Clustering Countries on Integrated Socio-Economic Data

Project presentation

Anna Fabbri

Data Mining and Machine Learning

Problem description:

The aim of this project is to compare different unsupervised learning techniques on their capability of clustering world countries. This is done on the basis of real world socio economic data consisting in an integration of two datasets which required heavy data preprocessing.

The question(s) it tries to answer are:

1. Is there a way at all to categorize world countries in different clusters considering multiple aspects of their social, geographical and economic aspects, and to show evolution in time?
2. Which algorithm is able to give more significant results?

Dataset description:

The project integrates two datasets:

1. World Development Indicators (WDI) database, published by the World Bank; a comprehensive collection of global development data, it includes over 1,500 indicators (link here)
2. World Happiness Report, a survey of the state of global happiness that ranks countries by how ‘happy’ their citizens perceive themselves to be (link here)

The rationale was combining material development indicators of different origins with “holistic” wellbeing measures.

References:

The project takes inspiration from Saraiva, C., & Caiado, J. (2025). Global development patterns: A clustering analysis of economic, social and environmental indicators.
This original study, though, emphasizes the aspect of sustainability and employs a single algorithm. The project adopts a more exploratory approach by integrating datasets with different perspectives and testing multiple algorithms in order to try to identify when meaningful clusters can be obtained.

Preprocessing steps:

- **WDI attribute selection**

Macrocategory	Total	80%	90%	Attr. Sel.
Economic Policy & Debt	357	63	25	PCA
Health	250	125	78	PCA
Private Sector & Trade	151	36	3	PCA
Environment	144	82	50	PCA
Public Sector	132	2	0	Manual
Financial Sector	55	6	1	Manual
Social Protection & Labor	142	57	1	PCA
Education	156	4	2	Manual
Gender	14	1	0	Manual
Infrastructure	36	10	0	Manual
Poverty	24	0	0	NO
Misc.	27	0	0	NO
Trade	24	0	0	NO
Employment and Time Use	1	0	0	NO

```
Explained variance by component:
PC1: 0.1712 (0.1712 cumulative)
PC2: 0.1191 (0.2903 cumulative)
PC3: 0.0883 (0.3786 cumulative)
PC4: 0.0537 (0.4323 cumulative)
PC5: 0.0492 (0.4816 cumulative)
```

```
Top contributing indicators per component:

PC1:
  GDP per capita, PPP (constant 2021 international $) (-0.287)
  GDP per capita, PPP (current international $) (-0.287)
  GNI per capita, PPP (current international $) (-0.281)
  Services, value added per worker (constant 2015 US$) (-0.270)
  GNI per capita, Atlas method (current US$) (-0.268)
  GDP per capita (constant 2015 US$) (-0.250)
  Industry (including construction), value added per worker (constant 2015 US$) (-0.244)
  GDP per capita (current US$) (-0.238)
  Exports of goods and services (% of GDP) (-0.206)
  Agriculture, forestry, and fishing, value added (% of GDP) (+0.204)
```

Preprocessing steps:

- **WDI attribute selection**
- **WHR attribute selection**
- **Dataset integration**
- **Final dimension reduction**
- **Missing values cleaning and final normalization**

End result: 24 attributes, 1 dataset per year (2005/2022)

Reference year for clustering exploration: 2017

Algorithms chosen:

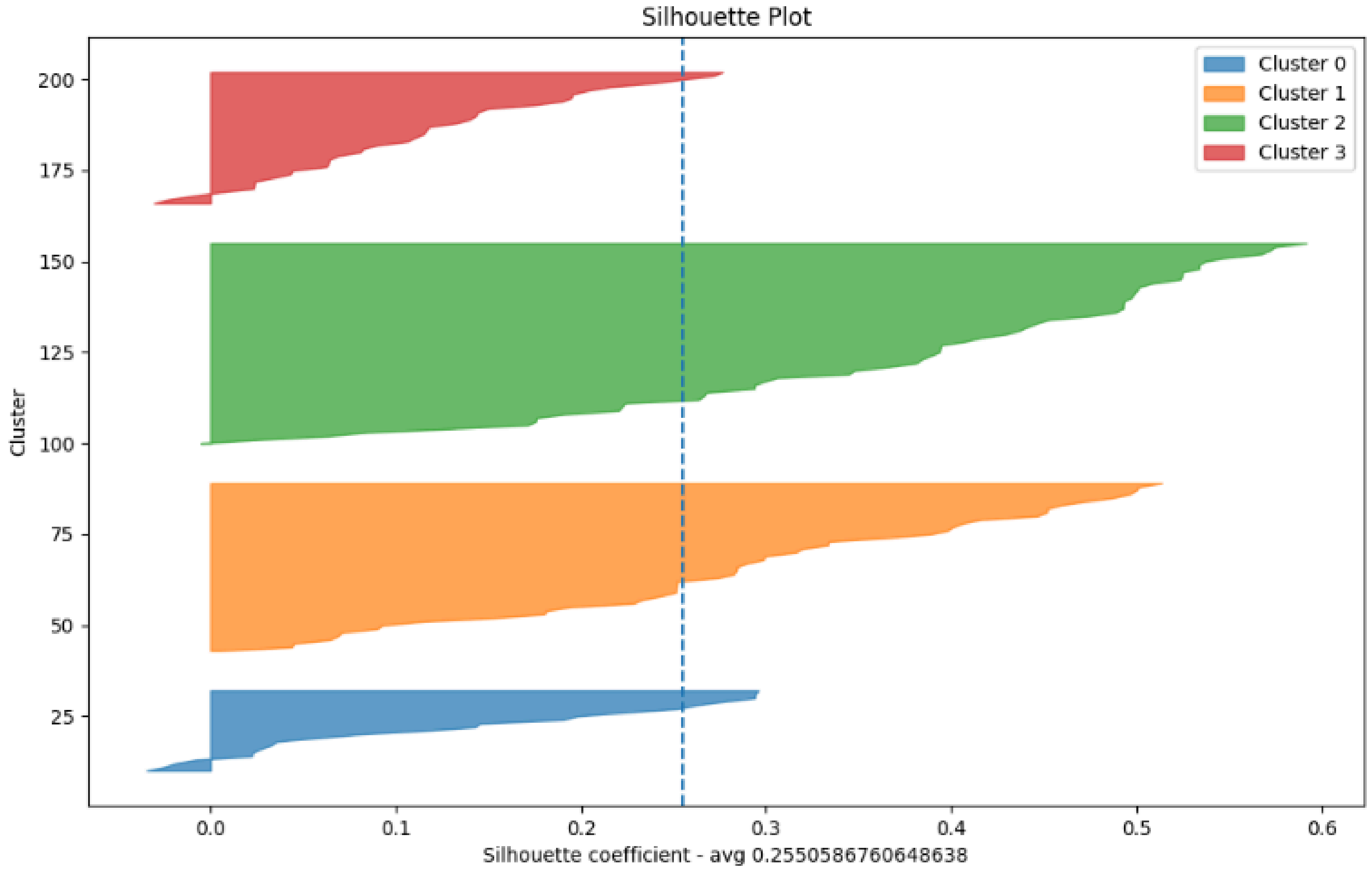
- K-means
- CLARA (K-medoids)
- DBSCAN
- Agglomerative Hierarchical

Multiple algorithms to compare different clustering techniques.

Clustering tendency measured
with Hopkins Statistic (30 iterations)

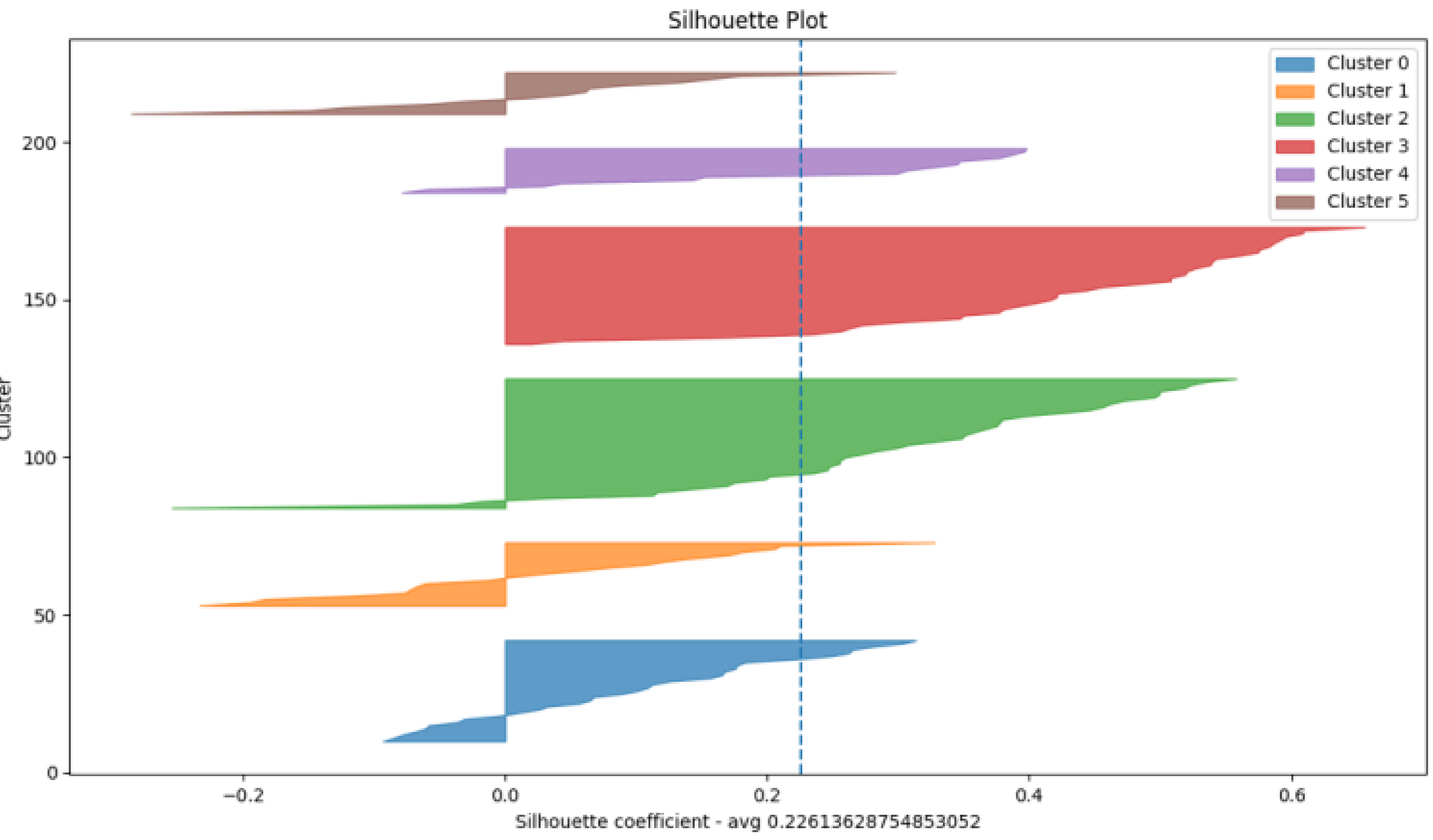
	Hopkins
Mean	0.7740
Std:	0.0169
Min:	0.7424
Max	0.7988

K-Means:



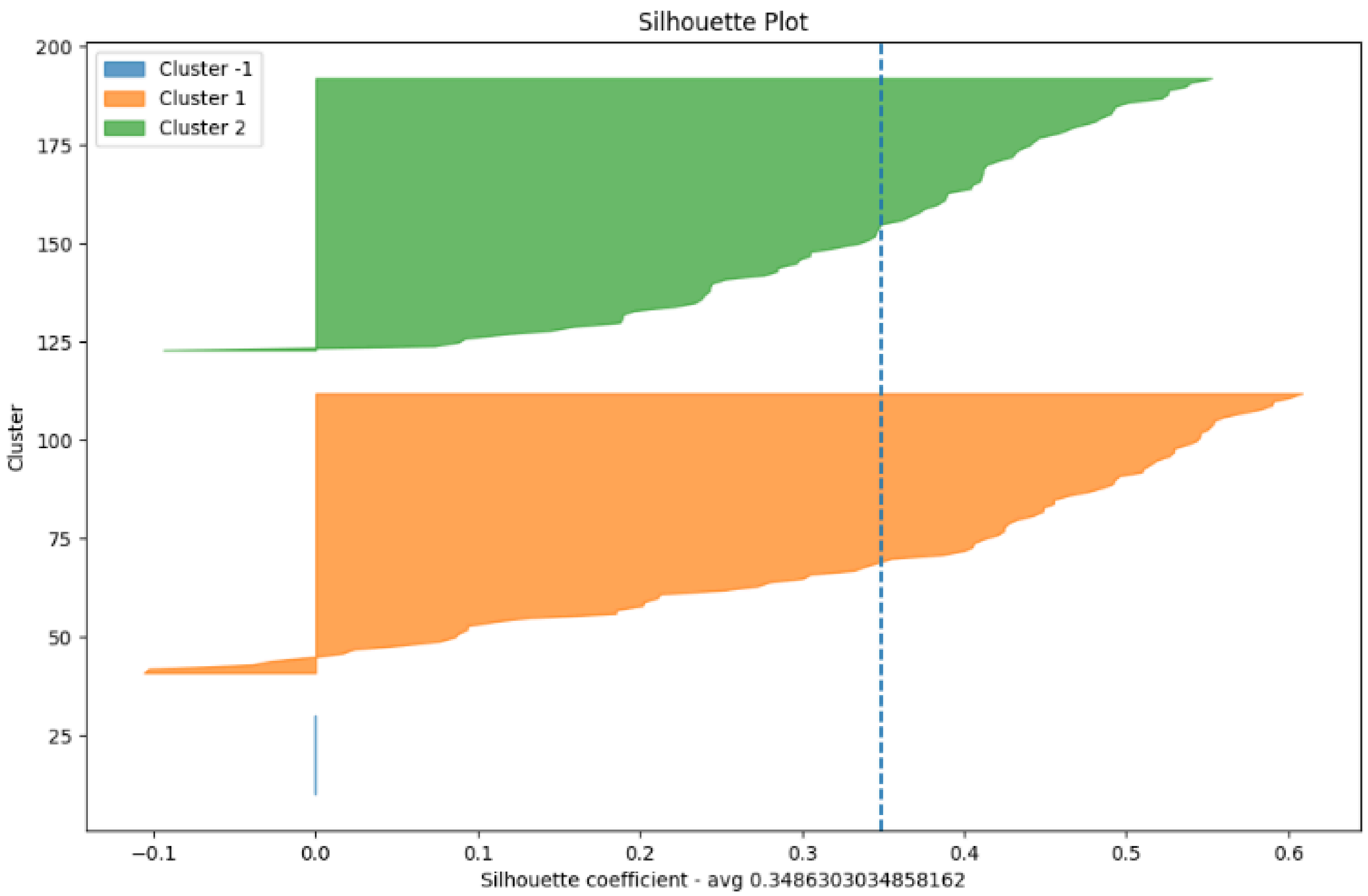
- **Cluster 0:** Good energy infrastructure, institutional weaknesses, high
- perceived corruption, unstable economy
- **Cluster 1:** Economically developed, urbanized countries.
- **Cluster 2:** Low income, limited digitalization and urbanization, high vulnerable employment.
- **Cluster 3:** Energy infrastructure in place but unsustainable, with significant social fragility.

CLARA:



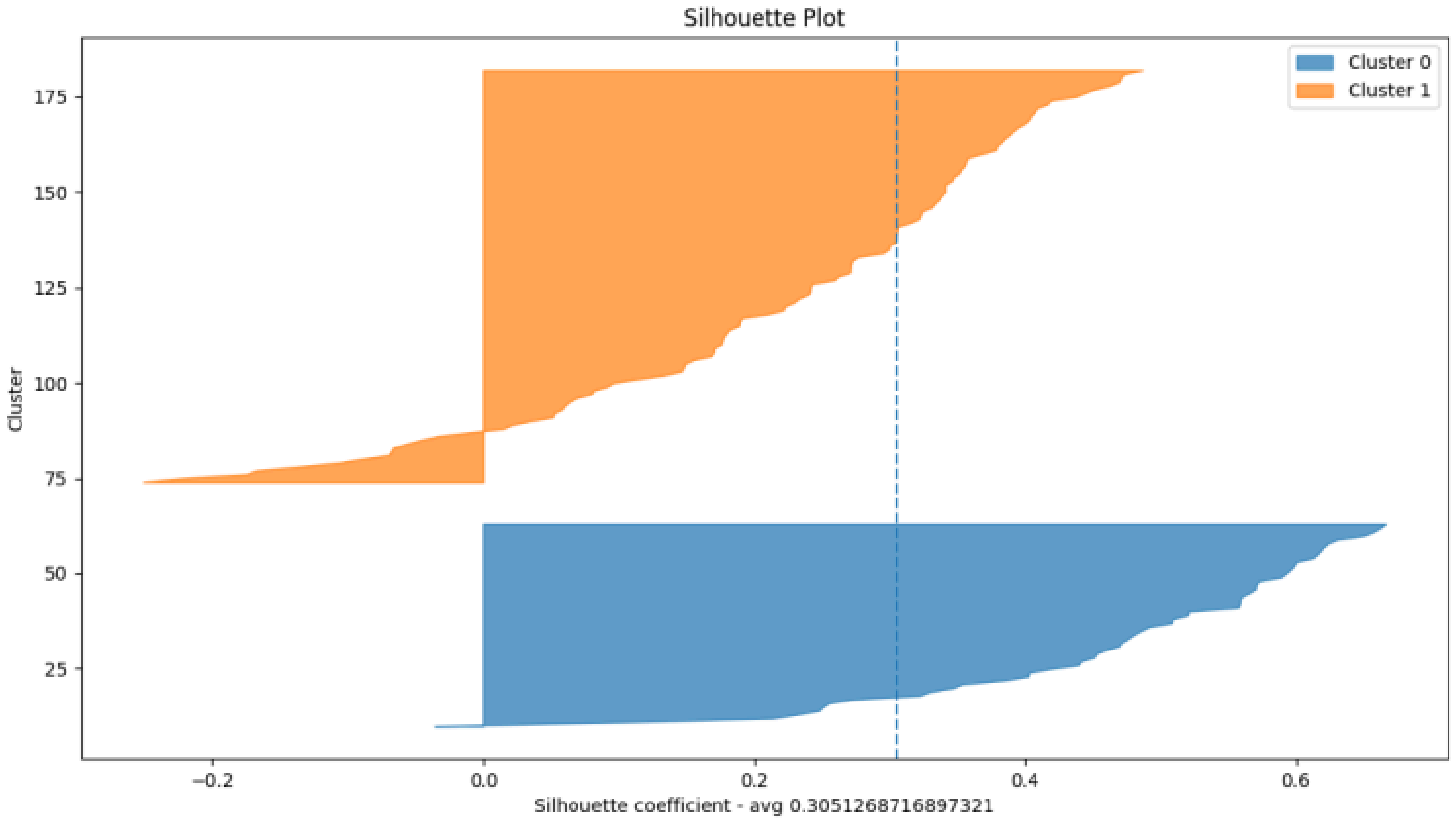
- **Cluster 0:** Strong energy access, low job participation, limited generosity, high corruption concerns.
- **Cluster 1:** Low income and industrial output, rural, high social support and generosity, widespread corruption concerns.
- **Cluster 2:** Economically strong and connected, low vulnerable jobs, high life satisfaction.
- **Cluster 3:** Economically weak and less connected, low urbanization, high renewables, mixed well-being.
- **Cluster 4:** High energy access and urbanization, limited renewables and fragile social cohesion.
- **Cluster 5:** Economically weak, low exports and digital access, fragile social mood.

DBSCAN:



- **This clustering configuration returned 2 clusters and 15 outlier countries**.
- **Cluster 1:** Low gdp per capita, low % of individuals using the internet, low industry value added per worker, low % of population ages 15-64, high vulnerable employment.
- **Cluster 2:** high access to electricity, high % of individuals using the internet, high % of population 15-64, low vulnerable employment.

HIERARCHICAL:

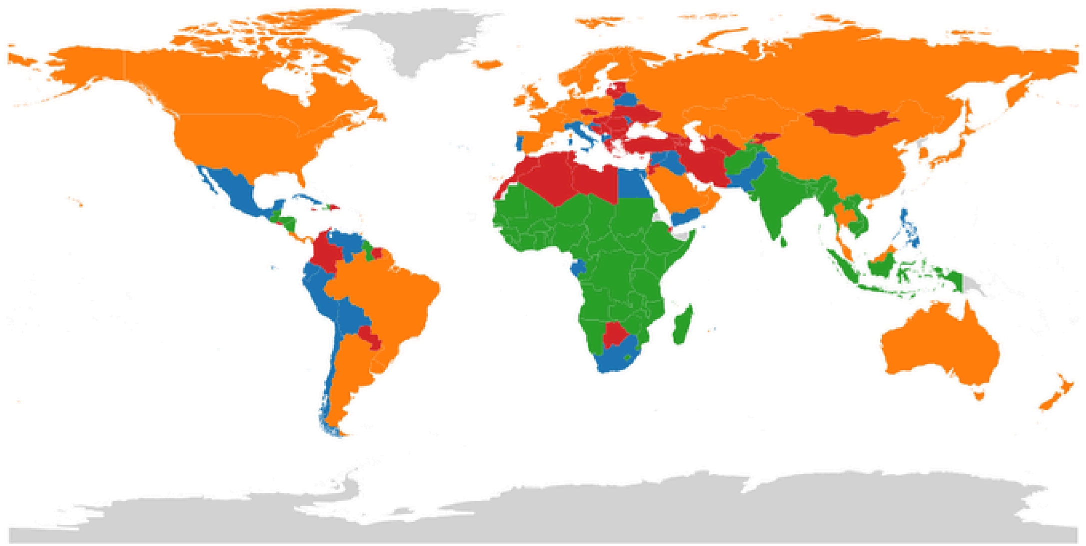


- **Cluster 0:** Economically weak, low digital and industrial activity, high employment vulnerability, limited social support.
- **Cluster 1:** High electricity access, more digitally connected, generally low vulnerable employment.

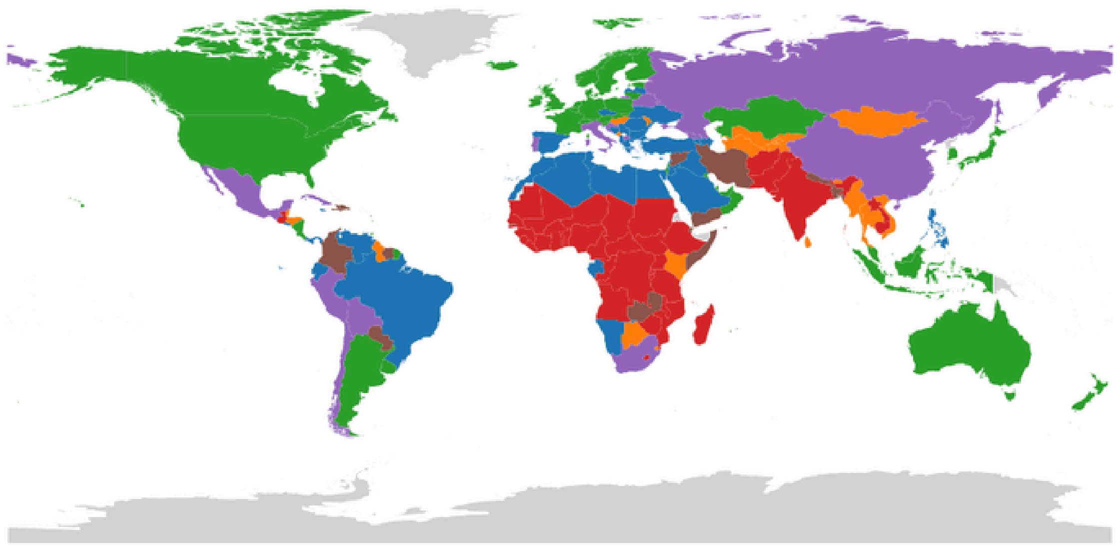
Results:

Algorithm	N. Clusters Found	Silhouette Score
Kmeans	4	0.255
Clara	6	0.226
DBSCAN	2	0.348
Hierarchical (complete linkage)	2	0.305

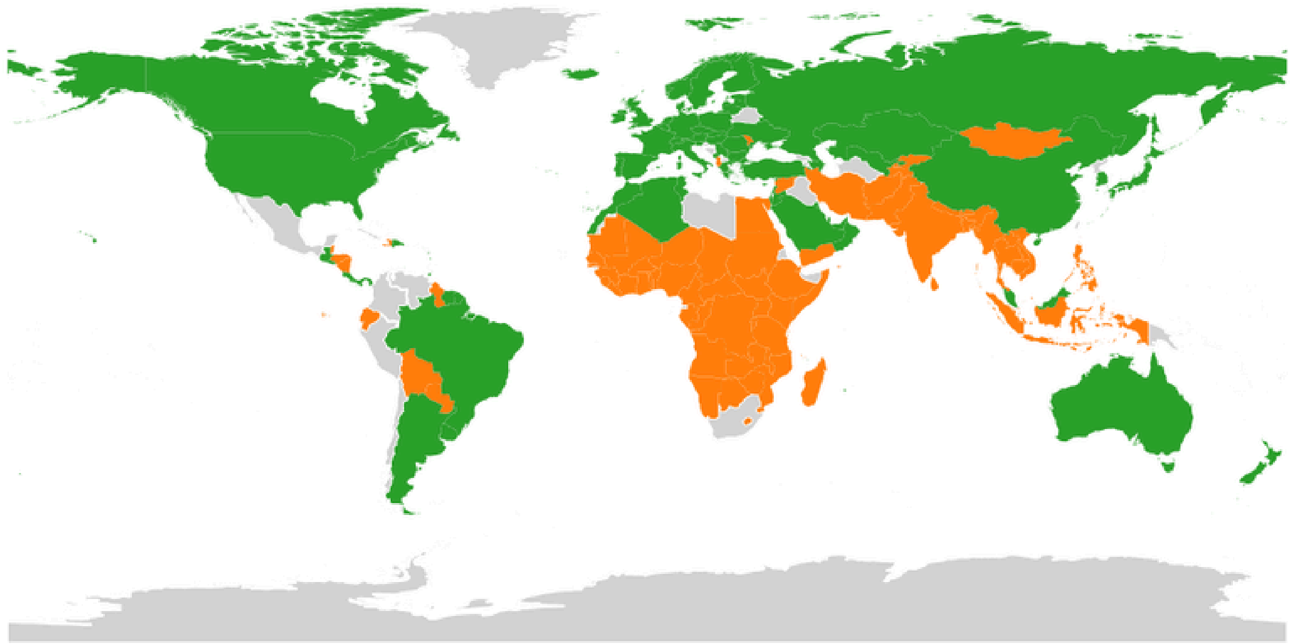
K-MEANS - Cluster 2017



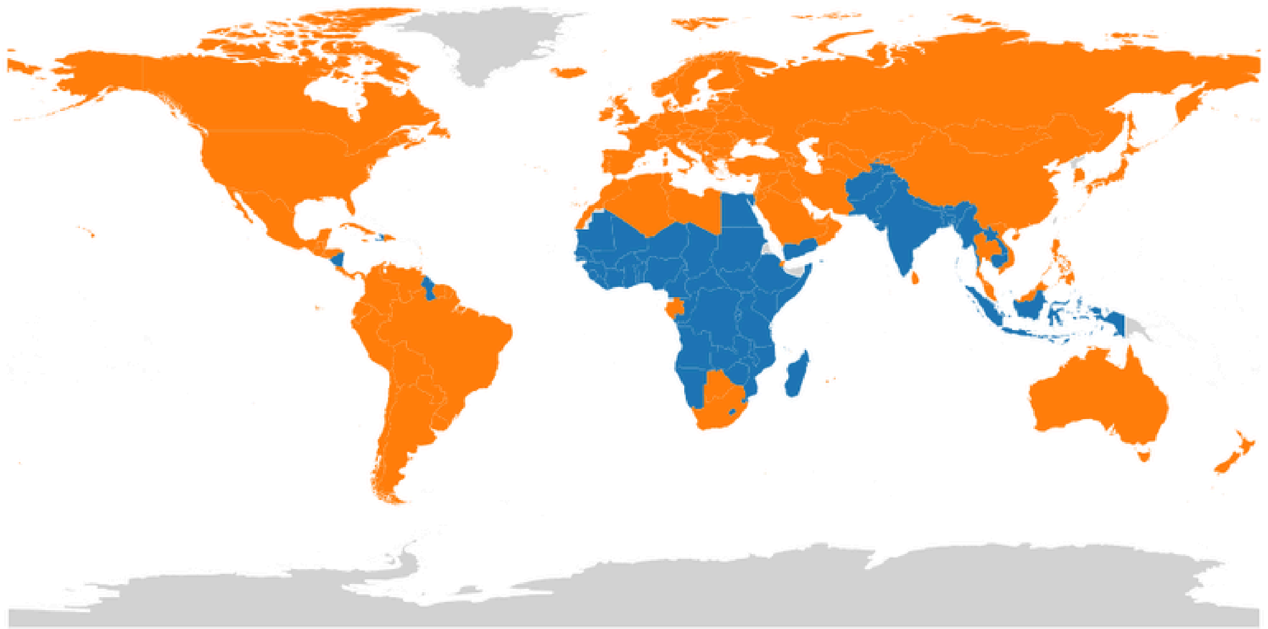
CLARA - Cluster 2017



DBSCAN - Cluster 2017



HIERARCHICAL - Cluster 2017



Temporal stability: K-MEANS

Table 11: Transition matrix for K-Means clusters (K=4)

From \To	0	1	2	3
0	**0.469**	*0.183*	0.066	*0.283*
1	0.064	**0.853**	0.005	0.078
2	0.024	0.007	**0.955**	0.014
3	*0.131*	0.104	0.018	**0.747**

Clusters not perfectly stable and interpretable over time. Some degree of mobility, particularly around Cluster 0, intrinsic to the structure of the data.

Temporal stability: CLARA

	0	1	2	3	4	5
0	**0.691**	0.059	*0.106*	0.003	0.097	0.044
1	*0.116*	**0.618**	0.053	0.075	0.069	0.069
2	0.087	0.031	**0.821**	0.006	0.048	0.007
3	0.006	0.035	0.010	**0.849**	0.011	0.089
4	*0.282*	*0.109*	*0.158*	0.040	**0.396**	0.015
5	*0.106*	0.097	0.040	*0.239*	0.027	**0.491**

Table 15: Average transition matrix for CLARA clustering with 6 clusters. Each entry (i, j) indicates the probability that a country in cluster i in a given year moves to cluster j in the following year.

structurally sufficiently coherent but more nuanced and more dynamic than the K-Means alternative.

Temporal stability: DBSCAN and HIERARCHICAL

	outl.	1	2
outl.	**0.309**	0.217	*0.474*
1	0.047	**0.838**	0.115
2	0.088	0.118	**0.794**

Table 19: Average transition matrix for DBSCAN clustering with 2 clusters.

Table 23: Transition matrix

From / To	Cluster 0	Cluster 1
Cluster 0	0.724	0.276
Cluster 1	0.204	0.796

Final results:

- **K-Means:** best balance between interpretability and temporal stability.
- **CLARA:** finer distinctions, with volatility.
- **DBSCAN:** robust two-cluster structure, successfully detected anomalies (e.g., post-2008 crisis).
- **Hierarchical clustering:** highlight rich–poor divide, showed sensitivity to yearly normalization.
- **No definitive “perfect” clustering emerged: the curse of dimensionality persists.**