroids to the global centroid [18]. To obtain the index, the sum of the squared distances between clusters (BSS) is calculated and defined as shown in

formula 4,

$$BSS = \sum_{k=1} K n_k |C_k - C|^2 \quad (4)$$

where n_k is the number of observations in cluster k, C_k is the centroid of cluster k, C is the centroid of the data set and K is the number of clusters [13].

On the other hand, the sum of squared distances within each cluster (WSS, Within-Cluster Sum of Squares) can be obtained by means of formula 5,

$$WSS = \sum_{i=1}^{n_k} |X_{ik} - C_k|^2 \quad (5)$$

where n_k is the number of observations in cluster k and X_{ik} is the observation i in cluster k.

Thus, the Calinski-Harabasz index can be defined as shown in formula 6,

$$CH = \frac{\frac{BSS}{K-1}}{\frac{WSS}{N-K}} \quad (6)$$

where N is the total number of observations.

5. Data

The data are part of a diagnosis carried out by the authors, validated with statistical software and expert criteria of the Universidad de las Ciencias de la Cultura Física y el Deporte ``Manuel Fajardo'' (UCCFD), in the year 2023. The diagnosis collects data from 528 Cuban citizens from all provinces of the country and all age groups over 15 years old.

The questionnaire was based on the FANTASTICO questionnaire developed by the Canadian University of McMaster [1]. The questionnaire was submitted to two rounds of expert analysis in order to adapt it to our country. At the end of both rounds of analysis, the experts agreed in more than 80% of the criteria to eliminate 2 items and to redefine one question, which allows the questionnaire to be adapted to the Cuban context.

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were performed on this questionnaire using the statistical tool IBM SPSS Statistics version 23[8-14].

To start the AFE, the Kaiser Meyer-Olkin (KMO) and Bartlett's test of sphericity tests are performed, verifying the relationships between variables and their significance in the instrument under analysis [6]. A value of KMO= .767 was obtained, which is considered an acceptable value. Bartlett's test of sphericity shows that $\chi 2$ (378) = 2939.7 ($p < .000$). These results confirm that the factor analysis is feasible. Its development is carried out through the principal component's procedure and the Varimax factorial rotation method. The reliability of the measurement instrument was determined by obtaining a Cronbach's Alpha = .742 and Macdonal's Omega = .723[11].

The CFA is carried out determining the model fit indicators: chi-square ratio over degrees of freedom ($\chi 2$/g.l.): 2.14, associated likelihood level (CMIN/DF): 2.17, moderate fit index (NFI): .91, comparative fit index (CFI): .88, Tucker-Lewis index (TLI): .90, goodness-of-fit index (GFI): .89, parsimonious normed fit index (PNFI): .82, root mean square residual (RMCR): .076, root mean square error of approximation (RMSEA): .085 and Akaike information criterion (AIC): 689.92. The method used for this analysis was the maximum likelihood method.

Taking into account the results of the CFA, the level of adjustment of the factors defined as a result of the whole validation process described so far can be satisfactorily evaluated in the theoretical order and with the statistical tests carried out, all of which shows the validity of the construct.

6. Results

6.1. Application of the PCA method

First, a dimensionality reduction was performed on the dataset so that it could be plotted. The PCA method was applied to reduce the dataset to 3 dimensions and then it was plotted using the plotly library for rendering 3D graphics, as shown in Figure 1.

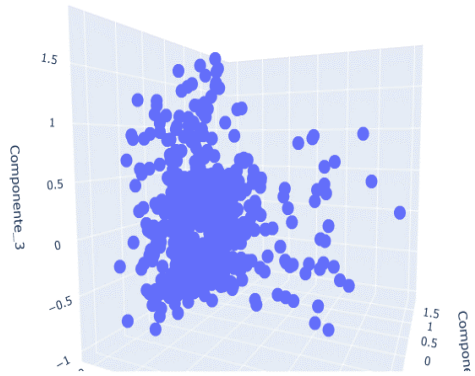

Fig. 4. *The figure shows the data spaced in three dimensions*

6.2. Execution of the algorithms

We proceeded to the calculation of the elbow method to define an efficient k for the execution of the K-Means algorithm [17].

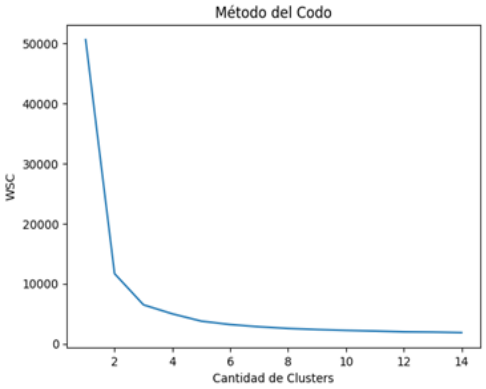

Fig. 5. *Plot resulting from the application of the elbow method to the dataset*

After obtaining the results of the application of the elbow method, it was decided to apply the algorithm with a k= 5, with which 5 clusters are formed. This decision is taken considering that the fifth cluster the function makes the inflection point. After this, it begins to converge definitively.

In addition, the necessary parameters were defined for the other two algorithms: in the case of DBSCAN, the epsilon parameters (ε=3) and the number of initial points to start the clustering (min_samples=3) were defined. And in the case of the HDBSCAN algorithm, the minimum exemplary parameters to assemble a cluster(min_cluster_size=5) were defined as well as the metric, which in this case was the cosine distance (metric='cosein').

The proposed algorithms were applied and the instances were grouped and distributed as shown in the following table.

Table 1

Shows the groups created after the execution of the algorithms and some unplaced elements considered noisy (-1).

Groups	K-MEAS	DBSCAN	HDBSCAN
-1	-	10	23
0	291	151	290
1	26	49	15
2	41	288	151
3	19	4	14
4	151	3	35
5	-	15	-
6	-	3	-
7	-	5	-

As can be seen, the DBSCAN and HDBSCAN algorithms were not able to group all the elements, missing 10 and 23 respectively, which were catalogued as noise. On the other hand, DBSCAN created 8 clusters, three more than the other two algorithms, showing a lower capacity to optimize the culturing in the dataset.

The elements of the dataset grouped after the execution of the algorithms are shownbelow in their respective graphs.

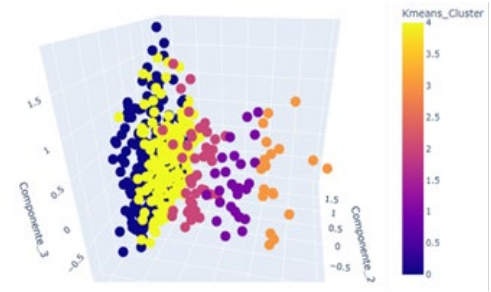

Fig. 6. *The figure shows the distribution of the dataset elements in groups after the application of the K-Means algorithm*

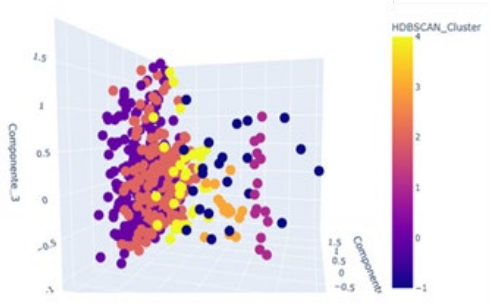

Fig. 7. *The figure shows the distribution of the dataset elements in groups, and some considered noise after the application of the HDBSCAN algorithm*

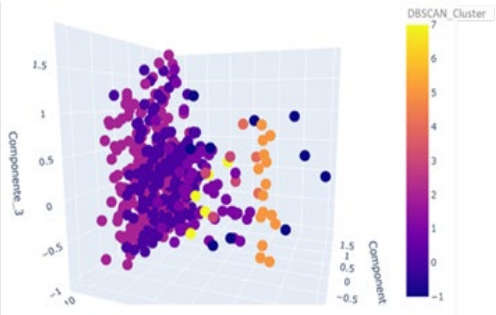

Fig. 8. *The figure shows the distribution of the dataset elements in groups, and some considered noise after the application of the DBSCAN algorithm*

Finally, internal validation metrics were applied to examine the clustering results as shown in Table 2.

Table 1

The values of the metrics for each of the algorithms are shown

Metrics	K-MEAS	DBSCAN	HDBSCAN
SC	0.50	0.47	0.51
DB	0.68	1.34	1.65
BBS	1685.54	436.27	670.50

As can be seen in relation to the Silhouette index, the three algorithms cohere their groups relatively well, HDBSCAN being the best performing,

although K-Means has practically the same result. With respect to the Davies-Bouldin index, the best performer is the K-Means algorithm, which improves the other two considerably, and finally with respect to the Calinski-Harabasz index, the K-Means algorithm again improves the other algorithms, almost tripling the result of HDBSCAN and quadrupling that of DBSCAN, the latter being the worst performer.

7. Conclusions

The research is based on the comparison of the performance of the K-Means, DBSCAN and HDBSCAN algorithms in terms of the conformation of groups of Cuban citizens according to their habits and lifestyles, considering for its study the data collected from a diagnosis carried out on 528 citizens.

After the application of the algorithms, the results were subjected to evaluation through internal validation metrics, which showed that the K-Means algorithm performs better clustering for this data set in terms of cohesion, separation and similarity of the groups.

References

1. Betancurth Loaiza, D.P., Vélez Álvarez, C., Jurado Vargas, L.: C*ontent validation and adaptation of the Fantastico questionnaire by Delphi technique*. In: Revista Salud Uninorte, Vol. 31(2), 2015, p. 214-227.
2. Diego-Cordero, R., Fernández-García, E., Romero, B.B.: *Use of ICT to promote healthy lifestyles in children and adolescents: the case of overweight*. In: Revista Española de Comunicación en Salud, Vol. 8, 2017, no.1, p.79-91. https://doi.org/10.20318/recs.2017.3607
3. Dorado Valín, A.: *Analysis of the Impact of Distance Measurements on Dimensionality Reduction Techniques.* 2023, available at: https://ruc.udc.es/dspace/handle/2183/33866
4. Esteban, A., Zafra, A., Ventura, S.: *Comparative Study of Dissimilarity Measures for Multi-Instance Clustering.* In: 2020 IEEE Congress on Evolutionary Computation (CEC), 2021.
5. Fan, C., Lai, X., Wen, H., Yang, L.: *Coal and gas outburst prediction model based on principal component analysis and improved support vector machine*. In: Geohazard Mechanics, vol. 1(4), 2023, p.319-324. https://doi.org/10.1016/j.ghm.2023.11.003
6. García-Hernández, A., González-Ramírez, T.: *Construction and validation of a questionnaire to assess student satisfaction with mathematics learning materials.* In: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality, 2018, p. 134-138. https://doi.org/10.1145/3284179.3284204
7. Hanafi, N., Saadatfar, H.: *A fast DBSCAN algorithm for big data based on efficient density calculation. Expert Systems with Applications*. Vol.203, 2022, 117501. https://doi.org/10.1016/j.eswa.2022.117501
8. Howard, M.C.: *A systematic literature review of exploratory factor analyzes in management*. In: Journal of Business Research. Vol. 164, 2023,

113969. https://doi.org/10.1016/j.jbusres.2023.113969

9. Lenssen, L., Schubert, E.: *Medoid Silhouette clustering with automatic cluster number selection*. In: Information Systems, Vol.120, 2024, 102290. https://doi.org/10.1016/j.is.2023.102290
10. MacQueen, J.: *Some methods for classification and analysis of multivariate observations.* In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, Volume 1: Statistics: Vol. 5.1 (pp. 281-298). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods- for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992
11. Malkewitz, C.P., Schwall, P., Meesters, C., et al.: *Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound.* In: Social Sciences & Humanities Open, Vol.7(1), 2023, 100368. https://doi.org/10.1016/j.ssaho.2022.100368
12. Melo-Martínez, H.A., Rosario-González, J.P., Bennasar-García, M. I.: *Use of ICT and its influence on healthy lifestyles in students*. In: Polo del Conocimiento, Vol.8(5), 2023, Article 5. https://doi.org/10.23857/pc.v8i5.5551
13. Ning, Z., Chen, J., Huang, J., et al: *An improved k-means clustering algorithm with a weighted distance and a novel internal validation index*. In: Egyptian Informatics Journal, Vol.23(4), 2022, p. 133-144. https://doi.org/10.1016/j.eij.2022.09.002
14. Price, L.R.: *Confirmatory factor analysis: Foundations and extensions*. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), International Encyclopedia of Education (Fourth Edition) (p. 607-618). 2023, Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10016-8
15. Ramirez Gomez, C.: *Classification and detection of topics on Twitter*: *Case study of the 2022 Colombian presidential elections*. 2023, http://repository.javeriana.edu.co/handle/10554/65484
16. Ros, F., Riad, R., Guillaume, S.: *PDBI: A partitioning Davies-Bouldin index for clustering evaluation*. In: Neurocomputing, Vol.528, 2023, p.178-199. https://doi.org/10.1016/j.neucom.2023.01.043
17. Schubert, E.: *Stop using the elbow criterion for k-means and how to choose the number of clusters instead.* In: ACM SIGKDD Explorations Newsletter, Vol.25(1), 2023, p.36-42. https://doi.org/10.1145/3606274.3606278
18. Wang, X., Xu, Y.: *An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index.* In: IOP Conference Series: Materials Science and Engineering, Vol.569(5), 2019, 052024. https://doi.org/10.1088/ 1757-899X/569/5/052024
19. Wang, Z., Zhang, G., Xing, Xet al.: *Comparison of dimensionality reduction techniques for multi-variable spatiotemporal flow fields.* In: Ocean

Engineering, Vol. 291, 2024, 116421. https://doi.org/10.1016/j.oceaneng.2023.116421

20. Wang, Z., Zhou, Y., Li, G.: *Anomaly Detection by Using Streaming K-Means and Batch K-Means*. In: the 5th IEEE International Conference on Big Data Analytics (ICBDA), 2020, p.11-17. https://doi.org/10.1109/ICBDA49040.2020.9101212
21. Zhang, Y., Wang, G., Wang, X., et al.: *TOC estimation from logging data using principal component analysis. Energy Geoscience*. Vol.4(4), 2023, 100197. https://doi.org/10.1016/j.engeos.2023.100197