Unsupervised Learning: Principal Component Analysis(PCA) and Clustering Analysis of Water-Related Indicators

1. Introduction

This study applies unsupervised learning to group countries based on key water-related indicators. K-means, hierarchical clustering, and DBSCAN were used, together with Principal Component Analysis (PCA) for dimensionality reduction. All features were averaged over 2012–2021 to emphasize recent trends and reduce short-term fluctuations. Multiple feature combinations and clustering parameters were evaluated, with the best configuration selected based on internal validity metrics.

2. R Libraries and Tools Used

Package	Purpose
tidyverse	Data wrangling and visualization
cluster	Computes silhouette scores for clustering evaluation
factoextra	PCA and cluster visualization
clusterSim	Calculates the Davies-Bouldin index for cluster evaluation
dbscan	Density-based clustering (DBSCAN) and outlier detection
rnaturalearth & rnaturalearthdata	Retrieves and manages natural Earth geographic data
sf	Spatial data handling for map visualization
stringr	String manipulation (used in feature set handling)
scales	Data rescaling for metric normalization

3. Clustering Methodology

3.1. Feature Selection & Normalization

The purpose of clustering was defined as grouping countries based solely on water-related characteristics, with the aim of uncovering underlying patterns in water usage, availability, and stress—independent of direct economic or demographic influences.
Accordingly, the following indicators were selected as potential clustering features:

- Available Water Resources
- Water Productivity
- Total Withdrawals
- Sectoral Withdrawals Proportions (Industry, Agriculture, Domestic)
- Precipitation
- Water Stress
- Natural Disasters

As will be discussed later, almost all possible subsets of these features were examined in the clustering process.

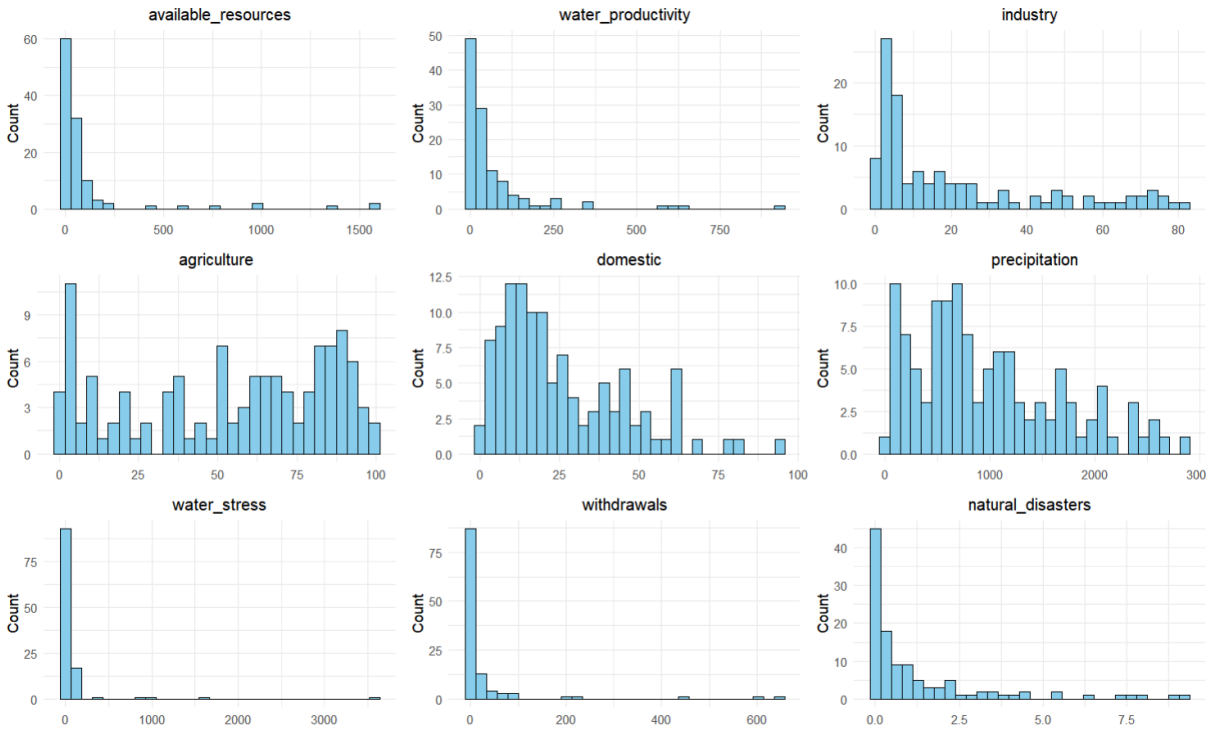


Figure - Distribution of Selected Features (2012–2021 Average)

The above figure presents the distribution of the selected water-related indicators based on their 2012–2021 averages.
As visualized in the feature-wise histograms, most variables exhibit non-normal, right-skewed, and highly heterogeneous distributions. Several features, including available water resources, withdrawals, and natural disasters, show extreme outliers and heavy tails, while others like agriculture and precipitation suggest multi-modal patterns.
These characteristics point to a dataset that may challenge traditional clustering methods relying on assumptions of convexity, uniform density, or balanced variance across clusters. Accordingly, clustering algorithms capable of handling irregular shapes, density variation, and outlier sensitivity might be more appropriate in this context.

```
> for (features in feature_sets) {
+   scaled_data <- scale(aggregated_data[, features, drop = FALSE])
```

Code Snippet – Normalization of Clustering Features

The selected clustering features were normalized (scaled) to ensure comparability across variables and to prevent any single feature from dominating the clustering results due to differences in scale.

3.2. Model Selection

Three clustering algorithms were selected for this study: K-means, hierarchical clustering, and DBSCAN. These models were chosen based on their widespread use, complementary strengths, and alignment with the interpretability goals of the analysis.
Although the data distributions presented earlier were skewed, non-spherical, and heterogeneous—conditions that are not ideal for K-means—this method was included as a baseline. Its simplicity, efficiency, and ease of implementation make it a valuable reference point. Moreover, K-means remains a widely adopted standard in exploratory clustering, offering a quick overview of global grouping tendencies.
Hierarchical clustering was selected to address structural limitations observed during the K-means exploration. In some clusters, a majority of countries came from two geographically distinct regions—for example, Europe and Africa—suggesting internal regional heterogeneity. Further subdivision could reveal more meaningful groupings. Hierarchical clustering supports this by capturing multi-level relationships and enabling flexible interpretation through dendrograms, without requiring a predefined number of clusters.
DBSCAN was included for its robustness in handling irregular shapes and outliers—both clearly present in the data. Earlier feature distributions showed skewed, multi-modal, and non-spherical patterns, making conventional centroid-based methods less effective. Unlike K-means and hierarchical clustering, DBSCAN relies on local density rather than global distance metrics or a fixed number of clusters. This allows it to identify arbitrarily shaped groupings and exclude noise points, making it well-suited for the heterogeneous distributions found in cross-national water-related datasets.
Together, the selected models offer a well-rounded approach: K-means provides a baseline, hierarchical clustering reveals nested structures, and DBSCAN captures non-linear patterns and outliers—all aligned with the study's goal of uncovering meaningful country groupings based on water-related characteristics.
Although Gaussian Mixture Models (GMM) might be considered appropriate for this dataset—given their ability to handle non-spherical clusters and overlapping group structures—they were excluded from the model selection. While GMM may offer advantages in identifying transitional or mixed-pattern countries, its soft clustering approach, which assigns probabilistic rather than fixed labels, was not aligned with the study's goal of producing clear and interpretable groupings for policy use. The decision was based not on performance concerns, but on the reduced practicality of applying probabilistic assignments in a real-world policy context. Hard clustering methods such as K-means, hierarchical clustering, and DBSCAN were prioritized for their clearer interpretability.

3.3. Principal Component Analysis

```
> pca_all <- prcomp(scale(aggregated_data), scale. = FALSE)
```

```
> pca_summary <- summary(pca_all)
> print(pca_summary)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6    PC7
PC8     PC9
Standard deviation     1.6557 1.3350 1.1158 1.1082 0.87717 0.79418 0.6163
0.47229 0.01969
Proportion of Variance 0.3046 0.1980 0.1383 0.1364 0.08549 0.07008 0.0422
0.02478 0.00004
Cumulative Proportion  0.3046 0.5026 0.6410 0.7774 0.86289 0.93297 0.9752
0.99996 1.00000
```

Code Snippet – Dimensionality Reduction: PCA Variance Analysis

Principal Component Analysis (PCA) was conducted to assess the variance contributions of each principal component and, based on this information, to determine an appropriate number of components for dimensionality reduction prior to clustering. The analysis showed that the first four components captured approximately 77.7% of the total variance, and the first six explained over 93%, indicating that a reduced representation using the first 4 to 6 components retains most of the information in the original feature space.

3.4. Grid Search

Clustering tasks inherently lack a clear ground truth, making it difficult to assert with certainty that any one model or configuration is universally optimal. Therefore, this study exhaustively evaluated all possible combinations of feature subsets, clustering algorithms (K-means, hierarchical, DBSCAN), algorithm-specific parameters, and the number of principal components used for dimensionality reduction. This was implemented through fully nested loops, ensuring that no plausible configuration was left untested within the defined search space.

```
> feature_sets <- list()
> combo_id <- 1
> for (k in 2:length(raw_feature_sources)) {
+  combos <- combn(raw_feature_sources, k, simplify = FALSE)
+  for (combo in combos) {
+    feature_sets[[paste0("Set_", combo_id)]] <- combo
+    combo_id <- combo_id + 1
+  }
+ }
```

Code Snippet – Feature Subset Generation for Grid Search

As part of the grid search, all possible combinations of water-related features from size 2 to the full set, were systematically generated. One-variable sets were excluded—not due to performance concerns, but to ensure interpretability. Clusters based on a single feature would lack dimensional context and likely be monotonous, making it difficult to derive meaningful conclusions for analysis or policy interpretation.

```
> num_pc_options <- 2:8
> epsilon_values <- seq(0.0, 2.0, by = 0.2)
> minpts_values <- 2:8
> cluster_counts <- 2:4
> results <- data.frame()
```

Code Snippet – Parameter Ranges for Grid Search

Based on the previous PCA results, which showed that the first 4 to 6 components explained approximately 78% to 93% of the total variance, principal components from 2 to 8 were tested to allow some margin around this core range.
For DBSCAN, initial experiments tested a wide parameter range — eps values from 0 to 4, and minPts values from 2 to 12 — to account for potential variability in cluster density across countries. However, well-performing configurations (particularly those with silhouette scores above 0.5) were consistently concentrated within the narrower range of eps = 0 to 2 and minPts = 2 to 8. Based on these observations, the parameter space was refined in the final grid search to enhance computational efficiency and reproducibility.
The number of clusters for K-means and hierarchical clustering was limited to 2–4, since one cluster offers no useful grouping, and having too many clusters could make interpretation difficult.

4. Results and Interpretation

4.1. Clustering Configuration Selection

	Method	Feature_Set	PCs	Param	Num_Clusters	Silhouette	DB_Index	Labels
579	Hierarchical	available_resources, withdrawals	2	2	2	0.9154100	0.28347158	1, 1, 1,....
2803	DBSCAN	water_stress, withdrawals	2	eps=1.8, minPts=2	2	0.9112008	0.24136186	1, 1, 0,....
2804	DBSCAN	water_stress, withdrawals	2	eps=1.8, minPts=3	2	0.9112008	0.24136186	1, 1, 0,....
2818	KMeans	water_stress, withdrawals	2	3	3	0.9112008	0.42321758	2, 2, 1,....
2821	Hierarchical	water_stress, withdrawals	2	3	3	0.9065226	0.20548475	1, 1, 1,....
2810	DBSCAN	water_stress, withdrawals	2	eps=2, minPts=2	2	0.9065226	0.26617159	1, 1, 1,....
2811	DBSCAN	water_stress, withdrawals	2	eps=2, minPts=3	2	0.9065226	0.26617159	1, 1, 1,....
576	KMeans	available_resources, withdrawals	2	2	2	0.9058169	0.53377466	2, 2, 2,....
2782	DBSCAN	water_stress, withdrawals	2	eps=1.2, minPts=2	2	0.9014863	0.11988371	1, 1, 0,....
2789	DBSCAN	water_stress, withdrawals	2	eps=1.4, minPts=2	2	0.9014863	0.11988371	1, 1, 0,....
2796	DBSCAN	water_stress, withdrawals	2	eps=1.6, minPts=2	2	0.9014863	0.11988371	1, 1, 0,....
2817	KMeans	water_stress, withdrawals	2	2	2	0.8971020	0.34276329	2, 2, 2,....
2820	Hierarchical	water_stress, withdrawals	2	2	2	0.8971020	0.34276329	1, 1, 1,....
562	DBSCAN	available_resources, withdrawals	2	eps=1.8, minPts=2	2	0.8904171	0.25632549	1, 1, 1,....
563	DBSCAN	available_resources, withdrawals	2	eps=1.8, minPts=3	2	0.8904171	0.25632549	1, 1, 1,....

Showing 1 to 15 of 148,736 entries, 8 total columns

Figure – Top Clustering Results by Silhouette Score

This figure shows a subset of the 148,736 clustering configurations, sorted in descending order of silhouette score. Although several top results show silhouette scores above 0.9 — suggesting high internal cohesion — further inspection revealed significant cluster size imbalance. In some cases, for example, the model produced two clusters where one cluster contained only two countries, while the remaining countries were all grouped into the other

cluster. Such results, despite their high internal metrics, do not yield meaningful or interpretable insights.
This also suggests a limitation of widely used clustering evaluation metrics, silhouette score and DB index, which do not account for cluster size balance.
To address this, a new cluster balance metric called MaxMinRatio was introduced, which measures the ratio between the largest and smallest cluster sizes (excluding noise points). Configurations with a MaxMinRatio above 10 were considered unbalanced and were filtered out, ensuring that only well-structured and interpretable clustering results were retained for further evaluation.
In addition, configurations with more than 20% missing or unassigned labels were excluded to ensure sufficient coverage of countries in the analysis.

```
> # Normalize Silhouette and DBI, then compute composite score
> results_clean$Silhouette_Norm <- rescale(results_clean$Silhouette, to = c(0, 1))
> results_clean$DBI_Norm_Inv <- 1 - rescale(results_clean$DB_Index, to = c(0, 1))
> results_clean$CompositeScore <- 0.5 * results_clean$Silhouette_Norm + 0.5 * results_clean$DBI_Norm_Inv
```

Code Snippet – Composite Clustering Score Based on Silhouette and DB Index

To build a reliable ranking of clustering configurations, a composite score combining the silhouette score and the Davies–Bouldin index (DBI) was constructed. This was necessary because a high silhouette score does not always correspond to a low (desirable) DBI, and vice versa. Also, these two metrics were selected for their complementary focus: silhouette captures separation and cohesion, while DBI emphasizes intra-cluster compactness relative to inter-cluster separation.
Both metrics were normalized to a 0–1 scale, with DBI inverted via (1 − normalized DBI) to ensure that higher values consistently indicated better clustering performance. Then, an equal weighting of 0.5 was applied to balance both perspectives without introducing subjective bias.
While alternative scoring schemes are possible, this design prioritizes interpretability and neutrality, which are critical in unsupervised model selection.

	Method	Feature_Set	PCs	Param	Num_Clusters	Silhouette	DB_Index	Labels	MaxMinRatio	MissingRate	Silhouette_Norm	DBI_Norm_Inv	CompositeScore
1	DBSCAN	available_resources, industry	2	eps=0.4, minPts=8	2	0.6664886	0.3776357	1, 1, 1,...	6.133333	0.06956522	0.8201664	0.9284760	0.8743212
2	DBSCAN	industry, water_stress	2	eps=0.2, minPts=7	2	0.6074117	0.2164016	1, 1, 0,...	7.727273	0.16521739	0.7774863	0.9614308	0.8694585
3	DBSCAN	industry, withdrawals	2	eps=0.4, minPts=8	2	0.6333523	0.3739639	1, 1, 1,...	6.714286	0.06086957	0.7962271	0.9292265	0.8627268
4	DBSCAN	available_resources, industry, withdrawals	2	eps=0.4, minPts=8	2	0.6268690	0.3836774	1, 1, 1,...	6.000000	0.08695652	0.7915432	0.9272411	0.8593922
5	DBSCAN	water_productivity, industry	2	eps=0.4, minPts=3	2	0.6150827	0.3684267	0, 1, 1,...	6.769231	0.12173913	0.7830282	0.9303582	0.8566932
6	DBSCAN	water_productivity, industry	2	eps=0.4, minPts=4	2	0.6020309	0.3606021	0, 1, 1,...	6.615385	0.13913043	0.7735989	0.9319575	0.8527782
7	DBSCAN	water_productivity, industry	2	eps=0.4, minPts=5	2	0.6020309	0.3606021	0, 1, 1,...	6.615385	0.13913043	0.7735989	0.9319575	0.8527782
8	DBSCAN	available_resources, industry, withdrawals	3	eps=0.4, minPts=8	2	0.6071918	0.3901438	1, 1, 1,...	5.933333	0.09565217	0.7773274	0.9259195	0.8516234
9	DBSCAN	water_productivity, industry	2	eps=0.4, minPts=6	2	0.5930882	0.3582433	0, 1, 0,...	6.538462	0.14782609	0.7671383	0.9324396	0.8497890
10	DBSCAN	water_productivity, industry	2	eps=0.4, minPts=7	2	0.5930882	0.3582433	0, 1, 0,...	6.538462	0.14782609	0.7671383	0.9324396	0.8497890
11	DBSCAN	available_resources, industry, water_stress	3	eps=0.4, minPts=8	2	0.5852545	0.3762419	1, 1, 0,...	5.733333	0.12173913	0.7614788	0.9287609	0.8451198
12	DBSCAN	water_productivity, industry, withdrawals	2	eps=0.4, minPts=7	2	0.5596253	0.3114408	0, 1, 1,...	5.200000	0.19130435	0.7429629	0.9420056	0.8424843
13	DBSCAN	water_productivity, industry, withdrawals	2	eps=0.4, minPts=8	2	0.5596253	0.3114408	0, 1, 1,...	5.200000	0.19130435	0.7429629	0.9420056	0.8424843
14	DBSCAN	water_productivity, industry, withdrawals	2	eps=0.4, minPts=4	2	0.5688889	0.3564821	0, 1, 1,...	5.533333	0.14782609	0.7496554	0.9327996	0.8412275
15	DBSCAN	industry, agriculture	2	eps=0.4, minPts=7	2	0.5516598	0.3002723	0, 1, 1,...	6.307692	0.17391304	0.7372082	0.9442884	0.8407483

Showing 1 to 15 of 4,547 entries, 13 total columns

Figure - Top Clustering Results After Applying All Filters, by Composite Score

A total of 4,547 configurations remained after filtering from the original 148,736 results, applying constraints on cluster balance (MaxMinRatio < 10) and missing label rate (≤ 20%).

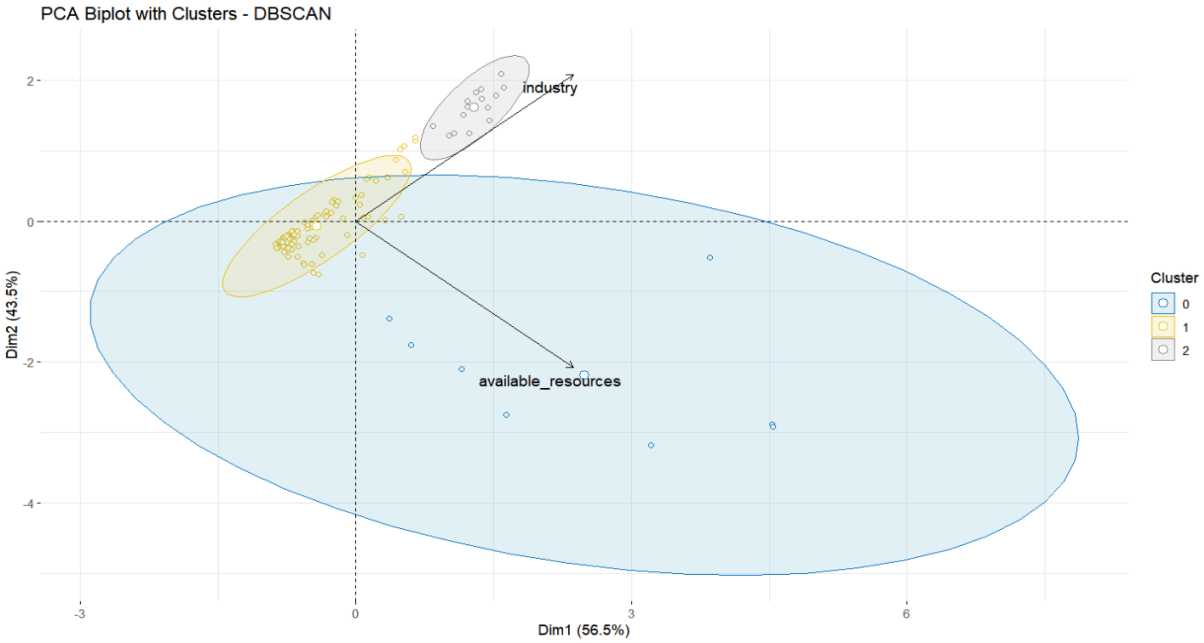
Notably, all top-ranked configurations based on the composite score were produced by DBSCAN. This outcome may be partly explained by its ability to detect arbitrarily shaped clusters and handle outliers — features that align well with the structure of the data.
As shown in the earlier section on Feature Selection & Normalization, country-level water indicators exhibit highly uneven distributions, non-convex groupings, and frequent regional outliers, reflecting the diversity of geographic, climatic, and economic conditions. These characteristics tend to favor density-based methods over centroid-based ones.
However, the dominance of DBSCAN is also shaped by evaluation constraints. For example, when the thresholds on MaxMinRatio and missing label rate were tightened to 5, the top configurations shifted toward K-means and hierarchical clustering. Therefore, while DBSCAN performed best under the applied conditions, its selection reflects both empirical results and structural compatibility with the data.
The highest-ranked configuration among the filtered results was selected for in-depth analysis and interpretation in the subsequent sections.

4.2. Evaluation and Visualization

```
> cat("Best Clustering Configuration:\n")
Best Clustering Configuration:
> cat("Silhouette Score:", round(best_result$Silhouette, 4), "\n")
Silhouette Score: 0.6665
> cat("DB Index:", round(best_result$DB_Index, 4), "\n")
DB Index: 0.3776
> cat("MaxMinRatio:", round(best_result$MaxMinRatio, 2), "\n")
MaxMinRatio: 6.13
> cat("Missing Label Rate:", round(best_result$MissingRate * 100, 2),
"%\n")
Missing Label Rate: 6.96 %
```

Code Snippet – Summary of the Best Clustering Configuration

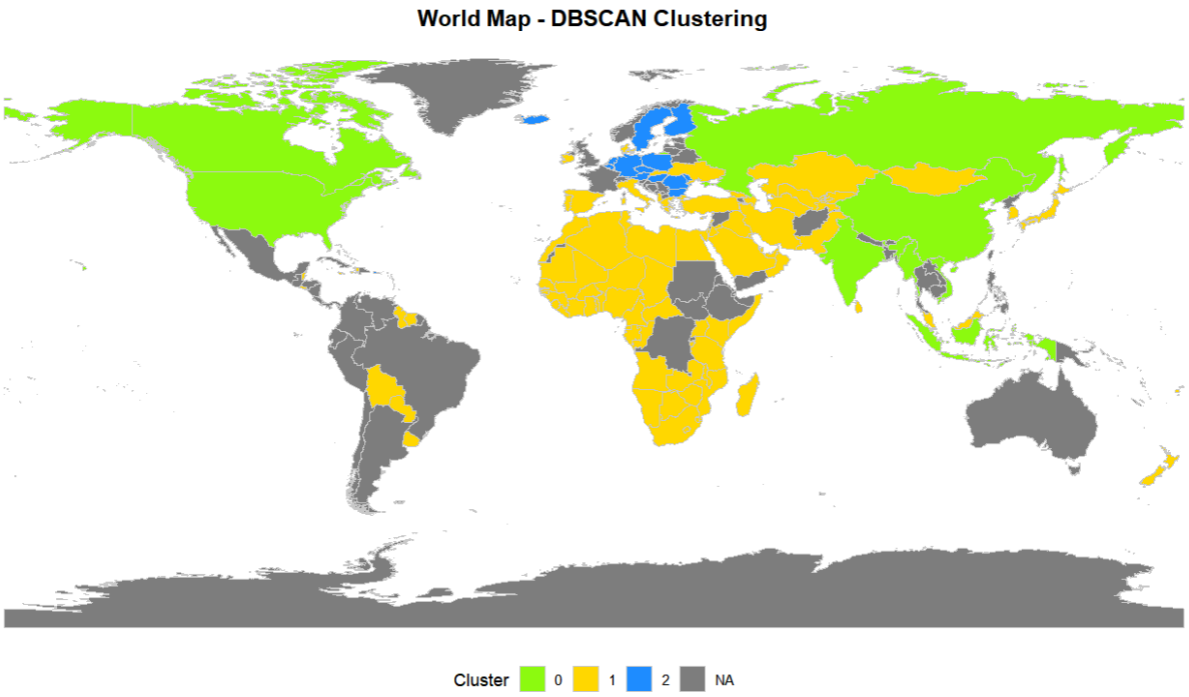
The best clustering configuration achieved a silhouette score of 0.6665, which indicates a moderate level of cohesion within clusters and separation between them. The DB index was 0.3776, a relatively low value that generally reflects compact and well-separated clusters. The MaxMinRatio was 6.13, meaning the largest cluster was about six times the size of the smallest; while this is within the threshold of 10, it suggests some level of imbalance. The missing label rate was 6.96%, indicating that a small portion of countries were not assigned to any cluster. Although all metrics fall within acceptable ranges, the imbalance in cluster sizes and the presence of unassigned data points may be worth considering when interpreting the results.



The figure shows the result of the best clustering configuration, projected onto the first two principal components (PC1 and PC2), which together explain 100% of the variance, as the analysis was conducted on two features: available water resources and industry. Each point represents a country, colored by its assigned cluster label.

- Cluster 1 (yellow) appears tightly grouped, indicating relatively strong internal cohesion among countries in this group.
- Cluster 2 (gray) forms a smaller but distinct group, positioned in the direction of higher values in the industry dimension.
- Cluster 0 (blue) consists of points labeled as noise by DBSCAN — countries that were not assigned to any cluster under the chosen parameter settings.

The plot provides a visual summary of how the two selected features contribute to differentiating countries into distinct groups under the chosen clustering method.
The final model, based on only available water resources and industry withdrawals, was not manually selected for interpretability, but rather emerged as the top-performing configuration after applying filtering criteria on cluster balance and label completeness.
Notably, this low-dimensional model also facilitated intuitive interpretation and effective 2D visualization, which was not consistently achievable with higher-dimensional alternatives. While more complex models were considered during the grid search, they often led to fragmented or less coherent cluster structures. In this sense, the final configuration represents a fortunate convergence of statistical performance and semantic clarity.
Additionally, although DBSCAN labeled some countries as noise (Cluster 0), their geopolitical and economic relevance warranted cautious interpretation. These observations were not treated as tightly grouped clusters but were still examined due to their potential policy implications.



In addition to the PCA projection, a geographic map of cluster assignments provides further insight into regional patterns.

- Cluster 1 (yellow) is concentrated in Sub-Saharan Africa, the Middle East, Central Asia, and parts of Latin America, possibly reflecting countries with similar agricultural reliance, infrastructure limitations, or climatic stress factors.

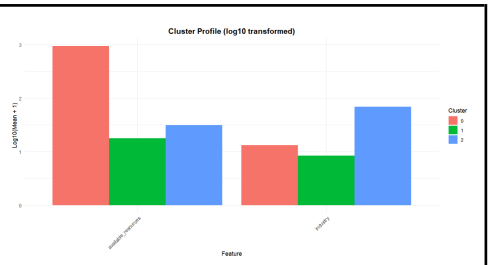
- Cluster 2 (blue) is mostly composed of Northern and Central European countries, forming a compact and contiguous region. This grouping could be associated with relatively high industrial development or consistent water governance practices.

- Cluster 0 (green) consists of countries labeled as noise by DBSCAN, including the United States, Canada, Russia, China, India, and several others across Southeast Asia and Eastern Europe. Although these countries were not assigned to any cluster under the current DBSCAN configuration, they will be individually discussed in the following section based on their feature characteristics.

- Countries shown in gray were previously excluded from clustering due to missing data. The missingness is notably concentrated in parts of Africa, Latin America, and Oceania, suggesting potential gaps in global data reporting or coverage.

While no assumptions about the exact drivers can be made from geography alone, this spatial distribution suggests that the clustering reflects more than random separation — and warrants further exploration through feature-based profiling in the following section.

4.3. Cluster Profiles

`> View(cluster_summary)`
`> cluster_summary`
`# A tibble: 3 × 3`
`  Cluster available_resources industry`
`  <fct>  <dbl>  <dbl>`
`1 0  1030.  25.7`
`2 1  38.9  12.4`
`3 2  46.2  68.9`
	

Average feature values by cluster(Table)	Cluster Profiles (Log-Scaled, Barplot)

The table and bar plot above shows the average values of the two selected features — Available Water Resources and Industry Withdrawals — for each cluster. These profiles provide insight into the distinct characteristics that define each group:

Cluster	Characteristics
Cluster1	Characterized by low available water resources (38.9) and low industrial use (12.4), this cluster likely includes countries with limited water supply and relatively low levels of industrialization. Many of these countries are located in Sub-Saharan Africa, the Middle East, and parts of Central and South Asia, suggesting environments where water scarcity and underdeveloped infrastructure may play important roles in shaping usage patterns.
Cluster2	This group is defined by moderate available resources (46.2) and the highest industrial use (68.9) among all clusters. The profile suggests countries with limited natural water availability but strong industrial activity, potentially supported by efficient management or heavy reliance on imports and infrastructure. These countries are mainly located in Europe, possibly reflecting developed economies with structured industrial sectors and controlled water use.
Cluster0	This group consists of countries originally labeled as noise by DBSCAN, including several large and globally influential states. Although not tightly clustered, they display relatively high water availability (1030.0) and moderate industrial use (25.7), suggesting a loosely shared profile. Given their economic and geopolitical weight, they are included here for reference despite their exclusion from the formal clustering structure.

This interpretation provides contextual understanding of the clustering outcome and serves as a foundation for potential policy recommendations tailored to the water resource profiles of each cluster.
However, it is important to acknowledge that while clustering reveals consistent patterns, it does not explain the underlying causes. As an unsupervised method, it identifies what exists, not why it exists. Accordingly, interpreting clusters as policy groups assumes internal homogeneity that may not fully hold.
Policy suggestions derived from cluster membership should therefore be treated cautiously, especially in the absence of causal inference and control for unobserved factors.

5. Discussion on Policy Implications

Drawing from the broad patterns observed across clusters, several exploratory policy considerations can be outlined.
While these groupings provide a useful lens for comparative analysis, any policy application must be context-sensitive and account for country-specific conditions beyond the scope of this clustering exercise.

[Cluster 1: Low Resource, Low Industrial Demand]

This group consists of countries with generally limited water availability and relatively low industrial withdrawals. While specific national contexts vary, the combination suggests potential vulnerabilities in both water access and economic capacity related to industrial use.

Tentative policy directions may include:

- Improving access to basic water infrastructure.
- Strengthening local water resilience planning.
- Aligning future industrial development with available water resources.

[Cluster 2: High Industrial Use with Limited Resources]

Countries in this group exhibit relatively high industrial water demand despite having only moderate to limited water availability. This configuration may reflect effective infrastructure and management, but also indicates a potential imbalance between usage and long-term sustainability.

Policy considerations may include:

- Enhancing water use monitoring in industrial areas.
- Supporting circular water practices such as recycling and reuse.
- Reviewing allocation frameworks to balance industrial needs with ecological and social priorities.

[Cluster 0: Resource-Rich with Moderate Industrial Activity (Unclustered Group)]

This group includes countries that were not formally clustered but share a general pattern of abundant water resources and moderate industrial withdrawals. A limited interpretation is offered here as a reference.
Possible considerations — highly dependent on country-specific contexts — include:

- Preserving resource security in anticipation of future industrial or demographic shifts.
- Promoting efficient technologies to avoid complacency in water-abundant systems.
- Strengthening monitoring in resource-intensive sectors to safeguard long-term sustainability.

Overall, this clustering framework highlights how data-driven groupings can guide region-specific policy interventions, ensuring that water governance aligns with actual usage patterns and resource constraints. However, these insights are exploratory rather than prescriptive and should be tailored to national-level contexts, given the absence of causal inference and the influence of unobserved political, economic, and geographic factors.

6. Conclusion

This study applied unsupervised learning techniques to identify meaningful clusters among countries based on water-related indicators, focusing on available water resources and industrial withdrawals. Principal Component Analysis (PCA) was used for dimensionality reduction, and an extensive grid search explored various clustering algorithms, feature subsets, and parameter combinations.

DBSCAN yielded the highest-performing configuration under the defined evaluation criteria, which combined silhouette score and Davies–Bouldin index into a composite metric. Additional constraints on cluster balance and label coverage were applied during post-processing to ensure interpretability and robustness. However, this outcome was shaped by the chosen thresholds and scoring design, and alternative methods such as K-means or hierarchical clustering may perform better under different assumptions.

The final model revealed three distinct groups, each with its own water-use profile. While one of these groups consisted of countries initially labeled as noise by DBSCAN, it was retained for reference due to its geopolitical and economic importance.

Although the study explored potential policy implications associated with the observed patterns, such recommendations remain exploratory in nature. Clustering reveals what patterns exist, not why they occur. Therefore, any use of these groupings for policy design must be approached with caution, tailored to national-level contexts, and supplemented by more detailed causal and institutional analysis.

Overall, the study demonstrates how a rigorous unsupervised learning framework — when properly validated and interpreted — can uncover latent structure in complex global datasets and inform further inquiry into context-specific water management strategies.