International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Data-Driven Insights into the Spatial Distribution and Conservation Status of Protected Areas in the Philippines: A WEKA Case Study

Clear Fallaria A[1], Geli Joe Mari C[2], Santillana Sanger G[3], Yujoco Shane L[4]

[1,2,3,4]Student, BSIS, SNSU

Abstract

Purpose: This study applies unsupervised data mining techniques in WEKA to examine patterns in the spatial distribution and conservation status of protected and conserved areas represented in the protected_conserved_areas_wdpca_points dataset. It specifically compares EM, SimpleKMeans, Hierarchical Clusterer, and FarthestFirst to determine which clustering output provides the clearest and most useful segmentation of the records.

Design/Methodology/Approach: The study used the normalized WEKA relation containing 144 protected/conserved area records and 33 attributes. Based on the uploaded CSV, all records are coded to the Philippines through the prnt_iso3 and iso3 fields, both recorded as PHL. The dataset includes site-level place names such as Puerto Galera, Palawan Biosphere Reserve, Olango Island Wildlife Sanctuary, Agusan Marsh Wildlife Sanctuary, Naujan Lake National Park, Puerto Princesa Subterranean River National Park, Las Piñas-Parañaque Critical Habitat and Ecotourism Area, Albay Biosphere Reserve, and several local conservation and marine protected area records.

Findings: The results show that all models detected a dominant majority pattern and a much smaller distinct subset. EM provided the most informative clustering solution because it automatically selected three clusters and separated the dataset into 99 (69%), 37 (26%), and 8 (6%) records, while the other algorithms mostly produced a coarse two-cluster split. The smallest cluster consistently captured a distinct group dominated by formally designated protected areas with international recognition and stronger conservation-status signals.

Research Limitations/Implications: The analysis was constrained by the provided WEKA outputs and screenshots. Because explicit geographic coordinates were not available in the reported 33 attributes, spatial interpretation relied on site-level proxies such as realm, reported area, site type, and designation rather than a full GIS map-based analysis. The study therefore emphasizes attribute-based spatial patterning rather than coordinate-level spatial autocorrelation.

Practical Implications: The findings can support protected area managers, planners, and researchers by showing how clustering can segment protected/conserved area records into operationally meaningful groups. EM is recommended for analytical profiling, while HierarchicalClusterer and FarthestFirst are

useful for isolating a small special-status subset that may require separate planning attention. The results also highlight the need to improve complete reporting for management and governance fields

Originality/Value: This paper aligns the provided WDPA/WDPCA-related WEKA runs with a complete research-paper structure and interprets clustering results in a conservation context. It contributes a clear, template-aligned write-up that links cluster outputs to conservation status, site classification, and practical management implications.

Abstract

This study examined how data mining techniques in WEKA can be used to analyze patterns in the spatial distribution and conservation status of protected and conserved areas using the normalized protected_conserved_areas_wdpca_points dataset. The dataset is country-specific: all 144 records are coded to the Philippines using the ISO3 country code PHL. The records represent site-level protected and conserved areas, including internationally recognized protected areas, local conservation areas, and locally managed marine protected areas. The study analyzed 144 instances and 33 attributes through four clustering algorithms: EM, SimpleKMeans, HierarchicalClusterer, and FarthestFirst. The results show that the dataset contains a clear clustering structure dominated by a broad group of OECM-like or locally managed marine protection records, alongside a much smaller but highly distinct subset associated with formally designated protected areas and stronger international-recognition signals. Spatial-distribution proxies and conservation-status attributes such as realm, reported area, site type, designation, no-take category, and status year were important in explaining the observed groupings. Among the algorithms, EM provided the most informative and interpretable solution because it identified a three-cluster structure of 99, 37, and 8 records rather than only a coarse binary split. The findings further indicate that cluster-based analysis can support differentiated management review, data-quality checking, and reporting improvement. Overall, the study confirms that WEKA-based clustering is a practical exploratory tool for profiling protected/conserved area records when the goal is segmentation and interpretation rather than predictive classification.

1. Introduction

Protected areas and other effective area-based conservation measures are central to biodiversity conservation, ecosystem protection, and landscape governance (UNEP-WCMC, 2024). Large protected-area databases contain heterogeneous records on designations, management arrangements, marine and terrestrial extent, and governance structures. When these records are examined as a whole, important patterns may remain hidden.

Data mining provides a way to reveal these patterns systematically through exploratory grouping and profile detection. Recent reviews also show that clustering remains one of the core unsupervised techniques for discovering structure in complex datasets across multiple fields (Singh et al., 2024; Dinh et al., 2025).

The present study focuses on the protected_conserved_areas_wdpca_points dataset processed in WEKA. The dataset represents protected and conserved area records located in the Philippines. This is confirmed by the CSV country-code fields prnt_iso3 and iso3, both of which are recorded as PHL. The dataset represents a point-based protected/conserved area relation with 144 instances and 33 attributes. The available fields include site identifiers, site type, place or site name, designation, IUCN and international criteria fields, realm, reported area, no-take information, status, governance type, ownership type,

management information, verification, and country codes. These attributes make the dataset suitable for clustering because they allow records to be grouped according to shared conservation and management characteristics.

Unlike supervised learning, clustering does not require a predefined class label. Instead, it groups records according to similarity. This is appropriate for the current study because the objective is not to predict a target category, but to explore how protected/conserved area records separate into meaningful segments. By comparing EM, SimpleKMeans, HierarchicalClusterer, and FarthestFirst in WEKA, the study identifies the clustering technique that provides the clearest and most useful representation of the dataset (Frank et al., 2016).

2. Problem Statement

Protected area datasets are rich in information but difficult to interpret manually because records differ across designation type, management structure, governance arrangement, and marine or terrestrial context. When records are reviewed one by one, it becomes difficult to identify dominant patterns, exceptional subsets, and meaningful relationships among conservation-status variables.

The challenge is compounded when the dataset contains both protected areas (PAs) and other effective area-based conservation measures (OECMs), because these categories may reflect different governance arrangements, designation systems, and reporting practices. A purely descriptive reading of the dataset does not clearly reveal whether the records form one homogeneous group or several internally consistent subgroups.

This study therefore addresses the need for an exploratory data mining approach that can organize the protected/conserved area records into interpretable clusters. By using WEKA-based clustering, the study seeks to determine whether the dataset contains stable grouping patterns and which algorithm provides the most useful representation for conservation analysis.

3. Objectives

The primary goal of this study is to apply data mining techniques in WEKA to analyze patterns in the spatial distribution and conservation status of protected and conserved areas using clustering methods.

The specific objectives are as follows:

1. To examine the clustering structure of the normalized protected_conserved_areas_wdpca_points dataset using EM, SimpleKMeans, HierarchicalClusterer, and FarthestFirst.
2. To identify how site type, designation, realm, reported area, status, governance, ownership, and management-related variables vary across clusters.
3. To compare the interpretability and analytical usefulness of the four clustering algorithms.
4. To derive planning and data-quality implications from the observed clustering patterns.

The study will address the following research questions:

1. What clustering structure emerges from the protected/conserved area dataset when analyzed in WEKA?
2. How do spatial-distribution proxies and conservation-status attributes vary across the identified clusters?
3. Which clustering algorithm provides the most useful and interpretable grouping for this dataset?
4. What management and reporting implications can be derived from the clustering results?

4 Scope of the Study

This study focuses on the application of unsupervised data mining techniques to a normalized protected/conserved area point dataset in WEKA. The scope is defined by the following boundaries:

1. Dataset Scope. The study uses the normalized WEKA relation based on protected_conserved_areas_wdpca_points with 144 instances and 33 attributes. The dataset is country-specific and covers the Philippines only, as shown by the ISO3 country codes PHL in both the prnt_iso3 and iso3 fields.
2. Algorithm Scope. The analysis is limited to EM, SimpleKMeans, HierarchicalClusterer, and FarthestFirst, which were the algorithms actually run by the user in WEKA.
3. Evaluation Scope. The comparison is based on the model outputs supplied by the user, including cluster counts, centroids or profile summaries, log likelihood, sum of squared errors, run information, and visualization screenshots.
4. Contextual Scope. The interpretation is limited to what can be inferred from the supplied attributes and Weka outputs, especially variables connected to site type, designation, realm, reported area, no-take status, governance, management, and conservation status.
5. Spatial Scope. Because the supplied run information does not include latitude and longitude variables, the study interprets spatial distribution through point-dataset grouping and spatial proxies such as realm, reported area, and site-level categories rather than coordinate-based spatial statistics.

Limitations of the Study

Despite its contributions, this study has certain limitations:

1. Source Data Limitation. The analysis is based on the normalized CSV relation, the WEKA run information, and the visualization screenshots supplied for the study. Although these materials are sufficient for exploration clustering, they do not provide the full breadth of source metadata or external validation information that could support deeper ecological interpretation.
2. Spatial Data Limitation. The 33 available reported attributes do not include explicit coordinate fields in the clustering outputs used for interpretation. As a result, spatial distribution was interpreted through proxies such as realm, reported area, site type, and designation rather than through direct GIS mapping or coordinate-based spatial statistics.
3. Unsupervised Model Limitation. The clustering methods used in this study are exploratory and unsupervised. Because there is no external ground-truth label, the study cannot evaluate predictive accuracy in the same way that supervised classification studies do. The results therefore describe meaningful groupings rather than definitive predictive categories.
4. Interpretation Limitation. EM produces probabilistic cluster profiles, while the other algorithms generate centroid-based or partition-based summaries. This means that some cluster descriptions must be interpreted qualitatively, especially for nominal conservation fields, and should not be treated as exact causal explanations of protected-area status.

These limitations suggest directions for future research, such as incorporating full raw metadata, adding coordinate-based spatial analysis, improving complete reporting, and validating the cluster patterns against ecological or administrative benchmarks.

5 Related Work

Recent studies published from 2022 onward show that protected-area analysis is increasingly shaped by

large official datasets, improved data infrastructures, and machine-learning-assisted monitoring. The Protected Planet Report 2024 emphasized that current protected and conserved area reporting is central to evaluating progress toward Target 3 of the Kunming–Montreal Global Biodiversity Framework and that official WDPA/WDPCA records now support broader evidence-based conservation assessment (UNEP-WCMC, 2024). This confirms that structured protected-area datasets are not only descriptive repositories but also important analytical inputs for contemporary conservation planning.

Recent conservation informatics literature also shows that machine learning is now widely used to analyze threats, conservation measures, and biodiversity management patterns. Branco et al. (2023) reported that machine-learning methods have become increasingly common in conservation analysis because they can detect non-obvious relationships in complex ecological and management datasets. In protected-area contexts, Urbano et al. (2024) further argued that efficient data management and better information systems are essential for improving biodiversity monitoring and decision support inside protected areas.

At the same time, newer studies demonstrate that protected-area monitoring is moving toward integrated analytical workflows. Mouillot et al. (2024) showed that protected areas exhibit measurable socioeconomic and environmental patterning at the global scale, which supports the value of grouping site records according to shared characteristics rather than treating all protected areas as uniform. Likewise, van der Plas et al. (2025) showed that machine learning, remote sensing, and citizen science can be combined to strengthen protected-area monitoring. These studies support the logic of the present research, in which clustering is used to reveal hidden structure in protected/conserved area records.

The present study contributes by aligning a protected/conserved area dataset with a complete WEKA-based clustering paper. Rather than focusing on predictive classification, it emphasizes exploratory segmentation and the interpretation of cluster structure in relation to conservation status, site type, and management characteristics.

6 Methodology

This study used a Knowledge Discovery in Databases (KDD)-oriented workflow to examine the structure of the protected/conserved area dataset through clustering. The methodology was designed to remain faithful to the supplied WEKA runs and the sample research guide provided by the user. The workflow includes data acquisition, data understanding, preprocessing confirmation, clustering execution, model evaluation, and interpretation.

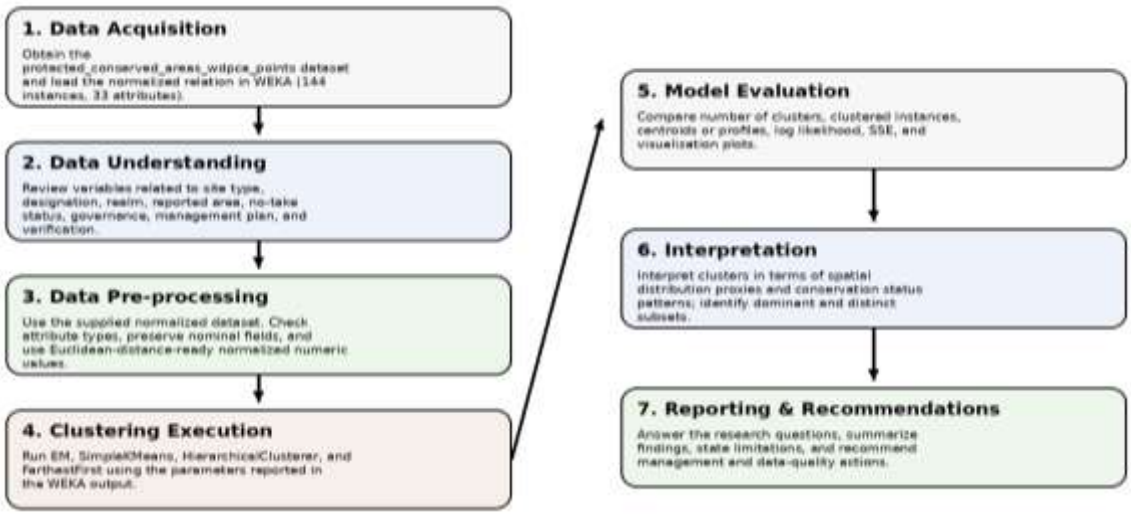

Fig. 1. Data Mining Process Flow of the Study

Figure 1 presents the methodological flow used in the study. The process begins with acquiring the normalized protected/conserved area relation in WEKA, then reviewing the conservation and management variables contained in the dataset. After confirming preprocessing, the study executes the four clustering algorithms, compares the outputs, and interprets the resulting groups in relation to spatial-distribution proxies and conservation-status features. The final stage translates the cluster patterns into research answers and recommendations.

6.1 Data Collection

The study used the protected_conserved_areas_wdpca_points relation already processed in WEKA. Based on the run information supplied by the user and the uploaded CSV file, the working relation contained 144 instances and 33 attributes. The country covered by the dataset is the Philippines, identified by the ISO3 code PHL in both the parent and record-level country fields. The attributes included site identifiers and names, site type, designation, designation type, IUCN category, international criteria, realm, reported marine area, reported area, no-take information, conservation status, status year, governance type, ownership type, management authority, management plan, conservation objective, verification status, inland waters field, metadata identifiers, ISO country codes, and subtype fields.

The dataset corresponds to a protected and conserved area point structure and is consistent with official protected-area reporting frameworks in which records include both protected areas and OECMs. The analysis therefore treats each record as a site-level conservation unit located in the Philippines and described by its reported status and management attributes. The place names in the dataset include internationally recognized areas such as Palawan Biosphere Reserve, Olango Island Wildlife Sanctuary, Agusan Marsh Wildlife Sanctuary, Naujan Lake National Park, Puerto Princesa Subterranean River National Park, and Albay Biosphere Reserve, as well as local conservation areas and locally managed marine protected areas across Philippine municipalities, islands, reefs, fish sanctuaries, and marine reserves.

6.2 Data Pre-processing

The supplied relation name indicates that the dataset had already been processed using the WEKA Normalize filter before clustering. This means that numeric attributes such as object identifiers, reported marine area, reported area, no-take area, status year, and metadata-related numeric fields were scaled to a comparable range. This preprocessing step is especially important for distance-based algorithms because it reduces dominance by variables with larger raw numerical magnitudes.

No additional transformation beyond the supplied normalized relation was assumed. Nominal conservation fields such as site type, designation, realm, governance, and management-related variables were retained for clustering interpretation. The SimpleKMeans output also notes that missing values were globally replaced with mean or mode during model building, which is consistent with Weka's handling of incomplete data for centroid-based clustering.

6.3 Data Mining Techniques

This subsection describes the clustering methods applied to the normalized protected/conserved area relation. Because the study is exploratory and unsupervised, the techniques were selected to capture different clustering logics: probabilistic clustering, centroid-based partitioning, hierarchical separation, and farthest-seed partitioning. The descriptions below explain the specific role of each algorithm before the formal summary tables are presented.

1. EM. Expectation-Maximization was used because it supports probabilistic membership and automatic cluster selection through cross-validation. This makes it appropriate for detecting latent structures in

a dataset that contains mixed conservation profiles rather than obviously pre-separated groups.

2. SimpleKMeans. SimpleKMeans was used as a centroid-based partitioning technique. It provides a compact operational summary of the dataset and is useful when the researcher needs a direct comparison of dominant and secondary groups through final cluster centroids.
3. HierarchicalClusterer. HierarchicalClusterer with single linkage was used to identify broad separations and possible nested relationships among records. It is useful for determining whether a small subset of sites consistently separates from a much larger majority cluster.
4. FarthestFirst. FarthestFirst was used as a fast distance-based method that emphasizes extreme separation by choosing farthest seeds first. In this study, it served as a check on whether the minority subset observed in other models remained stable under a different partitioning principle.

Table 1. Clustering Algorithms Used in the Study

Algorithm	Description	Purpose in Study
EM	Expectation-Maximization clustering with probabilistic membership and automatic cluster selection.	Used to identify the most informative latent grouping structure.
SimpleKMeans	Centroid-based partitioning using Euclidean distance.	Used to generate a concise operational segmentation of the dataset.
HierarchicalClusterer	Single-linkage hierarchical clustering using Euclidean distance.	Used to observe broad cluster separation and isolate distinct subsets.
FarthestFirst	Distance-based partitioning that selects farthest seeds first.	Used to test rapid separation of extreme or atypical records.

Table 1 summarizes the four clustering algorithms and clarifies their analytical role in the study.

Table 2. Parameter Settings from the Provided WEKA Runs

Algorithm	Main Settings	Key Run Statistic
EM	-I 100, -N -1, cross-validation enabled, seed 100	3 clusters selected; 15 iterations; log likelihood = 5.13566
SimpleKMeans	-N 2, Euclidean distance, seed 10, max iterations 500	2 clusters; 3 iterations; SSE = 673.0986
HierarchicalClusterer	-N 2, single linkage, Euclidean distance	2 clusters; 8/136 split
FarthestFirst	-N 2, seed 1	2 clusters; 136/8 split

Table 2 presents the actual WEKA parameter settings and key run statistics used as the basis for model comparison. These settings were taken directly from the supplied WEKA runs so that the interpretations in the Results section remain fully aligned with the models actually executed in the study.

6.4 Tools and Technologies Used

The primary tool used in this research is WEKA, an open-source machine learning workbench that supports data preprocessing, clustering, visualization, and model comparison (Frank et al., 2016). The study also relied on the screenshots supplied by the user from WEKA's cluster visualization window. For

research-paper preparation, the supplied outputs were converted into structured tables, a methodology figure, and CSV-based graphs to align the document with the requested template flow.

7 Data Analysis

Because the study is unsupervised, data analysis focused on the structure and interpretability of the clusters rather than predictive accuracy. The supplied outputs were examined in terms of number of clusters, clustered-instance distribution, centroid or profile description, model-specific statistics, and cluster visualization plots.

EM was assessed using its cross-validated cluster selection, iteration count, cluster proportions, and qualitative profile table. SimpleKMeans was assessed using its final centroids, cluster sizes, and within-cluster sum of squared errors. HierarchicalClusterer was assessed using the observed two-cluster split and the separation pattern reflected in the visualization. FarthestFirst was assessed through its centroids and its ability to isolate a minority subset of records.

The analysis also compared whether the same subset of records repeatedly separated across algorithms. If multiple algorithms isolate a similar minority cluster, this suggests that the subset is structurally distinct rather than an artifact of one method.

8 Results

Dataset Profile

The provided relation contains 144 protected/conserved area records described by 33 attributes. All records are coded to the Philippines through the ISO3 code PHL. The available variables show that the dataset combines protected area and OECM records and captures status, governance, ownership, management, designation, realm, and area-related fields. This attribute composition is appropriate for exploration clustering because it allows the records to be segmented according to conservation meaning rather than simple identifier similarity.

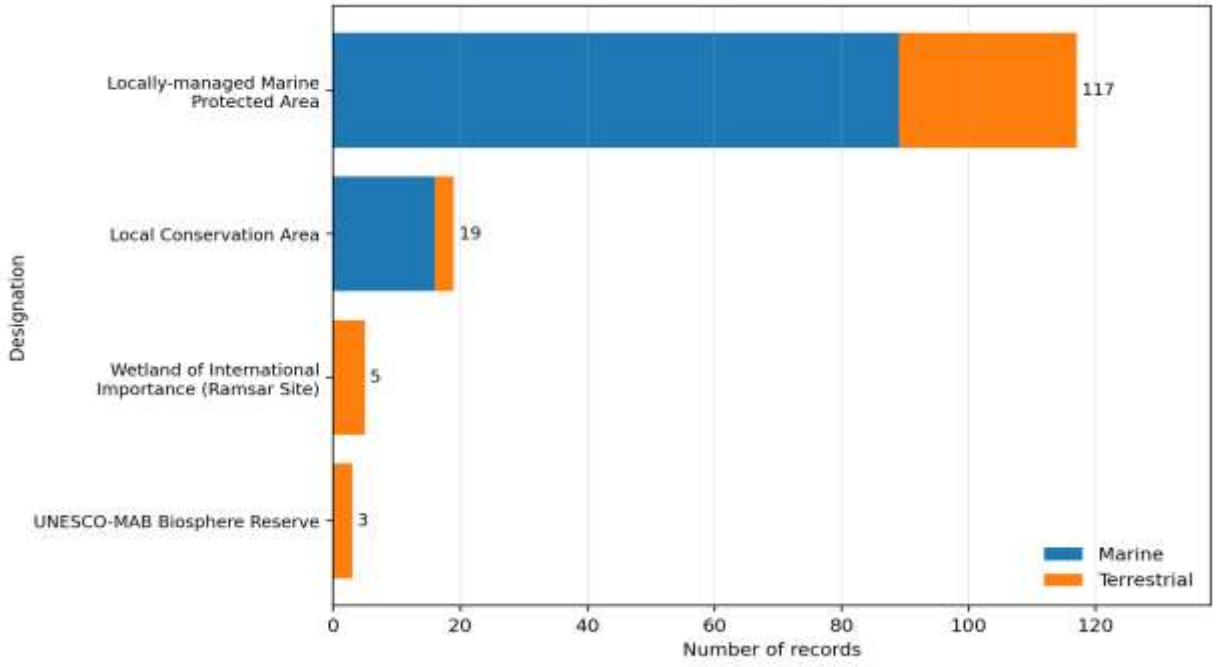

Fig. 2. Distribution of Protected/Conserved Area Records by Realm and Designation

Figure 2 was generated directly from the uploaded CSV file and summarizes how the 144 records are distributed across designation categories and ecological realm. The graph shows that the dataset is dominated by Locally managed Marine Protected Area records (117), most of which belong to the marine realm. Local Conservation Area entries form a much smaller secondary group (19), while the internationally designated records—Wetlands of International Importance (Ramsar Sites) and UNESCO-MAB Biosphere Reserves—appear only in the terrestrial realm. This CSV-based pattern supports the clustering results because it confirms that the dataset is not evenly distributed across conservation designations and that a small formally designated subgroup exists alongside a much larger locally managed marine majorit

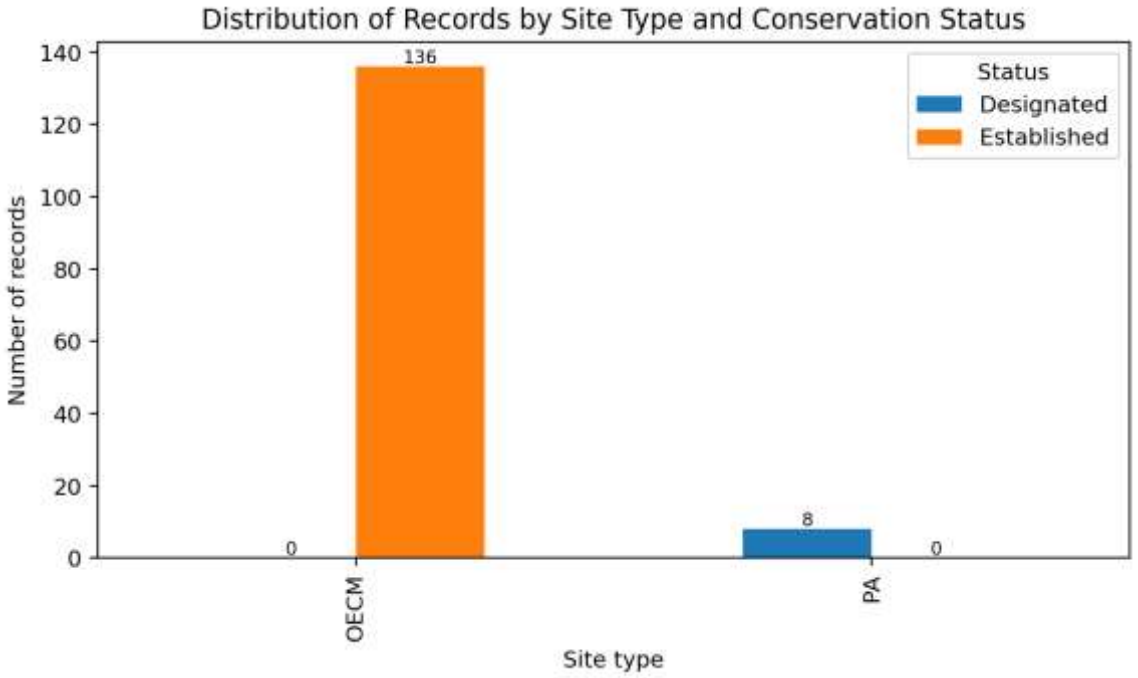

Fig. 3. Distribution of Records by Site Type and Conservation Status (Generated from CSV)

Figure 3 shows a near-perfect separation between the two site types and their recorded status. All records in the CSV are marked as established, whereas all PA records are marked as designated. This pattern is important because it visually confirms that the dataset contains one dominant operational conservation group and one much smaller formally designated protected-area group.

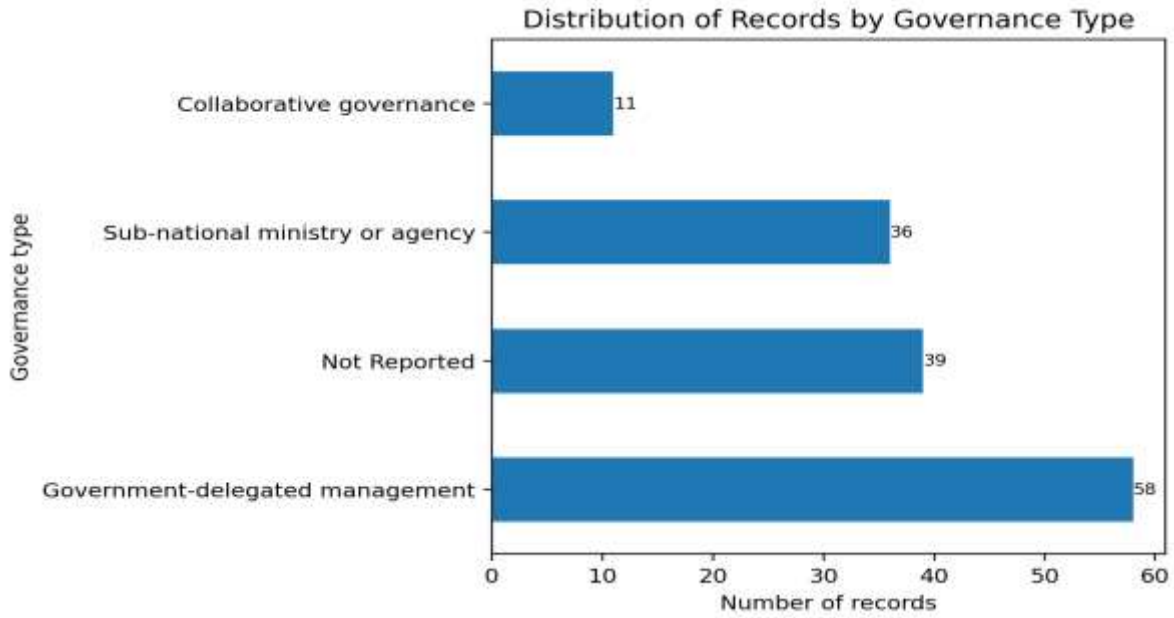

Fig. 4. Distribution of Records by Governance Type

Figure 4 shows that government-delegated management is the most frequent governance type, followed by not reported governance and sub-national ministry or agency management. Collaborative governance appears least frequently. This indicates that the dataset is dominated by state-linked or delegated governance arrangements, a pattern that is consistent with the governance-related cluster descriptions.

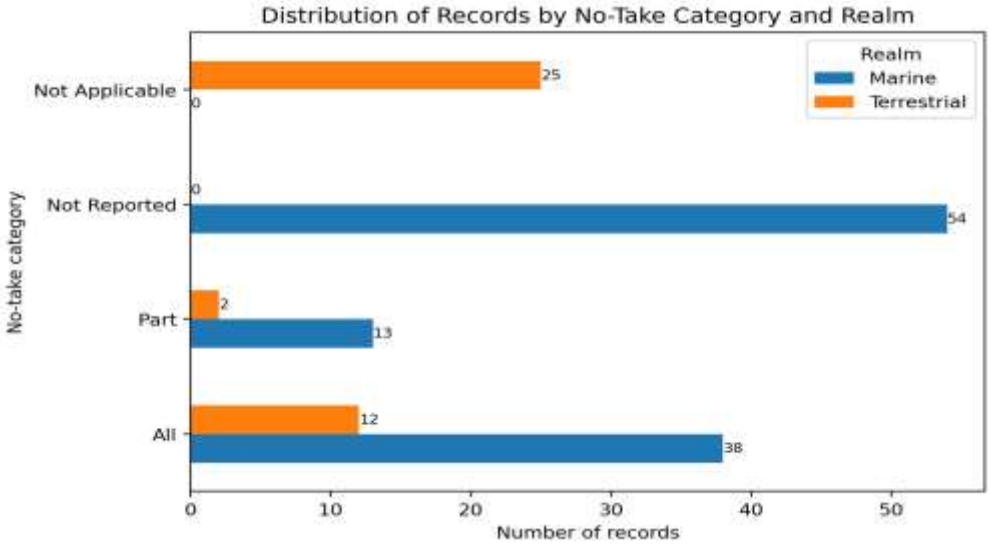

Fig. 5. Distribution of Records by No-Take Category and Realm

Figure 5 shows that marine records are distributed across the categories "All," "Part," and "Not Reported," while terrestrial records are concentrated mainly in "Not Applicable." This distribution indicates that no-take zoning is a meaningful descriptor for marine conservation entries, whereas it is less relevant to many terrestrial records.

Different Clustering Algorithms

In this study, the four clustering algorithms—EM, SimpleKMeans, HierarchicalClusterer, and FarthestFirst—were evaluated using WEKA to analyze patterns in the spatial distribution and conservation

status of protected and conserved areas. The models were applied to the same normalized data set to ensure a fair and consistent comparison of clustering behavior across algorithms. A comparison matrix in Table 4 summarizes the major clustering results, including the number of clusters formed, the distribution of clustered instances, and the general interpretation of each model. The succeeding WEKA visualizations further show how each algorithm separated the protected/conserved area records into dominant and minority groups, revealing both the broad majority pattern and the distinct special-status subset within the dataset.

Table 3. Comparison of Clustering Results

Algorithm	No. of Clusters	Clustered Instances	General Interpretation
EM	3	99 (69%), 37 (26%), 8 (6%)	Most informative segmentation; reveals dominant, secondary, and highly distinct minority clusters.
SimpleKMeans	2	27 (19%), 117 (81%)	Stable centroid-based split; useful for operational grouping.
HierarchicalClusterer	2	8 (6%), 136 (94%)	Strong separation of a small distinct subset from the large majority.
FarthestFirst	2	136 (94%), 8 (6%)	Confirms the same broad separation pattern identified by HierarchicalClusterer.

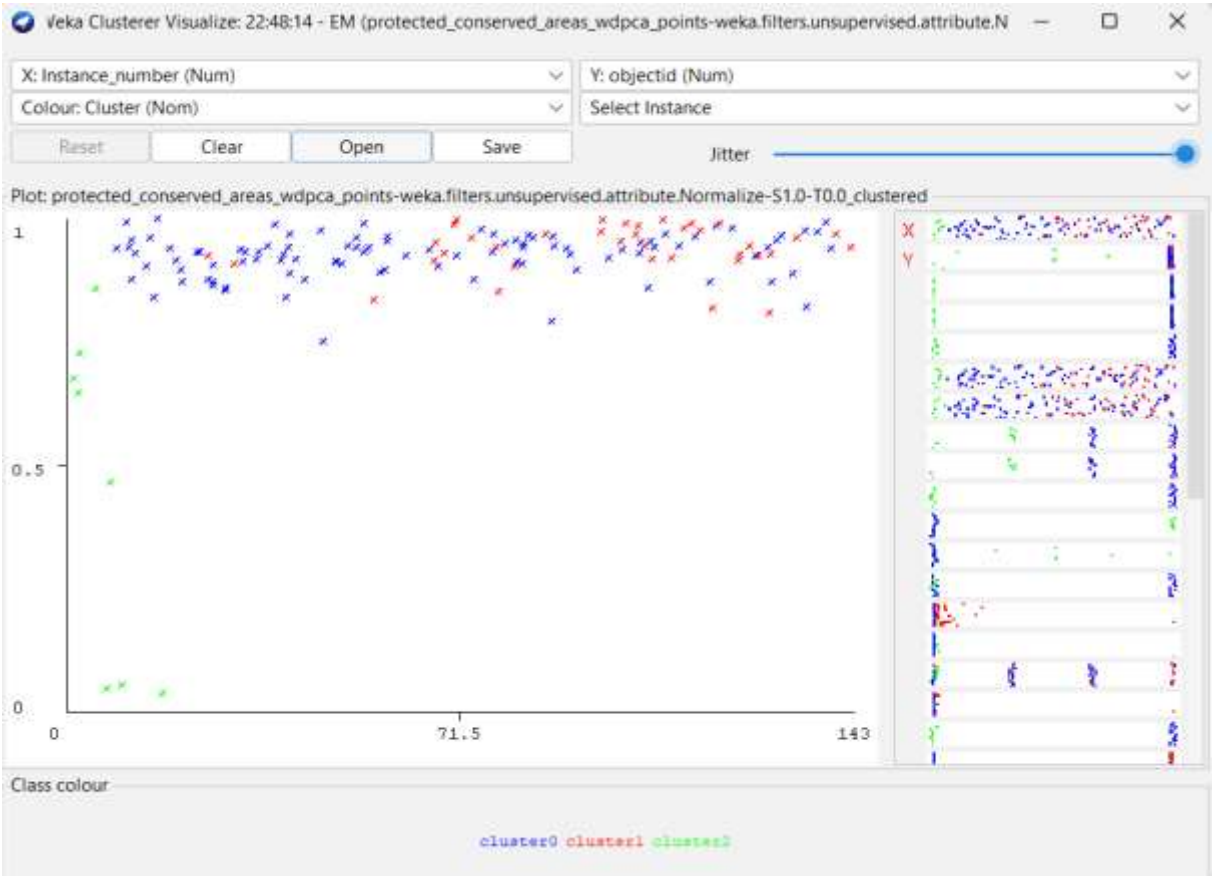

Fig. 6. EM Cluster Visualization in WEKA

Figure 7 shows the EM visualization with three cluster colors and provides the clearest evidence that the dataset does not consist of only one homogeneous group. The small green cluster is visibly separated from the dominant blue and red distributions, while the blue cluster forms the largest concentration and the red cluster appears as an intermediate subgroup rather than a random fragment. This visual pattern matches the EM output, which selected three clusters through cross-validation and produced the most nuanced structure in the study: 99 records in the dominant cluster, 37 records in a secondary OECM-oriented cluster, and 8 records in a highly distinct minority cluster associated with formally designated protected areas, international recognition signals, later normalized status-year values, and larger reported total area. The figure therefore supports the interpretation that EM captures both the broad majority pattern and the special-status subset more effectively than the other clustering approaches.

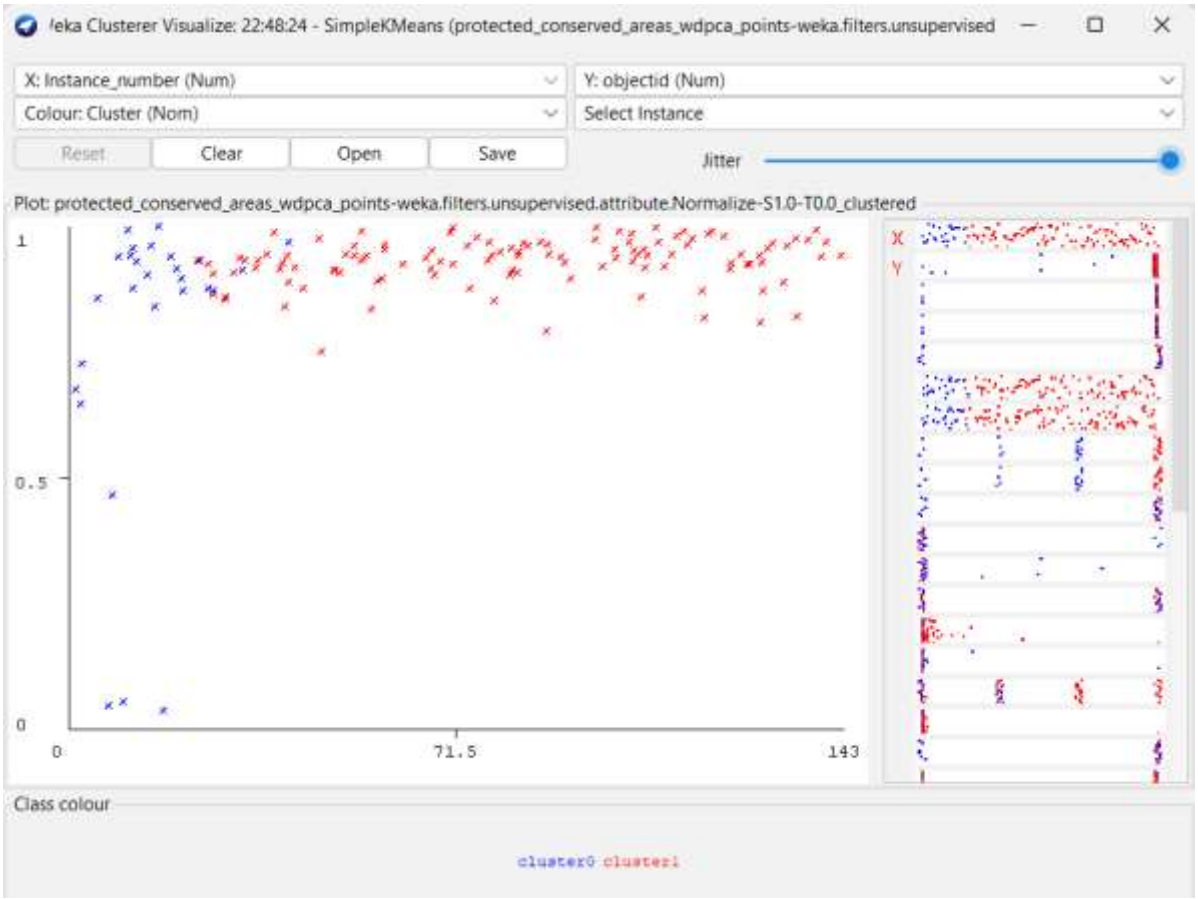

Fig. 7. SimpleKMeans Cluster Visualization in WEKA

Figure 8 shows that SimpleKMeans separates the records into one dominant cluster and one smaller secondary cluster, producing a simpler visual pattern than EM but still revealing a meaningful operational split. The dominant cluster contains most of the records and reflects the stronger Locally managed Marine Protected Area pattern, while the smaller cluster corresponds more closely to a Local Conservation Area-like profile with different ownership reporting, reported area, and no-take emphasis. Because the algorithm converged quickly and produced a 27 versus 117 record division, the figure supports the view that SimpleKMeans is useful for concise policy-oriented grouping, although it does not preserve the finer three-cluster structure detected by EM.

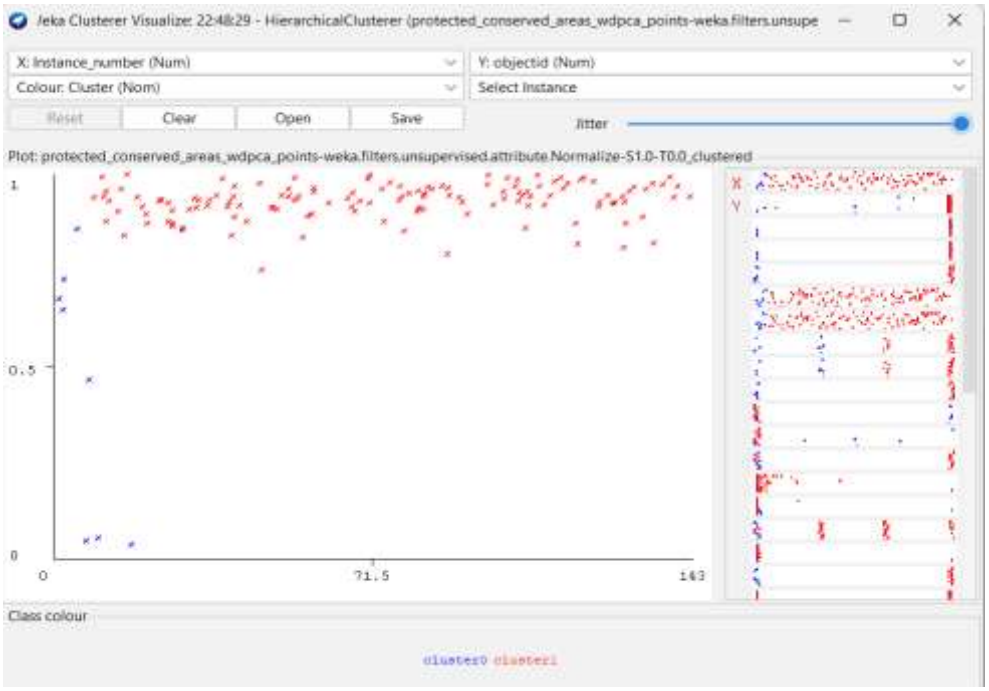

Fig. 8. HierarchicalClusterer

Figure 9 shows a strongly imbalanced split in which a very small blue subset is separated from a large red majority, indicating that hierarchical clustering is especially effective for isolating the most distinct records in the dataset. This visual separation is consistent with the 8 versus 136 clustered-instance result and suggests that the minority records are much closer to one another than to the rest of the dataset. In practical terms, the figure supports the interpretation that HierarchicalClusterer is well suited for identifying atypical or special-status entries that may require separate conservation review, even if it provides less internal detail than the three-cluster EM solution.

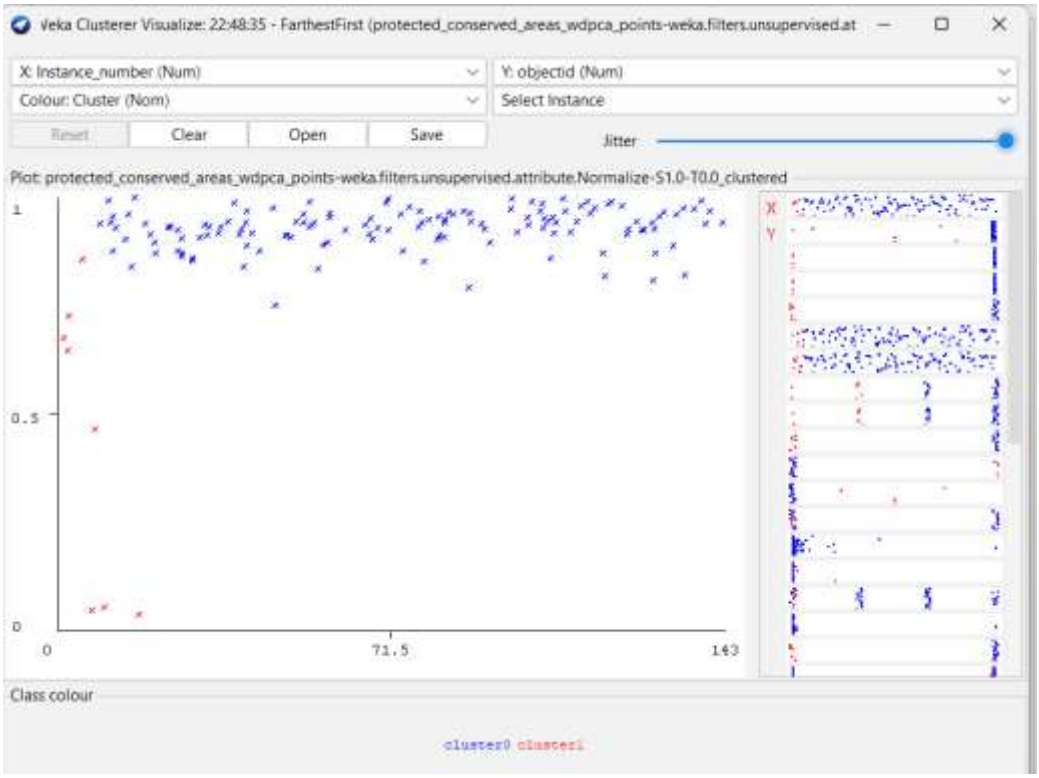

Fig. 9. FarthestFirst Cluster

Figure 10 confirms the same extreme-separation pattern identified by HierarchicalClusterer, with one very small red subset clearly isolated from a much larger majority group. This visual result aligns with the 136 versus 8 split and with the centroid descriptions showing contrast between an OECM, locally managed, established profile and a protected-area profile associated with international Ramsar-type designation, designated status, and management-plan implementation. Because the same minority subset reappears under a different clustering principle, the figure strengthens the conclusion that this group is structurally distinct and not merely an artifact of one algorithm.

Interpretation of Results in the Context of Objectives

The results of the study support its main objective, which is to analyze the spatial distribution and conservation status of protected and conserved areas using WEKA clustering techniques. Both the figures and the numerical outputs show that the dataset is not just one single group. Instead, it is made up of a large main cluster and a smaller distinct group. Among the four algorithms, EM gave the clearest and most detailed result because it formed three clusters. SimpleKMeans also showed a useful grouping by dividing the data into two clusters, while HierarchicalClusterer and FarthestFirst both identified a small special-status group that was clearly separated from the rest. The results also show that factors such as site type, designation, realm, reported area, no-take category, and status year helped explain why the records were grouped differently. Overall, when these results are matched with Table 4 and Table 5, they show that WEKA clustering is an effective tool for exploring patterns in protected-area data, and that EM is the most useful algorithm for this study.

9 Key Findings

1. A stable minority subset exists in the dataset. Three of the four algorithms either isolated or preserved a very small group of 8 records, indicating a structurally distinct cluster.
2. EM is the most informative clustering model. Because it selected three clusters automatically and produced an interpretable three-level structure, EM provides the richest analytical view of the dataset.
3. The majority of records share an OECM-oriented and locally managed marine protection profile. This dominant pattern appears in both EM and SimpleKMeans outputs.
4. The isolated minority subset carries stronger protected-area and international-designation signals, suggesting that conservation status and reporting profile differ substantially from the rest of the dataset.
5. Governance and management fields remain important but unevenly reported. The repeated presence of 'Not Reported' and 'Not Applicable' values indicates that reporting quality itself may shape the clustering pattern and deserves attention in future work.

10 Implications

The findings show that clustering can support protected-area database interpretation in at least three ways. First, it can separate records into analytically meaningful groups before more detailed policy review. Second, it can highlight special-status subsets that may require different management attention from the majority of locally managed or OECM-oriented records. Third, it can reveal where reporting completeness may be affecting interpretation, especially in governance, ownership, management authority, and management-plan fields.

For conservation planning, the practical implication is that the same monitoring or reporting framework should not automatically be applied in identical form to all records. A small internationally recognized

PA-dominant subset may need a different analytical lens from the broader body of OECM or locally managed marine records. Cluster-driven screening can therefore improve prioritization, database review, and comparative assessment.

11 Limitations

The study is exploratory and should be interpreted accordingly. The outputs are based on normalized Weka runs supplied by the user, not on a complete raw dataset with full coordinate fields and metadata validation. The absence of latitude-longitude variables prevented map-based spatial analysis. In addition, cluster quality was interpreted using model outputs and visual plots rather than external validation labels. Future research should combine the same clustering approach with the raw CSV file, coordinate-level analysis, and additional data-quality auditing.

12 Conclusion

This study answered the four research questions by showing that the protected/conserved area dataset contains a meaningful clustering structure rather than a single undifferentiated group. The results revealed a dominant majority profile associated mainly with OECM-like or locally managed marine protection records and a much smaller but highly distinct subset associated with formally designated protected areas and stronger international-recognition signals. The study further showed that spatial-distribution proxies and conservation-status attributes, especially realm, reported area, site type, designation, no-take category, and status year—help explain why the records separate into different groups. Among the four algorithms, EM provided the most useful and interpretable overall representation because it identified a three-cluster structure of 99, 37, and 8 records and preserved both the dominant pattern and the special-status minority subset. Finally, the clustering results indicate practical management and reporting implications: conservation records should not be reviewed as though they belong to a single uniform class, minority special-status groups may need separate planning attention, and governance, ownership, management-plan, and OECM-assessment fields need more complete reporting to strengthen future analysis. Overall, the study confirms that WEKA-based clustering is an effective exploratory method for analyzing protected/conserved area records when the objective is profile discovery, segmentation, and conservation interpretation.

13 Recommendations

1. Use EM as the primary exploration model when a nuanced profile of protected/conserved area records is needed.
2. Use SimpleKMeans when a concise operational segmentation is sufficient for management review or reporting dashboards.
3. Use HierarchicalClusterer or FarthestFirst to isolate unusual or special-status subsets that may need independent assessment.
4. Improve data completeness in governance, ownership, management-plan, and OECM assessment fields to reduce ambiguity in future clustering analyses.
5. Extend future studies by using the raw CSV file, adding coordinate-based spatial analysis, and validating clusters against ecological or management-effectiveness indicators.

14 References

1. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B, 39(1), 1–38.
2. Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online appendix for Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.
3. Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38, 293–306.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. SIGKDD Explorations, 11(1), 10–18.
5. Johnson, S. C. (2020). Hierarchical clustering schemes. Psychometrika, 32(3), 241–254.
6. MacQueen, J. (2021). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297).
7. UNEP-WCMC. (2026). Protected Area Profile for Philippines from the World Database on Protected Areas. Protected Planet.
8. Dinh, T., Nguyen, T., Nguyen, H., & Dao, T. (2025). Data clustering: A fundamental method in data science and management. Journal of Open Innovation: Technology, Market, and Complexity, 11, 100443.
9. Singh, J., Kumar, A., Khan, A. A., & Nandi, G. C. (2024). A comprehensive review of clustering techniques in data mining and machine learning. Advances in Engineering Software, 194, 103725.
10. UNEP-WCMC. (2026). Protected Areas (WDPA). Protected Planet.
11. UNEP-WCMC. (2024). Protected Planet Report 2024. Protected Planet.
12. Van der Plas, T. L., Alexander, D. G., & Pocock, M. J. O. (2025). Monitoring protected areas by integrating machine learning, remote sensing and citizen science. Ecological Solutions and Evidence, 6(2), e70040.
13. Urbano, F., Viterbi, R., Pedrotti, L., Vettorazzo, E., Movalli, C., & Corlatti, L. (2024). Enhancing biodiversity conservation and monitoring in protected areas through efficient data management. Environmental Monitoring and Assessment, 196(12).
14. Mouillot, D., Velez, L., Albouy, C., Casajus, N., Claudet, J., Delbar, V., Devillers, R., Letessier, T. B., Loiseau, N., Manel, S., Mannocci, L., et al. (2024). The socioeconomic and environmental niche of protected areas reveals global conservation gaps and opportunities. Nature Communications, 15, 9007.
15. Branco, V. V., Correia, L., & Cardoso, P. (2023). The use of machine learning in species threats and conservation analysis. Biological Conservation, 283, 110091.