Statistical analysis of the impact of climate change on the development of the countries in the world

Author:
Jugoslav Stojcheski
stojcheski.jugoslav@students.finki.ukim.mk

Supervisor:
Ljupco Kocarev, Academician PhD
lkocarev@manu.edu.mk

Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, Republic of Macedonia

09.11.2018

Abstract

Mankind nowadays lives better than in any given period of time in the past and is developing itself with the highest rate so far. *The World Bank* estimates that in the period from 1990 to 2015, the global literacy rate increased by 11% (from 75% in 1990 to 86% in 2015), and in the same period the global extreme poverty rate decreased by 26% (from 36% in 1990 to 10% in 2015). [1] The predictions by the major world organizations say that these trends are expected to continue in the future and that the extreme poverty will be eradicated by 2030, but meanwhile warn that climate change is an acute threat to the development of the countries in the world.

This paper applies machine learning techniques to analyze and reveal which measures (according to the measures included in the *UNDP*'s *Human Development Report 2018*) are the best descriptors of the development of the countries in the world and to what extent climate change affects their development.

Contents

1 Introduction 3
- 1.1 Description of the problem . . . 3
- 1.2 Difference between drought and aridity . . . 3
- 1.3 Aridity Index . . . 3

2 Datasets 4
- 2.1 Human Development Report 2018 [UNDP] . . . 4
  - 2.1.1 Brief description . . . 4
  - 2.1.2 Problematic attributes . . . 5
  - 2.1.3 Problematic countries . . . 5
- 2.2 CRU CY v4.01 [Climatic Research Unit @ University of East Anglia] . . . 5
  - 2.2.1 Brief description . . . 5
  - 2.2.2 Derivation of the measurement values for Aridity Index . . . 6
  - 2.2.3 Problematic countries . . . 6

3 Data preprocessing 6
- 3.1 ISO standardization of country names . . . 6
- 3.2 Standardization of attribute values . . . 6

4 Approach to solving the problem 7
- 4.1 Selection of time period for work . . . 7
- 4.2 Selection of attributes for work . . . 7
- 4.3 Dismissing the attribute groups . . . 8
- 4.4 Finding the significance of attributes by applying the Connection Weights algorithm 8
  - 4.4.1 Architecture of the neural network . . . 8
  - 4.4.2 Connection Weights . . . 9
- 4.5 Formation of the graphs . . . 10
  - 4.5.1 Calculation of distance between two countries . . . 10
- 4.6 Analysis of the graphs . . . 11
  - 4.6.1 Weighted Stochastic Block Model . . . 11
  - 4.6.2 Affinity Propagation . . . 13

5 Presentation of the obtained results 14
- 5.1 Results obtained by applying the Connection Weights algorithm . . . 14
- 5.2 Results obtained by applying WSBM . . . 16
- 5.3 Results obtained by applying the Affinity Propagation method for grouping . . . 20
- 5.4 Comparison of the results obtained by applying the WSBM and the Affinity Propagation method . . . 21

6 Summary 24

References 25

7 Appendix 26
- 7.1 A list of problematic attributes of the HDR data set . . . 26

1 Introduction

According to many of the great world organizations, the future ahead (both close and distant future) brings us many challenges. The *United Nations* predict that in 2050 the world's population will be between 9.4 and 10.2 billion people (nowadays this figure is about 7.7 billion) and that in that year more than 5 billion people could suffer from water shortages due to the growing demand for water and climate change. The report of the *World Water Forum* of 2018 says that people use about 4600 cubic kilometers of water each year. The global water demand has increased 6-fold over the past 100 years and this trend continues to grow at a rate of about 1% each year. Water demand is expected to increase fastest in developing countries, and climate change will have an additional negative impact on their development.

1.1 Description of the problem

Basically, this paper seeks the connection between the development of the countries in the world and the climate in which those countries exist. The idea is to find out the significance of each measure that measures the development of the countries in the world and to use their quantitative values to find out to what extent two countries differ in their level of development. Once a quantitative measurement of differences between the countries according to the level of development is done, they should be grouped into blocks, such that countries belonging to a same block would have similar characteristics.

Hence, the posed question is: *Does the grouping of the countries according to their level of development corresponds significantly to the grouping of the countries according to the values of their Aridity Index as an indicator of the climate of a country?* This question could be answered affirmatively, if a significant number of countries belonging to a same block (grouped by level of development) have similar values for their *Aridity Index*. In other words, the smaller the difference of the values of *Aridity Index* between the countries belonging to a same block, the more the *Aridity Index* measure would be a common feature of those countries and it could determine their level of development.

1.2 Difference between drought and aridity

Although the words "drought" and "aridity" have similar meanings and are very often related as synonymous, technically they represent two different phenomenons. [11]

Drought is a phenomenon occurring in a certain period of time in which abnormal dry weather prevails long enough to cause serious hydrological imbalance. Usually, the term "drought" refers to the moisture balance in one region and it occurs in a short period of time. In other words, if the supply of water is less than the demand for water for some period of time (there is a serious hydrological disorder), then that time period is said to be abnormally dry, i.e. that drought occurred as phenomenon. However, this phenomenon is temporary, and once the dry period has ended, the climate returns to normal.

Aridity is a phenomenon that represents the degree to which the climate lacks adequate moisture to promote life in one region. Basically, this phenomenon takes into account the balance between water supply (precipitation) and the water demand (evapotranspiration) over a longer period of time. If the water demand is greater than the water supply (on average), then the climate is said to be arid. Unfortunately, aridity as phenomenon is considered permanent and (until now) no case of climate change to normal after the occurrence of this phenomenon is found.

1.3 Aridity Index

Aridity Index is a numerical indicator of the level of dryness of the climate in a given region. This indicator identifies the regions suffering from water scarcity, a condition that can seriously affect the effective use of the land for activities such as farming and livestock breeding. The boundaries that define the classification according to different degree of aridity [5] and the percentage of Earth's territory covered by that degree of aridity is shown in Table 1. The distribution of the arid regions on Earth is given in Figure 1.

Classification	Aridity Index	Global land area (%)
Hyperarid	$0.00 < \mathrm{AI} < 0.05$	7.5%
Arid	$0.05 < \mathrm{AI} < 0.20$	12.1%
Semi-arid	$0.20 < \mathrm{AI} < 0.50$	17.7%
Dry subhumid	$0.50 < \mathrm{AI} < 0.65$	9.9%

Table 1: Classification of Earth's regions according to their degree of aridity and the percentage of land coverage according to that classification.

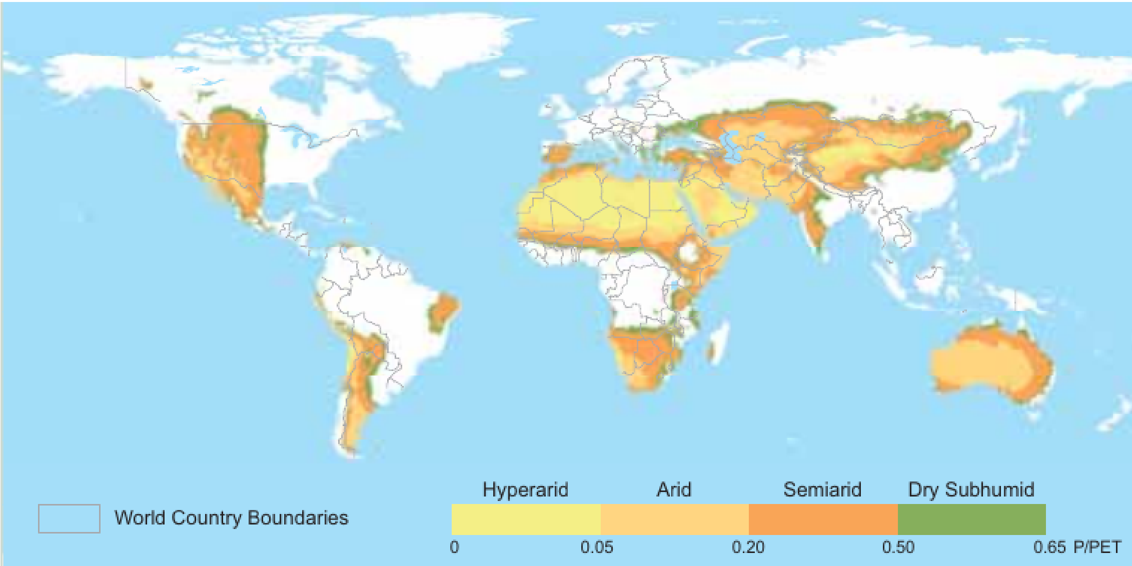

Figure 1: Distribution of arid regions on Earth.

2 Datasets

Within this paper, two datasets are used. The UNDP's *Human Development Report* dataset contains information on the level of development of countries in the world, and the other dataset (*CRU CY* by the *Climatic Research Unit of the University of East Anglia*) contains information on the climate of the countries in the world. A brief description of both datasets follows.

2.1 Human Development Report 2018 [UNDP]

2.1.1 Brief description

The *Human Development Report 2018* (hereinafter: *HDR*) dataset contains information on up to 196 states[1] and 140 attributes divided into 12 groups, as well as 1 attribute that is derived from them (*Human Development Index*). The groups of attributes (and the number of attributes that the group contains, in parentheses) are the following:

- *Demography* (9)
- *Education* (24)
- *Environmental sustainability* (9)
- *Gender* (23)
- *Health* (15)

[1] There are missing values for some attributes and countries.

- *Human Security* (7)
- *Income/Composition of Resources* (6)
- *Inequality* (12)
- *Mobility and Communication* (6)
- *Socio-economic sustainability* (12)
- *Trade and Financial Flows* (5)
- *Work, employment and vulnerability* (12)

In order to avoid the view that the development of countries is a consequence strictly of their economic development, the values of the *Human Development Index* are calculated as the average achievement in three key dimensions of human development: long and healthy life, education, and a decent standard of living. The *Human Development Index* is calculated as geometric mean of the normalized indices for each of the three dimensions. As a result of this calculation, two countries may have the same value for their *Gross National Income*, but also to have a different value for the *Human Development Index*. An illustrated overview of the way the *Human Development Index* value is calculated is presented in Figure 2.

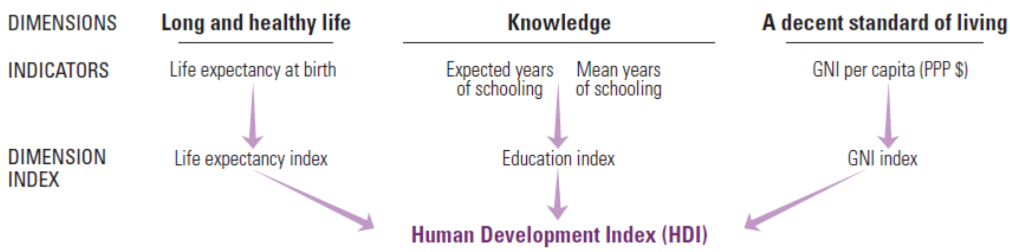

Figure 2: The way in which the *Human Development Index* value is calculated.

2.1.2 Problematic attributes

The *HDR* dataset contains 18 attributes for which only one record in a given time period is present. These attributes are removed from the dataset, because they cannot be used as information that extends over a period of time. After their removal from the dataset, there are total of 123 attributes left (122 independent + *Human Development Index*). The complete list of attributes that are removed from the dataset is given in the Appendix.

2.1.3 Problematic countries

There are 3 countries in the *HDR* dataset for which there are no entries in the *CRU* dataset (presented in subsection 2.2): *Guangdong* (in *China*), *Micronesia (Federated States of)* and the *State of Palestine*. There are also 7 countries that do not have records for their *Human Development Index*, and they are: *Marshall Islands*, *Monaco*, *Nauru*, *North Korea*, *San Marino*, *Somalia* and *Tuvalu*.

2.2 CRU CY v4.01 [Climatic Research Unit @ University of East Anglia]

2.2.1 Brief description

The *Climate Research Unit CY v4.01* [10] dataset (hereinafter: *CRU*) contains information for up to 292 states/territories[2] and 10 attributes. The attributes that exist in this dataset (as well as their metric units, in parentheses) are:

[2]There are attributes for which no measurements for all countries/territories exist, or there are measurements only from a specific time point onwards.

- *Cloud Cover* (%): percentage of the sky obscured by clouds when observed from a particular location; [11]
- *Diurnal Temperature Range* (°C): difference between maximum daily and minimum daily temperature;
- *Ground Frost Frequency* (days): number of days on which the upper layer of the soil was frozen; [2]
- *Maximum Temperature* (°C);
- *Mean Temperature* (°C);
- *Minimum Temperature* (°C);
- *Potential Evapotranspiration* (mm/day → mm/month): amount of evaporation that occurred if sufficient water quantity was available.
- *Precipitation* (mm/month): any product of condensation of atmospheric water vapor which falls under the influence of gravity (i.e., rainfalls etc.);
- *Rainy Days* (days): number of precipitation days over one period of time (month, quarter, year);
- *Vapour Pressure* (hPa): atmospheric pressure that occurs as a result of evaporation.

2.2.2 Derivation of the measurement values for Aridity Index

Since this dataset does not contain explicit information on the *Aridity Index*, its values must be derived from the information on *Precipitation* and *Potential Evapotranspiration.* Theoretically, the equation for calculating the *Aridity Index* values is as follows:

$$\text{Aridity Index} = \frac{\text{Precipitation}}{\text{Potential Evapotranspiration}}$$

In order to be able to derive the measurement values for the *Aridity Index*, the values of the *Potential Evapotranspiration* measurements need to be transformed from mm/day to mm/month to equalize the metric unit with the measurements for *Precipitation.*

2.2.3 Problematic countries

From all the countries that are present in the *HDR* dataset, the following countries are missing all the records for their *Aridity Index* in the *CRU* dataset: *Kiribati*, *Maldives*, *Malta*, *Marshall Islands*, *Nauru*, *St. Lucia* and *Tuvalu.*

3 Data preprocessing

3.1 ISO standardization of country names

For easier merging of both datasets (*HDR* and *CRU*), since they come from different sources, the *ISO 3166-1 alpha-3* standard is used for naming the countries.

3.2 Standardization of attribute values

- *Max-Min* standardization of attribute values represents transformation of the values of each attribute to the interval [0, 1] by applying the following equation:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

  - x: attribute value **before** standardization
  - x': attribute value **after** standardization

– $min(x)$: **minimum** of all values for the attribute

– $max(x)$: **maximum** of all values for the attribute

- *Z-score* standardization of attribute values represents transformation of the values of each attribute by using their corresponding mean (μ) and standard deviation (σ) values. The calculation of the standardized value (x') is done by applying the following equations:

$$x' = \frac{x - \mu}{\sigma} \qquad \mu = \sum_{n=1}^{N} \frac{x_n}{N} \qquad \sigma^2 = \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{N - 1}$$

4 Approach to solving the problem

4.1 Selection of time period for work

Figure 3 shows the number of countries, the number of attributes, and the density (number of present values expressed in %) of the attributes for which there are records in a given year. As the most favorable years for work, the period between 2010 and 2016 is chosen because the dataset for those years has (at least one attribute that has) data for all 195 countries, from 102 to 116 total number of attributes and relatively high density (from 77% to 82%).

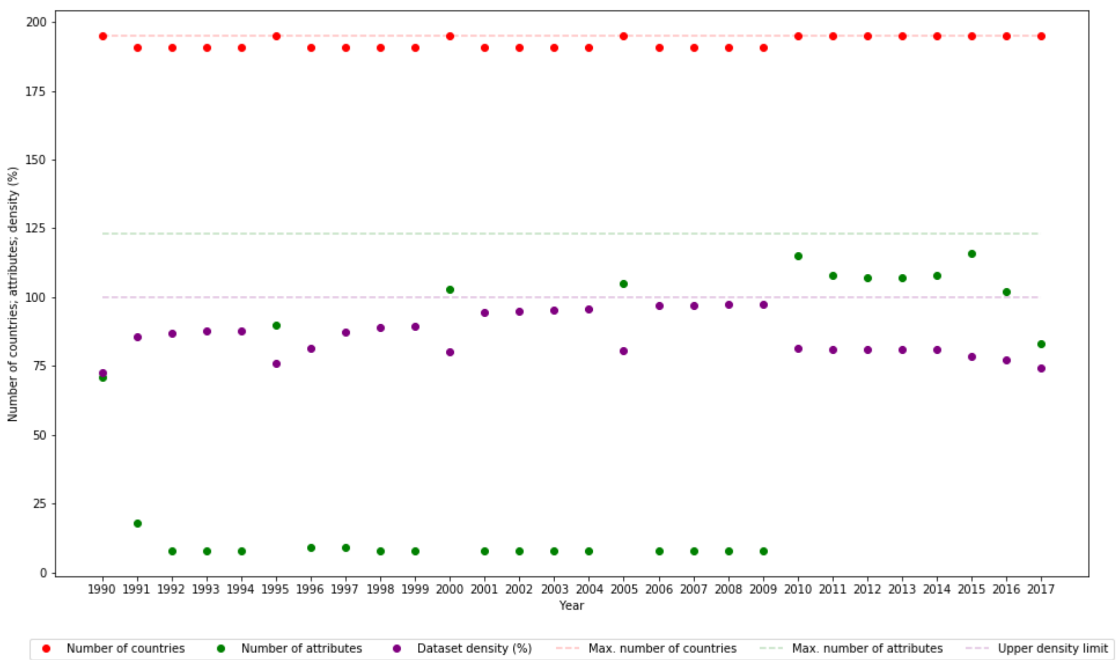

Figure 3: Number of states, attributes, and density (%) for each year in the period from 1990 to 2017 for the *Human Development Report 2018* dataset.

4.2 Selection of attributes for work

Out of the total of 140 attributes that exist in the *HDR* dataset, there are 98 mutual attributes for all years between 2010 and 2016, so it is chosen to use them in order to make a comparison of the values of significance of the attributes obtained by applying the *Connection Weights* algorithm (described in subsubsection 4.4.2) in two arbitrary years, since any change in the set of attributes would also affect the values of the attributes' significance. As expected, this step causes loss of information that the discarded attributes carry, because some attribute that is not taken into account may be of great significance, and this action may cause misinterpretations of the results.

In the remainder of the text, we work only with these attributes, such that the *Human Development Index* is used as an "output" attribute (dependent variable), and the remaining 97 attributes are used as "input" attributes (independent variables).

4.3 Dismissing the attribute groups

In both most recent versions of the *HDR* dataset (*HDR 2016* and *HDR 2018*), a crossing of some attributes from one group to another group can be noticed. For example, the attributes *Violence against women ever experienced, intimate partner (%)* and *Violence against women ever experienced, nonintimate partner (%)* in the version of the dataset from 2016 belong to the *Human Security* attribute group, and in the version of the dataset from 2018 belong to the *Gender* attribute group. Therefore, we can conclude that the attributes do not belong to a particular group of attributes uniquely, so the division of attributes into mutually exclusive groups is completely dismissed as a consequence of this phenomenon.

4.4 Finding the significance of attributes by applying the Connection Weights algorithm

One of the most important things that we want to learn about the *HDR* dataset is the individual significance of each attribute for each year. The theoretical equation (described in subsection 2.1) which calculates the *Human Development Index* value uses only few attributes from the dataset to calculate the theoretical value. In our case, the interest is focused on how to obtain the *Human Development Index* value, if all the attributes of the *HDR* dataset are included in the calculation. One of the ways in which this can be done is by applying the *Connection Weights* algorithm [14] on neural networks. The essence of this procedure is to calculate (predict) the value of the "output" attribute (*Human Development Index*), given the values of the "input" attributes. A description of the architecture of the neural network, as well as the way the *Connection Weights* algorithm works follows.

4.4.1 Architecture of the neural network

First of all, the architecture of the neural network is described, through which the application of the *Connection Weights* algorithm reveals the value of significance of each attribute. Learning the significance of each attribute can be modeled by using a two-layer fully-connected feed-forward neural network trained by applying the *backpropagation* algorithm [17] in which:

- the input layer ($\mathbf{x}$) is composed of 97 neurons, representing all input attributes (independent variables);
- the hidden layer ($\mathbf{h}$) is composed of 12 neurons[3];
- the output layer ($\mathbf{y}$) is composed of 1 neuron, representing the output attribute (*Human Development Index*).

As already mentioned, the network is fully-connected, that is, there are connections between each pair of neurons in two adjacent layers. In addition, a so-called bias neuron for the hidden and the output layer is included.

The weight of the connections between the input layer ($\mathbf{x}$) and the hidden layer ($\mathbf{h}$) are elements of a matrix denoted as $W_{x,h}$ that has dimensions (97, 12), and the weight of the connections between the hidden layer ($\mathbf{h}$) and the output layer ($\mathbf{y}$) are elements of a matrix (more precisely, vector) $W_{h,y}$ that has dimensions (12, 1). The values of the bias neuron in the hidden layer is an element denoted as b_h, and the value of the bias neuron in the output layer is an element denoted as b_y.

[3]The number of neurons in the hidden layer is a hyperparameter of the architecture of a neural network. In our case, the number of neurons in the hidden layer is chosen to be 12, because that is the number of attribute groups that exist in the *HDR* dataset.

As an activation function, the *sigmoid* activation function is used, defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

As a cost function, the *mean squared error* (*MSE*) is used, defined as:

$$C(\vec{y}, \vec{t}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2 \tag{2}$$

- N: total number of countries
- $\vec{y}$: output values of the neural network for the *Human Development Index*
- $\vec{t}$: target values for the *Human Development Index*

Other parameters (and their specific values in our case) relevant to the neural network are:

- (Constant) learning rate: 1;
- Maximum number of iterations: 100,000.

4.4.2 Connection Weights

One way to discover the significance of each attribute within a single dataset is by applying the *Connection Weights* [14] algorithm to the neural network.

Before iterating the *Connection Weights* algorithm, it is necessary to transform the values of the input attributes (independent variables) with the *Z-score* standardization method, and the values of the output attribute (dependent variable) with the *Max-Min* standardization method, methods described in subsection 3.2. The values of the output attribute are transformed with the *Max-Min* standardization method to match the range of the output values produced by the *sigmoid* activation function in the output layer.

After performing these transformations, it is possible that some of the values of the records for the input attributes have a "true" value of 0. This is a problem, because these values will not affect the input signal (will not participate) in the calculations made by the neural network. This problem can be easily captured through the following example.

Example: Let's say that the value of an arbitrary input attribute $x_k = 0$. Thus, the calculations of the value of an arbitrary neuron h_j (before "activation") give that

$$\begin{aligned} h_j &= \sum_{i=1}^{|X|} w_{i,j} \cdot x_i + b_h \\ &= w_{1,j} \cdot x_1 + w_{2,j} \cdot x_2 + \cdots + w_{k,j} \cdot x_k + \ldots + w_{|X|,j} \cdot x_{|X|} + b_h \\ &= w_{1,j} \cdot x_1 + w_{2,j} \cdot x_2 + \cdots + w_{k,j} \cdot 0 + \ldots + w_{|X|,j} \cdot x_{|X|} + b_h \end{aligned}$$

In order to avoid this problem, the entries for input data having a "true" value of 0 are replaced with an arbitrary value close to 0.[4]

Another problem that may occur among input data is missing values in the dataset.[5] If there is no appropriate method for imputation, it is desirable that these values do not have any effect when calculating the values of the neurons in the neural network (as well as the significance of the attributes), so these values are replaced with a "fake" value of 0. Here, it is important to note that the order of solving these two problems is important, because it may happen to replace some "fake" values of 0 with an arbitrary value close to 0, which is an undesirable occurrence.

[4] In our case, the value of $1 \cdot 10^{-9}$ is chosen.

[5] In our case, about 20% of data is missing for each year from 2010 to 2016.

Also, if there is no appropriate method for imputation, items that miss values for the output attribute are dismissed from the dataset during the training process of the neural network, that is, these items are not taken into account when calculating the significance of the attributes.

Before iterating the neural network, the initial weights of the network connections are initialized with values drawn from a uniform continuous distribution defined on the interval $[-c, +c]$ ($c \in \mathbb{R}$)[6]. Also, between two consecutive iterations during training of the neuronal network, the items in the dataset are mixed in a random manner.

During the process of optimization, (in theory) it is important that the neural network converges to the global minimum of the cost function. However, in practice this is almost impossible to obtain, so one of the ways to approximate the values of significance of the attributes (that would be obtained when the neural network would reach the global minimum) is by combining the values of significance of the attributes obtained by combining multiple local minimum (rather than choosing one of them as "closest" to the global minimum), thus calculating the average value of the attributes that are acquired as a result of the application of the *Connection Weights* algorithm for multiple random starts. The values of significance of the attributes ($\vec{\mu}$) for one random start are calculated by the following equations:

$$\vec{\mu} = |W_{x,h} \cdot W_{h,y}| \qquad\qquad \mu_x = |\sum_{h=1}^{|H|} W_{x,h} \cdot W_{h,y}| \tag{3}$$

In Equation 3, the absolute value of the sum (used in the *Connection Weights* algorithm) is calculated rather than the sum of absolute values (used in the *Garson's* algorithm [7]), because this way of calculation gives a better estimate of significance of the attributes. To demonstrate the superiority of the *Connection Weights* algorithm over the *Garson's* algorithm, comparisons between the two algorithms on a synthetically-generated dataset has been done. [15]

The whole process is repeated for each year of the selected time period (from 2010 to 2016), taking the values for all 195 countries and all 98 attributes (97 input attributes and 1 output attribute) that appear as mutual in this time period.

4.5 Formation of the graphs

By using the significance of the attributes ($\vec{\mu}$) obtained by applying the *Connection Weights* algorithm to neural networks, an undirected weighted graph for each year is formed, such that each node represents a single country, and the value of the weight of each edge between two nodes in the graph is the distance between those two countries. The greater the value of the edge weight between those two countries, the more those two countries are different from each other.

4.5.1 Calculation of distance between two countries

In order to be able to compare the values of distances between countries as values calculated for two distinct years, it would be desirable the values of the attributes and the values of distances to be limited to a fixed interval of values. One of the most intuitive intervals is the interval $[0, 1]$, so the values of the attributes are mapped to this interval by applying the *Max-Min* standardization method, and as a measure of distance that meets this condition the *Minkowski* distance is used, defined as:

$$M_p(\vec{c_1}, \vec{c_2}) = \left(\sum_{a \in A} |c_{1,a} - c_{2,a}|^p \right)^{\frac{1}{p}} \tag{4}$$

where $\vec{c_1}, \vec{c_2}$ are the values of the attributes of two arbitrary countries from the set of countries C, the parameter $p \in \mathbb{R}$ and A represents the set of attributes mutual[7] to these two countries.

[6] In our case, the value of $c = 0.3$ is chosen.

[7] The term "mutual attributes" refers to attributes for which there are records for both countries in the dataset.

However, each attribute in the HDR dataset does not have same significance. Therefore, in order to take into account the significance of the attributes, normalized weights $\frac{\mu_a}{(\sum_{a\in A}\mu_a)}$ are appended to Equation 4 in order calculate the distance between two countries. After this enrichment the equation looks like this:

$$M_p(\vec{c}_1,\vec{c}_2,\vec{\mu}) = \left(\sum_{a\in A}\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}|^p\right)^{\frac{1}{p}} \tag{5}$$

When choosing the value of the parameter p one should take into account the following two important properties of the equation that calculates the *Minkowski* distance:

$$\lim_{p\to+\infty}\left(\sum_{a\in A}|c_{1,a}-c_{2,a}|^p\right)^{\frac{1}{p}} = \max_{a\in A}|c_{1,a}-c_{2,a}|$$

$$\lim_{p\to-\infty}\left(\sum_{a\in A}|c_{1,a}-c_{2,a}|^p\right)^{\frac{1}{p}} = \min_{a\in A}|c_{1,a}-c_{2,a}|$$

which expressed within Equation 5 look like this:

$$\lim_{p\to+\infty}\left(\sum_{a\in A}\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}|^p\right)^{\frac{1}{p}} = \max_{a\in A}\left\{\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}|\right\}$$

$$\lim_{p\to-\infty}\left(\sum_{a\in A}\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}|^p\right)^{\frac{1}{p}} = \min_{a\in A}\left\{\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}|\right\}$$

Given that Equation 5 already introduces values that alter the significance of the attributes (relative to the initial equalized values), any value of the parameter $p\neq 1$ would additionally alter the values of significance of the attributes, a phenomenon that is not desirable at all. Therefore, the value of the parameter p is set to 1, and the distance between two arbitrary countries $\vec{c_1},\vec{c_2}$ is calculated as a normalized weighted sum of differences across all attributes mutual to those two countries by applying the following equation:

$$M_1(\vec{c}_1,\vec{c}_2,\vec{\mu}) = \sum_{a\in A}\frac{\mu_a}{(\sum_{a\in A}\mu_a)}\cdot|c_{1,a}-c_{2,a}| \tag{6}$$

4.6 Analysis of the graphs

In order to answer the question posed in subsection 1.1, grouping the countries (represented as nodes in the graph) into blocks[8] and comparing whether the *Aridity Index* values for countries that belong to a same block of countries are similar to each other is done. To provide the grouping of the countries, the *Weighted Stochastic Block Model* [3] algorithm and the *Affinity Propagation* [6] method are used.

4.6.1 Weighted Stochastic Block Model

The *Weighted Stochastic Block Model* [3] (hereinafter: *WSBM*) is a model that learns the distributions that form the edges between two blocks of nodes in the graph by applying the variational Bayes method. Specifically, the *WSBM* models the weight of each edge between two nodes in the graph as a draw from a distribution (which belongs to the parametric-exponential family of distributions) whose parameters depend only on the blocks in which these two nodes belong. All nodes belonging to a same block are stochastically equivalent[9]. This characteristic denotes their

[8] We use the words blocks, groups and clusters interchangeably.

[9] Two random variables X_1 and X_2, defined in the probability space $(\Omega,\mathcal{F},\mathrm{P})$, are stochastically equivalent if $\Pr(X_1=X_2)=1$.

equivalent role in generating the graph structure and they maintain the same probabilistic connectivity to the rest of the graph.

Within the model the following parameters are defined:

- K is the parameter that represents the number of blocks. This parameter determines the complexity of the model and its optimal value is determined by the value of log-probability (i.e. the optimal number of blocks corresponds to the maximum value for log-probability);
- z is a vector with length $|C|$ and represents the vector of belongings of the nodes by blocks. The value of the element $z_i \in \{1, \ldots, K\}$ represents the block to which the node i belongs;
- θ is a matrix with dimensions $K \times K$ which contains the parameters that determine the distribution of the edge weights between two blocks. The value of the element $\theta_{i,j}$ represents the parameters that specify that govern the weight distribution of the (z_i, z_j) edge bundle;
- A is a matrix with dimensions $|C| \times |C|$ which contains the values of the edge weights in the graph. It is determined by the parameter θ_{z_i,z_j} and depends only on the blocks in which the nodes i and j belong. The value of the element $A_{i,j}$ is the edge weight between node i and node j. Theoretically, the value of the weight of an arbitrary edge $A_{i,j} = 0$ can have three different meanings:
  1. absence of an edge
  2. an edge that exists, but has weight 0
  3. missing data, i.e. an unobserved interaction

The distributions that model the existence of the edges and the edge weights in the graph belong to the parametric-exponential family of distributions:

$$\Pr(A \mid z, \theta) \propto \exp\Big(\sum_{i,j} T(A_{i,j}) \cdot \eta(\theta_{z_i,z_j})\Big)$$

The graphical model of the *WSBM* is shown in Figure 4.

By selecting different appropriate pairs of functions (T, η), a different stochastic block model (*SBM*) can be specified, whose values of the edge weights are drawn from the distribution belonging to the parametric-exponential family of distributions. If the pair (T_e, η_e) denotes the family of distributions that determine the existence of an edge in the graph and the pair (T_w, η_w) denotes the distribution family that determines the value of a single edge weight, then they can be combined in likelihood function with parameter $\alpha \in [0, 1]$ which determines their relative importance in inference

$$\log \Pr(A \mid z, \theta) = \alpha \sum_{i,j \in E} T_e(A_{i,j}) \cdot \eta_e(\theta^{(e)}_{z_i,z_j}) + (1 - \alpha) \sum_{i,j \in W} T_w(A_{i,j}) \cdot \eta_w(\theta^{(w)}_{z_i,z_j}) \qquad (7)$$

where E is a set of observed interactions, and W is a set of weighted edges ($W \subset E$). Various hidden latent structures can be obtained for different values of the parameter α. In the case when $\alpha = 1$, the model (called *SBM*) ignores the information on the edge weights and treats their existence as binary (either the edge exists or does not exist). In the case where $\alpha = 0$, the model

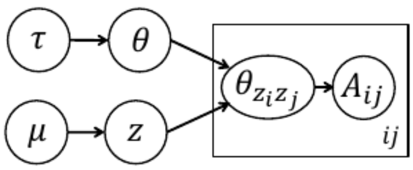

Figure 4: Graphical model of the *WSBM*. Each edge weight $A_{i,j}$ is distributed according to the corresponding parameter θ_{z_i,z_j} for each observed interaction (i, j). In a variational Bayes inference scheme, the latent parameters z and θ are modeled as random variables that are distributed according to μ and τ, respectively.

(called *"pure" WSBM*) treats the absence of an edge as an unobserved interaction and uses only the information on the edge weights. In the case when $0 < \alpha < 1$, Equation 7 combines both information (the existence of the edges and the edge weights).

Parameters that completely determine the *WSBM* are z and θ. Learning these parameters is done by applying a variational Bayes method in which parameters are treated as random variables and an appropriate prior distribution π is set for the parameters z and θ. If the prior distribution is $\pi(z,\theta) = \Pr(z,\theta)$, then the posterior distribution $\pi^*(z,\theta) = \Pr(z,\theta \mid A)$ can be calculated as:

$$\pi^*(z,\theta) \propto \Pr(A \mid z,\theta)\pi(z,\theta)$$

The introduction of the prior distribution π prevents the posterior distribution π^* from overfitting to the degenerate maximum likelihood solution. However, the analytical calculation of the posterior distribution (in general) is difficult, so $\pi^*(z,\theta)$ is approximated with a factorizable distribution $q(z,\theta) = q_z(z)q_\theta(\theta)$. The choice of the approximate distribution q is done by minimizing the *Kullback-Leibler* (*KL*) divergence with respect to the posterior distribution π^*, defined as:

$$D_{\mathrm{KL}}(q \mid\mid \pi^*) = -\int q \log \frac{\pi^*}{q}$$

After calculating the posterior distribution, the beliefs about the values of the parameters z and θ can be accepted as satisfactory or the algorithm can proceed further to calculate the point estimate in order to minimize the expected loss of the posterior distribution with respect to a given loss function. After learning the parameters z and θ, the generating process of formation a weighted graph with the *WSBM* is as follows:

- Each node i is assigned to a block z_i to which the node belongs;
- For each pair of blocks (k,k'), a parameter $\theta_{k,k'}$ is assigned to that edge bundle;
- For each edge (i,j), the value of the edge weight $A_{i,j}$ is derived from a distribution belonging to the parametric-exponential family of distributions (T,η) with parameters θ_{z_i,z_j}.

As a result of this process, a graph is formed in which each pair of nodes is potentially connected with an edge having a real-valued weight.

4.6.2 Affinity Propagation

Affinity Propagation [6] is a grouping method that forms blocks by sending messages between the data points until the whole process converges. The intensity of the messages exchanged between the data points are changed based on simple equations that seek after the minimum of an appropriately-selected energy function[10].

As input data this method receives a matrix of similarities between the data points, where the similarity $s(i,k)$ indicates how well the data point k fits to be an exemplar[11] of the data point i. Usually, the similarities between two data points are calculated as a *negative Euclidean distance*, which for two arbitrary data points x_i and x_k is defined as:

$$s(i,k) = -||x_i - x_k||^2$$

The *Affinity Propagation* method is especially interesting because the number of groups does not have to be predefined and it arises naturally from the dataset on which this method is applied. Before iterating, this method treats each item of the dataset as potential exemplars equally. However, if necessary, it is possible to set weights to the data points (by defining the preference parameter) so that they have a greater/lesser chance of being selected as exemplars.

Within this method there are two different types of messages that are exchanged between the data points, and they are:

[10] An energy function is a measure of stability of a system's configuration in which the system appears to be.
[11] Exemplar is a center of a group that is selected from the items of a dataset.

1. A message of "responsibility" $r(i,k)$ sent from the data point i to the data point k (as a potential exemplar of the data point i) that tells how well the data point k fits to be the exemplar of the data point i, taking into account the other data points that can be its exemplar.

$$r(i,k) \leftarrow s(i,k) - \max_{k';\ k' \neq k} \{a(i,k') + s(i,k')\} \tag{8}$$

2. A message of "availability" $a(i,k)$ sent from the data point k (as a potential exemplar of the data point i) to the data point i that tells how appropriate is the data point i to select the data point k for its exemplar, considering the "support" the data point k receives from other data points to be selected as their exemplar.

$$a(i,k) \leftarrow \min \left\{0,\ r(k,k) + \sum_{i';\ i' \notin \{i,k\}} \max\{0,\ r(i',k)\}\right\} \tag{9}$$

$$a(k,k) \leftarrow \sum_{i';\ i' \neq, k} \max\{0,\ r(i',k)\} \tag{10}$$

At the beginning, the values of the "availability" messages are set to 0 ($\forall_{i,j}\ a(i,k) = 0$) and the values of the "responsibility" messages are calculated by applying Equation 8. Then, the values of the "availability" messages are calculated by applying Equation 9 and Equation 10, and the intensity of the messages change in the next step $(t+1)$ by applying the following equations:

$$r_{t+1}(i,k) = \delta \cdot r_t(i,k) + (1-\delta) \cdot r_{t+1}(i,k)$$

$$a_{t+1}(i,k) = \delta \cdot a_t(i,k) + (1-\delta) \cdot a_{t+1}(i,k)$$

The whole procedure is iterated until the system reaches a state of stability. In addition, when changing the values of messages between two steps, it is important that they are "damped" by a so-called "damping" factor ($\delta \in [0,1]$) to avoid numerical oscillations that may occur in some cases. An illustrated example of how the *Affinity Propagation* method works and the method of sending messages is given in Figure 5 and Figure 6, respectively.

5 Presentation of the obtained results

In this chapter, we present the results obtained by applying the *Connection Weights* algorithm, the *WSBM*, and the *Affinity Propagation* method described in subsubsection 4.4.2, subsubsection 4.6.1 and subsubsection 4.6.2, respectively.

5.1 Results obtained by applying the Connection Weights algorithm

Figure 7 presents the values of the cost function of the neural networks in the last 10 iterations for 20 random starts in the time period from 2010 to 2016. As shown in the figure, the values of the cost function of the neural networks in the time period from 2010 to 2014 are quite stable and they fluctuate in the interval $[0,\ 3 \cdot 10^{-10}]$. The values of the cost function of neural networks in 2015 have small deviations and these values fluctuate in the interval $[0,\ 3 \cdot 10^{-8}]$. The values of the cost function of the neural networks in 2016 have relatively large deviations and they fluctuate in the interval $[0,\ 1 \cdot 10^{-7}]$. However, since all values of the cost function in the last iteration are less than $1 \cdot 10^{-7}$, they are considered small enough to conclude that the local minimum of the neural networks are good-enough approximators of the global minimum and that the values of significance of the attributes that are obtained by applying the *Connection Weights* algorithm are valid.

Figure 8 shows the average values of significance of the attributes for each year in the period from 2010 to 2016. As shown in the figure, the 3 most important attributes in this time period are:

1. *Estimated gross national income per capita, female (2011 PPP$)*
2. *Gross national income (GNI) per capita (2011 PPP$)*
3. *Mean years of schooling (years)*

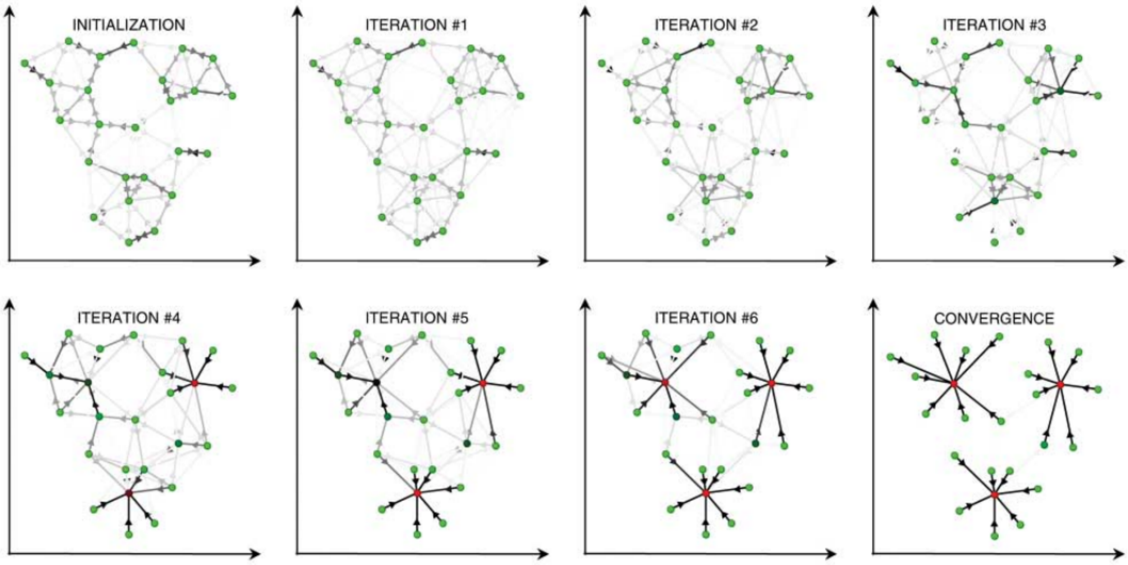

Figure 5: An example of how the *Affinity Propagation* method works through multiple iterations.

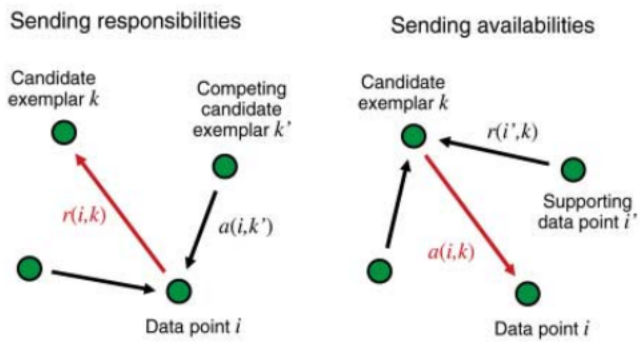

Figure 6: Illustration of how the procedure of sending messages within the *Affinity Propagation* method works.

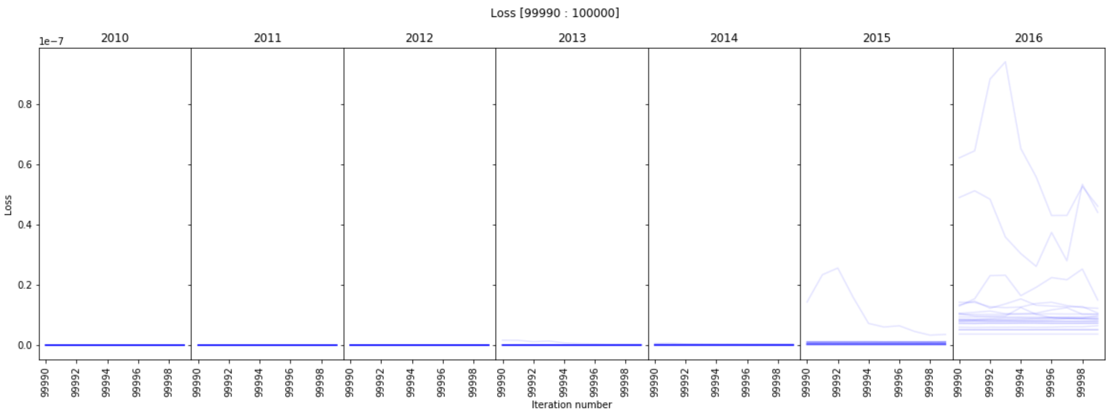

Figure 7: Values of the cost function in the last 10 steps for 20 random starts of the neural network in the time period from 2010 to 2016.

A detailed overview of the values of significance of these attributes (for each attribute individually) in relation to the iterations of the neural networks is shown in Figure 9, Figure 10 and Figure 11, respectively. According to the figures, the values of significance of the attributes *Gross national income (GNI) per capita (2011 PPP $)* and *Mean years of schooling (years)* are fairly stable for all years in the period from 2010 to 2016, i.e. their average value in the limit converges to a particular known value. The values of significance of the attribute *Estimated gross national income per capita, female (2011 PPP $)* are fairly stable for all years in the period from 2010 to 2014, and a growing trend can be noticed[12] for the values of significance of this attribute in 2015 and 2016.

5.2 Results obtained by applying WSBM

Before presenting the obtained results, we will briefly set the parameters that we use within the *WSBM*, i.e. the "mixing" parameter α, as well as the distributions that determine the existence of the edges in the graph and their respective weights.

In our case, the edges that connect two nodes in the graph represent a distance between two countries, i.e. value that represents a normalized weighted sum of differences between those two countries calculated by applying Equation 6. In addition, it is known that each pair of countries has at least one mutual attribute (which enables the calculation of the distance between those two countries), so this leads to the conclusion that the occurrence of the value 0 as an element in the matrix A represents an edge that exists but has a weight of 0. This fulfills the second case (of the cases written in subsubsection 4.6.1) and entails the value of the "mixing" parameter α to be 0.

The value of the "mixing" parameter $\alpha = 0$ neglects the part of Equation 7 which includes the calculation of the distribution that determines the existence of an edge in the graph and completely relies on the distribution that determines the values of the edge weights in the graph, which entails that the *"pure" WSBM* is applied. Additionally, by doing the *Kolmogorov-Smirnov* test of distributions, it turns out that the values of the edge weights in the graph are drawn from gamma distributions and only these distributions are used to model the (weights of the) edges in the graph. A comparison of the histograms of the values of the edge weights in the graphs with their theoretical gamma distribution in the time period from 2010 to 2016 are shown in Figure 13.

Figure 14 presents the dependence of the log-probability value for different number of blocks (K). The maximum value of log-probability at the level of one year gives the optimum value for the number of blocks (K) for that year. In addition, it can be noticed that different values are obtained for different years. Thus, for 2010 and 2011, the optimal number of blocks in which the nodes should be grouped by applying the *WSBM* is 7, and for the years from 2012 to 2016, the optimum number of blocks is 6.

Furthermore, the results for all years in the period from 2010 to 2016 are shown in Figure 12. The following types of plots are presented in the following columns:

1. column: a heatmap showing the edge weights between the nodes in the graph divided into blocks marked by black rectangles;
2. column: scatter plot showing the value of the *Human Development Index* of each country (vertically) divided into blocks (horizontally);
3. column: scatter plot showing the value of the *Aridity Index* of each country (vertically) divided into blocks (horizontally);

In this figure, it can be noticed that the countries are well grouped into blocks according to the values for their *Human Development Index* (i.e., there are no huge overlaps between the groups according to this measure), which is expected to occur, because the distance values between the countries are derived from the *HDR* dataset. However, by comparing the grouping of the countries according to the values for their *Aridity Index* (shown in the third column of Figure 12), huge overlaps between the groups can be noticed. Hence, it can be concluded that the *Aridity Index* is

[12] In the limit it is expected that these values will converge to some particular value, but the final value in our case is not known. Therefore, the average of the values of significance in the last iteration is chosen as a final value.

Figure 8: Average value of normalized significance of the attributes in the time period from 2010 to 2016.

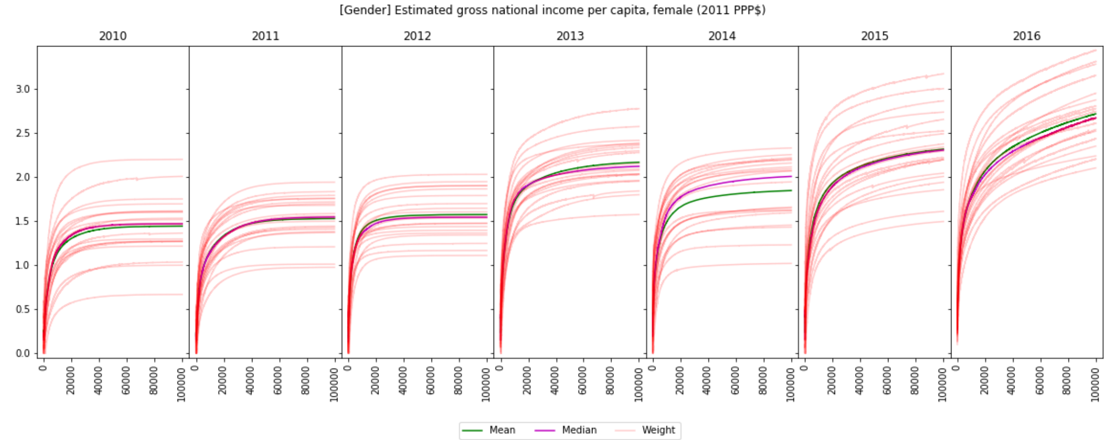

Figure 9: Values of significance of the *Estimated gross national income per capita, female (2011 PPP $)* attribute in relation to the iterations of the neural networks for each year in the period from 2010 to 2016.

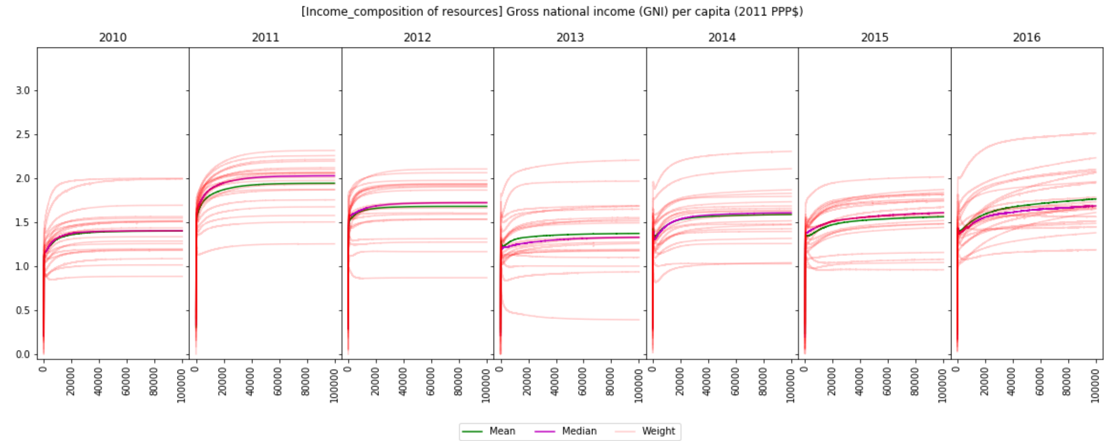

Figure 10: Values of significance of the *Gross National Income (GNI) per capita (2011 PPP $)* attribute in relation to the iterations of the neural networks for each year in the period from 2010 to 2016.

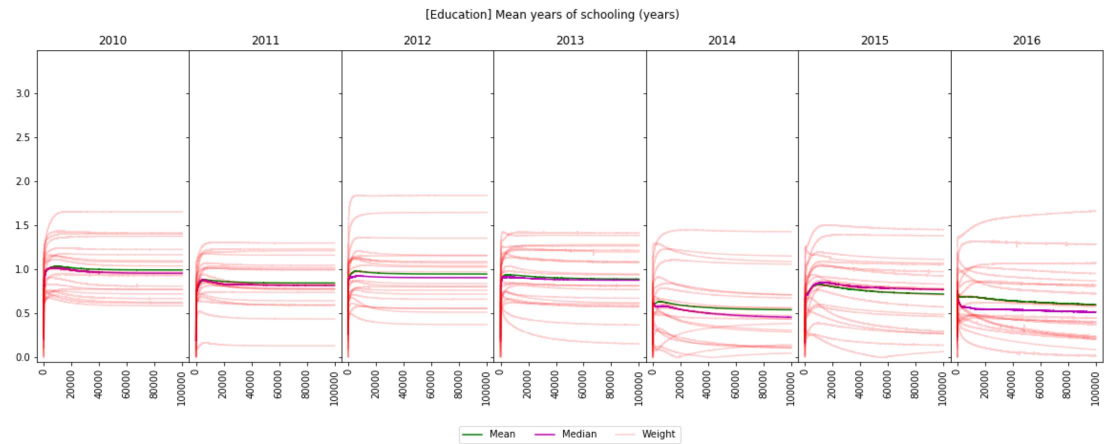

Figure 11: Values of significance of the *Mean years of schooling (years)* attribute in relation to the iterations of the neural networks for each year in the period from 2010 to 2016.

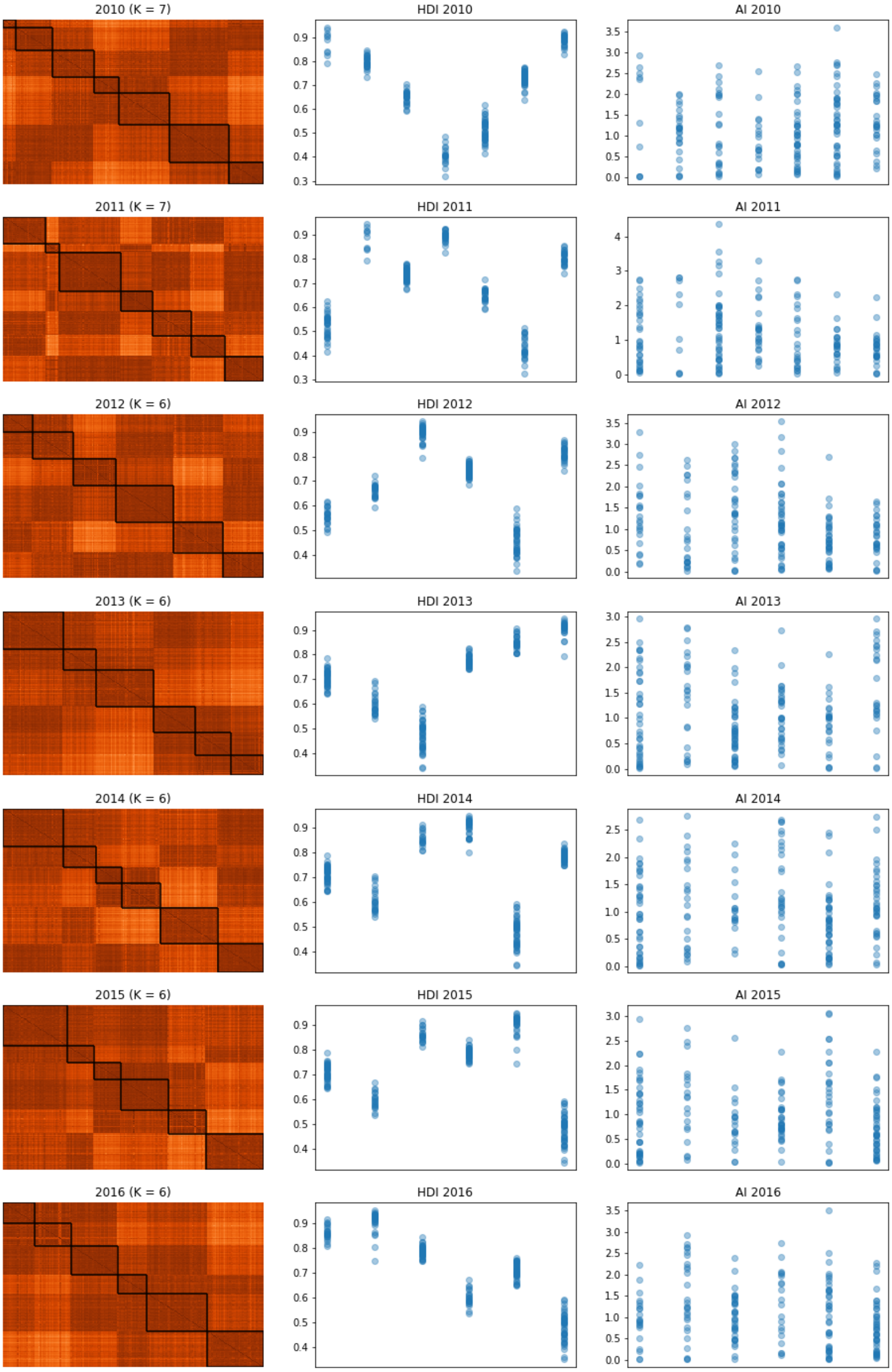

Figure 12: Results obtained by applying the *WSBM* for each year in the period from 2010 to 2016.

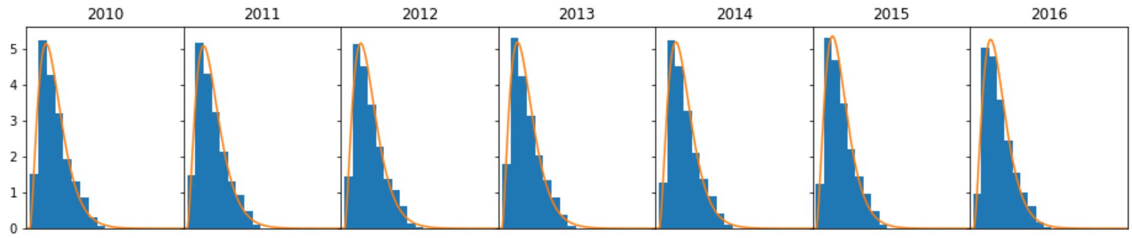

Figure 13: Comparison of the histograms of the edge weights in the graph and their theoretical gamma distributions in the time period from 2010 to 2016.

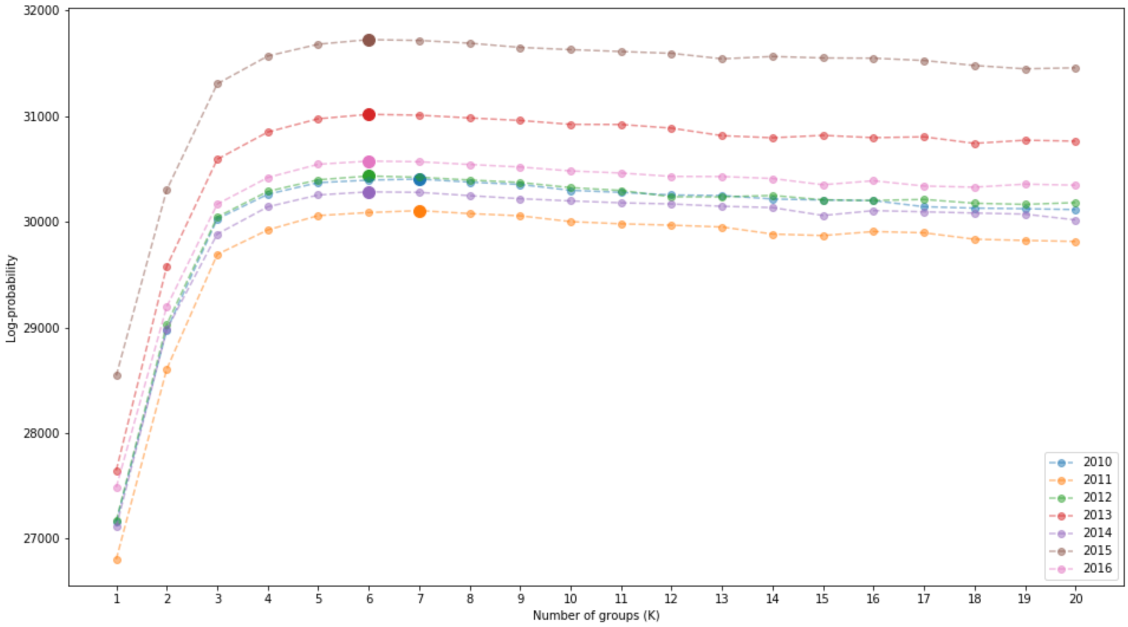

Figure 14: Log-probability values for different number of blocks (K) obtained by applying the *WSBM* for each year in the period from 2010 to 2016. The larger points indicate the optimal value for that year.

not a mutual feature of the countries belonging to a same block, i.e. that there is no relationship between the level of development of the countries with the climate in which those countries exist.

5.3 Results obtained by applying the Affinity Propagation method for grouping

The *Affinity Propagation* method has two parameters that can produce different groupings by changing their value:

- an preference parameter that gives a greater/lesser chance for data points to become exemplars;
- the "damping" factor δ, which makes a balance between the share of the message values calculated in the previous step (t) and the message values calculated in the current step $(t + 1)$.

In this case, given that there is no prior information about which data points should have an advantage over other data points for them to become exemplars, the values of the preference parameter are not taken into account and all data points are treated equally, i.e. they have equal chances to become exemplars.

In order to obtain more accurate results, it is desirable the method to be iterated for different values of the "damping" factor δ. In our case, the "damping" factor is taken to be $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$[13]. The results obtained for the optimum values of the "damping" factor for each year are shown in Figure 15.

Similar to the results obtained by applying the *WSBM* (presented in subsection 5.2), the results obtained by applying the *Affinity Propagation* method tell that the countries are well grouped into blocks according to the values for their *Human Development Index* (i.e. there are no huge overlaps between the groups according to this measure), and that by comparing the grouping of the countries according to the values for their *Aridity Index* (shown in the third column of Figure 15), huge overlaps between the groups can be noticed. The conclusion that can be derived here is the same as the conclusion derived by applying the *WSBM*, i.e. that the *Aridity Index* is not a mutual feature of the countries belonging to a same block, i.e. that there is no relationship between the level of development of the countries with the climate in which those countries exist.

5.4 Comparison of the results obtained by applying the WSBM and the Affinity Propagation method

In *Schaeffer*'s review article on graph clustering [18], several measures for calculating the quality of groupings have been defined. One of the measures (proposed by *Newman* [13]) is the measure of *modularity*. In theory, there are several formulations for the same measure, depending on whether the graph is weighted and whether it is used as a measure of minimization or maximization. In our case, the *modularity* is calculated as follows:

$$M(C_1,\ \dots,\ C_K) = \sum_{i=1}^{K} \epsilon_{i,i} - \sum_{\substack{i \neq j \\ i,j \in \{1,\dots,K\}}} \epsilon_{i,j} \tag{11}$$

where the sum of the values of the edge weights ($\epsilon_{i,j}$) between the two groups is calculated as:

$$\epsilon_{i,j} = \sum_{\substack{\{u,v\} \in E \\ u \in C_i, v \in C_j}} w(u,v)$$

Also, it is important to note that the value of the weight of each edge is calculated only once.

Since the values of the edge weights in the graph are defined as distances between the nodes, the *modularity* (in our case) is used as a measure of minimization. This means that the smaller the value of *modularity* (calculated by applying Equation 11), the better the grouping of the countries into blocks, since it is desirable that the sum of values of the edge weights between the nodes belonging to a same block (calculated as $\sum_{i=1}^{K} \epsilon_{i,i}$) to be as small as possible, and the sum of the values of edge weights between the nodes belonging to two different blocks (calculated as $\sum_{\substack{i \neq j \\ i,j \in \{1,\ \dots,\ K\}}} \epsilon_{i,j}$) to be as large as possible.

The numerical and graphical representation of the *modularity* values obtained by applying the *WSBM* and the *Affinity Propagation* method are shown in Table 2 and Figure 16, respectively. It can be noticed that the *Affinity Propagation* method produces better results than the *WSBM*. One of the reasons for this phenomenon may be the difference in goals to which both methods strive, because the primary purpose of the *WSBM* method is to learn the distributions that model the edges in the graph in order to generate a graph with the learned distributions, while the primary purpose of the *Affinity Propagation* method is to group data points into blocks.

[13]Theoretically, the "damping" factor δ can take any value between 0 and 1, but the implementation of the method in `Python` (more precisely, within the `scikit-learn` package) allows only values between 0.5 and 1 to be used.

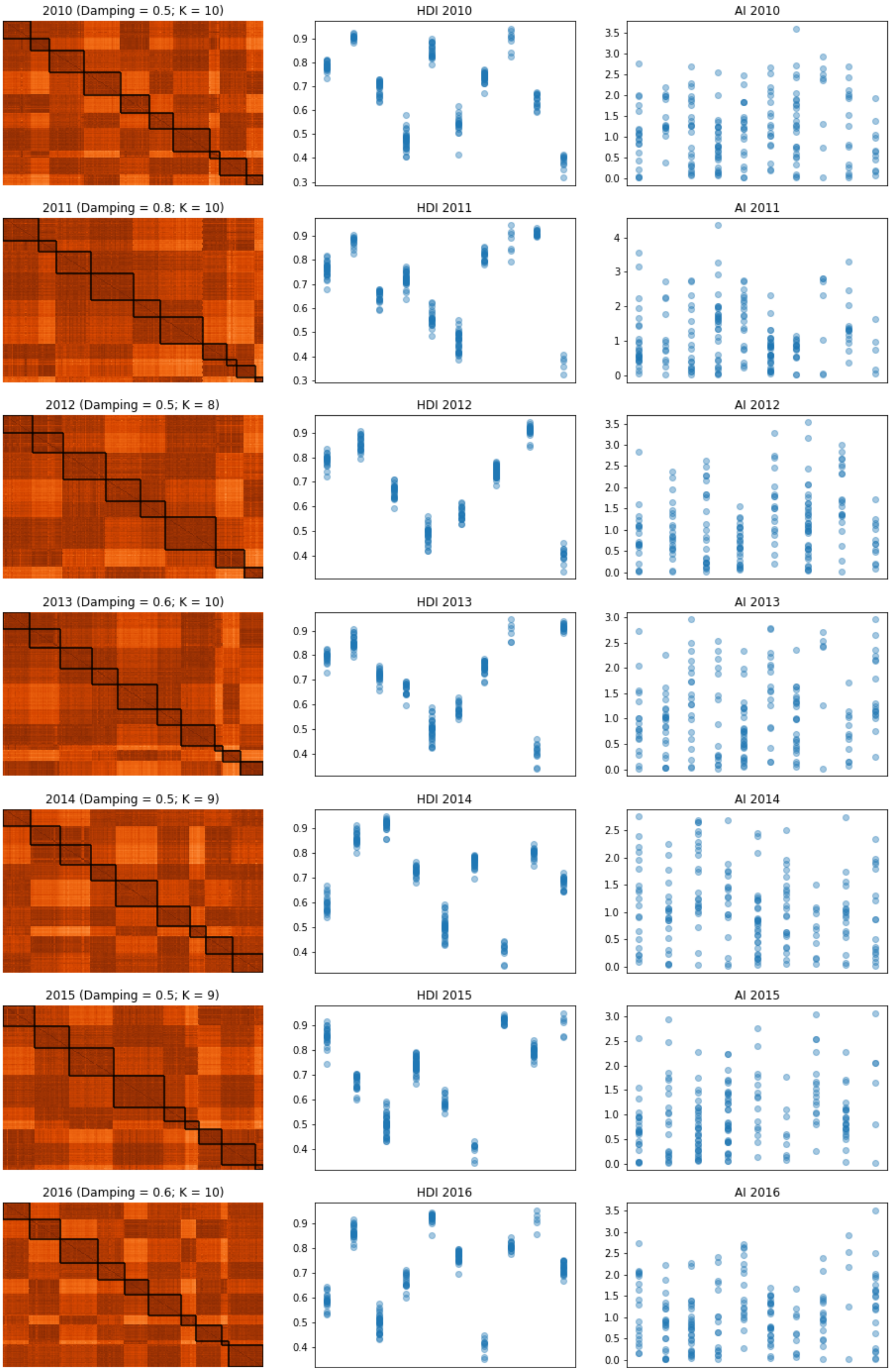

Figure 15: Results obtained by applying the *Affinity Propagation* method for each year in the time period from 2010 to 2016.

Since the quality of grouping by applying the *Affinity Propagation* method is better than the quality of grouping by applying the *WSBM*, the overall optimal results correspond completely to the optimal results obtained by the *Affinity Propagation* method. These results are described in subsection 5.3 and graphically presented in Figure 15.

	2010	2011	2012	2013	2014	2015	2016
AP ($\delta = 0.5$)	**-3059.26**	-2986.44	**-2967.14**	-2958.66	**-3031.68**	**-2864.01**	-2933.43
AP ($\delta = 0.6$)	-3050.67	-2986.44	-2943.08	**-3025.65**	-2991.35	-2857.72	**-3023.12**
AP ($\delta = 0.7$)	-2953.90	-2982.44	-2965.37	-2946.95	-2991.35	-2862.40	-2957.98
AP ($\delta = 0.8$)	-3049.25	**-3054.75**	-2940.22	-2946.95	-2988.05	-2857.72	-2957.98
AP ($\delta = 0.9$)	-3034.91	-2985.72	-2940.22	-2946.95	-2988.05	-2857.72	-2957.98
WSBM	-2872.83	-2917.01	-2823.18	-2763.56	-2775.51	-2653.51	-2751.96

Table 2: Numerical representation of the *modularity* values calculated for the best results obtained by applying the *Affinity Propagation* method (for different values of the "damping" factor δ) and the best overall results obtained by applying the *WSBM* for all years from 2010 to 2016.

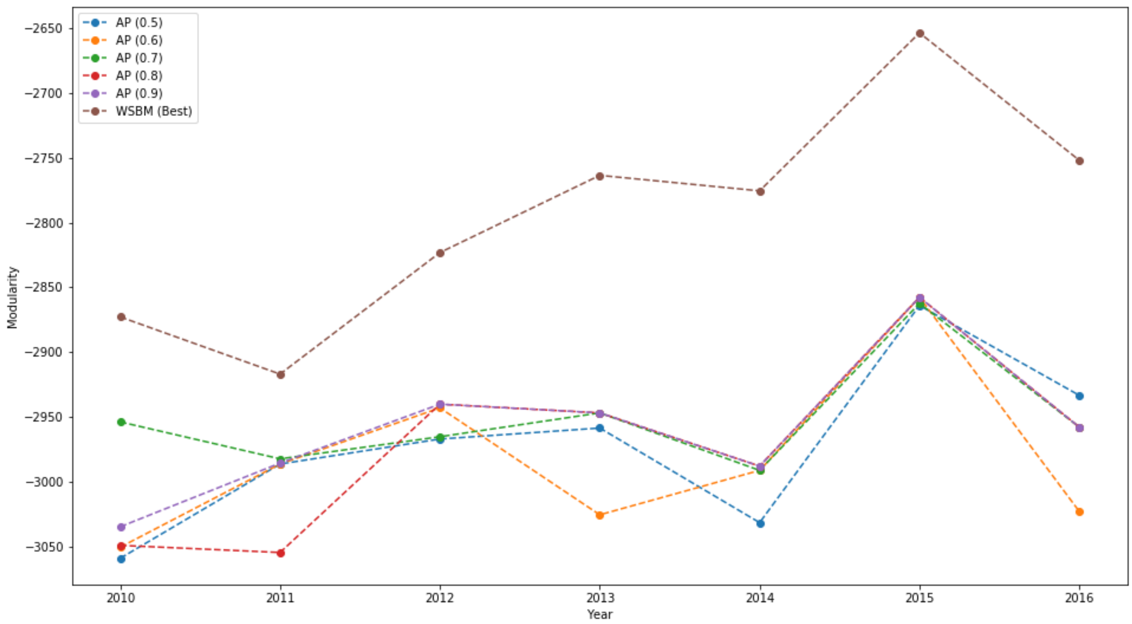

Figure 16: Graphical representation of the *modularity* values calculated for the best results obtained by applying the *Affinity Propagation* method (for different values of the "damping" factor δ) and the best overall results obtained by applying the *WSBM* for all years from 2010 to 2016.

6 Summary

Regarding the methods and algorithms applied in this paper, some alterations to the proposed workflow could be applied in attempt to achieve better results. For example, a longer iteration of the neural networks could be done in order to obtain more accurate values for the attributes' significance; *ReLU* [8, 9] could be used as an activation function for the nodes in the hidden layer of the neural networks; kernel methods could be used to calculate the values of distance between two countries; *Chinese Restaurant Process* [4, 16] or some other nonparametric Bayesian method [12] could be used to group the nodes into blocks; the graphs for many years could be combined together into a 3D tensor (formed as: (year, country, country) $\rightarrow$ distance) and methods for tensor decomposition could be applied to analyze it, etc.

From the results obtained by applying the *WSBM* [3] and the *Affinity Propagation* [6] method (described in subsection 5.2 and subsection 5.3, respectively), a negative answer to the question posed in subsection 1.1 can be given, i.e. the impact of climate change as a significant factor on the development of the countries in the world cannot be confirmed. As a main fact that justifies this conclusion, one can take into account that there are countries that have high values for their level of development and low values for their *Aridity Index* (e.g. *Australia*, *Israel*), as there are countries that have low value for their level of development, and high values for their *Aridity Index* (e.g. *Liberia*, *Sierra Leone*).

Given that the *HDR* dataset is composed of multiple datasets created by several major world organizations and its intention is strictly to measure the development of the countries in the world, it can be said that at this point it is the most relevant dataset that could describe the impact of climate change on the development of the countries in the world. This leaves an open question as to whether new measures should be introduced in the future that could better describe the impact of climate change and could thus direct the countries to make decisions that would avoid catastrophic consequences for the population living in those territories.

References

[1] Decline of global extreme poverty continues but has slowed: World bank. http://www.worldbank.org/en/news/press-release/2018/09/19/decline-of-global-extreme-poverty-continues-but-has-slowed-world-bank.

[2] *International meteorological vocabulary.* Secretariat of the World Meteorolog. Organization, 1992.

[3] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2014.

[4] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[5] Saeid Eslamian and Faezeh A Eslamian. *Handbook of Drought and Water Scarcity: Environmental Impacts and Analysis of Drought and Water Scarcity.* CRC Press, 2017.

[6] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[7] G David Garson. Interpreting neural-network connection weights. *AI expert*, 6(4):46–51, 1991.

[8] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947, 2000.

[9] Richard HR Hahnloser and H Sebastian Seung. Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in Neural Information Processing Systems*, pages 217–223, 2001.

[10] IPDJ Harris, Philip D Jones, Timothy J Osborn, and David H Lister. Updated high-resolution grids of monthly climatic observations–the cru ts3. 10 dataset. *International journal of climatology*, 34(3):623–642, 2014.

[11] Ralph E Huschke et al. Glossary of meteorology. 1959.

[12] Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian nonparametric data analysis.* Springer, 2015.

[13] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[14] Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2):135–150, 2002.

[15] Julian D Olden, Michael K Joy, and Russell G Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4):389–397, 2004.

[16] Jim Pitman et al. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course, 2002.

[17] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[18] Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

7 Appendix

7.1 A list of problematic attributes of the HDR data set

- *Education* (1)
  - *Proportion of schools with access to the Internet (%)* | 2008-2013
- *Environmental sustainability* (1)
  - *Forest area, change (%)* | 1990/2015
- *Gender* (4)
  - *Child marriage, women married by age 18 (% of women ages 20–24 years who are married or in union)* | 2003-2017
  - *Female share of graduates in science, mathematics, engineering, manufacturing and construction at tertiary level (%)* | 2007-2017
  - *Violence against women ever experienced, intimate partner (% of female population ages 15 and older)* | 2005-2018
  - *Violence against women ever experienced, nonintimate partner (% of female population ages 15 and older)* | 2005-2018
- *Health* (1)
  - *Child malnutrition, stunting (moderate or severe) (% under age 5)* | 2010-2016
- *Human Security* (3)
  - *Birth registration (% under age 5)* | 2006-2017
  - *Homeless people due to natural disaster (average annual per million people)* | 2007/2017
  - *Prison population (per 100,000 people)* | 2004-2015
- *Inequality* (3)
  - *Income inequality, Gini coefficient* | 2010-2017
  - *Income inequality, Palma ratio* | 2010-2017
  - *Income inequality, quintile ratio* | 2010-2017
- *Mobility and Communication* (1)
  - *Internet users, female (% of female population)* | 2013-2017
- *Socio-economic sustainability* (2)
  - *Income quintile ratio, average annual change (%)* | 2005/2017
  - *Overall loss in HDI value due to inequality, average annual change (%)* | 2010/2017
- *Work, employment and vulnerability* (2)
  - *Child labour (% ages 5-17)* | 2010-2016
  - *Old-age pension recipients (% of statutory pension age population)* | 2006-2016

Note: 2008-2013 indicates that there is one record in that period, and 1990/2015 indicates that the value of the record is the average value for that period.