Università di Pisa

Artificial Intelligence and Data Engineering

Clustering Countries on Integrated Socio-Economic Data

A Comparative Analysis

Author: Anna Fabbri
Course: Data Mining and Machine Learning
Academic Year: 2025/2026

Contents

1 Introduction

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, together, keep different domains of information, from geographical to demographic, political, social, economical and data gathered with a more "holistic" approach. The question(s) this project tries to answer are:

- 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?
- Which algorithm is able to let the most interesting clusters emerge?

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.

All clustering techniques were implemented and integrated into a Python application with a lightweight graphical user interface, designed to facilitate the execution of the algorithms and the collection of all the necessary outputs.

1.1 Methodology

The datasets were chosen due to the different nature of the data they represent. The data preprocessing phase focused mainly on dimensionality reduction, which was done at different levels. After the data preprocessing phase, cluster tendency was examined. Whether the dataset showed clustering tendencies or not, the aim of the project remained to try to find the clustering algorithm most able to delineate different groups between countries.

Since the datasets exhibit a temporal structure, in order to properly apply clustering techniques the decision was made to reason in a cross-sectional manner by fixing a reference year.

By treating each country as a single observation in a given year, clustering algorithms can focus on structural differences in socio-economic indicators rather then temporal fluctuations. At the end of the clustering algorithm choice phase, the clustering algorithm deemed most promising for producing meaningful clusters is run for the remaining years, allowing for a comparison of cluster compositions across different periods. Additionally, the developed application allows the user to dynamically select the year of interest, the algorithm and its hyperparameters, enabling interactive exploration of clustering results for any available year in the dataset.

2 Datasets

2.1 World Development Indicators

The first dataset used is the World Development Indicators (WDI) database, published by the World Bank. It is a comprehensive collection of global development data, providing key economic, social, and environmental statistics. It includes over 1,500 indicators covering more than 200 countries and territories, with data spanning several decades. The indicators are sourced from reputable national and international agencies, ensuring high-quality, consistent, and comparable data.

The dataset contains the following columns:

Name of the attribute	**Description**
Country Name	Name of the country
Country Code	ISO code identifying the country
Indicator Name	Name of the indicator
Indicator Code	Code identifying the indicator
1960	Value of the indicator in the year 1960
The same structure applies to the following columns, which report indicator values for all subsequent years from 1961 to 2024.	

Table 1: Description of the WDI dataset attributes

There are 1513 indicators for each country, spanning from GDP to health and education statistics. Many of them, from looking at the descriptions, seem to be highly correlated on each other. Dimensionality reduction will be one of the main challenges of preprocessing.

2.2 World Happiness Report

The World Happiness Report is a survey of the state of global happiness that ranks countries by how 'happy' their citizens perceive themselves to be. This dataset originates from the World Happiness Report and was accessed via Models Demystified: A Practical Guide from Linear Regression to Deep Learning by Michael Clark and Seth Berry. It combines wellbeing data with surveys made directly to citizens of all countries surveyed and finds a happiness score and ranking based on six main variables:

- Having someone to count on
- Log GDP per capita
- Healthy life expectancy
- Freedom to make life choices

- Generosity
- Freedom from corruption

The dataset contains the following columns:

NAME OF THE ATTRIBUTE	**MEANING**	**TYPE OF ATTRIBUTE**
Country	The country name	Text
Year	The year of the survey	Numeric – Interval scaled
life_ladder	The happiness score	Numeric
log_gdp_per_capita	The log of GDP per capita	Numeric
social_support	The social support score	Numeric
healthy_life_expectancy _at_birth	The healthy life expectancy at birth	Numeric
freedom_to_make_life _choices	The freedom to make life choices score	Numeric
generosity	The generosity score	Numeric
perceptions_of_corruption	The perceptions of corruption score	Numeric
positive_affect	The positive affect score	Numeric
negative_affect	The negative affect score	Numeric
confidence_in_national _government	The confidence in national government score	Numeric
happiness_score_sc	The happiness score – scaled version of the feature	Numeric

Table 2: Description of the WHR dataset attributes

These two datasets were chosen together because one of the main aims of the project was to try to find a way to characterize countries based on different aspects, ranging from more material characteristics to more "holistic" ones.

3 Preprocessing

The WDI Dataset, as we have already seen, originally spans a significantly higher number of years than the WHR Dataset. In order to combine them effectively, we should limit our analysis to the years where both datasets are present. Another characteristic of the WDI dataset is its huge amount of indicators. Regarding the WHR dataset, a number of data is missing. The steps that have to be taken are the following:

- the WDI dataset's dimension in relation to the number of indicators (which, when we will perform the cross-sectional clustering, will be considered attributes) must be drastically reduced;
- missing data must be fixed.

3.1 Coarse WDI Dimensionality Reduction

The first steps that were taken in order to take down the WDI dataset enormous dimension were:

- eliminating all observation years in which WHR data were not collected (this was done manually by opening the csv and eliminating the columns related to the target years)
- eliminating all attributes whose number of empty values exceed 20% of the total.

before proceeding with the second step, an exploratory analysis of the WDI dataset attributes was conducted by grouping them in thematic macrocategories. In the file WDISeries.csv there is the whole list of indicators along with their in-depth explanation, which aided in this process.

All in all, the presence of 87 fine-grained indicator categories was revealed, which were subsequently aggregated into 14 macro-categories in order to obtain a higher-level view of the dataset structure.

The dataset was split according to macrocategories in order to find the coverage of each attribute pertaining to each macrocategory, the goal there being trying to keep a representation of as many macrocategories als possible in de final dataset.

This initial pruning gave the following result per macrocategory:

The steps leading to this were performed in Python whilst the step of selecting threshold-passing attributes was done manually in order to take this opportunity to look at the remaining attributes, since in order to properly select relevant attributes a bit of domain knowledge and human eye is still needed.

By viewing this data, different strategies per each macrocategory were devised in order to choose the most significant attributes (the ones that have the best coverage, are most representative of the macrocategory and iyeld more information):

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

Table 3: Summary of threshold-passing attributes by macrocategory.

- For macrocategories "Private Sector and Trade", "Public Sector", "Financial Sector", "Social Protection and Labor", "Education", "Gender" and "Infrastructure" (all the macrocategories which have less than 10 attributes that surpass the 80% coverage threshold and less than 5 attributes surpassing the 90% coverage threshold) the attributes were chosen manually;
- For macrocategories "Economic Policy and Debt", "Health", "Private Sector and trade", "Environment", "Social Protection and labor", Principal Component Analysis was applied to aid the choice.

Selecting attributes manually a problem was encountered woth the macrocategory "public sector": namely, the only two attributes that passed the treshold were not representing data that was deemed as significant as the attribute right below the treshold. a decision was made to take that into account instead.

3.1.1 Principal Component Analysis as an exploratory tool

To prepare data for explorative PCA we take the data that we already split by attribute macrocategory and:

- section it to only consider one specific year and pivot it to become a Country x Indicator dataset
- impute missing values indicator per indicator by using attribute mean
- perform a z-score normalization per indicator. (the z-score normalization will be performed at the very end when we have the final dataset)

Macrocategory	**Initial # indicators**	**min. 80% coverage**	**Final # selected**
Economic Policy & Debt	357	63	3
Health	250	125	3
Private Sector & Trade	151	36	2
Environment	144	82	4
Public Sector	132	2	1
Financial Sector	55	6	1
Social Protection & Labor	142	57	3
Education	156	4	1
Gender	14	1	1
Infrastructure	36	10	1
Poverty	24	0	0
Misc.	27	0	0
Trade	24	0	0
Employment and Time Use	1	0	0

Table 4: Summary of indicators and number of selected attributes by macrocategory.

PCA was not used as a dimensionality reduction technique per se, but rather to guide feature selection by identifying the most informative indicators within each macro-category. Instead of directly using principal components, which are linear combinations of multiple variables and may be difficult to interpret substantively, original indicators with high loadings on the leading components were selected. The selection was guided also by the significance of relevant attributes and their descriptions in the dataset.

We finally end up with 20 attributes. The dataset is then built as follows: Data is split per year. Only selected attributes are kept for each year. Missing data is imputed in the same fashion as for the PCA preparation step, and normalized in the same way.

Looking at the WHR dataset briefly, the decision was made to eliminate the already normalized versions of the attributes to normalize the whole dataset in a homogeneous way. The attribute log_gdp_per_capita had almost duplicates in the reduced WDI dataset, so a choice was made to remove it, both to reduce dimensionality and to focus on the more "holistic" attributes that this dataset provides.

3.2 Data Integration and finer dimension reduction

Data was normalized and then integrated on a year-by year basis. It was observed that the WHR dataset has many "structural" missing values, that is to say, many countries were not surveyed every year nor ever. These missing values were left unaltered. The only choice made was to remove all countries that were never surveyed for the WHR.

After integrating the two datasets a correlation analysis bewteen the attributes was performed, in order to see whether some attributes were redundant.

Attr. 1	**Attr. 2**	**# Y.**
Access to electricity (% of population)	Age dependency ratio, young (% of working-age population)	7
Age dependency ratio, young (% of working-age population)	Individuals using the Internet (% of population)	3
Age dependency ratio, young (% of working-age population)	Population ages 0-14 (% of total population)	18
Age dependency ratio, young (% of working-age population)	Population ages 15-64 (% of total population)	17
Age dependency ratio, young (% of working-age population)	healthy_life_expectancy_at_birth	4
GDP per capita, PPP (current international $)	Services, value added per worker (constant 2015 US$)	8
Individuals using the Internet (% of population)	Population ages 0-14 (% of total population)	7
Individuals using the Internet (% of population)	Vulnerable employment, total (% of total employment) (modeled ILO estimate)	3
Industry (including construction), value added per worker (constant 2015 US$)	Services, value added per worker (constant 2015 US$)	10
Merchandise exports by the reporting economy (current US$)	Merchandise imports by the reporting economy (current US$)	18
Population ages 0-14 (% of total population)	Population ages 15-64 (% of total population)	12
Population ages 0-14 (% of total population)	healthy_life_expectancy_at_birth	5

Table 5: Couplees of attributes with high correlation, number of years involved.

This table shows the results. The following decisions were made:

- Age dependency ratio, young (% of working-age population) appears in many cases, very redundant, is removed
- Services, value added per worker (constant 2015 US$) appears in a few cases, considered redundant, is removed
- Between Merchandise exports by the reporting economy (current US$) and Merchandise imports by the reporting economy (current US$) only the former is kept
- Population ages 0-14 (% of total population) also appears in a few cases and will be removed

3.3 Normalization and Missing Values

The dataset, after this final merge and ulterior reduction, was normalized. Before normalizing, a missing value strategy was revised, because of WHR having many missing values. What was decided was that since all countries have at least 1 observation, fill missing values with the median of recorded values during the years in the country with an added random jitter.

3.4 Final dataset

The final dataset contains 24 attributes per country for each year, all normalized with z-score, and with an additional column to signal outliers. The year with less initially missing values imputated to the lack of WHR is 2017, which is the year that will be used for cluster exploration. It is the most "organic" one. The interface still gives us the possibility to use every year at our disposal to cluster.

Attribute	Description
Proportion of seats held by women in national parliaments (%)	Women in parliaments are the percentage of parliamentary seats in a single or lower chamber held by women.
Compulsory education, duration (years)	Duration of compulsory education is the number of years that children are legally obliged to attend school.
Inflation, GDP deflator (annual % growth)	Inflation as measured by the annual growth rate of the GDP implicit deflator shows the rate of price change in the economy as a whole. The GDP implicit deflator is the ratio of GDP in current local currency to GDP in constant local currency.
Individuals using the Internet (% of population)	Individuals who have used the Internet (from any location) in the last 3 months.
Military expenditure (% of GDP)	Military expenditure by country as percentage of gross domestic product.
GDP per capita, PPP (current international $)	This indicator provides values for gross domestic product (GDP) expressed in current international dollars, converted by purchasing power parities (PPPs).
Industry (including construction), value added per worker (constant 2015 US$)	Value added is the total value of output produced and deducting the total value of intermediate consumption of goods and services used to produce that output.
Urban population (% of total population)	Urban population refers to people living in urban areas as defined by national statistical offices.
Access to electricity (% of population)	Access to electricity is the percentage of population with access to electricity.
Renewable energy consumption (% of total final energy consumption)	Renewable energy consumption is the share of renewables energy in total final energy consumption.
Land area (sq. km)	Land area is a country's total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones.
Population ages 15-64 (% of total population)	Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Merchandise exports by the reporting economy (current US$)	Total merchandise exports by the reporting economy to the rest of the world, as reported in the IMF's Direction of trade database.
Net migration	Number of immigrants minus the number of emigrants, including both citizens and noncitizens.
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	Contributing family workers and own-account workers as a percentage of total employment.
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	Employment to population ratio is the proportion of a country's population that is employed.
happiness_score	The average life evaluation score for the year in question.
social_support	If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?
healthy_life_expectancy_at_birth	Based on data from the World Health Organization Global Health Observatory.
freedom_to_make_life_choices	Are you satisfied or dissatisfied with your freedom to choose what you do with your life?
generosity	Have you donated money to a charity in the past month?
positive_affect	The national average of binary responses (0=no, 1=yes) about three emotions experienced on the previous day: laughter, enjoyment, and interest.
negative_affect	The national average of binary responses (0=no, 1=yes) about three emotions experienced on the previous day: worry, sadness, and anger.
perceptions_of_corruption	The average of two questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

Table 6: Final attributes with their descriptions.

4 Clustering

In this section first we assess the clustering tendency of the dataset, and then we try five different algorithms pertaining to four different clustering algorithm categories. This is done in order to seek the algorithm which gives us the most interesting clusters on the cross sectional data set that we deemed the most interesting, due to it being the one which had originally the least null values. after that we try to find the same clusters on data from the other years.

4.1 Clustering Tendency

To assess clustering tendency, Hopkins statistic was implemented. Hopkisn statistic works by sampling points from the dataset and from an uniform space. 30 trials were performed, both considering the original dataset and performing PCA beforehand (60 trials in total) and the results displayed a good clustering tendency of the dataset.

	Hopkins	Hopkins+PCA
Mean	0.7740	0.7386
Std:	0.0169	0.0246
Min:	0.7424	0.6808
Max	0.7988	0.7785

Five algorithms are chosen and implemented from scratch in Python in order to perform the clustering and in order to compare their final results and see what they show. Clustering will be performed on the full set of attributes to preserve interpretability, without reducing dimensionality beforehand. To visualize the results, several techniques will be adopted:

- Principal Component Analysis (PCA) will be applied only for visualization purposes, projecting the data onto 2 dimensions to explore and illustrate the structure of the resulting clusters;
- For each cluster, the mean and standard deviation of each attribute will be examined in order to interpret the cluster;
- Silhouette plot and score will be shown;
- "Clusterized" map of the world.

4.2 The algorithms

Five algorithms were chosen. The rationale with which every algorithm was chosen was to try to explore how different clustering techniques tackle data that walks the curse of dimensionality line.

4.2.1 K-means

The algorithm The K-Means clustering algorithm was chosen for its simplicity, and because it's one algorithm where we can easily see the evolution in time of the very same clusters that we will look for in the 2017 dataset. How it works: K-Means partitions the data into K clusters by iteratively minimizing the distance between each point and the centroid of its assigned cluster. K initial centroids are chosen randomly, then each point is assigned to the nearest centroid according to the selected distance metric. Centroids are then updated as the mean of the points assigned to each cluster. This process repeats until the centroid shift is below a tolerance threshold or the maximum number of iterations is reached.

The two distance metrics implemented were euclidean and cosine.

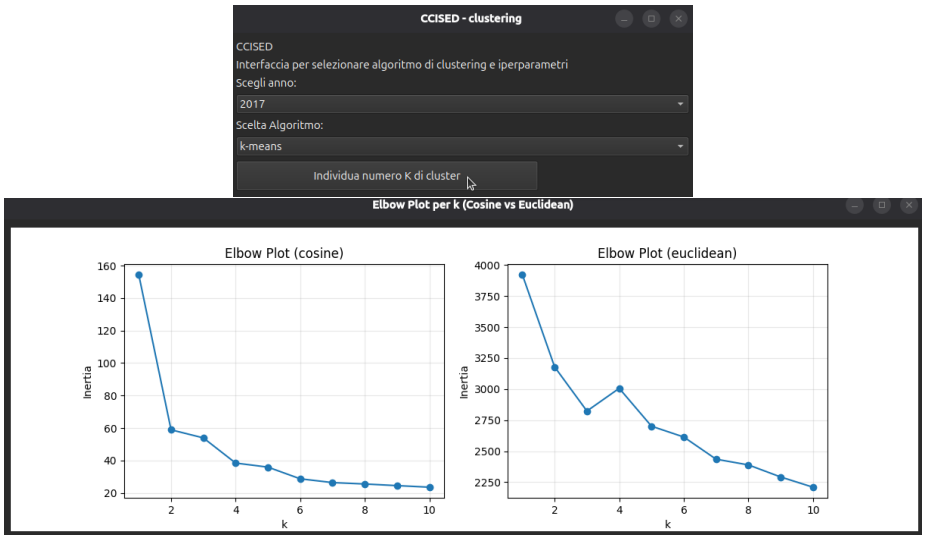

Figure 1: Elbow Plot: K-means

Elbow plot Elbow method was tried with both cosine and euclidean distance and the plot showed a more regular elbow with cosine distance, as expected because of the dimension not giving reliable euclidean distance results.

The elbow plot shows 4 and 6 to be possible values of k. the algorithm is then run with k = 4 and cosine distance as the distance metric, and later on with k = 6 and cosine distance as the distance metric.

K-Means with k = 4 The silhouette plot shows an average of 0,255. Since the number of dimensions is high, de variegate quality and amount of data and the fact that this is all based on real life data that models complex realities like whole countries are, we can expect our clusters to exist but not be the strongest clsuters to ever exist.

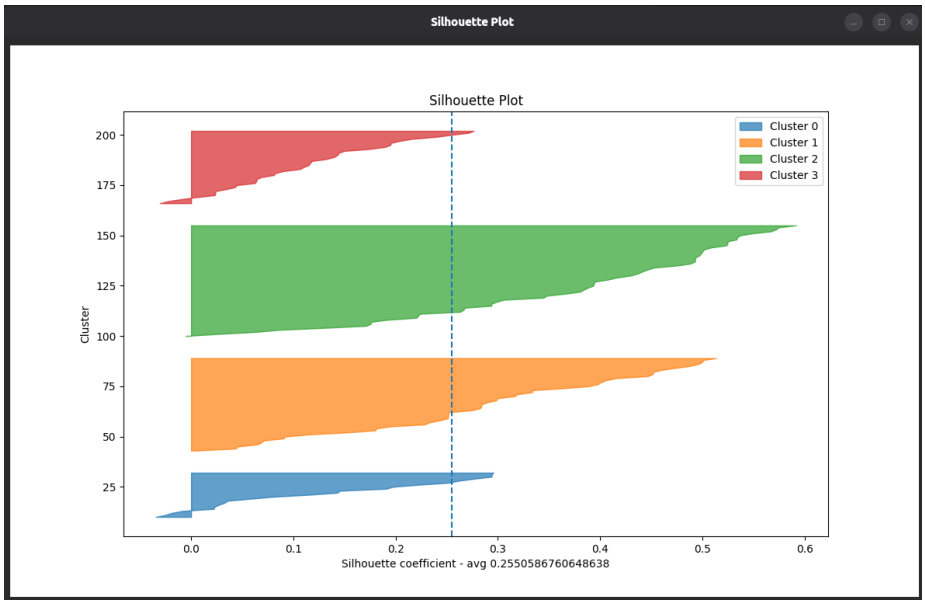

Figure 2: K means with k = 4: silhouette

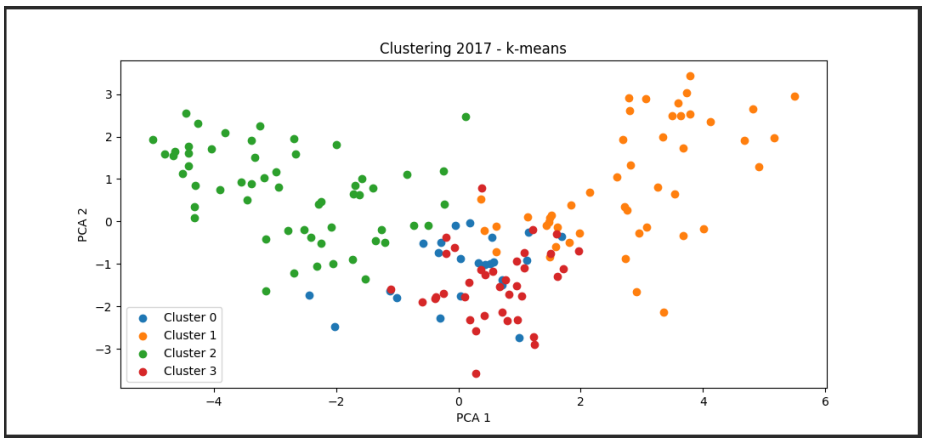

Figure 3: K means with k = 4: 2-dim PCA

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping, which we mainly see between clusters 0 and 3.

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the mean value.

	0	1	2	3
Access to electricity (% of population)	0.468 ± 0.223	0.612 ± 0.034	-1.043 ± 1.091	0.495 ± 0.348
Compulsory education, duration (years)	0.623 ± 1.116	0.264 ± 0.843	-0.581 ± 0.908	0.156 ± 0.859
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	-0.444 ± 1.031	0.383 ± 0.630	0.257 ± 1.122	-0.602 ± 0.826
GDP per capita, PPP (current international $)	-0.233 ± 0.425	1.160 ± 1.078	-0.769 ± 0.167	-0.156 ± 0.422
Individuals using the Internet (% of population)	0.017 ± 0.573	1.063 ± 0.436	-1.102 ± 0.497	0.306 ± 0.524
Industry (including construction), value added per worker (constant 2015 US$)	-0.255 ± 0.394	1.019 ± 1.354	-0.567 ± 0.188	-0.268 ± 0.301
Inflation, GDP deflator (annual %)	0.491 ± 2.623	-0.108 ± 0.076	-0.044 ± 0.320	-0.099 ± 0.075
Land area (sq. km)	-0.141 ± 0.263	0.400 ± 1.764	-0.131 ± 0.316	-0.214 ± 0.295
Merchandise exports by the reporting economy (current US$)	-0.149 ± 0.472	0.690 ± 1.644	-0.323 ± 0.194	-0.286 ± 0.158
Military expenditure (% of GDP)	-0.148 ± 0.675	0.144 ± 1.436	-0.221 ± 0.655	0.251 ± 0.896
Net migration	-0.429 ± 1.448	0.332 ± 1.226	-0.137 ± 0.718	0.051 ± 0.522
Population ages 15-64 (% of total population)	0.135 ± 0.667	0.719 ± 0.777	-0.931 ± 0.852	0.405 ± 0.446
Proportion of seats held by women in national parliaments (%)	0.433 ± 1.113	0.216 ± 1.001	-0.136 ± 1.085	-0.335 ± 0.629
Renewable energy consumption (% of total final energy consumption)	-0.438 ± 0.669	-0.564 ± 0.664	1.039 ± 0.782	-0.577 ± 0.472
Urban population (% of total population)	0.160 ± 0.785	0.850 ± 0.702	-0.994 ± 0.629	0.324 ± 0.584
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.195 ± 0.682	-0.877 ± 0.409	1.082 ± 0.723	-0.396 ± 0.496
happiness_score	-0.327 ± 1.094	0.757 ± 0.775	-0.407 ± 0.784	-0.163 ± 1.001
social_support	-1.269 ± 1.348	0.432 ± 0.889	-0.115 ± 0.658	0.397 ± 0.551
healthy_life_expectancy_at_birth	-0.111 ± 0.001	-0.110 ± 0.000	-0.112 ± 0.001	0.381 ± 2.089
freedom_to_make_life_choices	-0.280 ± 1.058	0.410 ± 0.943	-0.111 ± 0.982	-0.101 ± 0.852
generosity	-0.427 ± 0.653	0.413 ± 1.006	0.213 ± 1.019	-0.569 ± 0.817
perceptions_of_corruption	0.588 ± 0.272	-0.481 ± 1.034	0.047 ± 0.960	0.164 ± 1.086
positive_affect	0.165 ± 0.546	0.459 ± 0.710	0.152 ± 0.852	-0.932 ± 1.179

Figure 4: K means with k = 4: average and standard deviation for each attribute per cluster

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

In this clustering configuration, the salient attributes per cluster (high or low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): High access to electricity, high perception of corruption.

  Other important attributes that are visible from the heatmap are: extremely variable inflation, highly variable but very low social support.

  Good energy infrastructure, institutional weaknesses, high perceived corruption, and an unstable economy

- Cluster 1 (oranje): High access to electricity, high percentage of people using the internet, low vulnerable employment.

  Other important attributes that are visible from the heatmap are: highly variable but high GDP per capita and industry value added per worker. Variable but high percentage of urban population. Highly variable but on the lower side perception of corruption.

  Economically developed, digital, and urbanized, with low vulnerable employment.

- Cluster 2 (green): Low GDP per capita, low percentage of people using the internet, low urban population.

  Other important attributes that are visible from the heatmap are: Highly variable but low access to electricity. Variable but low percentage of population ages 15-64. Highly variable but on the lower side happiness score. Variable but high renewable energy consumption. Very high but variable vulnerable employment.

  Low income, limited digitalization and urbanization, with high vulnerable employment.

- Cluster 3 (red): High access to electricity, low renewable energy consumption.

  Other important attributes that are visible from the heatmap are: variable but low employment to population ratio, variable but on the higher side number of individuals using the internet, urban popluation and population aged 15-64. Variable but low generosity and positive affect.

  Energy infrastructure in place but unsustainable, with significant social fragility.

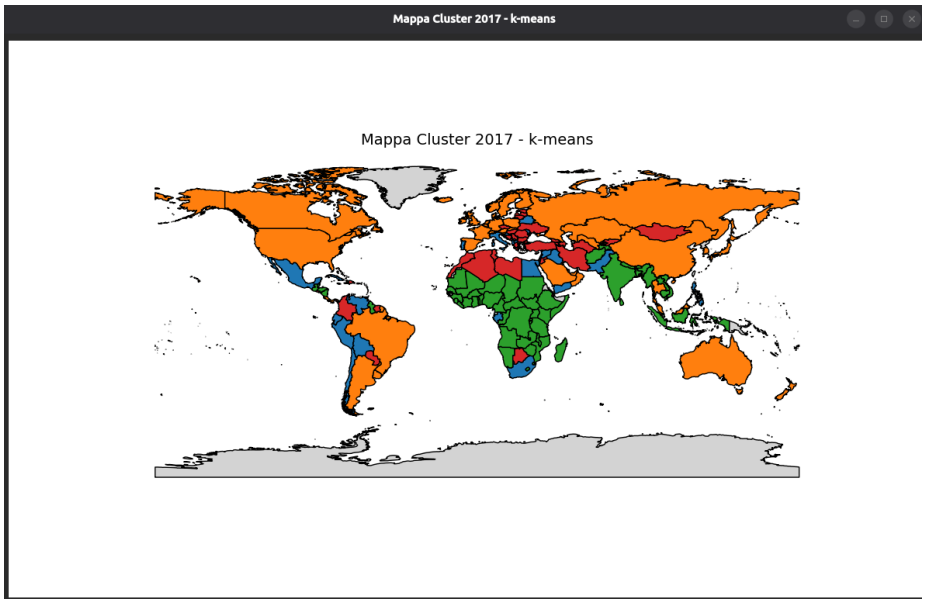

Figure 5: K means with k = 4: world plot

This data visualization is useful to see whether the division in clusters is completely random or not. In this situation we have:

- cluster 0 (blue): countries like Italy, Mexico, Portugal, Egypt, Chile

- cluster 1 (orange): countries like France, China, Australia, Russia, the US
- cluster 2 (green): countries like India, Senegal, Cameroon, Ghana, Vietnam
- cluster 3 (red): Countries like Morocco, Palestine, Ukraine, Iran, Libya

We can conclude that this clustering does not look totally random, as countries grouped together seem to share similar economic, social, and infrastructural characteristics rather than geographic proximity alone.

K-Means with k = 6 Now we repeat the analysis with K = 6 instead of 4.

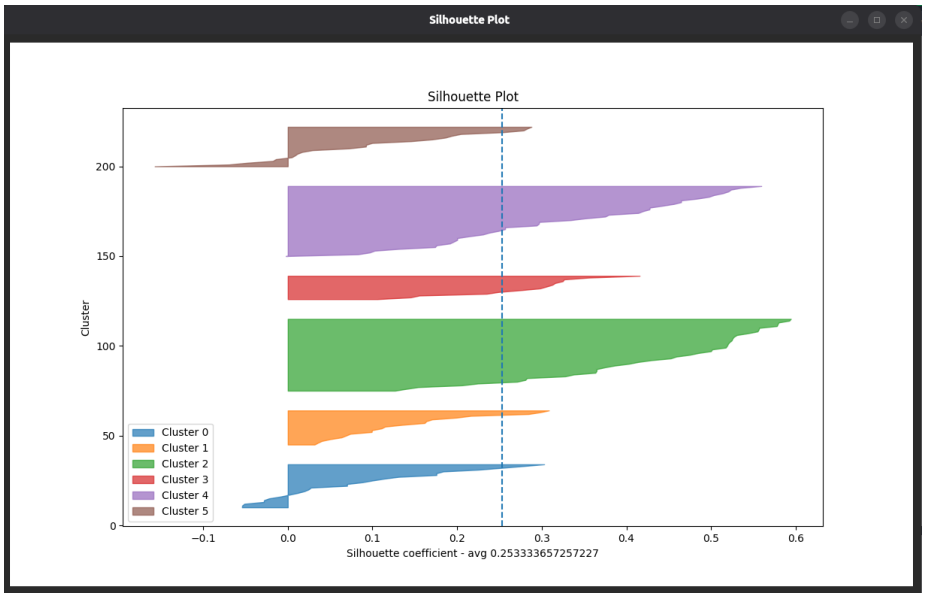

Figure 6: K means with k = 6: silhouette

The silhouette plot shows an average of 0,253, which is only slightly lower than what we obtained with k=4. The dimensions of clusters are very different and in particular we have a number of misclustered points in cluster 0 (the blue one) and cluster 5 (the brown one).

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping. since we do have more clusters there is some high overlapping at the center, whilst at the borders clusters appear to be more separated.

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the mean value.

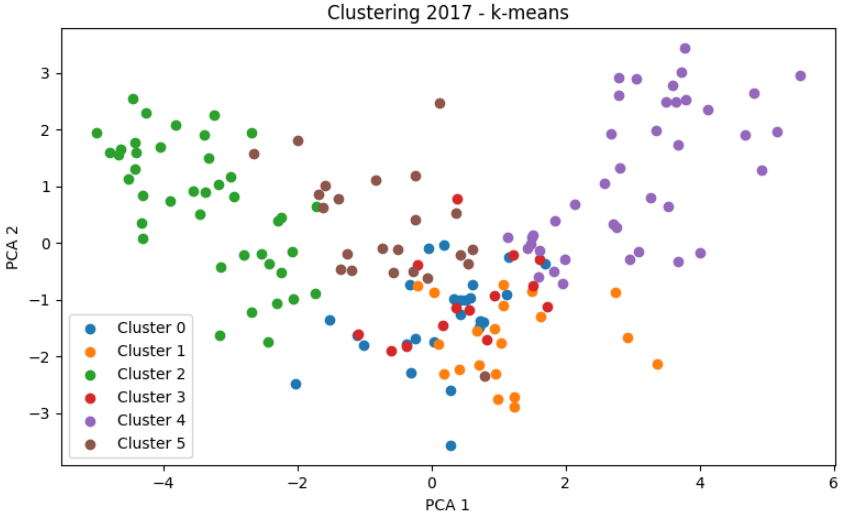

Figure 7: K means with k = 6: 2-dim PCA

	0	1	2	3	4	5
Access to electricity (% of population)	0.425 ± 0.392	0.494 ± 0.352	-1.460 ± 0.928	0.482 ± 0.388	0.617 ± 0.003	0.318 ± 0.408
Compulsory education, duration (years)	0.200 ± 0.949	0.492 ± 0.934	-0.712 ± 0.775	0.185 ± 1.183	0.256 ± 0.869	0.063 ± 1.051
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	-0.751 ± 0.985	-0.429 ± 0.824	0.213 ± 1.191	-0.264 ± 0.929	0.370 ± 0.679	0.325 ± 0.778
GDP per capita, PPP (current international $)	-0.186 ± 0.437	0.004 ± 0.548	-0.827 ± 0.130	-0.178 ± 0.473	1.338 ± 1.028	-0.531 ± 0.240
Individuals using the Internet (% of population)	0.111 ± 0.573	0.493 ± 0.421	-1.280 ± 0.412	0.325 ± 0.503	1.138 ± 0.359	-0.449 ± 0.580
Industry (including construction), value added per worker (constant 2015 US$)	-0.223 ± 0.395	-0.108 ± 0.483	-0.597 ± 0.192	-0.305 ± 0.199	1.203 ± 1.364	-0.490 ± 0.132
Inflation, GDP deflator (annual %)	-0.063 ± 0.123	0.526 ± 2.821	-0.024 ± 0.373	-0.086 ± 0.085	-0.113 ± 0.060	-0.092 ± 0.082
Land area (sq. km)	-0.172 ± 0.259	0.487 ± 1.993	-0.093 ± 0.334	-0.262 ± 0.247	0.188 ± 1.348	-0.225 ± 0.240
Merchandise exports by the reporting economy (current US$)	-0.179 ± 0.459	-0.117 ± 0.357	-0.340 ± 0.172	-0.337 ± 0.069	0.795 ± 1.750	-0.260 ± 0.260
Military expenditure (% of GDP)	-0.114 ± 0.584	1.258 ± 1.744	-0.187 ± 0.673	-0.039 ± 0.609	-0.098 ± 0.886	-0.431 ± 0.538
Net migration	-0.005 ± 0.445	-0.345 ± 1.094	-0.308 ± 1.224	0.087 ± 0.448	0.510 ± 1.192	-0.086 ± 0.266
Population ages 15-64 (% of total population)	0.095 ± 0.623	0.510 ± 0.483	-1.305 ± 0.608	0.447 ± 0.412	0.682 ± 0.810	0.313 ± 0.628
Proportion of seats held by women in national parliaments (%)	0.469 ± 1.077	-0.492 ± 0.712	-0.131 ± 1.113	-0.258 ± 0.610	0.395 ± 0.941	-0.373 ± 0.857
Renewable energy consumption (% of total final energy consumption)	-0.468 ± 0.659	-0.753 ± 0.396	1.234 ± 0.757	-0.293 ± 0.507	-0.554 ± 0.677	0.117 ± 0.777
Urban population (% of total population)	0.237 ± 0.664	0.626 ± 0.514	-1.024 ± 0.648	0.230 ± 0.483	0.976 ± 0.626	-0.816 ± 0.559
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.383 ± 0.570	-0.534 ± 0.584	1.325 ± 0.587	-0.259 ± 0.527	-0.935 ± 0.312	0.313 ± 0.640
happiness_score	-0.335 ± 1.049	-0.202 ± 1.072	-0.545 ± 0.741	-0.011 ± 1.106	0.756 ± 0.839	0.171 ± 0.689
social_support	-0.902 ± 1.389	0.292 ± 0.685	-0.337 ± 0.766	0.318 ± 0.840	0.475 ± 0.844	0.279 ± 0.637
healthy_life_expectancy_at_birth	0.231 ± 1.710	-0.111 ± 0.000	-0.112 ± 0.001	-0.111 ± 0.001	-0.110 ± 0.000	0.308 ± 2.010
freedom_to_make_life_choices	-0.863 ± 1.308	0.023 ± 0.546	-0.212 ± 0.848	0.389 ± 0.306	0.445 ± 0.864	0.412 ± 0.833
generosity	-0.575 ± 0.691	-0.876 ± 0.608	-0.047 ± 0.640	-0.366 ± 0.873	0.514 ± 0.957	0.818 ± 1.201
perceptions_of_corruption	0.641 ± 0.229	0.404 ± 0.782	0.058 ± 0.941	-0.378 ± 1.376	-0.639 ± 1.009	0.176 ± 0.919
positive_affect	0.033 ± 0.554	0.145 ± 0.376	-0.059 ± 0.906	-2.248 ± 0.804	0.403 ± 0.784	0.582 ± 0.351

Figure 8: K means with k = 6: average and standard deviation for each attribute per cluster

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

In this clustering configuration, the salient attributes per cluster (high or

low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): high perception of corruption.

  Other important attributes that are visible from the heatmap are: high access to electricity, low but variable employment to population ratio, high but highly variable proportion of seats held by women, low renewable energy consumption, low but highly variable social support, low but highly variable freedom to make life choices, low but highly variable generosity.

  High corruption and variable employment, low social support and renewable energy.

- Cluster 1 (orange): high population 15-64, high urban population, low renewable energy consumption, low generosity, high access to electricity.

  Other important attributes that are visible from the heatmap are: high amount of people using the internet, very high but very variable military expenditure, high urban population, low but variable vulnerable employment and generosity. high but very variable perception of corruption.

  High working-age and urban population, strong electricity, militarization and internet access, low generosity, variable corruption.

- Cluster 2 (green): low gdp per capita, low percentage of individuals using the internet, low industry value added per worker, high vulnerable employment.

  Other important attributes that are visible from the heatmap are: high but variable renewable energy consumption, low but variable access to electricity, low but variable population ages 15-64, low but variable urban population, low but variable happiness score.

  Economically weak, low digital and industrial activity, high vulnerable employment, mixed energy indicators.

- Cluster 3 (red): access to electricity and population using the internet on the higher side, very low positive affect.

  Other important attributes that are visible from the heatmap are: low merchandise exports, low industry value added per worker.

  Well-connected countries with low happiness, limited industry, and low merchandise exports.

- Cluster 4 (lilac): High access to electricity and amount of people using the internet. Low vulnerable employment.

  Other important attributes that are visible from the heatmap are: high but variable gdp per capita, high but very variable industry value

added per worker, high but very variable population aged 15.65, high urban population, high but variable happiness score.

High electricity and internet access, low vulnerable employment, high GDP and urbanization.

- Cluster 5 (brown): Low GDP per capita, high positive affect.

  Other important attributes that are visible from the heatmap are: low urban population, low industry value added per worker

  Poor, rural countries with strong positive emotions but weak industry.

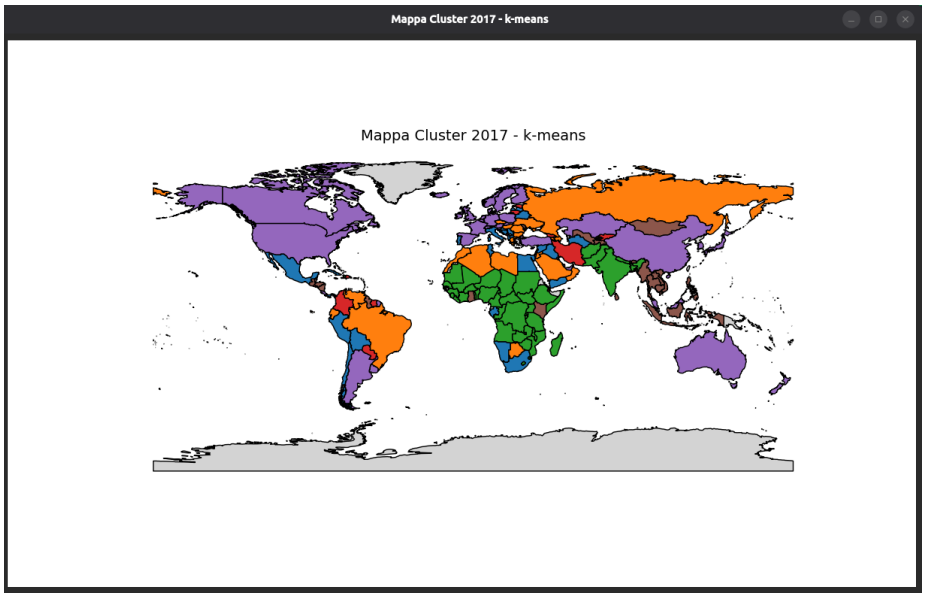

Figure 9: K means with k = 6: world plot

This data visualization is useful to see whether the division in clusters is completely random or not. In this situation we have:

- cluster 0 (blue): countries like Mexico, Italy, South Africa, Chile
- cluster 1 (orange): countries like Russia, Morocco, Brazil, Ukraine
- cluster 2 (green): countries like India, Sudan, Senegal, Pakistan
- cluster 3 (red): Countries like Iran, Colombia, Georgia, Paraguay
- cluster 4 (lilac): Countries like Canada, China, the Netherlands, New Zealand

- cluster 5 (brown): Countries like Mongolia, Thailand, Guatemala, Vietnam.

The conclusions are comparable with the former configuration. We can see how some clusters are similar, for example cluster 0 in both configurations, or cluster 4 of the 6 clusters configuration with cluster 1 of the 4 clusters configuration. Silhouette average scores are also comparable; the 4 clusters configuration has a more stable distribution of points in clusters.

4.2.2 K-medoids: CLARA

The algorithm CLARA (Clustering Large Applications) is an algorithm designed for efficiently clustering large datasets by sampling subsets of the data and applying the PAM (Partitioning Around Medoids) algorithm on each subset. How it works is that first a subset of the dataset is randomly sampled. The size of the subset is typically $min(40 + 2k, n)$, where k is the number of clusters and n is the total number of data points. Then, within the sampled subset, k initial medoids are selected randomly and iteratively refined within the subset by:

- Assigning each point in the subset to the nearest medoid based on a distance metric (Euclidean or cosine distance).
- Updating each medoid to the point in its cluster that minimizes the total distance to all other points in the cluster.
- Repeating until medoids converge or a maximum number of iterations is reached.

Once the medoids are determined from the subset, assign all points in the full dataset to the nearest medoid and compute the total cost (sum of distances from points to their medoids). The above process is repeated for several random subsets. The medoids and cluster assignments with the lowest total cost across all subsets are selected as the final clustering.

The Clara clustering algorithm was chosen for its ability to handle large datasets and to see whether it could also tackle high dimensional datasets as well, and because it's one algorithm where we can easily see the evolution in time of the very same clusters that we will look for in the 2017 dataset.

Elbow plot Same as with k-means, elbow method was tried with both cosine and euclidean distance and the plot showed a more regular elbow with cosine distance, as expected because of the dimension not giving reliable euclidean distance results.

The elbow plot shows 4 and 6 to be possible values of k. the algorithm is then run with k = 4 and cosine distance as the distance metric, and later on with k = 6 and cosine distance as the distance metric.

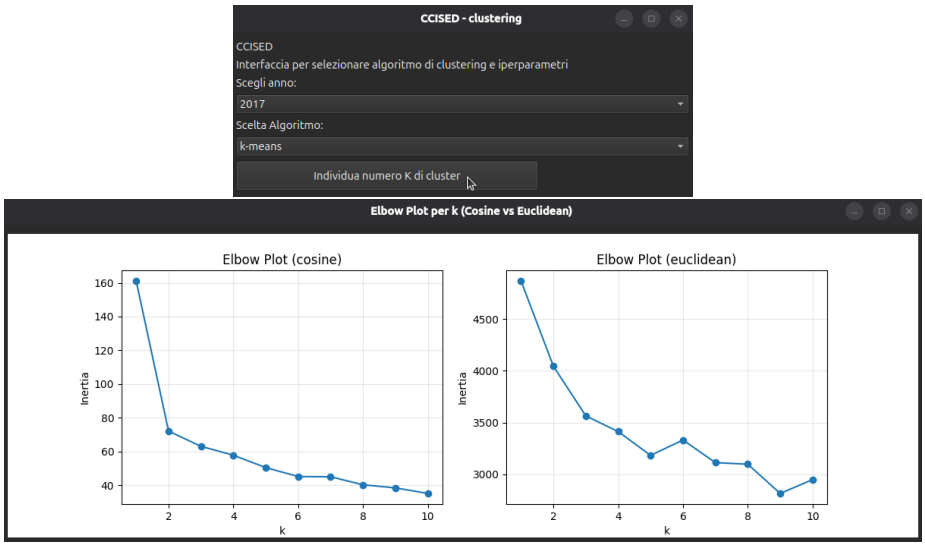

Figure 10: Elbow Plot: Clara

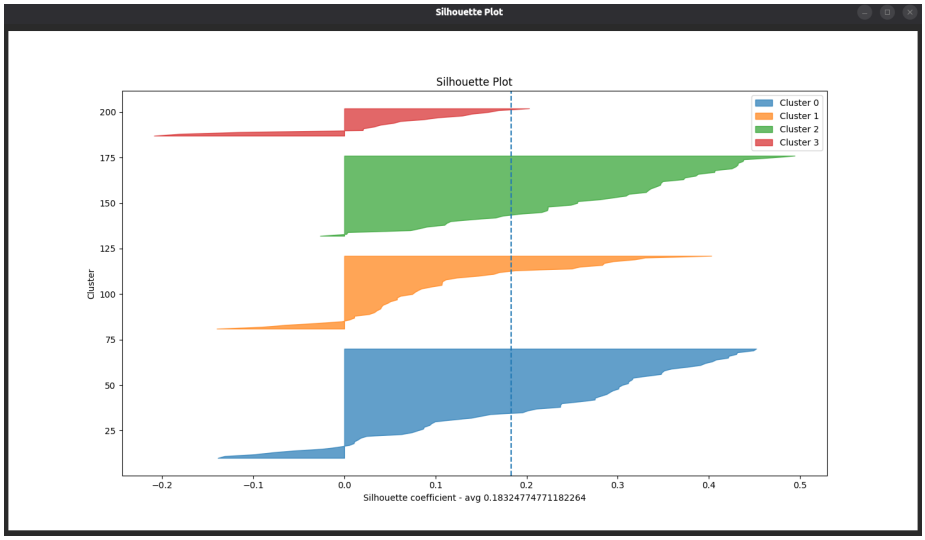

Figure 11: Clara with k = 4: silhouette

Clara with k = 4 The silhouette plot shows an average of 0,18. All clusters have an undesired amount of points which appear to be misclassified.

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping, which we mainly see between

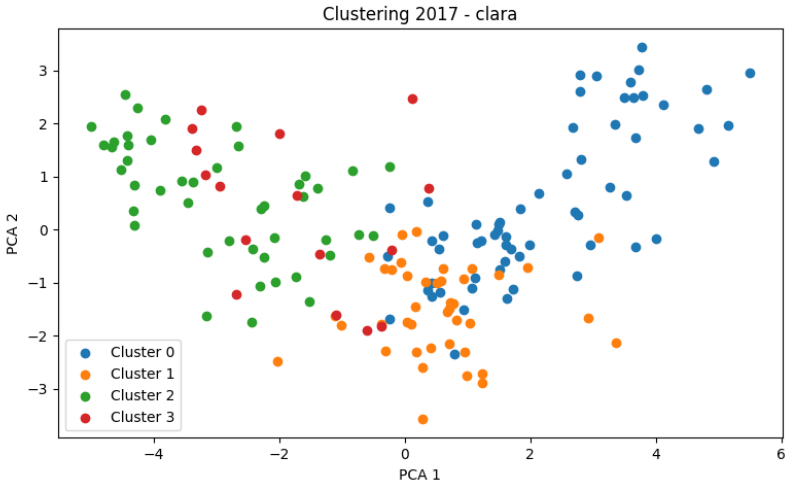

Figure 12: Clara with k = 4: 2-dim PCA

clusters 1, 2 and 3.

Statistiche cluster				
	0	1	2	3
Access to electricity (% of population)	0.599 ± 0.082	0.490 ± 0.284	-1.129 ± 1.111	-0.401 ± 0.968
Compulsory education, duration (years)	0.189 ± 0.827	0.502 ± 1.009	-0.443 ± 0.921	-0.769 ± 0.972
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	0.199 ± 0.716	-0.554 ± 0.989	0.414 ± 1.058	-0.507 ± 1.097
GDP per capita, PPP (current international $)	0.873 ± 1.087	-0.169 ± 0.470	-0.762 ± 0.173	-0.730 ± 0.199
Individuals using the Internet (% of population)	0.877 ± 0.568	0.224 ± 0.546	-1.117 ± 0.504	-0.783 ± 0.665
Industry (including construction), value added per worker (constant 2015 US$)	0.664 ± 1.325	-0.159 ± 0.489	-0.564 ± 0.202	-0.518 ± 0.200
Inflation, GDP deflator (annual %)	-0.107 ± 0.077	0.235 ± 1.968	-0.037 ± 0.357	-0.083 ± 0.044
Land area (sq. km)	0.155 ± 1.504	-0.011 ± 0.704	-0.112 ± 0.325	-0.231 ± 0.244
Merchandise exports by the reporting economy (current US$)	0.494 ± 1.500	-0.223 ± 0.290	-0.341 ± 0.164	-0.330 ± 0.154
Military expenditure (% of GDP)	-0.214 ± 0.695	0.609 ± 1.478	-0.169 ± 0.689	-0.250 ± 0.654
Net migration	0.338 ± 0.984	-0.156 ± 0.918	-0.182 ± 1.015	-0.376 ± 1.005
Population ages 15-64 (% of total population)	0.650 ± 0.687	0.303 ± 0.635	-1.052 ± 0.781	-0.311 ± 0.889
Proportion of seats held by women in national parliaments (%)	0.279 ± 0.917	-0.138 ± 0.999	-0.158 ± 1.034	-0.260 ± 1.118
Renewable energy consumption (% of total final energy consumption)	-0.553 ± 0.588	-0.585 ± 0.603	1.045 ± 0.764	0.687 ± 0.952
Urban population (% of total population)	0.595 ± 0.776	0.488 ± 0.673	-1.033 ± 0.616	-0.617 ± 0.713
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.815 ± 0.438	-0.264 ± 0.597	1.127 ± 0.740	0.631 ± 0.691
happiness_score	0.580 ± 0.814	-0.217 ± 1.061	-0.395 ± 0.797	-0.593 ± 1.020
social_support	0.280 ± 1.063	-0.193 ± 1.139	-0.205 ± 0.788	-0.039 ± 0.744
healthy_life_expectancy_at_birth	0.048 ± 1.234	0.098 ± 1.335	-0.112 ± 0.001	-0.111 ± 0.001
freedom_to_make_life_choices	0.380 ± 0.858	-0.387 ± 1.020	-0.255 ± 1.022	0.445 ± 0.417
generosity	0.204 ± 0.981	-0.547 ± 0.818	-0.033 ± 0.664	0.744 ± 1.552
perceptions_of_corruption	-0.321 ± 1.084	0.496 ± 0.574	0.329 ± 0.654	-0.991 ± 1.274
positive_affect	0.135 ± 1.045	-0.149 ± 0.843	0.304 ± 0.579	-1.027 ± 1.457

Figure 13: Clara with k = 4: average and standard deviation for each attribute per cluster

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the

mean value.

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

In this clustering configuration, the salient attributes per cluster (high or low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): High access to electricity, low vulnerable employment.

  Other important attributes that are visible from the heatmap are: high amount of individuals using the internet, low renewable energy consumption.

  High digital and energy access, low employment vulnerability, minimal renewable energy adoption.

- Cluster 1 (oranje): High access to electricity.

  Other important attributes that are visible from the heatmap are: low but highly variable employment to population ratio, high but variable percentage of people using the internet, high but variable perception of corruption, low but variable generosity.

  High energy access but with fluctuating employment, digitalization, corruption, and social support.

- Cluster 2 (green): Low GDP per capita, low percentage of people using the internet, low industry value added per worker, low urban population.

  Other important attributes that are visible from the heatmap are: low but variable access to electricity, low population 15-64, high but variable renewable energy consumption and vulnerable employment. low but variable happiness score, high but variable positive affect and perception of corruption.

  Low income, limited digitalization and urbanization, with high vulnerable employment and variable positive affect.

- Cluster 3 (red): Low GDP per capita, low industry value added per worker.

  Other important attributes that are visible from the heatmap are: low but very variable access to electricity, low but very variable compulsory education duration, low amount of people using the internet, low merchandise exports, high byt variable renewable energy consumption and vulnerable employment, low but variable urban population. high but variable generosity, low but extremely variable perceptions of corruption and positive affect.

Weak economy and exports, low urbanization, high vulnerable employment, unstable social indicators.

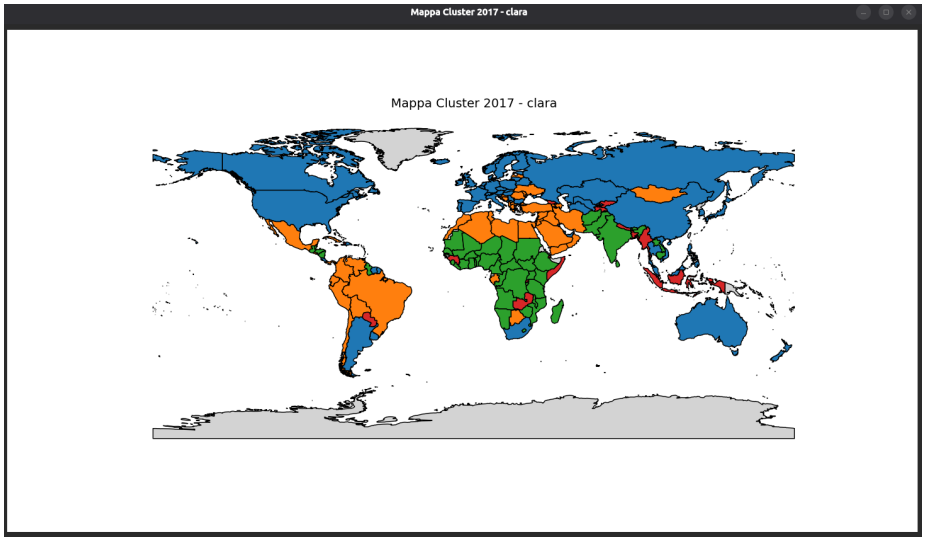

Figure 14: K means with k = 4: world plot

This data visualization is useful to see whether the division in clusters is completely random or not. In this situation we have:

- cluster 0 (blue): countries like Italy, China, USA, Spain, Russia
- cluster 1 (orange): countries like Mexico, Brazil, Egypt, Ukraine, Turkey
- cluster 2 (green): countries like India, Senegal, Cameroon, Ghana, Vietnam
- cluster 3 (red): Countries like Bangladesh, Gambia, Haiti, Indonesia, Paraguay

Clara with k = 6 The silhouette plot shows an average of 0,226; higher than the k = 4 case. Almost all clusters still have an undesired amount of points which appear to be misclassified.

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping. We see more than in the previous case, for all attributes apart from attribute 2.

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the mean value.

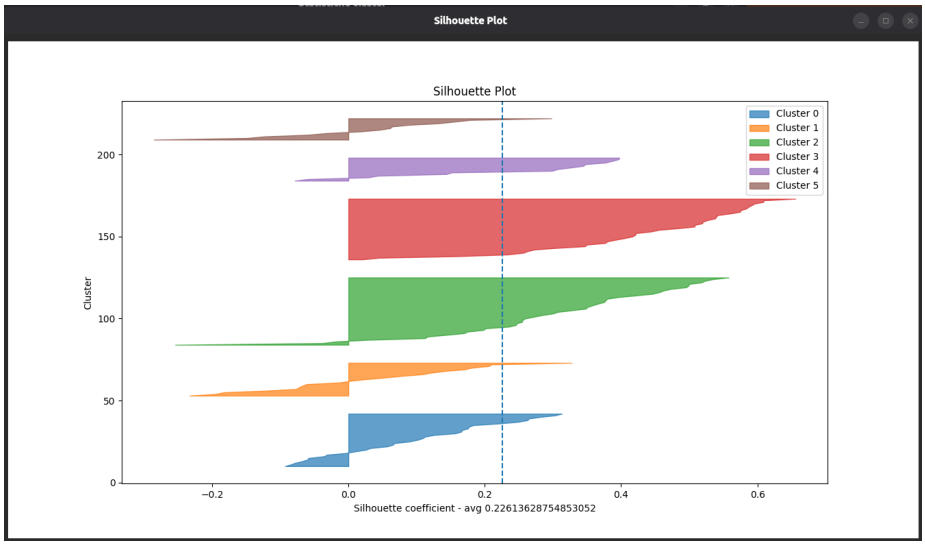

Figure 15: Clara with k = 6: silhouette

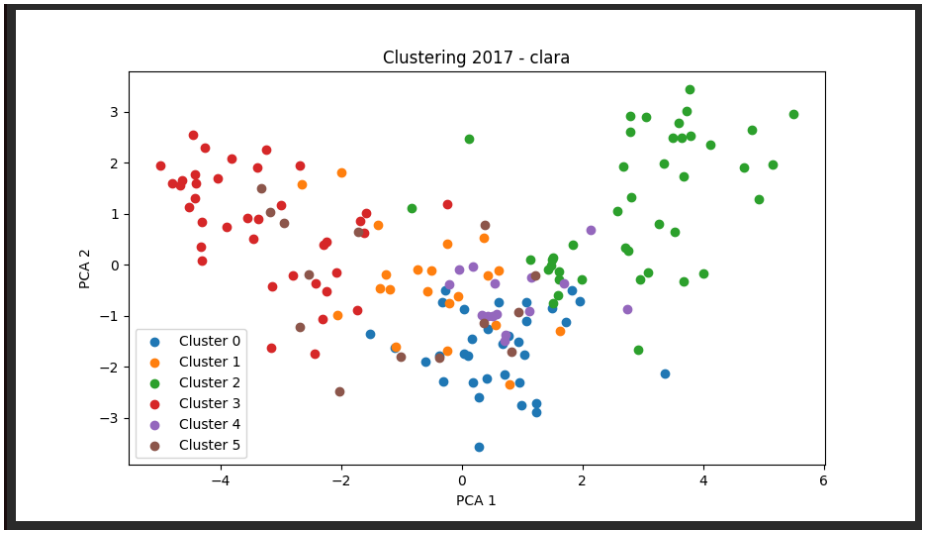

Figure 16: Clara with k = 6: 2-dim PCA

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

	0	1	2	3	4	5
Access to electricity (% of population)	0.489 ± 0.360	0.308 ± 0.493	0.602 ± 0.088	-1.414 ± 1.025	0.540 ± 0.164	-0.203 ± 0.834
Compulsory education, duration (years)	0.219 ± 1.017	0.001 ± 0.981	0.196 ± 0.876	-0.513 ± 0.927	0.521 ± 0.910	-0.274 ± 1.192
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	-0.644 ± 0.776	-0.077 ± 0.868	0.431 ± 0.624	0.491 ± 1.070	-0.063 ± 0.814	-0.926 ± 1.216
GDP per capita, PPP (current international $)	-0.115 ± 0.539	-0.451 ± 0.301	1.265 ± 1.064	-0.814 ± 0.139	-0.017 ± 0.416	-0.593 ± 0.301
Individuals using the Internet (% of population)	0.324 ± 0.490	-0.293 ± 0.689	1.082 ± 0.510	-1.240 ± 0.440	0.330 ± 0.404	-0.563 ± 0.770
Industry (including construction), value added per worker (constant 2015 US$)	-0.150 ± 0.485	-0.434 ± 0.201	1.109 ± 1.386	-0.597 ± 0.197	-0.230 ± 0.311	-0.431 ± 0.224
Inflation, GDP deflator (annual %)	0.287 ± 2.195	-0.093 ± 0.087	-0.114 ± 0.057	-0.023 ± 0.388	-0.090 ± 0.085	-0.031 ± 0.131
Land area (sq. km)	-0.059 ± 0.766	-0.250 ± 0.183	0.061 ± 1.146	-0.082 ± 0.339	0.714 ± 2.324	-0.191 ± 0.243
Merchandise exports by the reporting economy (current US$)	-0.200 ± 0.287	-0.274 ± 0.256	0.516 ± 1.286	-0.336 ± 0.178	0.573 ± 2.170	-0.343 ± 0.070
Military expenditure (% of GDP)	0.509 ± 1.345	-0.143 ± 0.619	-0.003 ± 1.173	-0.183 ± 0.695	-0.148 ± 0.733	-0.297 ± 0.602
Net migration	-0.217 ± 0.955	-0.009 ± 0.175	0.412 ± 1.125	-0.207 ± 1.104	0.184 ± 0.613	-0.344 ± 1.222
Population ages 15-64 (% of total population)	0.325 ± 0.555	0.338 ± 0.575	0.662 ± 0.817	-1.299 ± 0.599	0.535 ± 0.398	-0.321 ± 0.900
Proportion of seats held by women in national parliaments (%)	-0.062 ± 0.744	-0.576 ± 0.708	0.359 ± 0.995	-0.073 ± 1.101	0.698 ± 1.110	-0.606 ± 0.804
Renewable energy consumption (% of total final energy consumption)	-0.546 ± 0.618	-0.029 ± 0.812	-0.503 ± 0.696	1.186 ± 0.737	-0.567 ± 0.302	0.247 ± 1.165
Urban population (% of total population)	0.438 ± 0.636	-0.796 ± 0.697	0.907 ± 0.654	-1.061 ± 0.616	0.331 ± 0.566	-0.037 ± 0.719
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.406 ± 0.492	0.061 ± 0.628	-0.926 ± 0.374	1.346 ± 0.579	-0.461 ± 0.701	0.502 ± 0.688
happiness_score	-0.402 ± 1.166	-0.053 ± 0.910	0.845 ± 0.632	-0.457 ± 0.732	-0.091 ± 1.018	-0.224 ± 0.810
social_support	0.323 ± 0.569	0.477 ± 0.239	0.482 ± 0.825	-0.359 ± 0.790	-1.844 ± 1.171	-0.021 ± 0.781
healthy_life_expectancy_at_birth	0.149 ± 1.488	0.348 ± 2.103	-0.110 ± 0.000	-0.112 ± 0.001	-0.110 ± 0.001	-0.111 ± 0.001
freedom_to_make_life_choices	-0.409 ± 1.071	0.023 ± 1.089	0.484 ± 0.839	-0.188 ± 0.884	-0.022 ± 1.039	0.223 ± 0.602
generosity	-0.835 ± 0.642	0.627 ± 1.204	0.561 ± 1.047	-0.097 ± 0.485	-0.614 ± 0.635	0.298 ± 0.901
perceptions_of_corruption	0.487 ± 0.629	0.476 ± 0.413	-0.734 ± 1.046	0.179 ± 0.879	0.441 ± 0.823	-0.642 ± 1.232
positive_affect	-0.205 ± 0.888	0.253 ± 0.767	0.321 ± 0.992	0.208 ± 0.578	0.292 ± 0.448	-1.781 ± 1.130

Figure 17: Clara with k = 6: average and standard deviation for each attribute per cluster

In this clustering configuration, the salient attributes per cluster (high or low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): High access to electricity.

  Other important attributes that are visible from the heatmap are: low employment to population ratio, low vulnerable employment, low generosity, high eprception of corruption, high but variable urban population, low but variable renewable energy consumption.

  Strong energy access, low job participation, limited generosity, high corruption concerns.

- Cluster 1 (oranje): low gdp per capita, low industry value added per worker, low urban population, high social support, high perceptions of corruption.

  Other important attributes that are visible from the heatmap are: low but variable proportion of seats held by women in national parliaments, high but very variable generosity, high but variable access to electricity.

  Low income and industrial output, rural, high social support and generosity, widespread corruption concerns.

- Cluster 2 (green): High access to electricity and amount of people using the internet, low vulnerable employment.

  Other important attributes that are visible from the heatmap are: High but variable gdp per capita, high but highly variable industry value added per worker, high but variable population 15-64, high happiness score, high positive affect score.

Economically strong and connected, low vulnerable jobs, high life satisfaction.

- Cluster 3 (red): Low gdp per capita, low percentage of individual using the internet and industry value added per worker.

  Other important attributes that are visible from the heatmap are: low but variable access to electricity, high but variable employment to population ratio, low population ages 15-64, high renewable energy consumption, low urban population, low but variable happiness score, high but variable positive affect.

  Economically weak and less connected, low urbanization, high renewables, mixed well-being.

- Cluster 4 (lilac): high access to electricity, high population 15-64, low renewable energy consumption.

  Other important attributes that are visible from the heatmap are: high but variable urban population, very low social support, low generosity

  High energy access and urbanization, limited renewables and fragile social cohesion.

- Cluster 5 (brown): Low gdp per capita.

  Other important attributes that are visible from the heatmap are: low but extremely variable employment ratio, low but variable percentage of people using the internet, low merchandise exports, very low positive affect

  Economically weak, low exports and digital access, fragile social mood.

This data visualization is useful to see whether the division in clusters is completely random or not. In this situation we have:

- cluster 0 (blue): countries like Spain, Brazil, Egypt, Turkey, Ukraine
- cluster 1 (orange): countries like Mongolia, Kenya, Hungary, Honduras, Thailand
- cluster 2 (green): countries like Australia, France, Argentina, the US, Sweden
- cluster 3 (red): Countries like Madagascar, Tanzania, India, Pakistan, Senegal
- cluster 4 (lilac): Countries like Italy, Russia, China, Mexico, Chile, Peru
- cluster 5 (brown): Countries like Iran, Nepal, Paraguay, Syria, Somalia

The clustering does not seem totally random but also not totally coherent, especially with cluster 0 and 4.

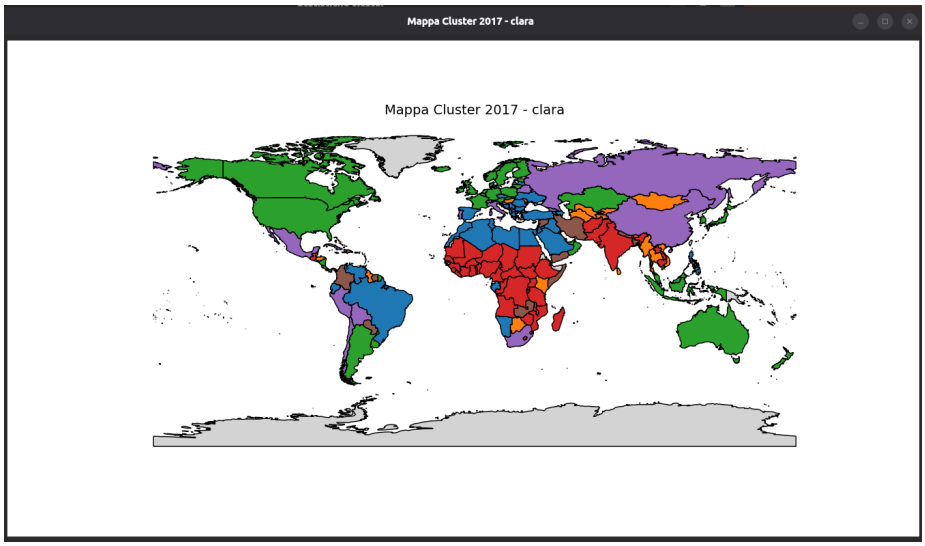

Figure 18: Clara with k = 6: world plot

4.2.3 DBSCAN

The algorithm DBSCAN relies on two key parameters:

- ε (eps): the radius that defines the neighborhood of a point
- `min_samples`: the minimum number of points required to form a dense region

A point is defined as a *core point* if at least `min_samples` points exist within its ε-neighborhood. Points that are reachable from a core point are assigned to the same cluster, while points that do not satisfy this condition are labeled as noise.

The implementation supports two distance metrics:

- **Euclidean distance**, suitable for spatial data
- **Cosine distance**, useful for high-dimensional vectors

The distance function is selected dynamically based on the `distance_metric` parameter. For each point p, the algorithm computes its ε-neighborhood:

$$N_\varepsilon(p) = \{q \mid d(p, q) \leq \varepsilon\}$$

This operation is performed by scanning all points in the dataset and selecting those within the specified radius. If a point is identified as a core point, a new cluster is created and expanded using a breadth-first search (BFS) strategy. All density-reachable points are iteratively added to the cluster. Points initially marked as noise may later be assigned to a cluster if they are found to be density-reachable. The algorithm outputs a label for each data point:

- Positive integers indicate cluster membership
- A label of -1 represents noise or outliers

Hyperparameters choice The hyperparameter choice was made as follows:

- `min_samples`: the rule of thumb is to choose this hyperparameter as min-samples $\geq M+1 where M is the number of features of the data. In this case, M = 24 so min\_samples = 25$;

- ε (eps): chosen by heuristic method consisting in plotting a k-distance graph, where k = MinPts. For each point compute the distance to its k-th nearest neighbour, sort the points based on their distances in descending order, and then identify the "elbow". This plot was tried for both distance metrics and at the end, a value between 0,50 and 0,65 was found to be promising. The end value was 0,62

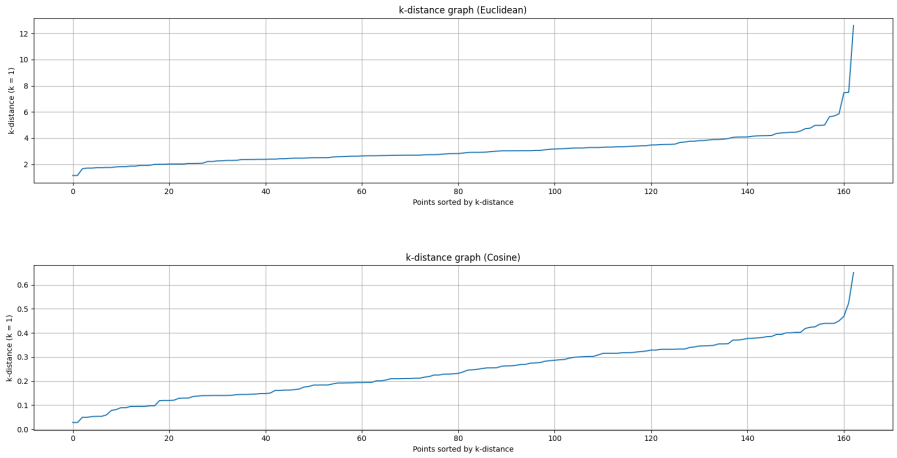

Figure 19: Eps plot

DBSCAN with eps = 0.62 and min$_s amples = 25$ Now we look at what happens when we try to cluster with the found hyperparameters.

The silhouette average score is higher than what we previously found.

Found 2 clusters and a relatively small amount of outliers.

This configuration of hyperparameters gave us 2 clusters and 15 outlier points.

- Cluster 1: Low gdp per capita, low amount of individuals using the internet, low industry value added per worker, low population ages 15-64, high vulnerable employment. Countries such as: Cameroon, India, Yemen

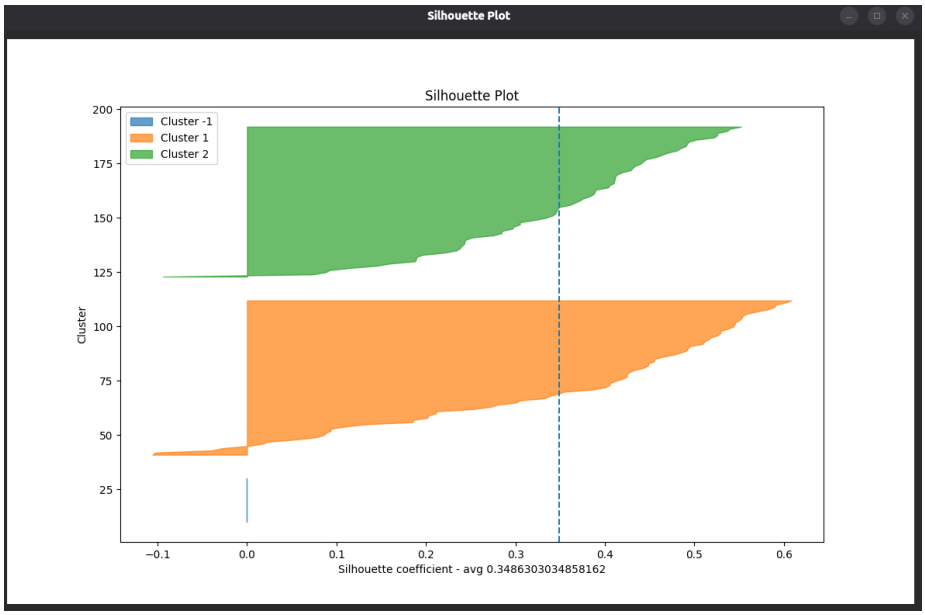

Figure 20: DBSCAN: silhouette

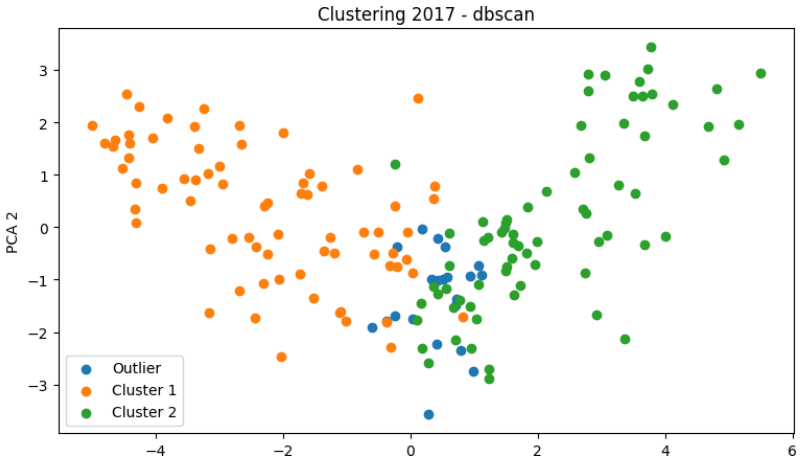

Figure 21: DBSCAN: PCA

- CLuster 2: high access to electricity, high amount of individuals using the internet, high amount of pop 15-64, low vulnerable employment. Countries such as: Kuwait, Singapore, Austria, China
- : outliers were countries such as Palestine, Armenia, Venezuela.

	1	2
Access to electricity (% of population)	-0.747 ± 1.131	0.606 ± 0.045
Compulsory education, duration (years)	-0.372 ± 0.938	0.314 ± 0.893
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	0.072 ± 1.194	0.054 ± 0.783
GDP per capita, PPP (current international $)	-0.703 ± 0.216	0.808 ± 1.054
Individuals using the Internet (% of population)	-0.914 ± 0.611	0.861 ± 0.533
Industry (including construction), value added per worker (constant 2015 US$)	-0.512 ± 0.270	0.615 ± 1.261
Inflation, GDP deflator (annual %)	-0.042 ± 0.290	-0.102 ± 0.079
Land area (sq. km)	-0.126 ± 0.302	0.176 ± 1.483
Merchandise exports by the reporting economy (current US$)	-0.313 ± 0.199	0.398 ± 1.422
Military expenditure (% of GDP)	-0.164 ± 0.659	0.198 ± 1.299
Net migration	-0.216 ± 0.946	0.230 ± 1.063
Population ages 15-64 (% of total population)	-0.718 ± 0.906	0.583 ± 0.684
Proportion of seats held by women in national parliaments (%)	-0.175 ± 1.097	0.133 ± 0.905
Renewable energy consumption (% of total final energy consumption)	0.724 ± 0.975	-0.548 ± 0.619
Urban population (% of total population)	-0.795 ± 0.756	0.739 ± 0.620
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	0.892 ± 0.764	-0.801 ± 0.431
happiness_score	-0.431 ± 0.847	0.530 ± 0.866
social_support	-0.104 ± 0.736	0.367 ± 0.882
healthy_life_expectancy_at_birth	-0.112 ± 0.001	-0.110 ± 0.000
freedom_to_make_life_choices	-0.109 ± 0.967	0.220 ± 0.937
generosity	0.155 ± 1.008	-0.052 ± 1.054
perceptions_of_corruption	0.147 ± 0.880	-0.304 ± 1.113
positive_affect	-0.024 ± 0.995	0.046 ± 1.046

Figure 22: DBSCAN: means and standard deviations among clusters

Some final considerations: this algorithm faces challenges when dealing with such a high-dimensional dataset, even though the data does have clsutering tendency, as we already demonstrated. Nevertheless, it was able to identify two clusters. While it struggles to uncover more fine-grained cluster structures, this limitation might be largely due to the nature of the data itself. Importantly, it achieved the highest average silhouette score observed so far, indicating that it provides the most promising clustering structure to date. Unfortunately, DBSCAN cannot directly search for the same clusters across different datasets due to its density-based nature. However, if DBSCAN proves to be the most promising algorithm for this dataset, it remains possible to apply it independently to each subsequent and precedent year, allowing us to investigate how the clustering structure evolves year by year.

4.2.4 Agglomerative Hierarchical

The Hierarchical clustering algorithm was chosen for its ability to capture nested structures in the data and because it does not require specifying initial centroids, unlike K-Means. It provides a tree-like structure (dendrogram) that allows us to observe cluster formation at different levels.

How it works: Hierarchical clustering starts with each point as a separate cluster. At each iteration, the two closest clusters are merged into a single cluster according to a chosen linkage criterion. This process continues until the

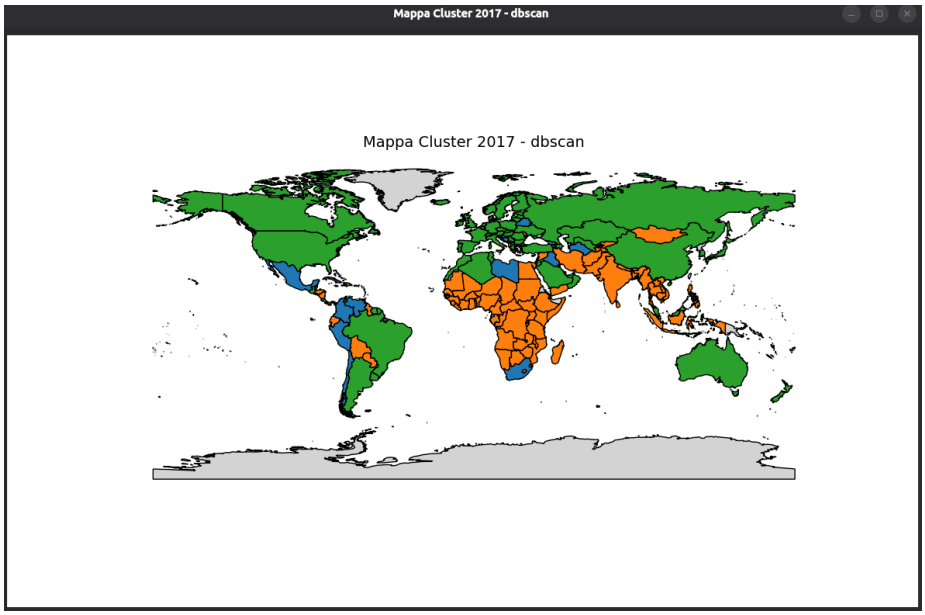

Figure 23: DBSCAN: world plot

desired number of clusters K is reached. The distance between clusters can be calculated in different ways:

- Single linkage: distance between the two closest points of the clusters.
- Complete linkage: distance between the two farthest points of the clusters.
- Average linkage: average distance between all points across the clusters.

Distances between points can be computed using either Euclidean or Cosine metrics. If Cosine distance is used, the dataset is first normalized using L2 normalization, which makes the cosine measure more meaningful. The algorithm maintains a distance matrix between all points and updates distances using a priority queue (min heap). After the merging process ends, each point is assigned a cluster label corresponding to the final clusters.

Dendrograms The first step was to build dendrograms to evaluate which linkage method performs best and to estimate the optimal number of clusters k. Ideally, we aim to obtain a balanced cluster structure.

We can clearly see how unbalanced the single linkage dendrogram is. This is due to the nature of single linkage, which link clusters based on the distance between two closest points of two clusters.

To select the optimal number of clusters in hierarchical clustering, we follow a quantitative procedure based on the dendrogram structure:

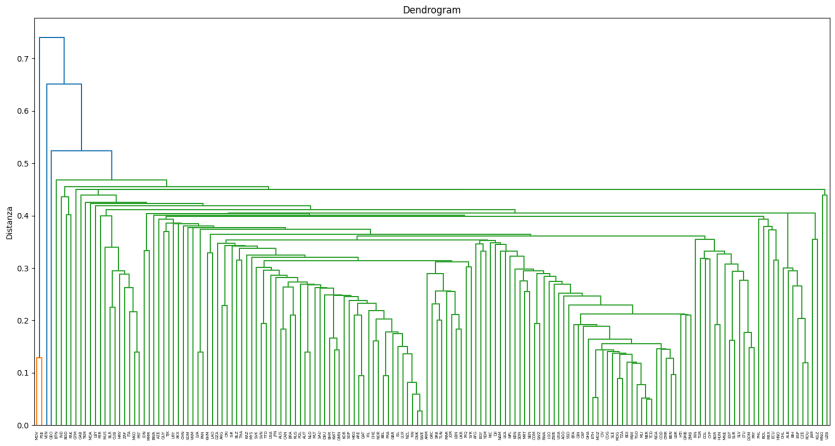

Figure 24: Dendrogram obtained selecting n. of clusters = 2 and linkage = single.

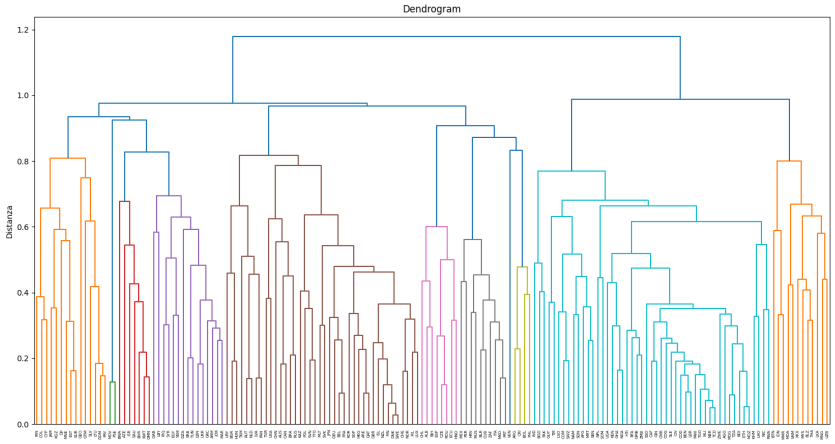

Figure 25: Dendrogram obtained selecting n. of clusters = 2 and linkage = average.

1. Let $Z \in R^{(n-1)\times 4}$ be the linkage matrix, where the third column $Z[:,2]$ contains the distances at which clusters are merged. Compute the consecutive differences:

$$\Delta_i = Z[i+1,2] - Z[i,2], \quad i = 1,\ldots,n-2$$

The index of the maximum difference indicates the largest separation be-

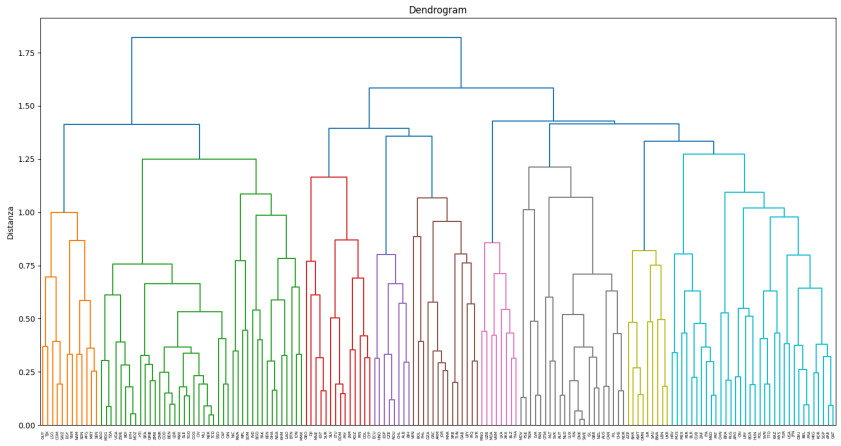

Figure 26: Dendrogram obtained selecting n. of clusters = 2 and linkage = complete.

tween clusters:

$$i_{\max} = \arg\max_i \Delta_i$$

The initial estimate of clusters is

$$k_{\text{gap}} = n - (i_{\max} + 1)$$

2. Apply a minimum cluster size constraint. Let $s_{\min}$ be the minimum number of points allowed per cluster. If any candidate k produces clusters smaller than $s_{\min}$, decrement k until all clusters satisfy the size constraint. The final number of clusters is denoted k_{opt}.

In both linkage complete and average this procedure showed 2 as the optimum number of clusters.

2 clusters, complete linkage The silhouette plot shows an average of 0,305 which is still a value on the lower end but higher than what we have previously seen. The orange cluster, cluster 1, has a higher amount of points which appear to be misclassified and drive down the average.

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping, but in this case the clusters show themselves to be not distanced but not overlapping.

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the mean value.

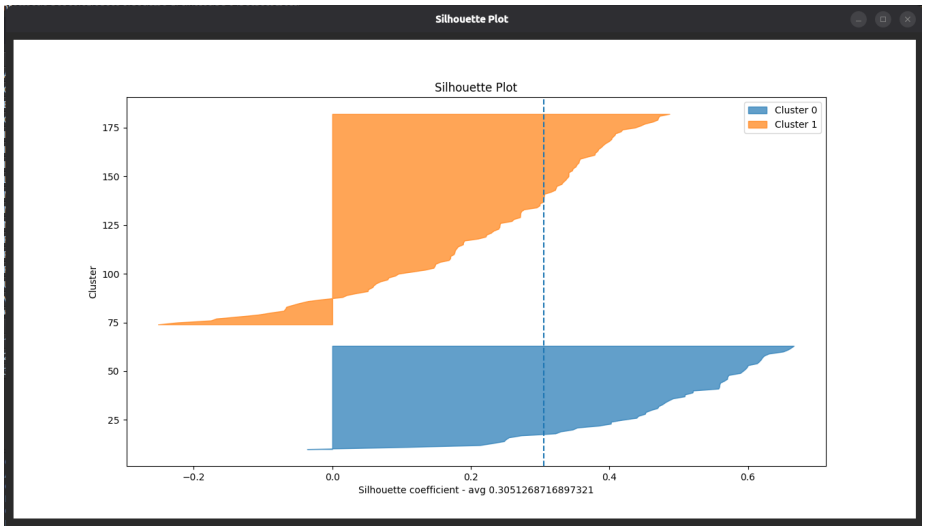

Figure 27: Complete linkage: silhouette plot and average.

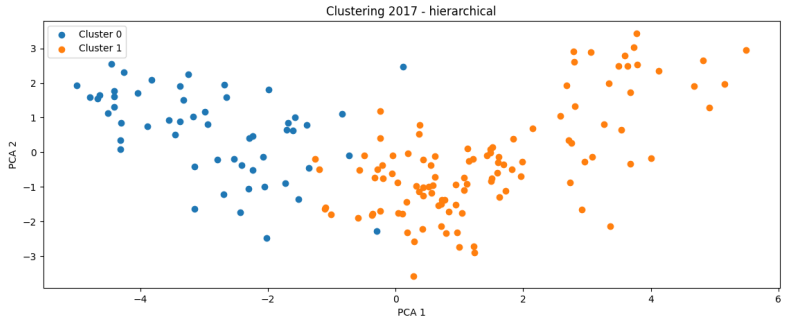

Figure 28: Complete linkage: 2d PCA visualization.

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

In this clustering configuration, the salient attributes per cluster (high or low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): low GDP per capita, low industry value added per worker, low percentage of individuals using the internet, low urban population.

	0	1
Access to electricity (% of population)	-1.105 ± 1.065	0.542 ± 0.229
Compulsory education, duration (years)	-0.637 ± 0.818	0.315 ± 0.942
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	0.142 ± 1.209	-0.071 ± 0.887
GDP per capita, PPP (current international $)	-0.776 ± 0.167	0.388 ± 1.025
Individuals using the Internet (% of population)	-1.149 ± 0.462	0.568 ± 0.654
Industry (including construction), value added per worker (constant 2015 US$)	-0.565 ± 0.194	0.283 ± 1.122
Inflation, GDP deflator (annual %)	-0.032 ± 0.328	0.017 ± 1.210
Land area (sq. km)	-0.105 ± 0.316	0.055 ± 1.208
Merchandise exports by the reporting economy (current US$)	-0.333 ± 0.170	0.168 ± 1.191
Military expenditure (% of GDP)	-0.189 ± 0.681	0.096 ± 1.124
Net migration	-0.262 ± 1.069	0.130 ± 0.952
Population ages 15-64 (% of total population)	-1.025 ± 0.785	0.506 ± 0.658
Proportion of seats held by women in national parliaments (%)	-0.133 ± 1.106	0.067 ± 0.951
Renewable energy consumption (% of total final energy consumption)	1.008 ± 0.866	-0.497 ± 0.626
Urban population (% of total population)	-1.007 ± 0.612	0.499 ± 0.759
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	1.120 ± 0.714	-0.553 ± 0.576
happiness_score	-0.415 ± 0.720	0.199 ± 1.064
social_support	-0.254 ± 0.774	0.120 ± 1.083
healthy_life_expectancy_at_birth	-0.112 ± 0.001	0.056 ± 1.228
freedom_to_make_life_choices	-0.190 ± 0.972	0.121 ± 0.971
generosity	0.178 ± 1.055	-0.084 ± 0.974
perceptions_of_corruption	0.081 ± 0.913	-0.043 ± 1.050
positive_affect	0.053 ± 0.857	-0.032 ± 1.073

Figure 29: Complete linkage: means and standard deviation of attributes cluster-wise.

Other important attributes that are visible from the heatmap are: low but variable access to electricity, low but variable duration of compulsory education, low merchandise exports, low but variable percentage of population 15-64, high but variable renewable energy consumption, high but variable vulnerable employment, low but variable happiness score and social support.

Economically weak, low digital and industrial activity, high employment vulnerability, limited social support.

Countries like: India, Tanzania, Congo, Egypt, Bangladesh

- Cluster 1 (oranje): High access to electricity.

  Other important attributes that are visible from the heatmap are: high but variable percentage of individuals using the internet, high but variable percentage of population ages 15-64, low amount of vulnerable employment.

  High electricity access, more digitally connected, generally low vulnerable employment.

  Countries like: Italy, Brazil, Russia, Australia, Canada.

This clustering is clearly not random, as it reflects really existing and noticeable underlying economic and infrastructural patterns among countries. However, the division is very coarse, as it primarily separates countries into two broad groups: one cluster (Cluster 0) that includes emerging economies or low-income countries with limited industrialization, low digital connectivity, and higher social and economic vulnerabilities, and a second cluster (Cluster 1) that groups together the remaining countries, which are more diverse and generally include higher-income, more urbanized, and better-connected nations. While this configuration captures some meaningful distinctions, it oversimplifies the global landscape and does not reflect the full spectrum of socio-economic differences between countries.

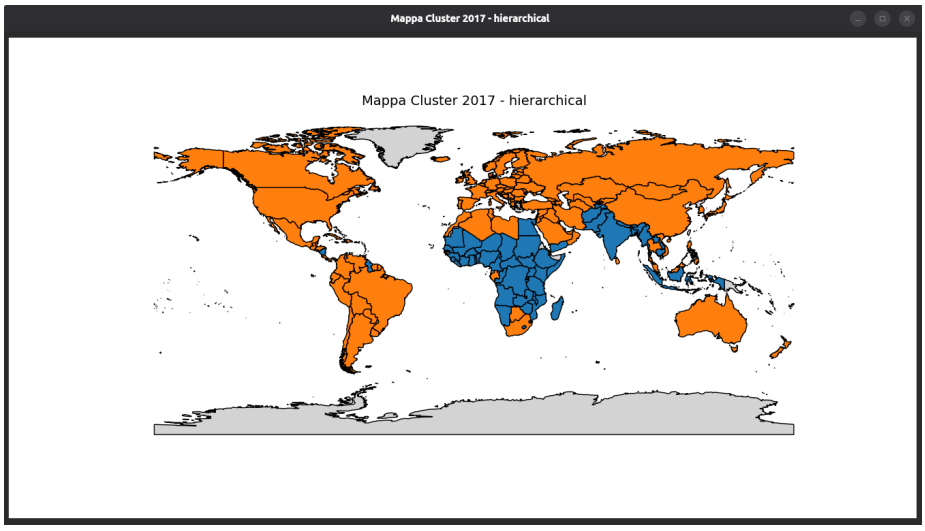

Figure 30: Complete linkage: world cluster plot.

2 clusters, average linkage The silhouette plot shows an average of 0,286 which is lower than the former clustering configuration. Both clusters have an amount of points which appear to be misclassified and drive down the average.

PCA Plot shows us, albeit visually and partially, whether there is some separation among the clusters. Given that this is a 2D reduction of a 25D dataset we can expect there to be some overlapping, but in this case the clusters show themselves to be only slightly overlapping.

This heatmap shows us the core of the clustering, that is, the differences in values between the attributes of the clusters and their dispersion around the mean value.

The salient attributes, which end up characterizing each cluster, are chosen by looking at each average and standard deviation. Given that the whole dataset is z-score normalized, the salient attributes are those that have an average close

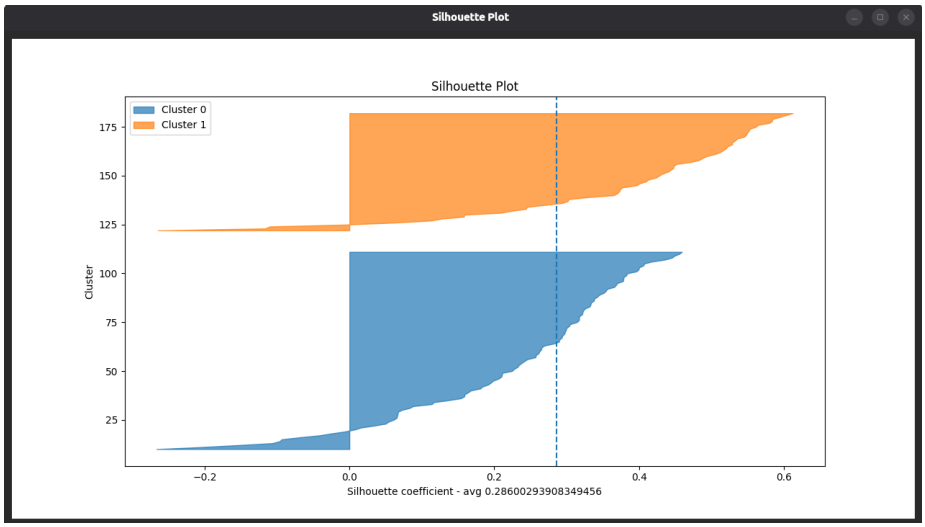

Figure 31: Average linkage: silhouette plot and average.

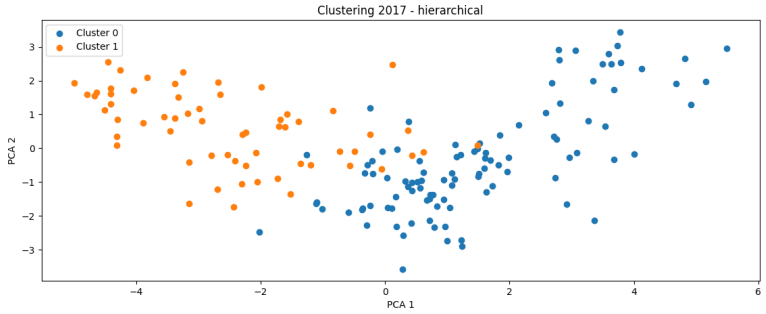

Figure 32: Average linkage: 2d PCA visualization.

or superior to 0.5, or close or inferior to -0.5, whilst having a standard deviation closer or inferior than 0.5 (mind that the standard deviation of the whole dataset has been normalized to 1).

In this clustering configuration, the salient attributes per cluster (high or low but compact - every attribute not listed is either close to the world average or high in variability) are:

- Cluster 0 (blue): High access to electricity.

  Other important attributes that are visible from the heatmap are: high but variable percentage of individuals using the internet, high but variable percentage of population ages 15-64, low amount of vulnerable employment.

	0	1
Access to electricity (% of population)	0.534 ± 0.245	-0.904 ± 1.145
Compulsory education, duration (years)	0.332 ± 0.939	-0.556 ± 0.864
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	-0.150 ± 0.926	0.251 ± 1.090
GDP per capita, PPP (current international $)	0.440 ± 1.037	-0.730 ± 0.226
Individuals using the Internet (% of population)	0.599 ± 0.641	-1.003 ± 0.641
Industry (including construction), value added per worker (constant 2015 US$)	0.334 ± 1.142	-0.553 ± 0.194
Inflation, GDP deflator (annual %)	0.026 ± 1.251	-0.042 ± 0.310
Land area (sq. km)	0.076 ± 1.245	-0.123 ± 0.310
Merchandise exports by the reporting economy (current US$)	0.182 ± 1.226	-0.299 ± 0.232
Military expenditure (% of GDP)	0.137 ± 1.146	-0.224 ± 0.662
Net migration	0.123 ± 0.981	-0.205 ± 1.022
Population ages 15-64 (% of total population)	0.472 ± 0.697	-0.793 ± 0.947
Proportion of seats held by women in national parliaments (%)	0.105 ± 0.963	-0.174 ± 1.060
Renewable energy consumption (% of total final energy consumption)	-0.515 ± 0.632	0.866 ± 0.920
Urban population (% of total population)	0.568 ± 0.707	-0.950 ± 0.653
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.607 ± 0.545	1.019 ± 0.743
happiness_score	0.165 ± 1.102	-0.289 ± 0.741
social_support	0.055 ± 1.139	-0.104 ± 0.725
healthy_life_expectancy_at_birth	0.068 ± 1.269	-0.112 ± 0.001
freedom_to_make_life_choices	0.083 ± 0.991	-0.090 ± 0.958
generosity	-0.154 ± 0.982	0.265 ± 0.997
perceptions_of_corruption	-0.086 ± 1.063	0.138 ± 0.893
positive_affect	-0.089 ± 1.098	0.138 ± 0.815

Figure 33: Average linkage: means and standard deviation of attributes cluster-wise.

High electricity access, more digitally connected, generally low vulnerable employment.

Countries like: Italy, Brazil, China, New Zealand, Canada.

- Cluster 1 (oranje): low gdp per capita, low industry value added per worker, low percentage of individuals using the internet, low urban population.

  Other important attributes that are visible from the heatmap are: low but variable access to electricity, low but variable duration of compulsory education, low merchandise exports, low but variable percentage of population 15-64, high but variable renewable energy consumption, high but variable vulnerable employment, low but variable happiness score and social support.

  Economically weak, low digital and industrial activity, high employment vulnerability, limited social support.

  Countries like: India, Tanzania, Congo, Egypt, Bangladesh

This clustering is very similar to the former one - the only difference being the order of the clusters is inverted - and just as the former one, it reflects really existing and noticeable underlying economic and infrastructural patterns among countries. However, the division is still very coarse.

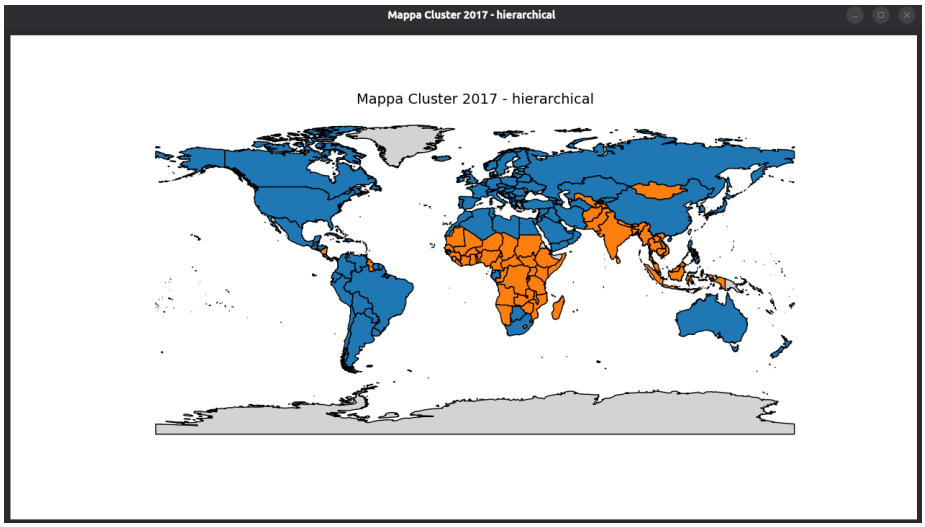

Figure 34: Average linkage: world cluster plot.

4.2.5 CLIQUE attempt

The CLIQUE (Clustering In QUEst) algorithm was included in the tool as an experimental approach to explore grid-based clustering in subspaces. Unlike K-Means, which partitions data in the full feature space, CLIQUE is designed to identify dense regions in arbitrary subspaces, which makes it theoretically attractive for high-dimensional datasets where relevant clusters may appear only across a subset of features.

How it works: CLIQUE first discretizes each dimension of the dataset into a fixed number of intervals, creating a multidimensional grid. Each point is assigned to a cell of this grid, and the algorithm identifies dense cells, i.e., those containing a minimum fraction of points. Starting from 1-dimensional subspaces, CLIQUE applies an apriori-like strategy to combine dense cells across multiple dimensions, generating candidate subspaces of higher dimension and identifying clusters as connected regions of dense cells. Points are assigned to clusters based on the cells they fall into, and outliers are labeled separately. To evaluate cluster quality, CLIQUE can compute silhouette scores based on the largest discovered subspace.

CLIQUE was implemented to test whether a grid-based subspace clustering algorithm could help uncover clusters that might exist only in particular combinations of indicators, potentially revealing structures that K-Means or other full-space algorithms could miss. The algorithm was fully implemented in Python and integrated into the tool with adjustable hyperparameters such as the number of grid intervals, the maximum subspace dimensionality, and the minimum fraction of points required to form a cluster.

However, the experiments showed that CLIQUE was not suitable for the

objectives of this project. In practice, it either labeled the majority of countries as outliers or produced clusters containing only very few countries at a time. This behavior makes the clustering results highly fragmented, non-reproducible across years, and unsuitable for deriving general insights about cross-country trends. Consequently, while CLIQUE is present in the tool for demonstration purposes and experimental use, its outputs were not considered for the main results.

4.3 Clusters found

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

For each clustering algorithm that yielded meaningful and interpretable clusters, the hyperparameter configuration producing the highest average silhouette score was selected as the reference setting. The specific configurations identified were: K-Means with K=4, CLARA with K=6, DBSCAN with ε=62 and min_samples=25, and Hierarchical clustering with 2 clusters using complete linkage. These settings correspond to the most structurally coherent clusters within the analyzed dataset, as quantified by silhouette-based cluster quality metrics.

5 Results

For algorithms with explicit cluster representatives, such as K-Means and CLARA, the centroids and medoids of the resulting clusters were preserved, to aid for the replication of the same cluster assignments across multiple years and making consistent longitudinal analysis and comparison possible.

For algorithms without fixed cluster centers, including DBSCAN and Hierarchical clustering, replication across years relied on re-executing the clustering procedure with the previously identified optimal hyperparameters. These methods do not provide explicit centroids or medoids, so maintaining the same ε, min_samples, or linkage configuration was the first step in ensuring that the structural characteristics of the clusters are comparable across different temporal snapshots.

5.1 Silhouette scores

Silhouette scores were recomputed for each year and for each algorithm to assess the validity of the clusterings. All values are low but positive, which is expected given the high dimensionality of the data and the generally low average silhouette scores observed in preliminary analysis. Most values fluctuate between 0.2 and 0.3, consistent with the initial results. The lowest scores were observed for Hierarchical clustering in 2006 and for DBSCAN in 2009.

Table 7: Silhouette Score per anno e algoritmo di clustering

Year	HIER	DBSCAN	KMEANS_4	CLARA_6
2005	0.319	0.220	0.268	0.214
2006	0.044	0.194	0.279	0.237
2007	0.266	0.192	0.248	0.200
2008	0.290	0.193	0.259	0.200
2009	0.201	0.009	0.251	0.197
2010	0.199	0.194	0.250	0.190
2011	0.074	0.182	0.243	0.183
2012	0.300	0.216	0.245	0.203
2013	0.244	0.215	0.235	0.191
2014	0.295	0.197	0.241	0.207
2015	0.290	0.202	0.246	0.178
2016	0.269	0.177	0.232	0.152
2017	0.305	0.199	0.255	0.226
2018	0.280	0.170	0.222	0.160
2019	0.268	0.213	0.237	0.183
2020	0.277	0.182	0.224	0.180
2021	0.250	0.209	0.230	0.164
2022	0.232	0.168	0.219	0.179

Mean ± SD: HIER = 0.245 ± 0.076, DBSCAN = 0.185 ± 0.056, KMEANS_4 = 0.244 ± 0.014, CLARA_6 = 0.191 ± 0.027

Given these results, which appear consistent with the preliminary analysis, it is reasonable to proceed with the year-by-year clustering analysis. We expect clusters obtained from CLARA are likely to be less robust, reflected in the generally lower silhouette scores. However, it is still informative to explore CLARA, since it is a k-medoids-based algorithm: when performing a "relative" analysis with z-score normalized data, each year's distribution is centered on that year, which may naturally lead to more fluctuating cluster assignments. This effect is amplified by the high dimensionality of the dataset, but the clustering can still reveal meaningful relative patterns across countries. The same reasoning can be done for DBSCAN, whose low mean silhouette score can though be at least in part be attributed to the anomalous value in the year 2009, and to its intrinsic difficulty in finding clusters with high dimensional data due to its nature.

5.2 K-means

Using K-Means, four clusters offering a compact and descriptive representation of countries across the world were identified. The following section provides detailed descriptions of these clusters.

- Cluster 0 (blue): **Good energy infrastructure, institutional weaknesses, high perceived corruption, and an unstable economy**
- Cluster 1 (orange): **Economically developed, digital, and urbanized, with low vulnerable employment.**
- Cluster 2 (green):

  Low income, limited digitalization and urbanization, with high vulnerable employment.
- Cluster 3 (red):

  Energy infrastructure in place but unsustainable, with significant social fragility.

This table shows the top 20 countries in terms of frequency within each cluster.

0	1	2	3
Cuba (18)	U. Arab Em. (18)	Angola (18)	Djibouti (18)
Ecuador (15)	Australia (18)	Burundi (18)	Lebanon (18)
Syria (14)	Austria (18)	Benin (18)	Maldives (18)
Iraq (13)	Belgium (18)	Burkina Faso (18)	Suriname (18)
Peru (12)	Canada (18)	Bangladesh (18)	Azerbaijan (17)
Venezuela (12)	Switzerland (18)	Bhutan (18)	Algeria (17)
Gabon (11)	Germany (18)	Centr. African Rep. (18)	Georgia (17)
Moldova (11)	Denmark (18)	Côte d'Ivoire (18)	Montenegro (17)
El Salvador (11)	Finland (18)	Cameroon (18)	Turkmenistan (17)
N. Macedonia (11)	U. K. (18)	D. Rep. of the Congo (18)	Bulgaria (16)
Dom. Rep. (10)	Hong Kong (18)	Congo (18)	Bosnia and Herz. (16)
Bolivia (10)	Ireland (18)	Comoros (18)	Greece (16)
Pakistan (7)	Iceland (18)	Ethiopia (18)	Jamaica (16)
South Africa (7)	Kuwait (18)	Guinea (18)	Latvia (16)
Portugal (7)	Luxembourg (18)	Gambia (18)	Morocco (16)
Albania (7)	Netherlands (18)	Haiti (18)	Mongolia (16)
Philippines (7)	Norway (18)	Kenya (18)	Palestine (16)
Croatia (6)	Oman (18)	Cambodia (18)	Ukraine (16)
Mexico (5)	Qatar (18)	Laos (18)	Turkey (16)
Egypt (5)	Singapore (18)	Liberia (18)	Botswana (15)

Table 8: Top 20 countries per cluster with counts (in parentheses). Columns 0–3 represent clusters.

The main countries listed appear, at least to an extent, to be consistent with the qualitative descriptions of the clusters to which they belong. It is worth noting that, while Clusters 1 and 2 clearly contain more than 20 countries that remain within the same cluster throughout the entire 18-year period (2005–2022), Clusters 0 and 3 (expecially Cluster 0) show a much higher degree of turnover. This pattern is further confirmed in the table below, which reports the countries with the highest number of cluster transitions: several of these are also among the most frequent members of Cluster 0 (for example, the Dominican Republic). The characteristics associated with Cluster 0 suggest that it might capture countries undergoing economic or institutional strain. Such conditions may be temporary or subject to structural change, which could explain the higher mobility observed within this cluster. Notably, Cuba is the only country that remains consistently in Cluster 0 across all years, possibly reflecting persistent structural constraints, including the long-standing United States embargo that inevitably weighs heavily on the country as a whole.

Table 9: Top 10 countries by number of cluster transitions in K-Means clustering ($K = 4$).

Country Code	Number of Transitions
Dominican Republic	15
Portugal	14
Paraguay	12
Mexico	12
Slovakia	12
Estonia	11
Spain	11
Czechia	11
Peru	11
Moldova	10

Examining the distribution of countries according to the number of cluster transitions, we observe that the average number of transitions is 3.8, with a median of 3 and a mode of 1. This suggests that, while a limited number of countries change clusters very frequently, the typical country experiences a moderate number of transitions over the 18-year period. The fact that the most common value is 1 indicates that many countries exhibit relatively stable clustering behavior, whereas the higher mean reflects the presence of a smaller group of highly volatile cases that increase the overall average.

Table 10: Distribution of countries by number of cluster transitions in K-Means clustering ($K = 4$).

Number of Transitions	Number of Countries
0	22
1	47
2	6
3	17
4	6
5	14
6	13
7	11
8	6
9	5
10	7
11	4
12	3
14	1
15	1

The temporal stability of the K-Means clustering with $K = 4$ indicates that most countries maintain relatively consistent cluster assignments over the 18-

year period. Although a subset of countries changes clusters multiple times, the overall distribution of transitions suggests moderate volatility rather than widespread instability. The average number of transitions per country is 3.8, with a median of 3 and a mode of 1. This implies that while a small group of countries exhibits relatively high mobility, the most common behavior is limited movement across clusters.

Yearly stability (defined as the proportion of countries remaining in the same cluster between consecutive years) ranges from 0.77 to 0.89, further confirming a generally robust structure. The average transition matrix reinforces this interpretation: diagonal probabilities are consistently high, indicating strong persistence within clusters, while off-diagonal entries reveal systematic but contained transitions, primarily involving Cluster 0.

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**

Taken together, these findings suggest that the chosen K-Means configuration produces clusters sufficiently stable and interpretable over time, although not perfectly so. Some degree of mobility, particularly around Cluster 0, appears intrinsic to the structure of the data rather than indicative of a fundamental weakness of the clustering solution.

5.3 Clara

CLARA identified 6 clusters, resulting in a more granular but less stable partition of the countries compared to the K-Means solution. Summary of the clusters identified:

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

- Cluster 5 (brown): **Economically weak, low exports and digital access, fragile social mood.**

This table shows the top 20 countries in terms of frequency within each cluster.

0	1	2	3	4	5
Algeria (18)	Belize (18)	U. Arab Em. (18)	Burundi (18)	Cuba (18)	Djibouti (17)
Libya (18)	Maldives (18)	Australia (18)	Benin (18)	Guyana (17)	Somalia (17)
Greece (17)	Sri Lanka (17)	Austria (18)	Burkina Faso (18)	Russia (16)	Nepal (16)
Bulgaria (16)	Vietnam (17)	Belgium (18)	Centr.Afr.Rep. (18)	South Africa (12)	Suriname (16)
Bosnia and H. (16)	Turkmenistan (16)	Canada (18)	Côte d'Iv. (18)	China (11)	Yemen (16)
Iraq (16)	Kyrgyzstan (15)	Switzerland (18)	Liberia (18)	Argentina (6)	Gambia (15)
Lebanon (16)	Thailand (14)	Denmark (18)	Madagascar (18)	United States (6)	Bangladesh (8)
Latvia (16)	Tajikistan (14)	Finland (18)	Mali (18)	Mexico (6)	Comoros (8)
Morocco (16)	Mongolia (13)	Hong Kong (18)	Malawi (18)	Croatia (5)	Afghanistan (8)
Palestine (16)	Mauritius (13)	Ireland (18)	Niger (18)	Italy (5)	Syria (8)
Turkey (16)	Kosovo (12)	Iceland (18)	South Sudan (18)	El Salvador (5)	Haiti (7)
Albania (15)	Paraguay (11)	Kuwait (18)	Uganda (18)	Saudi Arabia (5)	Sudan (6)
Armenia (15)	Uzbekistan (11)	Luxembourg (18)	Angola (17)	Colombia (4)	Myanmar (5)
Egypt (15)	Indonesia (10)	Netherlands (18)	DR Congo (17)	Jordan (4)	Nicaragua (5)
Romania (15)	Honduras (9)	Norway (18)	Ethiopia (17)	Bolivia (4)	India (4)
Serbia (15)	Panama (8)	New Zealand (18)	Mozambique (17)	Ecuador (3)	Gabon (4)
Slovakia (15)	Jamaica (6)	Oman (18)	Nigeria (17)	Spain (3)	Pakistan (4)
Tunisia (15)	Botswana (6)	Qatar (18)	Rwanda (17)	Namibia (3)	Georgia (3)
Azerbaijan (14)	Guatemala (6)	Singapore (18)	Chad (17)	Dominican Republic (3)	Cameroon (3)
Belarus (14)	Moldova (6)	Sweden (18)	Togo (17)	Malaysia (3)	Guatemala (3)

Table 12: Top countries per cluster (0–5) with counts in parentheses.

The top countries within each cluster broadly align with the qualitative descriptions previously provided, and this holds consistently across all six clusters. Overall, the clustering structure appears to capture a variety of country characteristics, rather than simply reproducing a binary "poor vs. rich" classification. The differentiation across clusters reflects multiple socio-economic and structural dimensions, suggesting that the partitioning captures more nuanced patterns in the data.

That said, the two most stable clusters are clearly those predominantly composed of high-income and low-income countries (Clusters 2 and 3). In these clusters, the socio-economic component of the description is particularly pronounced, and a large number of countries remain almost entirely stationary over the full time span. This indicates a strong structural persistence for countries at the two ends of the development spectrum.

By contrast, the other clusters (1, 4, and 5) display a more mixed composition. They contain only a small number of highly stationary countries — often including the medoid itself, which likely acts as an anchor point for the cluster — while the remaining members exhibit greater mobility. Cluster 0 also includes several countries that remain relatively close to one another over time, although turnover is more frequent compared to Clusters 2 and 3.

It is worth noting that the medoid-based approach, which centers clusters around representative countries, may in this specific application introduce some structural bias. By construction, certain countries exert a stronger anchoring effect on cluster composition, which could amplify local similarities and potentially increase sensitivity to relative shifts in the data.

At the same time, examining the most stationary countries within each cluster reveals several interesting and non-obvious pairings, suggesting that the clustering process uncovers meaningful structural affinities that are not immediately apparent from simple income-based classifications.

Table 13: Top 10 countries by number of cluster transitions in CLARA clustering ($K = 6$).

Country	Number of Transitions
Colombia (COL)	17
Peru (PER)	15
Argentina (ARG)	14
Czech Republic (CZE)	14
South Korea (KOR)	14
Myanmar (MMR)	13
Nicaragua (NIC)	13
Ecuador (ECU)	13
Botswana (BWA)	13
Kazakhstan (KAZ)	12

The average number of transitions increases to 5.78, compared to the K-Means configuration. The mode remains equal to 1, indicating that the most common outcome is still a single transition over the entire period. However, the median rises to 5, signaling a substantial upward shift in the central tendency of the distribution.

This pattern suggests that, overall, countries tend to move across clusters more frequently under the CLARA configuration. While a non-negligible group of countries remains relatively stable (as reflected by the unchanged mode), the higher mean and median indicate a broader and more systematic increase in cluster mobility. In other words, compared to K-Means, the six-cluster medoid-based solution appears to produce a more dynamic — and potentially less temporally stable — partition of countries over time.

Table 14: Distribution of countries by number of cluster transitions in CLARA clustering ($K = 6$).

Number of Transitions	Number of Countries
0	2
1	35
2	4
3	21
4	7
5	16
6	7
7	16
8	12
9	10
10	8
11	10
12	6
13	4
14	3
15	1
17	1

The average transition matrix for CLARA with 6 clusters shows a significant amount of off-diagonal mass, indicating that countries frequently switch clusters from year to year.

	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.

This result confirms the patterns that were already highlighted in the previous analysis. While diagonal entries remain dominant, indicating a certain degree of persistence, there is substantial off-diagonal mass, pointing to frequent year-to-year reassignments, especially in the already observed less stable clusters.

As anticipated, Clusters 2 and 3 show the highest stability, whereas Clusters 4 and 5, and to a lesser extent 0 and 1, display considerably greater mobility. This reinforces the earlier observation that the six-cluster configuration pro-

vides finer differentiation but at the cost of reduced temporal stability. Overall, the analysis consolidates the interpretation that the CLARA solution is still structurally sufficiently coherent but more nuanced and more dynamic than the K-Means alternative. It is also worth noting that the medoid-based approach might have slightly overfitted to the 2017 data, which could amplify transitions in clusters that are less structurally stable.

5.4 Hierarchical and DBSCAN

5.4.1 Evaluation of cluster stability

The clustering algorithms were applied to all years using the same hyperparameters, but it cannot be guaranteed that the clusters carry the same semantic meaning across different years. To provide an initial assessment of the clustering characteristics — albeit an imperfect one for these types of algorithms — the mean and standard deviation of each cluster for the year 2017 were evaluated during the hyperparameter selection and clustering process.

As a first step toward understanding the potential reusability of clustering results for subsequent years, the temporal evolution of cluster centroids was examined. These centroids were computed as the mean values of the attributes within each cluster for each year. By visualizing them in a heatmap, a preliminary sense of the stability of cluster structures over time can be obtained, providing valuable insight into whether the clusters maintain a coherent interpretation from one year to the next.

This approach is considered reasonable in this context for several reasons. First, the main clusters of interest are only two, which limits the risk of overgeneralization. Second, the clusters are based on socio-economic indicators that, while multi-dimensional, are relatively continuous and not strongly non-linear. As a result, using centroids to summarize cluster characteristics is appropriate and provides a practical approximation for evaluating temporal stability across years.

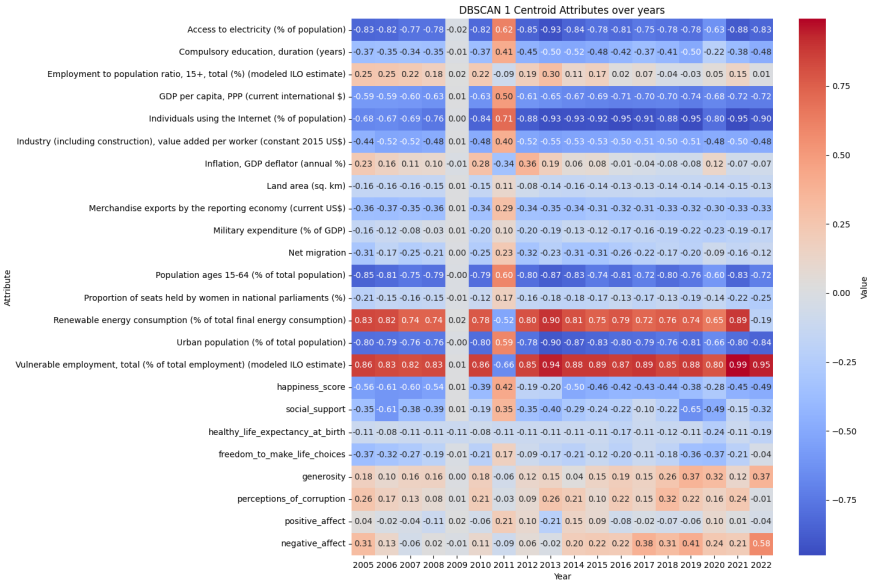

Figure 35: DBSCAN cluster 1 centroid over time

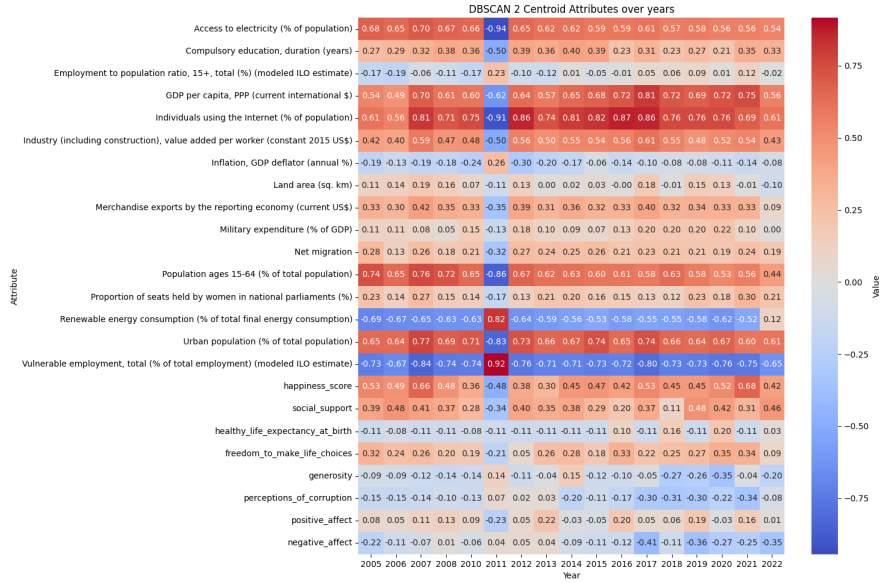

Figure 36: DBSCAN cluster 2 centroid over time

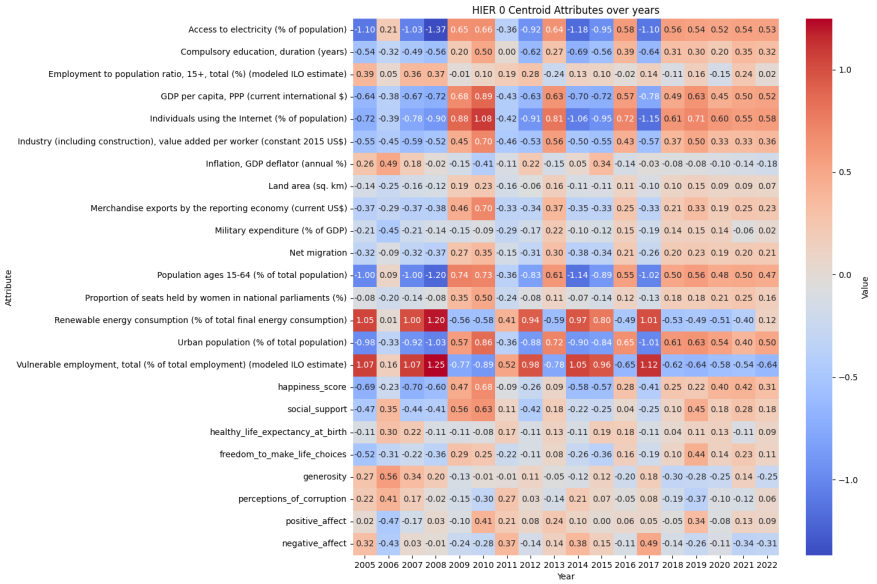

Figure 37: Hierarchical cluster 0 centroid over time

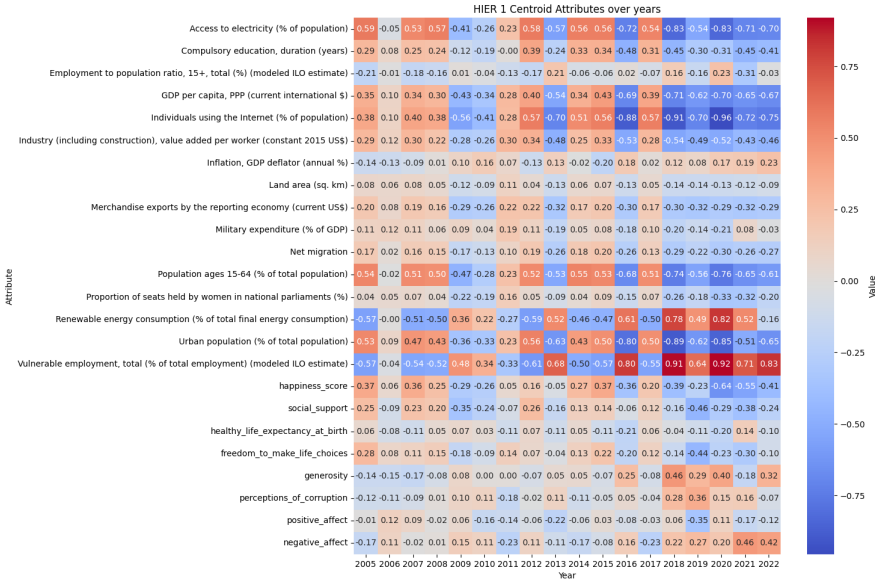

Figure 38: Hierarchical cluster 1 centroid over time

In DBSCAN, the clusters exhibit relatively stable centroids from year to year, with the exception of 2011 and considering that in 2009 only a single

cluster was present. In contrast, in Hierarchical clustering, the centroids show noticeable shifts over time, sometimes even "reversing" their relative positions. This suggests that, while DBSCAN produces clusters with more consistent attribute profiles across years, Hierarchical clustering may be more sensitive to temporal variations in the data.

5.4.2 DBSCAN

The clusters identified in 2017, which remain comparable throughout the observation period, are as follows:

- Cluster 1: Poorer, less industrialized countries.
- Cluster 2: Richer, more industrialized countries.
- : outliers were countries such as Palestine, Armenia, Venezuela.

0	1	2
Maldives (14)	Afghanistan (18)	United Arab Emirates (17)
Palestine (12)	Angola (18)	Argentina (17)
Turkmenistan (10)	Burundi (18)	Australia (17)
Brazil (8)	Benin (18)	Austria (17)
Venezuela (8)	Burkina Faso (18)	Belgium (17)
Cuba (8)	Bhutan (18)	Bulgaria (17)
Russia (8)	Central African Republic (18)	Bahrain (17)
Georgia (7)	Côte d'Ivoire (18)	Canada (17)
Kosovo (7)	Cameroon (18)	Switzerland (17)
Uzbekistan (6)	DR Congo (18)	Costa Rica (17)
Jordan (5)	Congo (18)	Germany (17)
Gabon (5)	Comoros (18)	Denmark (17)
Iran (5)	Ethiopia (18)	Finland (17)
Mongolia (5)	Ghana (18)	France (17)
Suriname (5)	Guinea (18)	United Kingdom (17)
Iraq (5)	Gambia (18)	Hong Kong (17)
Kyrgyzstan (5)	Haiti (18)	Ireland (17)
Thailand (4)	Kenya (18)	Iceland (17)
Armenia (4)	Cambodia (18)	Israel (17)
Bosnia and Herzegovina (4)	Laos (18)	Italy (17)

Table 16: Top countries per cluster (0–2) with counts in parentheses.

The countries listed in Cluster 0 largely consist of outliers or less consistently grouped nations, which tend to deviate from the main socio-economic patterns captured by Clusters 1 and 2. Many of these countries, such as the Maldives, Palestine, and Turkmenistan, have relatively small populations or unique political, geographic, or economic circumstances, which may explain their placement

in a separate cluster. This aligns with DBSCANs sensitivity to density, as these countries do not form dense clusters with others and are therefore treated as outliers or loosely connected members.

Regarding the remaining clusters, in both cases the top 20 countries per cluster remain relatively stable over time. It is worth noting, however, that in 2009 the cluster representing the "wealthier" countries did not appear at all. Interestingly, this coincides with the year following the 2008 global financial crisis, which likely caused significant disruptions in economic indicators and may have temporarily altered the structure of the clusters. Overall, aside from this anomaly, the core membership of the other clusters shows a high degree of consistency.

Many of the countries listed as transitioning here below frequently move between a main cluster and the outlier category. This behavior reflects their unique socio-economic, political, or geographic conditions, which prevent them from consistently fitting into dense clusters. Such transitions are therefore expected and illustrate DBSCAN's sensitivity to atypical or extreme cases rather than indicating an error in clustering.

Country	Number of Transitions
Iran (IRN)	15
Uzbekistan (UZB)	15
Bosnia and Herzegovina (BIH)	14
Suriname (SUR)	13
Egypt (EGY)	12
Palestine (PSE)	12
Thailand (THA)	12
Cuba (CUB)	12
Moldova (MDA)	12
Brazil (BRA)	12

Table 17: Top 10 countries by number of cluster transitions.

Number of Transitions	Number of Countries
0	48
2	2
3	38
4	5
5	11
6	3
7	13
8	11
9	7
10	7
11	8
12	6
13	1
14	1
15	2

Table 18: Distribution of countries by the number of cluster transitions.

The distribution of cluster transitions shows that the majority of countries experience few changes, with a mean of approximately 4.57 transitions over the observation period. The median number of transitions is 3, indicating that at least half of the countries change clusters three times or fewer. The mode is 0, meaning the most common scenario is for countries to remain in the same cluster throughout the period. Overall, while some countries exhibit frequent transitions, the typical country demonstrates relatively high temporal stability in DBSCAN clustering.

	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. 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. Bold numbers indicate the highest probability for each row, italic numbers highlight notable off-diagonal transitions.

The transition matrix confirms what was already observed: many outliers "oscillate" between being classified as outliers and rejoining one of the main clusters.

In general, DBSCAN identifies two coarse clusters and highlights which countries are stable within a cluster and which tend to fluctuate. While this global-level clustering captures fewer nuanced aspects of each country, it still provides a meaningful overview of stability and variability across the world.

5.4.3 Hierarchical

To enable the same analysis for the Hierarchical clustering results, clusters were recombined based on the centroid heatmap. Except for the year 2006, which shows a different configuration, the clusters are largely comparable across the observation period.

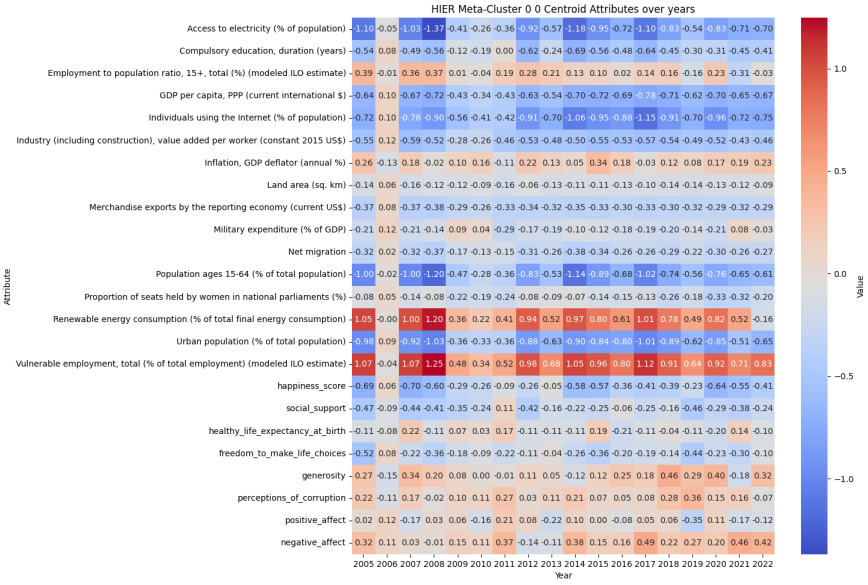

Figure 39: Hierarchical cluster 0 centroid recombined over time

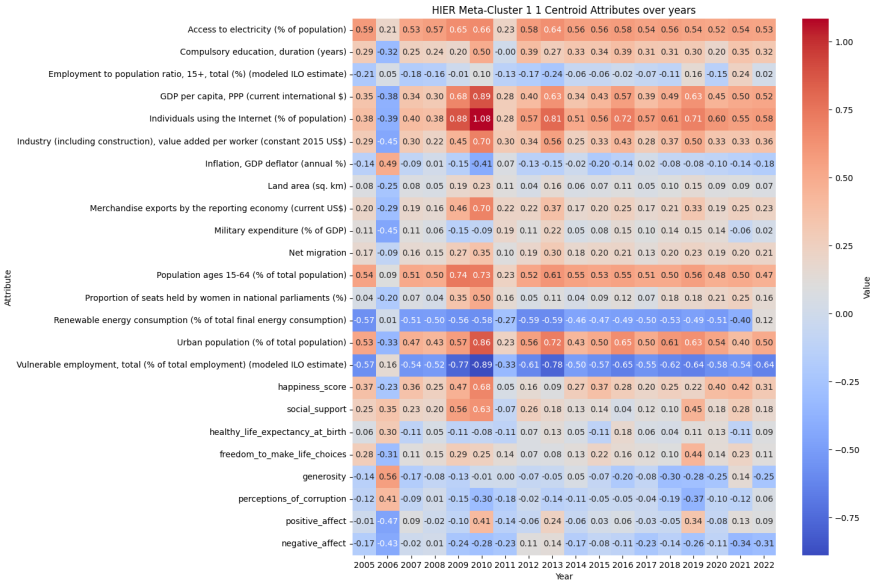

Figure 40: Hierarchical cluster 1 centroid recombined over time

The clusters that were originally found on the 2017 data showed these results

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

Table 20: Top countries per cluster

Cluster 0	Cluster 1
Gambia (18)	United Arab Emirates (18)
Burundi (17)	Belgium (18)
Benin (17)	Bulgaria (18)
Burkina Faso (17)	Belarus (18)
Central African Republic (17)	Switzerland (18)
Côte d'Ivoire (17)	China (18)
Cameroon (17)	Germany (18)
Democratic Republic of the Congo (17)	Denmark (18)
Republic of the Congo (17)	Finland (18)
Comoros (17)	France (18)
Ethiopia (17)	United Kingdom (18)
Haiti (17)	Hungary (18)
Kenya (17)	Ireland (18)
Liberia (17)	Iceland (18)
Madagascar (17)	Japan (18)
Mali (17)	South Korea (18)
Myanmar (17)	Kuwait (18)
Mozambique (17)	Luxembourg (18)
Niger (17)	Netherlands (18)
Nigeria (17)	Norway (18)

Looking at the top countries per cluster, it appears that the 20 richest countries tend to remain in the "rich" cluster across all years, while most poorer countries have at least occasionally been assigned to the other cluster. This behavior is likely influenced by the characteristics of Hierarchical Agglomerative Clustering with complete linkage, which tends to produce compact clusters, combined with the yearly z-score normalization of the data, making assignments relative to each year's distribution. Overall, the pattern highlights the stability of the most homogeneous groups while also reflecting some variability among less stable countries.

Table 21: Top 10 countries by number of transitions

Country	Number of Transitions
Philippines	13
Iraq	13
Egypt	13
Djibouti	12
Mongolia	12
Vietnam	12
Thailand	12
Kyrgyzstan	12
Belize	12
Suriname	11

Interestingly, the countries that frequently switch clusters under Hierarchical Agglomerative Clustering with complete linkage tend to be those that, in previous clusterings, occupied borderline positions in terms of socio-economic indicators. For example, countries like Iraq and Egypt, which sometimes belonged to the "poorer" cluster in CLARA, or Djibouti, the medoid of cluster 5, occasionally shift between clusters. This behavior may be related to the way complete linkage prioritizes maximum distances between clusters: in high-dimensional space, borderline countries can sometimes be associated with clusters that are relatively less typical for them.

Table 22: Distribution of transitions across countries

Number of Transitions	Number of Countries
0	26
1	1
3	47
4	4
5	29
6	3
7	22
8	6
9	11
10	3
11	2
12	6
13	3

The distribution of cluster transitions across countries shows that most countries experience a moderate number of changes. The average number of transitions is approximately 4.8, indicating that on average, countries switch clusters nearly five times over the observation period. The median is 5, consistent with the mean, while the mode is 3, showing that the most common number of transi-

tions is slightly lower than the average. This confirms that while some countries remain stable, a substantial fraction undergo multiple transitions, highlighting variability in cluster assignments over time.

Table 23: Transition matrix (example)

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

The transition matrix supports this observation. The probabilities of moving between clusters are nearly symmetric, with roughly 70–80 per cent of countries remaining in their original cluster from one year to the next, and 20–30 per cent switching. This pattern confirms that while the majority of countries exhibit stable cluster assignments, a notable minority of countries frequently transitions between clusters, consistent with the variability observed in the distribution of cluster transitions.

5.5 Final remarks

Silhouette scores more or less confirmed the clustering quality trasversally on all algorithms and years, showing non conforming value only on the year that was most problematic for Hierarchical and for the year in which DBSCAN only had one cluster.

In general, the four clustering approaches shows distinct strengths and limitations, which reflect both their methodological properties and the characteristics of the data.

K-Means ($K = 4$) produced clusters that are mostly reasonably stable over time and interpretable in terms of socio-economic and structural dimensions. The top countries per cluster align well with the qualitative descriptions, and the transition analysis indicates moderate mobility, mainly concentrated in Cluster 0.

CLARA ($K = 6$) offered a more granular partition, revealing subtler distinctions among countries, but this higher resolution came at the cost of temporal stability. Transition analysis and the average transition matrix indicate more frequent movement across clusters, particularly in Clusters 0, 1, 4, and 5, reflecting the dynamic nature of the medoid-based solution. Its results are more sensitive to small variations in the data and may be slightly overfitted to the reference year (2017).

DBSCAN identifies two broad clusters and clearly separates outliers. Its focus on density makes it robust for detecting anomalies and extreme cases, but the algorithm is less suited to capturing finer gradations among the majority of countries due to the high amount of dimensions. Low silhouette scores reflect the difficulty of defining compact clusters in high-dimensional socio-economic space, yet the transition patterns confirm that most countries remain largely stable, with outliers oscillating as expected.

Hierarchical clustering with complete linkage produces relatively interpretable clusters when recombined for temporal consistency. However, centroids show noticeable shifts over time, particularly for countries at the margins of the main clusters. This reflects the sensitivity of complete linkage to the maximum distances between members and the impact of yearly normalization. While the main clusters (rich vs. poor) remain broadly consistent, countries with intermediate or borderline socio-economic characteristics exhibit greater mobility.

Ultimately, no definitive clustering solution was found. Despite the significant dimensionality reduction performed, the dataset remains high-dimensional, and in hindsight, some variables could have been further eliminated, as they do not appear to contribute meaningfully to the clusters. This was intentionally left as is, to explore the extent of the "curse of dimensionality" in this context.

Nevertheless, CLARA and K-Means uncovered a few genuinely interesting relationships between countries, like persistent structures and transitional patterns. DBSCAN identified a stable division into two coarse clusters and successfully detected an anomaly in 2009, following the 2008 global financial crisis. Hierarchical clustering highlighted that even when using yearly z-score normalization to make clusters relative to each year, seemingly stable clusters can still shift over time.

A natural next step would be to track each country individually to detect recurring patterns, particularly using CLARA or K-Means where transient clusters exist, and to compare these patterns with real-world events. Another intriguing avenue would have been to implement CLIQUE for subspace clustering, though in practice it proved extremely challenging to obtain consistent clusterings across years and to visualize the results effectively.

Overall, while the high dimensionality and limited granularity constrained the emergence of perfect clusters, the study demonstrates that even in these conditions, meaningful and interpretable country-level patterns can still be extracted.