ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

July 5, 2020

Clustering algorithms on World Development variables in reduced dimension

---

Bachelor Thesis Bsc[2] Econometrics/Economics[1]

Lara Malinov - 449285

Supervisor: Velden, M. van de

Second assessor: Cavicchia, C.

Abstract

The past century saw a large association of well-being and development with the growth of production of goods. More recent indicators, such as the Human Development Index (HDI), changed how the world development is viewed. Nonetheless, it has also been subject to criticism as it does not consider aspects such as sustainability of political governance. This research aimed to make a categorization of the world development in 2018 using multidimensional country variables. As some statistics are unavailable for some countries and to remedy the curse of dimensionality, a probabilistic principal component analysis was applied to preserve the maximum variance of the development of countries. K-means, Spectral and Agglomerative Hierarchical clustering were used in combination with various k-selection criterions. Agglomerative Hierarchical clustering with three clusters proved to make country clusters with the strongest distinctions. The clusters are similar to the categorization of the development of countries by the International Monetary Fund (IMF). However, many Eastern European countries were clustered with the developed nations whereas more South Asian countries were clustered with the least developed countries as defined by the IMF.

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Simon Kuznets (1934) developed the Gross Domestic Product (GDP) after the Great Depression of the 1930s started which is the aggregated added value of the production of a nation. The measure was developed as a way to evaluate the growth of industries and see how policies would affect the production of the country, considering that the US unemployment rate reached 24,9% in 1933. It was meant to capture the production of individuals, companies and the government to establish the health of an economy. Following the Bretton Woods conference in 1944, the World Bank and the International Monetary Fund declared that the GDP would be used to evaluate different economies and compare them. The growth of aggregated added value of production became associated with increase in well-being. The reconstruction of the world and the thrive of the economies in the second half of the 20th century led to an increase in goods, with the wide adoption of televisions and other home appliances. More goods were synonymous with a better life. However, Kuznets warned in 1934 that 'the welfare of a nation can scarcely be inferred from a measurement of National income'. Indeed, the welfare of a population also depends on security, education, functioning of the government, physical and mental health of the citizens etc. that cannot be inferred from a measure of production. It can be argued that the growth of goods and the well-being of the population are intertwined as people who are employed can satisfy their basic needs and spend extra income that in a way finances the employment of others who can increase their standard of living as well. But in an era driven by technology and services whose prices cannot be accurately estimated, the GDP may be inaccurately used in political and monetary decisions and insufficient to assess the well-being of a nation.

While the limitations and the wide usage of the GDP still prevail, economists have tried to understand the economics of development in order to advise nations on their directory. In 1990, Amartya Sen developed the Human Development Index (HDI) which is a geometric mean between the Life Expectancy Index (LEI), the Education Index (EI) and the Income Index (II). It has been mainly used by the United Nations Development Programme (UNDP) to classify the countries, but it has also been subject to criticism. Indeed, the EI is an average between the mean years of education and the expected years of education, while in reality the mean years of education taken alone may be more representative of the general education of the population. Similarly, the LEI be 1 if the average age of the population reaches 85 and the II does not take into account inequalities in income. Despite, it has been fundamental in shifting the focus of governments from income policies

to well-being policies. For example, Bhutan declared in 1972 that they would aim to increase their Gross National Happiness (GNH) index instead of their aggregated value of production. The GNH considers aspects such as health, education, public governance, psychological well-being and living standards. Although other nations have also focused their policies on those aspects, the GDP is still the primary indicator used to evaluate the general health of an economy and the effectiveness of policies.

Nowadays, the United Nations and the International Monetary Fund (IMF) classify the countries as 'developed', 'developing' or 'least developed'. This classification, however, relies on the GDP, the degree of industrialization and other monetary indicators, that do not entirely reflect the living population. Similarly, the HDI is only taking three indices to reflect the development of a country. Countries like Oman and Qatar have a high income per capita due to their resources which classifies them as countries with very high development, but they also have among the lowest mean years of education in their category. As such, a metric using a few indicators cannot reflect the complexity of development. Further research is necessary to motivate the use of alternatives not only to reflect the development of countries in a simple way but to change the way well-being is associated with production. This also includes finding a way to process and convey the information in an efficient way.

Moreover, the improvements in computer science and statistics have led to the development of sophisticated methods that can handle and process more data as well as reveal tendencies and relations between them. Classification methods can be used to understand how countries are categorized. However, these analyses only depend on how the countries were classified and what variables were used. A way to categorize data that is not labeled is clustering which consists of grouping data points based on their similarity or closeness. This paper will attempt to cluster the development of countries in order to reconcile the need of using more variables and more elaborate methods to differentiate countries. In essence, this research aims to answer the folowing question:

How can the development of countries be clustered using multidimensional data?

Various statistics on the health, education, economy, sustainability and governance of countries are used. However, the lack of transparency of some countries and the data gathering effectiveness represent a hurdle in most research which has also motivated the use of fewer indicators. To circumvent, this it is possible to use data transformation techniques or ways to estimate the missing information. Therefore, a probabilistic principal component analysis (PPCA) is performed which

can preserve the pairwise variances in reduced dimensionality as well as handle missing data. This is motivated by the need to include most countries in the analysis and use as many relevant variables as possible. Different clustering methods will be used namely k-means, spectral and agglomerative hierarchical clustering in order to see how the variables of countries in reduced dimension should be clustered. These method require a pre-specified number of clusters and consequently, different selection criterions are used to determine it. While many indicators are based on the sum of squares of the clusters (CH, DB, Sil), Wang (2010) defines cross-validation methods to determine the optimal k by the least cluster instability. These will, together with other selection criterions, determine the optimal number of clusters dependent on the clustering technique. Finally, the three different clusterings will be analyzed and compared in order to determine what method can best illustrate the different categories of development using variables in reduced dimension.

First, a summary of the existing literature is provided, followed by a description of the data used and the application of PPCA. Next, the different clustering approaches are discussed as well as the different criterions to determine the number of clusters. Afterwards, the results are outlined and compared. Finally, a conclusion and discussion of the research is presented.

2 Literature Review

Assessing the development of a country is a sensitive matter. Due to its association with the production of goods, economists have tried to find more accurate measures and understand what is important towards the development of a country. With that comes a lot of responsibilities as they can indirectly influence policy making. Stephen Morse (2013) states that indicators are "simplifying complexity" so that they are understandable not only to politicians and experts but more importantly to the general population. He warns that the creators of indicators have "great power as they can influence high policy makers and managers" while the users do not fully know how they are calculated or what is included in the metric. This reflects the moral responsibilities that accompany the making and usage of indicators. Monni and Spaventa (2013) argue that policy makers choose their goals and then evaluate their policies using indicators. In their opinion this ultimately places "decision making power in the hands of economists and their theories" (p.229). Therefore, the research of economists is fundamental. It can change how development is viewed and what measures are appropriate to elevate the well-being of a nation. Overall, how an economists performs a research and creates indicators can have a great influence on policy making and the

conceptualization of development.

The creation of the HDI was revolutionary in some regard as it considered education and longevity as well as income of the population. It answered the criticism that growth in production does not equal growth in well-being and put the people back in the center of policy decisions (Bagolin, 2004). Nonetheless, it has itself been subject to criticism. On the one hand, Bagolin (2004) finds that the proponents see the HDI as an advancement compared to earlier measures because it is more relevant and helpful in public policy decision, as it considers multidimensional aspects of development. On the other hand, the opponents find that the HDI is misrepresenting development as it is restricted to the socio-economic aspect of life and does not include considerations on governance and civil liberties that also contribute to well-being. The construction of the metric is also criticized as the Gross National Product (GNP) is subject to approximations in developing countries and the life expectancy is not available in most of the less developed countries. They argue that there needs to be improvements in information gathering in order to be able to use such a metric. However, this is a process that can take time and what may be more relevant is how much we can do with the data we have at our disposal. This research is motivated to address the lack of transparency of some countries as well as the need to create an assessment of development that is multidimensional.

Meanwhile, an important consideration is what variables should be included in such an analysis. Hicks and Streeten (1979) at the time of their research found a lot of contradiction in the literature where the GNP is said to be correlated with economic and social indicators, but changes in GNP are less correlated to changes in basic needs. In their research they find that a combination of social and economic indicators are better at finding relationships with the GNP than social indicators alone, due to the fact that the relationship between social indicators and the GNP is non-linear. Therefore, social indicators will be used in combination with economic variables in this research as they can complementarily reflect the socioeconomic sphere of development. Reig-Martínez (2013) made a human Well-being Composite Index (WCI) that includes various indices of development such as income inequality, basic needs fulfillment, gender gap or governmental effectiveness. In his metric, every index can have more or less weight in the score depending on the country. This is particularly interesting as using multiple variables in his analysis showed that there are large differences in Europe and around the Mediterranean Basin. Especially, his research shows that countries that have a similar composite index can have different individual indices with respect to gender equality or governance for example. This shows that development is more complex than it

has been previously established. Countries with similar degree of development may be drastically different in some aspects and accounting both similarities and discrepancies is necessary in an analysis of the development of countries.

To make a categorization with multiple variables, cluster analysis can be used. It is a technique that groups together observations that are similar based on distance measures or affinities for example. For instance, Yorulmaz (2016) clusters health and socioeconomic variables of countries and compares it with the existing HDI categorization. He uses data on various aspects such as the CO2 emissions, share of parliament seats held by women or internet users rate. From the output of the dendrogram of agglomerative hierarchical clustering, he decides to model four clusters using k-means. The hierarchical clustering agglomerative groups together observations based on a linkage criterions. Since a dendrogram is tree shape, one can see how all the observations can be narrowed down to a couple of clusters. k-means uses centroids to make clusters. Each point is assigned to the cluster that has the closest centroid based on the Euclidean distance. The centroids are recomputed at every iteration until the clusters are stable. In the end, Yorulmaz (2016) finds that countries are in different categories than in the HDI categorization, although "the main characteristics of these clusters are similar to the relevant HDI country categories"(p.5). Similarly, Mrázová and Dagli (2008) use a fuzzy c-means algorithm on World Bank data to determine the membership degree of countries to a cluster. Fuzzy c-means is similar to k-means only that a data point can belong to multiple clusters and the membership degree to the clusters is defined. To determine the number of clusters, they use the partition entropy, partition coefficient and proportion exponent which are k selection criterions used in fuzzy cluster analysis. They find that the variables should be grouped in 7 clusters. Additionally, they show that the characteristics found in each cluster differ from the guidelines proposed by the World Bank to classify countries. This suggest that using different variables in clustering can help make more accurate clusters as well as identify patterns in how developed a country is or can be, when considering the membership degree to a cluster. Furthermore, Mylevaganam (2017) applies a principal component analysis (PCA) on the same individual indices of the HDI namely the Life Expectancy Index (LEI), the Education Index (EI) and the Income Index (II). Mylevaganam uses this technique such that the variance of the multidimensional data is captured along the principal axes. Next, he applies k-means clustering and he uses the elbow, the silhouette and the GAP statistics to determine the number of clusters. He finds that the countries should be clustered in four development categories. However, he suggests that further clustering algorithms should be considered. We can see that k-means clustering is a

popular choice used in clustering the countries. Therefore, it is interesting to see how other models may perform in comparison. Moving along, we will perform a similar analysis as Mylevaganam only with more variables on diverse aspects of development.

In summation, the way economists construct indices can have a high impact on public policy decisions. This motivates the use of more variables and more sophisticated methods as the GDP and HDI can scarcely reflect the development of a country. As such, this analysis will focus on clustering multiple variables to make a categorization of the development of countries.

3 Data

Since the aim is to have a multidimensional view of the development of countries, many variables are gathered to make the analysis. The United Nations Development Programme (UNDP) provides a yearly Human Development Report and the data from the 2019 report includes indicators on 195 countries using the information of 2018 or the last available data from surveys. The World Bank (WB) has an additional 1141 variables on 217 countries from 2018 in their 'World Development Indicators' database. The variables of UNDP and WB concern different aspect of a nation such as education, health, employment, basic needs fulfilment and gender inequalities of a population as well as indicators on the socioeconomic and environmental sustainability of an economy. These variables characterize the population and the economic activity, however another aspect of development regards the civic rights and liberties as well as the governmental effectiveness. The Economist Intelligence Unit publishes yearly a World Democracy Report where they score and rank countries based on the civil liberties, the functioning of the government, the electoral process and pluralism as well as the political participation and culture. To add a governmental and civic rights consideration to development the indices from 2018 are retrieved. A last important factor that has been gaining in importance the last decades is the environment. Economic growth cannot be sustained if it relies on depleting resources and harming the health of civilians. For this purpose Columbia University and Yale University have jointly with the World Economic Forum designed Environmental Performance indices. These are divided into two dimensions to inform on the performance of countries, namely the environmental health and the ecosystem vitality. The environmental health dimension considers aspects such as air quality and water sanitation while the ecosystem vitality dimension focuses on aspects like air pollution, biodiversity and forest loss. In sum, these four sources of information reflect the socioeconomic, health, educational, sustainable and political sphere of the development

of countries.

The aim is to use multiple variables to make a clustering of development. However, the more features we add per country, the more the dimensionality increases. This means that distances becomes exponentially large. Since many clustering algorithms are based on distance measures such as the Euclidean distance, the data should be reduced to a lower dimension. A popular option is a Principal Component Analysis (PCA). However, this would require to have a complete data set. Under this condition, too many variables would be lost and many countries would be left out of the analysis as some countries are less transparent and in other cases the data gathering process is not efficient. An alternative is to use Probabilistic Principal Component Analysis (PPCA) that preserves the dominant correlations and that in combination with an Expectation Maximization (EM) algorithm can handle up to 15 % of missing data. Therefore, all the variables having more than 15% of country data missing are excluded from the analysis as well as the countries and islands that have few available data such as Lichtenstein or Palau. As a result, the final data set contains 149 variables on 186 countries. A complete overview can be found in the Appendix.

A probabilistic approach to PCA was independently established by Tipping and Bishop (1999) and Roweis (1998). The method assumes that the distribution of the observed data is generated by a normal distribution in lower dimensions. The parameters of the distribution of the latent variable are approximated by maximum likelihood. The Expectation Maximization (EM) algorithm is used to iteratively approximate the parameters and is computationally efficient. The principal components are the eigenvectors of the covariance matrix for which the likelihood function is maximized. This is particularly useful as it is computationally efficient, it does not require a complete data set and it can remedy the curse of dimensionality for the analysis.

Three principal components are used to reflect various development variables in reduced dimensionality. The three principal components account for 40% of the explained variance and 26%, 9% and 5% respectively. As seen on Figure 1 and 2, the plot shows similarities with how the UNDP classifies development. The countries with 'very high human development' as defined by the HDI are on the upper left while the 'low human development' countries are on the upper right in Figure 2. The different degrees of development of countries appear to have a U-shape or crater-like shape (see Figure 3) with countries at the bottom thorn in between the different sides of the crater. Surprisingly, the HDI categories are not clearly defined as the colors are overlapping. This already suggests that the HDI is not adequate to reflect the degree of development of a country. Some 'very

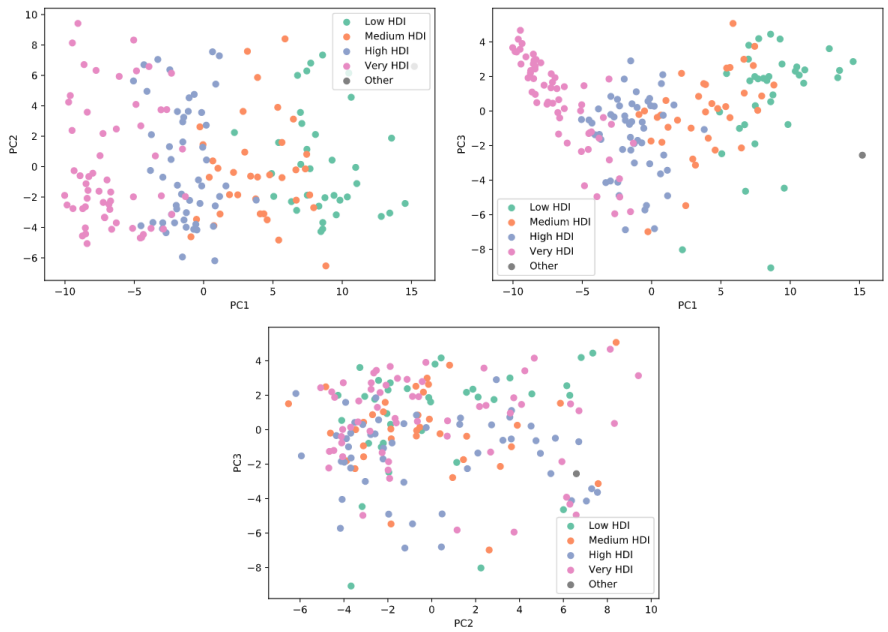

Figure 1: Scatter plot of the principal components of the countries by the HDI categories of the United Nations Development Programme

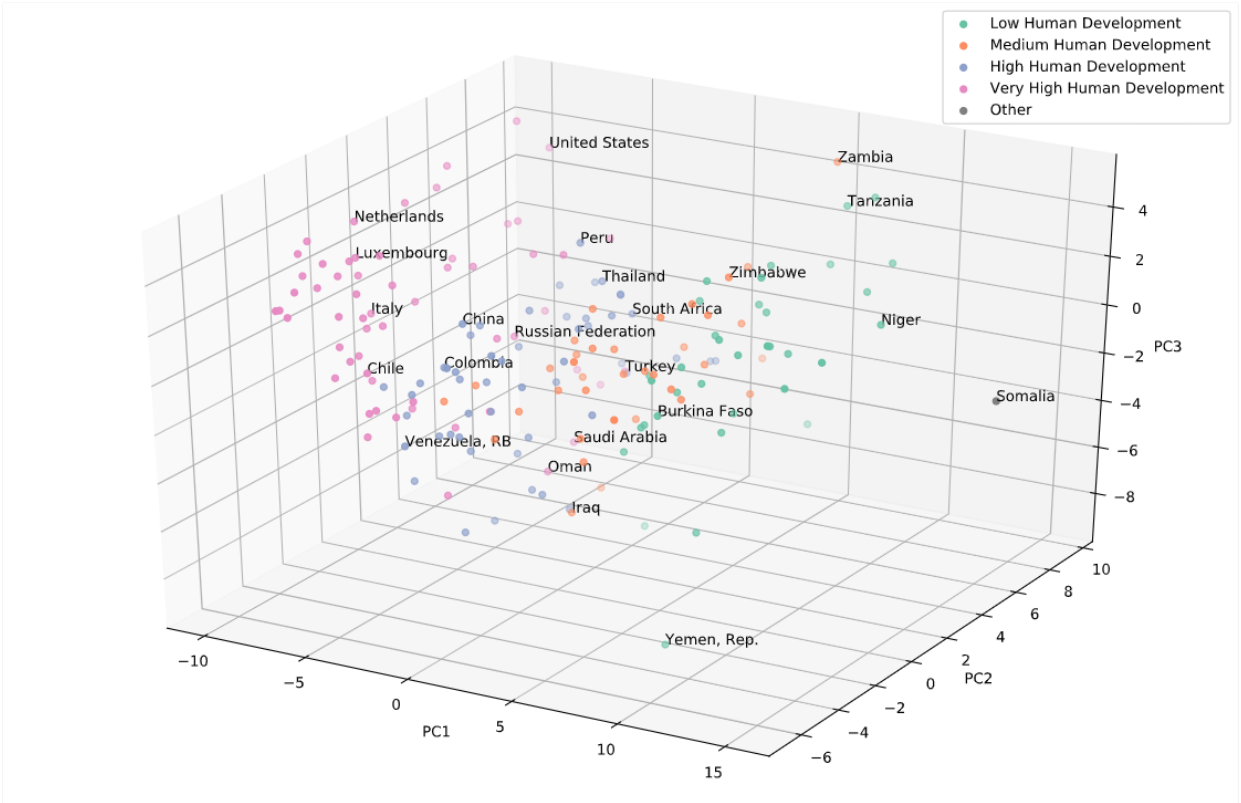

Figure 2: Plot of three principal components of the countries defined by the HDI categories of the United Nations Development Programme

high' human development countries like Saudi Arabia, Oman, Russian Federation are in between the 'high' and 'medium' human development countries. Since these countries have a lot of resources, the income per capita may inaccurately reflect the real income of the populations and their development. Moreover, the United States are on the opposite side of the European countries which also indicates dispersion among an HDI category. Peru seems closer to the 'very high human development' than Chile although Chile is the only South American country with a 'very high development' and is recognized as having the most advanced economy among them. Similarly, China seems closer to the 'very high development' countries and the Russian Federation to the 'high developed' countries. Two countries are completely apart, namely the Republic of Yemen and Somalia, and seem to be dissimilar to the other countries. Therefore, a clustering of this data may be a more accurate reflection of the current categorization of the development of countries as it is derived from diverse variables on the development of countries.

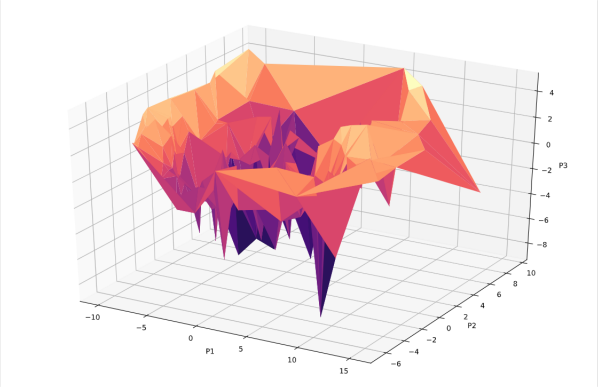

Figure 3: Triangular surface plot

4 Methodology

In this paper, we are interested in finding out how the development indicators in reduced dimension should be clustered. The aim of this research is to use multiple variables to make this classification of the development of countries and see how similar it is to the existing classification of international organization. Thus, three clustering methods are used namely k-means, Spectral and Agglomerative Hierarchical clustering. For each method, various k selection criterions are used to determine the optimal number of clusters. Finally, the cluster results can be compared in order to establish which one can be used on world development indicators in reduced dimension.

4.1 Clustering

k-means The first algorithm and probably the most popular clustering algorithm is k-means clustering. This method requires a pre-specified number k of clusters C_j for $j = 1 \dots k$. The aim is to assign a cluster C_j to $\mathbf{x_i}$ for $i = 1...n$. Each cluster C_j has a centroid $\mathbf{c_j}$, which is the center of mass or mean point of all the coordinates within the cluster. Every point $\mathbf{x_i}$ is assigned to a cluster C_j whose centroid is closest to $\mathbf{x_i}$. In other words, the aim is to minimizes the Euclidean distance between $\mathbf{x_i}$ and $\mathbf{c_j}$ which is defined in p-dimensions by:

$$||\mathbf{x_i} - \mathbf{c_j}|| = \sqrt{\sum_{j=1}^{p} (x_{ij} - c_{ij})^2} \tag{1}$$

After the variables have been assigned a cluster, the centroids are recalculated and the procedure is repeated until the clusters and centroids are stable, i.e. the same. Mathematically speaking, the aim is to minimize the objective function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} ||\mathbf{x_i} - \mathbf{c_j}||^2 \tag{2}$$

where $r_{ij} = 1$ if $\mathbf{x_i}$ is assigned to cluster C_j and zero otherwise. The goal is to find values for r_{ij} and $\mathbf{c_j}$ for which the objective function J is minimal. In practice, the k-means algorithm looks like the following:

Input: k

Method:

1. Initialization: determine the first k centroids
2. Assign $\mathbf{x_i}$ to the cluster with the closest centroids:

$$r_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_l ||\mathbf{x_i} - \mathbf{c_l}||^2 \text{ for } l = 1...k \\ 0 & \text{otherwise} \end{cases}$$

3. Recompute centroids : $\mathbf{c_j} = \frac{\sum_{i=1}^{n} r_{ij} \mathbf{x_i}}{\sum_{i=1}^{n} r_{ij}}$
4. Repeat 2 and 3 until the centroids are stable, i.e. the same.

Output: set of k clusters

The result of k-means clustering is sensitive to the initialization, namely the initial centroids in step 1. To circumvent this, the k-means++ algorithm was proposed by Arthur and Vassilvitskii

(2006). The algorithm works the same as k-means but has an additional step for initializing the first centroids. It starts by selecting a random point from the available data and computes its distance $D(x_i)$ to all other points. It chooses x_i as the next centroid such that the probability $\frac{D(x_i)^2}{\sum_{i=1}^{n} D(x_i)^2}$ is highest. This ensures that the point furthest away is chosen. The step is repeated with the newly chosen centroid and considers all the points that have not been marked as centroids. The initialization ends when k centroids have been selected to start the algorithm.

Spectral clustering In contrast to k-means, Spectral Clustering (Ng, Jordan, & Weiss, 2002) is a connectivity based method, meaning that points will be in the same cluster not because they are closest to a reference point but because they are closest to each other. The algorithm also requires a prespecified number k of clusters. The clusters are derived by performing k-means on the eigenvectors of the normalized Laplacian of the similarity matrix of the data. This explains why spectral clustering is also referred to as a graph-based clustering approach. In more details, one computes the gaussian similarities s_{ij} between all pairs of data:

$$s_{ij} = \exp(-\frac{||\mathbf{x_i} - \mathbf{x_j}||^2}{2\sigma}) \text{ for } i \neq j \tag{3}$$

Using the similarity, one constructs an affinity graph where the weights of the edges between to data points are equal to their similarities. From the graph, the adjacency matrix A can be derived where $A_{ij} = s_{ij}$ and $A_{ii} = 0$ which can also be called the affinity or similarity matrix. The diagonal degree matrix D is computed from the sum of the rows of the affinity matrix, hence $D_{ii} = \sum_{j=1} A_{ij}$. This leads us to computing the graph's normalized Laplacian which is a matrix representation of a graph.:

$$L = I - D^{-1/2} A D^{-1/2} \tag{4}$$

From which the eigenvalues λ_i and eigenvectors v_i can be derived:

$$Lv_i = \lambda_i v_i \text{ for } i = 1...n \tag{5}$$

The first k eigenvectors are put together in a matrix Y as columns and the rows are normalized to have unit length. Finally, k-means clustering, as explained previously, is performed on the rows of the new matrix and cluster C_j is assigned to $\mathbf{x_i}$ if row i of matrix Y belongs to C_j. sa

Hierarchical Clustering The last method that will be applied on the data is Agglomerative Hierarchical clustering. The approach will start by treating every point as a single cluster and in

a step-wise approach will group together clusters based on a linkage criterion. The linkage method that will be used in this analysis is Ward's method which consist of adding clusters together such that the within sum of squares is minimized at every step. Therefore, one adds two cluster A and B together if the merging cost is the smallest among all pairs of clusters:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} ||\mathbf{c_A} - \mathbf{c_B}||^2 \tag{6}$$

where $\mathbf{c_j}$ is the cluster center. This procedure results in a dendrogram, which is a tree-shaped diagram, where the individual observations are the leaves and the branches represents clusters. From the output of the dendrogram one usually decides on the number of clusters or where to horizontally "cut" the tree. However, this method is rather trivial and subjective to visual interpretations. Therefore, it will be interesting to subject this decision to the k-selection criterions.

4.2 K selection criterions

Even though clustering methods are a popular choice to categorize non-labeled data, they often require a prerequisite number k of clusters. In some cases the number can be deduced intuitively from the data or simply with prior knowledge. In this analysis, it is particularly interesting to see how mathematical methods agree with the classification of the UNDP from 'low' to 'very high human development' or the classification of the World Bank and the IMF, namely 'least developed', 'emerging economies' and 'developed economies'. As both are distinct in how they measure development, it is particularly interesting in how many clusters should the world development indicators be clustered. Wang (2010) considers cluster instability when he defines his cross-validation methods to determine the optimal choice of k. A selection of methods will be used to determine what is the optimal k in the range of 3 and 10 depending on the clustering algorithm.

4.2.1 Cross-validation

Wang (2010) uses cross-validation to define the optimal number of clusters. Cross-validation is used in model validation to evaluate if a model can accurately predict other points. Wang uses this idea to define cluster instability to find the number of clusters. His idea is built on the premise that two random observations should be consistently either in the same cluster or in different clusters. For cross-validation with voting, the data is permuted C times and for every permutation the data is split into three sets z_1^c, z_2^c and z_3^c respectively of size m, m and $n - 2m$ where $m = \frac{n}{3}$. The sets z_1^c and z_2^c are clustered $\psi_i^c = \Psi_i(z_i^c; k)$. Each observations of z_3^c is assigned a cluster from ψ_1^c and ψ_2^c.

The cluster instability increases when two pair of observations of z_3^c are not consistently clustered together or apart in each clustering ψ_1^c and ψ_2^c.:

$$\hat{s}^c(\Psi, k, m) = \sum_{2m+1 \leq i < j \leq n} V_{ij}^c(\Psi, z_1^c, z_2^c) \tag{7}$$

$$\text{with } V_{ij}^c(\Psi, z_1^c, z_2^c) = I[I\{\psi_1^c(\mathbf{x_i}^c) = \psi_1^c(\mathbf{x_j}^c)\} + I\{\psi_2^c(\mathbf{x_i}^c) = \psi_2^c(\mathbf{x_j}^c)\}] = 1]$$

For every permutation c, $\hat{k}^c$ is determine by the lowest cluster instability $\arg\min_{3 \leq k \leq 10} \hat{s}^c(\Psi, k, m)$. Then, the optimal k is defined by the mode of the different $\hat{k}^c$.

The second method, namely cross-validation with averaging, works similarly than the previous one, only that $\hat{s}^c(\Psi, k, m)$ is averaged over all C permutations for $k = 3...10$:

$$\hat{s}(\Psi, k, m) = C^{-1} \sum_{c=1}^{C} \hat{s}^c(\Psi, k, m) \tag{8}$$

Finally, the optimal choice of k is defined by $\arg\min_{3 \leq k \leq 10} \hat{s}(\Psi, k, m)$.

Wang (2010) in both his approaches assigns clusters of ψ_1^c and ψ_2^c to the observations of z_3. For k-means clustering, the variables are assigned to the cluster with the closest centroid. With methods like spectral clustering he suggest to use the k-Nearest Neighbor (KNN) algorithm to assign a cluster to the observations using 10 neighbors in total. However, this method does not always work as in some cases, the neighbors are equally split between two clusters. Therefore, whenever a cluster assignment results in a tie the procedure is repeated by leaving out the furthest neighbor until a cluster can be determined. Since there is no suggestion for agglomerative hierarchical clustering, the method used will be to assign a cluster to an observation such that the within sum of squares is minimal. This relates to the linkage criterion that groups two clusters such that the within sum of squares is smallest and therefore, it seems like the most appropriate method to use with agglomerative hierarchical clustering.

4.2.2 Comparison

Elbow method The first metric is the elbow method which consist in determining k from the plot of the sum of squarer errors. The optimal k is defined by the value where the rate of decrease starts decreasing from a clustering to one with more clusters also called the elbow point. This method motivates the finding of an equilibrium between the average dispersion and then number of clusters.

Caliński and Harabasz method The next metric was developed by Caliński and Harabasz (1974) and consist in finding the ratio of the between and within sum of squares that is maximized for a certain k. The formula takes the following form:

$$CH(k) = \frac{(n-k)B(k)}{(k-1)W(k)} \tag{9}$$

where $B(k)$ is the between sum of squares and $W(k)$ is the within sum of squares.

Davies-Bouldin Index The Davies-Bouldin (DB) index is a criterion that takes into account the inner cluster variety and the between cluster distances. The closeness of two clusters is defined by the sum of the clusters within sum of squares divided by the distance of their centroids.

$$R_{k_i,k_j} = \frac{W(k_i) + W(k_j)}{|c_{k_i} - c_{k_j}|} \tag{10}$$

For every cluster, the 'closest' neighbouring cluster is determined by the maximum R_{k_1,k_2} for $k_1/neqk_2$. The Davies-Bouldin index averages the statistic of the closest neighbor:

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_{k_i} \text{ where } R_{k_i} = \max_{i \neq j} R_{k_i,k_j} \tag{11}$$

Silhouette method The silhouette statistic is another known method developed by Kaufman and Rousseeuw (1990) to determine the number of k clusters. It is defined by

$$s(k) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \tag{12}$$

where $a(x_i)$ is the average distance to other points in the cluster and $b(x_i)$ is the average distance to points in the nearest cluster. The number k is determined by the clustering that maximizes the silhouette statistic $s(k)$.

Gap Statistic The gap statistic was developed by a team of researchers at Stanford (Tibshirani, Walther, & Hastie, 2001) which tries to approximate the distribution of errors. It will compare the innervariance of the clusters to a reference data set with a uniform distribution z_n:

$$GAP(k) = B^{-1}\sum_{b=1}^{B} \log W_k^b - \log W_k \tag{13}$$

where W_k is the within sum of squares distances with k clusters and W_k^b is the bth within sum of squares of the clustering of the reference data set. One chooses the smallest k for which $GAP(k)$ is within one standard deviation of the statistic for $k+1$, namely we choose k such that $GAP(k) \geq GAP(k+1) - s_{k+1}$, where s_{k+1} is the sample standard deviation of $\log W_k^{*b}$.

5 Results

5.1 Selection of k and Clusterings

To determine the optimal choice of k, different methods are used as well Wang's cross validation. In Table 1, the different methods choose four clusters for k-means, eight clusters for Spectral clustering and three clusters for Agglomerative Hierarchical clustering. This is not surprising as the methods cluster in a different way and are therefore not directly comparable. The Caliński-Harabasz index and the Silhouette statistic seem to be equivalent in how they determine the number of clusters in all three clustering methods. Cross-validation with voting and with averaging define consistently ten clusters which is the maximum number indicated among the criterions. The cross-validation method do not seem to agree with the other criterions on this data whereas Wang (2010) showed in his simulation that the cross-validations methods were more consistent is selecting then number of clusters. Moreover, the cross-validation methods are computationally inefficient as they have a time complexity of $O(Ckm)$ whereas most of the other criterions run in $O(k)$. While the idea of using cross-validation to determine k is innovative in comparison to the exiting criterions, it does not seem to be more efficient.

Table 1: Results of the k-selection criterions

	k-means	Spectral Clustering	Agglomerative Hierarchical
Elbow	6	8	5
CH	4	8	3
DB	4	5	9
Silhouette	4	8	3
GAP	4	3	3
CV_v	10	10	10
CV_a	10	10	10
	4	8	3

As a result, the three clusterings methods are applied on the data with the previously determined k (see Figure 4). Among the different clusterings, Agglomerative Hierarchical seems to have the most clearly separated clusters meaning that it may be the most appropriate method to cluster the development of countries. The clusters of k-means are similar to the ones of Agglomerative Hierarchical only that cluster 4 seems to span into cluster 1 and cluster 3. Hence, analysing the

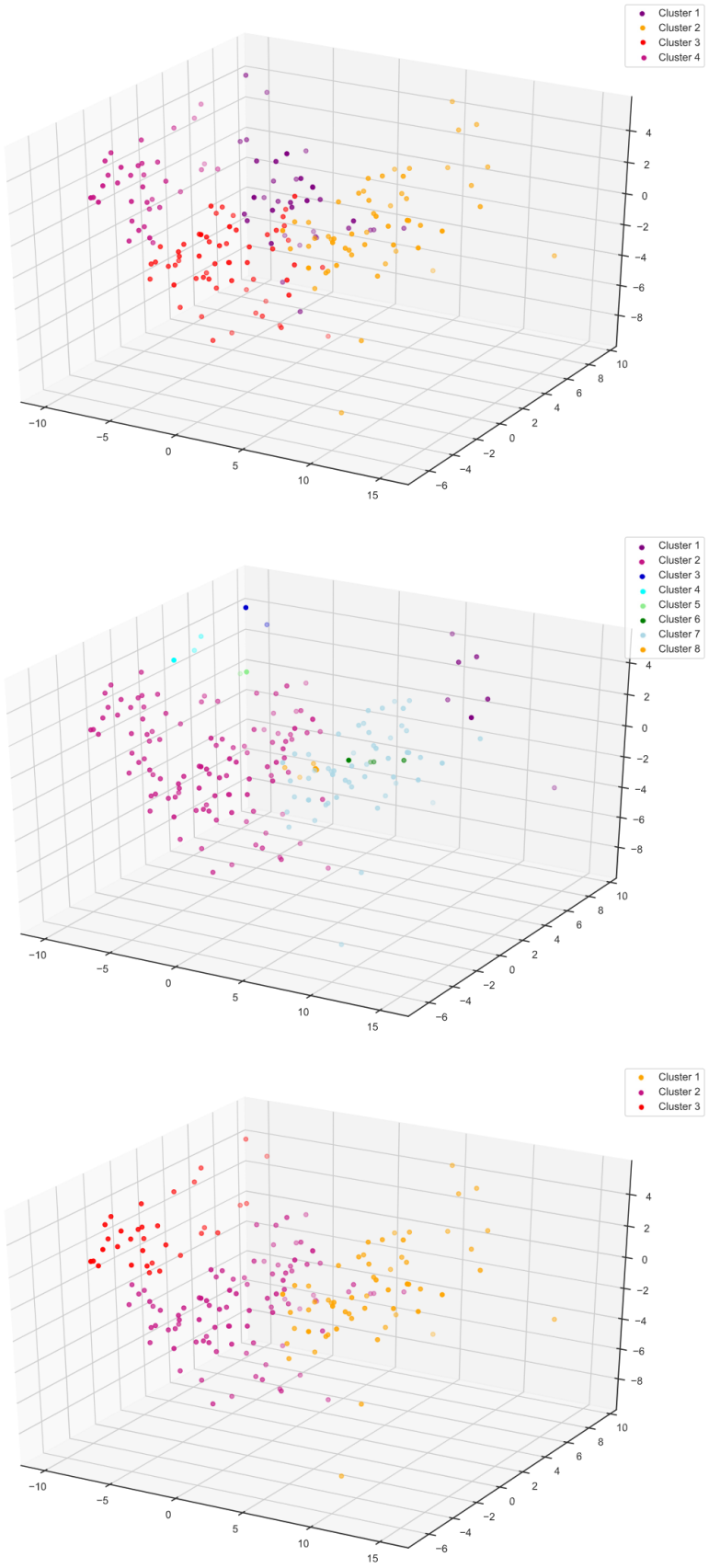

Figure 4: Clustering of k-means, Spectral Clustering and Agglomerative Hierarchical (Top to Bottom)

original variables may lead to interesting findings towards why cluster 4 is in its separate cluster and how it compares to the other clusters. Last but not least, the Spectral Clustering shows two large clusters and many smaller ones. This could be interpreted as two reference categories while the others are groups of countries that are very dissimilar to the reference categories. In all the clusterings, further analysis should provide more insights into whether the results can be interpreted or if the methods should not be used to cluster the development of countries.

5.2 Cluster Statistics

To understand if the algorithms made interesting and relevant clusterings it is important to analyze and compare them. Firstly, for k-means clustering, Figure 5 shows the visual representation of the clusters on a map. Cluster 1 was the most intriguing cluster as it seemed to span into cluster 2 and 3. Indeed the countries of cluster 1 have a lot of similarities with the countries of cluster 3 as they have similar mean and expected schooling years, a similar adolescent birth rate and around an average of 60% of their employment in services. They have among the highest average natural resource depletion as a percentage of GNI and one of the lowest Environmental Performance Index (EPI) along with cluster 2 and 3. The United States of America and United Kingdom are clustered

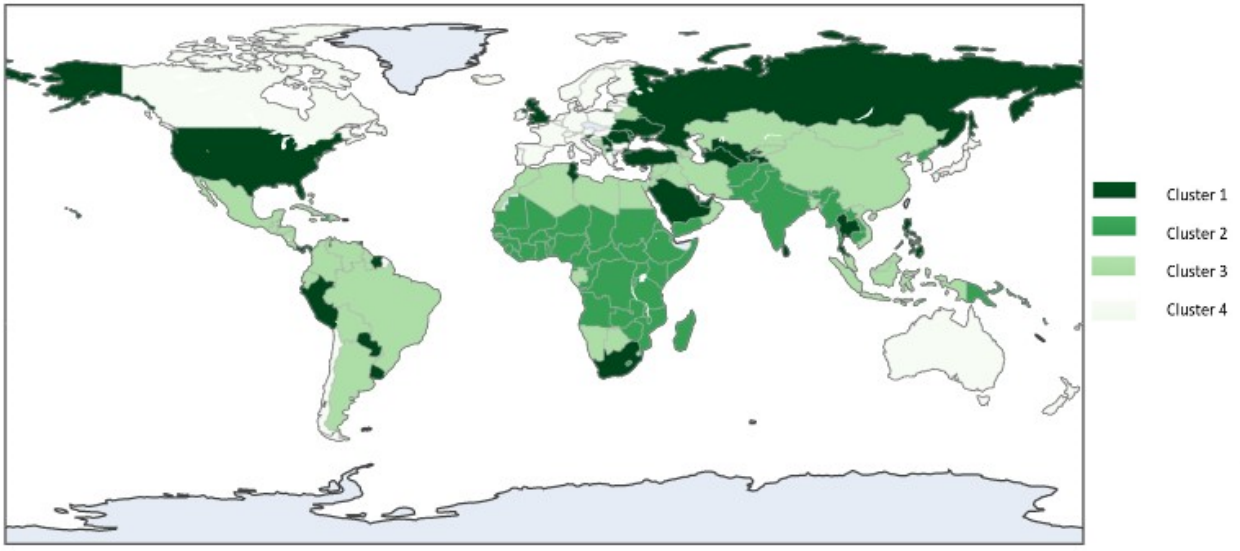

Figure 5: Map of the world by the k-means clusters

differently than the very highly developed countries and cluster 1 contains countries spanning from very high human development to medium human development as defined by the HDI. This is surprising and may indicate that k-means does not cluster the different degrees of development

efficiently. The k-means algorithm seems to make a cluster in parallel of cluster 2, 3 and 4 which have similarities to the categorization of the IMF (see Figure 8). The clustering seems to exhibit more interesting contrasts on the different continents than globally. For example in South America, Chile is recognized as being the most economically advanced country, whereas in the African continent it is South Africa. In South East Asia, Hong-Kong has a highly advanced economy and Thailand is mainly a tourism economy where tourism accounts for around 20% of the GDP. In summation, the clustering of k-means seems makes a cluster that has a lot of characteristics of cluster 2 and 3 and has, while the other three clusters share resemblance with the categorization made by the IMF. It is therefore interesting to see how different the countries of cluster 4 are from the others but it results in limited interpretations to categorize the development of countries on a global scale.

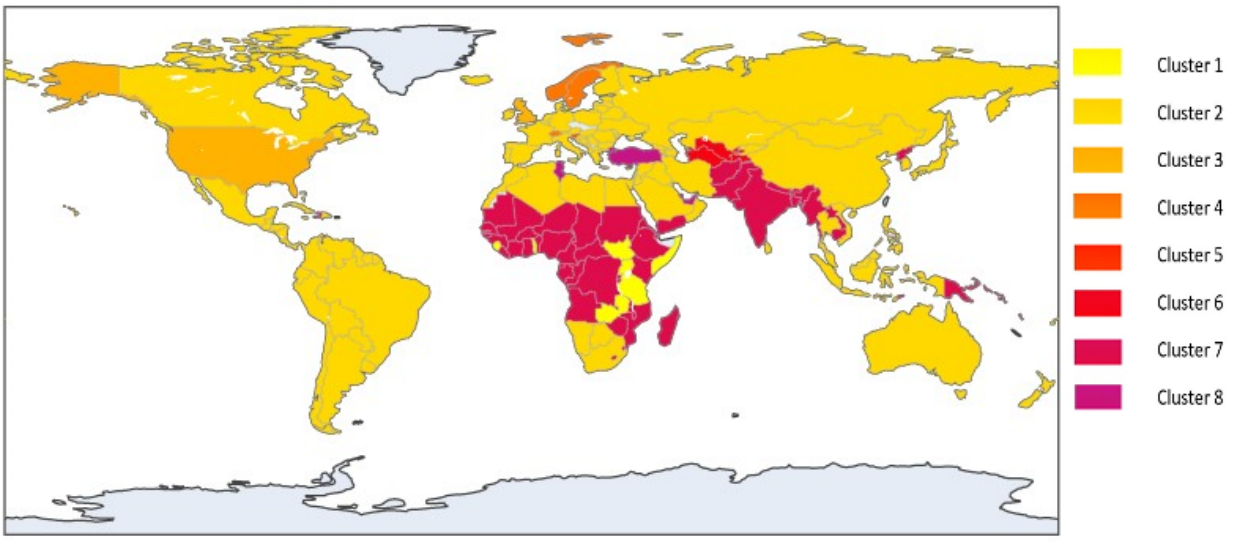

Figure 6: Map of the world by the Spectral Clustering clusters

The statistics of the clusters from Spectral clustering seem to also provide little evidence for interpretations as it contains multiple small clusters and two larger ones (see Figure 6). The Scandinavian countries and Switzerland are as well separately grouped from the other European nations but it is not possible to compare the two cluster as cluster 2 does not have dominant characteristics. Furthermore, there are more distinctions in Africa where countries like Tanzania and Zimbabwe are clustered differently than Cameroon or Namibia as they have younger population and higher years of education, but also higher mortality rates and less access to electricity, drinking water and sanitation services. The USA and the UK are again clustered together. However, most of the countries of the world are included in in cluster 2. Those include countries ranging from medium

to very high human development as defined by the UNDP. There are no apparent characteristics as there is too much variation in the variables of the countries of cluster 2. This clustering does not seem to have a pattern that can capture further strong differences. In hindsight, Spectral clustering should not be used to cluster the development of countries.

The map of the world clustered with the Agglomerative Hierarchical shows a lot of resemblances to the current classification by the United Nations. Indeed, cluster 1 on Figure 7 is characterized by the lowest mean and expected school years as well as the largest gap in education between men and women (see Table 2). They have the highest employment in agriculture, the median age of the population is 20 and they have the highest infant mortality rate. Their environmental performance is also quite low since their exposure to heavy metals is high. They have a poor average index for drinking water and sanitation because they have few wastewater treatment plants. On average, they have the highest positive forest area change and the highest forest area as a percentage of land is among the clusters. The population of the countries of cluster 2, in comparison, has an average median age of 30, 70% of the population has on average a secondary education and they are mainly employed in industries. It is the only cluster that has a negative average forest change. By contrast, the population of the countries in cluster 3 have the highest life expectancy and the average medium age is twice that of the countries of cluster 1. Around 90% of the population finishes their secondary education the average income is nearly 12 times that of the average income of cluster 1. The population is mostly urban, their employment is mainly in services, they have the highest average GNI for both male and female and they have around 900 times more internet servers per million people compared to the countries of cluster 1. They have the highest average ecosystem vitality index as they protect a lot of areas but they also have the largest average amount of natural resource depletion as a percentage of GNI. They have the highest emissions and air pollution and consequently the highest mortality rate attributable to household and ambient air pollution. Additionally, women are usually employed in services as in cluster 2 but they have the highest Women in Business and Law index and the lowest vulnerable female employment. In the countries of cluster 1 and 2, women are usually the most vulnerable to unemployment than men.

In all, the characteristics resemble the classification of the International Monetary Fund (IMF) (see Figure 8). However, India belongs to the sub-Saharan African countries often considered the least developed countries along with Pakistan, Zimbabwe, Nigeria, Cameroon, Papua New-Guinea and North Korea. Poland is in cluster 3 along with the other European countries while Lithuania, Greece and Cyprus are in cluster 2 mostly associated with the developing economies.

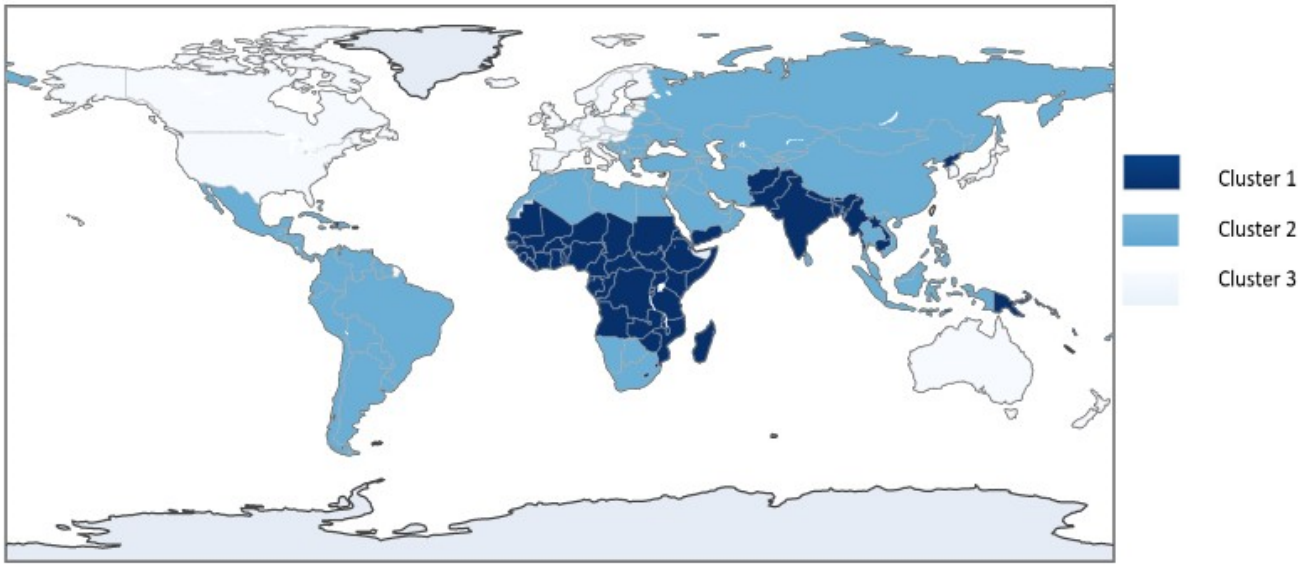

Figure 7: Map of the world by the Agglomerative Hierarchical clusters

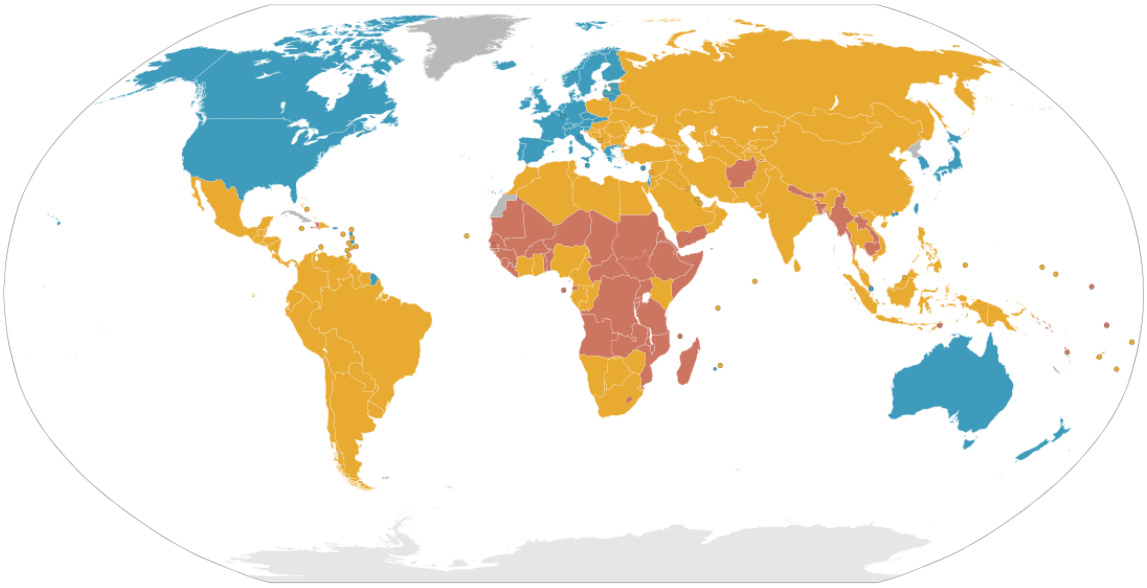

Figure 8: Map of the development of countries by the International Monetary Fund

Table 2: Selection of averages of the variables of the Agglomerative Hierarchical clusters

	Cluster 1	Cluster 2	Cluster 3
Median Age	20,57	31,79	42,09
Life expectancy at birth, female (years)	65,36	77,28	84,04
Life expectancy at birth, male (years)	61,72	71,83	78,90
Expected years of schooling, female (years)	9,63	14,09	17,56
Expected years of schooling, male (years)	10,52	13,47	16,72
Mean years of schooling, female (years)	4,00	9,39	12,17
Mean years of schooling, male (years)	5,76	9,60	12,37
Population with at least some secondary education, female (% ages 25 and older)	23,94	69,26	89,21
Population with at least some secondary education, male (% ages 25 and older)	34,45	70,58	91,19
Estimated gross national income per capita, female (2011 PPP $)	2571,72	11706,13	35018,88
Estimated gross national income per capita, male (2011 PPP $)	4508,13	22514,64	52750,64
Employment in agriculture (% of total employment)	50,24	17,06	3,28
Employment in industry (% of total employment)	12,89	23,19	21,66
Employment in services (% of total employment)	36,88	59,75	75,05
Women Business and the Law Index Score	64,78	71,90	93,97
Vulnerable employment, female (% of female employment)	79,97	28,30	7,71
Vulnerable employment, male (% of male employment)	65,07	27,52	10,84
Mortality rates, infant (per 1,000 live births)	46,78	13,77	3,13
Mortality rate attributed to Household and ambient air pollution (per 100,000 population)	83,25	93,34	101,45
Forest Area Change (%)	3,47	4,89	-2,23
Forest area (% of total land area)	29,65	34,95	25,00
Environmental Performance Index	42,57	57,92	74,70
Heavy Metals Exposure Index	35,12	52,52	82,50
Ecosystem Vitality Index	45,36	52,41	65,94
Natural resource depletion (% of GNI)	3,47	4,19	5,35
Secure Internet servers (per 1 million people)	40,18	5142,35	35054,13

In summary, Agglomerative Hierarchical clustering leads to a clustering that strongly resemblances to the classification of the IMF while the other algorithms find more regional differences. K-means clustering finds a fourth cluster that is in parallel with the other clusters and that makes. The clusters show more differences in South America, Asia and East Europe. Similarly, Spectral Clustering finds more clusters in Europe and Africa. Both may be considered jointly to see if they can present more precise different degrees of development but Agglomerative Hierarchical clustering is the most appropriate method to cluster the different degrees of development. In the year 2018, three different clusters are identified as having different economic and human development stages.

6 Conclusion

This paper aimed to find out how the development of countries should be clustered. The research motivated the use of a large amount of variables to determine clear contrasts in the development of countries. This was done using a probabilistic principal component analysis to represent the data in reduced dimensionality preserving the maximum variance among the observations. Different clustering algorithms were tried out, all of which required predefined number of clusters k. To determine the optimal choice of clusters, numerous criterions were used. Two of those method are cross validation methods developed by Wang (2010) that had not been used in hierarchical clustering before. Using the different k selection criterions four, eight and three clusters were respectively assigned to k-means, Spectral and Agglomerative Hierarchical clustering. Wang's cross validation methods did not perform well among the different criterions as it did not find the same number of clusters as the other criterions and consistently predicted 10 clusters while the other criterions found different number of clusters for each clustering algorithm. Also, since the computing time is really high in comparison to the other criterions, the cross-validation methods may not be recommended to determine the optimal choice of k as deduced from the research in this application. By contrast, Wang found that his cross-validation methods were more consistent in predicting the number of clusters in his simulation. This indicates that there may not be an underlying structure in the data that can be clustered. However, further real applications of Wang's cross-validation should give more understanding in whether his methods can be used to determine the optimal choice of k in practice. Nonetheless, for the sake of categorizing the countries in different development categories, the data of the development of countries in reduced dimension was clustered. The research found that k-means and Spectral clustering are inappropriate to cluster the development of country in comparison

to Agglomerative Hierarchical clustering. Since they defined smaller clusters and clusters that have different characteristics than the other clusters, it may be relevant for further research to look at clusterings regionally or within the defined world development clusters. Agglomerative Hierarchical clustering showed to have the most similarities with the established classification of the IMF with the exception of some countries. All the three clusters had differences in income, educational attainment, type of employments and environmental sustainability that characterize each development cluster. Although the results are similar to existing classifications, the novelty of this research was to use more variables to terminate the association of the development with the GDP or the HDI. Indeed, the classification or categorization of development needs to take into account other variables than monetary indicators of growth and industrialization. The academic and social relevance of this paper motivates the use of variables related to health, education, political and environmental sustainability to establish the differences in the development of countries. This is however constrained to the lack of transparency of some countries where official statistics are not always available. This makes the subject of establishing different degrees of development, in addition of using multiple variables, even more difficult. Nonetheless, data transformation techniques can be used. The downside is that it may lead to doubt whether the structure of data is accurately preserved and to what extent one can derive results and make conclusions. Further research is necessary to motivate the use of data transformation to. In this application, the data transformation seemed to preserve the pattern of the current classification of the UNDP. Its clustering showed to rival with the established development categories and allows to question the current classification of certain countries as the categories were overlapping. Given the difference within the categories and within the continents, further research should be done to establish the inner cluster and intra-region differences and enable further comparisons of different degrees of development. It may also be relevant to look at the evolution of the development of countries through time. Since all economies have been growing exponentially but started off with large inequalities, the richest countries are still the richest countries and the development of other nations can be likely overlooked and reduced when compared at a certain moment in time. Therefore, as development is not static, establishing categories for a certain year may not reflect the evolutionary character of the development of countries and an analysis throughout time using the different aspects of development may prove to be more relevant at making comparisons between countries.

References

Arthur, D., & Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. *Stanford*.

Bagolin, I. (2004). Human development index (hdi)-a poor representation to human development approach. *University of Rio Grande do Sul, PUCRS, Brazil*.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *1*(3), 21-27.

Hicks, N., & Streeten, P. (1979). Indicators of development: the search for a basic needs yardstick. *World development*, *7*(6), 567-580.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. A Wiley and Sons. New York.

Kuznets, S. (1934). National income, 1929-1932. *In National Income, 1929-1932*, 1-12. NBER.

Monni, S., & Spaventa, A. (2013). Beyond GDP and HDI: Shifting the focus from paradigms to politics. *Development*, *56*(2), 227-231.

Morse, S. (2013). *Indices and indicators in development: An unhealthy obsession with numbers*. Routledge.

Mrázová, I., & Dagli, C. H. (2008). Semantic clustering of the world bank data. *International Journal of General Systems*, *4*(37), 417-442.

Mylevaganam, S. (2017). The analysis of human development index (hdi) for categorizing the member states of the united nations (un). *Open Journal of Applied Sciences*, *7*(12), 661–690.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). In advances in neural information processing systems. *Neural Information Processing Symposium 2001*, 849-856.

Reig-Martínez, E. (2013). Social and economic wellbeing in europe and the mediterranean basin: Building an enlarged human development indicator. *Social Indicators Research*, *111*(2), 527-547.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *2*(63), 441-423.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *94*(4), 834–904.

Yorulmaz, Ö. (2016). Assessing the effects of various socio-economic and health indicators on hdi country categories. *Alphanumeric Journal*, *4*(1), 1-10.

Appendix

Table 3: Variables of the United Nations Development Programme

2019 Human Development Data - United Nations Development Programme	
Gender Development Index	GDP Annual growth (%)
Life expectancy at birth female (years)	Gross fixed capital formation (% of GDP)
Life expectancy at birth male (years)	General government final consumption expenditure (% of GDP)
Expected years of schooling female (years)	Domestic credit provided by financial sector (% of GDP)
Expected years of schooling male (years)	Consumer price index (2010=100)
Mean years of schooling female (years)	Employment to population ratio (% ages 15 and older)
Mean years of schooling male (years)	Labour force participation rate (% ages 15 and older)
Estimated gross national income per capita female (2011 PPP $)	Total unemployment (% of labour force)
Estimated gross national income per capita male (2011 PPP $)	Youth unemployment (% ages 15-24)
Maternal mortality ratio (deaths per 100000 live births)	Birth registration (% under age 5)
Adolescent birth rate (births per 1000 women ages 15–19)	Refugees by country of origin (thousands)
Share of seats in parliament (% held by women)	Homeless people due to natural disaster (average annual per million people)
Population with at least some secondary education female (% ages 25 and older)	Prison population (per 100000 people)
Population with at least some secondary education male (% ages 25 and older)	Suicide rate female (per 100000 people)
Labour force participation rate female (% ages 15 and older)	Suicide rate male (per 100000 people)
Labour force participation rate male(% ages 15 and older)	Average dietary energy supply adequacy (%)
Total (millions)	Exports and imports (% of GDP)
Average annual growth 2015/2020 (%)	Foreign direct investment net inflows (% of GDP)
Urban population (%)	Private capital flows (% of GDP)
Popultion under age 5 (millions)	Remittances inflows (% of GDP)
Population aged 15-64 (millions)	Net migration rate (per 1000 people)
Population aged 64+ (millions)	Stock of immigrants (% of population)
Median Age	International inbound tourists (thousands)
Dependency ratio of young age 0-14 (per 100 people ages 15–64)	Internet users total (% of population)
Dependency ratio of young age 65 and older (per 100 people ages 15–64)	Mobile phone subscriptions (per 100 people)
Total fertility rate (births per woman)	Physicians (per 10000 people)
Infants lacking immunization DPT (% of one-year-olds)	Hospital beds (per 10000 people)
Infants lacking immunization Measles (% of one-year-olds)	Pupil-teacher ratio primary school (pupils per teacher)
Mortality rates infant (per 1000 live births)	Rural population with access to electricity (%)
Mortality rates under five (per 1000 live births)	Population using at least basic drinking-water services (%)
Mortality rates female Adult (per 1000 people)	Population using at least basic sanitation services (%)
Mortality rates male Adult (per 1000 people)	Mandatory paid maternity leave (days)
Mortality rates attributed to noncommunicable diseases female (per 100000 people)	Renewable energy consumption (% of total final energy consumption)
Mortality rates attributed to noncommunicable diseases male(per 100000 people)	Forest area (% of total land area)
Incidence of tuberculosis (per 1000 people at risk)	Forest Area Change (%)
Healthy life expectancy at birth (years)	Natural resource depletion (% of GNI)
Current health expenditure (% of GDP)	Mortality rate attributed to Household and ambient air pollution (per 100000 population)
Population with at least some secondary education (% ages 25 and older)	Mortality rate attributed to Unsafe water sanitation and hygiene services (per 100000 population)
Gross enrolment ratio primary (% of primary school-age children)	Red List Index
Survival rate to the last grade of lower secondary general education	Gross capital formation (% of GDP)
GDP per capita (2011 PPP $)	[rgb] .961, .961, .961Concentration index (exports)]

Table 4: Variables of the World Bank

World Development Indicators 2018 - World Bank	
Life expectancy at birth female (years)	Merchandise exports to low- and middle-income economies in Latin America & the Caribbean (% of total merchandise exports)
Life expectancy at birth male (years)	Merchandise exports to low- and middle-income economies in Middle East & North Africa (% of total merchandise exports)
Foreign direct investment net inflows (% of GDP)	Merchandise exports to low- and middle-income economies in South Asia (% of total merchandise exports)
Access to electricity urban (% of urban population)	Merchandise exports to low- and middle-income economies in Sub-Saharan Africa (% of total merchandise exports)
Birth rate crude (per 1000 people)	Merchandise exports to low- and middle-income economies outside region (% of total merchandise exports)
Death rate crude (per 1000 people)	Merchandise imports (current US$)
Ease of doing business score (0 = lowest performance to 100 = best performance)	Merchandise imports from low- and middle-income economies in East Asia & Pacific (% of total merchandise imports)
Employers female (% of female employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies in Europe & Central Asia (% of total merchandise imports)
Employers male (% of male employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies in Latin America & the Caribbean (% of total merchandise imports)
Employers total (% of total employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies in Middle East & North Africa (% of total merchandise imports)
Employment in agriculture (% of total employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies in South Asia (% of total merchandise imports)
Employment in agriculture female (% of female employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies in Sub-Saharan Africa (% of total merchandise imports)
Employment in agriculture male (% of male employment) (modeled ILO estimate)	Merchandise imports from low- and middle-income economies outside region (% of total merchandise imports)
Employment in industry female (% of female employment) (modeled ILO estimate)	Merchandise trade (% of GDP)
Employment in industry (% of total employment) (modeled ILO estimate)	Population density (people per sq. km of land area)
Employment in industry male (% of male employment) (modeled ILO estimate)	Population growth (annual %)
Employment in services (% of total employment) (modeled ILO estimate)	Rural population (% of total population)
Employment in services female (% of female employment) (modeled ILO estimate)	Secure Internet servers (per 1 million people)
Employment in services male (% of male employment) (modeled ILO estimate)	Self-employed total (% of total employment) (modeled ILO estimate)
Foreign direct investment net outflows (% of GDP)	Terrestrial and marine protected areas (% of total territorial area)
Labor force female (% of total labor force)	Unemployment female (% of female labor force) (modeled ILO estimate)
Land area (sq. km)	Unemployment male (% of male labor force) (modeled ILO estimate)
Lower secondary school starting age (years)	Urban population growth (annual %)
Merchandise exports (current US$)	Vulnerable employment female (% of female employment) (modeled ILO estimate)
Merchandise exports to low- and middle-income economies in East Asia & Pacific (% of total merchandise exports)	Vulnerable employment male (% of male employment) (modeled ILO estimate)
Merchandise exports to low- and middle-income economies in Europe & Central Asia (% of total merchandise exports)	Wage and salaried workers total (% of total employment) (modeled ILO estimate)
	[rgb] .961, .961, .961Women Business and the Law Index Score (scale 1-100)]

Table 5: Democracy Index and Environmental Performance Index variables

Democracy Index Report 2018 - Economist's Intelligence Unit	Environmental Performance Indices - Columbia University and Yale University
Environmental Performance Index	Democracy index
Environmental Health Index	Electoral pluralism index
Air Quality Index	Government index
Water and Sanitation Index	Political participation index
Heavy Metals Exposure Index	Political culture index
Ecosystem Vitality Index	Civil liberties index
Biodiversity and Habitat Index	
Climate and Energy Index	
Air Pollution Index	
Water Resources Index	
Agriculture Index	

Table 6: K-means clusters

Cluster 1	Cluster 2		Cluster 3		Cluster 4
Moldova	Afghanistan	Niger	Albania	Iraq	Australia
North Macedonia	Angola	Nigeria	Algeria	Jamaica	Austria
West Bank and Gaza	Benin	Pakistan	Antigua and Barbuda	Jordan	Belgium
Panama	Bhutan	Papua New Guinea	Argentina	Kazakhstan	Bulgaria
Paraguay	Burkina Faso	Rwanda	Armenia	Kuwait	Canada
Peru	Burundi	Senegal	Azerbaijan	Kyrgyz Republic	Chile
Philippines	Cambodia	Sierra Leone	Bahamas, The	Lebanon	Croatia
Qatar	Cameroon	Solomon Islands	Bahrain	Libya	Cyprus
Romania	Central African Republic	Somalia	Bangladesh	Malaysia	Czechia
Russian Federation	Chad	South Sudan	Barbados	Maldives	Denmark
St. Lucia	Comoros	Sudan	Belarus	Mauritius	Estonia
St. Vincent and the Grenadines	Congo, Rep.	Tanzania	Belize	Mexico	Finland
Samoa	Congo, Dem. Rep.	Togo	Bolivia	Micronesia, Fed. Sts.	France
Sao Tome and Principe	Cote d'Ivoire	Uganda	Bosnia and Herzegovina	Mongolia	Germany
Saudi Arabia	Djibouti	Vanuatu	Botswana	Montenegro	Greece
Serbia	Equatorial Guinea	Yemen, Rep.	Brazil	Morocco	Hong Kong SAR, China
Seychelles	Eritrea	Zambia	Brunei Darussalam	Namibia	Hungary
Singapore	Ethiopia	Zimbabwe	Cabo Verde	Nicaragua	Iceland
Slovakia	Gambia, The		China	Oman	Ireland
Slovenia	Ghana		Colombia	Syrian Arab Republic	Israel
South Africa	Guinea		Costa Rica	Venezuela, RB	Italy
Sri Lanka	Guinea-Bissau		Cuba	Vietnam	Japan
Suriname	Haiti		Dominica		Korea, Rep.
Tajikistan	India		Dominican Republic		Latvia
Thailand	Kenya		Ecuador		Lithuania
Timor-Leste	Kiribati		Egypt, Arab Rep.		Luxembourg
Tonga	Korea, Dem. People's Rep.		El Salvador		Malta
Trinidad and Tobago	Lao PDR		Eswatini		Netherlands
Tunisia	Lesotho		Fiji		New Zealand
Turkey	Liberia		Gabon		Norway
Turkmenistan	Madagascar		Georgia		Poland
Ukraine	Malawi		Grenada		Portugal
United Arab Emirates	Mali		Guatemala		Spain
United Kingdom	Mauritania		Guyana		Sweden
United States	Mozambique		Honduras		Switzerland
Uruguay	Myanmar		Indonesia		
Uzbekistan	Nepal		Iran, Islamic Rep.		

Table 7: Spectral Clustering Clusters

Cluster 1	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 8
Sierra Leone	United Kingdom	Norway	Singapore	Samoa	Qatar
Somalia	United States	Sweden	Slovenia	Tajikistan	St. Vincent and the Grenadines
South Sudan		Switzerland		Turkmenistan	Tunisia
Tanzania				Uzbekistan	Turkey
Togo					United Arab Emirates
Uganda					
Zambia					

	Cluster 2			Cluster 7	
Albania	Colombia	Hungary	Afghanistan	Guinea-Bissau	Sudan
Algeria	Costa Rica	Iceland	Angola	Haiti	Timor-Leste
Antigua and Barbuda	Croatia	Indonesia	Bangladesh	India	Vanuatu
Argentina	Cuba	Iran	Benin	Kenya	Yemen, Rep.
Armenia	Cyprus	Iraq	Bhutan	Kiribati	Zimbabwe
Australia	Czechia	Ireland	Burkina Faso	Korea, Dem. People's Rep.	
Austria	Denmark	Israel	Burundi	Lao PDR	
Azerbaijan	Dominica	Italy	Cambodia	Lesotho	
Bahamas, The	Dominican Rep.	Jamaica	Cameroon	Liberia	
Bahrain	Ecuador	Japan	Central African Rep.	Madagascar	
Barbados	Egypt, Arab Rep.	Jordan	Chad	Malawi	
Belarus	El Salvador		Comoros	Mali	
Belgium	Estonia		Congo, Rep.	Mauritania	
Belize	Fiji		Congo, Dem. Rep.	Mozambique	
Bolivia	Finland		Cote d'Ivoire	Myanmar	
Bosnia and Herzegovina	France		Djibouti	Nepal	
Botswana	Georgia		Equatorial Guinea	Niger	
Brazil	Germany		Eritrea	Nigeria	
Brunei Darussalam	Greece		Eswatini	Pakistan	
Bulgaria	Grenada		Ethiopia	Papua New Guinea	
Cabo Verde	Guatemala		Gabon	Rwanda	
Canada	Guyana		Gambia, The	Sao Tome and Principe	
Chile	Honduras		Ghana	Senegal	
China	Hong Kong		Guinea	Solomon Islands	

Table 8: Agglomerative Hierarchical clusters

Cluster 1		Cluster 2		Cluster 3
Afghanistan	Sierra Leone	Albania	Lebanon	Australia
Angola	Solomon Islands	Algeria	Libya	Austria
Bangladesh	Somalia	Antigua and Barbuda	Malaysia	Belgium
Benin	South Sudan	Argentina	Maldives	Canada
Bhutan	Sudan	Armenia	Mauritius	Denmark
Burkina Faso	Tanzania	Azerbaijan	Mexico	Estonia
Burundi	Timor-Leste	Bahamas, The	Micronesia, Fed. Sts.	Finland
Cambodia	Togo	Bahrain	Moldova	France
Cameroon	Uganda	Barbados	Mongolia	Germany
Central African Republic	Vanuatu	Belarus	Montenegro	Hong Kong SAR, China
Chad	Yemen, Rep.	Belize	Morocco	Hungary
Comoros	Zambia	Bolivia	Namibia	Iceland
Congo, Rep.	Zimbabwe	Bosnia and Herzegovina	Nicaragua	Ireland
Congo, Dem. Rep.		Botswana		Israel
Cote d'Ivoire		Brazil		Italy
Djibouti		Brunei Darussalam		Japan
Equatorial Guinea		Bulgaria		Korea, Rep.
Eritrea		Cabo Verde		Latvia
Eswatini		Chile		Lithuania
Ethiopia		China		Luxembourg
Gabon		Colombia		Malta
Gambia, The		Costa Rica		Netherlands
Ghana		Croatia		New Zealand
Guinea		Cuba		Norway
Guinea-Bissau		Cyprus		Poland
Haiti		Czechia		Portugal
India		Dominica		Singapore
Kenya		Dominican Republic		Slovenia
Kiribati		Ecuador		Spain
Korea, Dem. People's Rep.		Egypt, Arab Rep.		Sweden
Lao PDR		El Salvador		Switzerland
Lesotho		Fiji		United Kingdom
Liberia		Georgia		United States
Madagascar		Greece		
Malawi		Grenada		
Mali		Guatemala		
Mauritania		Guyana		
Mozambique		Honduras		
Myanmar		Indonesia		
Nepal		Iran, Islamic Rep.		
Niger		Iraq		
Nigeria		Jamaica		
Pakistan		Jordan		
Papua New Guinea		Kazakhstan		
Rwanda		Kuwait		
Senegal		Kyrgyz Republic		