Data Science Lab, a.y. 2023/2024

Word Value Survey: a data science approach

Alice Brunazzi[1], **Alessandro Della Beffa**[2], **Daniele Lepre**[3] **and Sofia Turroni**[4]

[1]*Bachelor's degree in Banking and Finance, Università degli Studi Milano-Bicocca*
[2]*Bachelor's degree in Mathematics, Università degli Studi di Milano*
[3]*Bachelor's degree in Banking and Finance, Università degli Studi Milano-Bicocca*
[4]*Bachelor's degree in Statistics, Università degli Studi Milano-Bicocca*

This manuscript was compiled on August 3rd, 2024

Abstract

This research explores cultural patterns across nations using data from the World Values Survey (WVS) and a data science-driven approach. By employing hierarchical clustering techniques, the study identifies and analyzes clusters of countries based on their cultural indices, including attitudes toward authority, religion, autonomy, equality, and socioeconomic variables such as GDP per capita and political regime types. The findings provide insights into how cultural values vary globally and correlate with factors like politics, religious beliefs and social dynamics. The research offers a comprehensive understanding of the cultural foundations that shape modern societies, shedding light on global cultural trends and their implications for socioeconomic and political landscapes.

Keywords: *World Values Survey, Inglehart–Welzel, hierarchical clustering, culture*

Revised: September 9th, 2024

Contents

1. Introduction

The World Values Survey (WVS [1]) is a comprehensive, global research initiative that aims to understand the changing cultural values, beliefs, and norms across different societies worldwide. Managed by the World Values Survey Association, this long-term project has been collecting data since 1981 from more than 100 countries, encompassing over 90% of the world's population. The survey investigates a wide array of topics, including political engagement, trust in institutions, gender roles, social tolerance, attitudes towards religion, environmental concerns, and other aspects that define the socio-cultural fabric of different nations.

One of the most significant contributions of the WVS is the development of the Inglehart [2]–Welzel [3] Cultural Map, which provides a visual representation of how countries are positioned along two primary dimensions of cultural variation: *traditional vs. secular-rational values* and *survival vs. self-expression values*. These dimensions help to categorize countries based on their predominant cultural attitudes, highlighting both differences and similarities across regions and societies.

- The *traditional vs. secular-rational values* dimension reflects the extent to which societies emphasize religion, traditional family values, and authority.
- The *survival vs. self-expression values* dimension captures the transition from materialist to post-materialist values, focusing on self-expression, tolerance, and quality of life.

The description of the aforementioned values deserves to be reproduced *verbatim*: "traditional values emphasize the importance of religion, parent-child ties, deference to authority and traditional family values. People who embrace these values also reject divorce, abortion, euthanasia and suicide. These societies have high levels of national pride and a nationalistic outlook. Secular-rational values have the opposite preferences to the traditional values. These societies place less emphasis on religion, traditional family values and authority. Divorce, abortion, euthanasia and suicide are seen as relatively acceptable. (Suicide is not necessarily more common.) Survival values place emphasis on economic and physical security. It is linked with a relatively ethnocentric outlook and low levels of trust and tolerance. Self-expression values give high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality, and rising demands for participation in decision-making in economic and political life." [4]

1.1. Goal of the project

The objective of this report is to conduct an analytical exploration of cultural indices derived from the WVS data, with a particular focus on the Inglehart-Welzel indices. To achieve this, we will employ a hierarchical clustering to identify and analyze clusters of countries based on their cultural similarities and differences.

The goal of this study is to understand how these culturally derived clusters correlate with various external factors, such as geographical location (continent), political regime type, population size, and GDP per capita. By examining these correlations, the report seeks to provide insights into the broader social, political, and economic implications of cultural values and to contribute to a deeper understanding of how cultural dynamics influence global trends and societal development.

1.2. Research questions

The aim of this project is to find an answer to the following questions:

- How do cultural values vary across different countries?
- Is it possible to find the matrix of the similarities (or differences) between different regions and countries?
- How do socioeconomic and political factors impact societal attitudes and values?

2. Dataset description

The dataset used in this project is sourced from the official World Values Survey (WVS) website, specifically from the seventh wave [5] of data collection, which spans from 2017 to 2022. This wave includes data from 66 countries and territories, with the majority of surveys conducted between 2018 and 2020. Due to the impact of the COVID-19 pandemic, approximately a dozen countries conducted their fieldwork between 2021 and 2022. The most recent survey in this wave was completed in India in July 2023. The standardized questionnaire used for this data collection is available at WVS Documentation WV7.

According to WVS guidelines, each country is surveyed only once per wave, using random probability sampling methods to ensure representative samples of the adult population. The vast majority of these surveys were conducted through face-to-face interviews, either using Paper-and-Pencil Interviewing (PAPI) or Computer-Assisted Personal Interviewing (CAPI) as the primary data collection method.

The original dataset comprises 97,220 rows and 613 columns, reflecting a broad spectrum of responses across the participating countries. Due to the dataset's extensive nature, a careful selection of relevant columns was necessary to focus the analysis on key cultural indices and other pertinent variables.

The nations with the highest number of respondends, as presented in Figure 1, are Canada (CAN) with more than 4000 respondents, China (CHN), Indonesia (IDN) and United Kingdom (GBR), each of which have over 3000 partecipans. At the same time, countries such as Cyprus (CYP), New Zealand (NZL) and Argentina (ARG), the ones with a limited sample size and lower representation, provide 1000 respondents each.

2.1. Data enrichment

To enhance the quality and relevance of our dataset, we have enriched it by incorporating additional variables, all independently grouped in the *cartel1* dataset. These variables report external information on countries, such as GDP per capita [6], population and continent [7], and regime type [8] (for more details, see paragraph 3.2). Initially, these columns were meticulously compiled and consolidated from various sources to provide a comprehensive foundation. To ensure consistency and comparability with the initial dataset, we utilized the columns *ISO A3* (country's identification code) and *B_COUNTRY_ALPHA* as standardized identifiers. The GDP per capita metric offers a refined measure of economic output

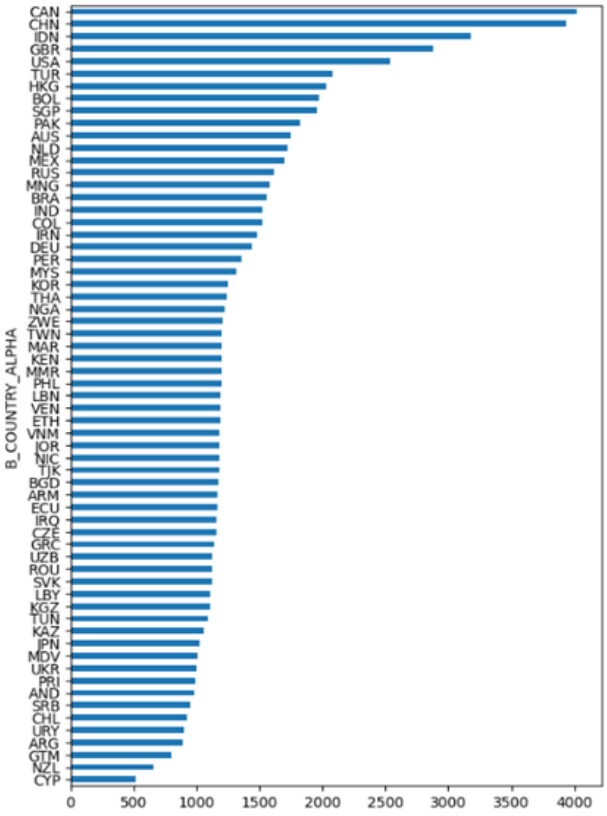


Figure 1. Distribution of respondents by country (see Appendix 1)

per individual, essential for assessing economic prosperity and standards of living across different nations. The population data, updated for the year 2022, provides insights population sizes, fundamental for socio-economic analysis. The continent variable categorizes countries by their respective geographical regions, facilitating comparative studies. Furthermore, the inclusion of regime type offers a perspective on the political contexts governing each country, highlighting how different government structures may impact economic performance, social progress, and stability.

2.2. Data cleaning process

To ensure a correct and consistent analysis, the dataset underwent a thorough cleaning process, during which certain adjustments were made to correctly label and categorize data:

- Some territories or regions were reconnected to their corresponding sovereign state from which they had been distinguished. For example, data from Northern Ireland (NIR) was annexed to the United Kingdom category, and data from the Special Administrative Region of Macao (MAC) was associated with the People's Republic of China;
- The questionnaire's answers were grouped by country of the respondent;
- To maintain the integrity of the dataset and prevent errors during analysis, all rows (of the 8 columns analyzed) containing missing values were removed. As a result of this operation, the number of rows in the dataset was reduced from 97,220 to 89,940, representing a removal of approximately 7.5% of the data;

After the cleaning process, out of the initial 66 countries analyzed, only 63 of them were displayed.

3. Feature selection

The selection of specific columns from the World Values Survey dataset was driven by the need to focus on key indicators that are most relevant to the analysis at hand, that will be discussed in the following paragraphs. Therefore, we chose to focus on a subset of columns

that reflect core indices, while excluding variables that, although related, do not contribute significantly to the primary objectives of the study. This selective approach ensures that the analysis remains focused and manageable, allowing for more accurate and meaningful clustering and comparison across countries. By definition, if there are no missing responses, the indices are constructed by averaging the values associated with three related questionnaire variables; in cases where one such variable is missing (the question has not been answered), a weighted average is used, and an approximated, linearly estimated value of the missing variable is computed instead.

3.1. Indices

All of the following variables takes values in [0, 1], as a result of each being the average of three discrete sub-indices ranging in [0, 1].

The first index is *Defiance*, built from three questions revolving around agreement on: 'Greater respect for authority', 'The goal of my life is to make my parents proud' and 'How proud are you to be of nationality of your country?'. Low values indicate high respect, fascination or pride for authorities.

The *Disbelief* index aims to measure the relevance that a religion holds in the life of the respondent. The questions asked to form the index are: 'Would you say that religion is important in your life?', 'Would you define yourself as a religious person?' and 'How often do you pray?'. Low values indicate firm religious beliefs and participation.

The *Relativism* index is structured to assess the respondent's moral and ethic stance on both minor and major forms of dishonesty or rule-breaking. Specifically, it gauges the respondent's attitude toward: fare evasion, tax evasion and corruption via bribe. Low values indicate high sense of righteousness and moral integrity with respect to such matters.

The *Scepticism* index aims to measure trust in state institutions, namely armed forces, police, and justice system. Low values indicate high confidence in these public institutions.

The *Autonomy* index evaluates the respondent's perspective on the importance of teaching and autonomy in raising children, by assessing three key qualities: independence, imagination, and obedience. High values indicate a high consideration of the teaching of such values, thus stressing the importance of achieving maturity.

The *Equality* index measures the respondent's views on gender equality across three dimensions: the beliefs that men have more right to a job than women, that men make better political leaders than women, and that a university education is more important for boys rather than girls. Low values indicate discriminatory, male-oriented gender opinions in social matters.

The *Choice* index assesses the respondent's stance on personal freedoms and social choices, focusing on three areas: homosexuality, abortion, and divorce. Each of these aspects is rated on a scale that ranges from 'never justifiable' (0) to 'always justifiable' (1).

The *Voice* index evaluates both the individual and collective priorities regarding the role of citizens participation and influence in political matters. Questions aim to identify the respondent's first and second most important topics for themselves and for their country in the upcoming ten years, with options ranging from economic growth and strong armed forces to increased public influence in decision-making and enhanced beauty of cities and countryside. High values indicate active participation in relevant political issues and freedom of expression as priorities for the individual or the country; low values indicate priorities other than these.

3.1.1. Indices distribution

Figure 2 shows a series of histograms depicting the distribution of several indices derived from the World Values Survey dataset, each providing insights into the respondents' values and beliefs across various dimensions.

Defiance shows higher density towards 0 and a long flat tail towards 1, indicating that a substantial portion of respondents exhibits a profound respect for authority. *Disbelief* shows an evident polarization towards the extremes, which translates into strong, opposed and similarly likely opinions on the importance of religion in one's life, with no possible middle grounds. *Relativism* and *Autonomy* are fundamentally discrete distributions with only four values. It can only be said that *Relativism* has high extremes and lower central values, as opposed to *Autonomy*, but further interpretation in unsuited for such distributions. Values of *Scepticism* are fairly evenly distributed, only growing slightly denser around the left-central values; levels of trust in public institutions thus greatly vary among the respondents, but are nonetheless moderate for the most part. *Equality* is irregularly distributed but still leaning towards central and high values, thus only reflecting a perceptible tendency among respondents to support aspects of gender equality. *Choice* has a regularly decreasing distribution towards 1, though not nearly as pronounced as that of *Defiance*; in a word, although the tendency among the respondents is one of openness to freedom of personal choices, adverse opinions remain widespread. Finally, *Voice* shows unevenly distributed values with isolated peaks close to 0 and 0.5, and does not allow to distinctly identify behaviours other than varied priorities concerning citizens' participation and influence in political decisions.

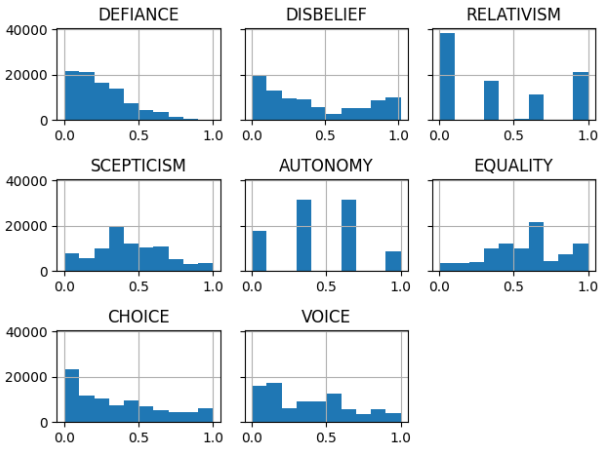


Figure 2. Indices Distribution

The correlation matrix of the indices has objectively low absolute values of linear correlation between any two of them, the highest being *Choice-Disbelief* (0.45) and *Choice-Equality* (0.42), and the rest being less than 0.35. Overall, this shows the eight indices as non-overlapping representatives of a wide spectrum of themes.

3.2. Other features

In order to further analyze the results computed by the cluster analysis, other measures were extracted both by the initial WVS dataset and the integrated *cartel1* dataset. These measures are:

- *RegimeType2022*, categorical, is the countries' Economist Democracy Index as of 2022; it takes four possible values: 'Flawed democracy' (e.g. United States), 'Full democracy' (Canada), 'Hybrid regime' (Turkey) and 'Authoritarian' regime (Russia);
- *GDPpercapita2022*, numeric, indicates the gross domestic product per capita in dollars ($) for 2022, as reported by the World Bank Group;
- *Q274*, numeric integer, answers the question: 'How many children do you have?';
- *Q275R*, numeric integer, indicates the highest education level;
- *Q289*, numeric integer, denomination of major religious groups.

To read all the possible answers see Appendix 3.

4. Clustering

Since the dataset to be clustered has 63 rows (countries), the choice of algorithm immediately fell on hierarchical clustering, which for such a limited number of rows—and regardless of the number of columns—can be easily visualised with a dendrogram. Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively; the root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. For more details, see the scikit-learn documentation on hierarchical clustering [9]. In Python, hierarchical clustering is implemented by the *AgglomerativeClustering* object (*AgglomerativeClustering* documentation [10]) and the main parameters are either the number of clusters or the maximum distance threshold, once defined; the choice of this distance is defined by the 'linkage' parameter, i.e. the criterion to be minimised in order to separate the sets of observations. Experiments have been carried out with all available methods ('ward', 'complete', 'average' and 'single'), and the best defined result in terms of separation of observations is obtained with Ward's method (based on the within-cluster variance). More details on the clustering criteria can be found in the previous *AgglomerativeClustering* documentation. On the basis of the dendrogram obtained (see visualization [11] and example [12]), we preferred to consider a variable number of clusters (minimum 7, maximum 9), to be precisely defined later according to the needs of interpretation. The 9 clusters obtained are shown in Figure 3, and the possible groupings according to the hierarchy are, in ascending order of distance, 4-7 and 1-8.

The table of average indices within the clusters (9 rows, 8 columns) provided us with the basic numerical information for interpreting the clustering; with this, using multidimensional scaling (MDS), we created a two-dimensional representation of the clusters from the original space of dimension 8. In general, MDS is a technique used for analysing similarity or dissimilarity data. In Metric MDS, the distances between output two points are set to be as close as possible to the similarity or dissimilarity data. In the non-metric version, the algorithms will try to preserve the order of the distances. For more details see scikit-learn documentation on multidimensional scaling [13]. Having experimented with both types of MDS and having found that the results are compatible, if not overlapping, we report here the map of the non-metric MDS (Figure 4), in which the bubbles have an area proportional to the number of respondents in the corresponding cluster.

In addition to multidimensional scaling, we decided to generate a self-organising map (SOM) of the 63 countries in terms of the 8 indices. In general, the SOM represents a distribution of input data items using a finite set of models, which are automatically associated with the nodes of a two-dimensional grid in an orderly fashion such that more similar models become automatically associated with nodes that are adjacent in the grid. For more details, see this reference. Our intention was not so much to create a new clustering after the hierarchical one, but to use the natural two-dimensional arrangement of the SOM points (the countries) to confirm or refute the compactness and isolation of the already known clusters. Since the SOM in its original formulation by its creator Teuvo Kohonen was implemented in R (Kohonen package documentation [14]), this part of our work was also carried out in R. Following a common practice described for example here, we set up a 6×6 square grid, so that we have approximately $5\sqrt{n}$ nodes ($n = 63$ is the number of observations). Figure 5 shows the SOM distribution of countries identified by the number of clusters.

4.1. Other experiments

At a later time, we decided to test another clustering algorithm on the dataset to compare it with the hierarchical one, allowing us to evaluate the results of different techniques and determine which one could best suit our problem. Our choice was on the k-means algorithm as it is one of the most widely used algorithm in clustering, its implementation is very simple and has a low computational complexity. Specifically, we used the Python library KMeans and its function *KMeans* which takes as input the number of clusters we would like to use for the analysis and the random state (we used value 42) for the reproducibility.

K-means clustering, as many other clustering techniques, is subject to the curse of dimensionality, meaning that high-dimensional data cause the objects to become equidistant, making it more challenging to find meaningful distances and clusters. Hence it is important to try and avoid this by performing a dimensional reduction. In order to do so, we implemented the technique called UMAP (Uniform Manifold Approximation and Projection, [15]), which is similar to t-SNE and Laplacian eigenmaps, as it derives from Belkin and Niyogi's work on Laplacian eigenmaps. It is particularly suited for clustering problems in that it aims to preserve the local distances among data rather than the global distances; such characteristic makes this technique useful for tasks where we want to identify local groups of data and visualize them in a low-dimensional space. To perform UMAP in Python we used the umap library, and its function *UMAP* which takes as input four parameters: *n_neighbors* (the number of neighbouring points used in local approximations of the manifold structure; default 15, then changed to 10 because of the few observations); *n_components* (the number of UMAP dimensions, set to 2 for obvious visualization reasons); *random_state* (the reproducibility index, default 42); and *metric* (the metric used to measure distance in the input space: in this case the euclidean distance).

Another important matter while performing a k-means clustering is choosing the optimal number of clusters as it is not defined a priori. There are many techniques that can be used to detect the optimal number of clusters, and since we are willing to compare our clustering results through the Silhouette index, we can also use it to choose the appropriate number of clusters. This method simply runs the clustering as many times as the different numbers of clusters we set, then computes the Silhouette index at each step, and finally chooses the optimal number.

The Silhouette index [16] is a useful tool to evaluate the goodness of a clustering algorithm as it does not need a training set to evaluate the results. The Silhouette value, expressed as $s(x_i)$, for the point x_i is defined as:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left(b(x_i), a(x_i)\right)} \tag{1}$$

Where x_i is a data point in the cluster π_k, $a(x_i)$ is the average distance of x_i to all other elements in the cluster π_k (within dissimilarity), and $b(x_i)$ is the minimum of the average distance from x_i among all clusters. The value of the Silhouette index can range between -1 and 1. The greater the Silhouette of an object is, the higher the likelihood to be clustered in the correct group, while observations with negative values of this index are likely to be clustered in the wrong group. In Python, the Silhouette index can be computed using the function *silhouette_score* from the library sklearn.metrics.

Number of clusters	Silhouette index
2	0.54
3	0.50
4	0.56
5	0.58
6	0.59
7	0.60
8	0.57
9	0.59
10	0.56

Table 1. Silhouette indices

Results of the iterative process showed in table (1) lead us to the conclusion that, formally, the optimal number of clusters for the k-

means algorithm is 7; differences in the Silhouette index with smaller but close numbers are minimal, hence confirming the first impression between 7 and 9.

4.2. Final considerations

We do not elaborate further on the description of *k*-means clustering, since we developed it as a parallel exercise to the hierarchical algorithm, which we ultimately preferred. Indeed, agglomerative clustering seems to distinguish differences between clusters more accurately since it does not require dimensionality reduction like UMAP to be visualized. Keeping the data in their original space allows the preservation of all variables, avoiding sacrificing the importance of those that might appear more marginal. Moreover, constructing a dendrogram based on the agglomerative clustering allows to explore the hierarchical structure and later decide how many clusters to use. We therefore limit ourselves to pointing out some reassuring similarities between the two techniques, despite the inevitable differences. We can affirm that there are no substantial differences in the distribution of countries within clusters between the two techniques, all variations that exist between the clusters identified by *k*-means are minor and occur within the same sub-branch of the hierarchical structure produced by the agglomerative clustering. This indicates that the core structure of the clusters remains consistent, even if the specific assignments of some countries may differ slightly between the two approaches. In particular, two clusters turn out to be perfectly identical in both techniques: one consisting of Andorra, Australia, Canada, Germany, the United Kingdom, the Netherlands, New Zealand, Uruguay, and the United States (cluster 5); the other consisting of Bolivia, Colombia, Ecuador, Guatemala, Mexico, Nicaragua, Peru, and Venezuela (cluster 6).

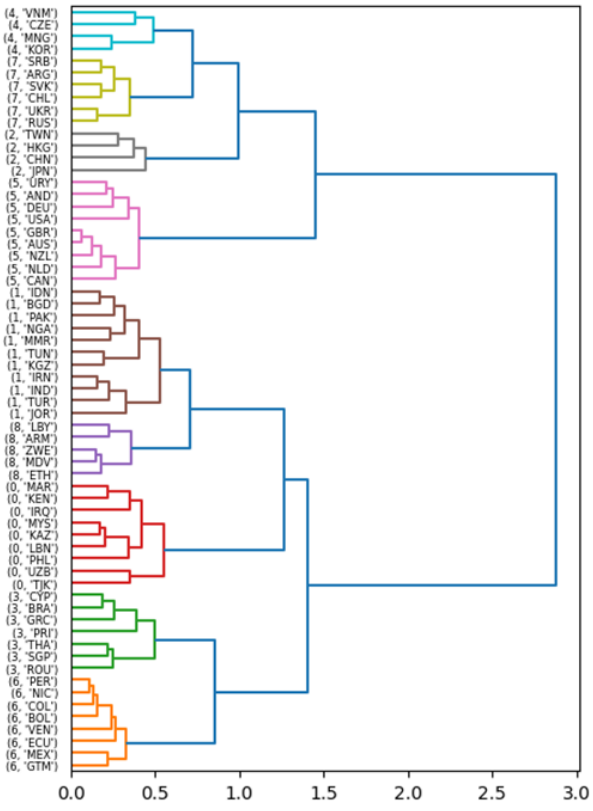


Figure 3. Dendrogram with the nine resulting clusters (see Appendix 1)

5. Results

Figure 3 provides a useful visualization of the hierarchical clustering of countries, the dendrogram, based on their similarities across the selected indices. Specifically, the dendrogram helps identify groups of countries by showing an overview of the relative distance of each object from any other.

The countries appear divided into the 9 clusters as follows:

- Cluster 0 includes Morocco, Kenya, Iraq, Malaysia, Kazakhstan, Lebanon, the Philippines, Uzbekistan, and Tajikistan; it represents 11.89% of the analyzed population;
- Cluster 1 includes Iran, Pakistan, Turkey, Bangladesh, Myanmar, Indonesia, Tunisia, Kyrgyzstan, Nigeria, and India; it represents 19.05% of the analyzed population;
- Cluster 2 includes China, Hong Kong, Japan, and Taiwan; it represents 9.15% of the analyzed population;
- Cluster 3 includes Brazil, Cyprus, Greece, Puerto Rico, Romania, Singapore, and Thailand; it represents 9.51% of the analyzed population;
- Cluster 4 includes Czechia, South Korea, Mongolia, and Vietnam; it represents 5.78% of the analyzed population;
- Cluster 5 includes Andorra, Australia, Canada, Germany, the United Kingdom, Netherlands, New Zealand, Uruguay, and the United States; it represents 18.89% of the analyzed population;
- Cluster 6 includes Bolivia, Colombia, Ecuador, Guatemala, Mexico, Nicaragua, Peru, and Venezuela; it represents 12.16% of the analyzed population;
- Cluster 7 includes Argentina, Chile, Russia, Serbia, Slovakia, and Ukraine; it represents 7.27% of the analyzed population;
- Cluster 8 includes Armenia, Ethiopia, Libya, Moldavia, and Zimbabwe; it represents 6.34% of the analyzed population;

The following Figure 4 reports the aforementioned visualization of the multidimensional scaling of the average indices values for the nine clusters (see paragraph 4).

Finally, here follows the self-organizing map (Figure 5) of the single countries aimed at comparing the goodness of the hierarchical clustering (see the end of paragraph 4).

5.1. Results' interpretation

To understand the cultural basis that supports the clustering process, each cluster underwent an analysis, at first using the indices used to generate them, and the using other indicators that could provide better insights (see paragraph 3.2).

5.1.1. Cluster 0

Cluster 0 is predominantly Asian (in the broad sense, 78%; African for the remaining 22%), 78% of these clusters present authoritarian or hybrid regimes ('Potemkin democracies'); on average, the cluster is sparsely populated (around 40 million inhabitants) and poor (GDP per capita around $5000).

Defiance is the lowest index (0.16), but is also in line with the low values of the other clusters (between 0.11 and 0.44); the combination of *Relativism* (0.60) and the other indices (between 0.26 and 0.46) distinguishes cluster 0 from all the others. Cluster 0 is thus characterised by a permissiveness (justifiability) of dishonest behaviours (*Relativism*), by a rejection of certain civil liberties (*Choice*) and by a heartfelt religiosity (*Disbelief*, 57% Muslims); in addition, great value is placed on authority or, more generally, on a sense of belonging (*Defiance*).

Cluster 0 appears isolated on the edges of the multidimensional scaling map (MDS), and its points are confirmed close together on the self-organising map (SOM). In these two plots, the closest (similar) clusters are 1 and 3.

5.1.2. Cluster 1

Cluster 1 is almost entirely Asian (in the broadest sense, 82%), consisting of 81% authoritarian countries and hybrid regimes; on average, the cluster is very populous (around 230 million inhabitants) due to the presence of India, but poor (GDP per capita around $4000).

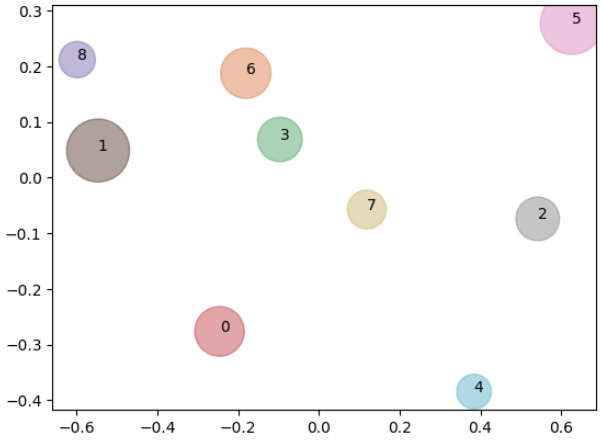


Figure 4. Multidimensional scaling (MDS) of clusters, proportional to the number of their respondents

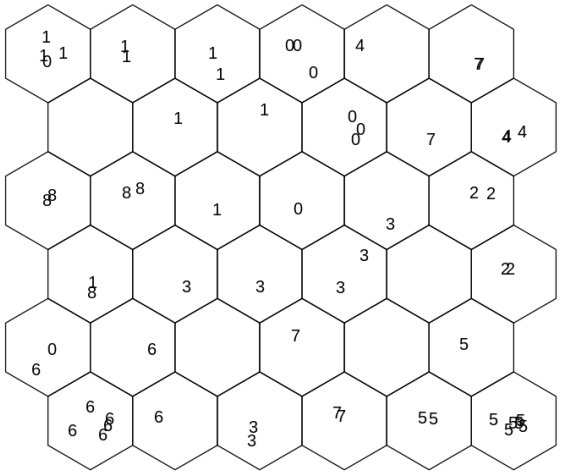


Figure 5. Self-organizing map (SOM) of countries identified by their cluster label

Choice (0.14) is the lowest index (also in comparison with the other clusters), followed by *Defiance* (0.16) and *Disbelief* (0.21); the remaining indices, ranging from 0.31 to 0.40, uniquely distinguish cluster 1 by the absence of indices greater than 0.5 (i.e., closer to 1 than to 0). Cluster 1 is thus interpreted in the same way as cluster 0, but is distinguished from it by greater intransigence against freedom of choice (*Choice*), greater religious fervour (*Disbelief*) and, above all, a greater sense of righteousness (*Relativism*, 0.37). Additionally, around 76% of the population identify themselves as Muslims, and the cluster has one of the lowest level of average education (1.7 out of 3).

Cluster 1 appears at the edge of the MDS and its points are grouped (but not compactly) in the SOM. Both views show that cluster 8 is the most similar, followed by 0 and, less clearly, 3.

5.1.3. Cluster 2

Cluster 2 is entirely Asian (100%); on average, it is the most populous (around 230 million inhabitants) due to the presence of China, and the second richest (GDP per capita around $31,000).

Relativism (0.31) and *Voice* (0.32) are the lowest indices, followed by *Scepticism* (0.35), which is also the lowest among the clusters; on the other hand, *Autonomy* (0.69) and *Disbelief* (0.71) have the highest value in cluster 2 (predominance of non-religious individuals, around 68%). *Defiance* (0.45) is the highest among clusters, meaning that resistance to traditional authority, albeit still moderate, is higher in this cluster than anywhere else. Cluster 2 is characterised by moderate

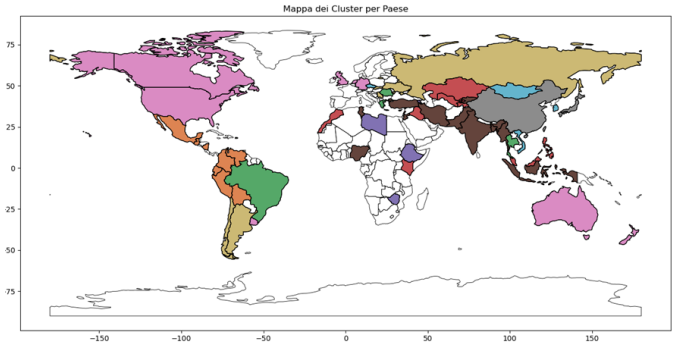


Figure 6. World map highlighted by cluster membership (see Appendix 1)

trust in state institutions (*Scepticism*), the importance attached to children's education and maturity (*Autonomy*) and a strong disinterest in religion (*Disbelief*). Education levels are fairly evenly distributed, with approximately 30% of the population in each educational category (low, medium, and high), with an average level of 2 (medium level of education).

Cluster 2, marginal in the MDS, is in the opposite position to 0 and 1, and its (few) points in the SOM are compact, although not diametrically opposed to those of 0 and 1. Quite isolated from any other cluster, the MDS shows 4 and 5 as closest to 2, and the SOM confirms this.

5.1.4. Cluster 3

Cluster 3 consists of 86% flawed democracies; on average, the cluster is sparsely populated (around 50 million inhabitants) and moderately wealthy (GDP per capita around $24,000).

Defiance (0.24) and *Relativism* (0.27) are the only indices below 0.34, just as *Equality* (0.65) is the only one above 0.44; with the exception of *Equality*, all the indices in cluster 3 have average values compared to their distribution among clusters. Cluster 3 is thus characterised by a positive attitude towards gender equality (*Equality*), moderate religiosity (*Disbelief*) and a strict integrity and sense of righteousness (*Relativism*). 27% of this group define themselves as Orthodox, 19% as Buddhists and 16% as Catholic, leading to a high variety of religions.

Cluster 3 is not isolated in the MDS, being one of the most central, nor is it compact in the SOM (Brazil and Puerto Rico in particular appear isolated); it is close to 6 and the scattered points of 7.

5.1.5. Cluster 4

Cluster 4 is 75% Asian; on average, the cluster is sparsely populated (around 40 million inhabitants).

Defiance (0.32) and *Voice* (0.33) are the only indices below 0.43; however, *Disbelief* (0.69) and *Relativism* (0.68) stand out, both because they are obviously higher than the others (below 0.52) and because *Relativism* is the highest among clusters. Cluster 4 is uniquely characterised by religious disinterest (*Disbelief*, 64% atheists) and loosest sense of justice and righteousness (*Relativism*). *Autonomy* (0.5) is just moderate, though children have statistically high levels of instruction (average 2.2 out of 3).

In the MDS, cluster 4 is marginal and isolated, diametrically opposed to 8; in the SOM, it is loosely and indistinctly mixed with some items from 7.

5.1.6. Cluster 5

Cluster 5 is 44% European, 33% American and 22% Oceanic; it consists of 87% full or otherwise flawed democracies. On average, the cluster is sparsely populated (around 60 million inhabitants) and the richest (GDP per capita around $47,000).

Defiance (0.29) and *Relativism* (0.33) are the lowest indices; *Equality* (0.80), *Choice* (0.68) and *Disbelief* (0.66) are the highest; *Equality*, *Choice* and *Voice* (0.51) are the highest of all clusters. Overall, cluster 5 stands out for the highest levels of agreement with gender equality

(*Equality*) and civil liberties (*Choice*), otherwise placing similarly at 2 with regard to religion (*Disbelief*, about 50% atheists), and child-rearing (*Autonomy*). Countries in this cluster also have the lowest average number of children per family (2.8) and the highest educational levels (average 2.4 out of 3).

Cluster 5 is isolated on the fringes of the MDS and the SOM, where it appears compact; cluster 2, although distant, is the closest and appears so in both the MDS and the SOM.

5.1.7. Cluster 6

Cluster 6 is entirely South American, 63% of which are hybrid or otherwise authoritarian regimes. On average, the cluster is among the least populous (around 40 million inhabitants) and poorest (GDP per capita around $7000).

Defiance (0.13) is the lowest index (the only one below 0.29) and the second lowest overall; *Scepticism* (0.65) is the highest, even compared to the other clusters, followed immediately by *Equality* (0.63) and detached from the others (all below 0.42). Cluster 6 is particularly characterised by unconditional respect or fascination for the authorities (*Defiance*) and, at the same time, by the lowest level of trust in state institutions (especially the armed forces, *Scepticism*); it should also be noted that it has a clearly egalitarian view of the role of women (*Equality*). Moreover, *Autonomy* (0.31) has the second-lowest value among clusters, but one of the highest number of children per household (4.7). 75% of the cluster's members have the lowest level of education, with (average 1.92 out of 3).

Cluster 6 is closer to 3 than to any other cluster in the MDS, and remains compactly distinguishable in the lower left-hand corner of the SOM, where the points of 3, however scattered, are indeed the closest.

5.1.8. Cluster 7

Cluster 7 is made up of two-thirds flawed democracies and is on average sparsely populated (around 40 million inhabitants) and moderately poor (GDP per capita around $11000).

Defiance (0.28) is the only index below 0.35, while *Equality* (0.60) and *Scepticism* (0.54) are the only ones above 0.46. On the whole, cluster 7 occupies average positions compared to the others, and is therefore not very distinctive; following the description of cluster 6, it is only possible to observe rather low expectations and trust in state institutions (*Scepticism*) and a more egalitarian mentality with regard to the role of women (*Equality*). It is difficult to explain the tendency to be more open to individual freedoms (*Choice*, 0.43), such as homosexuality, than e.g. 6 (0.29). The average level of education in this cluster is comparably high (average 2.3 out of 3).

Cluster 7 is central in the MDS and therefore equidistant from several other clusters, the closest being still 3. In the SOM, the points of 7 are scattered and rather central, consistently mixed with those of 3.

5.1.9. Cluster 8

Cluster 8 is 60% African and 40% Asian, composed by 75% authoritarian or otherwise hybrid regimes; on average, it is one of the least populous clusters (around 30 million inhabitants) and a poor one (GDP per capita around $5000).

The cluster is immediately recognisable by some of the lowest indices. *Defiance* (0.11), *Relativism* (0.15), *Choice* (0.15) and *Disbelief* (0.17) are the lowest and, with the exception of *Choice*, the absolute lowest indices among the clusters; *Equality* (0.53) is the highest index. Overall, cluster 8 is defined by the greatest respect (or awe) for authority (*Defiance*), the greatest intransigence towards crime (*Relativism*), the greatest rejection of progressive individual freedoms of choice (*Choice*), and the greatest importance given to religion (*Disbelief*, 47.8% Muslims and 29% Orthodox). The education level is the second-lowest (average 1.9 out of 3)

Cluster 8 is isolated on the edge of the MDS, close to 1 and 6 and opposite, among others, 5; in the SOM it occupies a circumscribed and marginal position on the left, between 1 and 6.

6. Further analysis: impact of socioeconomic and political factors

6.1. Regime type

The analysis of the indices across different regime types (authoritarian, hybrid, flawed democracy, and full democracy) reveals, as shown in 10.1, Appendix 2, distinct patterns in societal values concerning authority, personal freedom, and civic engagement. On average, authoritarian regimes exhibit the highest level of awe for authority (lowest *Defiance*), the lowest consideration for women roles in society (lowest *Equality*), and the most radical refusal of individual liberties of choice (lowest *Choice*); this is especially true when compared to the corresponding average values of fully democratic countries, which are the highest in all of these categories. Full democracies emerge as the most progressive, both in terms of the aforementioned indices, and of others such as *Voice* and *Disbelief*. Flawed democracies and hybrid regimes present a more nuanced picture, occupying an intermediate space, albeit often more aligned with full democracies rather than authoritarian regimes.

6.2. GDP per capita

The correlation matrix (10.2 in Appendix 2) reveals several key insights into the relationships between GDP per capita and various sociocultural indices. Notably, there is a strong positive correlation between GDP per capita and *Choice* (0.744), and between GDP per capita and *Equality* (0.668), suggesting that wealthier nations tend to embrace a progressive stance on social matters (e.g. homosexuality, abortion, and role of women in society). Similarly, GDP per capita is positively associated with *Disbelief* (0.641), indicating that higher economic development often corresponds to lower religious influence in individuals' lives.

Furthermore, GDP per capita is moderately correlated with *Autonomy* (0.581), *Defiance* (0.586), and *Voice* (0.506). These findings highlight the intricate interplay between economic prosperity and progressive societal values, suggesting that wealthier nations are also likely supporters of individual freedoms, gender equality, and reduced religious adherence. Conversely, *Relativism* and *Scepticism* have (very) weak negative correlations with GDP per capita (-0.187 and -0.196, respectively), suggesting that economic wealth has less, if any, influence on moral flexibility and mistrust in public institutions.

7. Conclusion

Overall, we can confidently say that some clusters of countries clearly emerge. This is the case of cluster 5, the prototype of the affluent Western world: liberal, progressive and educated; an Anglophone bias is undeniable, so many countries of the Commonwealth of Nations (former British Empire) are included, while others that have always been culturally independent, such as India, are excluded. Also well defined, in the opposite direction, is the small cluster 8: repressive and intolerant, openly Muslim or Orthodox, and conservative. Similar to cluster 8, and therefore well defined, is cluster 1: with slightly less radical positions, it is still authoritarian and intransigent, and fervently Muslim. Clusters 0 and 4 are also characteristic: cluster 0, mostly Asian, including former Soviet republics, is authoritarian and poor, and has the least sense of justice; on the other hand, the small cluster 4, including the 'Western' Far East, is also more permissive, but absolutely non-religious. Clusters 2 and 6 are still quite recognisable: the all-Asian cluster 2 is wealthy and includes China; it is trusting, atheistic and devoted to education and child-rearing. Cluster 6 is Spanish colonial America, from Mexico to Peru: poor and politically unstable, subject to the power of authority, which it mistrusts, it is also the least educated cluster. Cluster 3 is diverse and ambiguous; it certainly includes European countries that are affluent (Greece, Cyprus, Romania) but not sufficiently wealthy, progressive or atheistic to fit into the Anglo-Western cluster (5). Cluster

7 includes Russia and Argentina, among others, and is difficult to classify.

7.1. Research questions

Through this analysis, we were able to identify 9 distinct clusters of countries, each characterized by specific cultural and socioeconomic values. The research questions guiding this analysis focused on three central inquiries:

How do cultural values vary across different countries? This study revealed significant global variations in cultural values, particularly in terms of authority, religion, autonomy, and gender equality. Countries in cluster 0 (e.g., Morocco and Iraq) are marked by a strong respect for authority and a deep adherence to religious values, reflecting more traditional societies. In contrast, nations in cluster 5 (e.g., the United States and Canada) exhibit high levels of gender equality and a diminished role for religion, illustrating the influence of progressive values. These differences demonstrate how deeply embedded societal norms are influenced by historical and religious traditions as well as the broader political structures within each nation.

Is it possible to map the similarities and differences between regions and countries? By employing hierarchical clustering techniques, this study effectively mapped the cultural proximity and divergence between countries. The analysis highlighted how countries can share common cultural attributes despite being far apart, while others, although geographically close, may diverge significantly. For instance, cluster 3 (Brazil, Greece, Singapore) displays a moderate balance between tradition and modernity, while cluster 2 (China, Japan, Taiwan) stands out for its high levels of *Autonomy*. The clustering revealed that geography does not always dictate cultural alignment; instead, shared historical, political, and religious experiences play a larger role in shaping the values of different nations.

How do socioeconomic and political factors impact societal attitudes and values? The analysis underscored the pivotal role of socioeconomic and political factors in shaping cultural values. Countries with higher GDP per capita and democratic regimes showed a pronounced tendency towards progressive values, including greater support for gender equality, civil liberties, and scepticism towards authority. Conversely, authoritarian countries, often with lower economic development, exhibited stronger adherence to religious values, respect for authority, and conservative views on social issues. For example, clusters 1 (Pakistan, Turkey) and 8 (Ethiopia, Zimbabwe) exhibit deeply entrenched religious values and social hierarchies.

The relationship between education and cultural values was also evident. Countries with higher levels of education generally support more progressive values, such as gender equality and freedom of choice. On the other hand, nations with larger families tend to adhere more closely to traditional values.

This research offers a comprehensive overview of global cultural trends, answering the research questions by demonstrating that cultural values are not only shaped by religious and historical factors but are also significantly influenced by the political and economic contexts of each country. The insights gathered from the clustering analysis illustrate the complexity of global cultural dynamics, providing a deeper understanding of the forces that shape societal development across different regions. Ultimately, this study demonstrates that a country's culture is the result of a nuanced interaction between religion, politics, education, and economic conditions, offering valuable insights into the foundations of modern societies.

8. Contact us

You can contact us through:
- a.brunazzi@campus.unimib.it
- a.dellabeffa@campus.unimib.it
- d.lepre@campus.unimib.it
- s.turroni@campus.unimib.it

References

[1] WVS, *World values survey*, 2024. [Online]. Available: https://www.worldvaluessurvey.org/wvs.jsp.

[2] Wikipedia, *Ronald inglehart*, 2024. [Online]. Available: https://it.wikipedia.org/wiki/Ronald_Inglehart.

[3] Wikipedia, *Christian welzel*, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Christian_Welzel.

[4] WVS, *Indices*, 2024. [Online]. Available: https://www.worldvaluessurvey.org/WVSContents.jsp.

[5] WVS, *Seventh wave*, 2024. [Online]. Available: https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

[6] W. bank, *Gdp per capita (current us)*, 2024. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

[7] Wikipedia, *Stati per popolazione e continente*, 2022. [Online]. Available: https://it.wikipedia.org/wiki/Stati_per_popolazione.

[8] Wikipedia, *The economist democracy index*, 2024. [Online]. Available: https://en.wikipedia.org/wiki/The_Economist_Democracy_Index.

[9] Scikit-learn, *Cluster*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html.

[10] Scikit-learn, *Agglomerativeclustering*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.htm.

[11] Scikit-learn, *Visualization of cluster hierarchy*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#visualization-of-cluster-hierarchy.

[12] Scikit-learn, *Plot hierarchical clustering dendrogram*, 2024. [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py.

[13] Scikit-learn, *Multi-dimensional scaling (mds)*, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/manifold.html#multi-dimensional-scaling-mds.

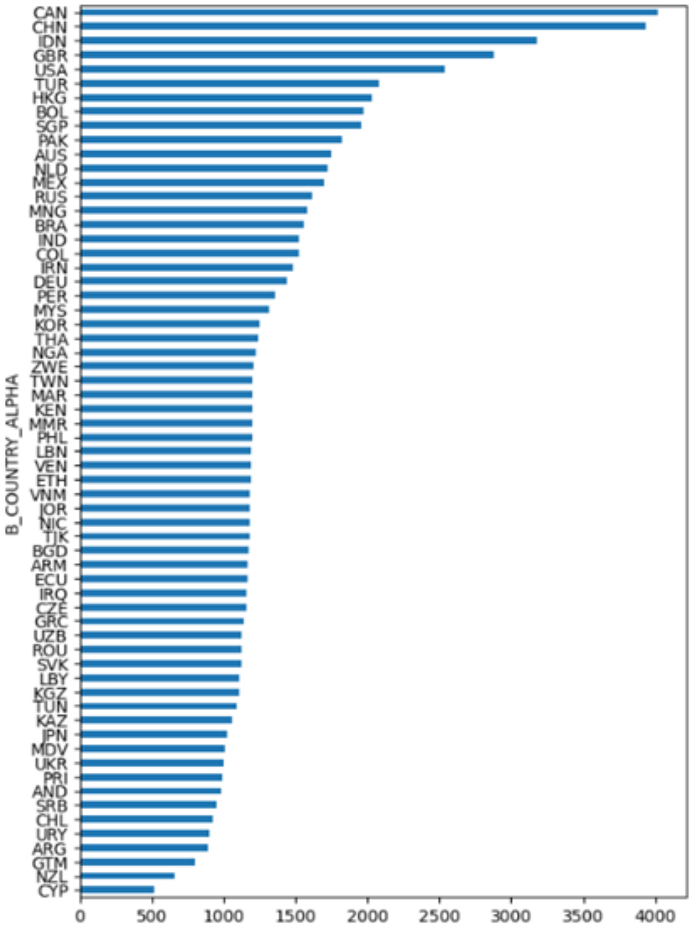
[14] R. Wehrens and J. Kruisselbrink, *Kohonen: Supervised and unsupervised self-organising maps*, 2023. [Online]. Available: https://cran.r-project.org/web/packages/kohonen/kohonen.pdf.

[15] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

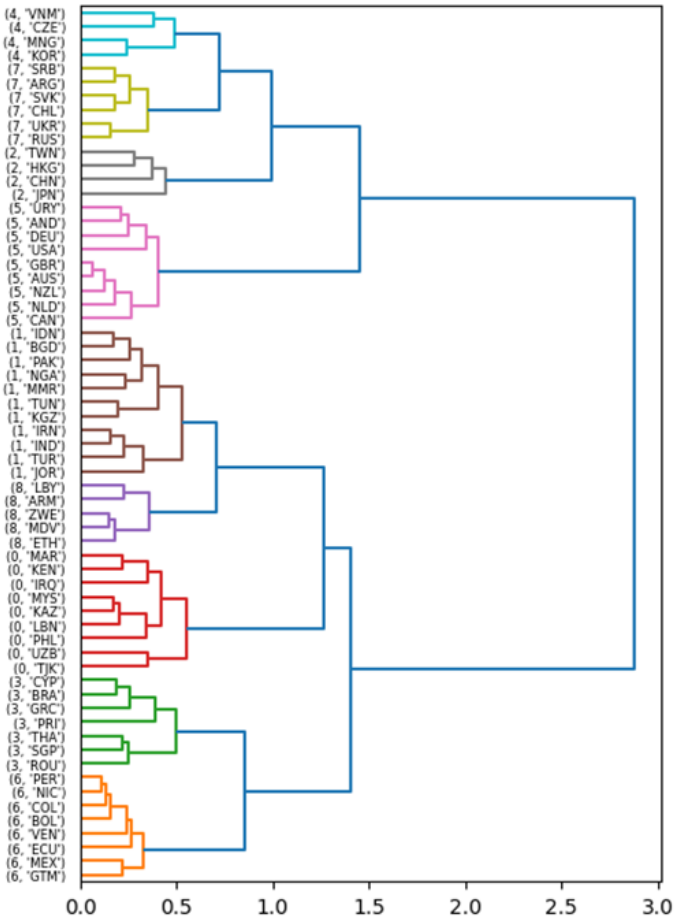
[16] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering", *Entropy*, 2021.

9. Appendix 1: images

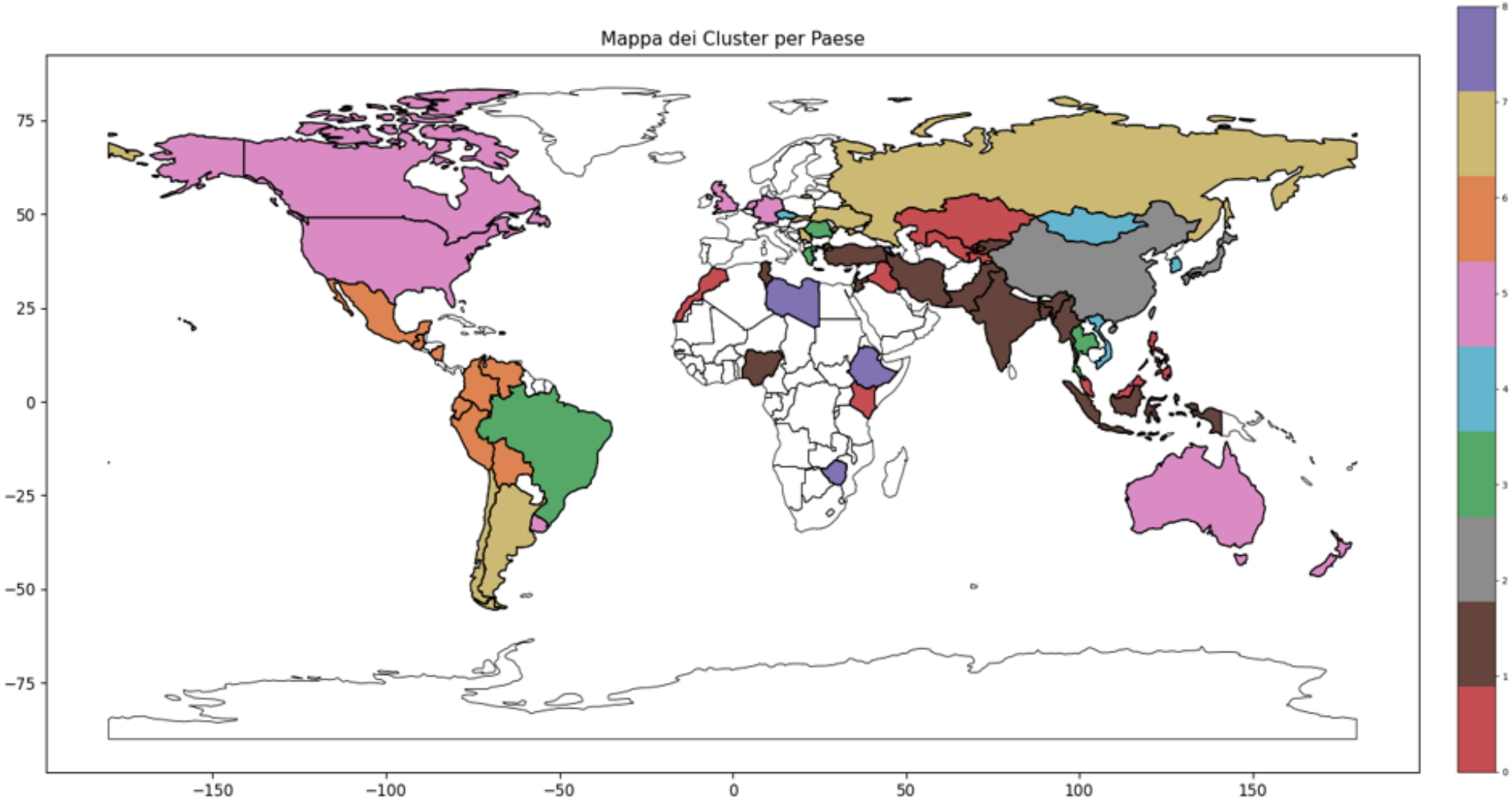
9.1. Distribution of respondents by country



9.2. Dendrogram



9.3. World map


Mappa dei Cluster per Paese
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
8
7
6
5
4
3
2
1
0

10. Appendix 2: tables

10.1. Regime type

	Defiance	Disbelief	Relativism	Scepticism	Autonomy	Equality	Choich	Voice
Authoritarian	0.159094	0.330826	0.396590	0.400821	0.386690	0.456688	0.207333	0.303393
Flawed democracy	0.242278	0.385011	0.430978	0.456746	0.444683	0.590320	0.369543	0.387474
Full democracy	0.343912	0.640854	0.341158	0.385497	0.577553	0.728224	0.605134	0.441584
Hybrid regime	0.181510	0.297031	0.403657	0.523864	0.375619	0.551146	0.263117	0.347508

10.2. GDP per capita

	GDPpc	Defiance	Disbelief	Relativism	Scepticism	Autonomy	Equality	Choice	Voice
GDPpercapita2022	1.000000	0.586052	0.641884	-0.186984	-0.195592	0.581160	0.668036	0.744416	0.506176
Defiance	0.586052	1.000000	0.717891	0.108146	-0.145461	0.709424	0.366063	0.588374	0.209120
Disbelief	0.641884	0.717891	1.000000	0.165090	-0.143807	0.684757	0.525475	0.765175	0.363937
Relativism	-0.186984	0.108146	0.165090	1.000000	0.009851	0.110927	-0.202649	0.077417	-0.018216
Scepticism	-0.195592	-0.145461	-0.143807	0.009851	1.000000	-0.418716	0.259074	-0.016743	0.135554
Autonomy	0.581160	0.709424	0.684757	0.110927	-0.418716	1.000000	0.301037	0.488977	0.230682
Equality	0.668036	0.366063	0.525475	-0.202649	0.259074	0.301037	1.000000	0.812301	0.662791
Choice	0.744416	0.588374	0.765175	0.077417	-0.016743	0.488977	0.812301	1.000000	0.668453
Voice	0.506176	0.209120	0.363937	-0.018216	0.135554	0.230682	0.662791	0.668453	1.000000

11. Appendix 3: other metrics breakdown

Q274: Have you had children? If yes, how many?

- 0 – No children
- 1 – 1 child
- 2 – 2 children
- 3 – 3 children
- 4 – 4 children
- 5 – 5 children
- 6 – 6 children
- 7-25 – 7 or more children
- -1- – Don't know
- -2- – No answer
- -4- – Not asked
- -5- – Missing; Not available

Q275R: Highest educational level (recoded into 3 groups)

- 1 – Lower (ISCED 0, ISCED 1, ISCED 2)
- 2 – Middle (ISCED 3, ISCED 4)
- 3 – Higher (ISCED 5, ISCED 6, ISCED 7, ISCED 8)
- -1- – Don't know
- -2- – No answer
- -4- – Not asked
- -5- – Missing

Q289: Do you belong to a religion or religious denomination? If yes, which one?

- 0 – Do not belong to a determination
- 1 – Roman Catholic
- 2 – Protestant
- 3 – Orthodox (Russian/Greek/etc.)
- 4 – Jew
- 5 – Muslim
- 6 – Hindu
- 7 – Buddhist
- 8 – Other Christian (Evangelical/Pentecostal/Fee church/etc.)
- 9 – Other