TUZLULUK DAĞILIMLARININ BELİRLENMESINDE KÜMELEME ALGORİTMALARININ ETKİNLİĞİNİN DEĞERLENDİRİLMESİ

Perihan Karaköse[1]

Accepted: 11 Temmuz 2025
DOI: 10.47118/somatbd.1712594

ÖZET

Bu çalışmada, farklı kümeleme algoritmalarının tuzluluk verisi üzerindeki performansları karşılaştırılarak, karmaşık mekânsal desenleri sınıflandırmadaki başarıları incelenmiştir. Analizlerde KMeans, Agglomerative Clustering, DBSCAN, MeanShift, Birch, MiniBatch KMeans ve Spectral Clustering algoritmaları kullanılmıştır. Performans değerlendirmesi için siluet skoru temel ölçüt olarak kullanılmıştır. Elde edilen sonuçlara göre, MeanShift algoritması 0.79 siluet skoru ile en iyi sonucu verirken, KMeans ve MiniBatch KMeans algoritmaları 0.38 skoru ile orta düzeyde başarı göstermiştir. Agglomerative Clustering 0.34, Birch 0.31, DBSCAN 0.28 skoru elde etmiş, Spectral Clustering ise -0.35 skoruyla en düşük performansı sergilemiştir. Sonuçlar, özellikle heterojen ve sürekli değişim gösteren okyanus verilerinde yoğunluk adaptif yöntemlerin (örneğin MeanShift) üstün performans sunduğunu göstermektedir. Çalışmada kullanılan deniz suyu tuzluluk verisi, NOAA World Ocean Database'den temin edilmiştir (https://www.ncei.noaa.gov/products/world-ocean-database).

Anahtar Kelime: Deniz Suyu Tuzluluğu, Kümeleme Analizi, Yapay Zeka

ASSESSING THE EFFECTIVENESS OF CLUSTERING ALGORITHMS IN IDENTIFYING SALINITY DISTRIBUTIONS

ABSTRACT

In this study, the performances of various clustering algorithms were compared on salinity data to evaluate their effectiveness in classifying complex spatial patterns. The clustering methods applied included KMeans, Agglomerative Clustering, DBSCAN, MeanShift, Birch, MiniBatch KMeans, and Spectral Clustering. The silhouette score was used as the primary evaluation metric. According to the results, the MeanShift algorithm achieved the best performance with a silhouette score of 0.79, while KMeans and MiniBatch KMeans showed moderate success with scores of 0.38. Agglomerative Clustering, Birch, and DBSCAN yielded silhouette scores of 0.34, 0.31, and 0.28, respectively, whereas Spectral Clustering exhibited the poorest performance with a negative score of -0.35. These findings highlight that density-adaptive methods like MeanShift are particularly effective for analyzing heterogeneous and continuous oceanographic data. The sea water salinity dataset used in this study was obtained from the NOAA World Ocean Database (https://www.ncei.noaa.gov/products/world-ocean-database).

Keywords: Sea Water Salinity, Clustering Analysis, Artificial Intelligence

[1] Bartın Üniversitesi, Bartın Meslek Yüksekokulu, Bartın, TÜRKİYE, email: pkarakose@bartin.edu.tr ORCID:0000-0002-8894-6997

1. INTRODUCTION

Clustering analysis is an effective data mining method used to identify meaningful structures within large datasets (Jain and Dubes, 1988; Xu and Wunsch, 2005). It is particularly utilized in oceanography to classify the distributions of physical parameters such as temperature and salinity (Maze et al., 2017). In this study, a diverse set of clustering algorithms was deliberately selected to represent different clustering strategies—including partition-based, hierarchical, density-based, and graph-based approaches—to comprehensively assess their performance on ocean salinity data. The K-means algorithm is one of the most popular methods due to its computational simplicity, fast execution, and ability to generate interpretable clusters (Ahmed, Seraj and Islam, 2020; Chong, 2021). It was included as a baseline due to its widespread usage in environmental data analysis. However, it can perform poorly on clusters with non-spherical shapes and in datasets with significant noise (Arbelaitz et al., 2013). Agglomerative (hierarchical) clustering was selected because it does not require a pre-specified number of clusters and provides a detailed representation of data structure through dendrograms. This method offers particular advantages when working with multiscale oceanographic data, where the number of distinct water masses may not be known a priori (Ranzinger et al., 2024). DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is known for its robustness to noise and ability to detect clusters of arbitrary shape. Given the heterogeneous and dynamic nature of ocean salinity distributions, DBSCAN provides a density-adaptive approach suitable for capturing irregular patterns (Singh, Girdhar, and Dahiya, 2022). The Mean Shift algorithm identifies clusters by locating the modes of a density function and can determine the number of clusters automatically. This non-parametric nature makes it particularly valuable when the true number of water masses or salinity zones is unknown, which is often the case in exploratory oceanographic studies (Aidilof et al., 2025). Birch (Balanced Iterative Reducing and Clustering using Hierarchies) was included due to its ability to efficiently process large datasets in a single scan. Its hierarchical structure allows it to retain summary statistics, making it suitable for streaming or high-volume satellite-derived salinity datasets (Wahyuningrum et al., 2021). MiniBatch K-means, a faster variant of K-means, was chosen for its computational efficiency on large-scale data. Its inclusion allows assessment of how much performance is gained or lost when optimizing for speed over accuracy in ocean monitoring tasks (Hanji & Hanji, 2023). Spectral Clustering operates on similarity matrices and can capture non-convex clusters through graph partitioning methods. It is particularly useful when underlying patterns are better understood through network relationships rather than geometric proximity, which can apply to salinity gradients influenced by ocean currents and bathymetric constraints (Berahmand et al., 2022). To evaluate the accuracy of clustering results and the quality of the clusters, the Silhouette score was used. This metric provides an objective comparison of algorithm performance by measuring the balance between inter-cluster separation and intra-cluster cohesion (Topaloğlu, 2024).

With the recent rise in the application of artificial intelligence, clustering methods have become increasingly widespread in environmental sciences. In this study, seven widely used clustering algorithms were systematically selected and compared on the same salinity dataset, each chosen for its unique strengths and relevance to the complex structure of oceanographic data.

2. MATERIAL AND METHODS

In this research, multiple clustering algorithms were implemented and evaluated to explore the structural characteristics of salinity datasets. The clustering processes were conducted using Python programming language and its associated machine learning libraries. The performance of each method was assessed through silhouette score metrics to objectively compare the quality of the clustering results. A detailed description of the applied algorithms is presented below.

2.1. K-means Clustering

K-means clustering is a partitioning method that aims to divide a dataset into a predefined number of clusters by minimizing intra-cluster variance. Initially, random centroids are selected, and each data point is assigned to the nearest centroid based on Euclidean distance. Centroids are updated iteratively until convergence is achieved. Despite its simplicity and computational speed, K-means may struggle to accurately cluster data with non-spherical shapes or significant noise.

2.2. Agglomerative Clustering

Agglomerative clustering, a bottom-up hierarchical method, begins by treating each observation as an individual cluster. At each step, the two clusters with the smallest distance are merged until a single cluster encompassing all observations is formed or until the desired number of clusters is reached. This approach reveals the nEsted structure of data and does not require specifying the number of clusters in advance, making it suitable for complex datasets with varying scales.

2.3. DBSCAN

DBSCAN is a density-based clustering algorithm that groups together data points closely packed in space and identifies points in low-density areas as noise. Unlike K-means, it can discover clusters of arbitrary shape and does not necessitate specifying the number of clusters beforehand. Its resilience to noise and effectiveness in handling datasets with varying densities make it advantageous for analyzing heterogeneous oceanographic data.

2.4. Mean Shift Clustering

Mean Shift is a non-parametric clustering method that identifies clusters by shifting data points towards the nearest peak of the data density. It operates without requiring the number of clusters to be predetermined, adapting instead to the underlying structure of the data. This property makes it particularly useful for datasets where natural cluster boundaries are not well-defined, such as variations in salinity measurements across ocean regions.

2.5. Birch Clustering

Birch (Balanced Iterative Reducing and Clustering using Hierarchies) is an efficient clustering algorithm designed for large-scale datasets. It constructs a Clustering Feature (CF) tree that summarizes the data compactly and then applies clustering on this summary structure. Birch performs particularly well when memory resources are limited and offers rapid processing while maintaining a reasonable level of accuracy.

2.6. MiniBatch K-means

MiniBatch K-means is an extension of the K-means algorithm that updates cluster centroids based on small random subsets (mini-batches) of the data rather than the full dataset. This significantly speeds up the convergence process, making it particularly advantageous when working with massive datasets. Although it introduces slight approximations compared to standard K-means, it achieves comparable clustering quality with greatly reduced computational cost.

2.7. Spectral Clustering

Spectral clustering transforms the original data into a new space using the eigenvectors of a similarity matrix derived from the data points. By clustering in this transformed space, it captures complex relationships between points that are not evident in the original feature space. Spectral clustering is especially powerful for identifying non-convex and intricate cluster structures, providing a flexible approach suitable for analyzing complex environmental data distributions.

2.8. Silhouette Score Calculation

To assess the effectiveness of the clustering algorithms applied, the silhouette score was chosen as the primary evaluation metric. The silhouette score evaluates how well each data point fits within its own cluster (cohesion) relative to other clusters (separation), offering a holistic view of the clustering quality.

For each data point i, two metrics are computed:

- $a(i)$: the average distance between point i and all other points within the same cluster (intra-cluster distance),

- $b(i)$: the average distance between point i and the closest point in any other cluster (nearest-cluster distance).

The silhouette coefficient $s(i)$ for each point is caluculated using the Equation 1.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

The silhouette coefficient $s(i)$ ranges from -1 to 1:

- Values close to 1 suggest that the point is well-matched to its own cluster and poorly matched to other clusters.

- Values near 0 indicate that the point lies on the boundary between two clusters.

- Negative values imply that the point may be incorrectly assigned to its current cluster.

The overall silhouette score for a clustering result is the average of the silhouette coefficients across all data points in Equation 2.

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{2}$$

where n represents the total number of data points.

In this study, the silhouette score was used to objectively evaluate the clustering performance of various algorithms applied to the salinity dataset. A higher silhouette score indicates a more well-defined and meaningful clustering outcome.

3. Data SET

In this study, two distinct datasets—bottle.csv and cast.csv—were utilized for the analysis of salinity and other physical parameters. These datasets are based on measurement data provided by the NOAA National Centers for Environmental Information (NCEI) as part of the World Ocean Database (WOD) (NOAA, 2025). To enhance computational efficiency and reduce processing load, a random sample of 10,000 observations was extracted from each dataset. The bottle.csv dataset comprises in-situ measurements of seawater samples collected at various depths and geographical locations using bottle samplers. It includes a range of parameters such as temperature, salinity, pressure, and chemical composition. This dataset serves as a critical source for evaluating vertical ocean profiles and the spatial distribution of physical characteristics. Conversely, the cast.csv dataset provides supplementary metadata for each sampling event, including geographical coordinates (latitude and longitude), sampling dates and times, and sampling depths. This metadata facilitates the accurate spatial contextualization of the physical parameters under investigation. Prior to analysis, both datasets underwent preprocessing procedures. Irrelevant columns were removed, and selected variables were normalized to a range between 0 and 1 using the MinMaxScaler method. The integration of bottle.csv and cast.csv has established a robust foundation for clustering analyses aimed at characterizing the structural variability of salinity and classifying heterogeneity within oceanic environments. A representative spatial visualization of the sampled data points is presented in Figure 1. This 2D plot displays the geographic distribution of salinity values across the selected region, where the color scale corresponds to salinity levels. The figure effectively illustrates the spatial variability and gradient of salinity within the study area.

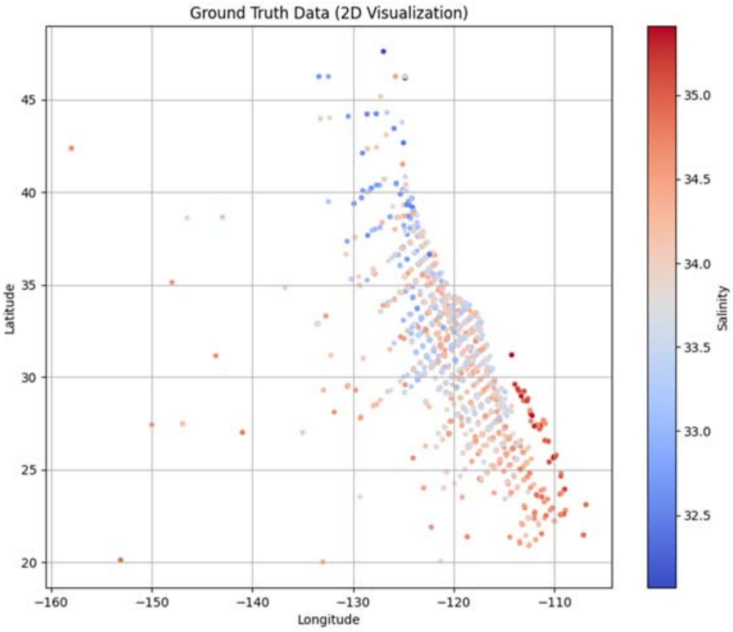
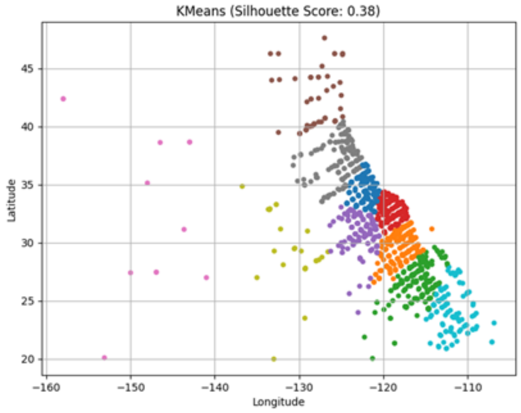


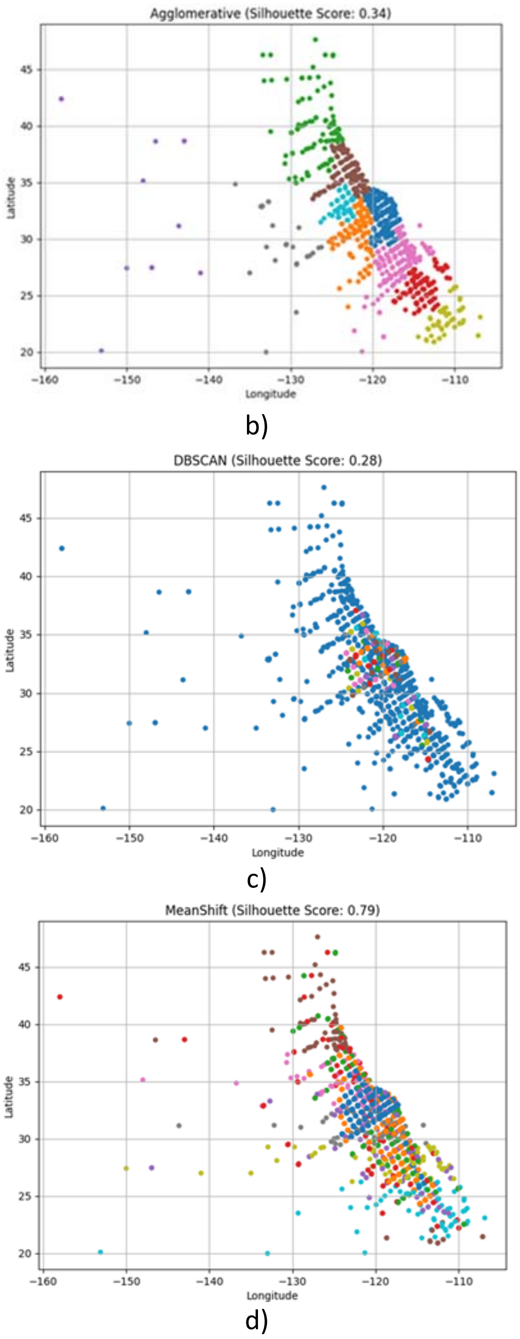
Figure 1. Ground Truth Salinity Data Map

4. SIMULATION AND DISCUSSIONS

In Figure 2.a, the results of the KMeans clustering algorithm are presented. The method appears to generate compact and well-defined clusters, particularly in regions with dense sampling. The silhouette score of 0.38 indicates a moderate level of internal cohesion and separation between clusters. The algorithm successfully captures the underlying structure of the spatial salinity distribution, especially around the central region. However, in more sparsely populated areas, the cluster boundaries become less distinct. In Figure 2.b, the clustering output from the Agglomerative Clustering algorithm is illustrated. While the overall cluster configuration resembles that of KMeans, the silhouette score is slightly lower (0.34), suggesting less effective inter-cluster separation. Some clusters are elongated and less compact, particularly toward the northern and southern extents of the study region. This observation suggests that while the method is suitable for capturing hierarchical relationships, its performance may be limited when dealing with high variability in spatial density. In Figure 2.c, the DBSCAN algorithm reveals a scattered distribution of clusters, accompanied by a significant number of noise points. This density-based approach identifies dense regions as clusters and labels the rest as outliers, resulting in a silhouette score of 0.28. Although DBSCAN is capable of identifying arbitrary-shaped clusters, its performance in this application appears constrained due to the variable spatial distribution of sampling points. The prevalence of noise may reflect limitations in capturing continuous spatial gradients. In Figure 2.d, MeanShift clustering demonstrates superior performance with the highest silhouette score (0.79) among all methods evaluated. The algorithm effectively identifies meaningful cluster centers based on data density and provides a highly coherent segmentation of the spatial domain. The smooth transition between clusters and their alignment with geographic gradients of salinity suggest that MeanShift is particularly well-suited for this type of oceanographic data. In Figure 2.e, the Birch clustering algorithm yields moderately distinct groupings, with a silhouette score of 0.31. While clusters in the central regions are relatively compact, peripheral areas display more dispersed and weakly defined boundaries. Birch is known for its efficiency on large datasets, but in this case, its ability to differentiate between clusters in sparsely populated areas appears limited. In Figure 2.f, MiniBatch KMeans produces a cluster structure nearly identical to standard KMeans, achieving an identical silhouette score of 0.38. This algorithm offers faster computation by processing mini-batches of data, which is advantageous for large-scale applications. The consistency in performance and clustering quality indicates its viability as a faster alternative to KMeans without significant loss in accuracy.



a)



Agglomerative (Silhouette Score: 0.34)

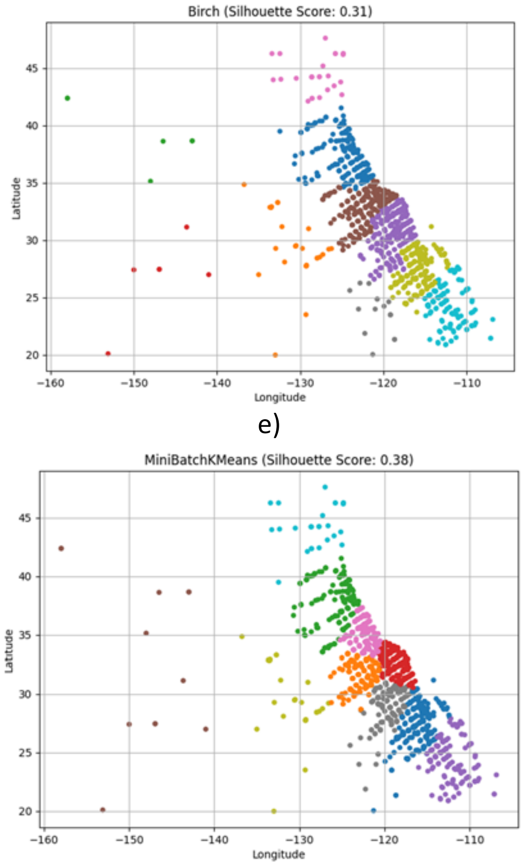
b)

DBSCAN (Silhouette Score: 0.28)

c)

MeanShift (Silhouette Score: 0.79)

d)



e)

f)

Figure 2. Spatial data clusters generated using different clustering algorithms. (a) KMeans (Silhouette: 0.38). (b) Agglomerative Clustering(Silhouette: 0.34). (c) DBSCAN (Silhouette: 0.28). (d) MeanShift(Silhouette: 0.79). (e) Birch (Silhouette: 0.31). (f) MiniBatch KMeans (Silhouette: 0.38).

5. CONCLUSIONS

In this study, the performances of various clustering algorithms were compared on salinity data to evaluate their ability to classify complex spatial patterns. Salinity was selected as the primary variable due to its fundamental role in determining marine physical processes. The clustering algorithms applied included KMeans, Agglomerative Clustering, DBSCAN, MeanShift, Birch, MiniBatch KMeans, and Spectral Clustering. Each algorithm was implemented following identical preprocessing steps (data normalization and sampling), and their performance was assessed using the silhouette score as the primary evaluation metric. According to the silhouette score results: The MeanShift algorithm achieved the highest silhouette score of 0.79, producing coherent clusters that best reflected the natural gradients of salinity data. KMeans and MiniBatch KMeans attained silhouette scores of 0.38, indicating moderate success, particularly in densely sampled regions. Agglomerative Clustering yielded a silhouette score of 0.34, producing clusters similar to KMeans but more dispersed. DBSCAN, due to its density-based structure, formed irregular clusters but achieved a lower silhouette score of 0.28, which can be attributed to the high number of noise points.

Birch produced relatively compact clusters in the central areas but more scattered clusters at the periphery, with a silhouette score of 0.31. Spectral Clustering exhibited the poorest performance, with a negative silhouette score of -0.35, indicating its inadequacy in successfully separating the clusters within the dataset.

These findings clearly demonstrate that the choice of clustering algorithm is critical in the analysis of environmental spatial datasets. Particularly for continuous and heterogeneous oceanographic data, density-adaptive methods such as MeanShift provide superior results.

For future studies, it is recommended to explore more advanced and flexible clustering approaches. For instance, Gaussian Mixture Models (GMM) can provide probabilistic soft clustering, which may better reflect gradual transitions in oceanographic parameters. In addition, deep learning-based methods such as autoencoder-based clustering (e.g., DEC – Deep Embedded Clustering) can be utilized to capture non-linear structures in high-dimensional spatiotemporal data. The integration of such methods may enhance the interpretability and accuracy of clustering results, especially in large-scale and multi-variable ocean datasets. Ultimately, future research should focus on developing customized clustering frameworks that account for the physical characteristics and spatial dynamics of marine data, potentially combining data-driven and physics-informed approaches for more robust spatial classification.

KAYNAKLAR

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645–678.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Radcenco, M. L., Lenca, P., ... & Le Goff, C. (2017). Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. Progress in Oceanography, 151, 275–292.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

Chong, B. (2021). K-means clustering algorithm: a brief review. Academic Journal of Computing & Information Science, 4(5), 37–40.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243–256.

Ranzinger, M., Heinrich, G., Kautz, J., & Molchanov, P. (2024). Am-radio: Agglomerative vision foundation model reduce all domains into one. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12490–12500).

Singh, H. V., Girdhar, A., & Dahiya, S. (2022, May). A literature survey based on DBSCAN algorithms. In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 751–758). IEEE.

Aidilof, H. A. K., Rosnita, L., Kurniawati, K., & Ikhwani, M. (2025). Clustering of data monitoring water quality using mean-shift clustering method. Journal of Computer Science, Information Technology and Telecommunication Engineering, 6(1).

Wahyuningrum, T., Khomsah, S., Suyanto, S., Meliana, S., Yunanto, P. E., & Al Maki, W. F. (2021, December). Improving clustering method performance using K-means, mini batch K-means, BIRCH and spectral. In 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 206–210). IEEE.

Hanji, S., & Hanji, S. (2023, January). Towards performance overview of mini batch K-means and K-means: Case of four-wheeler market segmentation. In International Conference on Smart Trends in Computing and Communications (pp. 801–813). Singapore: Springer Nature Singapore.

Berahmand, K., Mohammadi, M., Faroughi, A., & Mohammadiani, R. P. (2022). A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix. Cluster Computing, 25(2), 869–888.

Topaloğlu, F. (2024). Saldırı tespit sistemlerinde K-Means algoritması ve Silhouette metriği ile optimum küme sayısının belirlenmesi. Bilişim Teknolojileri Dergisi, 17(2), 71–79.

NOAA National Centers for Environmental Information (NCEI). (2025). *World Ocean Database*. National Oceanic and Atmospheric Administration. Retrieved from https://www.ncei.noaa.gov/products/world-ocean-database