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 Packages 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.

```
> 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 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.

3.3. 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.

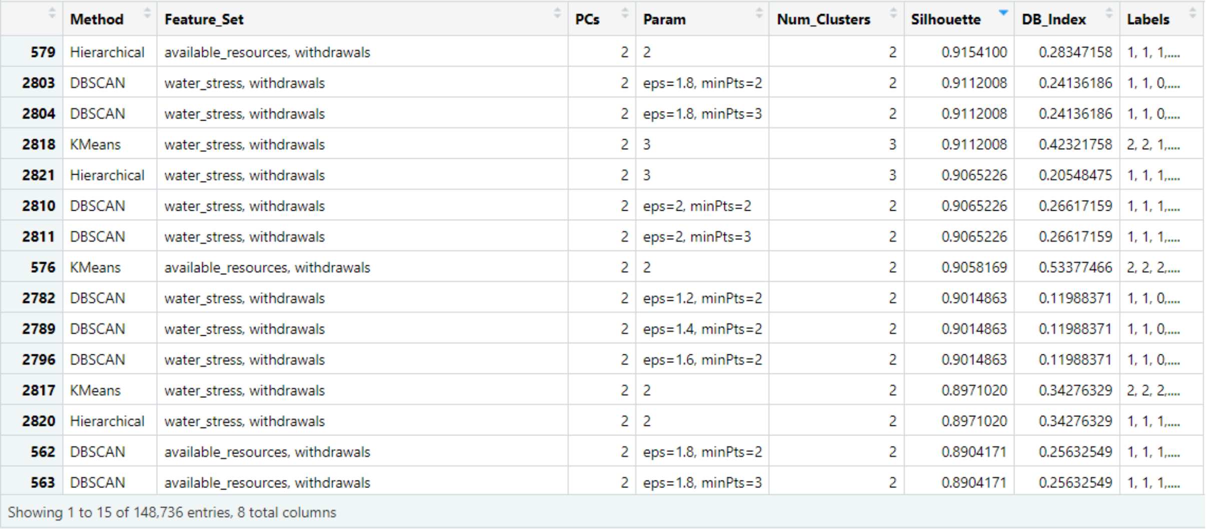
```
> num_pc_options <- 2:8
> epsilon_values <- seq(0.0, 2.0, by = 0.2)
> minpts_values <- 2:8
> cluster_counts <- 2:4
> results <- data.frame()
```

Code Snippet – Parameter Ranges for PCA, DBSCAN, and K-Means 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, eps values ranged from 0.0 to 2.0, and minPts from 2 to 8. These ranges were set by giving uniform margins around the default values. The appropriateness of these values was assessed based on the results, as will be discussed later.
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 can 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.

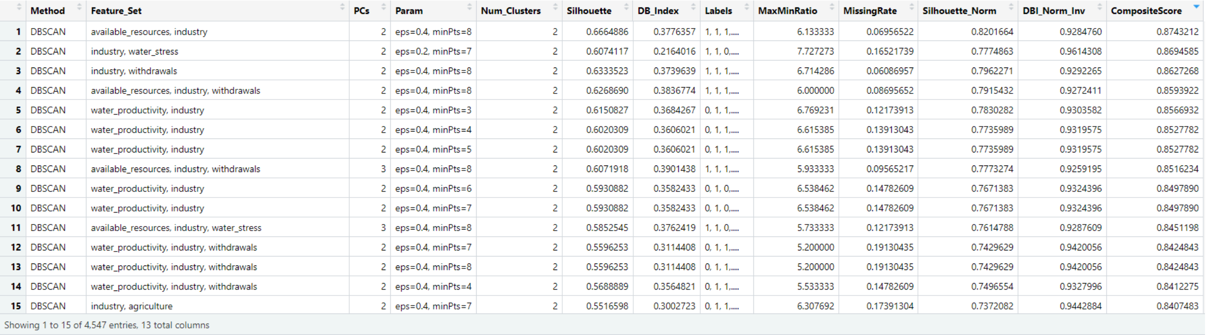
To address this, a new cluster balance metric called MaxMinRatio was introduced, which measures the ratio between the largest and smallest cluster sizes (excluding noise points). Configurations with a MaxMinRatio above 10 were considered unbalanced and were filtered out, ensuring that only well-structured 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 silhouette score and DB index was constructed. This was necessary because a high silhouette score does not always correspond to a low (and better) DB index, and vice versa. Both metrics were first normalized to a 0–1 scale. The DB index was then inverted (1 − normalized DBI) so that higher values consistently indicated better clustering quality. Finally, the two normalized scores were averaged to produce a single composite score used for ranking.



	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%). Interestingly, all top-ranked configurations based on the composite score were produced using the DBSCAN algorithm. This may be partially explained by DBSCAN's ability to handle outliers and its flexibility in identifying clusters of arbitrary shapes, which can be advantageous given the structure of the data. However, this outcome is also influenced by the filtering criteria applied, particularly the thresholds on missing label rate and cluster size balance (MaxMinRatio). For instance, when both thresholds are tightened to 5, most of the top-performing configurations shift toward K-means and hierarchical clustering. This indicates that the dominance of DBSCAN in the current results is partly a function of the evaluation criteria rather than an inherent superiority of the algorithm. As a result, while DBSCAN performed well under the given conditions, it cannot be affirmatively considered the best method.
The highest-ranked configuration among the filtered results was selected for in-depth analysis and interpretation in the subsequent sections.

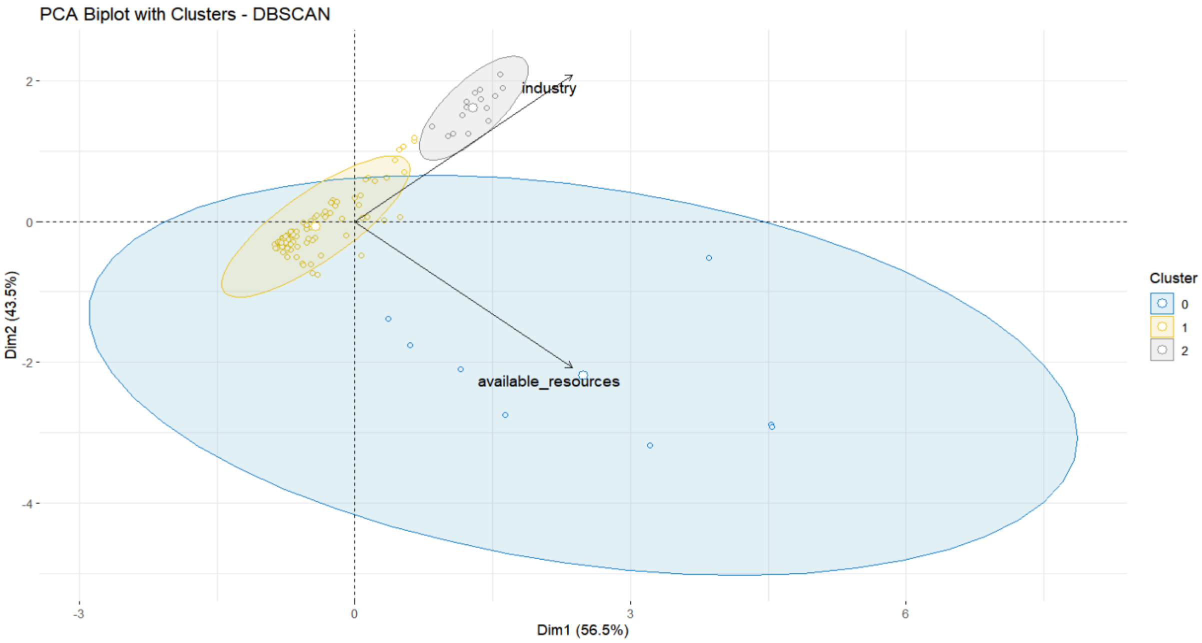
4.2. Evaluation and Visualization

```
> 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.

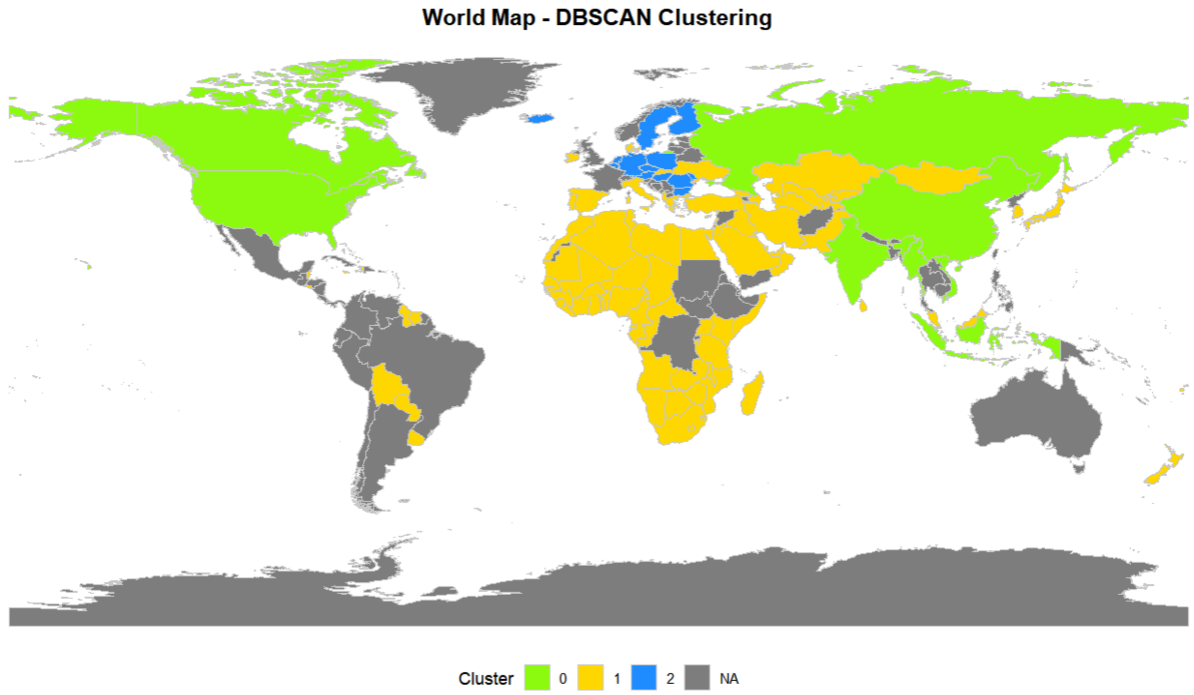
The MaxMinRatio was 6.13, meaning the largest cluster was about six times the size of the smallest; while this is within the threshold of 10, it suggests some level of imbalance. The missing label rate was 6.96%, indicating that a small portion of countries were not assigned to any cluster. Although 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 DBSCAN clustering, 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.



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.

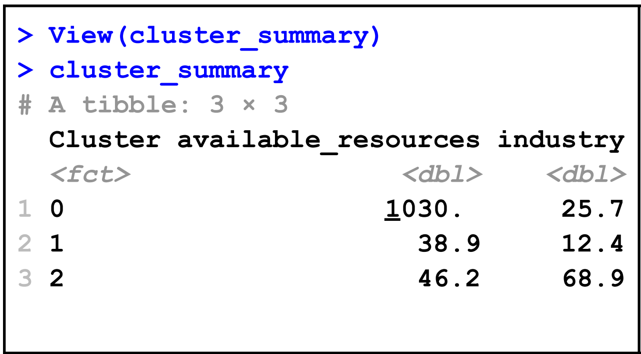
- Cluster 2 (blue) is mostly composed of Northern and Central European countries, forming a compact and contiguous region. This grouping could be associated with relatively high industrial development or consistent water governance practices.

- Cluster 0 (green) includes a wide range of countries such as the United States, Canada, Russia, China, India, and several in Southeast Asia and Eastern Europe. While this group was labeled as noise by DBSCAN — meaning these countries did not belong to any dense cluster under the current parameters — their spatial pattern and global importance justify further interpretation. They may share common traits in terms of scale, resource abundance, or water management diversity.

- 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.

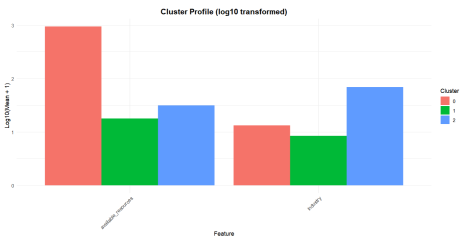
While no assumptions about the exact drivers can be made from geography alone, this spatial distribution suggests that the clustering reflects more than random separation — and warrants further exploration through feature-based profiling in the following section.

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 cluster shows very high available water resources (1030.0) and moderate industrial use (25.7). It includes several large and globally influential countries. While the group was originally labeled as noise by DBSCAN, the internal average suggests that these countries may share a broad profile of resource abundance and moderate industrial activity, though they were not tightly grouped in the clustering structure. Their inclusion here still warrants attention due to their regional and economic significance.

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.

4. Discussion on Policy Implications

Based on the distinct profiles and regional distribution of the clusters, several policy-relevant insights emerge. While specific strategies must be tailored to each country's context, the following implications offer a general direction for water resource management within each cluster.

[Cluster 1: Low Resource, Low Industrial Demand]

This group is characterized by low available water resources and low industrial water use. These countries may face constraints due to limited natural water availability and underdeveloped industrial infrastructure. Policy directions may include:

- Enhancing water access and basic infrastructure to ensure reliability under resource scarcity.
- Supporting climate resilience strategies in water-stressed areas.
- Planning industrial development alongside sustainable water supply systems to avoid future bottlenecks.

This cluster may benefit from targeted investment to improve both water security and industrial capacity in tandem.

[Cluster 2: Limited Resources with High Industrial Activity]

Cluster 2 includes countries with moderate to low water availability but the highest level of industrial water use among all clusters. This imbalance may indicate pressure on existing water systems or reliance on strong infrastructure and management to meet industrial demand. Relevant policy measures include:

- Strengthening water monitoring in industrial zones to ensure efficient use and prevent overextraction.
- Promoting circular water use strategies, such as recycling and reuse, within the industrial sector.
- Ensuring that water allocation policies balance industrial needs with environmental sustainability and public supply.

These countries must carefully manage their industrial growth to avoid exacerbating water stress in the face of limited resources.

[Cluster 0: Resource-Rich with Moderate Industrial Activity]

Countries in this cluster have very high levels of available water resources and moderate industrial water use. Relevant policy considerations include the following. However, given the relatively large internal variance within this cluster, policy approaches should be adapted at the country level to account for specific conditions and needs:

- Maintaining existing resource security while preparing for shifts in industrial or population-driven demand.
- Encouraging water-efficient technologies in the industrial sector to prevent long-term inefficiencies.
- Monitoring resource-intensive sectors to ensure sustainable extraction aligned with national development goals.

These countries have the potential to lead in sustainable water governance, but vigilance is necessary to preserve their resource advantages.

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.

5. Conclusion

This study applied an unsupervised learning approach to group countries based on two key indicators: available water resources and industrial water use. Dimensionality reduction was performed using Principal Component Analysis (PCA), and clustering configurations were explored using a grid search across various methods and parameters. After evaluation using silhouette score, Davies–Bouldin index, cluster balance (MaxMinRatio), and missing label rate, the final configuration was selected from the DBSCAN algorithm.

The analysis identified three groups of countries with distinct patterns of water availability and industrial demand. One cluster consisted of resource-abundant countries with moderate industrial usage, another showed limited resources and low industrial activity, and the third had moderate resources but high industrial demand. These results were visualized through a 2D PCA plot and a global map, and further examined using cluster profiles. Although one group was originally labeled as noise by DBSCAN, its geographic and economic relevance justified its inclusion in the analysis, though its less cohesive structure warrants cautious interpretation.

Based on these results, policy implications were proposed to reflect the needs and constraints of each group — ranging from resource protection and infrastructure planning to efficiency-focused industrial policy.

While DBSCAN yielded the most favorable results under the applied criteria, it is important to note that the outcome was shaped by filtering thresholds and an artificially created evaluation metric. Other methods, such as K-means or hierarchical clustering, may outperform DBSCAN under different conditions.

Overall, the study demonstrates that when appropriately validated and interpreted, unsupervised learning can help reveal underlying patterns in global water-use behavior and support more targeted, evidence-based policy development.