LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Department of Economics

Data Analytics and Business Economics

DABN01: Master Thesis

May 2022

Exploring Greenhouse Gas Emissions and socio-economic factors for climate change mitigation: A worldwide clustering analysis

Author:

Anna Pasini

Supervisor:

Simon Reese

Abstract

In response to the pressing need to address climate change and reduce global greenhouse gas (GHG) emissions, this study implemented Gaussian Mixture Models Clustering to detect the levels of GHG emissions and related socio-economic factors in 174 countries. To handle the panel data, Principal Component Analysis was conducted to achieve dimension reduction. Based on the algorithm, countries were grouped into four clusters according to the development of similar features from 2001 to 2018. The algorithm performed well, yielding an average silhouette value of 0.54, indicating a clear assignment of data points to reference clusters with minimal uncertainty. With the exception of one cluster containing only seven countries, all countries were equally divided among the clusters, allowing for the capture of potential peculiarities. The clusters were arranged in order of GHG emissions per capita, from the highest to the lowest. The results reveal that the cluster with the highest GHG emissions per capita also exhibited the highest levels of GDP per capita, consumption of fossil fuels, imports and exports, internet usage, and urbanization. Conversely, high GHG emissions per capita coincided with low renewable electricity and energy outputs, as well as male and female unemployment rates. Identifying patterns and similar socio-economic structures among countries enables collaboration and the implementation of unified measures and regulations, making climate change mitigation more efficient and effective.

Keywords**:** GHG emissions, Climate Change, Panel Data, Gaussian Mixture Models (GMMs) Cluster Analysis, PCA

Acknowledgements:

I would like to express my deepest gratitude to my thesis supervisor, Simon Reese. His help and support have been instrumental in shaping this research and guiding me towards its completion. His contributions and teachings have significantly influenced my work and personal growth.

Thanks to Emilie, Emilie, and Tamar, my lovely companions on this academic journey and experience. They have made this process both enriching and enjoyable.

I dedicate this thesis to my family. Thank you for making this possible and thank you for your love, the most important thing I have in life.

Contents

1. Introduction

"Global warming is caused by the concentration of greenhouse gases (GHGs) emissions" (Williston, 2018).

Global warming has become the most important environmental problem in the world. All countries face tremendous pressure to reduce greenhouse gas (GHG) emissions (Feng et al., 2017). The term GHGs includes 6 different types of gases: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs) and sulfur hexafluoride (SF_6) (Kyoto Protocol, 1997). Among these, more than half of the total greenhouse emissions is primarily anthropogenic or human-induced. From 1997 to the present, annual global CO_2 emissions have increased, so analysis of the main drivers of global changes in GHG emissions has become extremely essential and urgent. Consequently, countries have been making sustained efforts to reduce their emissions.

Although, countries have a non-uniform distribution of energy resources, and different socio-economic histories and conditions. For this reason, economic development and carbon emissions are imbalanced throughout the world. Different territories and cultures consequently require different measures, interventions, and targets but also resources to become more sustainable.

Confronted with the growing requirements of cutting GHG emissions all around the world, this paper hereby wants to identify similarities and patterns, but also differences, between countries primarily through the use of an Unsupervised Machine Learning method namely Clustering Analysis. By using this technique, the subjects in question, i.e., the countries, will be divided into groups so that the countries that will be drawn from the same group are as similar to each other as possible, while those assigned to different groups are dissimilar. Thus, Clustering Analysis is the appropriate choice for categorizing countries with similar carbon emission characteristics and facilitating the exploration of other features and possible socio-economic influences of each category and their causes.

The intention is to include in the Cluster Analysis both variables more related and theoretical to the level of emissions, but also variables related to social and economic development, such as Unemployment Rate, Share of Urban and Rural Population, Percentage of women in Parliament, and Internet Access. This paper will explore the feasibility of comparing the level of sustainability in terms of the level of GHG emissions per capita of different countries considering a comprehensive set of indices related to economic, energy, environmental, and social conditions.

All these parameters are essential in evaluating countries' performance regarding Sustainable Development, as the one outlined by the United Nations Member States' 2030 Agenda for Sustainable Development. Specifically, Goals 10 and 13 of 'Goals and Targets from the 2030 Agenda" are: "Reduce inequality within and among countries" (Goal 10) and "Take urgent action to combat climate change and its impacts" (Goal 13) (SDG-Tracker.org, website, 2018).

To accomplish Goals 10 and 13, a data set dedicated to the study of Sustainable Development Goals, such as the 'World Sustainability Dataset', has been used in this analysis. All the variables involved and their relation to GHG emissions will be explained, in addition to the use of Machine Learning techniques, data preparation and the specific implementation of Cluster Analysis.

In the face of the growing and urgent need to take action against climate change worldwide, this paper aims to answer a series of research questions: which countries are experiencing an increase in greenhouse gas emissions? What is the relationship between countries' economic developments and greenhouse gas emissions? How do different sectors affect GHG emissions? Which countries have similar levels of emissions? And countries with similar emission levels also have similar causes of emissions? To answer these questions Principal Component Analysis (PCA) and Gaussian Mixture Models (GMMs) Cluster Analysis will be employed to investigate the contribution of different socio-economic and intrinsic reasons to GHG growth and identifying countries with similar characteristics and reasons for emissions.

Through the results, it will be possible to identify the emission patterns of countries, and this will provide information in order to make informed policy decisions, develop strategies to mitigate the drivers of global emissions and identify opportunities for collaboration among countries.

In summary, all the analyses in this paper will provide a deeper understanding of the complexities of sustainability and contribute to the current literature on GHG emissions and the climate change fight.

The remained of this thesis is structured as follows: In Sections 2, 3 and 4 the methodology employed in this thesis will be detailed. This will include a comprehensive literature review that supports the analysis conducted. The data sources and algorithms used in the clustering process will also be explained. Section 5 will present the findings of the GMMs Clustering, which resulted in the classification of the countries into different clusters based on their GHG emissions and socio-economic factors. The subsequent section, Section 6, will delve into a thorough discussion of the results obtained. It will examine the characteristics of each cluster. Section 7 will address the limitations encountered during the analysis and highlight potential areas

for future research. Lastly, Section 8 will draw conclusions from the analysis and present the key findings of the thesis. It will summarize the main insights gained from the clustering approach and provide a concise summary of the implications and potential policy recommendations derived from the study.

2. Literature Review

To understand the factors influencing a country's level of greenhouse gas (GHG) emissions, I reviewed existing studies that analyze the relationship between GHG emissions and socio-economic indicators, as well as the sectoral distribution of emissions. Addressing climate change globally requires countries to determine their GHG emissions levels (Feng et al., 2017). Connecting ecological change to economic prospects and social prosperity has perpetually been a difficult task. Comprehending and measuring sustainability encourages this route.

For the purpose of this paper, I focused on machine learning techniques such as linear regressions and looked to previous papers on Cluster Analysis at regional levels for guidance on applying clustering to the data in this thesis. However, it is important to note the limited research available globally, particularly regarding the inclusion of socio-economic variables in emissions studies.

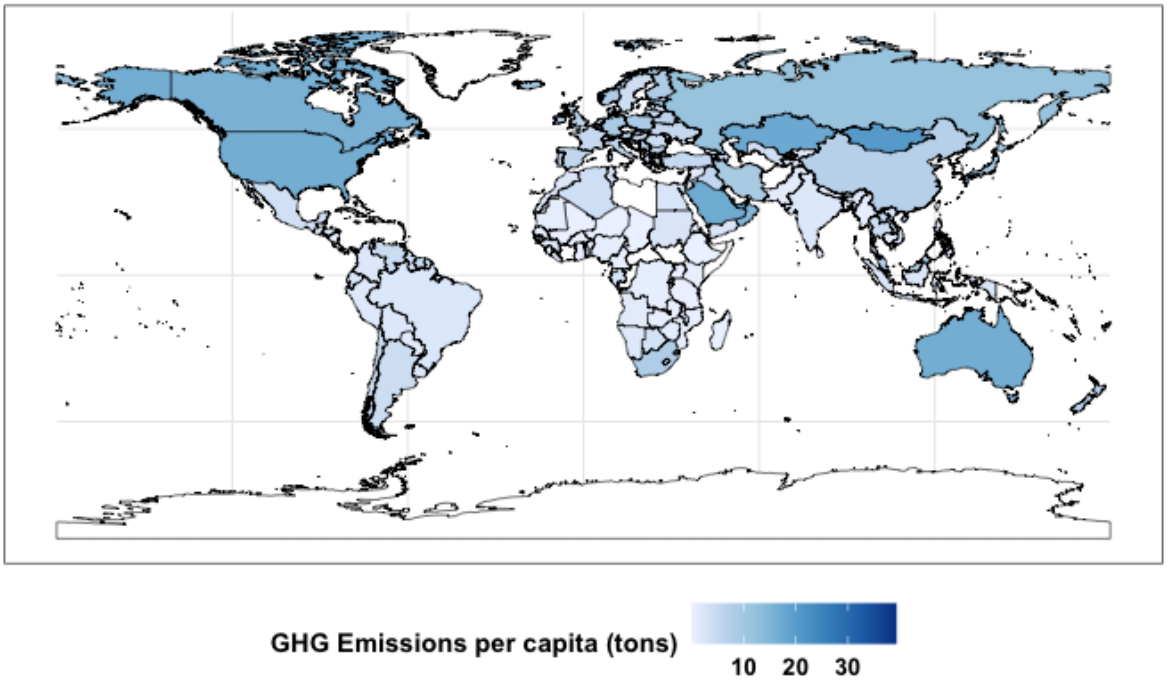

Figure 1: Worldwide GHG Emissions per capita (in tons) in 2018

Figure 1 shows the level of GHG emissions per capita in countries around the world in 2018.

An important aspect to consider is how emissions are distributed worldwide. Coulter (1989) demonstrated a highly uneven Gini coefficient of 80.9 for emission

distribution between countries, indicating significant inequality. A coefficient score of 100 indicates that one country is responsible for all emissions. This highlights the stark disparity between countries in terms of their respective contributions to climate change. It also raises the question of whether this disparity can be further explored and understood through cluster analysis. By using cluster analysis, it can be investigated whether countries with similar emission patterns and characteristics can be grouped together, thereby providing insights into the underlying factors contributing to the disparities in emissions.

The main countries responsible for these emissions are China (30%), the United States (15%), Europe (10%), and India (6.5%) (PBL Netherlands Environmental Assessment Agency, 2015). According to the PBL Netherlands Environmental Assessment Agency, certain patterns among countries were identified. In 2020, the United States (-8.5%), European Union (EU-27) (-8.4%), India (-3.9%), Russian Federation (-4.9%), and Japan (-6.3%) experienced decreasing GHG emissions from fossil fuel combustion and industrial non-combustion processes. On the contrary, China was the only country to witness a 1.5% increase in GHG emissions that year (PBL Netherlands Environmental Assessment Agency, 2022).

Clustering methods can identify patterns and group countries based on similar characteristics related to their emission levels. By analyzing various factors and attributes, clustering techniques provide insights into the heterogeneity among countries and their emissions profiles. This allows for a deeper understanding of the diverse characteristics that may contribute to variations in emission levels across different clusters. Previous literature has explored different factors related to emissions, with machine learning methods widely employed to predict future CO_2 emissions and analyze a country's sustainability performance (Mardani et al., 2020). The main sources of total CO_2 emissions are associated with human activities in different sectors including transportation, electricity production, industry, commercial and residential, agriculture, land use and forestry (Mardani et al., 2020).

In examining the environmental and energetic performance of European countries, Cucchiella et al. (2017) emphasize the importance of sustainability, which is closely linked to emissions reduction, renewable energy use, and energy efficiency. Hence, this thesis model incorporates indicators such as the share of renewable electricity output and the share of renewable energy consumption.

While much of the literature on environmental and energy topics often focuses solely on specific indicators, neglecting economic and societal aspects, it is important to recognize that sustainability encompasses not only the environmental dimension but also the economic and societal pillars (Çağlar & Gürler, 2021). Ignoring economic and societal factors can lead to an incomplete understanding of the relationship

between environmental indicators and overall sustainability. For instance, the unemployment rate of a country can significantly impact its level of greenhouse gas (GHG) emissions (Lou, Lin & Li, 2022). A high unemployment rate tends to dampen economic activity, resulting in reduced energy consumption and production, ultimately leading to lower GHG emissions. Furthermore, individuals facing unemployment may adopt more energy-efficient lifestyle choices, such as utilizing public transportation instead of private vehicles or practicing energy conservation in heating and cooling. Therefore, the objective of this thesis is to expand on existing research by undertaking a comprehensive analysis that incorporates various economic relationships.

In particular, the choice of this topic for the thesis was prompted by Mardani et al.'s (2020) research, in which CO_2 emissions were predicted using GDP and energy consumption in the Group of Twenty (G20) nations. Therefore, they suggested broadening the research to include various variables like renewable energy consumption, electricity consumption, urbanization, financial development, fossil fuels consumption, import and export structures, and other related variables in the different sectors and industries. While their paper focused on G20 nations, this research employs the clustering method to analyze other countries as well.

According to the United Nations (2023), climate change is primarily caused by several factors:

One of the main contributors is power and heat generation, which involves burning fossil fuels to produce electricity and heat. While renewable sources provide a portion of global electricity, the majority still comes from coal, oil, or gas.

Industrial activities, such as manufacturing and mining, also contribute to emissions as they rely on fossil fuels for energy. Machines used in these processes often run on coal, oil, or gas.

Deforestation is another significant cause, as cutting down trees releases stored carbon. This is mainly driven by the conversion of land for agriculture and the creation of pastures.

Transportation, including cars, trucks, ships, and planes, heavily relies on fossil fuels, particularly gasoline and diesel, leading to substantial greenhouse gas emissions. The demand for transportation energy is projected to increase significantly in the future.

Buildings, both residential and commercial, consume a significant amount of electricity globally. Heating and cooling contribute to greenhouse gas emissions, and the energy demand for these purposes is rapidly rising.

Our consumption patterns and lifestyles also play a crucial role in climate change. The choices we make in electricity usage, travel, diet, waste generation, and consumption of goods impact greenhouse gas emissions. The wealthiest segment of the population has a greater responsibility, as the top 1% collectively produces more emissions than the bottom 50%.

These factors collectively contribute to climate change, highlighting the need for sustainable practices, renewable energy adoption, and responsible consumption to mitigate its effects.

Previous research has indicated that the increase in the level of greenhouse gas emissions is mainly driven by the growth of consumption per capita, economic productivity and population (Arto & Dietzenbacher, 2014) (Althor, Watson & Fuller, 2016). Levinson and Taylor (2008) conducted an analysis of the pollution haven effect, aiming to uncover its causes and implications in different countries. Import and export structure emerged as a crucial factor influencing emission levels. The authors found that countries heavily reliant on imports of goods and services tend to have higher emission levels. This can be attributed to the increased energy consumption associated with the production and transportation of imported goods. On the other hand, countries that are more self-reliant and export a larger quantity of goods and services generally exhibit lower emission levels. Additionally, the authors discovered that the impact of imports and exports on emissions varies across different countries. For instance, while an increase in imports led to higher emissions in countries like Canada, East Asia, EU-27, and Russia, countries such as China, the United States, and Australia experienced the opposite effect, where increased imports coincided with lower emissions. This discrepancy can be attributed to the influence of other economic factors. One such factor is the growing consumption relative to production in developed countries, leading to a widening current account deficit. These global trade imbalances have translated into imbalances in emissions as well.

Furthermore, Liu, Guo, and Xiao (2019) conducted research on global emissions and found that the primary driving factor is the exponential growth of the economy. However, they also highlighted the significant impact of technological innovation and energy efficiency in decreasing emissions. The factors contributing to greenhouse gas emissions vary between countries, with developing countries experiencing a larger impact from investment effects and developed countries showing more prominent export effects. Previous Cluster Analysis studies have also revealed a relationship between the growth of GHG emissions and gross domestic output in various countries, with developing countries experiencing positive effects on emissions changes due to the importance of agriculture (Li et al., 2019). Implemented Cluster Analysis attributed the growth of GHG emissions also to fossil fuel type (coal, oil, or gas), type of demand (consumption or investment), and country group (developed or developing) (Jiang &

Guan, 2016).

Additionally, the use of renewable energy sources and energy efficiency plays a crucial role in achieving better energy goals and contributing to the reduction of greenhouse gas emissions (International Renewable Energy Agency, 2019). Energy efficiency is essential for enhancing socio-economic and environmental infrastructure, which is desirable for advancing sustainable development reforms in all countries. It can help reduce energy imports and fossil fuel use. The transition to a sustainable energy system often involves transforming local communities, where cultivating a shared vision is a strength.

To sum up, several studies in recent years have focused on the associations among energy consumption, emissions, and Gross Domestic Product (GDP). However, there is a lack of descriptive indicators characterizing emissions in the existing literature, and more indicators reflecting emission structure should be considered (Yu et al., 2011). Carbon dioxide (CO2), in fact, is linked not only to climate change, but also to political, social, and economic matters. This means that trying to reduce the effects of climate change is highly dependent on socioeconomic and politico-cultural elements and clustering approaches can be employed to investigate and evaluate the connections between various environmental, economic, and energy indicators of the cluster's member countries (Çağlar & Gürler, 2021).

Cluster Analysis applied to panel data has been extensively discussed in the literature, with various methodologies proposed. In this thesis, we adopt a methodology inspired by the work of Wang and Lu (2021). They introduced Principal Component Analysis (PCA) as an effective approach to reducing the dimensionality of panel data, followed by the application of Cluster Analysis. This approach enables us to capture meaningful patterns in the data and facilitates the clustering process.

3. Data Description

The data in this thesis is a Multivariable panel dataset covering 22 features that are observed on a yearly basis for 174 countries from 2001 to 2018. Most of the data was taken from a World Sustainability dataset available on Kaggle (Kaggle, 2022), which tracks the sustainability of 173 countries over 19 years. In addition to that, in order to include variables highly correlated to the level of emissions, sectoral GHG emissions have been obtained from the sectoral GHG emissions dataset (Climate Watch, 2022) including emissions of the six major GHGs from most major sources and sinks. Non-CO_2 emissions are expressed in CO_2 equivalents using 100-year global warming potential values from the IPCC Fourth Assessment Report. Lastly, from Energy Source per Capita dataset (Our World in Data, 2021) yearly fossil fuel consumption per capita from 2001 to 2018 for Countries was pulled.

Afghanistan, Andorra, Brunei, Djibouti, Grenada, Kiribati, Libya, Marshall Islands, Micronesia, Nauru, Palau, Papua New Guinea, Saint Kitts and Nevis, Samoa, Sao Tome and Principe, Somalia, South Sudan, Taiwan, Tuvalu, Holy See, State of Palestine, and Turkmenistan were not included in the analysis due to a lack of data about emissions per capita.

Tables 1 and 2 show the variables included in the model by emphasizing the general topic to which they correspond.

Table 1: Variables highly related to the emissions

Variable	*Units of measurement*	*Description*
Total greenhouse gas emissions per capita (including land use, land use change and forestry)	Tons (methane, nitrous oxide, and trace gases in CO_2-equivalents)	Total GHG emissions÷Total population
Fossil fuels consumption per capita	Kilowatt-hour, Wh	Total Fossil fuels÷Total population
Building Sector GHG emissions	Megaton, Mt	Expressed in CO_2-equivalents
Transportation Sector GHG emissions	Megaton, Mt	Expressed in CO_2-equivalents
Agriculture Sector GHG emissions	Megaton, Mt	Expressed in CO_2-equivalents
Industrial Processes Sector GHG emissions	Megaton, Mt	Expressed in CO_2-equivalents
Land use, land-use change, and forestry (LULUCF) GHG emissions	Megaton, Mt	Human-induced land use (settlements, commercial activities, land-use transformation, and forestry activities)

Table 2: Socio-economic variables

Variable	*Units of measurement*	*Description*
GDP Per Capita	Current Usd$	Total GDP÷ Total Population
Access To Electricity	% Of The Population	Population with access to electricity
Imports	% Of GDP	Imports Of Goods and Service
Exports	% Of GDP	Exports Of Goods and Service
Inflation	%	Annual inflation
Individuals Using The Internet	% Of The Population	Population with access to the Internet
Final Consumption Expenditure	% Of GDP	Private Consumption Expenditure + Government Consumption Expenditure
Gross National Expenditure	% Of GDP	Private Consumption Expenditure + Government Consumption Expenditure + Gross Domestic Investment
Renewable Electricity Output	% Of Total Electricity Output	Renewable Electricity ÷ Total Electricity
Renewable Energy Consumption	% Of Total Final Energy Consumption	Renewable Energy ÷ Total Energy Consumption
Seats Held by Women In National Parliaments	% Of Total Seats	Women in parliament ÷ Total Parliament Seats
Men Unemployment Rate	% Of Men Population	Unemployed Men ÷ Total Men Population
Women Unemployment Rate	% Of Women Population	Unemployed Women ÷ Total Women Population
Population Living In Urban Areas	% Of The Total Population	Population living in Urban areas ÷ Total Population
Population Living In Rural Areas	% Of The Total Population	Population living in Rural areas ÷ Total Population

However, the dataset contains several missing values. Precisely, the percentages of missing values for each variable in the dataset are illustrated in Appendix C in Figure C. The following section will outline the methodology used to address these missing values in this thesis.

4. Methodology of Algorithms

4.1 Missing Values Imputation

Building upon the information provided in the data description, it is important to address the issue of missing values in the dataset used for this thesis. This section will discuss the methodology utilized to solve the problem.

Missing values can be difficult to handle. Especially, the Cluster Analysis method that has been used for this research cannot operate with these values. Removing the rows that contain missing observations and performing the analysis on the complete rows is wasteful and depending on the fraction missing present in the used dataset, unrealistic. In this research, the imputation of missing values has been made by using a random forest method trained on the observed values of the dataset to predict the missing values (Stekhoven & Buhlmann, 2011). To be more specific, a random forest is trained for each individual feature, using the remaining features as inputs. Missing values in the feature for which the random forest was trained are imputed with model predictions. In total, the imputation of missing values variable by variable is repeated 10 times in order to arrive at updated, improved imputed values. Thus, it should be possible to use similar countries' socio-economic distinctions for the country for which data have not been collected or are not available to predict which characteristics that country will have.

4.2 Principal Component Analysis

The cluster analysis method of Multivariable panel data, such as the one used in this analysis, is more complex because of its complexity of data form (Wang & Lu, 2021). In Multivariable panel data, when performing sample clustering, the relevant definitions of the distance between samples and distance between classes cannot be directly applied if the number of variables is large, as in this dataset. Therefore, PCA (Principal Component Analysis) was used to reduce the dimensionality of a dataset. It finds a low-dimensional representation of a data set that contains as much as possible of the variation. Since not all the observations and features are equally noteworthy, PCA looks for a limited number of measurements that are as significant as could be expected. Every one of the measurements found by PCA is a standardized linear combination of the 22 (p) features (James et al., 2013). Since the first principal component accounts for most of the variation, only that one has been chosen so that the projected observations are as close as possible to the original observations.

Since variance is the only interesting element, each of the variables in the data matrix has been centered to have a mean zero and data has been standardized. In this analysis, in particular, the linear combination of the sample feature values is of the following form.

$$Z_{i1} = \varphi_{11}X_{i1} + \varphi_{21}X_{i2} + \cdots + \varphi_{p1}X_{ip} \qquad (4.2.1)$$

which has the largest sample variance.

The values of Z_{11} are known as the principal component scores.

The specific thinking and calculation steps of this paper's PCA are as follows:

(1) Standardize the data matrix, i.e.,

$$Z_{ij}(t) = \frac{X_{ij}(t) - \overline{X}_j}{\mathrm{var}(X_j)}$$

$$\text{Among them, } \overline{X}_j = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{ij}(t), \quad \mathrm{var}(X_j) = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{ij}(t) - \overline{X}_j)^2 \qquad (4.2.2)$$

(2) Perform a principal components analysis on the data matrix and extract just the scores of the 1° component, i.e., the first principal component direction of the data, that along which the observations vary the most;

(3) Transform the number of samples from N (number of countries x 18 Years, i.e., 3132) to NT i.e., the Number of Years (18).

All this results in a dataset in which for each country there is only one value, instead of 22, for every year, thus meaning the development of the ith-country considering all variables.

4.3 Gaussian Mixture Models Clustering

Lastly, Cluster Analysis was applied to the transformed and univariate dataset. Through the use of Clustering the aim is to partition the countries into distinct groups so that the countries within each group are quite similar to each other, while countries in different groups are quite different from each other. Firstly, the concept of 'observations to be similar or different' must be defined (James et al., 2013). In the context of Gaussian Mixture Models (GMMs) clustering, similarity refers to the measure of how close or alike two data points are in terms of their probability distribution. More specifically, the similarity between two data points is determined by

comparing their probabilities of belonging to each component of the GMM. In practical terms, the similarity between two data points in GMMs clustering can be expressed as the posterior probability of the data points belonging to a particular component of the GMM (Bishop, 2016). These probabilities are calculated using the Expectation-Maximization (EM) algorithm, which estimates the parameters of the GMM and assigns membership probabilities to the data points.

In summary, the steps of EM for Gaussian Mixtures are the following:

(1) Choose initial values for the means μ_k, covariances Σ_k and mixing coefficients π_k, and evaluate the initial value of the log-likelihood.

(2) **E step.** Use the current values for the parameters to evaluate the posterior probabilities.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{4.3.1}$$

(3) **M step.** Re-estimate the means, covariances, and mixing coefficients using the probabilities in the E step.

$$\begin{aligned} \boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\ \boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\ \pi_k^{\text{new}} &= \frac{N_k}{N} \end{aligned} \tag{4.3.2}$$

(3) Evaluate the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{4.3.3}$$

and check for convergence, i.e., when the change in either the log-likelihood function or in the parameters falls below a threshold equal to $1e^{-10}$. If the convergence criterion is not satisfied return to step 2 (Bishop, 2016).

5. Results

In order to reach (approximate) convergence with the EM algorithm, which requires numerous iterations and is computationally expensive, the K-means algorithm is run first to find an appropriate initialization. Here the GMMs clustering is used by setting the number of clusters as the optimal number of clusters suggested by the K-means algorithm, as shown in Figure 2.

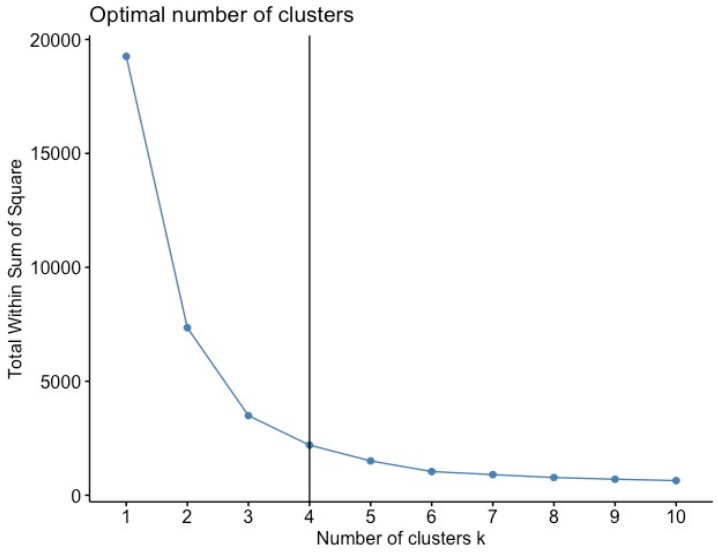

Figure 2: Optimal number of clusters

The elbow method is the technique used to determine the optimal number of clusters. By plotting the total sum of squared distances against the number of clusters it is possible to identify the elbow point where the rate of decrease in variance significantly diminishes, to which the optimal number of clusters corresponds (Bishop, 2016). After running an initial and general Gaussian model on the data and setting initial parameters, the best model was chosen according to the BIC criterion with 4 clusters.

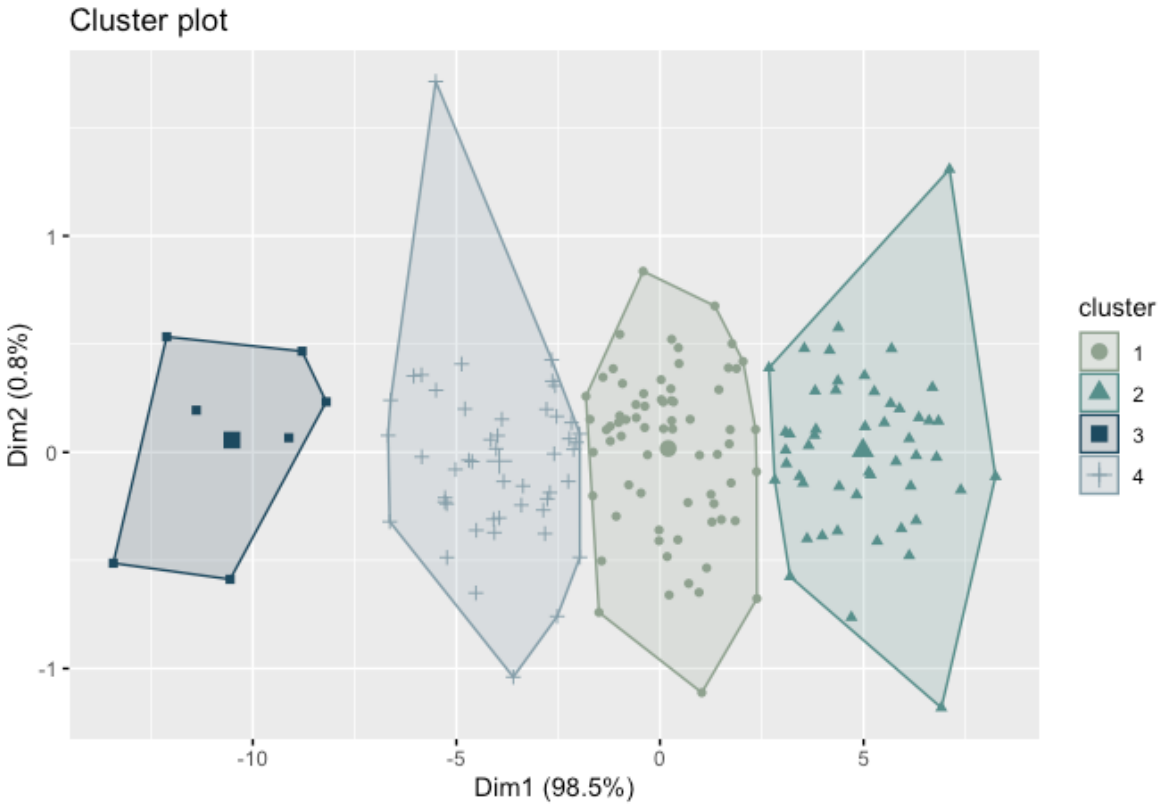

Figure 3: Cluster plot

The data that has been clustered has 18 dimensions, i.e., the development of all features from 2001 to 2018. In order to visualize the Clusters, the fviz_cluster reduces the data into two dimensions. The graph produced by fviz_cluster uses a Principal Components Analysis and projects the data onto the first two principal components. Those are the two dimensions that show the most variation in the data. The x-axis is the first principal component. Dimension 1 means that the first principal component accounts for 98.5% of the variation. The second principal component accounts for 0.8% of the variation, which is shown in the y-axis. So together they account for 99.3% of the total variation. In Figure 3, the Cluster means are plotted, which are represented by one larger symbol than the others. Respectively, for all Clusters. The shapes and sizes are fitted to the points in the four Clusters.

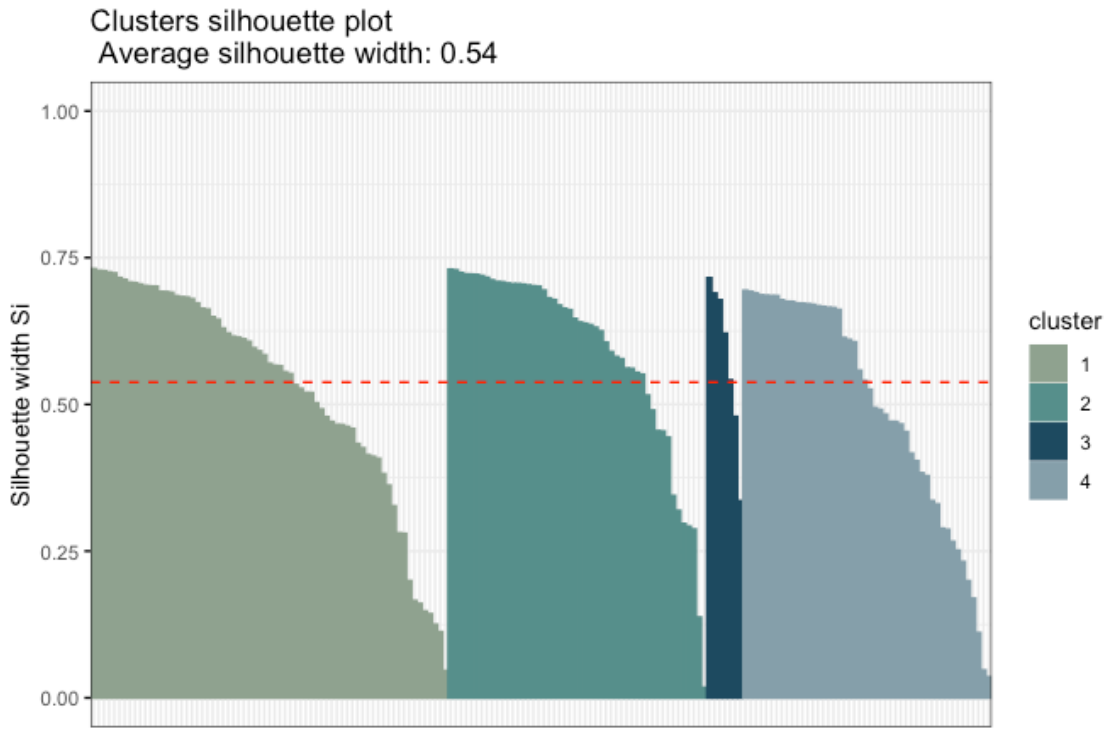

Figure 4: Silhouette plot

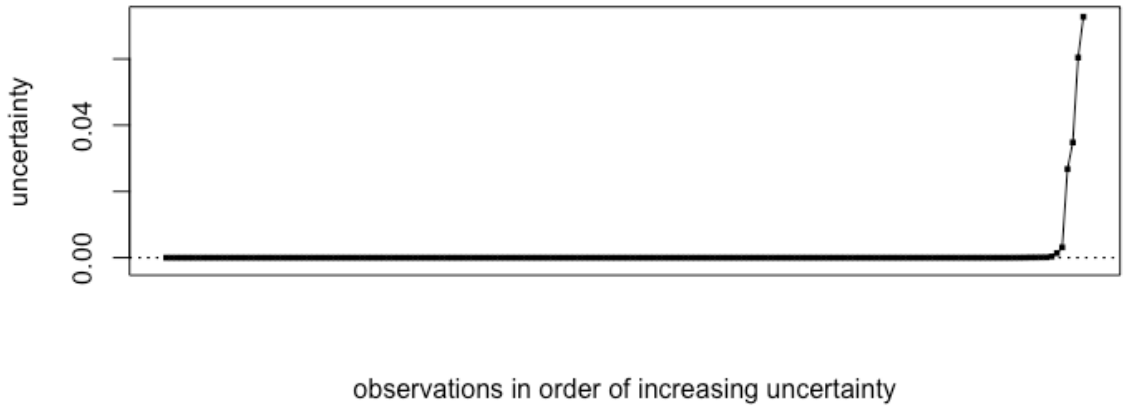

Figure 5: The uncertainty associated with the classification

In order to evaluate the clustering results, the silhouette plot (Figure 4) and the uncertainty of the assignment for each country to the respective Cluster (Figure 5) are shown. Figures 4 and 5 display a measure of how well each data point is assigned to its Cluster. The silhouette value ranges from -1 to 1. The closer the value is to one, the more the country is assigned to the correct cluster.

The values shown by the silhouette plot are calculated as follows:

$$Width = (distance_i - distance_j) / max(distance_i, distance_j) \quad (5.1)$$

where: "$distance_i$" is the average distance between a sample and all other points in the same cluster, "$distance_j$" is the average distance between a sample and all points in the nearest neighbouring cluster.

The average silhouette value, among the clusters, is equal to 0.54, which indicates that the assignment of the countries to clusters is clear. The same can be stated by looking at the Cluster plot, where the 4 groups are clearly separated. The silhouette also shows the distribution of the different clusters. While Clusters number 1,2 and 4 are evenly distributed, it shows that Cluster 3 is too narrow, which might indicate a capture of a group of outliers. Precisely, Cluster 1 is composed of 69 countries, Cluster 2 of 50 countries, Cluster 3 of 7 countries and Cluster 4 of 48 countries. Likewise, the values shown in Figure 5 represent the probability that a country is incorrectly assigned to its Cluster. The uncertainty in the GMMs clustering model is defined by using the Bayesian Information Criterion (BIC), which balances the trade-off between model fit and model complexity, with lower BIC values indicating better model performance. The clustering solution with the lowest BIC value was considered the most appropriate. The values then can assume values from 0 to 1. The subdivision then is certain, as the maximum value of uncertainty is 7%.

Thus, it can be concluded that the clustering separated countries with different development for almost all characteristics from 2001 to 2018 quite well into groups, while countries with comparable development were grouped together.

Table 3: The clustering results

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

6. Discussion

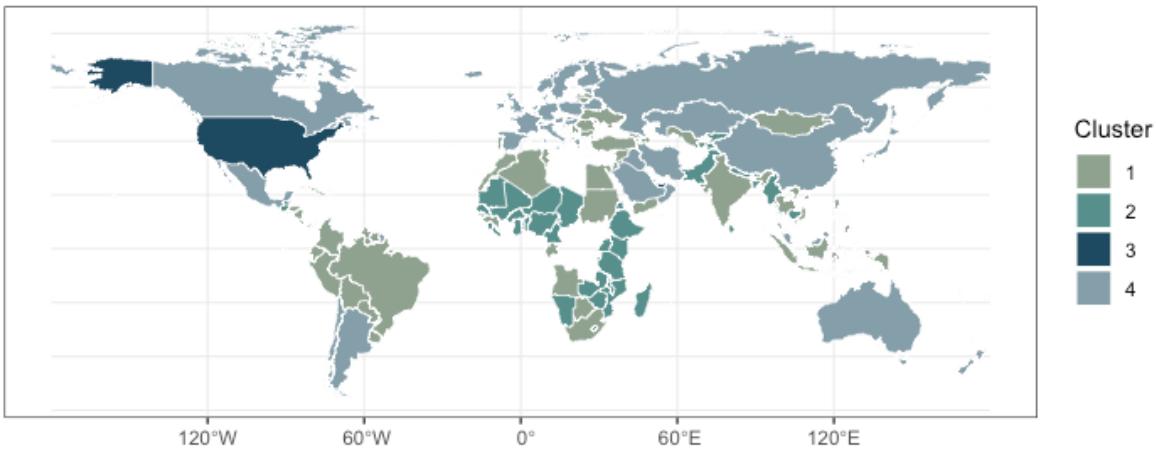

Figure 6: Clusters map

The geographical division of countries into Clusters is shown in Figure 6.

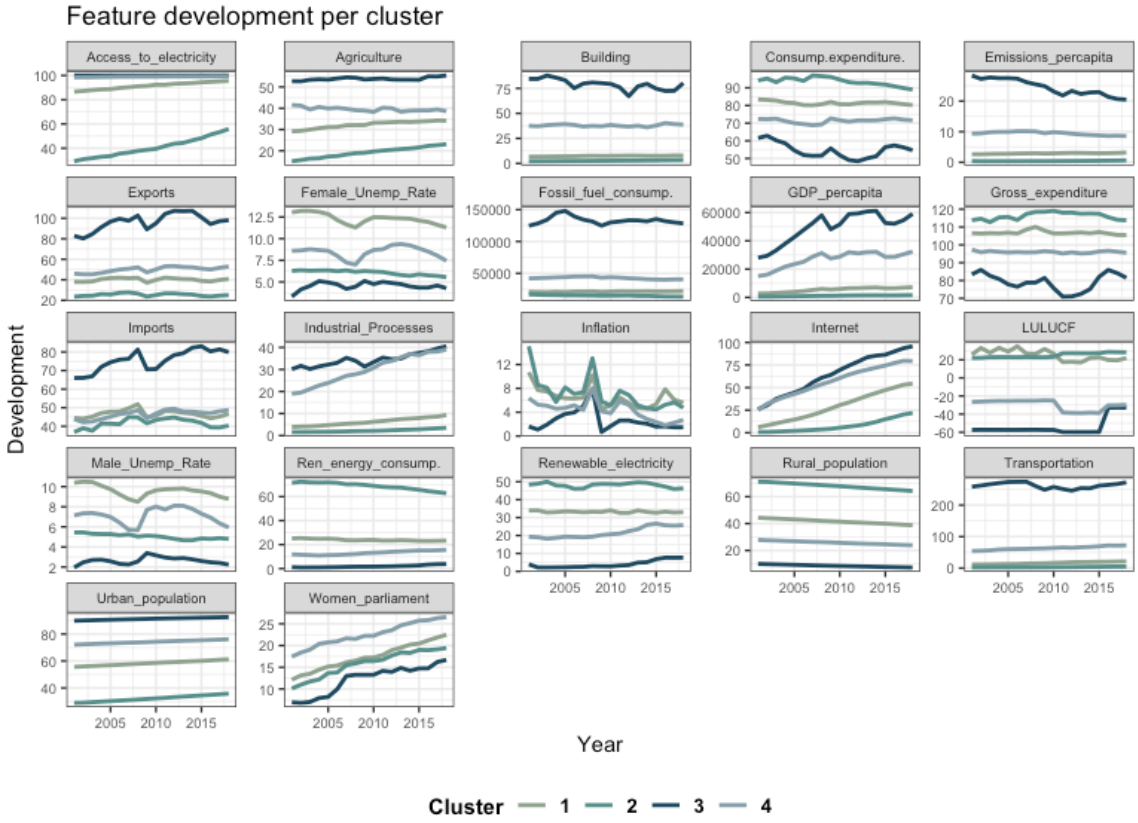

Figure 7: Features development per cluster

During the period between 2001 and 2018, the characteristics of worldwide countries' carbon emissions can be divided into four categories using the Gaussian Mixture models clustering. The characteristics and their effects can be analyzed as follows.

Bahrain, United States, United States Arabia, Luxembourg, Qatar, Singapore, and Kuwait, i.e., Cluster 3, registered the highest level of GHG gas emissions. In line with what the literature states and in light of the results of Liu, Guo, and Xiao (2019), Arto & Dietzenbacher (2014) and Althor, Watson & Fuller (2016)'s papers, a high level of emissions coincides with a high GDP per capita and with high consumption of fossil fuels. But it also coincides with high values for other macro indicators, such as Imports and Exports, Internet use and Urbanization. On the other hand, the Cluster with the lowest level of emissions is number 2, followed by number 1, with an average difference of 20 tons of emissions per capita with respect to Cluster 3's countries. The per capita emission of Qatar in Cluster 3, precisely 39.27 tons per capita in 2018, is more than 700 times that of The Democratic Republic of the Congo and Burundi (in Cluster 2), which are respectively only 0.03 and 0.05 tons/person (in 2018). As a consequence, Cluster 2 has also the lowest GDP per capita, fossil fuel consumption per capita, Exports, Imports, the share of individuals using the Internet, and Urban population.

The characteristics of the four Clusters are quite different, except for Cluster 1 and 2, which show some similarities in terms of emissions. As in the 2 Group, countries are mainly part of Africa and relatively close in space to countries of Group 1, the countries have a similar development of features from 2001 to 2018, except for Female and Male Unemployment rates and Access to electricity. In addition to the same development, they also share the same levels for, in particular, Emissions per Capita, Inflation, Share of Women in parliament, Building and Industrial Sectors and LULUCF emissions.

From 2001 to 2018 there was growth in Internet use for all Clusters. The same was true for renewable energy use, except for Cluster 2. In addition, Cluster 3 presents the highest level of emissions for all the sectors, except for the LULUCF one, for which it has, on the contrary, the lowest level.

Looking at Figure 7 it is evident that GHG emissions per capita and consumption expenditure are highly negatively correlated, precisely the correlation is equal to -0.46. The order of Clusters in terms of GHG emissions level is the opposite of the consumption expenditure one.

(1)

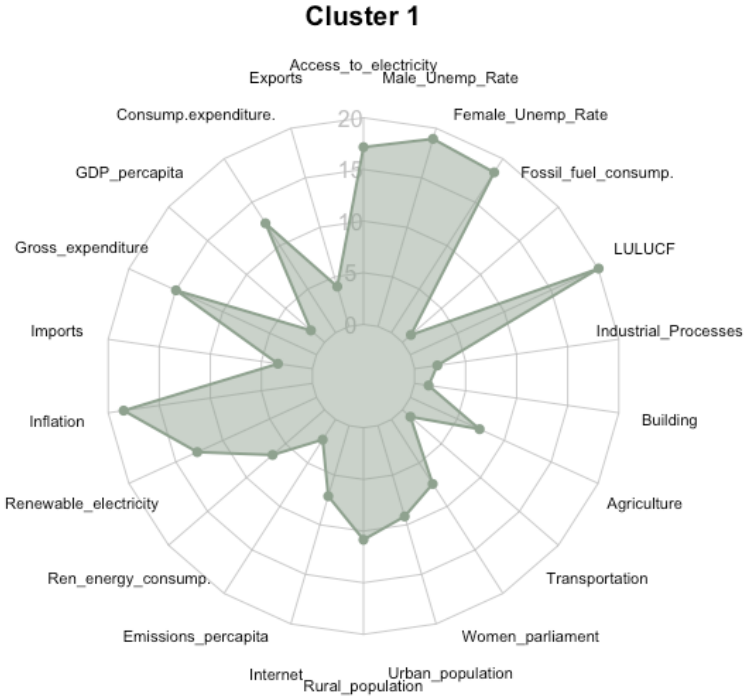

Figure 8: Cluster 1 characteristics: Feature means

The first Cluster is represented by a very extended area. It includes countries in South America, and Africa, as well as Europe, such as Romania and Ukraine, and Asia, such as India, Indonesia, and Thailand. In fact, this Cluster is the largest and includes the highest number of countries. However, countries in Group 2 can be defined as underdeveloped, poor, and low-GHG emissions per capita countries.

The emissions of countries in Cluster 1 are very similar to those in the group with the lowest emissions, i.e., Cluster 2. The difference on average between Cluster 2 and 1 is 2.5 tons of GHG emissions per capita. Among the countries in this group, The Republic of Congo has the lowest level of emissions, while South Africa has the highest, 0.46 and 8.7 tons per capita respectively.

The main reasons for the appearance of this characteristic can be summarized as follows: Firstly, there is the impact of a low average GDP per capita compared to Cluster 3 and 4 GDP per capita. This is equal to 5388.819 USD $.

Secondly, Land Use, Land Use Change and Forest GHG emissions are the highest. The countries of the first group are the ones characterized by more rural areas instead of urban areas, which results in a wider presence of land and forest.

Thirdly, there is the impact of transportation, industrial and building sectors' structure. In all three sectors, GHG emissions are lower than Cluster 3 and 4 sectors' emissions.

These Countries being underdeveloped can be identified in the highest rates of female and male unemployment, precisely 13% and 9.6%.

Finally, there is the impact of renewable electricity and energy structure. There is an important heterogeneity within the Cluster in terms of renewable sources. In particular, Antigua and Barbuda have the lowest proportion of both renewable electricity and energy structure, with a share of renewable electricity equal to 25% and of energy equal to 17%, while Albania, which has the highest proportion, has the two shares equal to 97% and 37%. From 2001 to 2018, there has been no improvement, that is, no growth in the use of renewable energy and electricity. This is also reflected in the emissions trend, as shown in Figure 7. In this cluster, in 2001, emissions amounted to 2.6 tons per capita, and renewable energy and electricity were 33.9% and 25% respectively. In 2018, in fact, the percentage of renewable energy dropped slightly, by 0.8 percentage points. Renewable electricity also dropped to 23% (-2%). The result on emissions is an increase to 3.2 tons per capita. The Republic of Congo and Vietnam instead registered a major decrease in both renewable electricity and energy structure (-39% and -32%) and so an increase in GHG emissions per capita (+0.4 and +1.48 tons per capita).

In conclusion, Cluster 1 countries face economic challenges, high reliance on land and forests, and a need for sustainable development strategies to address GHG emissions and promote economic growth. To begin, implementing policies and practices that promote sustainable land use and forest management is crucial. This involves preventing deforestation, promoting reforestation and afforestation, and supporting sustainable agricultural practices to reduce greenhouse gas emissions. Encouraging community-based forest management should also be prioritized. In addition, fostering economic diversification is essential to reduce dependence on land-based activities and create alternative employment opportunities. Developing sustainable industries, such as renewable energy and sustainable agriculture, can lead to economic growth and job creation. Supporting entrepreneurship, innovation, and access to finance are key components of this strategy. Promoting the adoption and expansion of renewable energy sources is another important step. This can be achieved through financial incentives, favourable regulatory frameworks, and capacity-building programs. Small-scale renewable energy projects, particularly in rural areas, should be encouraged to improve energy access and innovation. Facilitating technology transfer and knowledge exchange is critical. Cluster 1 countries

can benefit from the expertise and experiences of developed countries and international organizations.

(2)

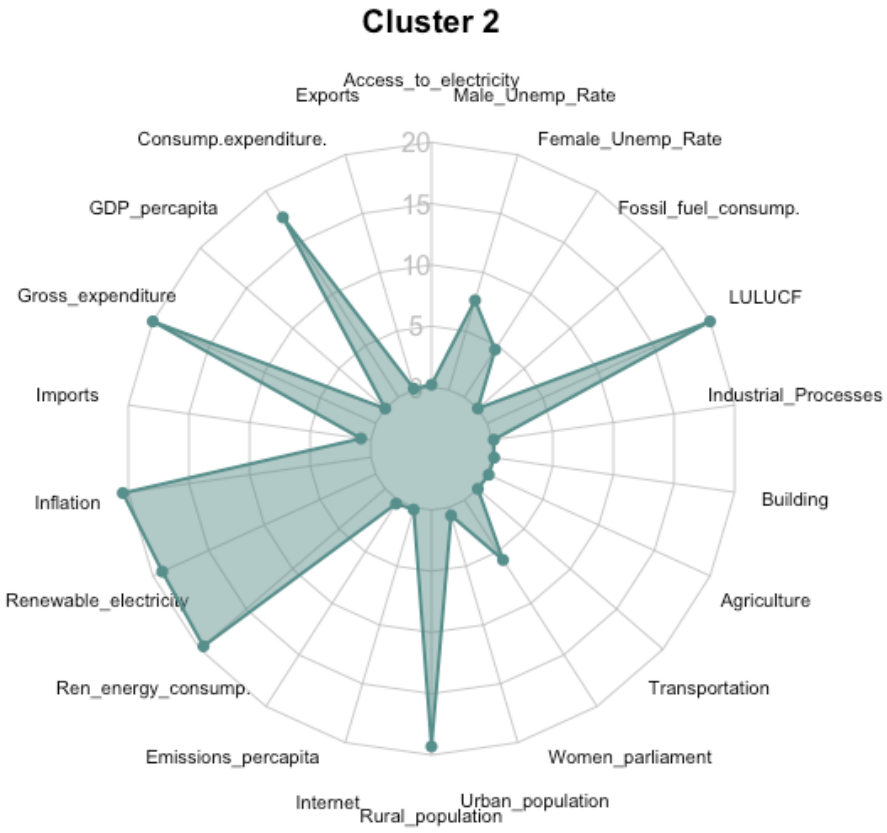

Figure 9: Cluster 2 characteristics: Feature means

Cluster 2 comprises the less polluting, poorest and most rural countries. Countries in Cluster 2 are the ones with the lowest GHG emissions per capita. Looking at the data, it can be clearly seen that the emission intensity per capita of GHG in these countries (around 0.42 tons) is much lower than in other countries. In addition to having the lowest level of GHG emissions per capita, Cluster 2 also has the lowest GDP per capita, compared to the other Clusters' countries, with exactly an average GDP per capita equal to 1120 (USD $). In 2018 Burundi, The Democratic Republic of the Congo and the Central African Republic were the countries which registered the lowest GDP per capita, i.e., around 271.75 $, and therefore the lowest GHG emissions per capita (0.03 tons). Consequently, we can state that Emissions per capita and GDP per capita are highly correlated in a positive way, precisely their correlation is 0.61. The type of GHG emission of Cluster 2 has the characteristics of very low Building, Transportation, and Industrial Processes emissions while high LULUCF activities, as illustrated in Figure 7.

The main causes of these characteristics are as follows. First, these countries are characterized by a very high share of the population living in rural areas, around an average of 68%, and only 32% of the population lives in urban areas. Consequently, less than half of the population has Access to Electricity, exactly 40%, and individuals who use the internet are just over 7% of the population, on average.

Secondly, fossil fuels consumption per capita is equal to 15072.26, which is the lowest among the four Clusters. In fact, as underlined by The United Nations (2023) burning fossil fuels to generate electricity and heat represents a major source of global emissions release.

Thirdly, the low level of GHG emissions per capita is also given by the lowest levels of Exports and Imports, 25.4% and 41% of GDP. The difference in Exports with the Cluster which has the highest level of GHG emissions is -71% for Cluster 2. According to the pollution haven hypothesis (Levinson & Taylor, 2018), countries with weak environmental regulations, such as Cluster 2's countries, attract the most polluting countries by moving their industries with stricter regulations to underdeveloped countries. In fact, stricter environmental regulations cause the cost of pollution-intensive inputs to rise and the comparative advantage of these countries for these types of goods to be reduced. This hypothesis shows that countries will export goods to less stringent jurisdictions. Therefore, Exports and Imports in Group 2 are significantly lower than in the Cluster with higher GHG emissions, and thus with higher costs to reduce their environmental impact.

Lastly, although the trend between 2001 and 2018 is very similar for Inflation for all the Clusters. This feature of Cluster 2, which is equal to 7.1 %, is the highest compared to the other Clusters, as can be observed in Figure 7. Higher inflation instability reduces environmental pollution (Xu et al., 2023). The data also shows that the correlation between Inflation and greenhouse gas emissions is negative, likewise is the correlation between Inflation and emissions from the transportation, building and industry sectors.

Given the Cluster 2 characteristics, the following policy recommendations can be made: Prioritize poverty alleviation strategies and social development initiatives to improve the well-being of the population. This can be achieved through financial support, capacity building, and technology transfer initiatives. Secondly, encourage and facilitate technology transfer from developed countries to Cluster 2 countries, focusing on sustainable development solutions. Also, promote the expansion of affordable and reliable access to information and communication infrastructures, such as modern communication systems, internet connectivity, computer networks, and various digital and mobile services, including rural areas population. All this can help

bridge the digital divide, enhance education and healthcare services, support agricultural productivity, and enable e-commerce and entrepreneurship opportunities.

(3)

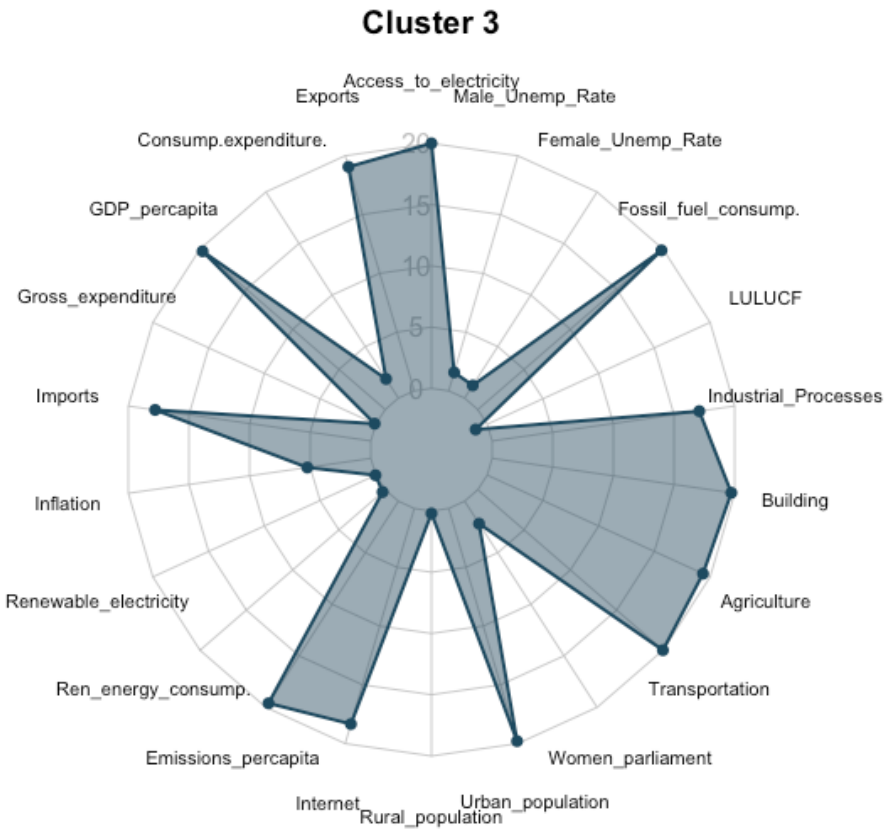

Figure 10: Cluster 3 characteristics: Feature means

Cluster 3 consists of the richest and most polluting countries. In particular, the GHG emissions of the third Cluster countries amount to more than 24.5 tons per capita on average. As shown in Figures 1 and 7, the GHG emission intensity in these countries is higher than in others. They also registered the highest GHG emissions in all sectors, except for the LULUCF one. In the analysis of the seven countries belonging to the third Cluster, it can be seen from Figure 7 that they also have the highest fossil fuel consumption per capita, which is equal to 133047.02 kWh and the highest GDP per capita, i.e., 49324.8$ on average, which more than 43 times higher than Cluster 2 countries' GDP per capita.

The countries in this group are mainly coastal. For example, in the United States, 30 out of 50 States have a coastline. Qatar has more than 560 km of coastline. This factor

has particular relevance to the levels of Exports and Imports. The presence of such characteristics in the seven countries has the main effect on Imports and Exports. They are the principal driving factors for such a high number of GHG emissions per capita.

In addition, Cluster 3 can also be defined as the most socially developed. Both unemployment rates, female and male, are lower than all other Clusters, 4.5% and 2.6%. Furthermore, countries in the third group also have 91% of the population living in urban areas, 64% of individuals have access to the Internet and the entire population has access to electricity (100% on average). One peculiarity, however, as shown in Figure 7, is that the Internet share has grown considerably from 2001 to 2018, precisely in 2001 it amounted to 26% while in 2018 to 95%. Among countries in Cluster 3, in 2018 Kuwait had the highest share of individuals using the Internet, reaching a 99.6%, while Luxembourg has the lowest, i.e., 97%.

On the contrary, the social development achieved through the use of the network is not reflected in gender inclusion. It is the Cluster with the lowest proportion of seats held by women in their national parliaments. The proportion is equal to 12%, while in Clusters 1,2 and 4 the proportions are respectively 17%, 15% and 22%.

Lastly, the characteristic of being the most polluting seven countries is mainly due to the lowest consumption of renewables in both the energy and electricity sectors, leading to high GHG emissions per capita, as Figure 10 shows. Note, however, that there has been an improvement in the use of renewable energy and electricity from 2001 to 2018. Even so, the increase was not enough, still in 2018, to move Cluster 3 up in position.

Consequently, Cluster 3 should focus on transitioning from fossil fuel-based energy systems to cleaner and renewable energy sources. Their targets focus on implementing and enforcing strict energy efficiency standards across sectors, including transportation, industry, and buildings. But also, promoting the use of energy-efficient technologies, encouraging energy conservation practices, and providing incentives for businesses and households to adopt energy-saving measures. This will help reduce energy consumption and, consequently, GHG emissions. Additionally, implementing policies that encourage the adoption of circular economy principles, including waste reduction, recycling, and responsible resource management can help accelerate the process toward sustainability.

This can be achieved through the implementation of policies that promote renewable energy investment, such as feed-in tariffs, tax incentives, and subsidies.

(4)

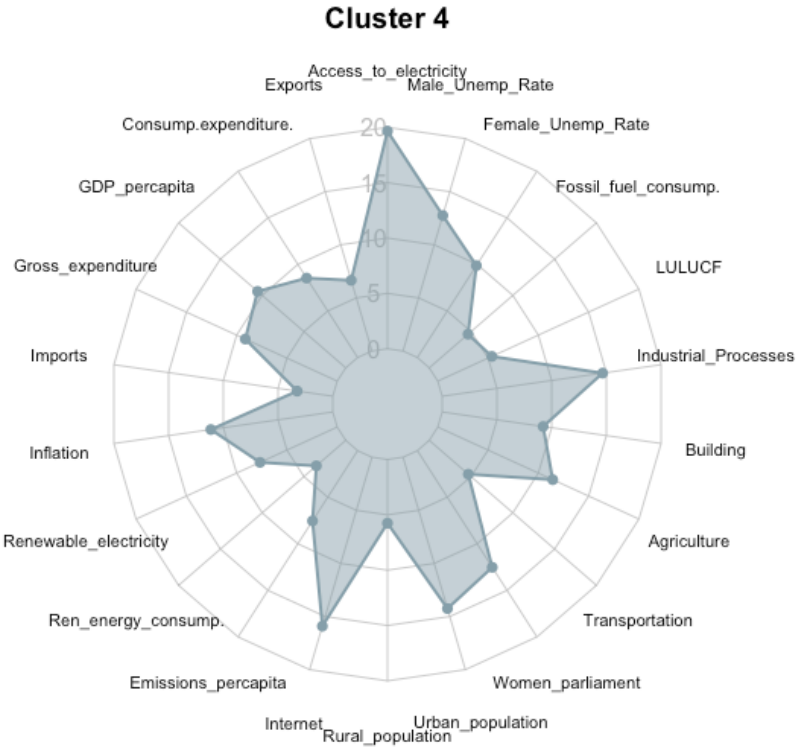

Figure 11: Cluster 4 characteristics: Feature means

Cluster 4 consists of rich but low-polluting countries. They have average GHG emissions per capita of 9.4 tons and a GDP per capita of $ 26895 on average. Cluster 4 is the third Cluster by the amount of GHG emissions per capita. Compared with Cluster 2, the GDP per capita of Cluster 4 is 24 times higher on average, but the GHG emissions per capita are only 9 times higher. In comparison, on the other hand, with Cluster 3, the most polluting Cluster, the GDP per capita is less than half and the emissions are almost three times higher.

A particularity of this Cluster is the growth of emissions in the Industrial Processes Sector from 2001 to 2018, as shown in Figure 7. In contrast to the other countries, which experienced slight and gradual growth, industrial emissions in group 4 countries increased very rapidly and by large amounts. In the period from 2001 to 2018, they have more than doubled in size, from 18 to 39 Mt. This strong growth, however, was also not seen in per capita GHG emissions, which remained almost constant during those years. The reason is a low positive correlation between GHG emissions per capita and Industrial sector emissions, which is equal to 0.08.

Instead, the main causes for this high increase in industrial emissions might be a sharp increase in both GDP per capita and individuals using the Internet. Firstly, the

average GDP per capita in 2018 (32384.57 USD$) is more than 2 times of 2001 GDP per capita (15065.62 USD$). Secondly, individuals using the Internet in 2001 were 25% of the population, while in 2018 the share was 79%.

Thirdly, the countries in the fourth category do not have abundant renewable electricity and energy outputs. The total renewable electricity amounts to 21.6% while renewable energy consumption is 13% of the total energy consumption. So that their intensity is not the highest.

Cluster 4, finally, consists of the countries which have the highest share of women in parliament, which is more than 22%. The correlation between GHG emissions and Women in parliaments is very low but positive. Figure 11 shows that unemployment rates between women and men are also almost equal, precisely 8.4% and 7%, surpassing Clusters 2 and 3. This could account for the high level of emissions in Cluster 4, as it is highly socially and economically developed. More than 98% of the population has access to electricity, 57% to the Internet, and 74% lives in urban areas.

Therefore, it can be said that this Cluster is the best performing in terms of the balance between development and economic growth, but also climate and gender sustainability. In fact, there is a large presence of European countries in Cluster 4, such as Sweden, Denmark, Austria, Finland, Ireland, Italy, France, and Germany, which, according to the research results of Cucchiella et al. (2017) have the highest sustainability indices in Europe. These can therefore determine and have a strong impact on the above-mentioned characteristics of this Cluster.

Given the rapid growth of emissions in the Industrial Processes Sector, Cluster 4 countries should focus on improving industrial efficiency and adopting cleaner production technologies. Implementing energy-efficient measures, promoting circular economy practices, and investing in sustainable manufacturing processes can help minimize emissions while maintaining economic growth. Based on their characteristics, cluster 4 countries can continue to showcase their successful balance between economic development, environmental sustainability, and gender equality. By further promoting renewable energy, enhancing industrial efficiency, strengthening climate and gender policies, promoting sustainable consumption and production, and engaging in collaboration and knowledge sharing, Cluster 4 countries can lead the way towards a sustainable and inclusive future.

Generalizing the results and characteristics of each individual Cluster, it can be concluded that in terms of GHG emissions per capita, the order of Clusters from the

one with lower to higher GHG emission levels is as follows: Cluster 2, Cluster 1, Cluster 4, and Cluster 3.

In addition, the same order can be seen with regard to GDP per capita, Fossil fuel consumption per capita, Share of renewable electricity and energy, and share of individuals using the Internet. That is, the Cluster with the lowest GHG emissions coincides with the one which has the lowest level of GDP per capita, fossil fuel consumption per capita and share of individuals using the Internet, on the contrary, which has the highest percentage of renewable energy and electricity.

7. Limitations and Future Research

A significant limitation in this research is the unavailability of historical and current data for certain countries, as well as the absence of key social, economic, and environmental indicators. The original dataset contains numerous missing values, particularly for variables like renewable energy consumption, population Internet usage, and per capita fossil fuel consumption. Although a predictive algorithm, such as random forest, was employed to fill in the dataset, the absence of observed data hinders the reliability of the model. The volatility of input data introduces uncertainty, potentially compromising the accuracy of the results, as there may be discrepancies between the model's predicted information and the actual reality due to non-homogeneity.

Future research could explore the incorporation of additional variables into the model to enhance its comprehensiveness and inclusiveness. Notably, this thesis is limited by the absence of weather-related and territorial composition variables, which play crucial roles in comprehending the factors influencing a country's emission levels. For instance, including weather variables such as temperature, precipitation, or wind patterns could provide insights into the impact of climate conditions on emissions. Similarly, incorporating territorial composition variables, such as natural resource distribution, in addition to the included land use patterns and urbanization levels, would enable a more comprehensive understanding of a country's available resources, opportunities, and challenges in relation to emissions. By considering these variables, future research can achieve a more nuanced analysis and foster greater awareness, facilitating the development of more effective strategies to address emissions. Overall, considering these variables enhances the power of the model by capturing additional factors that directly or indirectly influence GHG emissions, resulting in more accurate projections and a better understanding of the underlying dynamics.

Another significant limitation of this research, as well as studies on sustainability and climate change causes and effects, is the paucity of literature concerning the economic and social dimensions of emissions. As revealed in the literature review, the interplay between sustainability, greenhouse gas (GHG) emissions, and a country's economic and social structure is profound. While strategies, especially those focused on energy efficiency, predominantly rely on technological advancements and the implementation of taxes, they must also be harmonized with a country's economic framework. Consequently, the lack of comprehensive data and inadequate data collection emerge as a pressing challenge. A cohesive and interconnected dataset that encompasses market dynamics, energy efficiency, social considerations, organizational

processes, systemic perspectives, and economic concepts is imperative to address the current knowledge gap.

As for possible future research, the first expansion is to include all countries of the world in the model, so that the analysis is comprehensive, and collaboration can proceed in which there are no country exclusions. This allows for a comprehensive view of the world.

Expanding the analysis to the present time highlights the dynamic nature of efforts towards mitigating the earth's surface temperature. In recent years, advancements in technologies, digitization, and the rapid dissemination of news and communication have brought about significant changes and improvements worldwide, starting in 2018 and continuing to the present year. Moreover, numerous governments have implemented measures and regulations aimed at promoting energy conservation and waste reduction in both private households and businesses. These developments have the potential to impact the results of Cluster Analysis and foster new collaborations. It is conceivable that countries, previously assigned to different clusters, may have grown more similar due to the adoption of shared practices and policies. Thus, considering the current context enables a more accurate understanding of the evolving dynamics and potential shifts in international cooperation towards addressing climate change and promoting sustainability.

It would be beneficial then to collect more current data for all countries around the world.

Lastly, it would be valuable to explore the application of various models of GMMs clustering or different clustering methods, such as k-means or hierarchical clustering, to the data. This comparative analysis would provide a deeper understanding of the relationships between variables and unveil additional causes and effects of GHG emissions. Furthermore, by comparing the results obtained from different clustering methods, it would be possible to validate and confirm the reliability of the model selected and presented in this thesis, should the outcomes align. This comprehensive approach would enhance the robustness of the findings and offer valuable insights into the complexities of the emission patterns and their underlying factors.

8. Conclusion

This paper examines the results of an unsupervised machine learning algorithm applied to create clusters of the worldwide countries' forward GHG emissions and socio-economic factors.

The following conclusions can be drawn.

Firstly, considering inter-clusters and within-clusters distances, the proposed GMMs clustering model can partition samples effectively. During the period between 2001 and 20018, the characteristics of the World's countries' GHG emissions and sustainable features can be divided into four categories.

The results of the analysis show that the most important indicators affecting the characteristics of countries' GHG emissions per capita are GDP per capita, fossil fuel consumption per capita, Exports and Imports, the share of individuals living in urban areas and using the Internet and renewable electricity and energy consumption. The main reasons for these appearances are that burning fossil fuels is one of the major causes of emission release and GDP per capita is the most positively correlated feature to GHG emissions per capita. In addition, global GHG emissions vary between developed and developing countries. In developed countries, adjustments to reduce pollution are more restrictive and costly than in developing countries, resulting in pollutants being moved to these locations. For this reason, exports and imports are relevant in the analysis.

On the contrary, in the classification, the role of the proportion of seats held by women in parliament is not apparent. In fact, from 2001 to 2018, the presence of women in national parliaments continued to grow, however, this growth has had no effect on GHG emission levels, since they have remained roughly constant. Thus, its impact on the countries' characteristics of GHG emission is not significant. In fact, there would be no significant change in clustering results if this indicator was removed.

Consumption and total gross expenditures are highly negatively correlated to GHG emissions per capita, precisely -0.46 and -0.44.

Furthermore, it is important to note that certain sectors, specifically Building, Transportation, and Land use, land use change, and forestry (LULUCF), have a more significant impact on greenhouse gas (GHG) emissions per capita compared to the Agriculture and Industrial sectors. The analysis reveals that clusters characterized by high emissions per capita tend to exhibit elevated levels of GHG emissions from Building and Transportation activities, while their LULUCF emissions remain relatively

low. Conversely, although both the Agriculture and Industrial sectors experience considerable changes in emissions, the corresponding fluctuations in emissions per capita remain comparatively low.

The characteristics of the four Clusters can be summarized as follows:

(1) Cluster 1 comprises a diverse group of countries from South America, Africa, Europe, and Asia. These countries are characterized as underdeveloped, with low GDP per capita, higher rates of unemployment, and lower GHG emissions per capita compared to other clusters. However, countries of Custer number 1 could be set to be similar to those of the second Cluster. They exhibit higher emissions from land use and forest activities. The Cluster shows limited growth in renewable energy adoption. Overall, Cluster 1 represents countries facing economic challenges, high reliance on land and forests, and a need for sustainable development strategies to address GHG emissions and promote economic growth.

(2) Cluster 2 consists of the least polluting, poorest, and predominantly rural countries. Their low levels of GHG emissions per capita can be attributed to factors such as limited industrialization, high rural population percentages, and low energy consumption. However, these countries also face issues of poverty, limited access to essential services, and weaker environmental regulations. Collaboration and support from the international community are crucial to help these countries transition to sustainable development paths, ensuring that their progress is inclusive and environmentally responsible. By addressing poverty, promoting renewable energy access, and facilitating technology transfer, Cluster 2 countries can work towards reducing their environmental impact while improving the well-being of their populations.

(3) Cluster 3 represents a group of countries characterized as the richest and most polluting. These countries exhibit the highest GHG emissions per capita, driven by their heavy reliance on fossil fuels and strong economic status. They have higher emissions in most sectors. Cluster 3 countries are primarily coastal, which impacts their import and export levels. Furthermore, Cluster 3 is socially developed, with lower unemployment rates and a high per cent of the population living in urban areas. Conversely, it has the lowest proportion of women holding seats in national parliaments. However, the road to decarbonization for the US is in the right direction. US renewable electricity surpassed coal in 2022 for the first time ever (We Don't Have Time, 2023). The Inflation Reduction Act greatly impacted the clean energy sector, with renewable energy like solar power, making progress across the country.

(4) Cluster 4 comprises wealthy countries with low levels of pollution. Cluster 4 countries demonstrate a balance between economic development, climate sustainability, and gender equality. Despite their high GDP per capita, their GHG emissions per capita remain lower than expected, suggesting a more efficient use of resources. These countries demonstrate high levels of prosperity while actively addressing climate change and promoting gender equality.

Based on the analysis of clustering characteristics and their causes in the GHG emissions of countries, the major policy recommendations on setting emission reduction targets for different countries are as follows. The targets can be set to vary from Cluster to Cluster, and countries belonging to the same Cluster can have the same target and measures. The target of the countries of cluster 4 could be set as lower than other Clusters' targets. The presence of sustainable practices and strong governance contributes to its positive performance in various sustainability indices. Their goal is to achieve and increasingly strive for emission neutrality and use more and more renewable resources. Cluster 2 countries can strive towards sustainable development, and improve living conditions for their populations, instead of concentrating on reducing GHG emissions. Cluster 3 countries should significantly reduce their GHG emissions, diversify their energy sources, promote sustainable development, and address gender inequality. Lastly, countries in cluster 1 can reduce dependence on land-based activities and create alternative employment opportunities or implement policies and practices that promote sustainable land use and forest management.

Collaboration with particularly developing countries, such as the ones in cluster 4, to share knowledge, technologies, and financial resources, would support clusters 1,2 and 3 countries in the transition towards low-carbon and sustainable development pathways. Furthermore, economic measures, such as carbon taxes and emissions trading schemes, could be used to incentivize the reduction of GHG emissions (Xiao et al., 2017).

In sum, the order of targets for reducing greenhouse gas (GHG) emissions per capita of the four clusters can be stated as follows: the third cluster has the highest priority for reducing GHG emissions per capita, followed by the fourth cluster. The first cluster has a lower priority compared to the third and fourth clusters, and the second cluster has the lowest priority among the four groups.

The final goal of this thesis is to point out the role of sustainability in each country from different perspectives, such as the environmental and energetic policy, and, at the same time, compare countries to support future strategic choices. This analysis is, therefore, useful as a decision-making tool.

References

Althor, G., Watson, J. E. M. & Fuller, R. A. (2016). Global Mismatch between Greenhouse Gas Emissions and the Burden of Climate Change, *Scientific Reports*, [e-journal] vol. 6, no. 1, Available Online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4742864/ [Accessed 20 March 2023]

Ao, X. & Wei, E. (2022). Review of Panel Data Clustering Methods and Applications, *IEEE Xplore*, Available Online: https://ieeexplore.ieee.org/document/9820005#:~:text=Cluster%20analysis%20with%20panel%20data%20can%20better%20reflect [Accessed 30 March 2023]

Arto, I. & Dietzenbacher, E. (2014). Drivers of the Growth in Global Greenhouse Gas Emissions, *Environmental Science & Technology*, vol. 48, no. 10, pp.5388–5394

Bishop, C. M. (2016). Pattern Recognition and Machine Learning, Springer

Çağlar, M. & Gürler, C. (2021). Sustainable Development Goals: A Cluster Analysis of Worldwide Countries, *Environment, Development and Sustainability*, vol. 24, no. 8593–8624

Climate Watch. (2022). | Greenhouse Gas (GHG) Emissions | Climate Watch, *Www.climatewatchdata.org*, Available Online: https://www.climatewatchdata.org/ghg-emissions?end_year=2019&start_year=1990 [Accessed 15 January 2022]

Cucchiella, F., D'Adamo, I., Gastaldi, M., Koh, S. L. & Rosa, P. (2017). A Comparison of Environmental and Energetic Performance of European Countries: A Sustainability Index, *Renewable and Sustainable Energy Reviews*, [e-journal] vol. 78, no. 1364-0321, pp.401–413, Available Online: https://www.sciencedirect.com/science/article/pii/S1364032117305804 [Accessed 14 April 2023]

European Commission. (2023). Quadro 2030 per Il Clima E L'energia, *Climate.ec.europa.eu*, Available Online: https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2030-climate-energy-framework_it [Accessed 14 April 2023]

Feng, T., Yang, Y., Xie, S., Dong, J. & Ding, L. (2017). Economic Drivers of Greenhouse Gas Emissions in China, Renewable and Sustainable Energy Reviews, vol. 78, no. 1364-0321, pp.996–1006

International Renewable Energy Agency. (2019). About IRENA, Available Online: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019/Apr/IRENA_Global_Energy_Transformation_2019.pdf [Accessed 14 May 2023]

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R, Springer

Klein, C. A. M. de, Pinares-Patino, C. & Waghorn, G. C. (2008). Greenhouse Gas Emissions., Environmental impacts of pasture-based farming, pp.1–32

KMO Function - RDocumentation. (2023). *Www.rdocumentation.org*, Available Online: https://www.rdocumentation.org/packages/EFAtools/versions/0.4.4/topics/KMO [Accessed 12 May 2023]

Levinson, A. & Taylor, M. S. (2008). UNMASKING the POLLUTION HAVEN EFFECT, International Economic Review, vol. 49, no. 1, pp.223–254

Li, W., Yang, G., Li, X., Sun, T. & Wang, J. (2019). Cluster Analysis of the Relationship between Carbon Dioxide Emissions and Economic Growth, *Journal of Cleaner Production*, vol. 225, no. 0959-6526, pp.459–471

Li, X., Hipel, K. W. & Dang, Y. (2015). An Improved Grey Relational Analysis Approach for Panel Data Clustering, *Expert Systems with Applications*, [e-journal] vol. 42, no. 23, pp.9105–9116, Available Online: https://www.sciencedirect.com/science/article/pii/S0957417415005266 [Accessed 30 March 2023]

Liu, D., Guo, X. & Xiao, B. (2019). What Causes Growth of Global Greenhouse Gas Emissions? Evidence from 40 Countries, *Science of The Total Environment*, [e-journal] vol. 661, no. ISSN 0048-9697, pp.750–766, Available Online: https://www.sciencedirect.com/science/article/abs/pii/S0048969719302165 [Accessed 20 January 2023]

Lou, X., Lin, Y. & Li, L. M. W. (2022). Predicting Priority of Environmental Protection over Economic Growth Using Macroeconomic and Individual-Level Predictors: Evidence from Machine Learning, *Journal of Environmental Psychology*, vol. 82, no. 0272-4944, p.101843

Lu, H. and Huang, S., 2011. Clustering panel data. In SIAM international workshop on data mining held in conjunction with the 2011 SIAM international conference on data mining (pp. 1-10).

Ma, N., Shum, W. Y., Han, T. & Lai, F. (2021). Can Machine Learning Be Applied to Carbon Emissions Analysis: An Application to the CO2 Emissions Analysis Using Gaussian Process Regression, *Frontiers in Energy Research*, vol. 9, no. 2296-598X

Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M. & Cavallaro, F. (2020). A Multi-Stage Method to Predict Carbon Dioxide Emissions Using Dimensionality Reduction, Clustering, and Machine Learning Techniques, *Journal of Cleaner Production*, vol. 275, no. 0959-6526, p.122942 [Accessed 28 April 2023]

Montiel, I. (2018). Pollution Havens, *The SAGE Encyclopedia of Business Ethics and Society*, vol. 929, p.2576

Our World in Data. (2021). Fossil Fuel Consumption per Capita, *Our World in Data*, Available Online: https://ourworldindata.org/grapher/fossil-fuels-per-capita?tab=table&time=2000..2018 [Accessed 21 January 2022]

PBL. (2015). Trends in Global CO2 Emissions: 2015 Report, *PBL Netherlands Environmental Assessment Agency*, Available Online: https://www.pbl.nl/en/publications/trends-in-global-co2-emissions-2015-report [Accessed 14 April 2023]

PBL Netherlands Environmental Assessment Agency. (2022). Trends in Global CO2 and Total Greenhouse Gas Emissions; 2021 Summary Report, *PBL Netherlands Environmental Assessment Agency*, Available Online: https://www.pbl.nl/en/publications/trends-in-global-co2-and-total-greenhouse-gas-emissions-2021-summary-report [Accessed 14 April 2023]

Rink, K. (2022). Best Practices for Visualizing Your Cluster Results, *Medium*, Available Online: https://towardsdatascience.com/best-practices-for-visualizing-your-cluster-results-20a3baac7426 [Accessed 31 March 2023]

Ritchie, Roser, Mispy & Ortiz-Ospina. (2015). Measuring Progress towards the Sustainable Development Goals - SDG Tracker, Our World in Data, Available Online: https://sdg-tracker.org [Accessed 31 March 2023]

Ritchie, Roser, Mispy & Ortiz-Ospina. (2018). Global Indicator Framework for the Sustainable Development Goals and Targets of the 2030 Agenda for Sustainable Development, Available Online: https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%20refinement_Eng.pdf [Accessed 17 April 2023]

Stekhoven, D. J. & Buhlmann, P. (2011). MissForest--Non-Parametric Missing Value Imputation for Mixed-Type Data, *Bioinformatics*, vol. 28, no. 1, pp.112–118

TrueCue. (2022). World Sustainability Dataset, *Www.kaggle.com*, Available Online: https://www.kaggle.com/datasets/truecue/worldsustainabilitydataset [Accessed 14 January 2023]

United Nations. (2023). Causes and Effects of Climate Change, *United Nations*, Available Online: https://www.un.org/en/climatechange/science/causes-effects-climate-change [Accessed 14 January 2023]

Wang, W. & Lu, Y. (2021). Application of Clustering Analysis of Panel Data in Economic and Social Research Based on R Software, *Academic Journal of Business & Management*, vol. 3, no. 10

We Don't Have Time. (2023). Renewables Surpass Coal in Record Year for Solar Energy, *We don't have time*, Available Online: https://app.wedonthavetime.org/posts/cc33d14d-541e-42a5-bdbe-201f0bd9f345?utm_source=linkedin&utm_medium=social&utm_campaign=palmetto-renewablescoal [Accessed 3 May 2023]

Xu, Y., Li, X., Yuan, P. & Zhang, Y. (2023). Trade-off between Environment and Economy: The Relationship between Carbon and Inflation, *Frontiers in Environmental Science*, vol. 11, no. 10.3389

Yu, S., Wei, Y.-M., Fan, J., Zhang, X. & Wang, K. (2012). Exploring the Regional Characteristics of Inter-Provincial CO2 Emissions in China: An Improved Fuzzy Clustering Analysis Based on Particle Swarm Optimization, *Applied Energy*, vol. 92, no. 0306-2619, pp.552–562

Zelei, X., Bangyi, L. & Sifeng, L. (2009). The Discussion on the Clustering Way Based on the Multi-dimensional Panel Data and Empirical Analysis. *Application Statistics and Management,* 28(05), 831- 838.

Appendices

Appendix A: R packages

For the imputation of missing values, MissForest package has been installed and the function *missForest()* was used, particularly it can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations. It yields an out-of-bag (OOB) imputation error estimate (Stekhoven & Buhlmann, 2011).

The function *princomp()* from the "stats" package was implemented for PCA analysis.

The following packages were needed while implementing the Gaussian Mixture Models algorithm: *cluster, ClusterR, factoextra.*

For radar plots specifically, *fmsb* package was used.

Appendix B: Descriptive statistics

Figure B shows the correlation matrix between all the features included in the model.

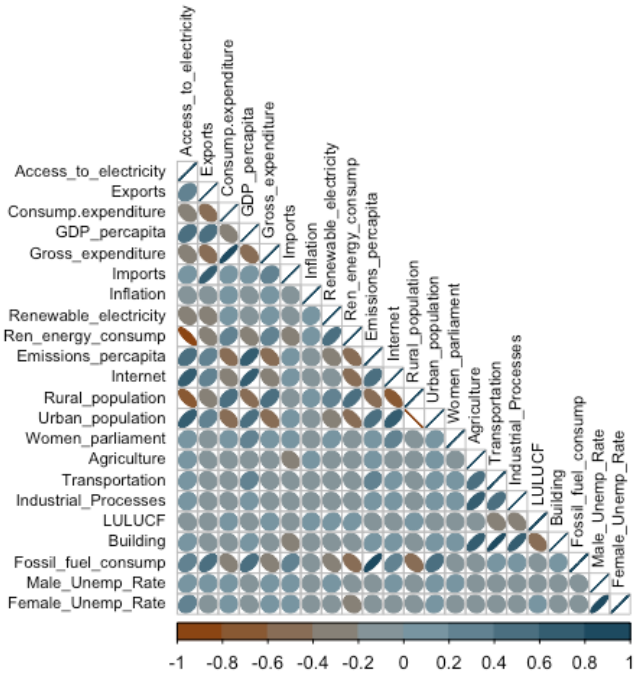

Figure B: Correlation matrix

Appendix C: Missing Values

In Appendix C the proportion of missing values presented in the original dataset before the implementation of the random forest procedure is illustrated. It also includes the setting of the missForest function for the imputation of missing values.

Fraction of features missing values

	%
Access_to_electricity	0.06
Exports	0.08
Consump.expenditure	0.10
GDP_percapita	0.04
Gross_expenditure	0.10
Imports	0.08
Inflation	0.09
Renewable_electricity	0.19
Ren_energy_consump	0.03
Emissions_percapita	0.09
Internet	0.12
Rural_population	0.04
Urban_population	0.09
Women_parliament	0.08
Agriculture	0.02
Transportation	0.03
Industrial_Processes	0.02
LULUCF	0.02
Building	0.03
Fossil_fuel_consump	0.56
Male_Unemp_Rate	0.05
Female_Unemp_Rate	0.05

Figure C: Proportion of missing values in the dataset for each feature

In the missForest function, the maximum number of iterations to be performed given the stopping criterion is not met beforehand was set equal to 10 and the number of trees to grow in each forest equal to 100.

Appendix D: PCA results

Figure D.1 shows the Kaiser-Meyer-Olkin (KMO) index. It is a measure of Sampling Adequacy (MSA) of factor analytic data matrices (Kaiser, 1970). The formula is based on the correlation between the features in question and the partial correlations. The index was calculated before running Principal Component Analysis, to test the adequacy of the dataset. The overall MSA in the dataset is equal to 71%, which indicates that the data is suitable for analysis.

Measure of sampling adequacy

	MSA for each item %
Access_to_electricity	79.70
Exports	55.56
Consumption_expenditure	92.10
GDP_percapita	88.87
Gross_expenditure	55.59
Imports	38.55
Inflation	75.50
Renewable_electricity	65.90
Renewable_energy_consumption	78.14
Emissions_percapita	83.68
Internet	87.11
Rural_population	78.13
Urban_population	78.13
Women_parliament	63.32
Agriculture	79.74
Transportation	64.25
Industrial_Processes	68.15
Land_Use_Change_Forestry	66.69
Building	63.84
Fossil.fuels.per.capita..kWh.	79.03
Unemployment_rate_male	51.85
Unemployment_rate_female	52.71

Figure D.1: Kaiser-Meyer-Olkin (KMO) index for each feature

Figure D.2 shows the total variance explained by each principal component of PCA.

The x-axis displays the principal component, and the y-axis indicates the percentage of total variance explained by each individual principal component. It is possible to see that the first principal component explains 28.4% of the total variation in the dataset.

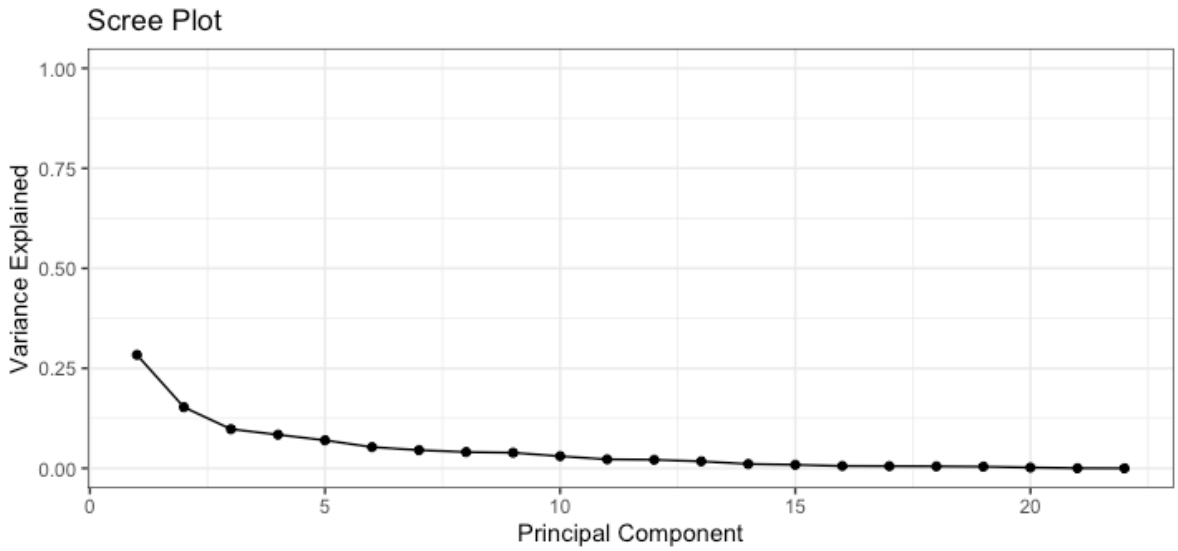

Figure D.2: Screeplot of PCA

Appendix E: GMMs Clustering summary

Table E: Model summary

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EVI (diagonal, equal volume, varying shape) model with 4
components:

 log-likelihood   n  df       BIC       ICL
      -4058.184 174 144 -8859.272 -8859.684

Clustering table:
 1  2  3  4
69 50  7 48
```

Appendix F: Map of Clusters

This section represents a zoom-in for each Cluster's geographical distribution.

Figure F.1: Cluster 1 map

Figure F.2: Cluster 2 map

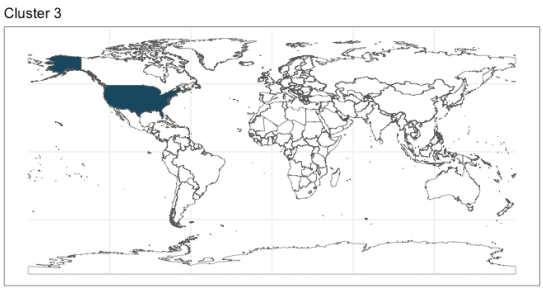

Figure F.3: Cluster 3 map

Figure F.4: Cluster 4 map